BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 3 2012, pages 366–372
doi:10.1093/bioinformatics/btr658

Systems biology Advance Access publication November 29, 2011

Gaussian process modelling for *bicoid* mRNA regulation in spatio-temporal Bicoid profile

Wei Liu and Mahesan Niranjan*

School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Bicoid protein molecules, translated from maternally provided *bicoid* mRNA, establish a concentration gradient in *Drosophila* early embryonic development. There is experimental evidence that the synthesis and subsequent destruction of this protein is regulated at source by precise control of the stability of the maternal mRNA. Can we infer the driving function at the source from noisy observations of the spatio-temporal protein profile? We use non-parametric Gaussian process regression for modelling the propagation of Bicoid in the embryo and infer aspects of source regulation as a posterior function.

Results: With synthetic data from a 1D diffusion model with a source simulated to model mRNA stability regulation, our results establish that the Gaussian process method can accurately infer the driving function and capture the spatio-temporal dynamics of embryonic Bicoid propagation. On real data from the `FlyEx` database, too, the reconstructed source function is indicative of stability regulation, but is temporally smoother than what we expected, partly due to the fact that the dataset is only partially observed. To be in line with recent thinking on the subject, we also analyse this model with a spatial gradient of maternal mRNA, rather than being fixed at only the anterior pole.

Contact: m.niranjan@southampton.ac.uk

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

Received on June 10, 2011; revised on November 4, 2011; accepted on November 25, 2011

1 INTRODUCTION

Patterning, during embryonic development, is thought to be regulated by a class of molecules known as morphogens, which propagate spatially and establish concentration gradients. As early as 1952, Turing (1952) hypothesized that a reaction-diffusion mechanism might form the basis of such differential concentrations. Wolpert (1969) further developed this view as a universal mechanism of spatial pattern formation arising from position-dependent cell fate. Subsequently, Driever and Nüsslein-Volhard (1988a, b) discovered the role of the Bicoid morphogen in early embryonic development of the fruit fly *Drosophila melanogaster*. Bicoid sets up a concentration gradient along the anterior-posterior (A–P) axis of the embryo and regulates several downstream gap genes including Hunchback and Krüppel, which are implicated in segmentation along the A–P axis (Driever and Nüsslein-Volhard, 1989; Ephrussi and Johnston, 2004; Houchmandzadeh *et al.*, 2002; Johnston *et al.*, 1989; Struhl *et al.*, 1989). The *bicoid* mRNA is maternal, deposited at the anterior pole of the embryo and translation of this mRNA is thought to begin immediately after egg deposition (Driever and Nüsslein-Volhard, 1989) with the resulting protein propagating towards the posterior end. Most of the computational and experimental work on the Bicoid morphogen is concerned with the establishment of the steady state concentration gradient, and the sensitivity with which a threshold on it may be sensed for downstream gene expression (Gregor *et al.*, 2005, 2007a, b; Hecht *et al.*, 2009; He *et al.*, 2010; Holloway *et al.*, 2006; Houchmandzadeh *et al.*, 2002; Löhr *et al.*, 2010). Systematic computational modelling of downstream gap gene circuits, and how their non-linear interactions result in segment boundary formation have also been published in Ashyraliyev *et al.* (2009); Jaeger *et al.* (2004); Reinitz and Sharp (1995), and include steady state exponential Bicoid expression profiles as a regulating input. In an alternate view of the dynamics of this system, Bergmann *et al.* (2007) suggested that much of the desirable decoding properties of the steady state profile can also be realized during the pre-steady-state stages.

A particularly novel insight into the process of Bicoid translation comes from the experimental work of Surdej and Jacobs-Lorena (1998). These authors suggested that the stability of the maternal mRNA may be systematically regulated; i.e. kept stable for a period of time during which mRNA is translated and morphogen synthesized, and subsequently rapidly killed off by some active processes. This observation, of course, matches our natural expectation as there is no need for the organism to continue to produce Bicoid protein beyond the point in time when it is decoded. However, to our surprise, modelling literature over the 30 years since the discovery of Bicoid ignore this possibility and assume a constant production rate at the anterior pole. We showed recently (Liu and Niranjan, 2009, 2011) that it is possible to model Bicoid production in a manner similar to Surdej and Jacobs-Lorena (1998)'s experimental findings and computationally extract the time at which mRNA decay begins, and the rate at which it is killed off, to match data measured on real fly embryos and archived in the `FlyEx` database (Pisarev *et al.*, 2009). We used an explicit model of mRNA stability regulation and a least squares fitting procedure between model output and observed data.

Bayesian inference has been shown to be useful in a range of applications including systems biology and bioinformatics. Successful examples range from the identification of gene regulatory networks by dynamic Bayesian networks

*To whom correspondence should be addressed.

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

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

(Husmeier, 2003; Peer *et al.*, 2001; Perrin *et al.*, 2003), network inference using informative priors (Mukherjee and Speed, 2008), inference of transcription regulation using a state space model (Sanguinetti *et al.*, 2006a, b), approximate inference using variational methods for the spatio-temporal Bicoid system (Dewar *et al.*, 2010) and parameter estimation using Monte Carlo simulations to understand stochastic dynamics of bacterial gene regulation (Wilkinson, 2011).

In this article, we build on these algorithmic foundations and apply the non-parametric probabilistic approach of Gaussian Process (GP) regression (Rasmussen and Williams, 2006) to address the problem of modelling Bicoid morphogen propagation. The GP model is the tool of choice for regression problems characterized by the need to model uncertainties and deal with unobserved data in a systematic way, both of which are aspects of our problem. The approach has been demonstrated in a wide range of applications in the machine learning literature. Specifically, Gao *et al.* (2008); Lawrence *et al.* (2007) and Honkela *et al.* (2010) studied interesting problems in biological modelling with GPs. The pioneering work on this is due to Lawrence *et al.* (2007) in which GPs were shown to be very effective in inferring target genes regulated by the tumour suppression transcription factor p53, providing an efficient alternative to the Monte Carlo approach advanced earlier by Barenco *et al.* (2006). Later publications have further confirmed the validity of the approach on other datasets of gene expression time series.

Apart from Dewar *et al.* (2010), the work on Bayesian methods mentioned above addressed purely temporal phenomena. Our work, on the other hand, is on a spatio-temporal problem and differs from Dewar *et al.* (2010) in the sense that we are interested in inferring the source driving function that corresponds to the stability regulation observed by Surdej and Jacobs-Lorena (1998). Alvarez *et al.* (2009) describe a similar attempt at combining a data-driven model (GP) with a latent process explaining the known physics of a system. They consider the heat equation, which is a simplified form of a diffusion system, to model the spatio-temporal profile of pollution. In this work, we derive the computational strategy needed to extend Lawrence *et al.* (2007)'s work to deal with spatio-temporal problems, and demonstrate its application to Bicoid regulation modelling using both synthetic data and real dataset from `FlyEx` database (Pisarev *et al.*, 2009). The results demonstrate the power of the method on synthetic and its limitations on real data. As such, ours is the first contribution that adapts the powerful algorithmic setting of non-parametric regression to tackle an important spatio-temporal inference problem in developmental biology.

2 METHODS

2.1 Bicoid spatio-temporal reaction-diffusion system

The partial differential equation, governing diffusion of Bicoid along a 1D anterior–posterior axis, we start from:

$$\frac{\partial}{\partial t}m(x,t)=D\frac{\partial^2}{\partial x^2}m(x,t)-\tau_p^{-1}m(x,t)+S_0f(t), \tag{1}$$

where $m(x,t)$ is the morphogen concentration as a spatio-temporal function, D, the diffusion constant and τ_p, the half-life of the morphogen protein. $f(t)$, the source, is the mRNA regulation function which we consider unknown and place a prior distribution over. A linear gain term S_0 is included to allow scaling of the data to match observations. There is an implicit assumption of a one to one mapping between mRNA regulation and the corresponding protein production, which we believe is justified as there is no evidence available of either post-transcriptional or post-translational regulation of Bicoid in this developmental system.

We consider the embryo as consisting of N cubes along the A–P axis in order to derive a linear dynamical system model from the continuous spatial diffusion equation. The basic idea for this stems from the work of Erban *et al.* (2007) (see later). With this discretization into N cubes, the chemical reactions involved in the diffusion process are as follows:

$$\text{Bicoid}_i \underset{d}{\overset{d}{\rightleftharpoons}} \text{Bicoid}_{i+1}, \quad \texttt{for} \quad i=1,2,\ldots,N-1 \tag{2}$$

$$\text{Bicoid}_i \xrightarrow{\tau_p^{-1}} \emptyset, \quad \texttt{for} \quad i=1,2,\ldots,N \tag{3}$$

$$\emptyset \xrightarrow{S_0 f(t)} \text{Bicoid}_1. \tag{4}$$

The first of these reactions, Equation (2), is Bicoid protein diffusion between neighbouring subvolumes with rate constant d, where d is given by $d=D/h^2$ and h is length of each cube. The second process in Equation (3) describes Bicoid protein degradation. Finally, Equation (4) is the translation of Bicoid proteins from the maternal mRNA with $f(t)$ being the latent function that needs to be inferred.

2.2 Linear dynamical system for Bicoid profile

We implemented a model in which source production occurs in a smaller first cube with length h_f ($5\mu m$—the length of a nucleus) and the other $N-1$ cubes ($h=10\mu m$) equally splitting the remaining A–P axis. Therefore, the rate constants are different between the first cube (d_f), where mRNA is produced and its stability regulated, and the other cubes (d):

$$d_f=D/(h_f h), \tag{5}$$

$$d=D/h^2. \tag{6}$$

In order to develop a linear dynamical system for Bicoid profile, following Erban *et al.* (2007), we rewrite the partial differential equation in Equation (1) as a system of ordinary differential equations for the morphogen concentration in each bin ($i=1,\ldots,N$):

$$\frac{\partial}{\partial t}m_1(t)=d_f(m_2(t)-m_1(t))-\tau_p^{-1}m_1(t)+S_0f(t), \tag{7}$$

$$\frac{\partial}{\partial t}m_i(t)=d(m_{i+1}(t)+m_{i-1}(t)-2m_i(t))-\tau_p^{-1}m_i(t), \tag{8}$$

$$\frac{\partial}{\partial t}m_N(t)=d(m_{N-1}(t)-m_N(t))-\tau_p^{-1}m_N(t). \tag{9}$$

Defining $\mathbf{m}(t)=[m_1(t),\ldots,m_N(t)]^T$, the linear dynamical system for Bicoid reaction-diffusion system is then vectorized as:

$$\frac{\partial \mathbf{m}(t)}{\partial t}=\mathbf{A}\mathbf{m}(t)+\mathbf{s}f(t), \tag{10}$$

where the spatial transition matrix $\mathbf{A}$ ($N\times N$) is defined by:

$$\begin{bmatrix} -(d_f+\tau_p^{-1}) & d_f & 0 & \cdots & 0 & 0 \\ d & -(2d+\tau_p^{-1}) & d & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & d & -(d+\tau_p^{-1}) \end{bmatrix}$$

and source production rate $\mathbf{s}=[S_0,0,\ldots,0]^T$.

The solution to Equation (10), $\mathbf{m}(t)$, giving the Bicoid spatio-temporal profile, is in terms of a matrix exponential and is given by:

$$\mathbf{m}(t)=\exp(t\mathbf{A})\mathbf{m}(0)+\int_0^t \exp((t-u)\mathbf{A})\mathbf{s}f(u)du, \tag{11}$$

where $\mathbf{m}(0)$ is zero at the beginning of the embryo development.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

2.3 Gaussian process modelling

We treat *bicoid* mRNA as a function drawn from a GP and extend it to the spatio-temporal application of Bicoid dynamical system. The GP prior for the latent mRNA regulation is defined by mean and covariance functions:

$$f(t)\sim N(0,k_{f,f}(t,t')), \tag{12}$$

where $k_{f,f}(t,t')$ is given by squared exponential covariance function with the length scale l:

$$k_{f,f}(t,t')=\exp\left(-\frac{(t-t')^2}{l^2}\right). \tag{13}$$

Because the right-hand side of Equation (11) is linear, as noted by Lawrence *et al.* (2007) in the content of modelling transcription regulation, $\mathbf{m}(t)$ turns out to be a multivariate function drawn from a GP:

$$\mathbf{m}(t)\sim N(\exp(t\mathbf{A})\mathbf{m}(0),\mathbf{K}_{\mathbf{m},\mathbf{m}}(t,t')). \tag{14}$$

The corresponding cross covariance function $\mathbf{K}_{\mathbf{m},\mathbf{m}}(t,t')$ is given by:

$$\mathbf{K}_{\mathbf{m},\mathbf{m}}(t,t')= \tag{15}$$

$$\int_0^t\int_0^{t'}\exp((t-u)\mathbf{A})\mathbf{s}(\exp((t'-u')\mathbf{A})\mathbf{s})^T k_{f,f}(u,u')dudu'.$$

The cross covariance function between $\mathbf{m}(t)$ and $f(t)$ becomes

$$\mathbf{k}_{\mathbf{m},f}(t,t')=\int_0^t\exp((t-u)\mathbf{A})\mathbf{s}k_{f,f}(u,t')du. \tag{16}$$

These expressions can be derived analytically and the derivation details are shown in Supplementary Material.

2.4 Predictive distribution

Let $\mathbf{f}_*$ be a vector of J_* values of the source function at equally spaced time points, and $\mathbf{m}_i^*$ be the corresponding protein profiles at these instances in time, within the i-th cube along the A–P axis. Concatenating these, we define $\mathbf{h}_*=[\mathbf{f}_*,\mathbf{m}_1^*,...,\mathbf{m}_N^*]^T$ corresponding to J_*+NJ_* test points. The corresponding training data consists of the morphogen values in the N cubes, taken at J points in time, and a fixed source value f, contained in the vector $\mathbf{h}=[f,\mathbf{m}_1,...,\mathbf{m}_N]^T$ of dimension $1+NJ$.

With the above notation, the mean and covariance of the posterior distribution are given by:

$$\bar{\mathbf{h}}_*^{\text{post}}=\mathbf{K}_{\mathbf{h}_*,\mathbf{h}}\mathbf{K}_{\mathbf{h},\mathbf{h}}^{-1}\mathbf{y}, \tag{17}$$

$$\mathbf{K}_{\mathbf{h}_*,\mathbf{h}_*}^{\text{post}}=\mathbf{K}_{\mathbf{h}_*,\mathbf{h}_*}-\mathbf{K}_{\mathbf{h}_*,\mathbf{h}}\mathbf{K}_{\mathbf{h},\mathbf{h}}^{-1}\mathbf{K}_{\mathbf{h},\mathbf{h}_*}. \tag{18}$$

Each covariance matrix in Equations (17) and (18) is partitioned across the source and N Bicoid intensities in different cubes. Illustrating this for $\mathbf{K}_{\mathbf{h}_*,\mathbf{h}}$,

$$\mathbf{K}_{\mathbf{h}_*,\mathbf{h}}=\begin{bmatrix}\mathbf{K}_{\mathbf{f}_*,f} & \mathbf{K}_{\mathbf{f}_*,\mathbf{m}_1} & \mathbf{K}_{\mathbf{f}_*,\mathbf{m}_2} & \cdots & \mathbf{K}_{\mathbf{f}_*,\mathbf{m}_N}\\ \mathbf{K}_{\mathbf{m}_1^*,f} & \mathbf{K}_{\mathbf{m}_1^*,\mathbf{m}_1} & \mathbf{K}_{\mathbf{m}_1^*,\mathbf{m}_2} & \cdots & \mathbf{K}_{\mathbf{m}_1^*,\mathbf{m}_N}\\ \mathbf{K}_{\mathbf{m}_2^*,f} & \mathbf{K}_{\mathbf{m}_2^*,\mathbf{m}_1} & \mathbf{K}_{\mathbf{m}_2^*,\mathbf{m}_2} & \cdots & \mathbf{K}_{\mathbf{m}_2^*,\mathbf{m}_N}\\ \vdots & \vdots & \vdots & \ddots & \vdots\\ \mathbf{K}_{\mathbf{m}_N^*,f} & \mathbf{K}_{\mathbf{m}_N^*,\mathbf{m}_1} & \cdots & \cdots & \mathbf{K}_{\mathbf{m}_N^*,\mathbf{m}_N}\end{bmatrix}$$

Therefore, the dimensions of the covariance matrices $\mathbf{K}_{\mathbf{h}_*,\mathbf{h}}$, $\mathbf{K}_{\mathbf{h},\mathbf{h}}$ and $\mathbf{K}_{\mathbf{h}_*,\mathbf{h}_*}$ are $(J_*+NJ_*)\times(1+NJ)$, $(1+NJ)\times(1+NJ)$ and $(J_*+NJ_*)\times(J_*+NJ_*)$, respectively. The observations, collected in vector $\mathbf{y}$ of dimension $1+NJ$, are assumed to be corrupted by additive noise as

$$y_i(t)=m_i(t)+e_i(t), \tag{19}$$

where $e_i(t)$ are drawn from $N(0,\sigma_i^2(t))$. Hence,

$$\mathbf{K}_{\mathbf{y},\mathbf{y}}=\mathbf{K}_{\mathbf{h},\mathbf{h}}+\Sigma, \tag{20}$$

$$\Sigma=\text{diag}(\sigma_f^2,\sigma_{11}^2,\ldots,\sigma_{1J}^2,\ldots,\sigma_{N1}^2,\ldots,\sigma_{NJ}^2). \tag{21}$$

With the hyperparameter vector $\boldsymbol{\theta}_h=[l,\sigma_f^2,\sigma_{11}^2,\ldots,\sigma_{1J}^2,\ldots,\sigma_{N1}^2,\ldots,\sigma_{NJ}^2]$, the likelihood is:

$$p(\mathbf{y}|\boldsymbol{\theta}_h)=\int p(\mathbf{y}|\boldsymbol{\theta}_\mathbf{h},\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta}_h)d\mathbf{f}. \tag{22}$$

The observations are taken from `FlyEx` integrated data. To get an estimate of the noise levels and the length scale l, we maximized the above likelihood following Rasmussen and Williams (2006). Estimates of these hyperparameters are given in Supplementary Material. When simulating synthetic data, we added zero mean Gaussian noise of SD $\sigma=0.1$, and assumed it to be known.

For model parameter estimation, we evaluated least squared error between Bicoid ODE system [Equations (7)–(9)] output (m^{model}) and measured intensities from `FlyEx` database (m^{data}) to estimate the parameters: diffusion constant D ($3\mu m^2/s$) and protein half-life τ_p (86min) (Liu and Niranjan, 2011):

$$Ls=\sum_{j=C11}^{C14A.8}\sum_{i=1}^{N}\left(S_0 m_{i,j}^{\text{model}}-m_{i,j}^{\text{data}}\right)^2. \tag{23}$$

Note we have estimated the hyperparameters using maximum likelihood and set the model parameters by least squares fitting. Ideally, one would like to exploit the elegance of the GP framework and estimate all these by maximum likelihood. To achieve this, we need analytical expressions for the covariance matrices and their derivatives. We discuss this further in the Supplementary Material.

Since the production concentration S_0 is independent of D and τ_p and is linear of the model output, we differentiate the Equation (23) with respect to S_0 and set it to zero. The calculation for S_0 is given by:

$$S_0=\frac{\sum_{j=C11}^{C14A.8}\sum_{i=1}^{N}m_{i,j}^{\text{model}}m_{i,j}^{\text{data}}}{\sum_{j=C11}^{C14A.8}\sum_{i=1}^{N}(m_{i,j}^{\text{model}})^2}. \tag{24}$$

3 RESULTS AND DISCUSSION

3.1 Inference of mRNA regulation function

We first assume that *bicoid* mRNA is localized and its regulation occurs only in the anterior pole, the first cube in our discretized model. Therefore, source production amplitude vector is given by

$$\mathbf{s}=[S_0,0,\ldots,0]^T. \tag{25}$$

We examine the performance of our GP approach on two synthetic datasets (Figs 1 and 2) and an experimental dataset (Fig. 3).

As in our previous work (Liu and Niranjan, 2009), to generate synthetic data, we implemented mRNA stability regulation by the function

$$f(t)=\delta(x)\left(\Theta(t)-\Theta(t-t_0)\right) \tag{26}$$

$$+\delta(x)\Theta(t-t_0)\exp\left(-\frac{t-t_0}{\tau_m}\right),$$

which characterizes mRNA, and hence the production of Bicoid protein, to be stable and constant to time t_0, followed by an exponential decay of time constant τ_m. δ is the Kronecker delta function and Θ is the Heaviside step function. The parameter values taken from our previous work, estimated by a least squares fit between model output and `FlyEx` measurements, are $t_0=144$ min and $\tau_m=9$ min. The dashed lines in Figures 1B, 2B and 3B show the true source function according to our hypothesis on regulation [Equation (26)].

Figure 1A shows the synthetic training data using the ODE system of Equation (7)–(9) with additive noise over the entire developmental period of $0\sim200$ min, with 20 equally spaced time points and 51 cubes along the A–P axis. Figure 1B shows the estimated mRNA regulation function, $S_0 f(t)$, from the GP approach. We see that the GP is able to recover the regulated source function quite well, though the resulting estimate is smoother. This is to

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

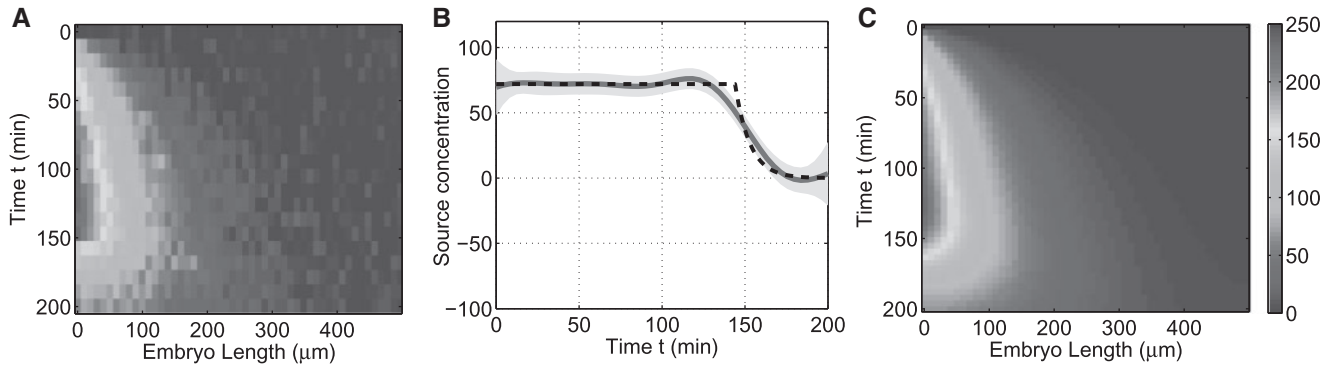

Fig. 1. Inferred *bicoid* mRNA regulation and spatio-temporal protein concentration from a synthetic dataset; time scale from 0 to 200 min in 51 cubes along A–P axis. (**A**) Training datasets from Bicoid reaction-diffusion ODE simulation with additive noise. (**B**) Inferred mRNA regulation function (solid line) with 95% confidence interval. Source function used in the simulation is shown with dashed line. (**C**) Posterior mean GP approximation of the spatio-temporal profile.

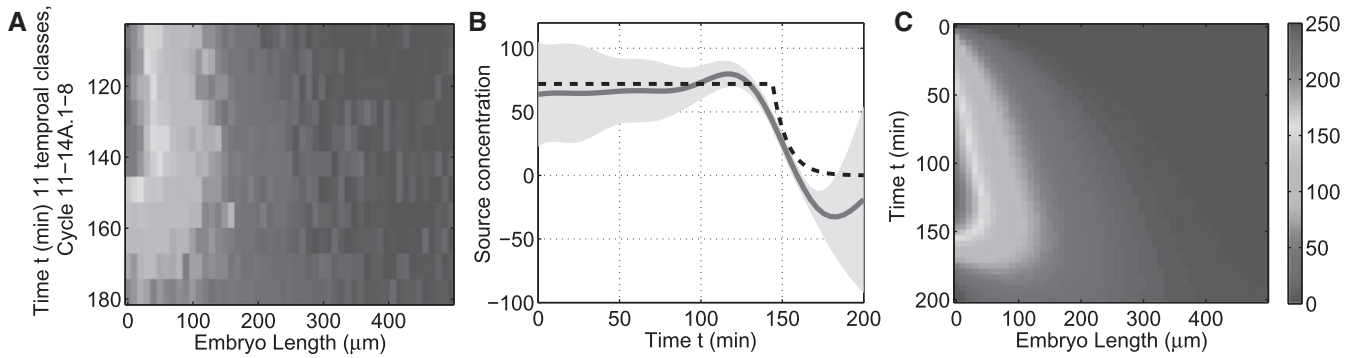

Fig. 2. Predicted results for *bicoid* mRNA and Bicoid spatio-temporal profile using only partial data (106–178 min): Cycles 11–13 and Cycle 14*A* class1–8. (**A**) partial data used in training. (**B** and **C**) Inferred source and spatio-temporal profiles, respectively.

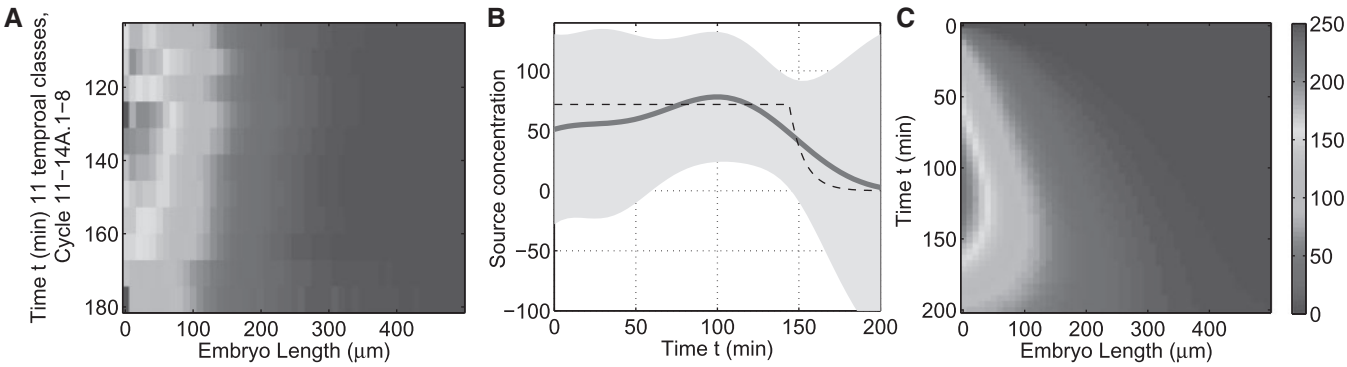

Fig. 3. Source function and Bicoid profile inferred from the `FlyEx` database. (**A**) Average profile of with time scale: Cycles 11–13 and Cycle 14*A* class 1–8. (**B**) Inferred source and (**C**) Bicoid concentration over the whole time scale (0–200 min). The dashed line in (B) is the assumed source function.

be expected from a GP model, for which a function with a sharp discontinuity will have very low likelihood in the prior. Still, the decay beyond 2 h is very rapid. The posterior mean of the inferred Bicoid concentration profile from the model is shown in Figure 1C. The temporal dynamics of a morphogen gradient being established and then killed off is clearly present in the model output.

In the above, shown in Figure 1, we have used the synthetic data over the full developmental time scale of interest. However, in the `FlyEx` dataset, we do not have measurements available over the whole timescale and the source has to be inferred from partial data, starting from 100 min. In order to simulate this situation with synthetic data, we ran our GP models with only the partial

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

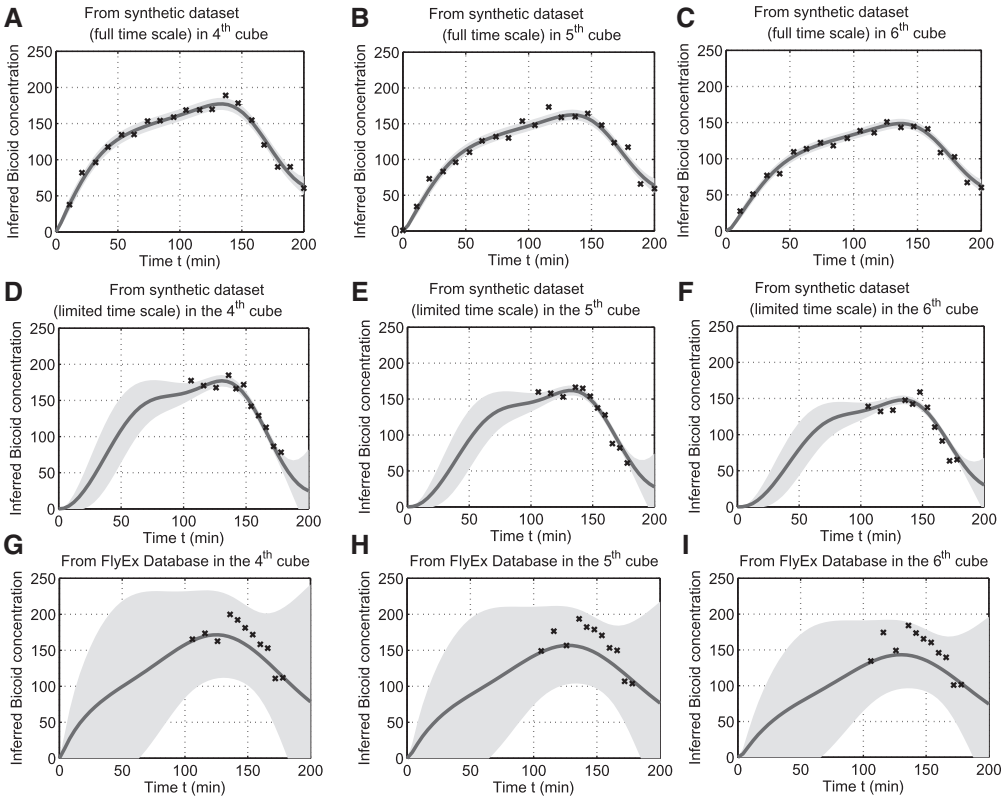

Fig. 4. Predicted temporal posterior distribution of Bicoid protein concentrations in individual cubes on different training datasets. The mean inference and 95% confidence intervals are shown with solid lines and grey area. The crosses represent three different training datasets shown in Figures 1–3. Panels (**A**–**C**) are inferred on synthetic dataset with full-time scale while the partial synthetic dataset is used in (**D**–**F**). Panels (**G**–**I**) are inferred on real dataset from FlyEx.

data, shown in Figure 2A as input. As expected, in Figure 2B, the credible interval is wider at the early stages where no data are present and narrow during 106 − 178 min. Still, the GP posterior of the morphogen profile captures the spatio-temporal dynamics well and contains the sharp post-peak decay.

Figure 3 shows the behaviour of the GP model on real data from FlyEx, with figure 3A, 3B and 3C showing the data, inferred source and model based spatio-temporal morphogen profile, respectively. We see that the reconstructed profile accurately reproduces the establishment of Bicoid gradient and subsequent decay. The inferred source function is smoother than our hypothesized model, but contains the basic elements of an approximately constant part and subsequent decay. The noisy nature of the data causes the resulting uncertainties to be very high. Noise levels inferred from a maximum likelihood setting [Equation (22)] is shown in Supplementary Figure S1. These are much higher than the variance of the additive noise we used to construct the synthetic data of Figures 1 and 2. Further, we note that the source of uncertainty in the data is not purely additive instrument noise. FlyEx measurements do not come from the observations on a single embryo. They are taken from populations of embryos, harvested at various stages of development. The effect of this is not modelled anywhere in our approach.

As noted, the GP inferred source functions are smoother than hypothesized by our model. A consequence of smooth functions fitting the data is also that the temporal point at which mRNA begins to decay starts earlier. The rapid change between mRNA being translated and killed off is not explicitly modelled in the GP approach. Such rapid changes may well be better modelled in a probabilistic framework that explicitly incorporates switching behaviour, such as the two-state Markov Jump process (also known as a telegraph process) considered in Sanguinetti *et al.* (2009).

Figure 4 shows the predicted temporal profiles of the Bicoid at three adjacent spatial points along the embryo.[1] The training data are also shown. The three rows in the figure correspond to cases illustrated in Figures 1–3. We see that at the level of the GP model generating the data, reasonably good fits are obtained. With real data, we see high uncertainties in regions where data are not present.

3.2 Non-localized maternal mRNA

Additionally, we show in Figure 5 cross sections of the spatio-temporal profiles, taken along the A–P axis at different developmental cycles. Here, we see that the exponential spatial decays of FlyEx measurements are faithfully captured by the GP model. We also observe that most of the mismatch between model output and measured data is towards the anterior part of the A–P axis. This mismatch motivates one to question the use of a highly localized point source as the input to the diffusion system.

[1]These cubes are chosen for illustration because these are locations where much of the variation is happening.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

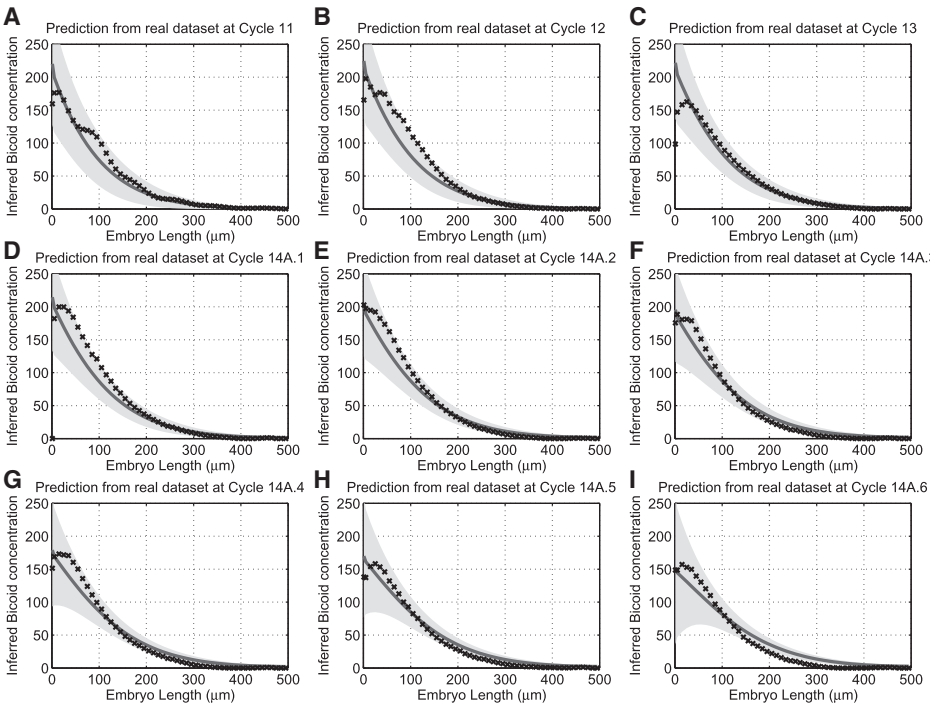

Fig. 5. Inferred posterior distribution of spatial Bicoid protein concentrations in the fixed time points from Cycle11 (**A**) to Cycle14A.6 (**I**) on the real dataset. The crosses show the observed protein spatial data in different developmental time points.

In the literature, Spirov *et al.* (2009) and Little *et al.* (2011) have discussed the scenario in which maternal *bicoid* mRNA itself has a spatial gradient (i.e. non-localized). While Spirov *et al.* (2009) considered this mRNA spatial profile to be the main determinant of morphogen gradient, Little *et al.* (2011) concluded that $>90\%$ of total *bicoid* mRNA is within the anterior 20% EL, and this alone is insufficient to establish the required spatial gradient of the protein in the time available over the length scale of interest. We also simulated this possibility in our GP models, with maternal mRNA being spatially distributed in the first 10 of the 50 cubes ($h=10\mu m$) with an initial exponentially decaying spatial profile. The corresponding results of the inferred mean source (now a spatio-temporal profile) and the GP mean morphogen profile are shown in Supplementary Figure S4B and C. We note that in this case, the onset of mRNA decay begins slightly earlier than for the point source in the first bin, which is to be expected since the mRNA spatial distribution contributes to the generation of morphogen upto 20%EL, and the destruction has to start earlier to compensate. As seen in Supplementary Figure S5, the fit to the data does improve with spatially distributed mRNA. We include this for completeness, showing that a GP model can be applied in a flexible way in this manner, but do not think that the results can resolve the differences discussed by Spirov *et al.* (2009) and Little *et al.* (2011).

4 CONCLUSION

In this contribution, we have shown that the non-parametric GP regression model can be effectively applied to the problem of inferring biologically useful information from the spatio-temporal distribution of the Bicoid morphogen in early Drosophila embryogenesis. Discretization of the spatial domain transforms the spatio-temporal problem into a dynamical system for which, with a GP prior imposed on the source, the solution can be obtained as a matrix exponential. With synthetic data obtained from a linear spatio-temporal dynamical system, our results show that the GP approach is able to accurately recover the driving input and model the Bicoid distribution. On real world data, our results also estimate a smoothed version of the driving input due to the data being available only during part of the developmental process, and yet the part of the source decay is fairly well estimated.

Conflict of Interest: none declared.

REFERENCES

Alvarez,M. *et al.* (2009) Latent force models. In van Dyk,D. and Welling,M. (eds) *Proceedings of The Twelfth International Conference on Artificial Intelligence and Statistics*. Clearwater Beach, Florida, USA. Vol. 5 of Journal of Machine Learning Research: Workshop and Conference Proceedings, pp. 9–16.

Ashyraliyev,M. *et al.* (2009) Gene circuit analysis of the terminal gap gene huckebein. *PLoS Comput. Biol.*, **5**, e1000548.

Barenco,M. *et al.* (2006) Ranked prediction of p53 targets using hidden variable dynamic modeling. *Genome Biol.*, **7**, R25.

Bergmann,S. *et al.* (2007) Pre-steady-state decoding of the Bicoid morphogen gradient. *PLoS Biol.*, **5**, 965–991.

Dewar,M.A. *et al.* (2010) Parameter estimation and inference for stochastic reaction-diffusion systems: application to morphogenesis in D. melanogaster. *BMC Syst. Biol.*, **4**, 21.

Driever,W. and Nüsslein-Volhard,C. (1988a) A gradient of bicoid protein in Drosophila embryos. *Cell*, **54**, 83–93.

Driever,W. and Nüsslein-Volhard,C. (1988b) The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. *Cell*, **54**, 95–104.

Driever,W. and Nüsslein-Volhard,C. (1989) The bicoid protein is a positive regulator of hunchback transcription in the early Drosophila embryo. *Nature*, **337**, 138–143.

Ephrussi,A. and Johnston,D.S. (2004) Seeing is believing: the Bicoid morphogen gradient matures. *Cell*, **116**, 143–152.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016

Erban,R. *et al.* (2007) A practical guide to stochastic simulations of reaction-diffusion processes. *Arxiv preprint arXiv:0704.1908*.

Gao,P. *et al.* (2008) Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities. *Bioinformatics*, **24**, i70–i75.

Gregor,T. *et al.* (2005) Diffusion and scaling during early embryonic pattern formation. *Proc. Natl Acad. Sci. USA*, **102**, 18403–18407.

Gregor,T. *et al.* (2007a) Probing the limits to positional information. *Cell*, **130**, 153–164.

Gregor,T. (2007b) Stability and nuclear dynamics of the Bicoid morphogen gradient. *Cell*, **130**, 141–152.

Hecht,I. *et al.* (2009) Determining the scale of the Bicoid morphogen gradient. *Proc. Natl Acad. Sci. USA*, **106**, 1710–1715.

He,F. *et al.* (2010) Shaping a morphogen gradient for positional precision. *Biophys. J.*, **99**, 697–707.

Holloway,D.M. *et al.* (2006) Analysis of pattern precision shows that Drosophila segmentation develops substantial independence from gradients of maternal gene products. *Dev. Dyn.*, **235**, 2949–2960.

Honkela,A. *et al.* (2010) Model-based method for transcription factor target identification with limited data. *Proc. Natl Acad. Sci. USA*, **107**, 7793–7798.

Houchmandzadeh,B. *et al.* (2002) Establishment of developmental precision and proportions in the early Drosophila embryo. *Nature*, **415**, 798–802.

Husmeier,D. (2003) Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, **19**, 2271–2282.

Jaeger,J. *et al.* (2004) Dynamic control of positional information in the early Drosophila embryo. *Nature*, **430**, 368–371.

Johnston,D. *et al.* (1989) Multiple steps in the localization of bicoid RNA to the anterior pole of the Drosophila oocyte. *Development*, **107**, 13–19.

Lawrence,N.D. *et al.* (2007) Modelling transcriptional regulation using Gaussian processes. In Schölkopf,B. *et al.* (eds) *Advances in Neural Information Processing Systems*, vol. 19. MIT Press, Cambridge, MA, pp. 785–792.

Little,S.C. *et al.* (2011) The formation of the Bicoid morphogen gradient requires protein movement from anteriorly localized mRNA. *PLoS Biol.*, **9**, e1000596.

Liu,W. and Niranjan,M. (2009) Matching models to data in modelling morphogen diffusion. In Džeroski,S. *et al.* (eds) *Proceedings of The Third International Workshop on Machine Learning in Systems Biology*. Helsinki University Printing House, Finland, pp. 55–64.

Liu,W. and Niranjan,M. (2011) The role of regulated mRNA stability in establishing Bicoid morphogen gradient in Drosophila embryonic development. *PLoS One*, **6**, e24896.

Löhr,U. *et al.* (2010) Bicoid - morphogen function revisited. *Fly*, **4**, 236–240.

Mukherjee,S. and Speed,T.P. (2008) Network inference using informative priors. *Proc. Natl Acad. Sci. USA*, **105**, 14313–14318.

Peer,D. *et al.* (2001) Inferring subnetworks from perturbed expression profiles. *Bioinformatics*, **17**(Suppl. 1), S215–S224.

Perrin,B.E. *et al.* (2003) Gene networks inference using dynamic Bayesian networks. *Bioinformatics*, **19**(Suppl. 2), ii138–ii148.

Pisarev,A. *et al.* (2009) FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic Acids Res.*, **37**, D560–D566.

Rasmussen,C.E. and Williams,C.K.I. (2006) *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, MA.

Reinitz,J. and Sharp,D.H. (1995) Mechanism of eve stripe formation. *Mech. Dev.*, **49**, 133–158.

Sanguinetti,G. *et al.* (2006a) A probabilistic dynamical model for quantitative inference of the regulatory mechanism of transcription. *Bioinformatics*, **22**, 1753.

Sanguinetti,G. *et al.* (2006b) Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, **22**, 2775.

Sanguinetti,G. *et al.* (2009) Switching regulatory models of cellular stress response. *Bioinformatics*, **25**, 1280–1286.

Spirov,A. *et al.* (2009) Formation of the Bicoid morphogen gradient: an mRNA gradient dictates the protein gradient. *Development*, **136**, 605–614.

Struhl,G. *et al.* (1989) The gradient morphogen bicoid is a concentration-dependent transcriptional activator. *Cell*, **57**, 1259–1273.

Surdej,P. and Jacobs-Lorena,M. (1998) Developmental regulation of bicoid mRNA stability is mediated by the first 43 nucleotides of the 3' untranslated region. *Mol. Cell. Biol.*, **18**, 2892–2900.

Turing,A.M. (1952) The chemical basis of morphogenesis. *Philos. Trans. R. Soc. London Ser. B*, **237**, 37–72.

Wilkinson,D.J. (2011) Parameter inference for stochastic kinetic models of bacterial gene regulation: a Bayesian approach to systems biology (with discussion). In Bernardo,J. M. *et al.* (eds) *Bayesian Statistics*, vol. 9. Oxford University Press, UK, pp. 679–705.

Wolpert,L. (1969) Positional information and the spatial pattern of cellular differentiation. *J. Theor. Biol.*, **25**, 1–47.

Downloaded from http://bioinformatics.oxfordjournals.org/ at :: on August 30, 2016