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, UK

Associate Editor: Prof. 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 one dimensional 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.

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 (1988b,a) 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 (St Johnston *et al.*, 1989; Driever and Nüsslein-Volhard, 1989; Struhl *et al.*, 1989; Houchmandzadeh *et al.*, 2002; Ephrussi and Johnston, 2004). 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 (Houchmandzadeh *et al.*, 2002; Gregor *et al.*, 2005, 2007a,b; Holloway *et al.*, 2006; Hecht *et al.*, 2009; Löhr *et al.*, 2010; He *et al.*, 2010). Systematic computational modelling of downstream gap gene circuits, and how their nonlinear interactions result in segment boundary formation have also been published in (Reinitz and Sharp, 1995; Jaeger *et al.*, 2004; Ashyraliyev *et al.*, 2009), 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 synthesised, 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 (Peer *et al.*,...
network inference using informative priors (Mukherjee and Speed, 2008), inference of transcription regulation using a state space model (Sanguinetti et al., 2006a), 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 paper, 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, Lawrence et al. (2007); Gao et al. (2008) and Honkela et al. (2010) studied interesting problems in biological modelling with Gaussian Processes (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 tumor 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 one dimensional anterior-posterior axis, we start from is:

$$\frac{\partial m(x, t)}{\partial t} = D \frac{\partial^2 m(x, t)}{\partial x^2} - \tau_p^{-1} m(x, t) + S_0 f(t), \quad (1)$$

where $m(x, t)$ is the morphogen concentration as a spatio-temporal function, $D$, the diffusion constant and $\tau_p$, 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-translation 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:

$$Bicoid_d \xrightarrow{d} Bicoid_{d+1}, \quad \text{for } i = 1, 2, \ldots, N - 1 \quad (2)$$

$$Bicoid_d \xrightarrow{\tau_p^{-1}} \emptyset, \quad \text{for } i = 1, 2, \ldots, N \quad (3)$$

$$\emptyset \xrightarrow{S_0 f(t)} Bicoid_d. \quad (4)$$

The first of these reactions, (Equation (2)), is Bicoid protein diffusion between neighbouring sub-volumes 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 function 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 height $h_f$ ($5\mu m$ – 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 ($d_1$), where mRNA is produced and its stability regulated, and the other cubes ($d$):

$$d_1 = D/(h_f h), \quad d = D/h^2. \quad (5)$$

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 ($i = 1, \ldots, N$):

$$\frac{\partial m_1}{\partial t} = d_f (m_2(t) - m_1(t)) - \tau_p^{-1} m_1(t) + S_0 f(t), \quad (7)$$

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

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

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

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

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

$$\begin{pmatrix}
-d_f + \tau_p^{-1} & d_f & 0 & \cdots & 0 \\
0 & -d_1 + \tau_p^{-1} & d & \cdots & 0 \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
0 & 0 & \cdots & -d + \tau_p^{-1}
\end{pmatrix}$$

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

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:

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

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

\[ f(t) \sim N(0, k_{f,f}(t,t')) \]

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

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

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

\[ m(t) \sim N(\exp(tA)m(0), K_{m,m}(t,t')). \]

The corresponding cross covariance function \( K_{m,m}(t,t') \) is given by:

\[ K_{m,m}(t,t') = \int_0^t \int_0^{t'} \exp((t-u)A)s(\exp((t'-u')A)s k_{f,f}(u,u')dudu'. \]

The cross covariances between \( m(t) \) and \( f(t) \) becomes

\[ k_{m,f}(t,t') = \int_0^t \exp((t-u)A)s k_{f,f}(u,u')du. \]

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

2.4 Predictive distribution
Let \( f_\star \) be a vector of \( J_x \) values of the source function at equally spaced time points, and \( m_\star \) be the corresponding protein profiles at these instances in time, within the \( i^{th} \) cube along the AP axis. Concatenating we define \( h_\star = [f_\star, m_\star', \ldots, m_\star'] \) corresponding to \( J_x \times NJ \) 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 \( h = [f, m_1, \ldots, m_N]^T \) of dimension \( 1 + NJ \).

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

\[ \mathbb{E}_{\text{post}} = K_{h,h}\Sigma^{-1}y, \]

\[ K_{h,h} = K_{h,h} - K_{h,m}\Sigma^{-1}K_{m,h}. \]

Each covariance matrix in Equation (17) and (18) is partitioned across the source and \( N \) Bicoid intensities in different cubes. Illustrating this for \( h, h_\star \):

\[ K_{h,h} = \begin{bmatrix} K_{f,f} & K_{f,m_1} & \ldots & K_{f,m_N} \\ K_{m_1,f} & K_{m_1,m_1} & \ldots & K_{m_1,m_N} \\ \vdots & \vdots & \ddots & \vdots \\ K_{m_N,f} & K_{m_N,m_1} & \ldots & K_{m_N,m_N} \end{bmatrix}. \]

Therefore, the dimensions of the covariance matrices \( K_{h,h} \) and \( K_{h,h_\star} \) are \((J_x + NJ_x) \times (1 + NJ)\) and \((J_x + NJ_x) \times (J_x + NJ_x)\) respectively. The observations, collected in vector \( y \) of dimension \( 1 + NJ \), are assumed to be corrupted by additive noise as:

\[ y(t) = m(t) + e(t), \]

where \( e(t) \) are drawn from \( N(0, \sigma_e^2) \). Hence,

\[ \mathbb{K}_{Y,Y} = K_{h,h} + \Sigma, \]

\[ \Sigma = \text{diag}(\sigma_1^2, \sigma_2^2, \ldots, \sigma_{J_x}^2, \ldots, \sigma_{N}^2). \]

With the hyperparameter vector \( \theta_h = [l, \sigma_1^2, \sigma_2^2, \ldots, \sigma_{J_x}^2, \ldots, \sigma_{N}^2]. \)

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

\[ s = [S_0, 0, \ldots, 0]^T. \]

We examine the performance of our Gaussian process approach on two synthetic datasets (Fig. 1-2) 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

\[ f(t) = \delta(x) \left( \Theta(t) - \Theta(t - t_0) \right) \]

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

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 \( \tau_m \). \( \delta \) is the Kronecker delta function and \( \Theta \) is the Heaviside step function. The parameter values taken from our previous work, estimated by a least squares fit between model output and \( F_{12} \) measurements, are \( t_0 = 144 \text{ min} \) and \( \tau_m = 9 \text{ min} \). The black dashed lines in Fig. 1B, 2B and 3B show the true source function according to our hypothesis on regulation (Equation (26)).
Fig. 1. Inferred *bicoid* mRNA regulation and spatio-temporal protein concentration from a synthetic dataset; time scale from 0 ∼ 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 (red line) with 95% confidence interval. Source function used in the simulation is shown with black 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): Cycle 11 ∼ 13 and Cycle 14A class 1 ∼ 8. A partial data used in training B and C inferred source and spatio-temporal profiles respectively.

Fig. 1A shows the synthetic training data using the ODE system of Equation ((7)-(9)) with additive noise over the entire developmental period of 0 ∼ 200 min, with 20 equally spaced time points and 51 cubes along the A-P axis. Fig. 1B shows the estimated mRNA regulation function, $S_0 f(t)$, from the Gaussian process 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 be expected from a Gaussian process model, for which a function with a sharp discontinuity will have very low likelihood in the prior. Still, the decay beyond 2 hours is very rapid. The posterior mean of the inferred Bicoid concentration profile from the model is shown in Fig. 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 Fig. 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 data, shown in Fig. 2A as input. As expected, in Fig. 2B, the credible interval is wider at the early stages where no data is 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.

Fig. 3 shows the behaviour of the GP model on real data from FlyEx, the three sub-figures A, B and C 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. We show in Fig. S1 (Supplementary Information), noise levels inferred from a maximum likelihood setting (Equation (22)). These are much higher than the variance of the additive noise we used to construct the synthetic data of Fig. 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 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.
Fig. 3. Source function and Bicoid profile inferred from the FlyEx database. A average profile of 50 bootstrapped data with time scale: Cycle 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.

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 red solid lines and grey area. The black crosses represent three different training datasets shown in Fig. 1, 2 and 3. Fig. 4A-C are inferred on synthetic dataset with full time scale while the partial synthetic dataset is used in Fig. 4D-F. Fig 4G-I are inferred on real dataset from FlyEx.

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).
Fig. 4 shows the predicted temporal profiles of the Bicoid at three adjacent spatial points along the embryo. The training data are also shown. The three rows in the figure correspond to cases illustrated in Figs. 1, 2 and 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 is not present.

3.2 Non localized maternal mRNA

Additionally, we show in Fig. 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.

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 = 10µm) 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 Information Fig. 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 Fig S5 (Supplementary Information), 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, yet, the part of the source decay is fairly well estimated.

REFERENCES


