Latent force models for describing transcriptional regulation processes in the embryo development problem for the Drosophila melanogaster

Juan David Vasquez Jaramillo, Mauricio A. Álvarez, Alvaro A. Orozco

Abstract— **In the embryo development problem for the Drosophila melanogaster, a set of molecules known as morphogens are responsible for the embryo segmentation. These morphogens are encoded by different genes, including the GAP genes, maternal coordination genes and pair-rule genes. One of the maternal coordination genes encodes the Bicoid morphogen, which is the responsible for the development of the Drosophila embryo at the anterior part and for the control and regulation of the GAP genes in segmentation of the early development of the Drosophila melanogaster. The work presented in this document, reports a methodology that tends to integrate mechanistic and data driven based models, aiming at making inference over the mRNA Bicoid from gene expression data at the protein level for the Bicoid morphogen. The fundamental contribution of this work is the description of the concentration gradient of the Bicoid morphogen in the continuous spatio-temporal domain as well as the output regression (gene expression at protein level) using a Gaussian process described by a mechanistically inspired covariance function. Regression results and metrics computed for the Bicoid protein expression both in the temporal and spatial domains, showed outstanding performance with respect to reported experiments from previous studies. In this paper, a correlation coefficient of $r = 0.9758$ against a correlation coefficient of $r = 0.9086$ is being reported, as well as a SMSE of 0.0303 ± 0.1512 against a SMSE of 0.1106 ± 0.5090 and finally reporting a MSLL of -1.7036 ± 1.3472 against -1.0151 ± 1.7669.**

I. INTRODUCTION

The comprehension of biological phenomena has dramatically changed in the last 60 years due to the integration of quantitative perspectives into life sciences, imposing physical laws for the sake of making sense of biological data. As a consequence different methodologies can be established to solve a problem in terms of quantitative models. From the point of view of molecular length scales, there have been significant contributions, mainly in the fields of mathematical modeling and scientific computing, projecting some of the qualitative discussions hold in the mid 1900's to a plane where dynamical modeling showed that biology was also a matter of physicists and mathematicians, presenting traditional discussions now in quantitative terms. The problem of embryonic development comes as one of the most needed and demanding challenges in mathematical biology [4].

As a biological model, the *Drosophila melanogaster* is almost a standard test bench for the comprehension of the biological, genetic and molecular basis of embryonic development [6], [7]. From the molecular perspective, it is believed that the embryonic development is governed by a set of molecules known as *morphogens*, which propagate spatially from the anterior pole of the *Drosophila* establishing a concentration gradient across the anterior-posterior axis [3], [20]. Since 1952, it is considered that the mechanism responsible for the concentration differentiation across the AP (anterior-posterior) axis in the *Drosophila* was a diffusion-reaction process [23]. Furthermore, the importance of the Bicoid mRNA and the Bicoid morphogen is discussed [14], and the role of the concentration gradient of the Bicoid morphogen in the Drosophila embryogenesis is also discussed [10], [11].

The idea behind the study of the morphogen gradients is to establish a relationship between these molecules and the proteins responsible for egg segmentation, in order to predict the interactions present in the biological process of embryonic development. For such novel purpose, different experimental techniques have been applied to the measurement of the morphogen's concentration, which include microarrays and mass spectrometry. Measurement of the morphogen's concentration could be done as well by means of indirect techniques, which is, measuring a variable which relation to the morphogen's concentration is known by means of deterministic expression or empirical procedures [3], [6]. According to [23], these molecular transport processes could be well model at a macroscopic scale using diffusion-reaction partial differential equations, which at the end can serve as a deterministic structure for performing such indirect measurements.

The transcriptional regulation process was approached using ordinary differential equations (ODE) together with Gaussian processes [8], [9]. The approach targeted the estimation of a transcription factor (excitation) using data from the mRNA concentration (response) based on an ODE that relates excitation and response. The approach later received the name of a *latent force model*, and was introduced in detail in [2]. Latent force models have been widely used in the context of transcriptional regulation [5], [19], [22]. Since the approach reported in [9] was only based on ODEs, it could only capture the dynamic part. The spatial variation was not examined, which is what really contributes to the study of the transport processes at the macroscale. A significant advance was made when transcriptional regulation of the Bicoid morphogen included the spatial dimension in the transcriptional regulation model

J. D. Vasquez is with the Department of Physics and Astronomy at Uppsala Universitet, Uppsala, Sweden, {juan.jaramillo@physics.uu.se}; M.A. Álvarez and A. A. Orozco are with the Department of Electrical Engineering, Universidad Tecnológica de Pereira, Pereira-Colombia, {malvarez, aaog}@utp.edu.co. This work was developed while JDV was still a Master student at Universidad Tecnológica de Pereira, Pereira-Colombia.

978-1-4244-7929-0/14/$26.00 ©2014 IEEE

[12]. This model is not completely accurate, though, given the fact that the model is discretized in the spatial coordinate.

This paper presents a novel method for integrating measurements with mechanistic models associated. The proposed model namely Latent Force Model, imposes a Gaussian process on the latent variable (input), and determines a new Gaussian process for the observation (output), where the covariance functions of the whole process are determined based on a mechanistic model, particularly, the diffusion reaction partial differential equation. This methodology enables simulations of complex biological processes within the context of embryo development, for targeting experimental design in control mutations in the Drosophila embryo development and for parametrizing a biological system from information in the macroscale with no specific regard to the nanoscale.

II. MATERIALS AND METHODS

A. Datasets

1) Synthetic Data: No source concentration is provided, hence, to reproduce experimental data, we assumed a non-homogeneous diffusion-reaction process [23]:

$$\frac{\partial \phi_\alpha(\overrightarrow{r},t)}{\partial t} = D_\alpha \overrightarrow{\nabla}^2 \phi_\alpha(\overrightarrow{r},t) + U(\overrightarrow{r},t), \qquad (1)$$

where $\phi_\alpha(\overrightarrow{r},t)$ is the concentration of the chemical specie denoted by α at position $\overrightarrow{r}$ and time t, D_α is the diffusion coefficient of the chemical specie denoted by α and $U(\overrightarrow{r},t)$ is an input field at position $\overrightarrow{r}$ and time t. According to [12], expression (1) must be modified, by introducing a destruction or reaction term $\tau_p^{-1} m(x,t)$ and by constraining the input concentration $U(\overrightarrow{r},t)$ to vary only with respect to time ($S_0 u(t)$) to fulfill the demand that the mRNA concentration must be localized in time and subsequent destruction of the morphogen molecule, as follows:

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

where $m(x,t)$ is the Bicoid morphogen concentration, D is the diffusion coefficient, τ_p is the proper time of source destruction, S_0 is the source sensitivity and $u(t)$ is the source concentration localized at the anterior pole of the *Drosophila* egg (note that with relation to expression (1), $\phi_\alpha(\overrightarrow{r},t) \rightarrow m(x,t)$, that is, the problem is treated as one dimensional, and also note that the synthesis is encoded in the sensitivity S_0).
To reproduce the experimental data properly using model (2), parameters D and τ_p must be measured *in situ*, and the partial differential equation must be first discretized in space to give a state equation of the form [21]:

$$\frac{\partial m(t)}{\partial t} = Am(t) + su(t), \qquad (3)$$

where A is the matrix representing the Laplacian discretization and s is the input parameter or sensitivity. Expression (3) can easily be solved using a source function that properly represents the localized mRNA concentration [12]. To solve expression (3), using such source function, the following procedure is used. The first step was generating the input $u(t)$ as proposed in [12] using the Heaviside function as follows:

$$u(t) = \delta(x)\left(h(t) - h(t-t_0)\right) + \delta(x)h(t-t_0)e^{-\frac{t-t_0}{\tau_m}},$$

where $h(t)$ is the step function and $\delta(x)$ is the impulse function. To solve the system (3), the following expression is employed

$$m(t) = \Phi(t,t_0)x(t_0) + \int_{t_0}^{t} \Phi(t,\tau)s(\tau)d\tau, \qquad (4)$$

with $\Phi(t,\tau)$ being the transition matrix [18].

2) Real database: The data set used is FlyEx[1]. FlyEx database is a database available online with 4716 images of 14 segmentation gene expression patterns obtained from 1579 embryos and 9500000 data records. Databases in FlyEx have available data for all segmentation genes in Drosophila early development in cycles 11-13 and for all temporal classes of cycle 14A [15], [16].

B. Latent Forces

A latent force model [1], [2] is basically a Gaussian process [17], with a covariance function that encodes the behavior of a physical system, in either the spatial domain or the spatio-temporal domain. In particular, the solution to the equation (2) is given by:

$$m(x,t) = s\int_0^t \int_0^l u(\tau)G(x,\xi,t-\tau)d\xi d\tau, \qquad (5)$$

where $G(\cdot)$ is the Green's function associated to the PDE. We assume that $u(t)$ follows a GP prior with zero mean and covariance function given by

$$K_{uu}(t,t') = \operatorname{cov}\left[u(t)u(t')\right] = exp\left[-\frac{t-t'}{\sigma_t^2}\right],$$

where σ_t is a parameter of the covariance model known as the length-scale. Given that equation (5) is a linear partial differential equation, $m(x,t)$ can also be drawn from a Gaussian process with covariance function $K_{mm}(x,x',t,t') = \operatorname{cov}\left[m(x,t)m(x',t')\right]$, that can be written in the following form

$$K_{mm}(x,x',t,t') = \mathbb{E}\left[Q_{mm}(x,t)P_{mm}(x',t')\right], \qquad (6)$$

where $Q_{mm}(x,t)$ and $P_{mm}(x',t')$ are given by:

$$Q_{mm}(x,t) = s\int_0^t \int_0^l u(\tau)G(x,\xi,t-\tau)d\xi d\tau$$

$$P_{mm}(x',t') = s\int_0^{t'} \int_0^l u(\tau)G(x',\xi,t'-\tau)d\xi d\tau$$

A solution for the above expression can be found. For details, see [2]. In order to make predictions and inference over the source function $u(t)$, the covariance function $K_{mu}(x,x',t,t')$ also has to be provided. This covariance function can be

[1] http://urchin.spbcas.ru/flyex/

calculated using $K_{mu}(x,x',t,t') = \text{cov}\left[m(x',t')u(t)\right]$, leading to

$$K_{um}(x,x',t,t') = \mathbb{E}\left[s\int_0^t\int_0^l u(t')u(\tau)G(x,\xi,t-\tau)d\xi d\tau\right] \tag{7}$$

A solution for the above expression can be found analytically. The solution appears in [2].
Once the joint covariance functions are specified using the analytic expressions obtained from solving the integrals in equations (6), and (7), we could use the standard Gaussian process formulae for the posterior over $u(t)$ and the predictive distribution over $m(x,t)$. For details see [17]. Parameters σ_t and S_0 can be estimated by maximizing the evidence [17], [13].

C. Methodology

The experiments performed for evaluating the proposed method in the mentioned context used synthetic and real data. Synthetic data was generated using expression (8) and it was used to determined a theoretical expected for the input and what kind of response would this excitation produce (See theoretical expected in dashed blue in figure 1). The real dataset (FlyEx database) was used in the training and validation processes for the latent force model. Different experiments were performed, and in each experiment data from a different embryo was tested. For the training process, all the samples from the database associated with an specific embryo were used. Whilst, the validation process in each case was performed using 20 samples in the space domain and 10 samples in the time domain. For the specific case of the latent force, since no data is provided, the training process was performed using three samples of synthetic data and the validation process was done using the theoretical expected. For each of the regression profiles, both in the time domain and in the spatial domain, the evaluation of the model was performed using metrics such as SMSE (standardized mean square error) and MSLL (mean standardized logarithmic loss) [17]. The main interest is in the regression profile in the spatial domain since it provides the necessary insight for the gradient concentration of the Bicoid morphogen.

III. Results and Discussion

The results reported in this paper are organized as follows: the latent force inference, the output regression and the performance metrics for the output regression. The main contribution presented in this paper is the inclusion of a time continuous - space continuous propagator for building kernel covariance functions and this is in contrast with the reported methodology in [12], which instead uses a time continuous - space discrete propagator for building the covariance functions.

In Figure 1, the result for the latent force inference is shown (in red), against the result reported in [12] (in magenta) and the theoretical expected (in dashed-blue). There are two important regions that need to be targeted in the inference, the stability regulation and the exponential decay region [23], [15], [16], [11]. Compared to the theoretical expected , the result reported in [12] fails to reproduce the stability regulation, however, with regards to the exponential decay tracking approximates fairly well. The proposed methodology reproduced the stability regulation regime with an oscillatory behavior during most of its duration and tracked the exponential decay region approaching much more to the theoretical expected as well as the estimated value of the rate constant in comparison with the result reported in [12].

In Figure 2, the output regression for a given input space is shown. The test data [15], [16] (in black), the output regression reported in [12] (in magenta) and the output regression (in blue) obtained from applying the proposed methodology. We also include the estimation of the uncertainty (in dashed-red). The results show that the outcomes of the model reported in this paper track the events in the transient regime with a smaller error approaching the observations) in respect with the result reported in [12], which exhibits a plain exponential decay with no other trend.

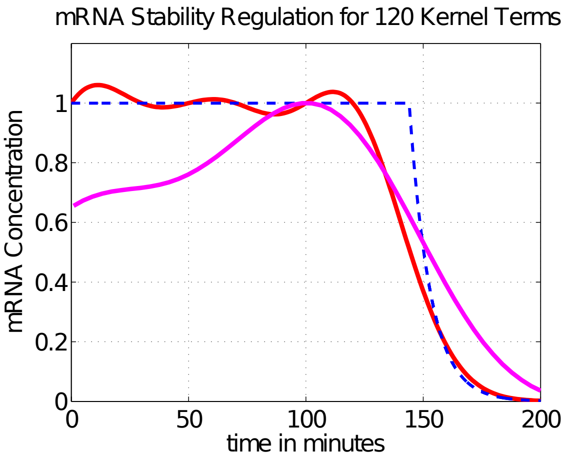

Fig. 1. Latent Force Inference Using 120 Kernel Terms to Approximate the Green's Function: Inferred Latent Force (red), Theoretical Expected (dashed blue), Inferred Result Reported by [12]

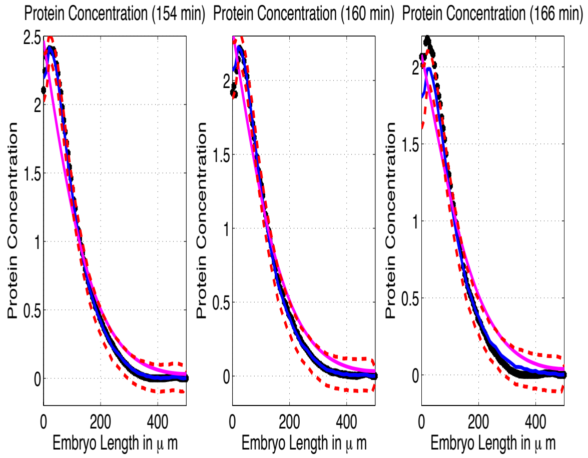

Fig. 2. Regression of the Output in the Spatial Domain: Observations (black), Output Regression (blue), Uncertainty of the Estimation (dashed red), Reported Result in [12] (magenta)

We know report the different performance measures evaluated. With respect to the correlation coefficient, the proposed methodology reported a correlation coefficient of $r = 0.9758$ against a correlation coefficient of $r = 0.9086$ calculated from the methodology reported in [12]. With respect to error measurement, the proposed methodology achieved a SMSE of 0.0303 ± 0.1512 against a SMSE of 0.1106 ± 0.5090 calculated from the reported methodology in [12]. Finally, for the MSLL, our method reported a value of -1.7036 ± 1.3472, contrasting with the value calculated from the methodology reported in [12] which came to be -1.0151 ± 1.7669.

IV. Conclusions

In this paper the results and analysis for the latent force inference corresponding to the Bicoid mRNA concentration in the Drosophila embryo development problem were presented. The main contribution of the proposed methodology is the inclusion of a continuous Green function with respect to time and space in the calculation of the covariance function for the Gaussian process. The proposed methodology outperformed the results reported in [12], approximating the stability regulation and exponential decay region with high accuracy with respect to the theoretical expected. With regards to the regression of the output, the achievements that this paper accounts for are the low uncertainty in the estimation of the predictions of the output and the precision in predictions for a given input space.

In general, the proposed methodology showed to perform better in the tasks of inference of the mRNA stability regulation and the Bicoid protein regression with respect to the results presented in [12].

V. Acknowledgments

The authors would like to thank professor Duberney Gaviria for the fruitful discussions, and Dr. Wei Liu for providing the necessary software for comparing results.

Appendix

The solution function in (4), can be approximated using the Euler-Runge-Kutta method as follows:

$$x_{k+1}^{j} = x_k^j + \frac{dt}{6}\left(\lambda_1^j(k) + \lambda_2^j(k) + \lambda_3^j(k) + \lambda_4^j(k)\right), \qquad (8)$$

where index k stands for the time stepping and index j stands for the space stepping. Furthermore,

$$\begin{aligned}
&\dot{x}(t) = f(t,x), \quad \Phi(t,\tau) = \mathscr{L}\left[(sI - A)^{-1}\right](t-\tau) \\
&\lambda_1^j(k) = f(t_k, x_k^j); \quad \lambda_2^j(k) = f(t_k + \frac{dt}{2}, x_k^j + \frac{dt}{2}\lambda_1^j(k)) \\
&\lambda_3^j(k) = f(t_k + \frac{dt}{2}, x_k^j + \frac{dt}{2}\lambda_2^j(k)) \\
&\lambda_4^j(k) = f(t_k + dt, x_k^j + dt\lambda_3^j(k)).
\end{aligned}$$

The parameters used in these simulations were obtained from [15] and [11].

References

[1] Mauricio A. Álvarez, David Luengo, and Neil D Lawrence. Latent Force Models. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 5, pages 9–16, 2009.

[2] Mauricio A. Álvarez, David Luengo, and Neil D Lawrence. Linear latent force models using Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2693–2705, November 2013.

[3] Jeffrey Alfred Drocco. *Biophysical studies of morphogen gradient formation in Drosophila melanogaster*. PhD thesis, Princeton University, 2011.

[4] Gabor Forgacs and Stuart A. Newman. *Biological Physics of the Developing Embryo*. Cambridge University Press, 2005.

[5] Pei Gao, Antti Honkela, Magnus Rattray, and Neil D Lawrence. Gaussian process modelling of latent chemical species : applications to inferring transcription factor activities. *Bioinformatics*, 24:70–75, 2008.

[6] Thomas Gregor, William Bialek, Rob R de Ruyter van Steveninck, David W Tank, and Eric F Wieschaus. Diffusion and Scaling During Early Embryonic Pattern Formation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18403–7, December 2005.

[7] Thomas Gregor, Eric F Wieschaus, Alistair P McGregor, William Bialek, and David W Tank. Stability and nuclear dynamics of the bicoid morphogen gradient. *Cell*, 130(1):141–52, July 2007.

[8] Neil Lawrence. Inferring Latent Functions with Gaussian Processes in Differential Equations, 2006.

[9] Neil D. Lawrence, Guido Sanguinetti, and Magnus Rattray. Modelling transcriptional regulation using Gaussian processes. In *NIPS*, pages 785–792, 2007.

[10] Shawn C Little, Gašper Tkačik, Thomas B Kneeland, Eric F Wieschaus, and Thomas Gregor. The formation of the Bicoid morphogen gradient requires protein movement from anteriorly localized mRNA. *PLoS biology*, 9(3):e1000596, March 2011.

[11] Wei Liu. *Machine Learning Approaches to Modelling Bicoid Morphogen in Drosophila Melanogaster*. PhD thesis, University of Southhampton, 2013.

[12] Wei Liu and Mahesan Niranjan. Gaussian process modelling for bicoid mRNA regulation in spatio-temporal Bicoid profile. *Bioinformatics (Oxford, England)*, 28(3):366–72, February 2012.

[13] Kevin Murphy. *Machine Learning: Probabilistic Perpective*. Adaptive Computation and Machine Learning Series, University of British Columbia, Canada., 2012.

[14] Theodore J Perkins, Johannes Jaeger, John Reinitz, and Leon Glass. Reverse Engineering the Gap Gene Network of Drosophila Melanogaster. *PLoS computational biology*, 2(5):e51, May 2006.

[15] Andrei Pisarev, Ekaterina Poustelnikova, Maria Samsonova, and John Reinitz. FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. *Nucleic acids research*, 37(Database issue):D560–6, January 2009.

[16] Ekaterina Poustelnikova, Andrei Pisarev, Maxim Blagov, Maria Samsonova, and John Reinitz. A database for management of gene expression data in situ. *Bioinformatics (Oxford, England)*, 20(14):2212–2221, april 2004.

[17] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge, MA, 2006.

[18] Daniel E Rivera. An Introduction to Mechanistic Models and Control Theory, 2007.

[19] Guido Sanguinetti, Neil D Lawrence, and Magnus Rattray. Gene expression Probabilistic inference of transcription factor concentrations and gene-specific regulatory activities. *Bioinformatics*, 22(22):2775–2781, 2006.

[20] Alexander Spirov, Khalid Fahmy, Martina Schneider, Erich Frei, Markus Noll, and Stefan Baumgartner. Formation of the bicoid morphogen gradient: an mRNA gradient dictates the protein gradient. *Development (Cambridge, England)*, 136(4):605–14, February 2009.

[21] Steven C. Chapra. *Applied Numerical Methods with Matlab for Engineers and Scientists*. McGraw Hill, third edition, 2002.

[22] Michalis K Titsias, Neil D Lawrence, and Magnus Rattray. Estimation of Multiple Transcription Factors using ODEs and Gaussian Processes. Technical Report April, University of MAnchester, Manchester, 2008.

[23] A. M. Turing. The Chemical Basis of Morphogenesis. *Philosophical Transactions of the Royal Society of London*, 237(641):37–72, 1952.