Meta Gene 14 (2017) 91–99

Contents lists available at ScienceDirect

Meta Gene

journal homepage: www.elsevier.com/locate/mgene

Evaluating the analytical distribution of *bicoid* gene expression profile

Hossein Hassani[a,*], Zara Ghodsi[b]

[a] Research Institute of Energy Management and Planning, No. 13, Ghods St., Enghelab Ave., Tehran, Iran
[b] Statistical Research Centre, Bournemouth University, 89 Holdenhurst Road, Bournemouth BH8 8EB, UK

ARTICLE INFO

Keywords:
Bicoid
Distribution
Drosophila melanogaster
Model
segmentation gene

ABSTRACT

Segmentation in *Drosophila melanogaster* starts with a key maternal input known as bicoid gene. The initial positional information provided by this gene induces the sequential activation of segmentation network. Therefore, an accurate mathematical model describing the gene expression profile of bicoid gene expects to provide essential insights into the gene cross-regulations presented in that network. The significantly stochastic, highly volatile and non-normal nature of the bicoid gene expression profile encouraged us to look for the best distribution function describing this profile. We exploit the use of fifty-four different powerful and widely-used distributions and conclude that FatigueLife(3P) fits the data more accurately than the other distributions. The reliability and validity of the results are evaluated via both simulation studies and empirical evidence thereby adding more confidence and value to the findings of this research.

1. Introduction

Bicoid[1] is a homeodomain transcription factor which plays a crucial role in patterning the head and thorax of *Drosophila melanogaster* during the embryogenesis stage (Driever and Nüsslein-Volhard, 1988; Frohnhöfer and Nüsslein-Volhard, 1986). It is widely accepted that embryos receiving different doses of *bcd* have differently sized anterior structures and in the absence of this morphogen, the anterior structures are replaced with the posterior regions (Berleth et al., 1988; Driever and Nüsslein-Volhard, 1988; Frohnhöfer and Nüsslein-Volhard, 1986).

Since the discovery of *bcd* in 1988, several models have been put forward to formulate the gradient of this morphogen (see, for example, Grimm et al., 2010, Spemann and Mangold, 2003, Wartlick et al., 2009). However, as the experimentally achieved gradient is highly volatile, the proposed models exhibited limited performance (Jaeger et al., 2004; Reinitz and Sharp, 1995). For example using the Synthesis Diffusion Degradation(SDD) model as the most frequently applied one, the time needed for attaining the steady state concentration profile is much longer than the protein lifetime and the length constant is much smaller than the length of the embryo (Ghodsi et al., 2015b).

Therefore, the extensive studies on molecular and functional features of this gradient have been continued and led to considerable improvements in different branches of developmental studies including embryogenesis, regional specification and metamorphosis (Müller et al., 2015).

Furthermore, finding a precise model for expression pattern of *bcd* also expects to give us a better understanding of an important developmental process known as canalisation (Staller et al., 2015). According to C.H. Weddington, an efficient way to unveil the exact canalisation process is to study the interaction between genes in a gene regulatory network (Waddington, 1942). Hence, to achieve a deeper understanding of gene-gene interactions, segmentation network in *Drosophila melanogaster* has been considered as a premier system for coupling experimental data and computational models (Hengenius et al., 2011; Jaeger et al., 2004; Papatsenko and Levine, 2011; Poustelnikova et al., 2004; Reinitz and Sharp, 1995).

Such studies are aimed at finding quantitative models that illustrate a mathematical picture of the protein concentrations produced by segmentation genes (among which *bcd* has a significant role as a valuable input to this network).

According to the hypotheses of these studies, if a model faithfully reproduces the wild type gene expression patterns then it would be possible to use that model to predict the genetic interactions of the segmentation network correctly. However, as can be seen in Fig. 1, due to the high volatility, heavy tail and lack of normality of the data, even modelling the Bcd as the simplest gene expression pattern of this network is not a simple task. The Bcd data characteristic is further discussed in Section 2.

It should also be noted that the expression of segmentation genes, especially *bcd*, are significantly stochastic with randomness in transcription and translation (Lecca et al., 2010). This stochasticity makes the modelling of the segmentation network considerably challenging.

* Corresponding author at: Department of Statistics, Payame Noor University, 19395-4697 Tehran, Islamic Republic of Iran.
E-mail address: hassani.stat@gmail.com (H. Hassani).

[1] In what follows, the italic lower-case *bcd* represents either gene or mRNA and Bcd refers to the protein.

http://dx.doi.org/10.1016/j.mgene.2017.07.014
Received 21 December 2016; Received in revised form 28 July 2017; Accepted 30 July 2017
Available online 09 August 2017

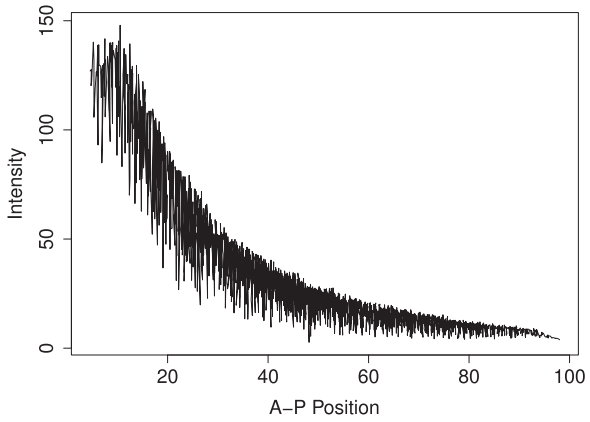

Fig. 1. A typical example of Bcd gene expression profile. Y-axis shows the fluorescence intensities and X-axis shows the position along the Anterior-Posterior (A-P) axis of the embryo.

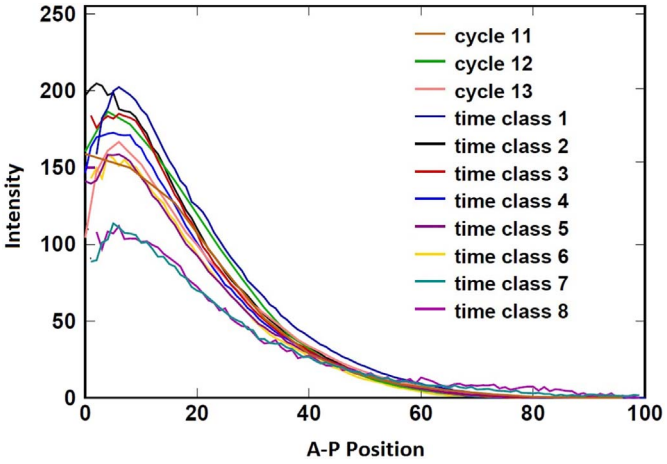

Fig. 2. The Bcd gradient along the embryo in different cleavage cycles and temporal classes.
Source: Figure adapted from Surkova et al. (2008).

The stochasticity is both controlled and exploited by cells and, as such, must be included in models of genetic networks (Shahrezaei and Swain, 2008). An effective way to grasp the function of a stochastic process is to drive the probability distribution of the process of interest. Moreover, a major problem in system biology is to determine which properties of the biochemical networks must be modelled to make accurate, quantitative predictions. Estimating the parameters of a distribution function can be a useful guide to find the characteristics of a gene expression pattern which allow to develop more valid predictions if included in a model.

Accordingly, this study seeks to evaluate the theoretical distribution of Bcd gene expression profile and to introduce the best statistical distribution describing this gradient. We have examined fifty four probability distributions. To validate the theoretical results, extensive simulation studies have been carried out. Analysing the real data set have also been performed on all the cleavage cycles in which Bcd is present in the embryo.

This development expects to open up the possibility of using statistical distribution to depict the characteristics of gene expression profiles and unveil the interaction networks in a dynamic multivariate system.

The remainder of this paper is organised as follows. Section 2 describes the data set applied in this study which is followed by a portrayal of the simulation study procedure. Section 3 describes the analytical methods adopted in this study. Section 4 summarises the empirical results and the paper concludes with a concise summary in Section 5.

2. Bicoid data

2.1. Real data

The quantitative *bcd* gene expression data in wild-type *Drosophila melanogaster* embryos was obtained from FlyEx database (Pisarev et al., 2009). This data set has been widely used as a valuable source of information for studying the dynamics of segment determination of early *Drosophila* development (Poustelnikova et al., 2004).

Data acquisition in this data set is based on the confocal scanning microscopy of fixed embryos immunostained for segmentation proteins. The applied antibody allows the visualisation of the Bcd proteins. In this study, the expression profiles were extracted from the nuclear intensities of 10% longitudinal and are unprocessed for any noise reduction methods. Similar to Ghodsi et al. (2015a,b), Hassani and Ghodsi (2014), we set to work with one-dimensional gene expression data. Hence, the second spatial coordinate (dorsoventral axis) has not been considered. In the achieved profiles, higher intensities imply greater presence of the Bcd protein.

Since the segment determination starts from cleavage cycle 10 and lasts to cleavage cycle 14A (when proteins synthesised from maternal transcripts begin to appear up to the onset of gastrulation) the data has been categorised to five main cycles of 10 to 14A. Additionally, as the cleavage cycle 14A is considerably longer in time, to facilitate the analysis, temporal classes 1 to 8 have been considered as the subgroups of this cleavage cycle. It should also be noted that each class of data contains a different number of embryos.

Since there is an undeniable variation in the pattern of Bcd in different cleavage cycles, it is critical to investigate whether a single distribution function can be of general use for Bcd profile or a different distribution should be defined for each cleavage cycle.

Fig. 2, illustrates the pattern of the Bcd profile for an individual embryo in cleavage cycle 11 to cleavage cycle 14, time class 8. It is of note that to depict the difference between the pattern of Bcd in different developmental cycles more precisely, a filtering step has been applied and the signals of the gene expression profiles extracted by Singular Spectrum Analysis (SSA) technique were used.

Table 1 presents the descriptive statistics of Bcd. As it can be seen, each cycle has a different number of embryos, and the length of the profiles obtained from each embryo is distinctive where a large series length indicates that a greater number of nuclei was presenting the fluorescence intensity. In other words, Bcd protein molecules were produced in a higher number of nuclei along the A-P axis.

For each cycle, average of variance, series length, mean, skewness and kurtosis are presented separately. Due to the considerable variation present in the data, median has been chosen as a measure of central tendency. The fourth column shows the variation of each profile within a cycle. For example, in time class 10, the minimum variance seen is 928.8, however, the maximum variance for this cycle is more than 2000. Hence, we are dealing here with two kinds of variation; within a cycle and between a cycle variation. The skewness was also tested, and the results confirm that there is a statistically significant skewness (at 5% level) indicating that almost all series have values towards the lower end in the series.

Determining whether the data is symmetric, left-skewed, or right-skewed is critical since a distribution which has the same shape as the profile under study would be expected to be a better candidate to fit the data.

Fig. 3 shows the histogram of Bcd profile. In plotting these histogram only one dimension of the data (the achieved fluorescence intensities) for two different individual embryos[2] has been used. As it is

[2] Histograms of cleavage cycles 10–13 and all time classes of cleavage cycle 14A are presented in Appendix B.