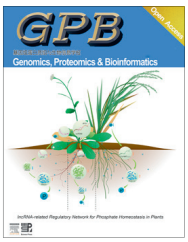
Genomics Proteomics Bioinformatics 13 (2015) 183–191

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com



METHOD

Bicoid Signal Extraction with a Selection of Parametric and Nonparametric Signal Processing Techniques

Zara Ghodsi [1,a], Emmanuel Sirimal Silva [1,b], Hossein Hassani [1,2,*,c]

[1] *The Statistical Research Centre, Bournemouth University, Bournemouth BH8 8EB, UK*
[2] *Institute for International Energy Studies (IIES), Tehran 1967743 711, Iran*

Received 11 August 2014; revised 22 January 2015; accepted 12 February 2015
Available online 18 July 2015

Handled by Fengfeng Zhou

KEYWORDS

Bicoid;
Drosophila melanogaster;
Signal extraction;
Signal processing

Abstract The maternal segmentation coordinate gene ***bicoid*** plays a significant role during ***Drosophila*** embryogenesis. The gradient of **Bicoid**, the protein encoded by this gene, determines most aspects of head and thorax development. This paper seeks to explore the applicability of a variety of **signal processing** techniques at extracting ***bicoid*** expression signal, and whether these methods can outperform the current model. We evaluate the use of six different powerful and widely-used models representing both parametric and nonparametric **signal processing** techniques to determine the most efficient method for **signal extraction** in ***bicoid***. The results are evaluated using both real and simulated data. Our findings show that the Singular Spectrum Analysis technique proposed in this paper outperforms the synthesis diffusion degradation model for filtering the noisy protein profile of ***bicoid*** whilst the exponential smoothing technique was found to be the next best alternative followed by the autoregressive integrated moving average.

Introduction

Morphogens are molecules which determine a cell's destiny in a concentration-dependent mode by governing the pattern of tissue development and the position of various specialized cell types within a tissue in the process of morphogenesis [1–3]. A classic example of morphogens is bicoid (*bcd*), which is the first known morphogen identified by Nsslein–Volhard in 1988 [1] and encodes a homeobox transcription factor (in what follows, the italic lower-case *bcd* represents either the gene or mRNA and Bcd refers to protein). *bcd* is localised at the anterior end of the egg during the oogenesis [2] and translation of *bcd* begins after fertilization. Consequently, Bcd distributes along the anterior-posterior (AP) axis of the egg, forming a concentration gradient [2]. Such diffusion of Bcd by regulating the production of the anterior structures determines the position

* Corresponding author.
E-mail: hhassani@bournemouth.ac.uk (Hassani H).
[a] ORCID: 0000-0002-0794-5168.
[b] ORCID: 0000-0003-3851-9230.
[c] ORCID: 0000-0003-0897-8663.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

http://dx.doi.org/10.1016/j.gpb.2015.02.006

and size of head and thorax of an adult fruit fly (http://highered.mcgraw-hill.com).

Several computational models have been published for Bcd gradient over the last three decades (see, for example, [3]). However, as the Bcd profile achieved by fluorescence antibodies technique is highly volatile, some proposed models, such as the simple synthesis diffusion degradation (SDD) model, only exhibited limited performance [4,5]. They fail to clearly explain some characteristics of the Bcd gradient, such as protein life time and length constant [3,6–7]. **Figure 1** shows a typical example of the Bcd gradient along the egg length at cleavage cycle 14(3), effect of noise (*i.e.*, fluctuations visible in Figure 1) in this gradient can be seen as the high volatile pattern. An initial look at the distribution suggests Bcd follows an exponential trend. However, owing to the high volatility seen in the series, the extraction of this signal is not a simple task.

SDD, which was formulated before the identification of *bcd* [4,8–11], is the most widely-accepted among the models used to explain Bcd diffusion pattern. SDD ich. As a relatively simple model, SDD follows an exponential curve [12]:

$$B = Ae^{-x/\lambda} \tag{1}$$

where A is the amplitude, x is distance from the anterior [13], and λ is the length parameter obtained by fitting an exponential model to the *bcd* intensity profile and computing the position at which the concentration has dropped to $1/exp$ of the maximal value at the anterior (at $x = 0$) [3]. However, this model is not fully consistent with all the experimental observations. For example based on [4], if Bcd molecules diffuse along the embryo with diffusion constant D and Bcd lifetime of τ, the concentration of Bcd in this model follows:

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

where, x and t represent positions along the egg and time, respectively, $S(x,t)$ is a source function describing the production of Bcd molecules, $m(x,t)$ is the formed concentration, and τ_p represents protein lifetime [14]. Nevertheless, when using this model the time needed for attaining the steady state concentration profile is much longer than the protein lifetime τ, whereas the length constant λ is much smaller than the length of the embryo. Moreover, pattern of Bcd expression established by any model should be flexible to different time scales, egg length, and embryos sizes [4,10].

Not only in developmental studies but also in all fields of genetic studies, signal extraction and noise reduction are regarded as important tasks since genetic data are often characterized by the existence of considerable noise. Many methods are utilized for signal extraction, such as machine learning algorithms [15,16] and different background removal techniques [17–19]. In this paper we evaluate the use of powerful and popular signal processing techniques which include both parametric and nonparametric methods to provide a sound extraction of Bcd signal. Our aim is to examine whether the selected signal processing models can provide a more accurate signal extraction of Bcd in comparison to SDD.

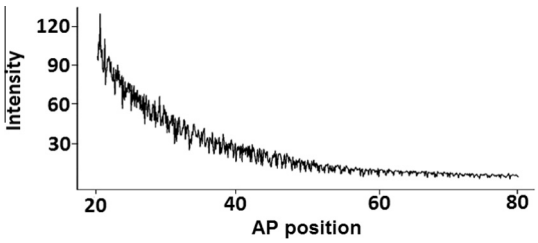


Figure 1 A typical example of noisy Bicoid
Y-axis shows the fluorescence intensities obtained from the attached fluorescence antibodies to the Bcd molecules and *X*-axis shows the position along the embryo.

The selection of models representing both parametric and nonparametric methods is important for several reasons. Firstly, as seen below, the residuals following signal extraction in Bcd are nonstationary. Secondly, parametric models rely on the underlying assumptions of normality and stationarity, and interestingly, SDD model is parametric. Thirdly, as noted in [20], for the parametric methods, assuming stationarity for the data, linearity of the model and normality of the residuals can provide only an approximation of the true situation. Therefore, a method that does not depend on these assumptions could be very useful for modelling and extracting the signal in Bcd data. Moreover, previous applications in solving signal extraction problems were taken into account when selecting models in this study. We use the SDD model as the overall benchmark as it is the most widely accepted approach for signal extraction in Bcd. We also consider the parametric autoregressive integrated moving average (ARIMA) [21], which has been applied for signal extraction in various fields both historically and recently (see for example, [22–24]). In addition, autoregressive fractionally integrated moving average (ARFIMA), which is mainly recognized as a parametric method suitable for long memory processes where the decay is slower than in an exponential process [24], is included for comparison as well. Other parametric models considered are state space models such as exponential smoothing (ETS), since SDD in itself follows an exponential curve [12]. Singular spectrum analysis (SSA) technique (like neural networks, NN) is a nonparametric signal processing model and does not rely on any assumptions [25]. The SSA technique was initially evaluated for gene expression [26] and has been previously applied for signal extraction [2,6,7,26–30]. Therefore, the models used in this paper include an optimized version of ARIMA [31], an ARFIMA model [31], ETS [32], a feed forward NN model [32], and SSA [33].

Gene expression can be traced either in time or space. The data points used in this study represent the intensity levels for the positions along the AP axis and are considered as a sequenced series. Therefore, one-dimensional gene expression data are used for signal extraction and the second spatial coordinate (Dorsoventral DV axis) has not been considered in this study. Moreover, it is important to note that this paper is not aimed at showing any particular technique to be universally best for modelling the Bcd signal. Instead we are mainly interested in showing how the selected signal processing techniques compare and compete against each other, and the widely