# Evaluating the Analytical Distribution of *bicoid* Gene Expression Profile

Zara Ghodsi* Hossein Hassani†

**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.

**Keywords:** *bicoid*; *distribution*; *Drosophila melanogaster*; *model*; *segmentation gene*.

## 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 [1, 2]. 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 [1–3].

Since the discovery of *bcd* in 1988, several models have been put forward to formulate the gradient of this morphogen (see, for example, [4–6]). However, as the experimentally achieved gradient is highly volatile, the proposed models exhibited limited performance [7, 9]. 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

*Statistical Research Centre, Bournemouth University,Fern Barrow, Poole, BH125BB, UK, e-mail: zghodsi@bournemouth.ac.uk.

†Institute for International Energy Studies (IIES), 65 Sayeh Street, Vali-Asr Avenue, Tehran, Iran, email: hassani.stat@gmail.com.

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

protein lifetime and the length constant is much smaller than the length of the embryo [8].

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 [10].

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 [11]. 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 [12]. 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 [7, 9, 13–15].

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 Figure 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.

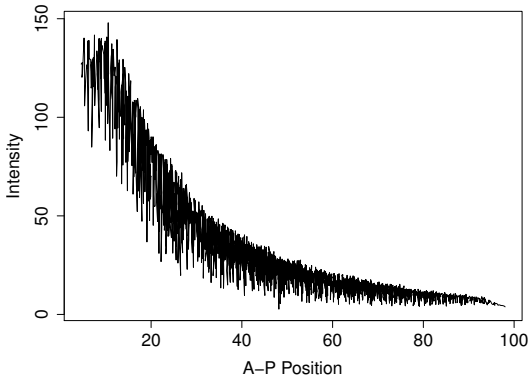



Figure 1: A typical example of Bcd gene expression profile. Y-axis shows the fluorescence intensities and X-axis shows the position along the Anterior-Posterior (A-P) axis of the embryo.

It should also be noted that the expression of segmentation genes, especially *bcd*, are significantly stochastic with randomness in transcription and translation [16]. This stochasticity makes the modelling of the segmentation network considerably challenging. The stochasticity is both controlled and exploited by

cells and, as such, must be included in models of genetic networks [17]. 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 fifity 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 [18]. This dataset has been widely used as a valuable source of information for studying the dynamics of segment determination of early *Drosophila* development [13].

Data acquisition in this dataset 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 [8, 19, 20], 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.

Figure 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.

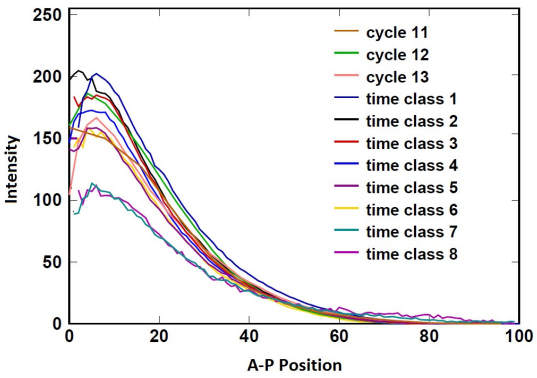



Figure 2: The Bcd gradient along the embryo in different cleavage cycles and temporal classes. Figure adapted from [21].

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, skewnwss 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 skweness 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.

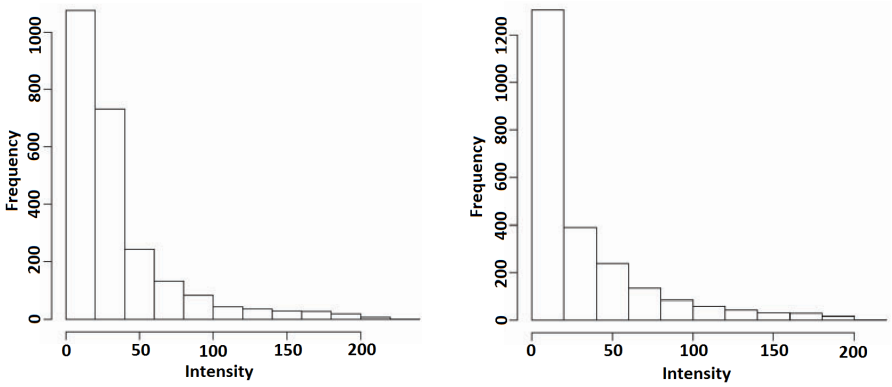
| Cycles | N | | Var. | length | Mean | Med. | Min. | Max. | Skew. | Kurt. |
|---|---|---|---|---|---|---|---|---|---|---|
| 10 | 7 | Min | 928.8 | 79.00 | 16.95 | 2.970 | 0 | 134.4 | 1.09 | 0.33 |
| | | Med | 1294 | 124.0 | 46.33 | 37.55 | 7.420 | 163.9 | 1.550 | 1.950 |
| | | Max | 2358 | 146.0 | 70.83 | 54.68 | 20.95 | 209.5 | 1.670 | 4.020 |
| 11 | 14 | Min | 693.5 | 238.0 | 34.41 | 21.39 | 3.67 | 152.91 | 1.190 | 0.300 |
| | | Med | 1780 | 288.5 | 46.30 | 26.92 | 6.160 | 185.8 | 1.460 | 1.120 |
| | | Max | 2999 | 308.0 | 77.07 | 63.64 | 20.96 | 223.2 | 1.860 | 2.980 |
| 12 | 31 | Min | 1160 | 394.0 | 35.15 | 17.63 | 1.570 | 147.20 | 0.66 | -1.27 |
| | | Med | 2422 | 524.0 | 51.09 | 28.41 | 7.110 | 206.4 | 1.420 | 0.980 |
| | | Max | 5224 | 607.0 | 165.5 | 174.6 | 87.62 | 239.7 | 1.860 | 2.850 |
| 13 | 98 | Min | 412.6 | 738.0 | 21.97 | 13.05 | 0 | 131.0 | 0.660 | -0.32 |
| | | Med | 1578 | 1276 | 42.08 | 26.21 | 4.740 | 197.8 | 1.810 | 2.701 |
| | | Max | 2795 | 1550 | 77.87 | 65.39 | 16.14 | 240.5 | 2.640 | 7.430 |
| 14(1) | 58 | Min | 705.4 | 1085 | 24.07 | 12.00 | 0 | 143.7 | 1 | -0.15 |
| | | Med | 2041 | 2257 | 43.25 | 24.73 | 4.110 | 223.67 | 1.890 | 2.880 |
| | | Max | 2968 | 2548 | 148.66 | 131.6 | 67.19 | 252.9 | 2.600 | 6.840 |
| 14(2) | 30 | Min | 1344 | 2043 | 32.18 | 16.54 | 0 | 190.9 | 1.680 | 2.250 |
| | | Med | 1921 | 2315 | 42.93 | 25.16 | 4.430 | 225.6 | 1.690 | 3.400 |
| | | Max | 1887 | 2678 | 51.64 | 36.20 | 11.84 | 245.63 | 2.710 | 7.240 |
| 14(3) | 38 | Min | 480.8 | 1642 | 17.62 | 6.460 | 0 | 147.0 | 0.680 | -0.66 |
| | | Med | 1490 | 2280 | 40.30 | 23.19 | 6.070 | 216.02 | 2.18 | 4.48 |
| | | Max | 2654 | 2783 | 138.9 | 125.9 | 56.07 | 252.7 | 2.560 | 6.540 |
| 14(4) | 28 | Min | 697.7 | 1741 | 33.49 | 16.39 | 0 | 170.8 | 1.390 | 1.110 |
| | | Med | 1578 | 2275 | 42.17 | 25.88 | 7 | 212.1 | 1.990 | 3.510 |
| | | Max | 2324 | 2520 | 55.33 | 42.64 | 13.91 | 234.6 | 2.250 | 5.110 |
| 14(5) | 25 | Min | 439.6 | 1707 | 22.87 | 13.69 | 0 | 113.0 | 0.520 | -0.90 |
| | | Med | 1195 | 2297 | 38.17 | 23.99 | 4.400 | 192.64 | 2.020 | 3.850 |
| | | Max | 2263 | 2453 | 154.0 | 137.7 | 71.08 | 236.60 | 2.270 | 5.450 |
| 14(6) | 29 | Min | 84.37 | 1583 | 27.85 | 15.64 | 0 | 134.20 | 0.980 | 0.820 |
| | | Med | 1131 | 2266 | 39.26 | 25.81 | 7.620 | 194.3 | 1.920 | 3.300 |
| | | Max | 2057 | 2584 | 93.40 | 83.62 | 40.43 | 235.2 | 2.460 | 6.700 |
| 14(7) | 15 | Min | 141.0 | 1535 | 17.48 | 12.70 | 0 | 81.58 | 0.670 | 0.090 |
| | | Med | 443.5 | 2109 | 40.00 | 34.94 | 8.140 | 133.6 | 1.670 | 2.110 |
| | | Max | 18060 | 2423 | 108.7 | 101.3 | 52.56 | 220.6 | 2.460 | 6.700 |
| 14(8) | 12 | Min | 397.9 | 1245 | 26.05 | 14.79 | 2.170 | 133.6 | 0.630 | -0.36 |
| | | Med | 636.2 | 1521 | 64.68 | 56.00 | 21.55 | 170.0 | 1.470 | 2.120 |
| | | Max | 818.0 | 2195 | 134.1 | 128.9 | 80.26 | 202.0 | 2.060 | 5.130 |

Table 1: Descriptive statistics of Bcd profile. N= Number of embryos studied per cycle (or time class), Var.= Variance in each profile, Length= Length of data in each expression profile, Mean=The average of intensity levels, Med.= Median of intensity levels, Min.= The minimum value of intensity levels, Max.= The maximum value of intensity levels, Skew. =Skewness, and Kurt.= Kurtosis.

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.

Figure 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 apparent, Bcd profile is asymmetric and notably right-skewed suggesting that it may be poorly characterised by its mean and variance. Further on, all distributions with a tail on the right side should fit more accurately to this profile.

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




(a) Time Class 14(3) (b) Time Class 14(4)

Figure 3: The histogram of Bcd related to time class 14(3) and 14(4).

## 2.2 Simulated Data

To find the best distribution function explaining the features of Bcd expression profile, relying on only the real data presents a number of challenges. Table 1, the second column, shows the number of embryos available in our data set separately in each cleavage cycle. As can be seen, the number of embryos ( from each, one Bcd expression profile is obtained) varies greatly in different cleavage cycles ranging between 7 in cleavage cycle 10 and 98 in cleavage cycle 13.

Furthermore, a close look at the real data suggests that the raw gene expression profile data is noisy and noise consisted of unknown structure. According to real data, the noise level also varies considerably between the expression profiles. Therefore, the appropriateness of a distribution should not only be examined for different developmental classes but should also be tested for different levels of noise existing in the real data.

Thus, in this paper, a series of simulated data is used to evaluate and rank the performance of different probability distribution functions.

To facilitate the study, three levels of noise including low, average and high volatile (in what follows, respectively assigned as noise level 1, noise level 2 and noise level 3) can be considered.

More importantly, as depicted in Figure 2, some of the expression profiles possess an initial curve in their pattern, whereas this curve is not present in the rest of the profiles. The initial curve attributes to the concentration of Bcd in nuclei at the anterior position along the AP axis which shows that Bcd concentration initially reaches a maximum value before decaying along the embryo. detecting the initial curve in some of the embryos suggests that the mechanism of gradient readout might be more complicated than reading a particular concentration of the morphogen. Therefore, for all different considered noise levels, the assessment conducted separately on the simulated profiles with and without the initial curve.

To start the simulation, an exponential curve drawn from the simple Synthesis Diffusion Degradation (SDD) model which is a common model for analysing the Bcd profile has been considered [1,22–24]. The concentration of Bcd in this model follows:

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

where $A$ is the amplitude, $x$ is distance from the anterior, and $\lambda$ 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 $x = 0$ ) [1,22–24].

To obtain a noisy series as close as possible to the real one, random errors $\epsilon$ of a normal distribution with zero mean and variance $\sigma_\epsilon^2$ with different amplitudes were added to various parts of this curve. This simulation was repeated $1,000$ times.

# 3 Empirical Results

## 3.1 Simulated Series

Finding the best distribution is defined as the procedure of selecting a statistical distribution that in average fits best to Bcd profiles of all cycles and therefore it is the most valid model to describe the data.

In total fifty–four distributions have been selected [3]. After fitting the distributions, it is necessary to determine whether a given distribution provides a reliable fit. Therefore, it is recommended to perform the goodness of fit tests to determine how well the distribution fits each profile.

In addition, to ascertain the most accurate distribution, statistical moments (mean, variance, etc.), tail probabilities and quantiles have also been calculated for each distribution and for distribution families with different sets of defined parameters. This process is performed for both simulated and real data.

The goodness of fit tests adopted here are among the most popular tests including Anderson-Darling (AD), Kolmogorov-Smirnov (KS), and Chi-square ($\chi^2$) tests. However, since the result provided by $\chi^2$ test was not consistent with the AD and KS tests at different cleavage cycles, the outcome of this test is not reported here.

AD test is the most widely accepted goodness of fit test, particularly for skewed series and therefore, it is considered as the primary criterion for decision making in this study [25,26]. It should be noted that in spite of having a same principal in the application, the implementation of the KS and AD tests is noticeably different [27].

The goodness of fit tests measure the distance between the actual data and the fitted distribution. Hence, after fitting different models, the distribution with the lowest statistic value will be rated as the best model and the rest of the models will be ranked down according to their test statistics. This approach enables us to easily compare the fitted models and determine the best fitted distribution.

[3] A complete list of all distributions applied in this study can be found in Appendix 1.

Tables 2 and 3 report the results of the simulation study respectively for the profiles with and without the initial curve following 1000 iterations. After each round of simulation, the distributions are ranked down based on their test statistics. The three best-fitted distributions are then assigned to those distributions which show the highest frequency in iteration and the lowest test statistic. This procedure is repeated for all different noise levels. According to the result shown in Table 2, FatigueLife(3P) distribution provides a good fit confirming by both KS and AD goodness of fit tests and is also well performed at all different noise levels.

Interestingly, with reference to the superior performance of FatigueLife(3P) distribution the outcome related to the profiles with the initial curve is consistent withTable 2, suggesting that FatigueLife(3P) is practical for both types of profiles. Therefore, there is no need to consider a separate distribution to formulate those series with the initial curve. However, the test statistics related to the performance of each distribution are lower in table 2, indicating that capturing a closer fit is more challenging in profiles with the initial curve.

| Noise Level | **Goodness of Fit Test** | | | |
|---|---|---|---|---|
| | Anderson-Darling | | Kolmogorov-Smirnov | |
| | Distribution | Statistic | Distribution | Statistic |
| 1 | FatigueLife(3P) | 1.55 | FatigueLife(3P) | 0.03 |
| | Log-Pearson3 | 2.99 | JohnsonSB | 0.04 |
| | Lognormal | 3.50 | Log-Pearson3 | 0.04 |
| 2 | FatigueLife(3P) | 1.64 | FatigueLife(3P) | 0.04 |
| | Lognormal | 3.16 | JohnsonSB | 0.04 |
| | Log-Pearson3 | 2.70 | Log-Pearson3 | 0.04 |
| 3 | FatigueLife(3P) | 1.36 | FatigueLife(3P) | 0.03 |
| | Log-Pearson3 | 2.21 | Burr(4P) | 0.06 |
| | Lognormal(3P) | 2.65 | Log-Pearson3 | 0.04 |

Table 2: The result of the simulated profiles without the initial curve.

| Noise Level | **Goodness of Fit Test** | | | |
|---|---|---|---|---|
| | Anderson-Darling | | Kolmogorov-Smirnov | |
| | Distribution | Statistic | Distribution | Statistic |
| 1 | FatigueLife(3P) | 2.68 | FatigueLife(3P) | 0.04 |
| | Log-Pearson3 | 3.58 | Log-Pearson3 | 0.05 |
| | Lognormal | 4.32 | Dagum | 0.05 |
| 2 | FatigueLife(3P) | 2.68 | FatigueLife(3P) | 0.05 |
| | Lognormal(3P) | 3.93 | Burr(4P) | 0.05 |
| | Burr(4P) | 3.57 | Log-Pearson3 | 0.05 |
| 3 | FatigueLife(3P) | 2.08 | FatigueLife(3P) | 0.04 |
| | Log-Pearson3 | 2.82 | Log-Pearson3 | 0.05 |
| | Lognormal(3P) | 3.55 | Burr(4P) | 0.05 |

Table 3: The result of the simulated profiles with the initial curve.

## 3.2 Bcd data

Next, to evaluate the reliability of the result obtained at the simulation step and primarily to examine the performance of the FatigueLife(3P) distribution, the analysis further conducted on the real data sets. Table 4 shows the outcome of this effort. Overall, the result of the application to real data appears to be consistent with the simulation findings and therefore confirms the validity of the results. As it can be seen, FatigueLife(3P) appears among the top three distributions for all cycles indicating the out-performance of this distribution. Figure 4 depicts the probability density function of FatigueLife(3P) distribution fitted to a Bcd profile of an embryo of time class 14(4).

| Cycle | **Goodness of Fit Test** | | | |
|---|---|---|---|---|
| | Anderson-Darling | | Kolmogorov-Smirnov | |
| | Distribution | Statistic | Distribution | Statistic |
| 10 | FatigueLife(3P) | 0.42 | FatigueLife(3P) | 0.06 |
| | Pearson6 | 1.24 | Burr | 0.07 |
| | Burr | 1.30 | Lognormal | 0.07 |
| 11 | Pearson5 | 0.82 | FatigueLife(3P) | 0.05 |
| | FatigueLife(3P) | 1.07 | Burr | 0.06 |
| | Lognormal(3P) | 1.10 | Lognormal | 0.08 |
| 12 | FatigueLife(3P) | 2.93 | Dagum(4P) | 0.05 |
| | Inv.Gaussian(3P) | 3.03 | FatigueLife(3P) | 0.06 |
| | Pearson5(3P) | 3.43 | Pearson5(3P) | 0.06 |
| 13 | Burr | 8.38 | Lognormal | 0.32 |
| | FatigueLife | 9.80 | Pearson6 | 0.06 |
| | Pearson6 | 11.41 | FatigueLife(3P) | 0.06 |
| 14(1) | Burr | 13.86 | Burr | 0.05 |
| | Pearson6 | 20.82 | FatigueLife(3P) | 0.07 |
| | FatigueLife(3P) | 26.78 | Lognormal | 0.08 |
| 14(2) | FatigueLife(3P) | 27.27 | Pearson6 | 0.06 |
| | Lognormal | 43.33 | FatigueLife(3P) | 0.08 |
| | Log-Logistic | 46.34 | Lognormal | 0.09 |
| 14(3) | Lognormal | 17.30 | Pearson6 | 0.06 |
| | FatigueLife(3P) | 25.22 | FatigueLife(3P) | 0.08 |
| | Gen.Gamma(4P) | 28.37 | Lognormal | 0.09 |
| 14(4) | Pearson5(3P) | 9.68 | Log-Logistic(3P) | 0.05 |
| | Pearson6 | 15.89 | Lognormal(3P) | 0.06 |
| | FatigueLife(3P) | 19.78 | FatigueLife(3P) | 0.07 |
| 14(5) | Burr | 8.43 | Burr(4P) | 0.04 |
| | Pearson6 | 14.33 | Log-Logistic(3P) | 0.05 |
| | FatigueLife(3P) | 25.71 | FatigueLife(3P) | 0.07 |
| 14(6) | Burr | 17.81 | Frechet | 0.06 |
| | FatigueLife(3P) | 24.51 | Pearson6 | 0.08 |
| | Lognormal | 67.79 | FatigueLife(3P) | 0.08 |
| 14(7) | Burr | 7.08 | Burr | 0.04 |
| | Log-Logistic | 17.32 | Pearson6 | 0.05 |
| | FatigueLife(3P) | 19.46 | FatigueLife(3P) | 0.06 |
| 14(8) | Burr | 3.32 | Inv.Gaussian(3P) | 0.06 |
| | FatigueLife(3P) | 5.57 | Gen.Gamma(4P) | 0.06 |
| | Log-Logistic(3P) | 7.75 | FatigueLife(3P) | 0.06 |

Table 4: A summary of the real data set distribution fitting result.

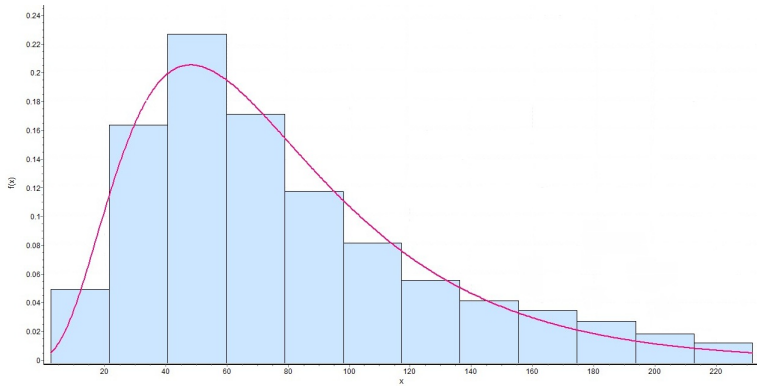



Figure 4: An example of FatigueLife(3P) distribution fitted to a Bcd profile.

Bcd molecules synthesised from the maternal transcripts begin to appear at cleavage cycle 10. Hence, the amount of these morphogens, in those areas where they are concentrated is lower amount for the time classes 10-12 making the length of the data and the expression level for those cycles considerably less than the other cycles. Therefore, the consistency in superior performance of the FatigueLife(3P) distribution in cycles 10-12 should be addressed as an important feature of this distribution.

Table 5 reports the most frequent distributions throughout all cleavage cycles in the real data. This result is reported separately for AD and KS test. From this table, it is evident that according to AD test, after FatigueLife(3P), Burr and Pearson are found to be the next best alternative distribution followed by Lognormal. This order is more or less similar for the KS test where Lognormal has the second rank after FatigueLife(3P).

Nevertheless, these distributions have attained a satisfactory level of fitness to the data, the highest reported frequency for the alternative distribution is much less than the frequency attributed to the FatigueLife(3P) highlighting the fact that the second distribution does not constantly outperform in all cycles and time classes. Moreover, the higher frequency for Pearson distribution was attained by two members of this family. Type V which is a three-parameter distribution represented by a curve and Type VI which is considered as a region between Gamma and Type V.

| Distribution | AD test | KS test |
|---|---|---|
| FatigueLife(3P) | 11/12 | 12/12 |
| Burr | 7/12 | 5/12 |
| Pearson | 7/12 | 6/12 |
| Lognormal | 5/12 | 7/12 |

Table 5: The number of times that the best-performed distributions have been recorded in the 12 different studied classes (cleavage cycle 10 to 14(8)).

# 4 Conclusion

Bcd gradient provides *Drosophila melanogaster* embryonic tissues with positional information by inducing target genes at different concentration thresholds. This gradient has been analysed quantitatively using different biophysical models such as steady-state and nuclear trapping models which mainly rely on production, diffusion and degradation, but no model has considered studying all its characteristics by analysing the distribution function of the gradient [6]. Here, we discuss how existing data on Bcd gradient fit different distribution functions and which function has the superior performance in explaining the features of this gradient.

To that aim, we evaluated the performance of fifty-four different distributions in describing the gradient of the Bcd protein molecules in *Drosophila melanogaster* embryos.

Having performed a comprehensive simulation study on all possible expression patterns with different levels of noise, our results suggest that FatigueLife distribution with three parameters outperforms the other distributions.

Thereafter, by conducting the analysis on real data of 385 embryos from different cleavage cycles, the outcome found to be consistent with the simulation study suggesting the superior performance of FatigueLife(3P) in comparison with the other distributions. It is of note that among all the evaluated distributions, FatigueLife(3P) is the only one consistently present at the three top outperforming distributions and in nearly all the cleavage cycles.

Moreover, the features of the distribution function which is determined to be the most reliable one for Bcd should be consistent with the nature of the Bcd. For example, as the intensity values representing the number of protein molecules cannot be negative, it is expected that a non-negative distribution such as FatigueLife performs better in this study.

We suggest that knowing the parameters of the attained distribution function, such as the shape, scale and location parameter, would help to identify and develop a better model of Bcd functioning as an initiator of segmentation network. It is of note that the aim of this paper is not to introduce the universally best-fitted model for Bcd gradient but is to propose the idea of studying the statistical distribution of gene expression profiles when analysing the gene-gene interactions in a gene regulatory network. This study lay the necessary groundwork for our ultimate goal of modelling a dynamic segmentation network and an automated mutant recognition system. Therefore, regarding future research, it would be insightful to investigate the parameters of FatigueLife(3P) and to find the statistical distribution of the other members of this network.

# References

[1] Driever, W. and Nsslein-Volhard, C., 1988. The bicoid protein determines position in the Drosophila embryo in a concentration-dependent manner. Cell, 54(1), pp.95-104.

[2] Frohnhfer, H.G. and Nsslein-Volhard, C., 1986. Organization of anterior pattern in the Drosophila embryo by the maternal gene bicoid. Nature,

324, pp.120-125.

[3] Berleth, T., Burri, M., Thoma, G., Bopp, D., Richstein, S., Frigerio, G., Noll, M. and Nsslein-Volhard, C., 1988. The role of localization of bicoid RNA in organizing the anterior pattern of the Drosophila embryo. The EMBO journal, 7(6), p.1749.

[4] Spemann, H. and Mangold, H., 2003. Induction of embryonic primordia by implantation of organizers from a different species. 1923. International Journal of Developmental Biology, 45(1), pp.13-38.

[5] Grimm, O., Coppey, M. and Wieschaus, E., 2010. Modelling the Bicoid gradient. Development, 137(14), pp.2253-2264.

[6] Wartlick, O., Kicheva, A. and Gonzlez-Gaitn, M., 2009. Morphogen gradient formation. Cold Spring Harbor perspectives in biology, 1(3), p.a001255.

[7] Reinitz, J. and Sharp, D.H., 1995. Mechanism of eve stripe formation. Mechanisms of development, 49(1), pp.133-158.

[8] Ghodsi, Z., Silva, E.S. and Hassani, H., 2015. Bicoid signal extraction with a selection of parametric and nonparametric signal processing techniques. Genomics, proteomics & bioinformatics, 13(3), pp.183-191.

[9] Jaeger, J., Blagov, M., Kosman, D., Kozlov, K.N., Myasnikova, E., Surkova, S., Vanario-Alonso, C.E., Samsonova, M., Sharp, D.H. and Reinitz, J., 2004. Dynamical analysis of regulatory interactions in the gap gene system of Drosophila melanogaster. Genetics, 167(4), pp.1721-1737.

[10] Mller, W., Hassel, M. and Grealy, M., 2015. Development and reproduction in humans and animal model species. Springer.

[11] Staller, M.V., Fowlkes, C.C., Bragdon, M.D., Wunderlich, Z., Estrada, J. and DePace, A.H., 2015. A gene expression atlas of a bicoid-depleted Drosophila embryo reveals early canalization of cell fate. Development, 142(3), pp.587-596.

[12] Waddington, C.H., 1942. Canalization of development and the inheritance of acquired characters. Nature, 150(3811), pp.563-565.

[13] Poustelnikova, E., Pisarev, A., Blagov, M., Samsonova, M. and Reinitz, J., 2004. A database for management of gene expression data in situ. Bioinformatics, 20(14), pp.2212-2221.

[14] Hengenius, J.B., Gribskov, M., Rundell, A.E., Fowlkes, C.C. and Umulis, D.M., 2011. Analysis of gap gene regulation in a 3D organism-scale model of the Drosophila melanogaster embryo. PLoS One, 6(11), p.e26797.

[15] Papatsenko, D. and Levine, M., 2011. The Drosophila gap gene network is composed of two parallel toggle switches. PLoS One, 6(7), p.e21145.

[16] Lecca, P., Ihekwaba, A.E., Dematt, L. and Priami, C., 2010. Stochastic simulation of the spatio-temporal dynamics of reaction-diffusion systems: the case for the bicoid gradient. J of Integrative Bioinformatics, 7, pp.150-182.

[17] Shahrezaei, V. and Swain, P.S., 2008. Analytical distributions for stochastic gene expression. Proceedings of the National Academy of Sciences, 105(45), pp.17256-17261.

[18] Pisarev, A., Poustelnikova, E., Samsonova, M. and Reinitz, J., 2009. FlyEx, the quantitative atlas on segmentation gene expression at cellular resolution. Nucleic acids research, 37(suppl 1), pp.D560-D566.

[19] Hassani, H. and Ghodsi, Z., 2014. Pattern recognition of gene expression with singular spectrum analysis. Medical Sciences, 2(3), pp.127-139.

[20] Ghodsi, Z., Hassani, H. and McGhee, K., 2015. Mathematical approaches in studying bicoid gene. Quantitative Biology, 3(4), pp.182-192.

[21] Surkova, S., Kosman, D., Kozlov, K., Myasnikova, E., Samsonova, A.A., Spirov, A., Vanario-Alonso, C.E., Samsonova, M. and Reinitz, J., 2008. Characterization of the Drosophila segment determination morphome. Developmental biology, 313(2), pp.844-862.

[22] Bergmann, S., Sandler, O., Sberro, H., Shnider, S., Schejter, E., Shilo, B.Z. and Barkai, N., 2007. Pre-steady-state decoding of the Bicoid morphogen gradient. PLoS Biol, 5(2), p.e46.

[23] Houchmandzadeh, B., Wieschaus, E. and Leibler, S., 2002. Establishment of developmental precision and proportions in the early Drosophila embryo. Nature, 415(6873), pp.798-802.

[24] Gregor, T., Bialek, W., van Steveninck, R.R.D.R., Tank, D.W. and Wieschaus, E.F., 2005. Diffusion and scaling during early embryonic pattern formation. Proceedings of the National Academy of Sciences of the United States of America, 102(51), pp.18403-18407.

[25] Anderson, T.W. and Darling, D.A., 1952. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics, pp.193-212.

[26] Scholz, F., 2013. adk: Anderson-Darling K-sample test and combinations of such tests.

[27] Engmann, S. and Cousineau, D., 2011. Comparing distributions: the two-sample Anderson-Darling test as an alternative to the Kolmogorov-Smirnoff test. Journal of Applied Quantitative Methods, 6(3), pp.1-17.

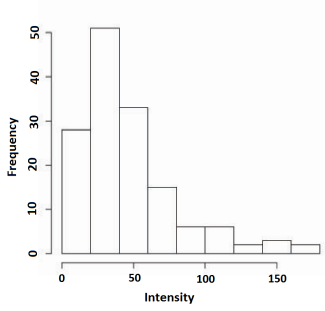
# Appendix 1. Probability Distributions

List of the probability distributions evaluated in this study:

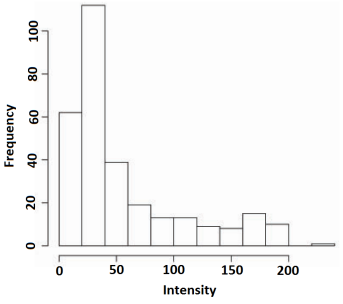
1. Beta
2. Burr (Burr Type 12, Singh-Maddala)
3. Cauchy (Lorentz)
4. Chi-Squared
5. Dagum (Burr Type 3, Inverse Burr)
6. Erlang
7. Error (Exponential Power)
8. Error Function
9. Exponential
10. F Distribution
11. Fatigue Life (Birnbaum-Saunders)
12. Gamma
13. Hyperbolic Secant
14. Inverse Gaussian
15. Johnson SB
16. Johnson SU
17. Kumaraswamy
18. Laplace (Double Exponential)
19. Levy
20. Logistic
21. Log-Gamma
22. Log-Logistic
23. Log-Pearson III (LP3)
24. Lognormal
25. Nakagami (Nakagami-m)
26. Normal
27. Pareto (first kind)
28. Pareto (second kind) (Lomax)
29. Pearson Type 5 (Inverse Gamma)
30. Pearson Type 6 (Beta Prime)
31. Pert
32. Power Function
33. Rayleigh
34. Reciprocal
35. Rice (Nakagami-n)
36. Student's t
37. Triangular
38. Uniform
39. Weibull
40. Gumbel (Extreme Value Type I)
41. Frechet (Extreme Value Type II)
42. Generalized Extreme Value (GEV)
43. Generalized Gamma
44. Generalized Pareto
45. Phased Bi-Exponential
46. Phased Bi-Weibull
47. Wakeby
48. Bernoulli
49. Binomial
50. Discrete Uniform
51. Geometric
52. Logarithmic
53. Negative Binomial
54. Poisson

# Appendix 2. The histogram of Bcd

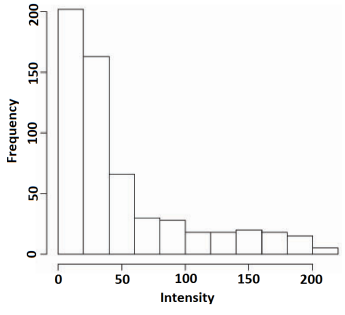
Presented below are the histograms of cleavage cycles 10-13 and all time classes of cleavage cycle 14A. It should be noted that each histogram is related to one particular embryo which has been selected as a sample representing that cycle or time class.



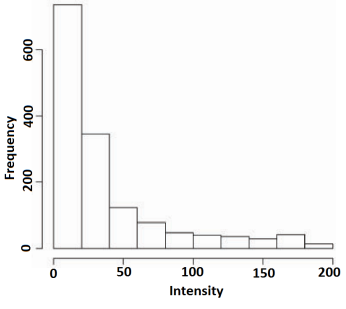

(a) Time Class 10



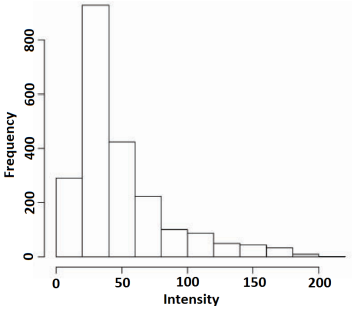

(b) Time Class 11



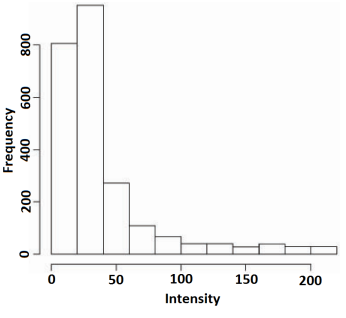

(c) Time Class 12



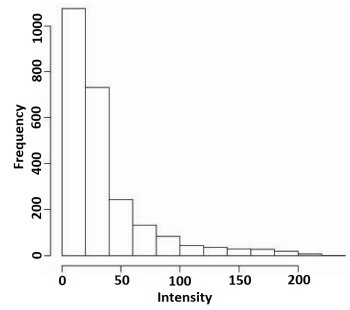

(d) Time Class 13



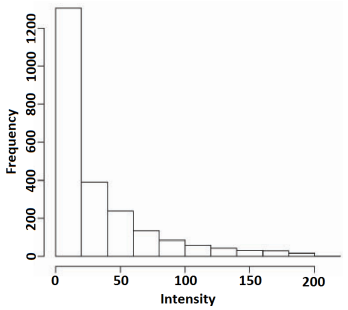

(e) Time Class 14(1)



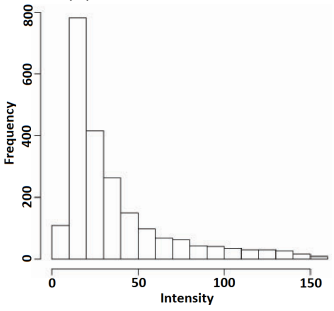

(f) Time Class 14(2)



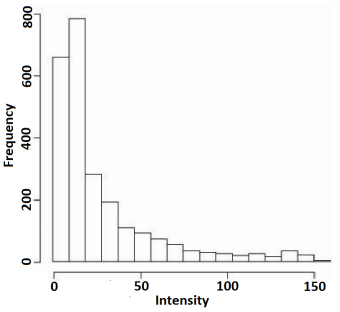

(g) Time Class 14(3)



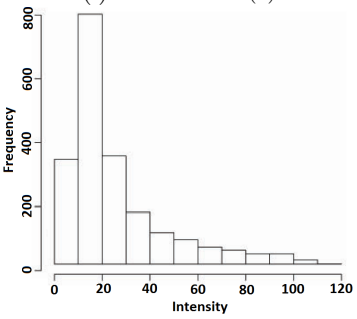

(h) Time Class 14(4)



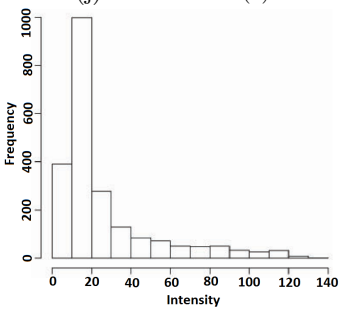

(i) Time Class 14(5)




(j) Time Class 14(6)




(k) Time Class 14(7)




(l) Time Class 14(8)

Figure 5: The histogram of Bcd related to cleavage cycles 10-13 and all the time class of cleavage cycle 14A.