A probabilistic framework for microarray data analysis: Fundamental probability models and statistical inference

Babatunde A. Ogunnaike, Claudio A. Gelmi, Jeremy S. Edwards

www.elsevier.com/locate/yjtbi

A probabilistic framework for microarray data analysis: Fundamental probability models and statistical inference

Babatunde A. Ogunnaike, Claudio A. Gelmi, Jeremy S. Edwards

PII: S0022-5193(10)00096-2
DOI: doi:10.1016/j.jtbi.2010.02.021
Reference: YJTBI 5874

To appear in: *Journal of Theoretical Biology*

Received date: 13 July 2009
Revised date: 7 December 2009
Accepted date: 12 February 2010

Cite this article as: Babatunde A. Ogunnaike, Claudio A. Gelmi and Jeremy S. Edwards, A probabilistic framework for microarray data analysis: Fundamental probability models and statistical inference, *Journal of Theoretical Biology*, doi:10.1016/j.jtbi.2010.02.021

A probabilistic framework for microarray data analysis: Fundamental probability models and statistical inference

Babatunde A. Ogunnaike[1§] (ogunnaike@udel.edu)

Claudio A. Gelmi[2] (cgelmi@ing.puc.cl)

Jeremy S. Edwards[3] (jsedwards@salud.unm.edu)

[1]*Department of Chemical Engineering and Delaware Biotechnology Institute, University of Delaware, Newark, DE, 19716, USA.*

[2]*Department of Chemical and Bioprocess Engineering, Pontificia Universidad Católica de Chile, Av. Vicuña Mackenna 4860, Santiago, Chile.*

[3]*Molecular Genetics and Microbiology, Cancer Research and Treatment Center, University of New Mexico Health Sciences Center, and Chemical and Nuclear Engineering, University of New Mexico, Albuquerque, NM, USA.*

[§]*Corresponding author (E-mail address: ogunnaike@udel.edu, Tel.: +1-302-831-4504, Fax: +1-302-831-1048)*

Abstract

Gene expression studies generate large quantities of data with the defining characteristic that the number of genes (whose expression profiles are to be determined) exceed the number of available replicates by several orders of magnitude. Standard spot-by-spot analysis still seeks to extract useful information for each gene on the basis of the number of available replicates, and thus plays to the weakness of microarrays. On the other hand, because of the data volume, treating the entire data set as an ensemble, and developing theoretical distributions for these ensembles provides a framework that plays instead to the strength of microarrays. We present theoretical results that under reasonable assumptions, the distribution of microarray intensities follows the Gamma model, with the biological interpretations of the model parameters emerging naturally. We subsequently establish that for each microarray data set, the fractional intensities can be represented as a mixture of Beta densities, and develop a procedure for using these results to draw statistical inference regarding differential gene expression. We illustrate the results with experimental data from gene expression studies on *Deinococcus radiodurans* following DNA damage using cDNA microarrays.

Keywords: Mixture models, Poisson distributions, Gamma distributions, Beta distributions, Gene Expression.

Introduction

Standard statistical analysis involves estimating (and drawing statistical inference about) p parameters using a data sample of size n, typically with $n >> p$. The precision of the estimates, as well as the power of the accompanying statistical tests of significance, are well-known to improve with increasing n. The converse is also true: reducing sample size n reduces estimation precision and the power of statistical hypotheses (other things being equal). The problem with microarray data analysis is well-documented: p (in this context the expression levels of thousands of genes to be estimated) is very large, while n, the number of independent replicates, from which these estimates are to be determined, is quite small. Many techniques have been proposed for extracting information about gene expression levels from microarray data, but this intrinsic characteristic of the data set (large p, small n) remains a major impediment (see, for example, Nadon and Shoemaker, 2002).

To complicate matters further, genes are known to operate, not independently as individuals, but in networks, suggesting that microarray data sets represent the collective behavior of a population best studied jointly. Still, the inference problem of primary concern is the quantitative determination of the differential expression level of individual genes represented in the data set: which ones are over-expressed, which are under-expressed—and by how much—and which are unchanged. None of the preceding factors change this primary focus, but they motivate us to consider analytical techniques that take into consideration the intrinsic characteristics of microarray data we have highlighted here. Our perspective in this paper is that by examining the entire data set as an ensemble and characterizing it as such, we are able to develop an analysis technique that actually plays to the strength of microarray technology.

The development of techniques for spot-by-spot analysis of microarray data has been the subject of much research effort; and many of the key results and the still unresolved issues have been discussed in several reviews (Allison et al., 2006; Dharmadi and Gonzalez, 2004; Nadon and Shoemaker, 2002; Quackenbush, 2001; Sebastiani et al., 2003).

A multitude of techniques have been proposed for testing for differential gene expression, each incorporating varying degrees of statistical rigor, ranging from the very simple (for example, based on ratios, or the so-called "fold-change" criterion), to the more sophisticated and computationally demanding approaches. One of the simplest, and most intuitive techniques for analyzing microarray data uses the "fold change" criterion to decide if the changes observed in a gene's mRNA-expression level between two distinct experimental conditions are truly significant or not. It is based on the ratio of measured expression levels, with significance indicated when this ratio (the so-called "fold change") exceeds a pre-specified threshold, typically 2.0. "Fold-change" analysis is an example of a gene-by-gene approach where expression data for genes on a microarray are treated and analyzed as unrelated individual measurements. Analyses based on ANOVA (Cui and Churchill, 2003; Kerr and Churchill, 2001) and *t*-test (Cui and Churchill, 2003; Troyanskaya et al., 2002) are other examples of gene-by-gene approaches.

An attractive alternative to the gene-by-gene approaches is provided by the class of techniques based on mixture models where the genes represented on the microarray are considered as consisting of two or more populations. In doing so, these techniques take advantage of the vast number of genes on any given array and explicitly recognize the fact that these genes are not

isolated entities: the expression level of a specific gene should affect, or share information with, its biological neighbor. Examples of mixture model approaches include mixture models for gene effect (Lee et al., 2000); empirical Bayes analysis (Efron et al., 2001; Kendziorski et al., 2003; Newton et al., 2001); fits of various probability densities to intensity ratios (Ghosh and Chinnaiyan, 2002), or to raw intensities (Hoyle et al., 2002; Steinhoff et al., 2003); and hierarchical mixture methods (Ghosh, 2004; Newton et al., 2004). Also several authors have used Beta distributions to analyze microarray data (Allison et al., 2002; Baggerly et al., 2001; Ji et al., 2005), mainly because of its mathematical convenience, not for reasons of biological plausibility. These methods are all based on empirical approximations fitted to observed data, with little or no possibility for mechanistic interpretation of the resulting distributions.

The objective of this paper is to present first-principles theoretical (as opposed to merely empirical) results for representing and characterizing microarray data as an ensemble. These results are then assembled into a theoretical framework for: (i) appropriately analyzing microarray data sets on the basis of this ensemble characterization; and (ii) drawing realistic inferences from the analysis. The proposed theoretical framework is them illustrated with an example experimental data set and in simulation. The experimental validation of the complete inference procedure will be presented in an upcoming publication.

A theoretical model of microarray measurements

Intensity measurements and gene expression levels

For each gene represented by a spot on a microarray, the expression level λ, corresponding to the number of cDNA molecules, is considered fixed, but unknown. It is estimated by measurements of fluorescent dye intensity.

For a fixed expression level, repeated fluorescent signal intensity measurements taken n separate and independent times, $I_j;\ j = 1,2,...,n$, will differ because of inherent variability. For fluorescent intensity measurements acquired from cDNA microarrays, it is typical to represent this fact as follows (Sebastiani et al., 2003):

$$I_j = a_j \lambda \tag{1}$$

where the uncertainty in the measurement is considered as having arisen primarily from a_j being a random proportionality constant, characteristic of the measurement device, relating the number of cDNA molecules to the observed intensity. In reality, it is more accurate to represent the raw intensity signal as:

$$I_j^{'} = a_j \lambda + b_j \tag{2}$$

where b_j is another random component corresponding to background noise; *i.e.* the intensity measurement values one would obtain even when there are no cDNA molecules present on the spot (Rocke and Durbin, 2001). Because of the non-negativity of intensity measurements, b_j has a non-zero mean value that is typically subtracted off to obtain the background-corrected

intensity measurement. Different technologies employ different techniques for such background correction. We note here that for high and medium intensity measurements, the contributions from b_j are relatively insignificant; appropriately accounting for the presence of this component is critical however at very low intensity measurements.

In what follows, we assume that the intensity signal has been appropriately background corrected so that Eq. (1) applies. Nevertheless, we will still pay due respect to the intensity magnitude.

Model development

Let I represent the observed fluorescent intensity corresponding to a total of z cDNA molecules; furthermore, we consider this intensity to be cumulative in the following sense:

$$I = I^{(1)} + I^{(2)} + \ldots + I^{(z)} \tag{3}$$

where $I^{(m)}$ is the intensity contribution from the mth molecule. We now refer to $z(I)$ as the number of molecules corresponding to the observed (background-corrected) total intensity measurement I. From this perspective, for a given observed I, z is an unknown random variable, an integer count.

The complementary element in this model development is the fluorescent intensity measurement obtained from k molecules, which, for purposes of this derivation, we will refer to as $y(k)$. From this intensity measurement perspective, given a fixed number k, y is an unknown (continuous) random variable.

Finally, let β represent the mean intensity observed per molecule, or conversely, its reciprocal, $\eta = 1/\beta$, represents the mean number of molecules per unit observed intensity. These quantities, characteristic of the measurement device (akin to a calibration factor), can be used to characterize the random behavior of both the continuous y and the discrete z and relate one to the other as we now show.

The problem at hand is to develop a probability model to characterize y as a function of k, the number of molecules responsible for the observed intensity signal measurement. We adopt a 2-step strategy: (i) begin with the complementary problem of characterizing $z(I)$; the discrete count of the number of molecules corresponding to the observed intensity, and (ii) convert the integer count to the continuous variable, y.

1. Characterizing *z(I)*. First-principles analysis and experimental confirmation (Ozbudak et al., 2002; Thattai and van Oudenaarden, 2001) have established that the steady state statistical properties of mRNA expression levels results in a Poisson distribution for z, *i.e.*

$$f(z) = e^{-\theta}\frac{\theta^z}{z!} \tag{4}$$

where θ is the mean number of molecules. Within the context of our current formulation, in relation to the observed intensity I, and given the characteristic parameter η as described above, Eq. (4) becomes:

$$f(z) = e^{-(\eta I)} \frac{(\eta I)^z}{z!} \tag{5}$$

which is now the probability model for characterizing *z*, the number of mRNA molecules in terms of *I*, the corresponding observed total intensity. Equation (5) provides an expression for computing the probability that *z*, the number of molecules responsible for generating the observed intensity *I*, takes on the integer values 0, 1, 2,…, ∞ given the population parameter η;

2. Convert count variable *z(I)* to continuous variable *y*. Let $y(k)$ be the total observed intensity for k molecules, *i.e.*

$$y(k) = I^{(1)} + I^{(2)} + \ldots + I^{(k)} \tag{6}$$

As we show in Appendix A, the probability distribution function for *y* given the probability distribution function for *z(I)* is:

$$f(y;\alpha,\beta) = \frac{1}{\beta^{\alpha}\Gamma(\alpha)} e^{-y/\beta} y^{(\alpha-1)} \tag{7}$$

The implications of this result are that the total fluorescent intensity measurement for each spot on the microarray possesses a Gamma distribution with the shape parameter α related directly to the number of cDNA molecules responsible for the observed intensity, y; the scale parameter β is the mean intensity per molecule.

It is important to note that Newton et al. (2001), invoked this model, strictly for convenience, but with a constant shape parameter. Also, Sebastiani et al. (2003), alluded to the fact that the Gamma distribution may be a more appropriate model for intensity data since a logarithmic transformation still leaves the data with unsatisfactory asymmetry.

A theoretical model of differential expression

At each spot on the microarray, the expression level of each represented gene is evaluated under two different conditions: 1 = "Test" conditions, 0 = "Reference" conditions (one labeled with Cy3 dye, the other with Cy5 dye). Let I_{i1} and I_{i0} be the fluorescent signal intensities measured from each spot i on a microarray, with I_{i1} obtained from the gene in question under test conditions, and I_{i0} from the gene under control conditions. Let λ_{i1} and λ_{i0} respectively represent the corresponding actual, but unknown, expression levels of the gene in question.

Then from Eq. (1), we have:

$$I_{i1} = a_{i1}\lambda_{i1} \tag{8}$$

$$I_{i0} = a_{i0}\lambda_{i0} \tag{9}$$

For various well-documented reasons (Dudley et al., 2002; Sebastiani et al., 2003) it is a near-impossible task to determine the absolute value of the expression levels λ_{i1}, λ_{i0} strictly from the observed microarray intensity measurements. Even so, determining by how much λ_{i1} differs from λ_{i0} without necessarily knowing the absolute value of each number is still considered extremely valuable information about the gene in question. While methods for determining absolute gene expression levels are being perfected for lower cost and higher-throughput (for

example, (Brenner et al., 2000; Mikkilineni et al., 2004) differential expression determination via standard microarray technology remains an important aspect of gene expression studies.

The customary measure of differential expression is the so-called fold change statistic defined as the ratio:

$$\rho_i = \frac{\lambda_{i1}}{\lambda_{i0}} \tag{10}$$

and the common practice is to base inference about ρ_i on the measured intensity signal ratio r_i defined as

$$r_i = I_{i1} / I_{i0} \tag{11}$$

It is important to observe from here however, that such a strategy is based on the (tacit) assumption that for both the Cy3 label signal intensity as the with the Cy5, the multiplicative factors are approximately the same. This may not be the case in general, especially since we have just shown that these multiplicative constants are related directly to the "mean intensity observed per molecule" a number that is clearly different for Cy3 and Cy5 fluorescent dye labels (This is potentially one source of the well-documented dye bias). Dye-swaps and data normalization are some of the tactics typically employed to minimize these problems.

Also, observe that if σ_a^2 is the variance associated with the random multipliers, then the observed variance in the signal intensity will be given by:

$$\sigma_I^2 = \lambda^2 \sigma_a^2 \tag{12}$$

so that the variance associated with signal intensity increases quadratically with expression level (and hence with intensity); or equivalently, the standard deviation varies linearly with intensity. This is in agreement with actual microarray data (Rocke and Durbin, 2001). For example, Figure 1 shows a plot of experimental data from gene expression studies in *Deinococcus radiodurans* following DNA damage (Tanaka et al., 2004) where the data variance is observed to increase with increasing intensity values. The implication for data analysis is that the statistical character of microarray measurements (especially measurement variance) is dependent on signal magnitude (Chen et al., 1997; Newton et al., 2001).

The most commonly employed strategy for addressing the preceding issue is a logarithmic transformation of the data to obtain:

$$\log I_{i0} = \log a_{i0} + \log \lambda_{i0} \tag{13}$$

turning the expression into the more familiar additive error form:

$$Y_i = X_i + \varepsilon_i \tag{14}$$

In most cases, it is now assumed that $\log(a_{i0})$ is normally distributed, and the analysis proceeds along these lines. However, it has been shown by various investigators (Rocke and Durbin, 2001; Sebastiani et al., 2003) that this assumption, even though convenient, is not necessarily valid. It would be valid if a lognormal density is an appropriate model for the intensity distribution.

However, we have just established that the Gamma density is more appropriate, not the lognormal.

An alternative transformation

The fan-shaped characteristics of raw cDNA microarray intensity data (illustrated by the example in Figure 1) suggests a strategy where the original data in the form of the ordered pair of measured signal intensities (y_{i1}, y_{i0}) for each spot i is converted from this "Cartesian" representation to the corresponding "polar" form (R_i, x_i) where R_i, the vector magnitude defined by:

$$R_i = \sqrt{y_{i1}^2 + y_{i0}^2} \tag{15}$$

is the "intensity magnitude", and x_i, the "fractional intensity" defined by:

$$x_i = \frac{y_{i1}}{y_{i1} + y_{i0}} \tag{16}$$

The following are important attributes of this coordinate transformation: (i) it naturally compensates for the inherent heteroskedasticity of microarray data since pie-shaped regions in Cartesian coordinates naturally map into rectangular regions in polar coordinates. For example, Figure 2 shows the effect of this transformation on the data in Figure 1; (ii) it normalizes the data efficiently without losing any of the original information. Observe that while x_i is dimensionless and naturally scaled between 0 and 1, R_i retains the information about intensity magnitude, thus preserving the two dimensional character of the original data; (iii) the characteristics of x_i are

such that values less than 0.5 indicate down regulation; $x_i = 0.5$ indicates no change and $x_i > 0.5$ indicates up-regulation.

Even though the traditional intensity ratio r is related to x_i according to the expression.

$$r_i = \frac{y_{i1}}{y_{i0}} = \frac{x_i}{1 - x_i} \tag{17}$$

we recommend x_i as a more convenient variable on which to base inference about differential expression for the following reasons: a probabilistic model can be derived for x_i from the model (as we will show shortly) while R_i provides a rational metric for partitioning the data (allowing us to separate low magnitude data from higher magnitude data).

Probability model for fractional intensity

If we represent as y_{i1}, the measured intensities, corresponding to gene i (on spot i) under test conditions, and y_{i0}, the value obtained for the same gene i under reference conditions, then the implications of the results given above are that:

$$y_{i1} \sim Gamma(\alpha_{i1}, \beta_1) \tag{18}$$

$$y_{i0} \sim Gamma(\alpha_{i0}, \beta_0) \tag{19}$$

allowing the equipment calibration parameter (mean intensity per molecule), β, to be dye dependent, but not gene dependent, so that we need only retain the subscripts 0 and 1 but not i. A well-known result from probability theory (*e.g.*, Evans et al., 1993) is that if Eqs. (18) and (19) are true then:

$$y_{i1} / \beta_1 \sim \ Gamma\left(\alpha_{i1}, 1\right) \tag{20}$$

$$y_{i0} / \beta_0 \sim \ Gamma\left(\alpha_{i0}, 1\right) \tag{21}$$

and the fractional variable, $\widetilde{x}_i$ defined by

$$\widetilde{x}_i = \frac{y_{i1} / \beta_1}{y_{i1} / \beta_1 + y_{i0} / \beta_0}$$

$$= \frac{y_{i1}}{y_{i1} + w y_{i0}} \tag{22}$$

where

$$w = \frac{\beta_1}{\beta_0} \tag{23}$$

possesses a Beta distribution, Beta(α_{j1}, α_{j0}), with a probability density function given by:

$$f(x) = \frac{\Gamma\left(\alpha_{i0} + \alpha_{i1}\right)}{\Gamma\left(\alpha_{i1}\right)\Gamma\left(\alpha_{i0}\right)} x^{\alpha_{i1}-1}\left(1-x\right)^{\alpha_{i0}-1} \tag{24}$$

Some of the consequences of these results for practical data analysis are now discussed.

Nominal data transformation

A subtle but important point about Eqs. (22) and (23) is that raw data consists of y_{i1}, and y_{i0} pairs only; the parameters β_1 and β_0 are unknown calibration parameters. In fact, this latter point is the reason why absolute estimates of expression levels are not possible from

microarray data. As noted earlier, the rationale for using the intensity ratios as the basis for drawing inference about differential expression is predicated on the tacit assumption that:

$$\beta_0 = \beta_1 \tag{25}$$

We will designate these customary assumptions as representing what we refer to as nominal conditions. Note that in reality these conditions may in fact not hold. The implications of departure from these nominal conditions, how to detect and correct for such departures will be discussed shortly.

For now, note that under these conditions, $w = 1$ and Eq. (22) indicates that the nominal fractional intensity defined by:

$$x_i = \frac{y_{i1}}{y_{i1} + y_{i0}} \tag{26}$$

obtained directly from raw data, possesses a Beta distribution, so that for each gene, the probability model is as shown in Eq. (24).

When nominal conditions do not hold, then in reality the computed fractional intensity x_i ought to be corrected to obtain $\tilde{x}_i$. The relationship between the corrected variable and the nominal fractional intensity x_i is easily established as the following dual expression:

$$\tilde{x}_i = \frac{x_i}{x_i + w(1 - x_i)} \tag{27}$$

and

$$x_i = \frac{w\tilde{x}_i}{w\tilde{x}_i + (1-\tilde{x}_i)} \tag{28}$$

Detecting and correcting for departure from nominal conditions

When nominal conditions do not hold, (*i.e.* $w \neq 1$) it is necessary to estimate w, and this is done as follows. Under conditions when there is no change in gene expression, the expected value of the nominal fractional intensity data, $E(X) = \mu_0$ should be theoretically equal 0.5. However, when $w \neq 1$ in (22), even though there is no change in gene expression, $E(X) \neq 0.5$, and hence any significant deviation of estimates of μ_0 from 0.5 provides an indication that nominal conditions do not hold. Given a value of $E(X) = \mu_0 \neq 0.5$, we are able to use Eq. (28) to estimate the value of w required so that the expectation of the corrected intensity $\tilde{x}_i$ is 0.5. It is easy to show that under these conditions:

$$\mu_0 = \frac{w}{w+1} \tag{29}$$

so that

$$w = \frac{\mu_0}{1-\mu_0} \tag{30}$$

Thus, from an estimate of the expected value of the nominal x for those genes showing no change in gene expression, we are able (i) to determine if nominal conditions hold and then (ii) correct the nominal fractional intensity for the entire data set using the estimate of w given above when nominal conditions do not hold.

Application to microarray data analysis

The theoretical implications of the results presented in Eqs. (24) and (26) are as follows: for a particular microarray dataset involving a total of N genes, each nominal fractional intensity $x_i = 1,2,\ldots,N$ is a random sample from its own (possibly unique) Beta distribution. Consequently, $f(x)$, the underlying theoretical distribution of (nominal) spot fractional intensities for the entire collection of genes, will be a mixture of overlapping Beta density functions consisting, in principle, of as many Beta densities as there are genes.

Because of the finite support [0,1] over which $f(x)$ is defined and the large number of genes involved, but most importantly as a result of the necessary "coalescence" of the distributions of genes with theoretically identical—or statistically similar—expression levels, in practice, the actual number of distinct and truly discernible Beta densities required to describe any specific collection of microarray data will be relatively few. Practically therefore, this suggests a characterization strategy in which the histogram of the (nominal) fractional intensities is modeled as a mixture of just a few Beta densities.

In particular, the distributions of all genes showing no differential expression will coalesce into one clearly discernible distribution, $f_0(x)$. If, at the barest minimum, the distributions of genes showing lower differential expression also coalesce into a single discernible distribution and the same happens for the genes showing higher differential expression, we may thus characterize the histogram of fractional intensity data (at the barest minimum) as a mixture of (at least) three Beta distribution functions, $f_1(x)$ for the data from the collection of genes

showing lower differential expression, $f_0(x)$ for those showing no differential expression, and $f_2(x)$ for those showing higher differential expression (Eq. (31)):

$$f(x) = \phi_1 f_1(x) + \phi_0 f_0(x) + \phi_2 f_2(x) \tag{31}$$

with $\phi_0 + \phi_1 + \phi_2 = 1$.

Alternatively, and more generally, this exercise may be considered as akin to an expansion of $f(x)$ in a set of basis functions, $f_i(x)$, that are beta densities. The total number of basis functions used to represent the true function depends on the level of desired "accuracy". The specific choice of 3 basis functions in Eq. (31) is based on the arguments advanced above: each of these functions can be related *directly* to the status of the gene in question.

Thus by fitting a mixture of Beta distributions to histogram data, we obtain a theoretical probabilistic characterization of the data where each contributing Beta distribution constitutes the population description of a category of genes with common differential expression attributes. Observe that the coefficients ϕ_0, ϕ_1 and ϕ_2 indicate the fraction of the entire population that belongs in each respective indicated category. In particular, for the fraction ϕ_0 of genes for which there is truly no differential expression, x possesses a Beta distribution that is perfectly symmetric, with mean = mode = median = $x^* = 0.5$. In some cases, it may be necessary to represent the data histogram with more than three Beta densities. Under such circumstances, the appropriate number of Beta densities required to represent the histogram structure adequately may be determined by employing such standard model selection criterion

as the Akaike's Information Criterion (AIC) or the Bayesian Information Criterion (BIC) in selecting among various possible models involving $n = 4, 5, 6, \ldots$ individual densities in the mixture model.

Finally, we note that Diaconis and Ylvisaker, 1985, have showed that *any* data distribution on the interval [0, 1] can be modeled as a finite mixture of Beta distributions, a theoretical result whose validity has been illustrated in a wide range of practical applications (Allison et al., 2002; Parker and Rothenberg, 1988; Skliar et al., 2007). In the current context, the primary implication of this result is that the probabilistic framework we have presented thus far does in fact extend beyond cDNA microarray data: data from Affymetrix GeneChips, and *any other* two-channel or one-channel array technology (old or new), once expressed in the form of fractional intensities, can also be analyzed using the results presented here.

Statistical Inference: Probability of expression status

The mixture model in Eq. (31) can now be used as the reference distribution for drawing rigorous statistical inference about the expression status of each gene as follows. The probabilities that each gene g_i with associated fractional intensity x_i is up-regulated, not differentially expressed, or down-regulated may be computed from Eq. (1), respectively as:

$$P_{up,i} = \mathrm{P}(g_i \in 2 \mid X{=}x_i) = \frac{\phi_2 \cdot f_2(x_i)}{f(x_i)} \tag{32}$$

$$P_{non,i} = \mathrm{P}(g_i \in 0 \mid X{=}x_i) = \frac{\phi_0 \cdot f_0(x_i)}{f(x_i)} \tag{33}$$

$$P_{down,i} = \mathrm{P}(g_i \in 1 \mid X{=}x_i) = 1 - P_{up,i} - P_{non,i} \quad (34)$$

These probabilities provide quantitative measures of the evidence contained in the microarray data regarding the true expression status of each gene. However, since each probability is itself computed from data x_i, the uncertainty inherent in the data is transmitted to these computed probabilities and must be quantified. As part of the inference framework we propose the following technique for quantifying these uncertainties and for converting these to a "confidence index" —a measure of the confidence associated with each computed probability.

Confidence index associated with the probability of expression status

Confidence limits around statistical estimates are generated from measures of uncertainty in the data employed for computing the estimates in the first instance; and naturally, such uncertainty measures are determined from replicates or other independent measures of pure variability. Microarray data are atypical in the sense that pure replicates, when available at all are often limited in number. Therefore, in recognition of the fact that there are very few situations where microarray data are generated from unreplicated experiments (Reid and Fodor, 2008), we treat this case first before dealing with the most common and important case, where the data are available in (limited) replicates.

***From unreplicated data*:** Under these circumstances, a measure of data uncertainty can be derived from considerations of the data generation technology itself. Under the assumption that

if the uncertainty in the intensity signal is proportional to the magnitude of the signal, the true but unknown values can be written as:

$$\tilde{y}_{i1} = y_{i1} \pm \gamma_1 \cdot y_{i1} \tag{35}$$

$$\tilde{y}_{i0} = y_{i0} \pm \gamma_0 \cdot y_{i0} \tag{36}$$

with y_{i1} as the actual measured intensity obtained from the gene in question under test conditions, and y_{i0} from the gene under control conditions. γ_0 and γ_1 are constants representing the extent of multiplicative uncertainty associated with the respective signal intensities. Such parameters are often supplied as part of instrument calibration characteristics.

From Eqs. (35) and (36), the true but unknown fractional intensity $\dot{x}_i$ is given by:

$$\dot{x}_i = \frac{\tilde{y}_{i1}}{\tilde{y}_{i1} + \tilde{y}_{i0}} = \frac{y_{i1} \pm \gamma_1 \cdot y_{i1}}{y_{i1} + y_{i0} \pm (\gamma_1 \cdot y_{i1} + \gamma_0 \cdot y_{i0})} \tag{37}$$

From standard propagation-of-error analysis (Taylor, 1997), and under the reasonable assumption of equal multiplicative uncertainties in each signal channel (*i.e.*, $\gamma_1 = \gamma_0 = \gamma$), the uncertainty propagation simplifies to:

$$\dot{x}_i \approx x_i \pm 2\gamma \cdot x_i \tag{38}$$

Each computed nominal fractional intensity x_i is thus located in the range $x_i - \Delta x_i < x_i < x_i + \Delta x_i$ with $\Delta x_i = 2\gamma x_i$, from which corresponding probability ranges can be computed for each gene:

($P^{-}_{down,i}$, $P^{+}_{down,i}$), ($P^{-}_{non,i}$, $P^{+}_{non,i}$) and ($P^{-}_{up,i}$, $P^{+}_{up,i}$). The superscripts '-' and '+' respectively denote the probabilities associated with the fractional intensity at the low end of the range, $x-\Delta x$, and at the high end $x+\Delta x$. The uncertainties in the fractional intensity data have thus been translated into uncertainties in the probabilities with ranges defined by $\Delta P_{down,i}$, $\Delta P_{non,i}$ and $\Delta P_{up,i}$, respectively.

Thus, associated with the nominal probabilities computed from Eqs. (32)-(34) are the uncertainties represented by the ranges of the probabilities of expression status, for any given γ. Clearly, the wider these ranges are, the greater the degree of uncertainty associated with each computed probability, reducing our confidence in these probabilities. Conversely, narrow ranges imply less uncertainty and hence increased confidence.

Because the probabilities of certain event must add up to 1, it is easy to establish that the elements of ΔP, the vector of uncertainties associated with the probabilities, must satisfy the following fundamental constraint:

$$\Delta P_{down,i} + \Delta P_{non,i} + \Delta P_{up,i} = 0 \tag{39}$$

so that a measure of the magnitude of the uncertainty vector defined by

$$S_i = \sqrt{\frac{\Delta P^2_{down,i} + \Delta P^2_{non,i} + \Delta P^2_{up,i}}{2}} \tag{40}$$

will be scaled between 0 and 1 (since the minimum value of the numerator term is 0 and the maximum value is 2). If we now define a confidence index as:

$$c_i = 1 - S_i \tag{41}$$

then observe that $0 \le c_i \le 1$, and a low value of c_i corresponds to low certainty regarding the true state of the expression status of the gene i, as indicated by the computed probabilities; conversely, a high value of c_i corresponds to a higher degree of confidence in what the computed probabilities imply about the expression status of the gene i.

***From replicated data*:** For each gene g_i, the fractional intensities computed from n replicates:

$$x_{i,1}, x_{i,2}, ..., x_{i,n} \tag{42}$$

are characterized by the usual average and the range (since typical microarray data seldom have enough replicates for determining a reliable estimate of standard deviation):

$$x_{i,\min} = \min\left(x_{i,1}, x_{i,2}, ..., x_{i,n}\right) \tag{43}$$

$$x_{i,\max} = \max\left(x_{i,1}, x_{i,2}, ..., x_{i,n}\right) \tag{44}$$

The nominal probabilities of expression status in Eqs. (32)-(34) are now computed for the average fractional intensities, and the associated uncertainty determined from the probabilities computed at the extremes, $x_{i,\min}$ and $x_{i,\max}$. The magnitude of the uncertainty vector in this case is defined as:

$$S_i = \sqrt{\frac{\Delta P_{down,i}^2 + \Delta P_{non,i}^2 + \Delta P_{up,i}^2}{d}} \tag{45}$$

where the scaling factor d depends on the number of replicates. It is easy to show that specifically, $d = 2$ if $n = 2$ (duplicates); otherwise $d = 3$ if $n \geq 3$ (triplicates or more). With S_i thus scaled between 0 and 1, the confidence index, as before, is computed using Eq. (41).

We now illustrate the application of the theoretical results we have presented thus far with an example involving experimental data.

Illustrative examples

Example 1: Experimental Data

The data set shown first in Figure 1 is from studies of approximately 3000 genes from *D. radiodurans* 0.5 hours after ionizing radiation (following DNA damage), using cDNA microarrays.

Data representation and partitioning: The polar coordinate transformed data is shown in Figure 2 where the intensity magnitude R is plotted against nominal fractional intensity x. The entire range of intensity magnitude is from $5.3 \cdot 10^4$ to $1.42 \cdot 10^7$. The data is sorted by intensity magnitude (from the lowest to the highest R value) and then partitioned into three groups containing an equal number of data points, ($5.3 \cdot 10^4 < R_{low} < 6.63 \cdot 10^5$); ($6.63 \cdot 10^5 < R_{med} < 1.27 \cdot 10^6$) and ($1.27 \cdot 10^6 < R_{high} < 1.42 \cdot 10^7$).

Nominal data characterization and model fits: A histogram of the data and the theoretical fit of mixture distributions are shown below. Figures 3 and 4 show the data and model fit for "Low *R*" data; Figures 5 and 6 for "Medium *R*", and Figures 7 and 8 for "High *R*" data. The probability model for the Low *R* group is:

$$f(x) = 0.18\, f_1(x) + 0.79\, f_0(x) + 0.03\, f_2(x) \tag{46}$$

where the Beta density parameters are, respectively, $\alpha_{11} = 64.0$, $\alpha_{10} = 98.8$, for $f_1(x)$; $\alpha_{01} = 98.9$, $\alpha_{00} = 120.2$, for $f_0(x)$; and $\alpha_{21} = 187.8$, $\alpha_{20} = 104.4$ for $f_2(x)$. In general, the parameters associated with the contributing Beta density $f_j(x)$ (for the collection of genes sharing this probabilistic description) are a_{j1}, corresponding to the effective comparative number of cDNA molecules under test conditions for the collection, and α_{j0} the counterpart value under reference conditions. The mixture model parameters can be estimated via a variety of approaches, for example, Expectation-Maximization (EM) or least squares. We chose the latter approach, the model's parameters by fitting the mixture model to the empirical cumulative distribution of the fractional intensity via least-squares.

The corresponding results for medium and high *R* values are shown in Appendix B. Note that for "High *R*" data, more than 3 Beta densities are needed to capture the indicated data characteristics. This latter observation underscores the importance of segregating the data by the intensity magnitudes. Not only are the qualitative characteristics of the data histograms in Figures 3, 5 and 7 visually different; quantitatively, the parameter values of the contributing Beta

densities, the associated mixture weights, ϕ_j's, and, in the case of the High *R* group, the actual number of Beta densities, are all different. A less discriminating analysis that lumps all the data together into one mass grouping will fail to detect the subtle characteristics of the lower intensity magnitude data which would tend to be overwhelmed by the more pronounced characteristics of the higher intensity magnitude data.

Bias detection and correction: Observe from each of the indicated $f_0(x)$ densities (either visually from the plots, or more quantitatively directly from the parameter estimates (α_{01}, α_{00})) that by definition of the mean of a Beta random variable, the mean values, given by:

$$\mu_0 = \frac{\alpha_{01}}{\alpha_{01} + \alpha_{00}} \tag{47}$$

deviate significantly from 0.5, with the immediate implication that nominal conditions do not hold for this data set. By estimating *w* for each data group, and using these to obtain bias-corrected fractional intensities $\tilde{x}_i$, a re-estimation of the mixture model now gives rise to the results shown for High *R* in Figures 9 and 10 (to conserve space, Low and Medium *R* results are omitted). The resulting model parameters are shown in Appendix B. Note the corrected characteristics of each $f_0(x)$ corresponding to the distribution of genes showing no change in gene expression.

The mixture beta distribution, *f(x),* developed above may now be used for drawing statistical inference regarding microarray fractional intensities.

Statistical inference. After fitting the mixture model to the data histogram, Eqs. (32)-(34) are then used to compute the probability that each gene with a fractional intensity x_i belongs to one of the three categories: down-regulated, up-regulated and not differentially expressed. A summary of the results is shown in Figure 11. The model fits to the data histograms are shown in the thick solid lines; the dashed lines and the thin solid lines show the respective probabilities of being down-regulated and up-regulated as a function of the fractional intensity x. For example, Fig. 11(a) and Fig. 11(b) show that the probability that genes with $x < 0.3$ are down-regulated (dashed line) is almost 1.0, while the probability that genes for which $x > 0.7$ are up-regulated (thin solid line) is also close to 1.0. And when the fractional intensity for a gene is 0.50 ($y_{i1} = y_{i0}$), not surprisingly the figures indicate that the probability that the gene in question is differentially expressed is extremely low, almost zero (for both lines).

A typical output of the analysis procedure is displayed in Table 1: the fractional intensity x and the corresponding probabilities of expression status, are shown for the minimum, maximum and average values of x. The last column of the table is the confidence index (c_i).

There are several ways in which the researcher may use the computed probabilities and the confidence index. For example, Figure 12 shows a histogram of the confidence index values associated with the computed probabilities of expression status. This provides a means of quantifying the overall "quality" of the experimental data. For instance note that the figure indicates that a substantial number (76.3%) of the analyzed genes have a confidence index greater than 0.80 associated with their computed probabilities of expression status —implying relatively low variability associated with the data in general.

However, the most important use of the computed results is in identifying and selecting differentially expressed (down-regulated or up-regulated) genes. By using as the cut-off criteria, threshold values for P_{down} or P_{up} appropriate to the study at hand, it is possible to select candidate genes satisfying the stipulated criteria. For example, to generate a list of down-regulated candidate genes, choosing a threshold value of 0.75 for P_{down} (indicating a desire to select only those genes that have a probability greater than or equal to 0.75 of being down-regulated, regardless of the precision associated with the computed probability) results in a list containing 218 out of 3,094 genes.

By imposing a threshold value on the precision of the computed probabilities, we can further narrow down the search for potentially down-regulated genes. Thus, for example, imposing a cut-off value of $c_i \geq 0.90$ reduces the number of candidate down-regulated genes from 218 to 129 genes. The researcher may now use a more precise, higher resolution experimental technique to determine the true expression status of this smaller subset of selected genes.

Example 2: A Simulation study

The purpose of this example is to complement Example 1 by assessing the performance of our approach in detecting differentially expressed genes when the "true" circumstances are known.

The study involves five microarray datasets simulated with statistical characteristics similar to the ones observed in real microarray data. Each simulated microarray dataset consisted of 3,000 genes with 750 genes differentially expressed: 300 genes were down-regulated (10% of the

genome), 450 genes were up-regulated (15% of the genome) and 2,250 genes were not differentially expressed (75% genome). The parameters of the Beta distributions were chosen to mimic very closely actual distributions of real microarray data (after the polar transformation) as observed in our laboratory. The main statistics of the three Beta distributions used to simulate the microarray data are summarized in Table 2.

Based on the probabilities of expression status, the top 300 down-regulated and the top 450 up-regulated genes were chosen as differentially expressed. Table 3 summarizes the total number of genes correctly identified as a function of the number of simulated microarrays (replicates). Next, the simulated microarray data was also analyzed using the nonparametric method Significance Analysis of Microarrays (SAM) (Tusher et al., 2001), a popular technique that does not require any assumption about the distribution of the microarray data. SAM does not produce probabilities; rather, it generates a sorted list of candidate genes given a specified False Discovery Rate (FDR). (The FDR associated with a decision rule is the proportion of false positives among all the genes initially identified as being differentially expressed (Wit and McClure, 2004)). The total number of genes correctly identified by SAM (with FDR = 3%) is also shown in Table 3 as a function of the number of available replicates.

We observe first, that the performance of the probabilistic approach is uniformly better than that obtained using SAM. This should come as no surprise: statistical inference based on distribution-free techniques general tend to be less powerful when the underlying data come from known distributions. Notwithstanding, the performance obtained with SAM is still quite good although it is unable to handle the case where no replicates are available. Second, and, again, not

surprisingly, both techniques show improved performance as the number of replicates increase. In particular, it is interesting to observe that with five replicates, the probabilistic approach identified 744 of the original 750 differentially expressed genes, a virtually-perfect precision of 99.6%; with SAM, 710 of the original 750 differentially expressed genes were identified, a still-impressive precision of almost 95%.

Summary and Conclusions

The three main results of this paper are now summarized:

1. The distribution of spot intensity follows a Gamma model as indicated in Eq. (7): the shape parameter, α, is related directly to the number of cDNA molecules responsible for the observed intensity, and the scale parameter, β, is the mean intensity per molecule–an equipment calibration parameter.
2. For modeling differential expression, a polar transformation of the raw spot intensity data (as in Eqs. (15) and (16)) yields the intensity magnitude R_i, (to be used for rationally segregating the data) and the fractional intensity x_i, which possesses a Beta distribution as in Eq. (24); the distribution parameters, α_{i1} and α_{i0} are related directly to the number of cDNA molecules under test and reference conditions, respectively.
3. The distribution of entire microarray fractional intensities is characterized, theoretically, by a mixture of overlapping Beta densities; in practice (and as demonstrated with an example) as a result of the coalescence of distributions of genes with similar expression levels, the discernible features of the histogram of real data may be captured by as few as three overlapping Beta densities, as shown in Eq. (31).

This last result has the most direct impact on the quest for reliable determination of differential expression for thousands of genes from noisy, potentially biased, microarray data, with few or no replicates. It indicates that, $f(x)$, the underlying probability density function of the fractional intensity, is a linear combination of Beta density basis functions providing the following important perspective of the data set: each contributing Beta density may be viewed as a

coalesced population description of a category of genes with common differential expression attributes. The resulting coefficients of the component Beta densities therefore indicate what fraction of the entire data set belong to the category represented by the Beta density in question. For example, for the case shown in Figures 3 and 4, with the model fit shown in Eq. (31), the implications are the following: the contributions from the class of potentially down-regulated genes, is about 18%; the contribution from the class of genes whose expression levels are potentially unchanged is 79%; the remaining 3% is the complement, the contribution from potentially up-regulated genes.

In conclusion, we observe that just as classical statistical inference is based on theoretical reference distributions (such as the Gaussian, *t*-, Chi-square, or the F- distributions) we have developed a methodology for drawing statistical inference using the mixture distribution *f(x)* as its theoretical basis: it consists of (i) a probability statement that a gene belongs to a category (the down-regulated, the up-regulated, or the no-change) and (ii) a degree of confidence associated with such probability statements, determined from the variability estimated from replicates, or else by propagation-of-error techniques when there are no replicates. The final outcome is an ordered triplet of results for each gene: a raw computed fractional (or relative) change in expression level, an associated probability that this number indicates lower, higher, or no differential expression (a category-membership probability) and a measure of confidence associated with the stated result. This analysis technique, which has been illustrated with a real experimental data set and also via simulation, provides an alternative approach specifically tailored to the inherent characteristics of microarray data.

In a future publication we will present an experimental validation of this probabilistic framework using data collected in our laboratory.

Acknowledgements

We would like to thank Dr. John R. Battista of the Louisiana State University for providing us the data on *D. radiodurans*. BAO and CAG acknowledge funding from the Delaware Biotechnology Institute; JSE acknowledges funding from the US Department of Energy Grant # DE-FG02-O1ER63200 and a subcontract grant from Genomatica.

Appendix A: Derivation of the pdf in Eq. (7)

To derive a probability distribution function for y as defined in Eq. (6), given the probability distribution function for $z(I)$, we begin by observing that $z(I)$ and $y(k)$ are related according to the following expression:

$$P[z(I) < k] = P[y(k) > I] \tag{48}$$

which, from the definition of the variables z and y, translates in words to: "if the z-count is not yet up to k then the observed intensity I must be less than y". In the language of probability, the event that z is less than k is equivalent to the event that the observed intensity I corresponding to z is less than y, and the probability of one event is identical to that of the other. From Eq. (48) we obtain

$$P[z(I)<k]=P[z(I)\le k-1]=\sum_{z=0}^{k-1}f(z) \tag{49}$$

which upon introducing Eq. (5). gives

$$P[z(I)<k]=\sum_{z=0}^{k-1}e^{-\eta I}\frac{(\eta I)^z}{z!} \tag{50}$$

The key result at this point is the following identity for the RHS of Eq. (50) above:

$$\sum_{z=0}^{k-1}e^{-\eta I}\frac{(\eta I)^z}{z!}=\frac{1}{(k-1)!}\int_{\eta I}^{\infty}e^{-z}z^{k-1}dz \tag{51}$$

To establish this result, first let

$$\tau(k)=\int_{a}^{\infty}e^{-z}z^{k-1}dz \tag{52}$$

from where it is easy to show via integration by parts, that

$$\tau(k)=a^{k-1}e^{-a}+(k-1)\cdot\tau(k-1) \tag{53}$$

a recursion equation that leads immediately, upon successive substitution and rearrangement, to

$$\tau(k) = (k-1)! \sum_{m=0}^{k-1} e^{-a} \frac{a^m}{m!} \tag{54}$$

establishing the result in Eq. (51).

We may now use this result in Eq. (48) to obtain:

$$\begin{aligned} P[y(k) > I] &= 1 - P[y(k \le I)] \\ &= \frac{1}{(k-1)!} \int_{\eta I}^{\infty} e^{-z} z^{k-1} dz \end{aligned} \tag{55}$$

or, from the definition of the cumulative distribution $F_y(I)$,

$$1 - F_y(I) = \frac{1}{(k-1)!} \int_{\eta I}^{\infty} e^{-z} z^{k-1} dz \tag{56}$$

Differentiating with respect to I using Leibnitz's rule for differentiating under the integral, we obtain

$$\frac{\partial F_y}{\partial I} = f(y) = \frac{\eta^k}{\Gamma(k)} e^{-\eta y} y^{(k-1)} \tag{57}$$

which may now be written in the familiar form:

$$f(y;\alpha,\beta)=\frac{1}{\beta^{\alpha}\Gamma(\alpha)}e^{-y/\beta}y^{(\alpha-1)} \tag{58}$$

as shown in Eq. (7).

Appendix B: Table of complete model results

For each data group, segregated according to the intensity magnitude *R,* the parameters (α_{j1}, α_{j0}) for the contributing probability densities, $f_j(x)$, and the associated mixture coefficient weights, ϕ_j (j = 0, 1, 2, 3) are shown in Table 4 for both nominal and corrected conditions. For each group, note the inequality of the nominal parameters (α_{01}, α_{00}) associated with $f_0(x)$; the indicated bias is eliminated upon correction.

Note also that bias correction has little or no effect on the mixture coefficient weights for Low and Medium *R* data, the influence of bias correction on these weights is more noticeable for the High *R* group.

References

Allison, D.B., Cui, X.Q., Page, G.P., and Sabripour, M., 2006. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics 7, 55-65.

Allison, D.B., Gadbury, G.L., Heo, M.S., Fernandez, J.R., Lee, C.K., Prolla, T.A., and Weindruch, R., 2002. A mixture model approach for the analysis of microarray gene expression data. Computational Statistics & Data Analysis 39, 1-20.

Baggerly, K.A., Coombes, K.R., Hess, K.R., Stivers, D.N., Abruzzo, L.V., and Zhang, W., 2001. Identifying differentially expressed genes in cDNA microarray experiments. Journal Of Computational Biology 8, 639-659.

Brenner, S., Johnson, M., Bridgham, J., Golda, G., Lloyd, D.H., Johnson, D., Luo, S., McCurdy, S., Foy, M., Ewan, M., Roth, R., George, D., Eletr, S., Albrecht, G.,

Vermaas, E., Williams, S.R., Moon, K., Burcham, T., Pallas, M., DuBridge, R.B., Kirchner, J., Fearon, K., Mao, J., and Corcoran, K., 2000. Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18, 630-634.

Chen, Y., Dougherty, E.R., and Bittner, M.L., 1997. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Journal of Biomedical Optics 2, 364-374.

Cui, X., and Churchill, G.A., 2003. Statistical tests for differential expression in cDNA microarray experiments. Genome Biology 4, 210.

Dharmadi, Y., and Gonzalez, R., 2004. DNA microarrays: Experimental issues, data analysis, and application to bacterial systems. Biotechnology Progress 20, 1309-1324.

Diaconis, P., and Ylvisaker, S., 1985. Quantifying prior opinion. In Bayesian Statistics 2 (series). Elsevier, New York.

Dudley, A.M., Aach, J., Steffen, M.A., and Church, G.M., 2002. Measuring absolute expression with microarrays with a calibrated reference sample and an extended signal intensity range. Proc Natl Acad Sci USA 99, 7554-7559.

Efron, B., Tibshirani, R., Storey, J.D., and Tusher, V., 2001. Empirical Bayes analysis of a microarray experiment. Journal of the American Statistical Association 96, 1151-1160.

Evans, M., Hastings, N.A.J., Peacock, J.B., and Hastings, N.A.J., 1993. Statistical distributions. J. Wiley, New York.

Ghosh, B., 2004. Mixture models for assessing differential expression in complex tissues using microarray data. Bioinformatics 20, 1663-1669.

Ghosh, D., and Chinnaiyan, A.M., 2002. Mixture modelling of gene expression data from microarray experiments. Bioinformatics 18, 275-286.

Hoyle, D.C., Rattray, M., Jupp, R., and Brass, A., 2002. Making sense of microarray data distributions. Bioinformatics 18, 576-584.

Ji, Y., Wu, C.L., Liu, P., Wang, J., and Coombes, K.R., 2005. Applications of beta-mixture models in bioinformatics. Bioinformatics 21, 2118-2122.

Kendziorski, C.M., Newton, M.A., Lan, H., and Gould, M.N., 2003. On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine 22, 3899-3914.

Kerr, M.K., and Churchill, G.A., 2001. Experimental design for gene expression microarrays. Biostatistics 2, 183-201.

Lee, M.L.T., Kuo, F.C., Whitmore, G.A., and Sklar, J., 2000. Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci USA 97, 9834-9839.

Mikkilineni, V., Mitra, R.D., Merritt, J., DiTonno, J.R., Church, G.M., Ogunnaike, B., and Edwards, J.S., 2004. Digital quantitative measurements of gene expression. Biotechnology and Bioengineering 86, 117-124.

Nadon, R., and Shoemaker, J., 2002. Statistical issues with microarrays: processing and analysis. Trends Genet 18, 265-271.

Newton, M.A., Noueiry, A., Sarkar, D., and Ahlquist, P., 2004. Detecting differential gene expression with a semiparametric hierarchical mixture method. Biostatistics 5, 155-176.

Newton, M.A., Kendziorski, C.M., Richmond, C.S., Blattner, F.R., and Tsui, K.W., 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. J Comput Biol 8, 37-52.

Ozbudak, E.M., Thattai, M., Kurtser, I., Grossman, A.D., and van Oudenaarden, A., 2002. Regulation of noise in the expression of a single gene. Nat Genet 31, 69-73.

Parker, R.A., and Rothenberg, R.B., 1988. Identifying Important Results From Multiple Statistical Tests. Statistics In Medicine 7, 1031-1043.

Quackenbush, J., 2001. Computational analysis of microarray data. Nature Reviews Genetics 2, 418-427.

Reid, R.W., and Fodor, A.A., 2008. Determining gene expression on a single pair of microarrays. BMC Bioinformatics 9.

Rocke, D.M., and Durbin, B., 2001. A model for measurement error for gene expression arrays. J Comput Biol 8, 557-569.

Sebastiani, P., Gussoni, E., Kohane, I.S., and Ramoni, M.F., 2003. Statistical challenges in functional genomics. Statistical Science 18, 33-60.

Skliar, D.B., Gelmi, C., Ogunnaike, T., and Willis, B.G., 2007. Interaction of 2,2,6,6-tetramethyl-3,5-heptanedione with the Si(100)-2 x 1 surface: Scanning tunneling microscopy and density functional theory study. Surface Science 601, 2887-2895.

Steinhoff, C., Müller, T., Nuber, U.A., and Vingron, M., 2003. Gaussian mixture density estimation applied to microarray data. Advances In Intelligent Data Analysis V Lecture Notes in Computer Science 2810, 418-429.

Tanaka, M., Earl, A.M., Howell, H.A., Park, M.J., Eisen, J.A., Peterson, S.N., and Battista, J.R., 2004. Analysis of *Deinococcus radiodurans's* transcriptional response to

ionizing radiation and desiccation reveals novel proteins that contribute to extreme radioresistance. Genetics 168, 21-33.

Taylor, J.R., 1997. An introduction to error analysis: the study of uncertainties in physical measurements. University Science Books, Sausalito, Calif.

Thattai, M., and van Oudenaarden, A., 2001. Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci USA 98, 8614-8619.

Troyanskaya, O.G., Garber, M.E., Brown, P.O., Botstein, D., and Altman, R.B., 2002. Nonparametric methods for identifying differentially expressed genes in microarray data. Bioinformatics 18, 1454-1461.

Tusher, V.G., Tibshirani, R., and Chu, G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of The National Academy of Sciences of The United States of America 98, 5116-5121.

Wit, E., and McClure, J.D., 2004. Statistics for microarrays: design, analysis, and inference. John Wiley & Sons, Chichester, England; Hoboken, NJ, USA.

Figures

Figure 1: Raw data from gene expression studies in *Deinococcus radiodurans* following DNA damage. Note the "conical" sector shape.

Figure 2: Raw data of Figure 1 after the polar coordinates transformation.

Figure 3: Histogram of Low *R* data and mixture model.

Figure 4: Contributing Beta densities for mixture model fit to data in Figure 3.

Figure 5: Histogram of Med *R* data and mixture model.

Figure 6: Contributing Beta densities for mixture model fit to data in Figure 5.

Figure 7: Histogram of High *R* data and mixture model.

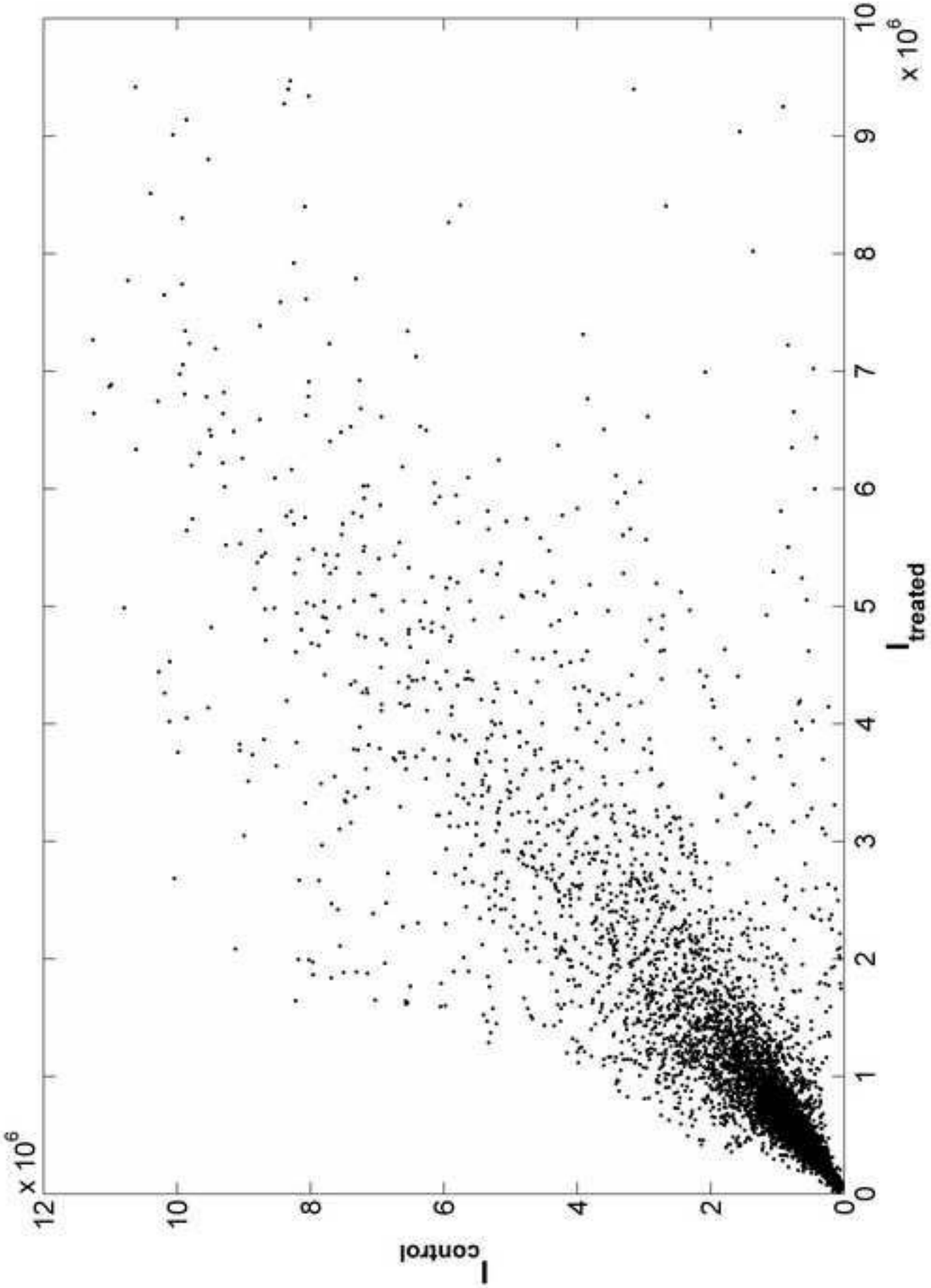
Figure 8: Contributing Beta densities for mixture model fit to data in Figure 7.

Figure 9: Histogram of bias-corrected High *R* data and mixture model.

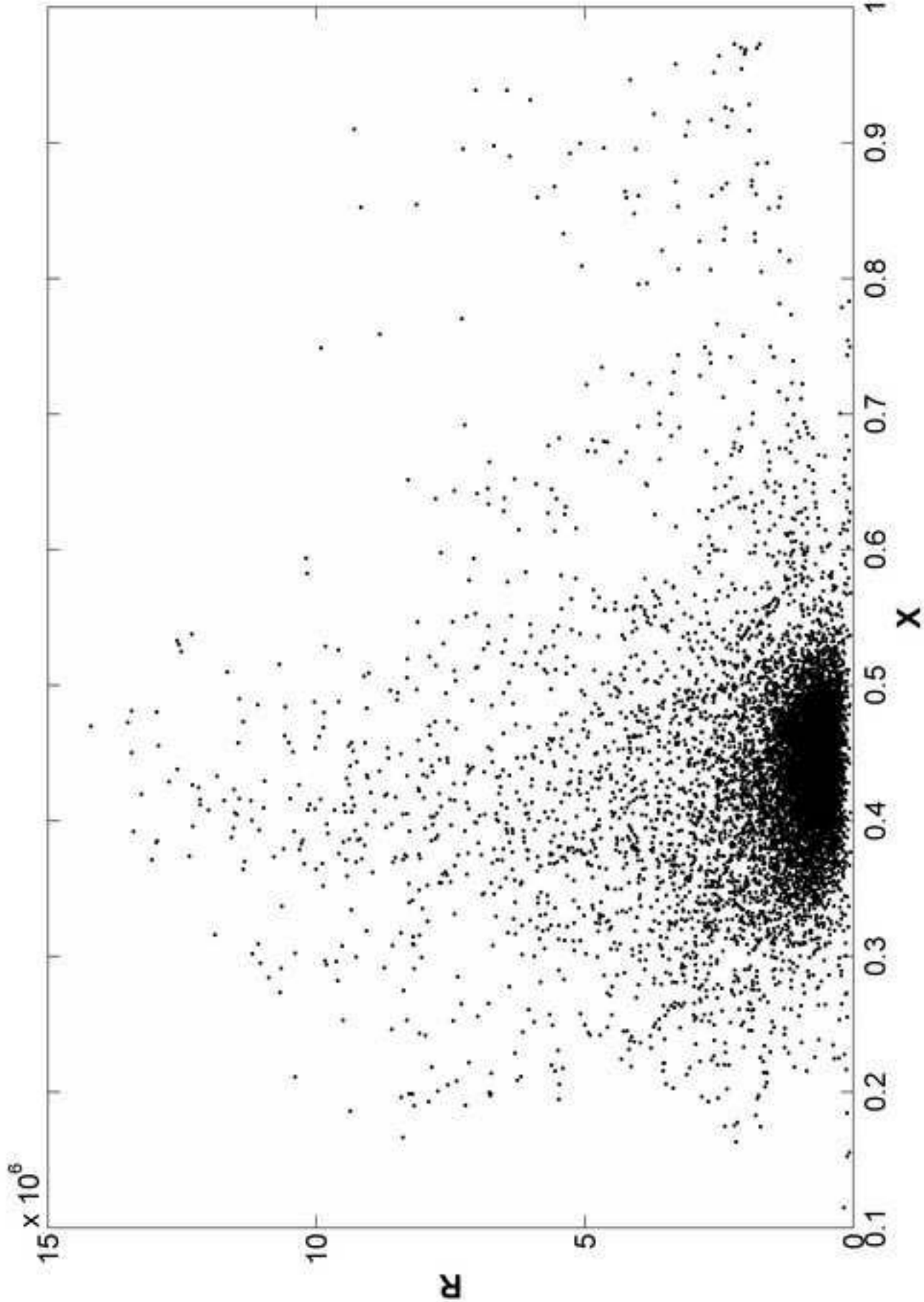
Figure 10: Contributing Beta densities for mixture model fit to corrected data in Figure 9.

Figure 11: Fitted mixture model (thick solid line) for the different intensity signal groups: (a) Low *R*; (b) Medium *R*; (c) High *R*. The probability of expression status as a function of the fractional intensity *x* is shown on the secondary *y* axis: P_{down} (thin solid line) and P_{up} (dashed line).

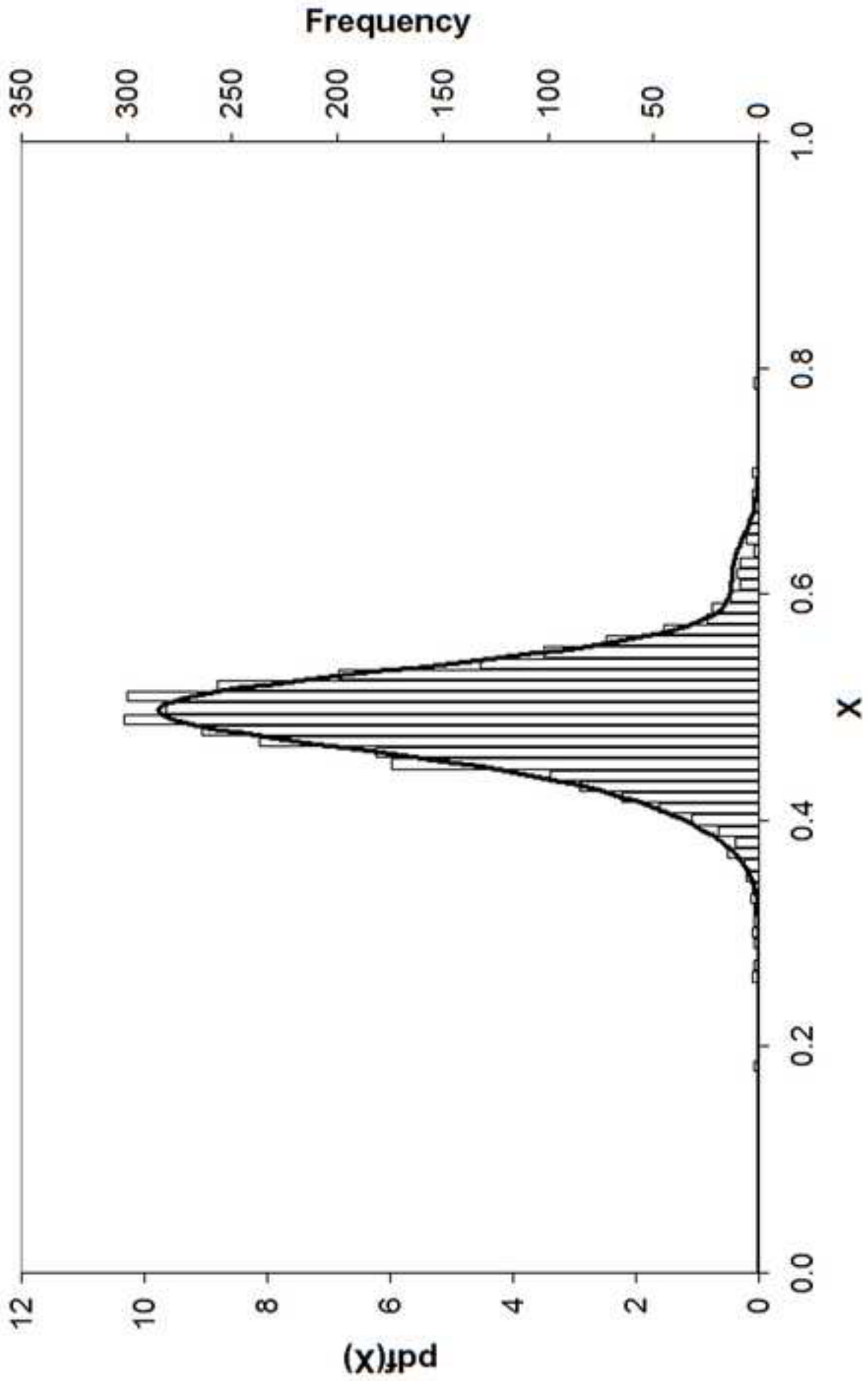
Figure 12: Histogram of the confidence index associated with the probabilities of expression status for the *D. radiodurans* case study. The histogram shows that 76.3% of the analyzed genes have a confidence index of 0.80 or greater associated with the computed probabilities, indicating that the variability associated with the computed probabilities is low overall.



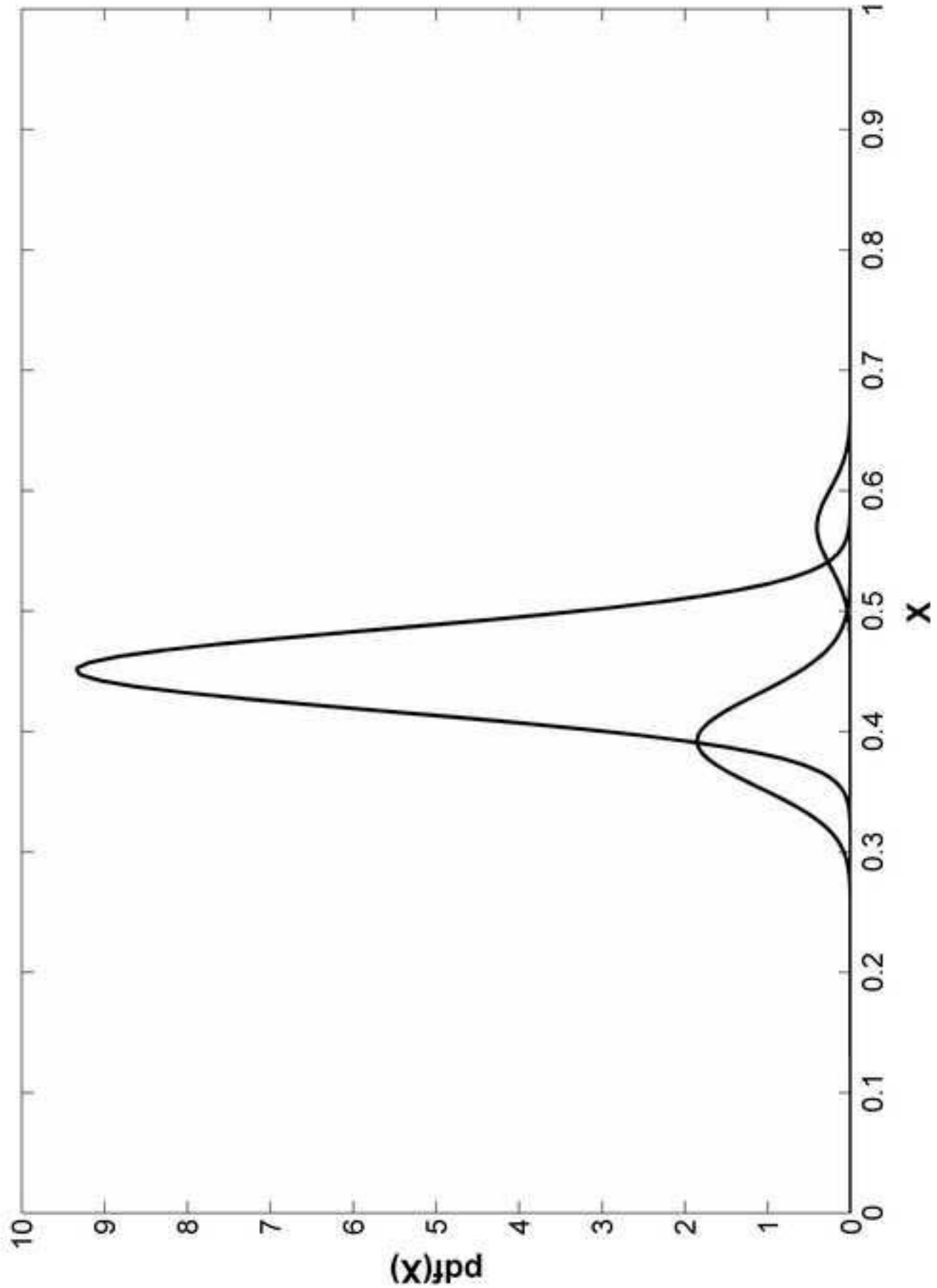
4. Figure 1



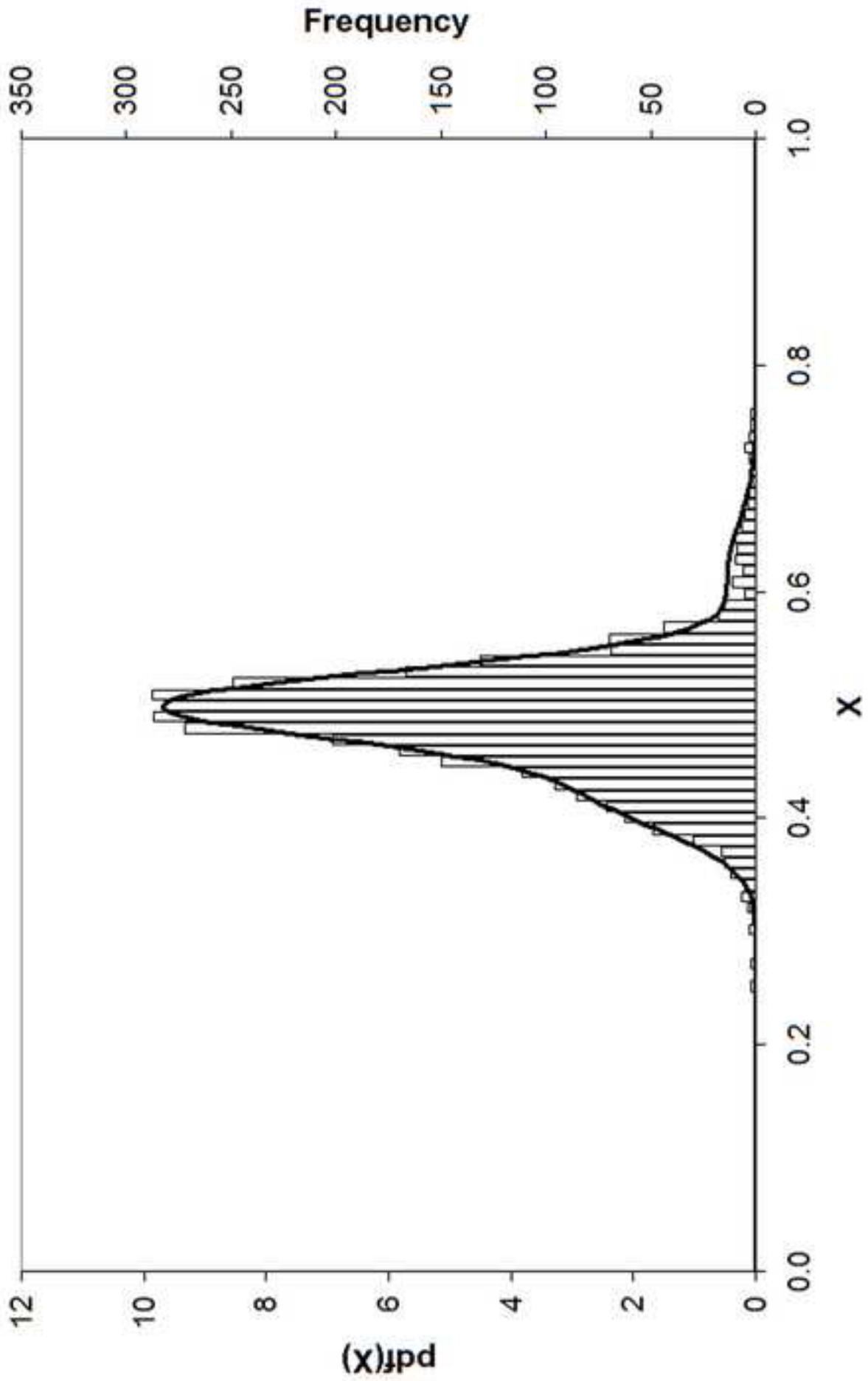
4. Figure 2



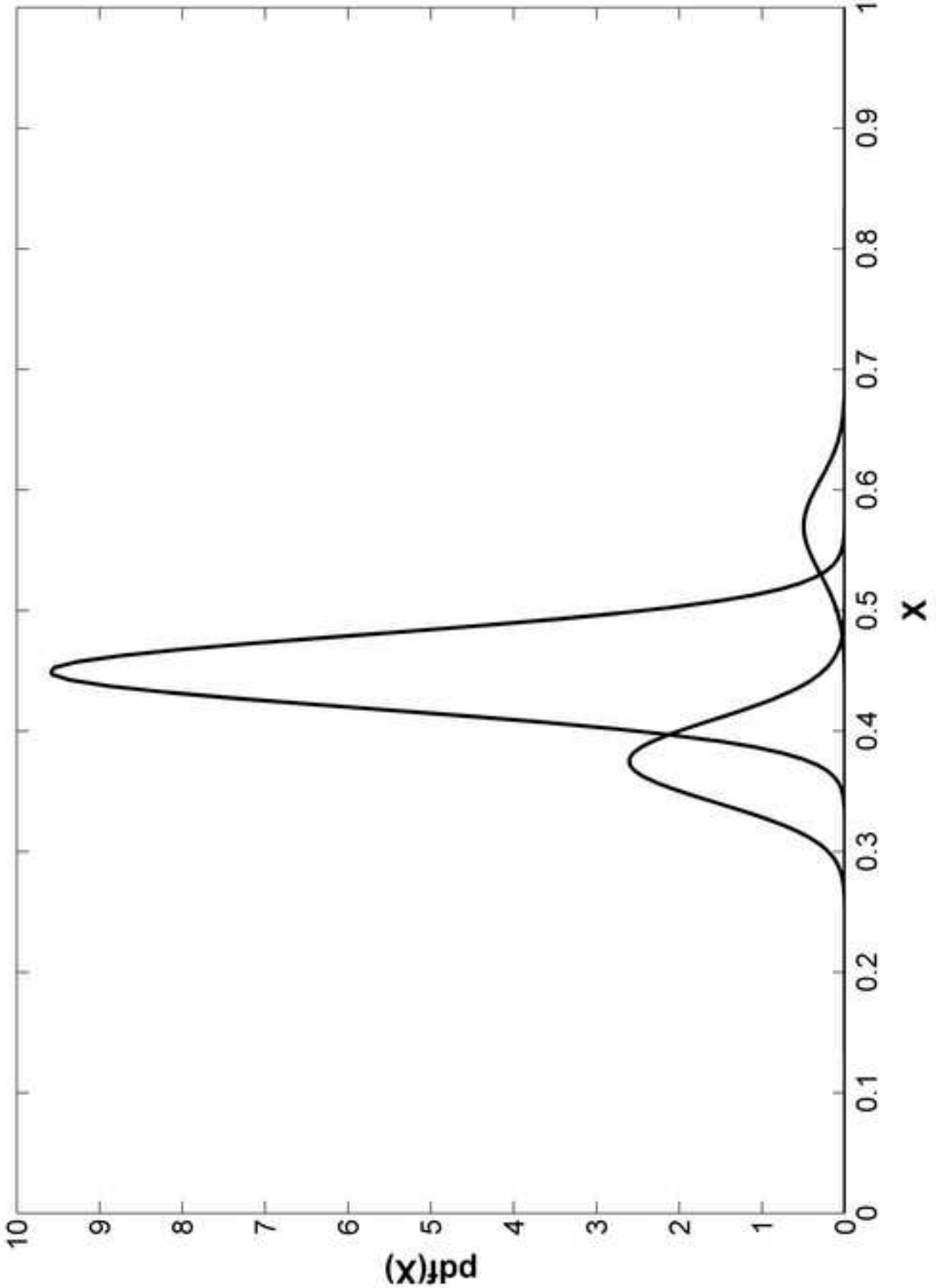
4. Figure 3



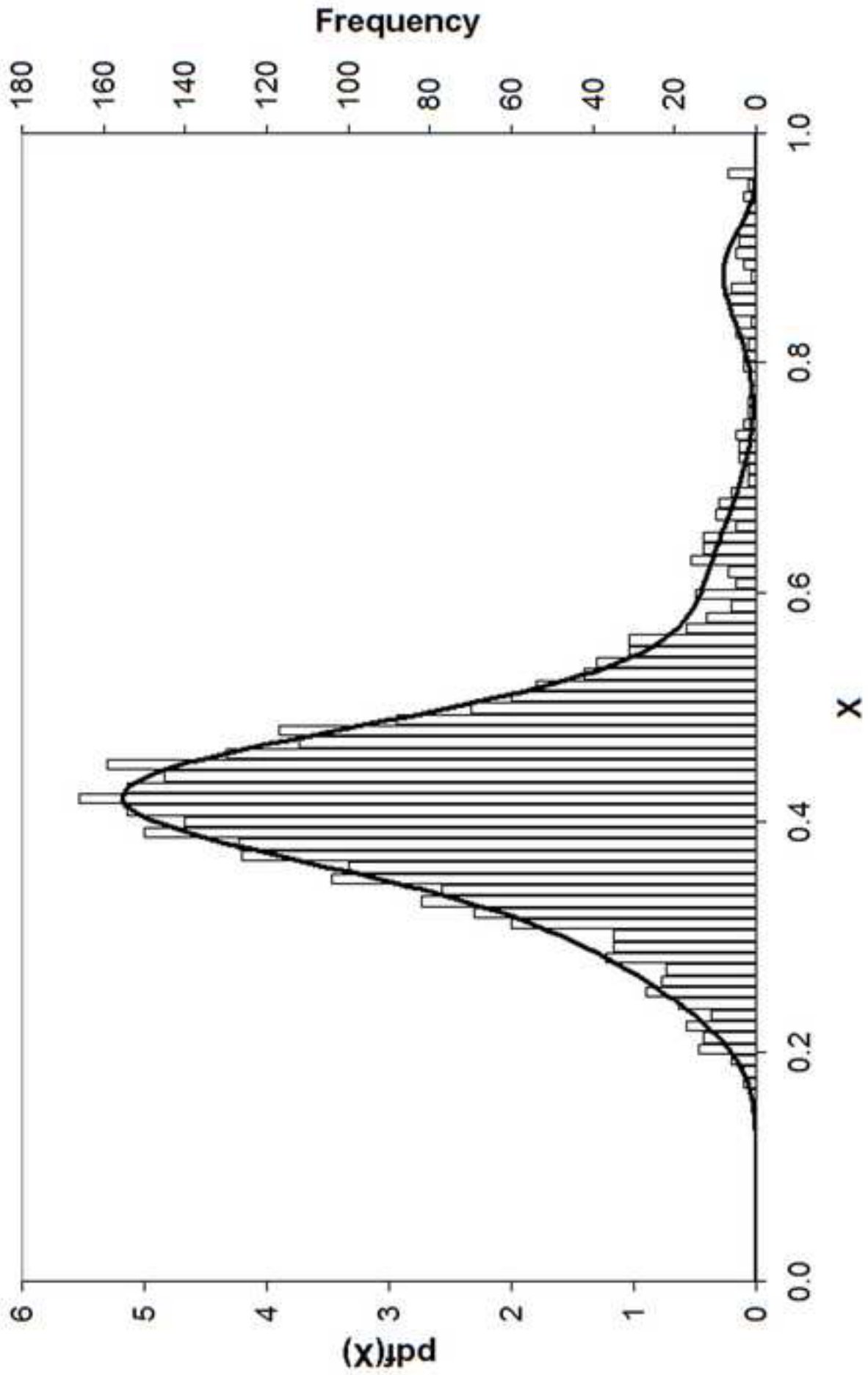
4. Figure 4



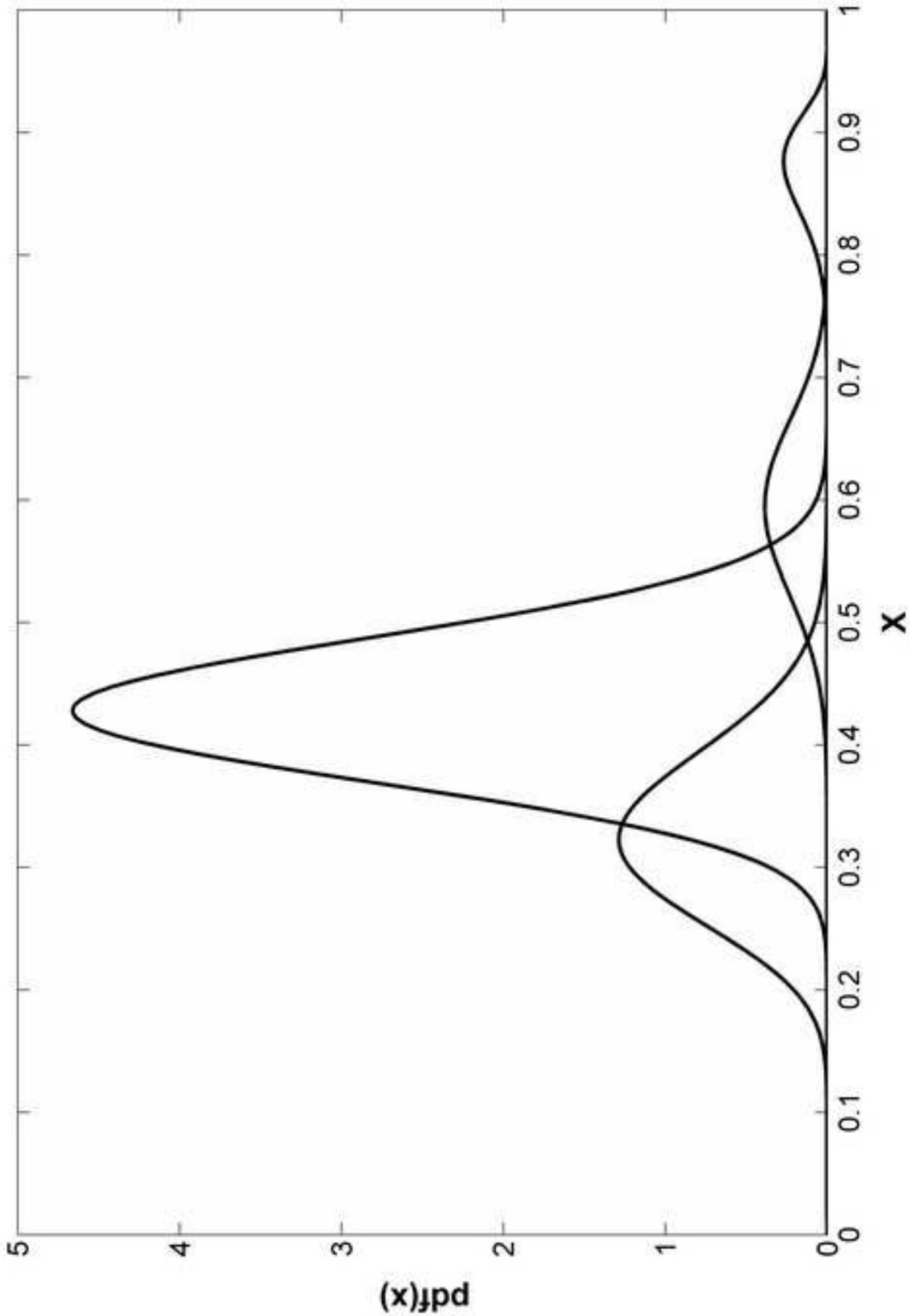
4. Figure 5



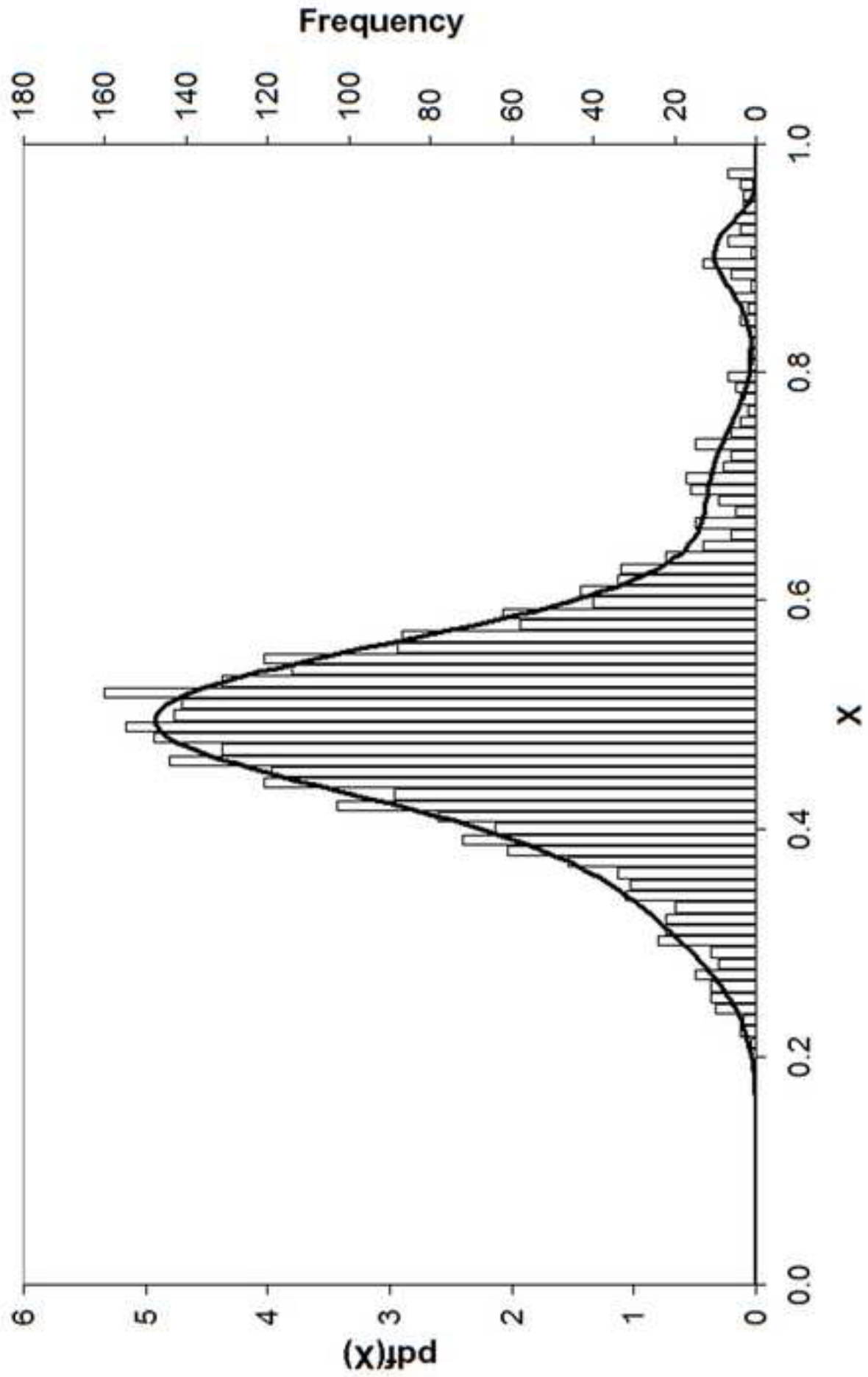
4. Figure 6



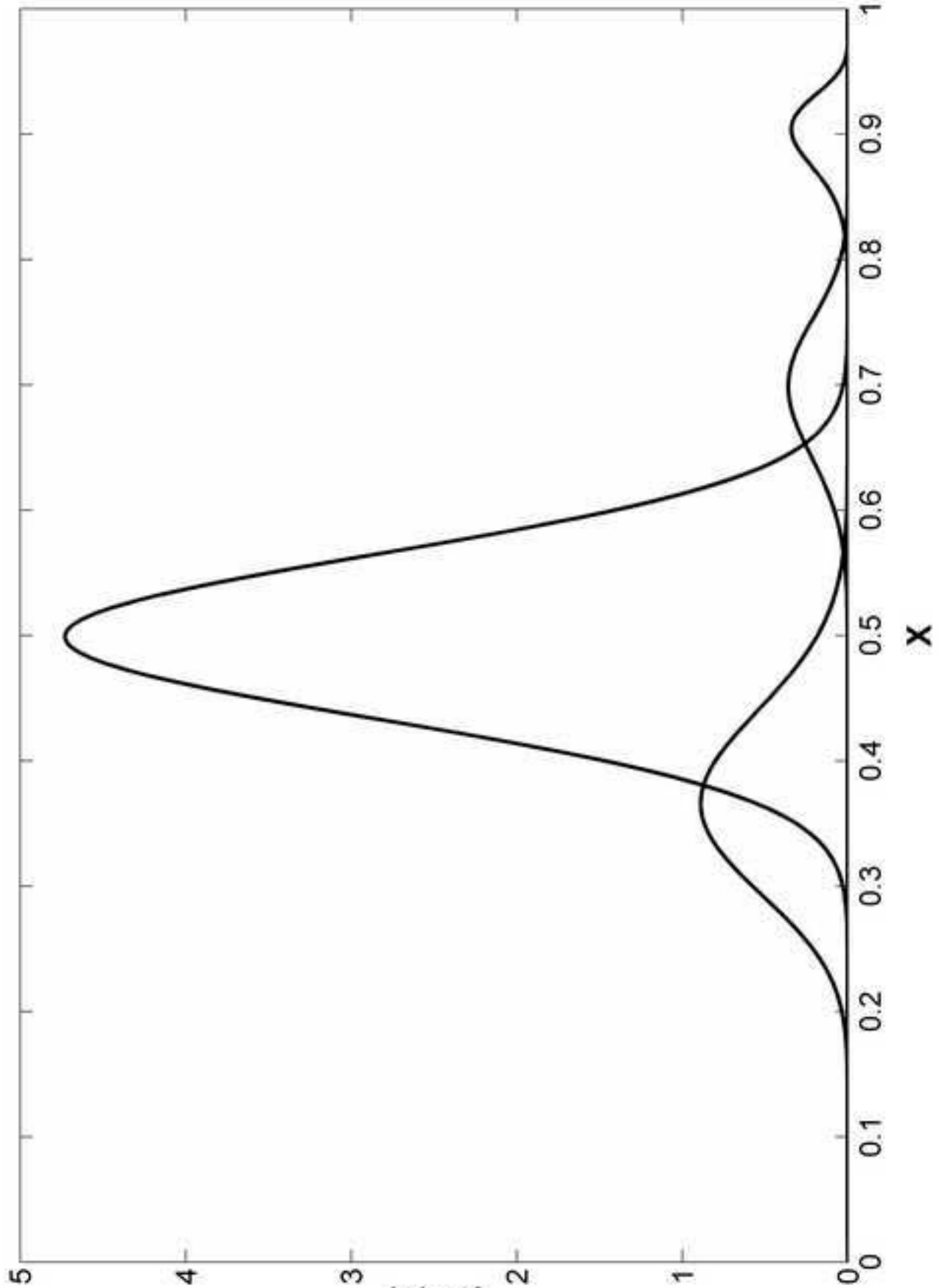
4. Figure 7



4. Figure 8

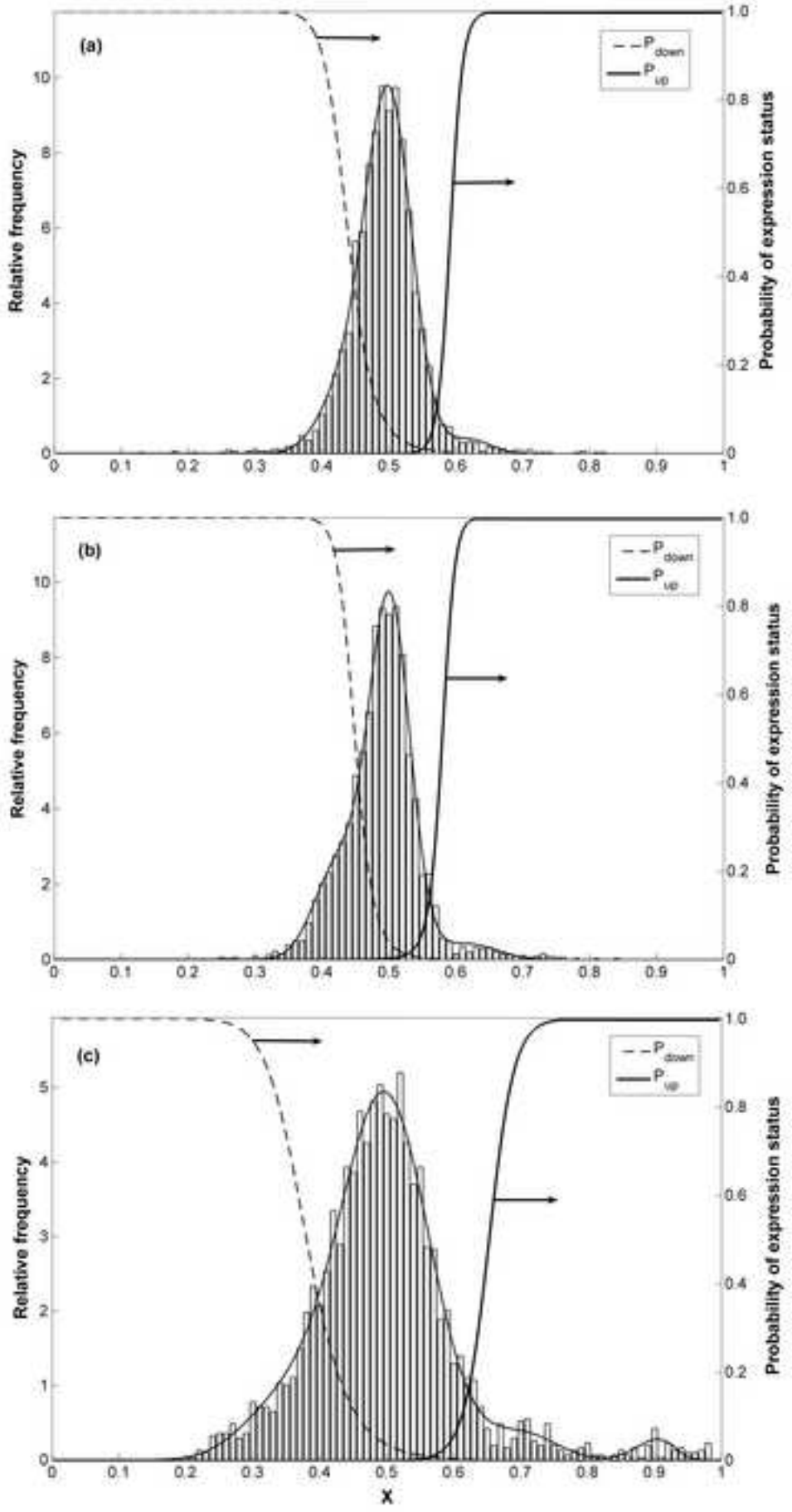


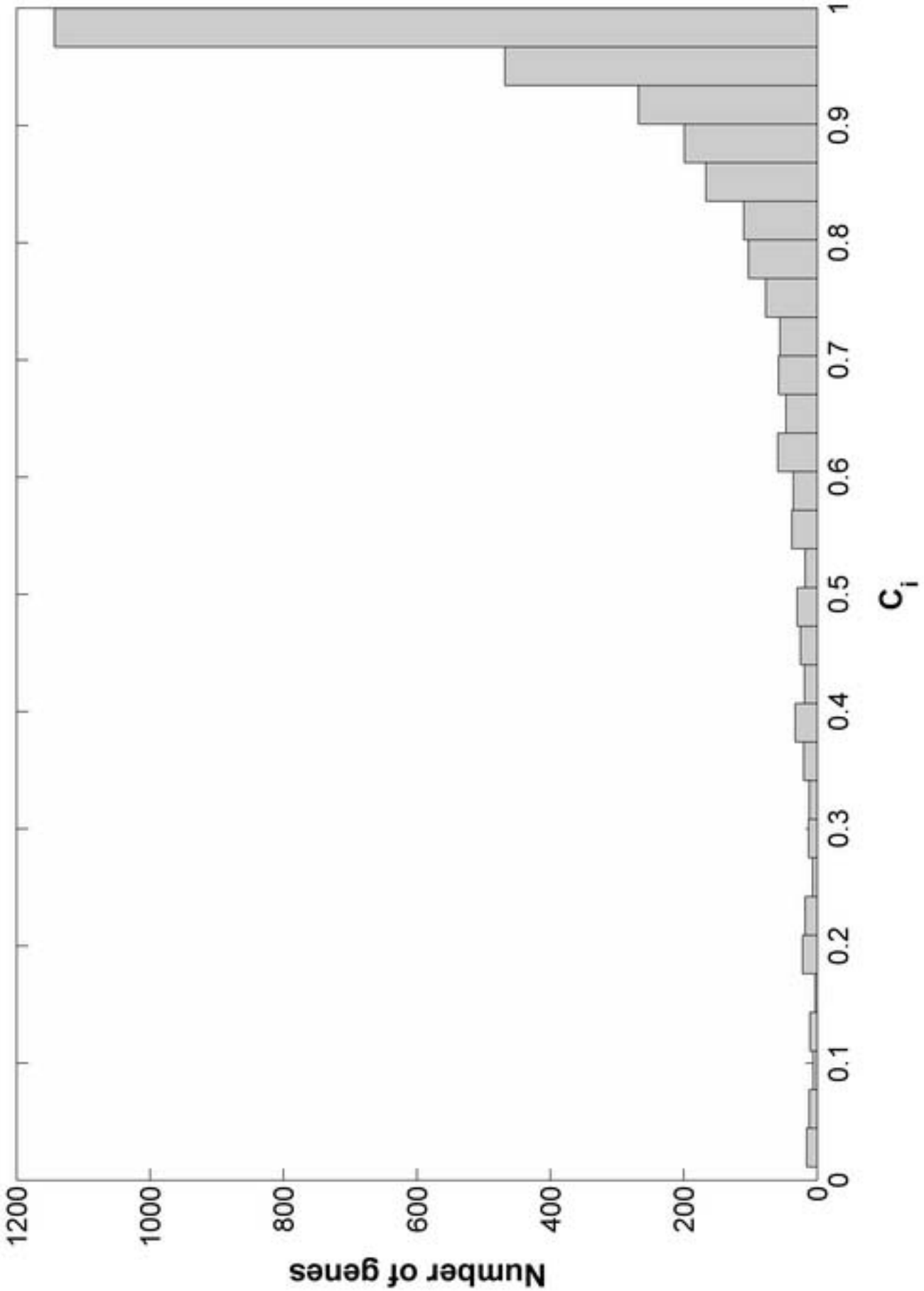
4. Figure 9



4. Figure 10

4. Figure 11





4. Figure 12

Table 1. Example of a typical output of the probabilistic framework.

Gene	Fractional intensity (x)			P_{down}			P_{non}			P_{up}			Confidence index (c_i)
	Min	Aver.	Max	Min	Aver.	Max	Min	Aver.	Max	Min	Aver.	Max	
arsenate reductase	0.49	0.51	0.52	0.01	0.02	0.04	0.96	0.97	0.98	0.00	0.01	0.01	0.97
cinA protein	0.89	0.90	0.90	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
citrate lyase, beta subunit	0.86	0.88	0.89	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
cytochrome P450-related protein	0.33	0.35	0.35	0.73	0.77	0.84	0.16	0.23	0.27	0.00	0.00	0.00	0.91
dihydroxy-acid dehydratase	0.68	0.69	0.69	0.00	0.00	0.00	0.07	0.09	0.12	0.88	0.91	0.93	0.96
DNA gyrase, subunit A	0.94	0.94	0.95	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
DNA helicase II	0.70	0.72	0.75	0.00	0.00	0.00	0.00	0.02	0.05	0.95	0.98	1.00	0.96
frnE protein	0.68	0.69	0.70	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
glutaryl-CoA dehydrogenase	0.39	0.41	0.44	0.48	0.80	0.96	0.04	0.20	0.52	0.00	0.00	0.00	0.61
glycerol uptake facilitator protein	0.38	0.40	0.41	0.49	0.81	0.99	0.01	0.19	0.51	0.00	0.00	0.00	0.59
glycerol-3-phosphate dehydrogenase	0.69	0.71	0.72	0.00	0.00	0.00	0.01	0.04	0.07	0.93	0.96	0.99	0.95
GMC oxidoreductase	0.47	0.49	0.51	0.04	0.09	0.19	0.81	0.91	0.96	0.00	0.00	0.00	0.88
Holliday junction DNA helicase	0.92	0.92	0.93	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
mocR protein	0.67	0.68	0.69	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
nitric oxide synthase-related protein	0.64	0.66	0.70	0.00	0.00	0.00	0.00	0.02	0.06	0.94	0.98	1.00	0.95
nitrogen regulatory protein P-II	0.64	0.66	0.68	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
nodulation protein-related protein	0.47	0.49	0.51	0.03	0.04	0.07	0.93	0.96	0.97	0.00	0.00	0.00	0.97
peptide methionine sulfoxide reductase	0.68	0.69	0.71	0.00	0.00	0.00	0.04	0.09	0.14	0.86	0.91	0.96	0.91
PhoH-related protein	0.66	0.67	0.69	0.00	0.00	0.00	0.09	0.25	0.35	0.65	0.75	0.91	0.79
phosphoribosylamine--glycine ligase	0.40	0.42	0.46	0.37	0.77	0.97	0.03	0.23	0.63	0.00	0.00	0.00	0.51
tellurium resistance protein TerA	0.77	0.78	0.78	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
thiophene and furan oxidation protein	0.72	0.73	0.74	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00
threonine synthase	0.36	0.38	0.39	0.94	0.97	0.99	0.01	0.03	0.06	0.00	0.00	0.00	0.95
tryptophan synthase, beta subunit	0.38	0.38	0.39	0.44	0.81	1.00	0.00	0.19	0.56	0.00	0.00	0.00	0.55
uroporphyrinogen decarboxylase	0.66	0.70	0.73	0.00	0.00	0.00	0.01	0.13	0.30	0.70	0.87	0.99	0.76
v-type ATP synthase, D subunit	0.60	0.61	0.62	0.00	0.00	0.00	0.04	0.11	0.25	0.74	0.89	0.96	0.82

Table 2. Main statistics of the distributions used to simulate the microarray data.

Status	Beta parameters (α, β)	Mean fractional intensity (*x*)	Standard deviation	Mean fold-change	Log_2(fold-change)
Down-regulated	(38, 62)	0.38	0.048	0.62	-0.69
Not differentially expressed	(75, 75)	0.50	0.041	1.00	0.0
Up-regulated	(62, 38)	0.62	0.048	1.63	0.70

Table 3. Total number of correctly identified genes by the probabilistic framework and SAM (the percentages of correctly identified genes are shown in parenthesis).

	1 array	**2 arrays**	**3 arrays**	**4 arrays**	**5 arrays**
Probabilistic framework	603 (80.4%)	688 (91.5%)	729 (97.2%)	743 (98.9%)	747 (99.6%)
SAM	*Requires replicates*	668 (88.7%)	699 (93.2%)	701 (93.3%)	710 (94.7%)

Table 4. Table of complete model results.

	Low *R*	**Medium *R***	**High *R***
$f_1(\alpha_{11}, \alpha_{10})$	(64.0, 98.8)	(76.0, 126.1)	(15.2, 30.7)
corrected	(69.3, 88.1)	(74.9, 100.3)	(16.9, 28.6)
$f_0(\alpha_{01}, \alpha_{00})$	(98.8, 120.2)	(121.5, 148.8)	(31.1, 41.3)
corrected	(106.9, 106.7)	(136.8, 136.1)	(30.1, 30.2)
$f_2(\alpha_{21}, \alpha_{20})$	(137.8, 104.4)	(95.6, 72.5)	(29.4, 20.3)
corrected	(150.3, 92.8)	(66.9, 42.4)	(50.3, 22.3)
$f_3(\alpha_{31}, \alpha_{30})$			(61.6, 9.5)
corrected			(96.3, 11.1)
ϕ_1 corrected	0.18	0.22	0.22
	0.18	0.24	0.16
ϕ_0 corrected	0.79	0.73	0.68
	0.79	0.71	0.77
ϕ_2 corrected	0.03	0.05	0.07
	0.03	0.05	0.05
ϕ_3 corrected			0.03
			0.02