Conference on Applied Statistics in Agriculture 1999 - 11th Annual Conference Proceedings

ANALYSIS OF NUCLEI FLUORESCENCE HISTOGRAMS USING NON-LINEAR FUNCTIONS OR WAVELETS

Susanne Aref

Maria Kocherginsky

Carrie A. Northcott

Lane A. Rayburn

See next page for additional authors

Recommended Citation

Aref, Susanne; Kocherginsky, Maria; Northcott, Carrie A.; and Rayburn, Lane A. (1999). "ANALYSIS OF NUCLEI FLUORESCENCE HISTOGRAMS USING NON-LINEAR FUNCTIONS OR WAVELETS," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1260

Author Information

Susanne Aref, Maria Kocherginsky, Carrie A. Northcott, and Lane A. Rayburn

ANALYSIS OF NUCLEI FLUORESCENCE HISTOGRAMS USING NON-LINEAR FUNCTIONS OR WAVELETS

Susanne Aref, Maria Kocherginsky,
Dept. of Statistics

Carrie A. Northcott, and Lane A. Rayburn
Dept. of Crop Sciences

Abstract

Histograms based on 5,000 nuclei from cells (Chinese hamster ovary cells, bone marrow cells) are used to determine the coefficient of variation (CV) of observations surrounding the highest peak. The cells are subjected to various treatments, for example exposure to herbicides. By eyeballing the histogram, an interval under the highest peak is determined. The CV calculated from the histogram on the eyeballed interval is the response variable in an ANOVA. To avoid the subjectivity of eyeballing the histogram, non-linear functions such as the Gaussian density function can be used to model the histogram. The CV may then be determined from the parameter estimates. In many experiments nonlinear functions modeling the histograms smooth away differences in CVs obtained this way, though visually the histograms appear to be different. Then nonlinear functions or wavelets can be used to obtain intervals for calculating CVs of the histograms restricted to these intervals. The nonlinear models require close initial values for each histogram, while the wavelets just require choice of wavelet and level of decomposition.

1. Introduction

In flow cytometry histograms of the spectra of DNA nuclei subjected to various treatments are used to obtain summary variables. These variables are then analyzed to determine differences or similarities of the treatments through an ANOVA. The summary variables are calculated from the histogram restricted to an interval around the highest peak. The determination of the interval is usually from eyeballing the part of the histogram associated with the highest peak. The mean and standard deviation based on the histogram on that interval are used to calculate the coefficient of variation, the CV (ratio between standard deviation and mean times 100). Usually one person determines all intervals. In histograms of certain types of nuclei there is one very well defined peak (Fig 1). In other types of nuclei there is a fairly pronounced secondary peak in the histogram. Here the largest peak appears to be less symmetric at the base,

the curve of the peak has slight bumps on the sides, consistently in replications of a treatment. The interval between peaks has higher counts than the intervals outside the peaks (Fig 2.). The eyeballing of the histogram to determine the interval becomes difficult and subjective. In some experiments the researcher determining intervals can identify histograms from different treatments (Fig. 3). Attempts to determine intervals blindly are therefore flawed.

The subjectivity of determining the interval by eyeballing the histograms is dissatisfying scientifically. Previous attempts to analyze this type of data include non-linear modeling, which sometimes smooth away differences, that visually appears to be present as in Fig. 4, where the control treatment looks different from the other treatments (shape-wise not location-wise). Another way to analyze such data is to use the Kolmogorov-Smirnov test (Young, 1977) on just two histograms. This does not take differences between histograms for the same treatment into account, which should be the measurement error for the determination of differences between treatments. A third way is to use a procedure of smoothing, translocating, and normalizing histograms from two different treatments. The resulting histograms are then subjected to t-tests in every channel with p-values showing where the differences occur followed by a classification argument (Bagwell et al., 1979). To generalize the method to several treatments is not easy.

The current method of using CVs is appealing in that once the measure is obtained the experiment can be analyzed using an ANOVA. The problem is that determining the defining interval is subjective. One way to amend that is to obtain parameter estimates from nonlinear function models of the histograms. Variables such as mean and CV are obtained from the parameter estimates. The usual ANOVA can then be carried out on these variable constructs. However, the smoothing from the models may smooth away more subtle differences. In such cases the histogram may still be modeled using nonlinear functions or be approximated by wavelets. If nonlinear functions are used, intervals are determined based on parameter estimates. If wavelets are fitted, the size of the wavelet fit at channels surrounding the highest peak are used to determine cut-off points and thus the end points of the interval. CVs are calculated from the raw data on this defining interval. Either way the interval is objectively determined, any particularities in the shape of the histograms are preserved, and a straightforward ANOVA can be carried out on the resulting CVs.

2. Experimental Methods

The data used in this paper are from two different experiments. Each experiment was set up using tanks in a randomized complete block design. On each day of laboratory analysis, nuclei from each treatment in a block was analyzed in random order. The different days are important since each day the flow cytometer-cell sorter must be realigned. In one of the experiments bone marrow cells were subjected to two different treatments, a control (non-exposure) treatment and a treatment of exposure to a known mutagen ara-C (cytosine β-D-arabinofuranoside). The other experiment was conducted on Chinese hamster ovary cells subjected to four different treatments, the herbicide atrazine at two levels, the known mutagen ara-C as a positive control, and non-exposure as a negative control.

Nuclei from 5,000 cells were stained with a fluorescent dye, fluorochrome (Propidium Iodide or PI). Based on the fluorescence of the stained nuclei excited by a laser, the amount of

fluorescence given off is recorded. These records fall into 256 channels that make up the histogram. A typical histogram will contain at least one distinct peak and some noise as in the bone marrow cells data (Fig 1). The histograms for certain types of cells are sometimes more complex with distinct second peaks and increased levels of noise as in the Chinese hamster ovary cells data (Fig 2). The first peak is the same as the G1/G0 peak and the second peak is the G2 peak. Commonly, the two extreme channels (1 and 256) have excess amounts of fluorescence. For the determination of intervals the two extreme channels are ignored by setting their channel counts to zero. Usually the variables of interest are the location of the highest peak or the CV of the highest peak. The histograms contain different total numbers of nuclei ($\leq$ 5,000) after the end channels have been set to zero. This does not impact location or CV. The CV is used rather than the standard deviation since the channels are wider with larger channel numbers due to a log transform to obtain the channel variable. For a more technical description see Taets et al, 1998.

3. Non-Linear Function Computational Methods

A non-linear function such as the Gaussian density can be used to model the censored histograms. In the case of a single peak histogram a simple approach using SAS PROC NLIN with a set of reasonable initial parameter values is fairly problem free. When the single peak histogram looks very symmetric the Gaussian density function is a good model for the histogram (Fig. 5). In the case of two peaks the focus is on the highest peak, as the experimenting scientist usually ignores the secondary peak. In the modeling of the histograms, it is harder to capture the second peak using non-linear modeling, but parameters for the primary peak do not change noticeably whether or not the secondary peak is modeled.

One problem with non-linear modeling is the over-smoothing of the histogram for certain types of experiments. Here differences that appear in the form of slight bumps on the sides of the peak and the elevated level of counts between peaks will not be picked up. In such cases the determination of a defining interval is the point of interest. One way is to look at a large enough interval around the highest peaks for all treatments and replications, that contains the highest peaks and near surroundings. A nonlinear model is then fitted to these censored histograms. The nonlinear model may be a Gaussian density type function (Fig 6) or closer-fitting model such as a combination of functions. One combination function that gives a closer fit consists of two exponential functions (one on each side of the peak) with a straight line connecting them (Fig 7). With SAS PROC NLIN this may be done including the estimation of the location of the two knots one for each of the exponential functions connecting to the piece of straight line by the derivative-free option (DUD). The estimated parameters are averaged over replications for each treatment and used to create the defining intervals. The intervals are placed according to the estimated location parameter of each histogram. Thus the location in a histogram is determined by the estimated parameter from the specific histogram, while the length of the interval surrounding the location estimate is the same within each treatment. The intervals defined by the two different fittings of non-linear functions are slightly different in width for the same factor and slightly different in location.

The exercise of fitting non-linear functions was only to determine an objective interval on which the usual response variable, the CV, can be calculated from the original data in the

histograms restricted to the these objective intervals. The CVs are then used as the response variable in an ANOVA. The experimental design contains days of performing the flow cytometry as blocks and the herbicides or other mutagens as treatments as mentioned in section 2.

4. Non-Linear Functional Forms

The Gaussian density function used for modeling a single peak histogram has the following form:

$$g(x) = B + C \exp[(x-A)^2/(2D^2)]$$

The parameter estimate of A is the mean or location of the largest peak. The CV is then determined from D/A. B is a nuisance parameter picking up ground level noise.

The results of fitting the Gaussian density function to six replicates each of a control and an ara-C treatment of bone marrow cells resulted in a significant difference between the CV calculated from the estimated parameters. The p-value was 0.0036 in a paired t-test and 0.0313 in a Wilcoxon signed rank test.

The other data set was six replications each of four different treatments of Chinese ovary cells. In this case differences were smoothed away. Both the Gaussian density function and the more close fitting function consisting of exponential functions connected by a line were used. The exponential functions were:

$$g(x) = B_1 + C_1\exp(-(A_1-x)/D_1) \qquad \text{for } x<A_1$$

$$g(x) = B_2 + C_2\exp(-(x-A_2)/D_2) \qquad \text{for } x>A_2$$

connected by the linear segment:

$$g(x) = B_2 + C_2 - (A_2-x)(B_2+C_2-B_1-C_1)/(A_2-A_1) \qquad \text{for } A_1<x<A_2.$$

Neither model was able to capture the difference between the control and the other treatments. Instead a defining interval was determined. From the estimated parameters the average for each treatment was used to create a defining interval around each location. For the Gaussian form the intervals were determined simply as $A \pm$ constant*D. For the piecewise function the interval was determined by $(A_2+A_1)/2 \pm (A_2-A_1 +$constant*$(D_1 + D_2))$. The intervals obtained in this fashion gave different results for different constants (from 2.0-4.5) in ANOVAs. The estimates of the spread parameters were not precise enough. To get better estimates of the spread, the lengths of the defining intervals were averaged over replications with D and the A_2-A_1 difference variables

averaged separately to ensure the proper placement around the highest peaks. Then CVs were calculated for the histograms restricted to the averaged defining intervals.

5. Wavelet Fitting

To obtain the nonlinear function fits, each histogram must be fitted separately using different initial parameter estimates. The method is rather elaborate. Another way of obtaining the intervals is to use wavelets ("small waves"), where initial settings are the same for all histograms. Wavelets are oscillatory over a very small temporal interval and decrease rapidly to zero outside that interval, contrasting with the behavior of Fourier basis functions that keep oscillating infinitely. Wavelets provide sets of basis functions and have zero net area, $\int\psi(t)dt = 0$. These characteristics result in wavelets being better in representing functions localized in both time and frequency. In particular, wavelets give better representations of functions with sharp spikes or edges using fewer terms than Fourier functions. Wavelets "turn the information of a signal into numbers - coefficients - that can be manipulated, stored, transmitted, analyzed, or used to reconstruct the original signal" (Hubbard, 1996).

6. Wavelet Computational Methods

The basic steps in wavelet constructions are:

1) select "mother" wavelet $\psi(t)$ from the different classes of known wavelets,
2) dilate and translate the mother wavelet in time to get a wavelet basis $\psi(at-b)$.

For the analysis of the data sets in this study wavelets were used to fit the largest peaks. The first step in the process is to transform the data using the wavelet basis. The data is de-noised as much as possible by using low order filters, such as the spline filter of order 1. In order not to over-smooth the fitting function, the decomposition is done only to the second level (Fig. 8). The coefficients obtained from the decomposition (Fig. 9) correspond to the location of the peak in the original data. Only the largest coefficients are of interest, since the smaller coefficients are considered to be noise.

Small coefficients are eliminated by compression. This can be done in two ways: either by keeping a fixed number of coefficients, or setting a threshold, below which the coefficients are discarded. Using the first approach the number of coefficients to keep was set to three. This choice was based on the coefficient plots in Fig. 9. In a) and b), of the largest coefficients two were very large, and the next largest one or two coefficients were distinctly larger than the remaining ones. In c) and d) there was one very large coefficient together with two somewhat smaller ones, and a gap between these and the rest of the coefficients. These patterns were similar for all replications. By keeping just the three largest coefficients, the compression ratio is 3:128 (or 2.3% of the coefficients). The next step is to reconstruct the data from the compressed coefficients by using the inverse wavelet transform. This denoises the data (since the compression ratio is small), and provides a rather coarse fit of the peak in the original data (Fig. 10).

Only the reconstructed data points with values above a small percentage of the largest reconstructed coefficient were kept. The points that are kept define the interval on which to compute CVs. This censoring was carried out using several different percentage cut-off values (between 0.6% to 0.9%) to check whether further analysis was robust to such changes. The cut-off percentage used was the same for the four treatments. A more robust estimate of the length of the defining interval was necessary, similar to the method using non-linear functions. So, the lengths of the defining intervals were averaged over the replications, and then centered at the midpoints of the initial defining interval.

7. Results

Different constants were used to obtain defining intervals with non-linear functions. Both functional fits, Gaussian and composite exponential, resulted in similar mean separations for the different constants see Table 1, a) and b). It appears that either function can be used, though the model based on the composite exponential function is slightly more stable, possibly due to a closer fit. The patterns of mean separations were almost identical for constants between 3.0 and 4.5 for the Gaussian function intervals and constants between 2.5 and 4.5 for the exponential composite function intervals. The results of the ANOVA were that the control was different from all other treatments, the two herbicide treatments were not different from each other, and the ara-C mostly was different from the two herbicide treatments.

Using different percentage cut-off points for the wavelet fits similar mean separation results occurred between 0.7% to 0.8%. At the 10% level the mean separation results were identical to the non-linear function approach, while at the 5% level the atrazine treatment at 20ppm was not different from the control (Table 1 c).

8. Conclusion

In flow cytometry CVs are used in an ANOVA to determine any differences in treatments. The CVs are calculated from histograms restricted to certain defining intervals. Eyeballing the histograms is commonly used to determine the defining intervals. An objective way to obtain the defining intervals is by the use of nonlinear modeling or wavelet fitting. Either way the length of the defining interval is the mean interval length for each treatment. The location of the interval is as determined by parameter estimates or wavelet coefficients.

In the case of nonlinear models the parameter estimates multiplied by factors from 2.5 to 4 provided similar ANOVA results at a 5% significance level, as did the wavelet models with cut-off points from 0.7% to 0.8% at a 10% significance level. There were differences between the control, the ara-C treatment, and the atrazine treatments, but not between the atrazine levels. These results were different from the results of an ANOVA performed on the CVs obtained by eyeballing the histograms. In the eyeballed CVs there was a difference between the control and the other treatments as in the analyses based on nonlinear modeling or wavelet fitting. There was no difference between ara-C and the two atrazine levels, while there was a difference in the two atrazine levels.

The methods used to obtain the CVs for further analyses are complicated, but at least are not subjective and do appear to give reasonable results. The non-linear estimation requires a close look at each of the histograms in order to get initial parameter estimates. The wavelet approach is more general and requires only decisions about the type of wavelet (here the spline filter of order 1), level of decomposition (here 2 levels), and how many coefficients to retain (here 3 coefficients). The defining intervals based on wavelets resulted in a mean separation that was more conservative than the mean separation obtained from using defining intervals based on non-linear functions. Further refinement of the wavelet method may give better results. The resulting analyses of the Chinese hamster ovary cells did not completely agree with the previous eyeballing technique. Thus the study also shows that sometimes this kind of data is very sensitive in determining the intervals on which the CVs are calculated. Especially for this type of data, results obtained using histogram eyeballing are suspect.

Acknowledgements

We thank Dr. Xuming He and Dr. Douglas Simpson, Dept of Statistics at UIUC, and Dr. Keith Muller, Department of Biostatistics, University of North Carolina, Chapel Hill, for discussions during the process of preparing this paper.

References

Bagwell, C.B., J.L. Hudson, and G.L. Irvin, III. 1979 Nonparametric Flow Cytometry Analysis. J. Histochemistry and Cytochemistry. 27:293-296.

Hubbard, Barbara. 1998. The World According to Wavelets. 320pp.

Littell, R.C., G.A. Milliken, W.W. Stroup, and R.D. Wolfinger. 1996. SAS System for Mixed Models. Cary, NC: SAS Institute Inc. 633pp

Mathematica Wavelet Explorer. 1996. Wolfram Research, Inc.

Taets, C., S. Aref, and A.L. Rayburn. 1998. The Clastogenic Potential of Triazine Herbicide Combinations Found in Potable Water Supplies. Environmental Health Perspectives. 106:197-201.

Young, I.T. 1977. Proof without Prejudice: Use of Kolmogorov-Smirnov Test for the Analysis of Histograms from Flow Systems and Other Sources. J. Histochemistry and Cytochemistry. 25:935-941.

Table 1. Chinese hamster ovary cells. Mean separation of CVs calculated on intervals defined by a) a Gaussian model, b) a combination of exponential models, c) wavelets, and d) eyeballing the histograms. Different letters signify significant differences at the 5% level in parts a), b), and d). Two significance levels (5% and 10%) are used in part c).

a) Intervals defined by Gaussian model

	constant										
treatment	2.00	2.25	2.50	2.75	3.00	3.25	3.50	3.75	4.00	4.25	4.50
ara-C	A	A	A	A	A	A	A	A	A	A	A
atr 20ppm	B	B	B	A	B	B	B	B	B	BC	B
atr 3ppm	AB	BC	B	A	BC	B	B	B	B	B	B
control	C	C	C	B	C	C	C	C	C	C	C

b) Intervals defined by Exponential models

	constant										
treatment	2	2.25	2.5	2.75	3.00	3.25	3.50	3.75	4.00	4.25	4.50
ara-C	A	A	A	A	A	A	A	A	A	A	A
atr 20ppm	A	AB	B	B	B	B	B	B	B	B	B
atr 3ppm	AB	B	BC	B	B	B	B	BC	B	BC	BC
control	B	C	C	C	C	C	C	C	C	C	C

c) Intervals defined by wavelets

d) eye-ball results

	percentage										
	0.65 %		0.70 %		0.75 %		0.80 %		0.90 %		
treatment	5%	10%	5%	10%	5%	10%	5%	10%	5%	10%	
ara-C	A	A	A	A	A	A	A	A	AB	AB	AB
atr 20ppm	A	AB	A	B	B	B	B	B	AB	AB	A
atr 3ppm	A	AB	A	AB	AB	B	AB	B	A	A	B
control	A	B	B	C	C	C	C	C	B	B	C

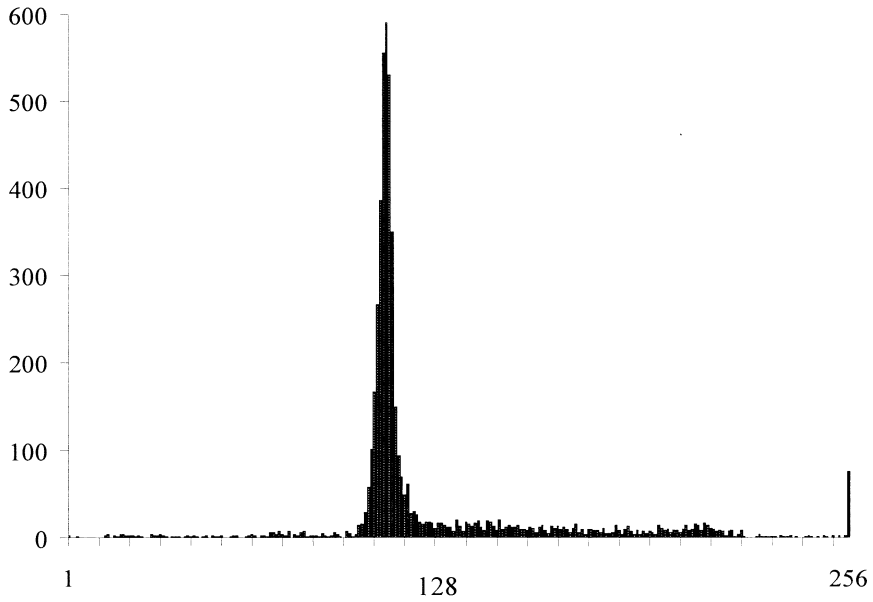

Figure 1. Bone marrow cells histogram.

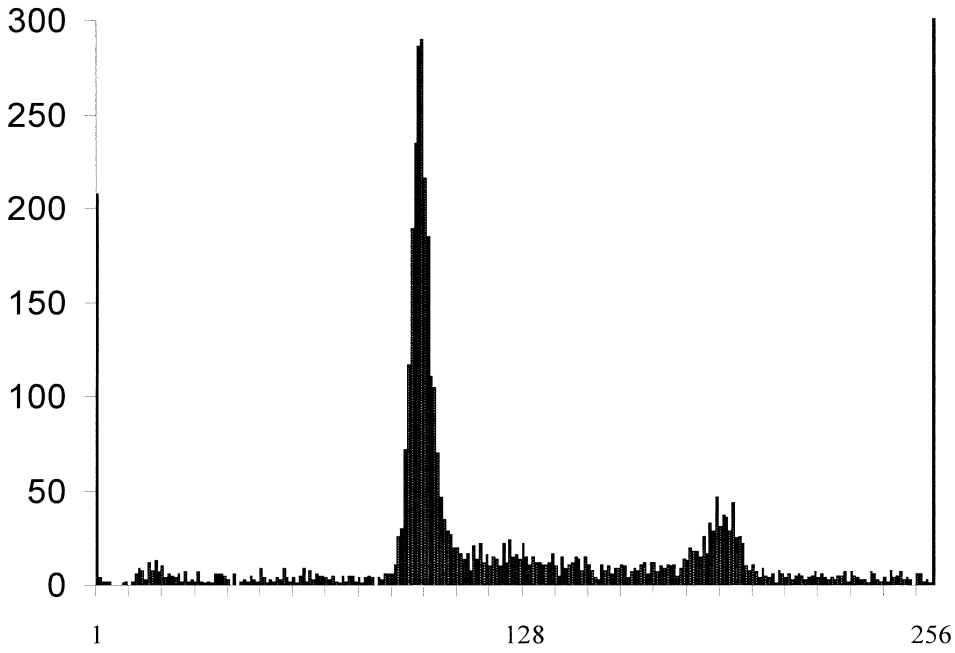

Figure 2. Chinese hamster ovary cells histogram.

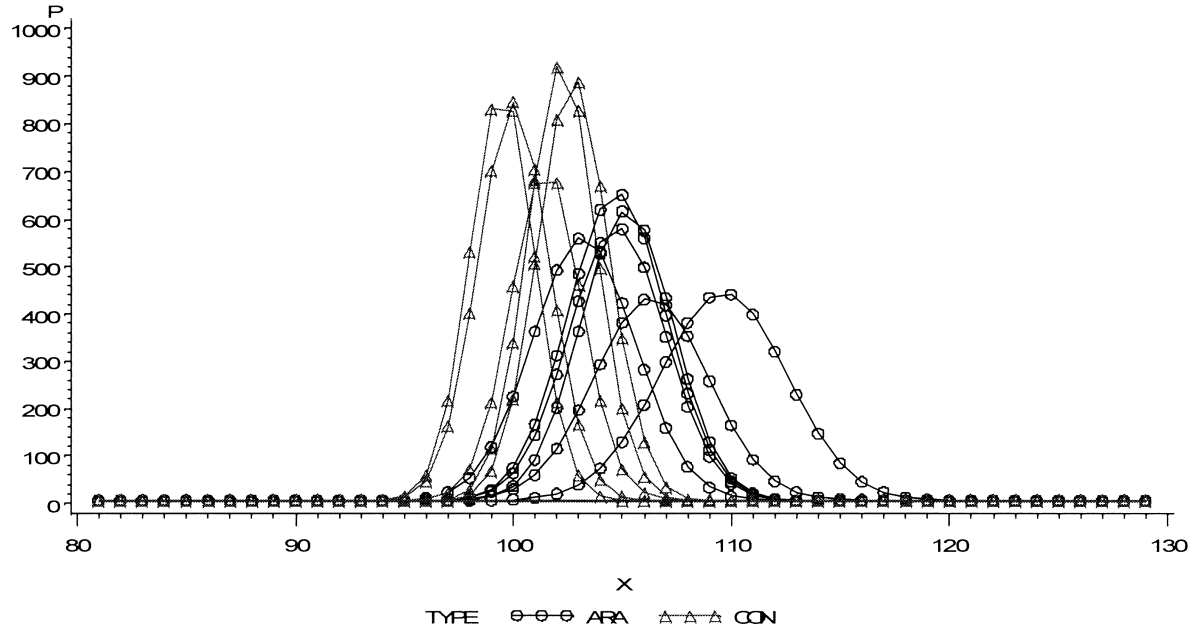

Figure 3. Bone marrow cells histograms for each replication.

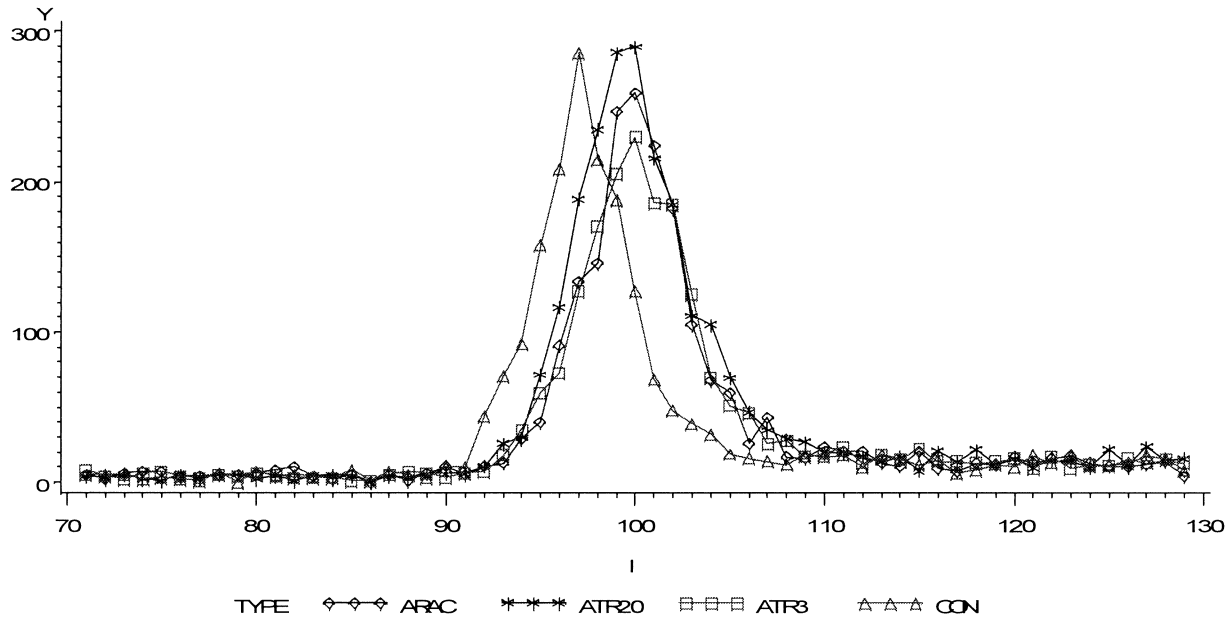

Figure 4. One replication of Chinese hamster ovary cell histograms for each treatment.

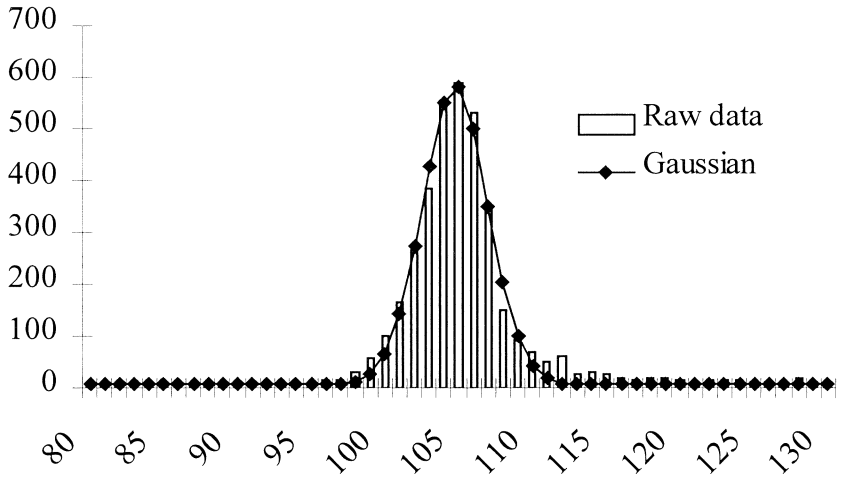

Figure 5. Bone marrow cells histogram and fitted Gaussian function.

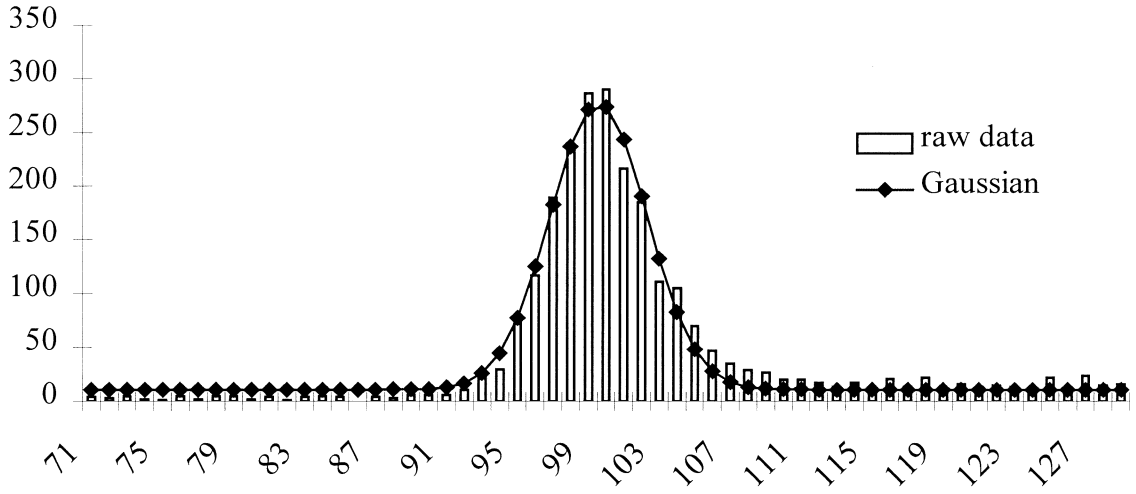

Figure 6. Chinese hamster ovary cells histogram and fitted Gaussian function.

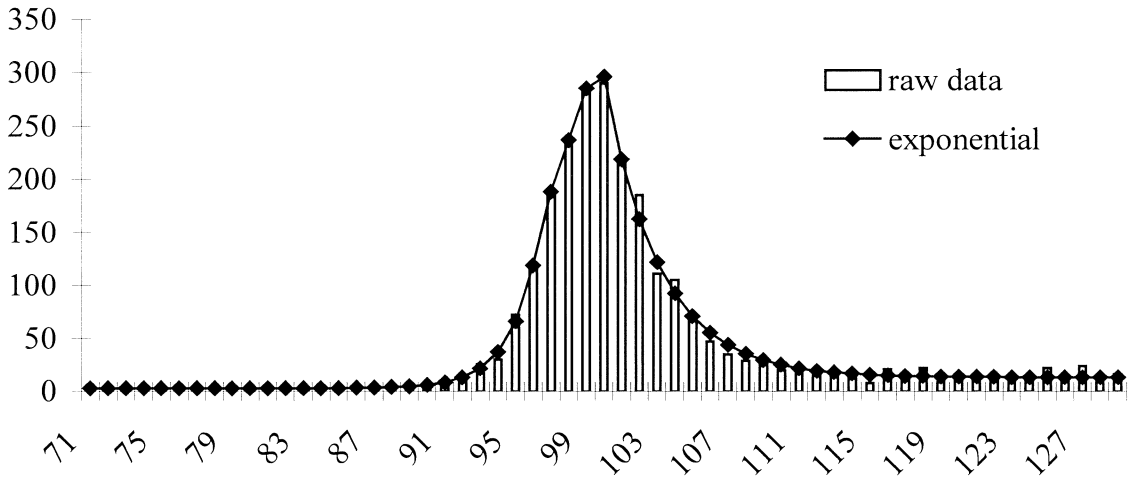

Figure 7. Chinese hamster ovary cells histogram and fitted composite exponential functions.

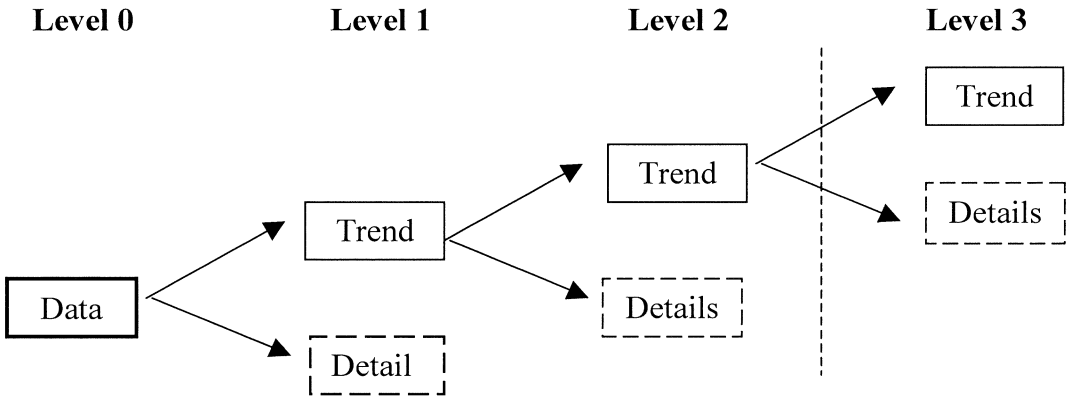

Figure 8. Wavelet decomposition to the second level.

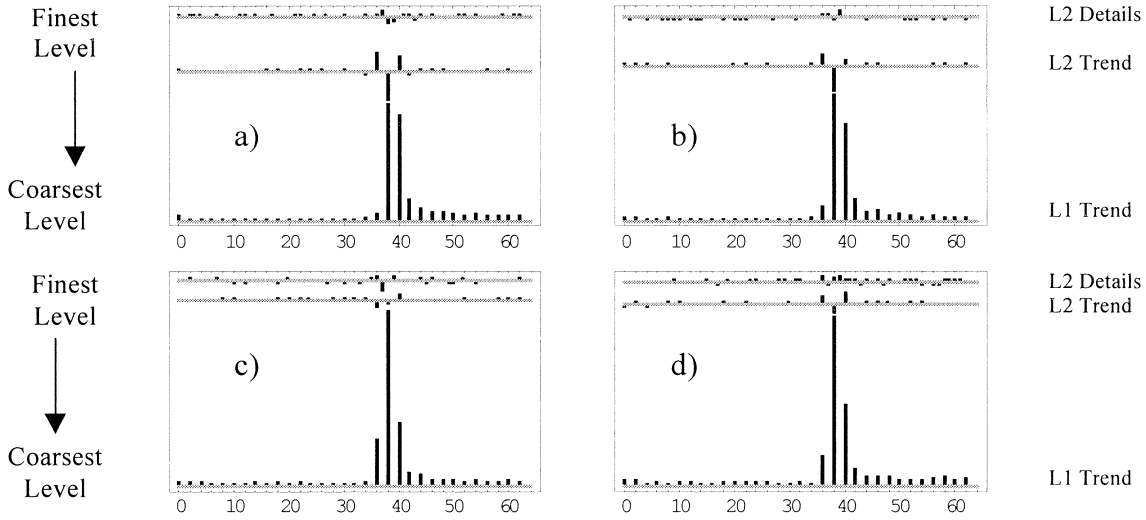

Figure 9. Coefficients from decomposition of the data in the wavelet basis, a) atrazine 20ppm, b) atrazine 3ppm, c) ara-C, and d) control.

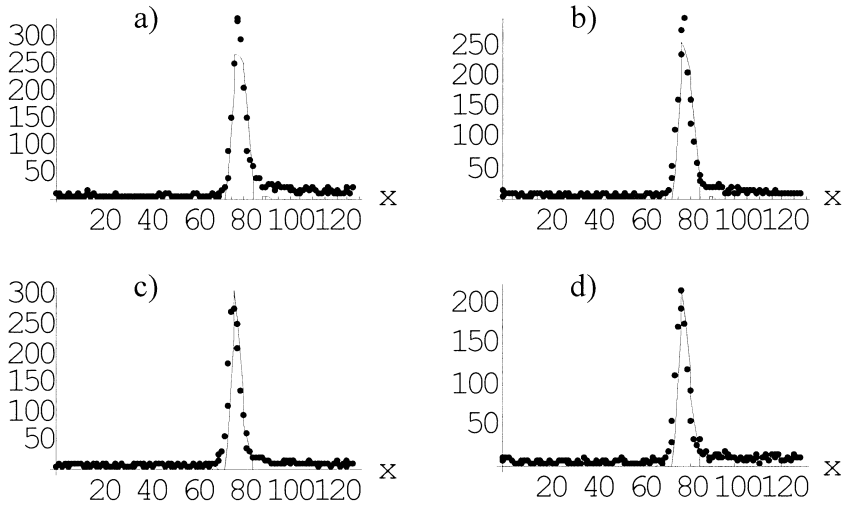

Figure 10. Reconstruction of the data from the compressed coefficients, a) atrazine 20ppm, b) atrazine 3ppm, c) ara-C, and d) control.