Genome-wide expression analysis of environmental stress in the cyanobacterium Synechocystis PCC 6803

Aguirre von Wobeser, E.

Publication date
2010

Chapter 3

A new normalization method for microarray data based on the generalized extreme value distribution

Eneas Aguirre von Wobeser[1], Bas W. Ibelings[2], Hans C.P. Matthijs[1] & Jef Huisman[1,3]

[1]*Aquatic Microbiology, Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Nieuwe Achtergracht 127, 1018 WS Amsterdam, The Netherlands*

[2]*Netherlands Institute of Ecology (NIOO-KNAW), Centre for Limnology, Rijksstraatweg 6, 3631 AC Nieuwersluis, The Netherlands.*

[3]*Corresponding author: Jef Huisman, phone: +31 20 5257085, fax: +31 20 5257064, Email: J.Huisman@uva.nl*

ABSTRACT

Microarray experiments produce data sets with thousands of values for each sample analyzed. Data sets from different samples generally show systematic differences due to experimental artefacts that may obscure the relevant biological signal. These artefacts need to be corrected by normalization of the data. Several normalization techniques exist, that vary both methodologically and conceptually. Here we present a new parametric normalization method that transforms microarray data with simple deterministic manipulations. This is achieved by modelling the distributions of the data sets using the generalized extreme value (GEV) distribution, and subsequent transformation of the data according to the estimated distributional parameters. The method normalizes microarray data to the same GEV distribution, while preserving the internal structure of the original data sets. This facilitates the determination of subtle changes in gene expression. GEV-based normalization is especially suitable for the detection of subtle changes in the expression of highly expressed genes. The method presents a good alternative for routine microarray data normalization.

INTRODUCTION

Microarray experiments typically compare the expression levels of thousands of genes of an organism under two or more experimental, histological or genetic conditions. These experiments produce complex data-sets, where besides random experimental noise, several systematic linear and non-linear artefacts compromise the relevant biological signal (Kerr and Churchill, 2001; Workman *et al.*, 2002). Therefore, an essential pre-processing step in microarray data analysis consists of normalizing the raw signals, to make them comparable (Quackenbush, 2002). Ideally, normalization should transform the raw data in a way that the final signal is proportional to the original concentration of mRNA target corresponding to each gene on the array, as if the number of cells used and the quantity of RNA hybridized on all arrays in the microarray were the same. In other words, normalization intends to uncover the 'true' changes in gene expression. The normalization strategy then consists of defining the characteristics that a normalized dataset should have and designing a suitable strategy to approach these characteristics as close as possible.

Several normalization strategies exist that vary both in the desired characteristics of the outcome and on the kinds of manipulations that they utilize. Most methods assume that the magnitude of the experimental error varies with the intensity of the signal (Quackenbush, 2002), for instance due to the chemical and physical properties of the dyes (Berger *et al.*, 2004). Removal of this artefact is termed within-array normalization. Lowess normalization (Yang *et al*. 2002), for example, explicitly removes trends associated with the mean signal intensity. These methods can be utilized as a first step before between-array normalization.

Between-array normalization seeks to obtain homogeneity in the distribution of signal intensities across different arrays. A simple approach consists of scaling the datasets from different arrays to have the same mean and standard deviation. Quantile-based methods extend this idea to adjust the data to a common distribution, keeping the original ranking of the data, for example using spline-interpolation (Workman *et al.* 2002) or using the mean of the data with the same rank in the different data-series (Bolstad *et al.*, 2003). In these

distribution-targeted methods, intensity-dependent biases are removed as a side-effect, rendering within-array normalization unnecessary. Another useful normalization method, known as variance stabilization (Huber *et al*., 2002), substitutes the commonly used log-ratios by a hyperbolic sine function that can be tuned to produce normalized data-series. Analysis of variance has also been used for microarray normalization purposes. In this approach, the effects of different dyes, different slides and other artefacts are estimated explicitly by ANOVA models and subsequently separated from the relevant gene expression signal (Kerr and Churchill, 2001; Wolfinger *et al.*, 2001).

Methods that adjust the data series of several arrays to the same distribution improve the ability to detect differential gene expression. The advantage of non-linear methods, like the quantile-based approach, may be that they maintain structural non-linearities in the data (Workman *et al.*, 2002). However, we have observed that forcing of data to the same target distribution in quantile-based methods can be too strict, and might result in over-fitting. Especially for genes at the extremes of the distribution (i.e., highly expressed genes), quantile-based transformations can alter the relation between data-points of different samples in a way that is not necessarily justified by the data.

As an alternative, parametric approaches have been proposed, where the distribution of signal intensities is described and transformed according to a parametric model (Siderov *et al.*, 2002; Konishi, 2004). However, existing parametric models seem to have their disadvantages. The model proposed by Siderov *et al.* (2002) uses two parameter sets to model the distribution, and the data are assigned to either one or the other parameter set. Therefore, data from the same array can be transformed by two different models, possibly introducing artefacts during the transformation. The recent model proposed by Konishi (2004) depends on the inclusion of a background parameter that may yield logarithms of negative numbers for signal intensities below this background value. This may result in the exclusion of many data-points, which are labelled as *not-detected*. This can be problematic if a gene is not expressed in one treatment but induced in another treatment, since the data for that gene would be lost in the not expressed condition, hampering the comparison between the treatments.

In this paper, we introduce a new parametric normalization method that transforms microarray data to a similar distribution using the Generalized Extreme Value (GEV) distribution. The generalized extreme value distribution is a distribution with very flexible tails (Kotz and Nadarajah, 2000; Coles, 2001). As its name implies, it is traditionally used to obtain accurate descriptions of the extreme values (*e.g.* minima and maxima) of distributions. Here, we exploit the flexibility of the GEV distribution to model the distributions of signal intensities in microarray datasets. Due to its flexible tails, the GEV distribution is particularly suitable for the accurate description of highly expressed genes.

MATERIALS AND METHODS

The Generalized Extreme Value distribution

We aimed to develop a normalization method that models the distribution of signal intensities obtained from microarray experiments, and that can transform the intensity data

from different microarrays to the same distribution. For this purpose, we searched for a distribution that fitted well to different microarray datasets. We tested several distributions, including the gamma distribution, the normal distribution and the log-normal distribution against datasets from 94 different microarray experiments. We found that the generalized extreme value (GEV) distribution was flexible enough to fit to the distribution of all microarray datasets. The GEV distribution, originally developed for extreme value theory, is a very flexible distribution comprising three different types of extreme value distributions known as the Gumbel, Fréchet and Weibull distributions (Kotz and Nadarajah, 2000; Coles, 2001). The GEV distribution is frequently applied to model extreme events in economy (Neftci, 2000), ecology (Gaines and Denny, 1993), meteorology (Katz, 1999; Palutikof et al, 1999), and earth sciences (Chowdhury 1991).

The probability density function of the GEV distribution is defined as:

$$y = f(x \mid \mu, \sigma, \gamma) = \left(\frac{1}{\sigma}\right)\left(1+\gamma\frac{(x-\mu)}{\sigma}\right)^{-1-\frac{1}{\gamma}} \exp\left[-\left(1+\gamma\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\gamma}}\right] \tag{1}$$

for

$$1+\gamma\frac{(x-\mu)}{\sigma} > 0 \tag{2}$$

The GEV distribution consists of three parameters. The location parameter μ displaces the distribution along the x-axis, the scale parameter σ modifies the spread of the distribution, and the shape of the distribution is modified by the shape parameter γ. The mean of the GEV distribution is:

$$E(x) = \mu - \frac{\sigma}{\gamma} + \frac{\sigma}{\gamma}\left[\Gamma(1-\gamma)\right] \tag{3}$$

and the variance is:

$$Var(x) = \frac{\sigma^2}{\gamma^2}\left(\Gamma(1-2\gamma) - [\Gamma(1-\gamma)]^2\right) \tag{4}$$

where Γ(z) is the gamma function. The skewness and the kurtosis are functions of γ only. Therefore, the shape of the GEV distribution is largely controlled by γ.

An example of the GEV distribution is shown in Figure 1A.

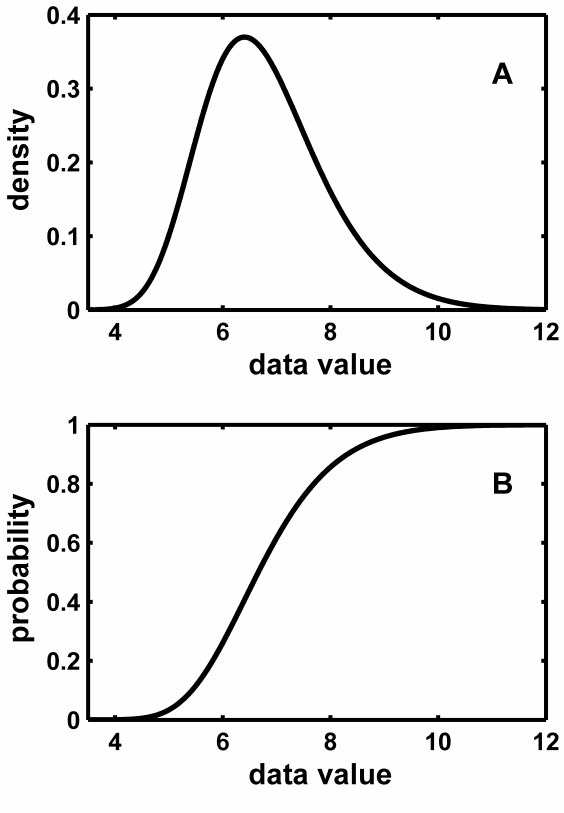

Figure 1. The generalized extreme value distribution. A. An example of a probability density function (PDF) based on the GEV distribution. B. The cumulative distribution function (CDF) of the same GEV distribution.

Our data consisted of the signal intensities of spots on microarrays. The signal intensities were presented on a natural logarithmic scale. The GEV distribution (Eq.1) was fitted to the distribution of the log intensities using maximum likelihood estimation (Prescott and Walden, 1980). Maximum likelihood estimation is a standard procedure for fitting distributions to data. It searches for the parameter values of μ, σ, and γ that maximize the probability that the modelled distribution is best described by the observed data. For our two-channel microarrays, data from each channel were treated as a separate dataset.

Between-array normalization

The next step consisted of transforming the datasets of all microarrays to the same GEV distribution. This requires definition of a target GEV distribution to which the data will be transformed, and a transformation method. We choose for a target GEV distribution, based on the median values of the estimated parameters as target μ_t, σ_t, and γ_t.

The transformation made use of the cumulative distribution function (CDF) of the GEV distribution. The CDF gives the probability that a random draw from the PDF is equal to or less than a given value x. In mathematical terms, the cumulative distribution function F(x) is related to the probability density function f(x) as follows:

$$F(x \mid \mu, \sigma, \gamma) = \int_{-\infty}^{x} f(x')dx' = \exp\left[-\left(1+\gamma\frac{(x-\mu)}{\sigma}\right)^{-\frac{1}{\gamma}}\right] \qquad (5)$$

Graphically, the CDF is an S-shaped function rising from 0 to 1 with increasing x (Figure 1B). Figure 2 shows how an x-value is mapped from the fitted GEV distribution to the target GEV distribution. This transformation is obtained by first calculating the CDF for the parameter values of the fitted GEV distribution, using Eq.5, and subsequently calculating the inverse of this CDF (known as the percent-point function) using the parameter values of the target GEV distribution.

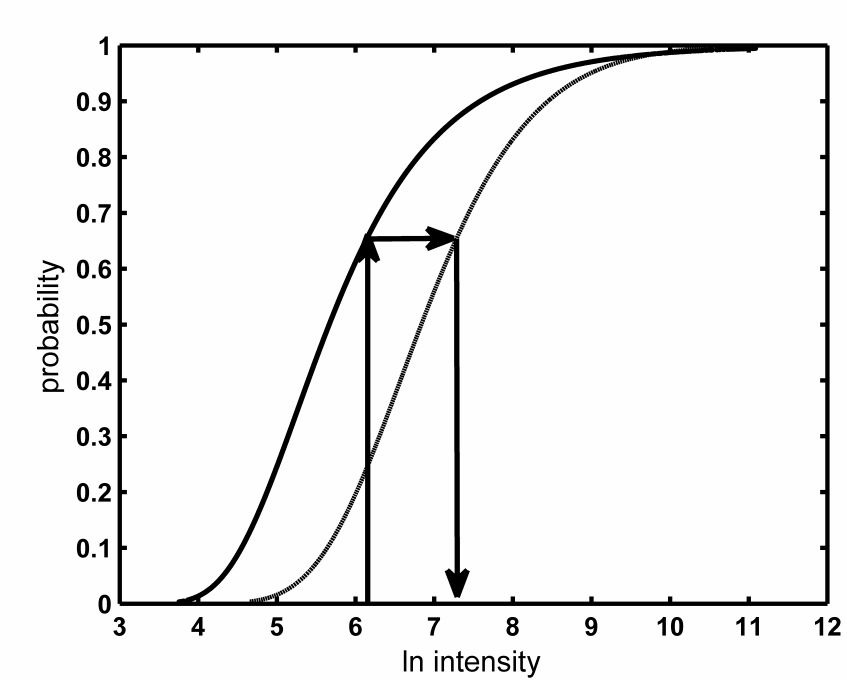

Figure 2. Mapping of the raw data to the target distribution using the cumulative distribution function (CDF). The graph shows the CDF of the raw data (dashed line), obtained after fitting of the data to the GEV distribution, and the CDF of the target distribution for normalization (solid line). The arrows illustrate the mapping from the log intensity in the raw data to the log intensity in the normalized data.

In view of Eq.5, this inverse-point function reads:

$$x_t = \frac{[-\ln(F(x))]^{-\gamma_t}}{\gamma_t}\sigma_t + \mu_t \tag{6}$$

where x are the original data and x_t are the transformed data, F(x) is defined by Eq.5, and μ_t, σ_t, and γ_t are the parameters of the target distribution.

Hence, inserting Eq. 5 into Eq. 6, the original data (expressed on natural log scale) were transformed according to the equation:

$$x_t = \frac{\left[1 + \gamma(x - \mu)/\sigma\right]^{\gamma_t/\gamma}}{\gamma_t}\sigma_t + \mu_t \quad (7)$$

Application of this transformation to different microarray datasets yields datasets that all fit the same target GEV distribution. The algorithm was implemented in the software package MATLAB 6.5 with Statistics Toolbox 4.0 (The Mathworks, USA). The implementation of the algorithm is shown in Appendix 3.

Experimental data

As test material, we used microarray data from a nitrogen starvation experiment with the freshwater cyanobacterium *Synechocystis* PCC 6803. The *Synechocystis* cells were grown in Erlenmeyer flasks with BG-11 medium, which is a nitrogen-rich growth medium specifically developed for cyanobacteria (Rippka *et al.*, 1979). The Erlenmeyer flasks were incubated at a temperature of 30°C under continuous illumination at 50 µmol photons m^{-2} s^{-1}. Subsequently, each of the cultures was split into two parts: one part was resuspended in nitrogen-free BG-11 medium to induce nitrogen starvation (treatment experiments), while the other part was resuspended in full BG-11 medium (control experiments). All experiments were run in triplicate. The cells of the control experiments were harvested by centrifugation and frozen at -80°C. The cells in the treatment cultures were harvested after 6, 12, 24 and 96 h of nitrogen starvation. After 96 h, nitrate was added to a final concentration of 20 mM, and subsequently the treatment cultures were harvested after 6 and 12 hours of recovery from nitrogen starvation. The harvested samples were hybridized to oligonucleotide microarrays. In total, this nitrogen starvation experiment consisted of 34 samples hybridized to 17 microarrays. Each microarray contained a control sample labelled with Cy5 and a treatment sample labelled with Cy3.

We also fitted the GEV distribution to microarray data from several other *Synechocystis* experiments. These datasets were based on a salt stress experiment (Chapter 2 of this thesis), a carbon stress experiment (Chapter 6 of this thesis; Eisenhut *et al.*, 2007), a nitrogen limitation experiment using continuous culture (Chapter 5 of this thesis) and a nitrogen limitation experiment using batch culture (Chapter 4 of this thesis). Combined with the nitrogen starvation experiment described in the previous paragraph, this yielded 94 data sets in total.

Microarray design

RNA extraction and labelling were performed as described in detail in Chapter 4 of this thesis. The labelled RNA was hybridized to custom-designed long oligonucleotide (45-60 bases) microarrays (Chapter 2 or this thesis), synthesized by Agilent using ink-jet technology (Hughes *et al.*, 2001). Each of the 3264 genes of *Synechocystis* PCC 6803 was represented in the microarray design by 1 to 4 different oligonucleotide probes dependent on the length of the ORF concerned. In total, the microarray consisted of 8091 probes representing the entire genome of *Synechocystis* PCC 6803. The full sequence and

annotation of *Synechocystis* PCC 6803 is available at CyanoBase (Kaneko *et al.,* 1996; http://www.kazusa.or.jp/cyano/). The microarrays were scanned at 10 micron resolution in an Agilent microarray scanner and the signal intensities of the spots were extracted using Feature Extraction Software 7.5 (Agilent Technologies). The signal intensities were transformed to natural logarithms.

RESULTS

The GEV distribution fitted well to all 94 data sets. As an illustration, examples of the fit of the probability density function (PDF) of the GEV distribution to three very different data sets are shown in Figure 3A-C. The similarity of the distribution of the raw data and the probability density function of the GEV distribution illustrate the suitability of the GEV distribution to model the data series. Figure 3D-F shows the same three data sets, after transformation of the data to a similar target distribution. Maximum likelihood estimates of the parameters γ, μ and σ of the GEV distribution for the 34 data sets of the nitrogen starvation experiment are given in Table 1.

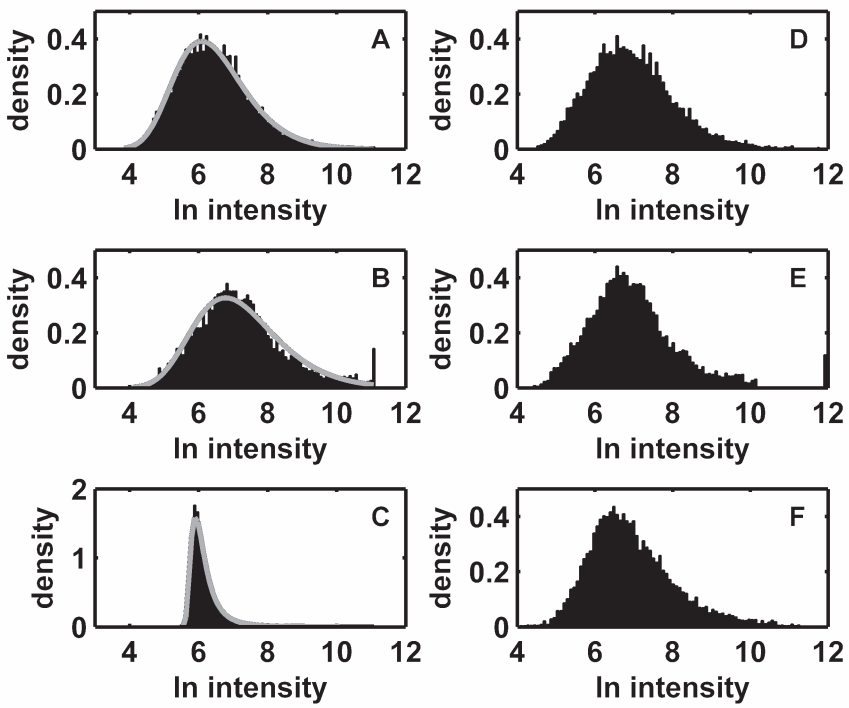

Figure 3. A-C) Three examples of the generalized extreme value distribution fitted to raw microarray data of the cyanobacterium *Synechocystis* PCC 6803. The microarray data are obtained from A) Control treatment in the nitrogen-starvation experiment. B) N-starvation treatment in the nitrogen-starvation experiment. C) Salt-stressed cells. The control treatment and N-starvation treatment in (A) and (B) were hybridized to the same microarray slide. D-F) The same microarray data sets as in A-C after normalization with the GEV distribution.

Data set	GEV parameters		
	γ	σ	μ
1	0.007 ± 0.015	0.839 ± 0.015	5.794 ± 0.020
2	-0.061 ± 0.015	0.964 ± 0.016	6.588 ± 0.023
3	-0.086 ± 0.014	1.017 ± 0.017	5.970± 0.025
4	-0.066 ± 0.015	1.021 ± 0.017	6.504± 0.025
5	-0.057 ± 0.015	0.976 ± 0.017	6.118± 0.024
6	-0.081 ± 0.015	1.082 ± 0.018	7.131± 0.026
7	-0.100 ± 0.015	0.971 ± 0.017	6.263± 0.024
8	-0.074 ± 0.015	1.064 ± 0.018	7.029± 0.026
9	-0.074 ± 0.016	0.978 ± 0.017	6.518± 0.024
10	-0.068 ± 0.015	1.026 ± 0.017	7.104± 0.025
11	-0.068 ± 0.016	1.096 ± 0.019	5.897± 0.027
12	-0.058 ± 0.015	1.155 ± 0.020	6.852± 0.028
13	-0.115 ± 0.015	0.978 ± 0.017	6.434± 0.024
14	-0.052 ± 0.015	1.048 ± 0.018	7.296± 0.025
15	-0.124 ± 0.014	0.980 ± 0.017	6.576± 0.024
16	-0.083 ± 0.015	1.087 ± 0.018	7.387± 0.026
17	-0.082 ± 0.016	1.031 ± 0.018	5.847± 0.025
18	-0.054 ± 0.015	1.144 ± 0.019	6.830± 0.028
19	-0.132 ± 0.015	0.995 ± 0.017	6.972± 0.024
20	-0.075 ± 0.015	1.035 ± 0.017	7.092± 0.025
21	-0.099 ± 0.014	0.945 ± 0.016	5.989± 0.023
22	-0.053 ± 0.015	1.129 ± 0.019	6.659± 0.027
23	-0.103 ± 0.014	1.066 ± 0.018	6.183± 0.026
24	-0.043 ± 0.015	1.113 ± 0.019	6.620± 0.027
25	0.001 ± 0.016	1.094 ± 0.019	5.952± 0.027
26	-0.061 ± 0.015	1.151 ± 0.019	6.897± 0.028
27	-0.019 ± 0.015	1.032 ± 0.018	6.133± 0.025
28	-0.070 ± 0.015	1.004 ± 0.017	6.419± 0.024
29	0.046 ± 0.018	0.985 ± 0.018	6.032± 0.024
30	-0.062 ± 0.016	1.096 ± 0.019	7.379± 0.027
31	0.024 ± 0.017	0.940 ± 0.017	5.210± 0.023
32	-0.041 ± 0.015	1.079 ± 0.018	6.519± 0.026
33	0.041 ± 0.017	0.879 ± 0.016	5.194± 0.022
34	-0.034 ± 0.015	0.913 ± 0.016	5.879± 0.022

Table 1. Estimated parameters γ, σ and μ of the Generalized Extreme Value distribution, obtained by fitting the GEV distribution to each of the 34 data sets of the nitrogen-starvation experiment using maximum likelihood estimation.

To assess whether normalization of the data using the GEV distribution was successful, we plotted so-called MA-plots (Smyth and Speed, 2003). For each spot on each slide, we calculated the mean log intensity *A*:

$$A = \frac{1}{2}\left({}^2\log I_T + {}^2\log I_C\right) \tag{8}$$

where I_T is the signal intensity of the treatment sample, and I_C is the signal intensity of the corresponding control sample. The change in gene expression in response to the treatment was expressed as the log ratio M:

$$M = {}^2\log\left(\frac{I_T}{I_C}\right) \tag{9}$$

Figure 4 shows a typical example of MA-plots before and after normalization of the data with the GEV distribution. The example is based on the nitrogen starvation experiment, using a treatment sample and its corresponding control sample after 24 hours of N starvation. In the MA plot of the raw data, many log ratios clustered below zero, and there was a slight tendency for the log ratios to decrease with mean log intensity (Figure 4A). In the MA-plot of the normalized data, the log ratios clustered around zero and were independent of the mean log intensity (Figure 4B). This illustrates that normalization of the data with the GEV distribution successfully removed potential bias in the log ratios.

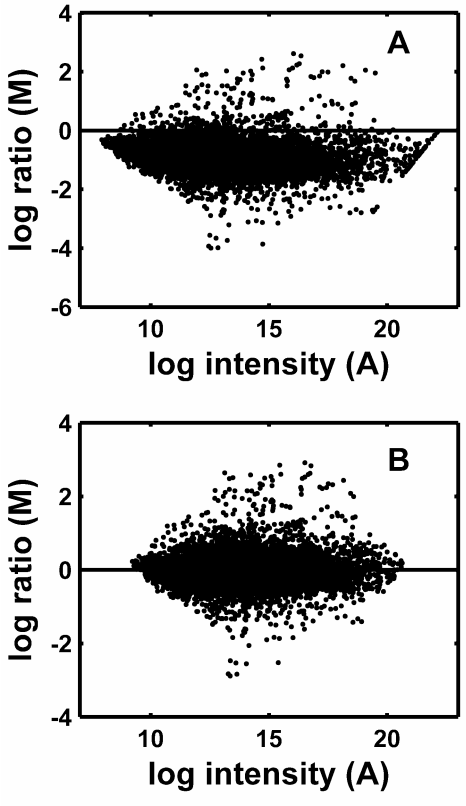

Figure 4. MA plots, showing the relation between the log ratio and the mean log intensity of the treatment and control data. (A) MA plot of the raw data. (B) MA plot of the same data after normalization using the GEV distribution. The treatment and control data were obtained from the nitrogen starvation experiment shown in Figure 3A,B.

Another requirement for a successful normalization method is that the internal data structure is preserved as much as possible. That is, genes with high expression levels in the raw data should, generally, also show high expression levels in the transformed data. To test how much the data structure was altered by the normalization process, we computed the correlation between the log intensity of the raw data and the corresponding normalized data (Figure 5). Product-moment correlation coefficients of raw versus normalized data were calculated

for each probe, using the signal intensities from the 17 treatment and 17 control samples of the nitrogen starvation experiment. Thus, each correlation coefficient was based on 34 data points. Since we covered the *Synechocystis* genome by 8091 probes, this resulted in 8091 correlation coefficients. The distribution of the correlation coefficients obtained by normalization with the GEV distribution is shown in Figure 5A. This shows that the correlation between the normalized and raw data is generally quite high, with a mean correlation coefficient of 0.52 and a median correlation coefficient of 0.58.

To compare the performance of normalization by the GEV distribution with other normalization methods, we performed a similar correlation exercise using the same raw data normalized by the variance stabilization method (Huber *et al*., 2002) and normalized by two quantile-based methods (Workman *et al.*, 2002; Bolstad *et al.*, 2003). The median of the correlation coefficients obtained after normalization by the GEV distribution was significantly closer to 1 than the median of the correlation coefficients obtained by the variance stabilization method, and were equal to the medians of the correlation coefficients obtained by the two quantile-based methods (Figure 5; Kruskal Wallis,test $p < 0.001$; $N = 24273$, $df = 2$, $\chi^2 = 128.1$). This showed that a nonparametric approach using a quantile-based normalization method and the new parametric approach using the GEV distribution were equally effective in conserving the information contained in the raw data.

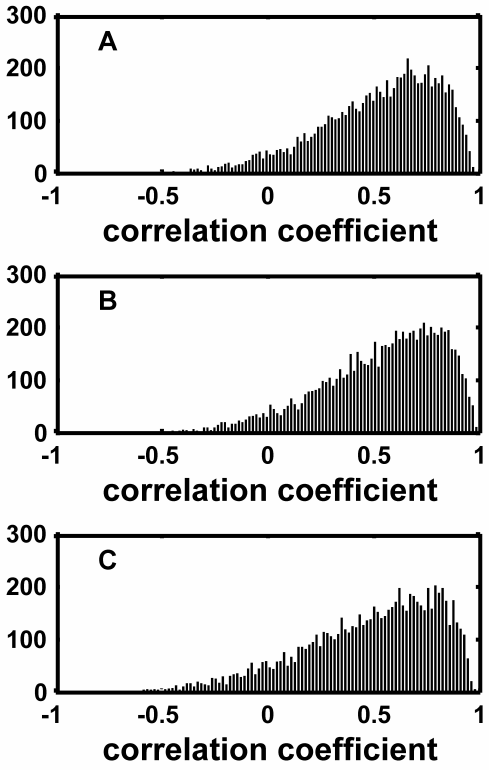

Figure 5. Distribution of the correlation coefficient between the log intensity of the raw data and normalized data. Normalization of the data was based on A) the GEV distribution, B) a quantile-based method using default settings (Bolstad *et al.,* 2003). The Bolstad *et al.* (2003) and the Workman *et al.* (2002) algorithms for quantile-based normalization produced essentially the same results. C) a variance stabilization method (Huber *et al*., 2002). The data were obtained from the nitrogen starvation experiment, consisting of 34 microarray data sets each covering the entire genome of the cyanobacterium *Synechocystis* PCC 6803. The correlation coefficients were calculated for each probe across all 34 microarrays (yielding 8091 correlation coefficients, with n = 34 for each correlation coefficient).

While normalization by both the quantile-based methods and the GEV-based method were equally effective in preserving the information contained in the data, there were differences in the individual genes. The quantile-based normalization method yielded slightly higher correlations for the lowest intensity data (Figure 6). These data include the signals of genes that were hardly or not expressed in the experiment. While some of these lowly expressed genes may have low magnitude expression changes with physiological relevance, detection of those expression changes would probably require large sample sizes even when they are normalized effectively. Conversely, the GEV-based normalization method yielded higher correlations for genes that had the highest expression levels in the raw data. These signals correspond to genes that generally result in highly abundant protein levels. Their expression changes are usually easy to detect, and they result in major physiological changes. A normalization method with good performance overall, and with a better ability to conserve the information in the raw data of these highly expressed genes is desirable, especially for experiments where highly expressed genes show differential expression.

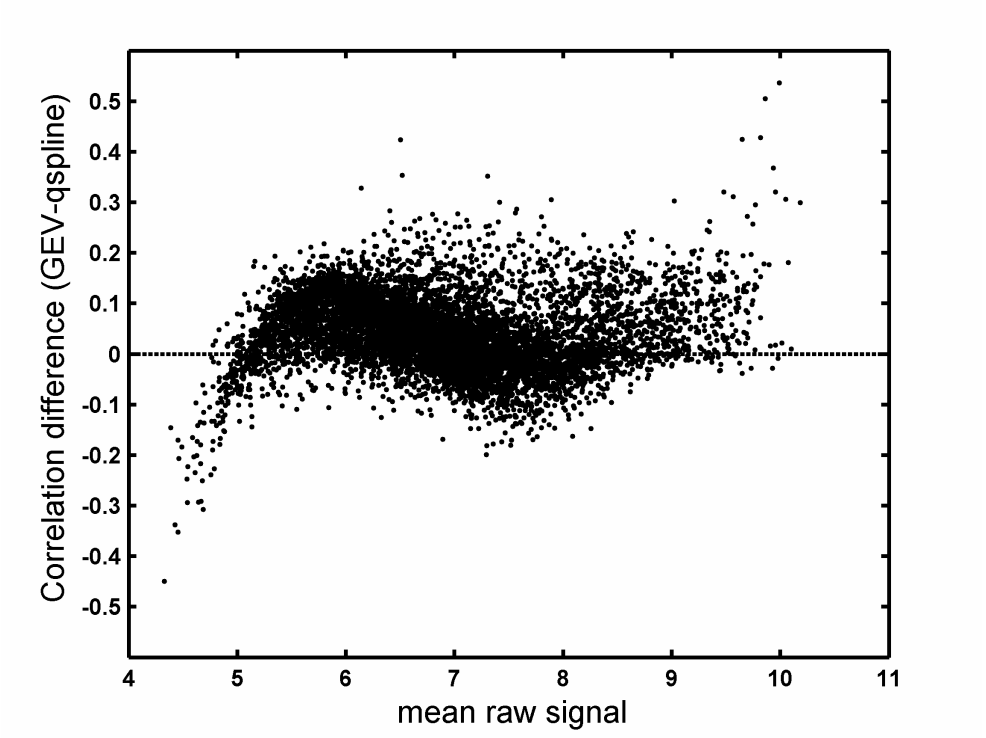

Figure 6. Differences in preservation of the structure of the raw data as a function of mean signal intensity using quantile-based normalization and the normalization method based on the GEV distribution. For both normalization methods, the correlation coefficient between the raw data and the normalized data was calculated for each probe. Then, the difference between the correlation coefficient obtained by GEV normalization and the correlation coefficient obtained by quantile-based normalization was taken. Positive values represent probes where the correlation was greater using the GEV-based method, and negative values represent probes where it was greater using the quantile-based method.

DISCUSSION

In this work, we have proposed a new alternative for microarray data normalization. The normalizing algorithm is based on the generalized extreme value distribution, and offers a parametric approach that adjusts the data to a similar GEV distribution between arrays. By applying a simple parametric transformation to each data-set, we avoided forcing the data to produce the target distribution exactly. Instead, the data were transformed to produce the target distribution well enough, while still allowing small variations particular to each data-set (Figure 3). These variations resulted in an improved conservation of the information contained in the original data for high signal genes, as compared to other methods (Figures 5 and 6). This improvement for highly expressed genes was at a cost for the lowest expressed genes, where the apparent conservation of information decreased. While highly expressed genes include many known important genes, which also produce high levels of proteins in the cells (like photosynthetic genes in cyanobacteria), the genes with lowest expression levels are mostly unknown and hypothetical genes, some of which might not have been expressed at all under the studied conditions. Therefore, it is arguably desirable to favour conservation of the information content of highly expressed genes.

The adequacy of the GEV distribution to conserve the information of highly expressed genes is a result of the general flexibility of its tails. This flexibility results from the presence of the shape parameter γ, that allows for many differently shaped distributions, including the Gumbel family (when γ tends to 0), the Fréchet family (when $\gamma > 0$) and the Weibull family of distributions (when $\gamma < 0$). Moreover, because the GEV distribution is constrained to the domain specified by equation 1b, it allows for both very sharp and smooth endings at the lower tail, both of which occur in microarray data series.

Many microarray data-series contain numerous moderately expressed genes but a few highly expressed genes residing in the upper tail of the log intensity distribution. Phycobilisome genes, for example, are amongst the highest expressed genes in Cyanobacteria growing under nitrogen-rich conditions (Chapters 4 and 5 of this thesis). Phycobilisomes are photosynthetic pigments rich in nitrogen that often give cyanobacteria their typical blue-green colour. However, changes in nitrogen availability can induce gradual changes in the expression of phycobilisome genes, because cyanobacteria tend to suppress the production of these pigments when nitrogen availability is reduced. Effective normalization of these highly expressed genes could thus yield detailed information on one of the most striking responses to nitrogen limitation, which results in visible colour changes of cyanobacteria due to reduced expression of their nitrogen-rich pigments. A normalization method tailored to the analysis of highly expressed genes, like the GEV-based algorithm presented here, can help exploit the information contained in these highly responsive genes. An experimental example of this approach is presented in Chapter 5 of this thesis.

In conclusion, we developed an algorithm to normalize microarray data that, similar to quantile-based normalization (Workman *et al*., 2002; Bolstad, 2003), results in data-series that fit the same distribution. Contrary to quantile-based methods, however, the parametric method presented here does not force the data to produce exactly the same target distribution, but preserves information contained in the distribution of the raw data. The method allows an improved analysis of highly expressed genes.

ACKNOWLEDGEMENTS

EAvW was financially supported by a scholarship from Consejo Nacional para la Ciencia y Tecnología (Mexico). The research of EAvW and JH was further supported by the Earth and Life Sciences Foundation (ALW), which is subsidized by the Netherlands Organization for Scientific Research (NWO). We kindly thank Dr. Timo Breit for helpful comments.

REFERENCES

Berger JA, Hautaniemi S, Järvinen A, Edgren H, Mitra SK, Astola,J. 2004. Optimized LOWESS normalization parameter selection for DNA microarray data. *BMC Bioinformatics* 5: 194.

Bolstad BM, Irizarri RA, Åstrand M, Speed T. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics 19: 185-193.

Chowdhury JU, Stedinger JR, Lu LH. 1991. Goodness-of-fit tests for regional generalized extreme value flood distributions. Water Resources Research 27: 1765-1776.

Coles S. 2001. An Introduction to Statistical Modeling of Extreme Values. Springer, Berlin.

Eisenhut M, von Wobeser EA, Jonas L, Schubert H, Ibelings BW, Bauwe H, Matthijs HCP, Hagemann M. 2007. Long-term response toward inorganic carbon limitation in wild type and glycolate turnover mutants of the cyanobacterium *Synechocystis* sp strain PCC 6803. Plant Physiology 144: 1946-1959.

Gaines SD, Danny MW. 1993. The largest, smallest, highest, lowest, longest, and shortest: extremes in ecology. Ecology 74: 1677-1692.

Huber W, von Heydebreck A, Sültmann H, Poustka A, Vingron M. 2002. Variance stabilization applied to microarray data calibration and to the quantification of differential expression. Bioinformatics 18: S96-S104.

Hughes TR, Mao M, Jones AR, Burchard J, Marton MJ, Shannon KW, Lefkowitz SM, Ziman M, Schelter JM, Meyer MR, Kobayashi S, Davis C, Dai H, He YDD, Stephaniants SB, Cavet G, Walker WL, West A, Coffey E, Shoemaker DD, Stoughton R, Blanchard AP, Friend SH, Linsley P. 2001. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. Nature Biotechnology 19: 342-347.

Kaneko T, Sato S, Kotani H, Tanaka A, Asamizu E, Nakamura Y, Miyajima N, Hirosawa M, Suqiura M, Sasamoto S, Kimura T, Hosouchi T, Matsuno A, Muraki A, Nakazaki N, Naruo K, Okumura S, Shimpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Tabata S. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA research 3: 109-136.

Katz RW. 1999. Extreme value theory for precipitation: sensitivity analysis for climate change. Advances in Water Resources 23: 133-139.

Kerr MK, Churchill GA. 2001. Statistical design and the analysis of gene expression microarray data. Genetics Research 77: 123-128.

Konishi T. 2004. Three-parameter lognormal distribution ubiquitously found in cDNA microarray data and its application to parametric data treatment. BMC Bioinformatics 5: 5.

Kotz S, Nadarajah S. 2000. Extreme Value Distributions: Theory and Applications. Imperial College Press, London.

Neftci SN. 2000. Value at risk calculations, extreme events and tail estimation. Journal of Derivatives. 2000: 1-15.

Palutikof JP, Brabson BB, Lister DH, Adcock ST. 1999. A review of methods to calculate extreme wind speeds. Meteorological Applications 6: 119-132.

Prescott P, Walden AT. 1980. Maximum likelihood estimation of the parameters of the generalized extreme value distribution. Biometrika 67: 723-724.

Quackenbush J. 2002. Microarray data normalization and transformation. Nature Genetics 32: 496-501.

Rippka R, Deruelles J, Waterbury JB, Herdman M, Stanier RY. 1979. Genetic assignments, strain histories and properties of pure cultures of cyanobacteria. Journal of General Microbiology 111: 1-61.

Siderov IA, Hosack DA, Gee D, Yang J, Cam MC, Lemicki RA, Dimitrov DS. 2002. Oligonucleotide microarray data distribution and normalization. Information Sciences 146: 67-73.

Smyth GK, Speed T. 2003. Normalization of cDNA microarray data. Methods 31: 265-273.

Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. 2001. Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology 8: 625-637.

Workman C, Jensen LJ, Jarmer H, Berka R, Gautier L, Nielsen HB, Saxild HH, Nielsen C, Brunak S, Knudsen S. 2002. A new non-linear normalization method for reducing variability in DNA microarray experiments. Genome Biology 3: research 0048.

Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed T. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Research 30:e15.