10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.1, June 4-6, 2007, Cancún, Mexico

NORMALISATION OF DNA ARRAY DATA TO FACILITATE THEIR USE IN BIOPROCESS DEVELOPMENT

Nicola Dawes and Jarka Glassey

*School of Chemical Engineering and Advanced Materials, Merz Court,
Newcastle University, Newcastle upon Tyne, NE1 7RU, UK*

Abstract: The application of DNA array is spreading at a very fast rate, although mainly in medical and systems biology studies. However, the recent FDA Process Analytical Technologies initiative provides a very exciting possibility of using transcriptomics, proteomics and metabolomics to gain deeper bioprocess understanding leading to more rapid process development and tighter control of critical process parameters. This contribution introduces a method of normalisation necessary for DNA data analysis usable in such circumstances – i.e. a limited number of experiments under varying operational conditions. We show that this normalisation technique provides more plausible data preparation for further data analysis.

Keywords: DNA array normalisation, transcriptomics, biotechnology, data analysis bioprocess development.

1.INTRODUCTION

DNA array technology provides the opportunity to study the many thousands of genes in an organism simultaneously. Gene expression in human tissues, cell lines and model organisms can be monitored to increase our understanding of the complex functional and metabolic pathways in such systems. This technology could also be used to enhance our knowledge of the effects of operating conditions upon production strains during process development. Although this technology has a promising potential, there are still issues to be addressed before reliable results can be obtained. Given the large variation in environmental conditions and/or production strains during process development an important issue is that of normalising the array data. This aligns the arrays to a common reference allowing direct comparison of different array hybridisation experiments, which may have been performed on different days or by different people.

The importance of normalisation of array data has been highlighted by a number of researchers in the past. For example, Edwards (Edwards, 2003) claims 'Normalization has profound effects on subsequent analysis, irrespective of the methodology used. Failure to normalize appropriately will generally lead to misleading conclusions.' An effective normalisation technique is one that reduces experimental variation or biases (noise) without affecting the measurement of the biological variation (signal). There are a number of well documented normalisation techniques ranging from simple scaling methods to more complex statistical approaches (Ballman *et al*, 2004, Quackenbush, 2002, Wu *et al*, 2005). 'Global' scaling methods are suitable for data sets where relatively few genes are expected to change between conditions and global array statistics such as median / mean expression levels can be used to scale the data. Statistical models require a good level of replication of experiments in order to give acceptable results (Kerr *et al*, 2000).

For some systems, normalisation is built into the experimental design using specific software for the DNA array system, such as the Affymetrix system. Some DNA array experiments, however, produce data that are not so easily normalised. This paper focuses on the normalisation of multi-condition time series gene expression data, generated using one-colour membrane macroarrays where the biological hypotheses being investigated are concerned with the interactions of genes involved in both the non-

specific and the specific stress response of the organism to phosphate limitation. This experimental design is very similar to the bioprocess development situation, where a number of operational conditions are varied in order to identify the optimum operating conditions. If DNA array data was to be useful in gaining deeper process understanding from such experiments, it is essential to develop a robust normalisation technique that would minimise the impact of noise without reducing the biological signal. Such a normalisation, referred to as Self-Consistent Set (SCS) normalisation, is proposed in this contribution and its robustness is tested in respect of the two user defined parameters required.

2. BIOLOGICAL SYSTEM

Data has been obtained from experiments where both the specific and non-specific response to phosphate stress have been investigated in a set of isogenic *Bacillus subtilis* mutants over time (Pragai and Harwood, 2002). The overall aim was to identify regulatory interactions between the σ^B-dependent general stress and Pho regulons in *B. subtilis*. Strains with null mutations in the key regulatory genes *sigB* and *phoR* were used to investigate the level of interaction between these two regulons. In total four strains were used: a wildtype strain (strain 168); *sigB*-null mutant; *phoR*-null mutant and a *sigB*-null, *phoR*-null (double) mutant. For a detailed description of the bacterial strains, plasmids, primers and medium used see Allenby et al (2005). Each strain was cultured in phosphate limiting conditions with typically four samples taken at specified times. These samples were processed and used in transcriptome analysis by hybridising to *B. subtilis* Panorama™ gene arrays (Sigma Genosys Biotechnologies Inc., The Woodlands, USA). The procedures of cell harvesting, RNA preparation, synthesis of radioactively labelled cDNA and hybridisation to the arrays as described by Eymann *et al* (2002) were followed. Arrays were exposed on a Fuji cassette for a pre-determined time. After exposure the cassette was scanned using a Storm phosphorimager to generate both .gel and .tiff image files. These digital images were imported into the software package ArrayVision™ to generate the data set.

3. NORMALISATION ISSUES

Currently the scientific literature mainly reports on more straightforward investigations of either single time points from a variety of strains / conditions or time profile of gene expression from a single strain using traditional methods of normalisation. However, functional genomics in particular will require an alternative approach to both experimental design and data analysis. Hence novel normalisation methods, such as the one proposed here, will become more appropriate.

Depending on the data array construction, the application of statistical modelling methods, which rely more heavily on data replication (Barash *et al.*, 2004), may be limited. A number of normalisation methods are based on the identification of a group of genes deemed to be invariant (Kepler *et al*., 2002), although often data from all the arrays in the experiment has to be compared to one array, taken as the baseline array. It is not logical to use a baseline approach with the *B. subtilis* data set due to biological variability between the strains as well as across the time trajectory within each strain as a result of growth and phosphate starvation. A similar argument would also apply in bioprocess development data. Therefore a new normalisation method, which does not require a selection of a baseline array, is proposed to identify a set of invariant genes globally, across all the arrays simultaneously as described below.

4. SCS NORMALISATION METHOD

Below is a mathematical description of the SCS algorithm. Figure 1 describes the process as a flow diagram.

For a ($m \times n$) data set where m is the number of genes (rows) and n is the number of arrays (columns), each element of the data set is g_{ij}, where i = 1 to m and j = 1 to n. For multi-strain time series data $n = s \times t$ where s = the number of strains from 1 to S and t = the number of time points from 1 to T.
Firstly, the contributions matrix C, is generated by dividing each gene expression value by the column total:

$$\left(c_i = \frac{g_i}{\sum_{i=1}^{m} g_i} \right)_j \qquad (1)$$

The average of each row of contributions is calculated and the top and bottom x% is disregarded.

$$R = \left\{ mx < rank\left(\frac{\sum_{j=1}^{n} c_j}{n} \right)_i < (m - mx) \right\} \qquad (2)$$

R is a vector of row numbers left once the top and bottom x% have been excluded. These row numbers are used to generate a new contributions matrix C_2 which is a subset of the matrix C. It is from this new (m-2mx) × n matrix that the initial SCS genes will be identified.

$$C_2 = (\forall R) C_2 \subseteq C \qquad (3)$$

For time point t:

$$SCS_i = \begin{cases} \left(\left|rank\left(C_{2(s1)}\right) - rank\left(C_{2(s2)}\right)\right|\right)_t < a \;\&\; \left(\left|rank\left(C_{2(s1)}\right) - rank\left(C_{2(s3)}\right)\right|\right)_t < a \\ \&\; \left(\left|rank\left(C_{2(s2)}\right) - rank\left(C_{2(s3)}\right)\right|\right)_t < a \end{cases} \quad (4)$$

In Equation 4, shown here there are three strains to consider, as the rank differences are calculated for each possible pairing of strains, the more strains there are, the more terms are needed in the equation. This is carried out for each time point to give SCS_1, SCS_2, SCS_3,……, SCS_T Then any gene that appears in all the SCS_t lists is deemed to be self-consistent across all strains and time points. These genes are then used to normalise the data by dividing each column of data by the sum of the SCS genes in that column.

$$\left(N_i = \frac{g_i}{\sum g_{SCS}} \right)_j \quad (5)$$

The process is then iterated k times by repeating each step from the calculation of the contributions until no change is seen between SCS_k and SCS_{k-1}.

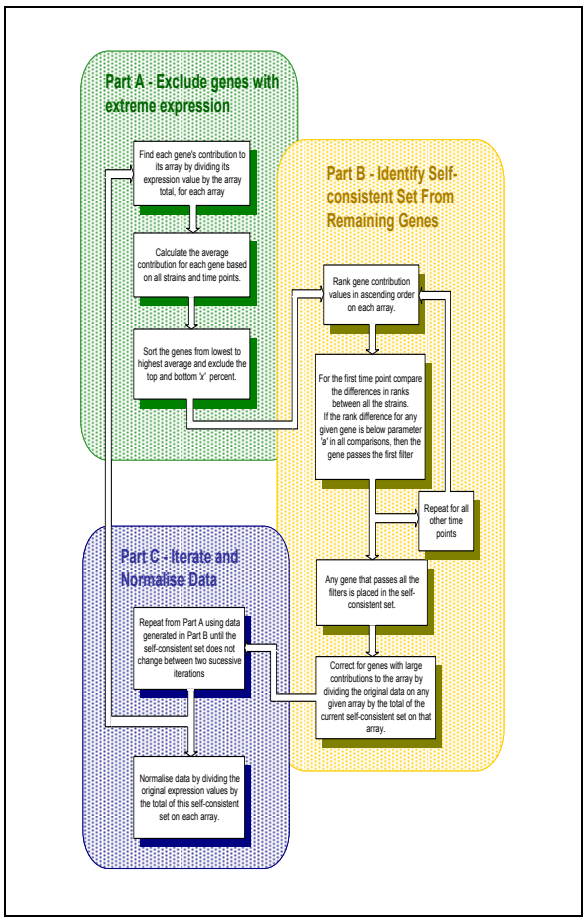

Fig. 1. Flow diagram of the SCS normalisation technique.

5. RESULTS

The number of SCS genes identified by the algorithm in the *B. subtilis* data set is largely dependant on the two user defined parameters, a, the absolute rank difference limit and x, the proportion of genes excluded. Figures 2 and 3 show how the number of genes identified as SCS changes depending on the values of a and x.

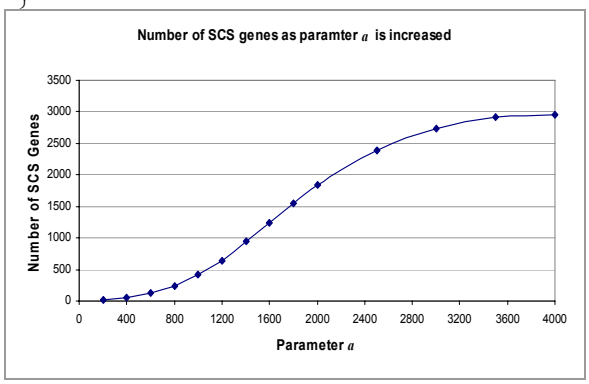

Fig. 2. Number of SCS genes identified with increasing value of parameter a.

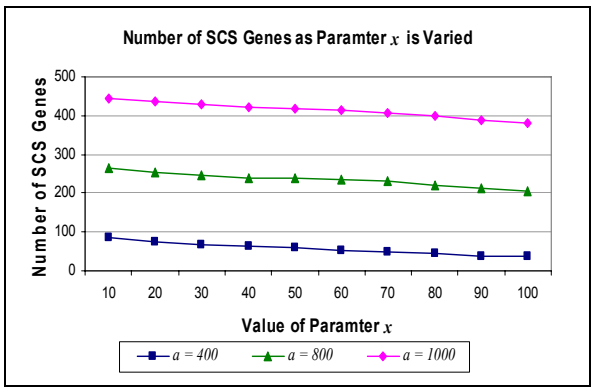

Fig. 3. Number of SCS genes identified with increasing value of parameter x for three settings of parameter a.

It is clear that parameter a has a greater influence over the size of the SCS than x. As a increases so does the number of genes which pass the filters and end up in the final SCS whereas when x is increased the stringency of the algorithm is increased as there are less potential SCS genes to start with. A wide range of a and x was investigated in this case to establish the sensitivity of the algorithm to these values, although it is clear that excessively large values of a result in an unrealistically large SCS gene sets.

The impact of increasing SCS set was assessed by a non-parametric Park score test (Park *et al*, 2001) which was used to assess the differential expression of the genes in strain-wise comparisons. In this case a comparison between sets obtained when parameter a values were set to 200, 400 and 600 was performed using the improvements in the number of differentially expressed genes (only data for genes scoring an extreme Park score of either 0 or 16 is shown here). Here 0 represents a gene overexpressed in strain 1 compared to strain 2, whereas a score of 16 refers to the opposite situation.

The bar chart in Figure 4 shows that there is no detectable improvement in using a = 600 (resulting in

132 SCS genes). There is a more notable difference in the number of genes with an extreme` Park score when a = 200. However this only results in 11 SCS genes, which is a rather low proportion of the total number of genes spotted on the array. Thus the value of a = 400 (63 SCS genes) was chosen for the future analysis of this data set.

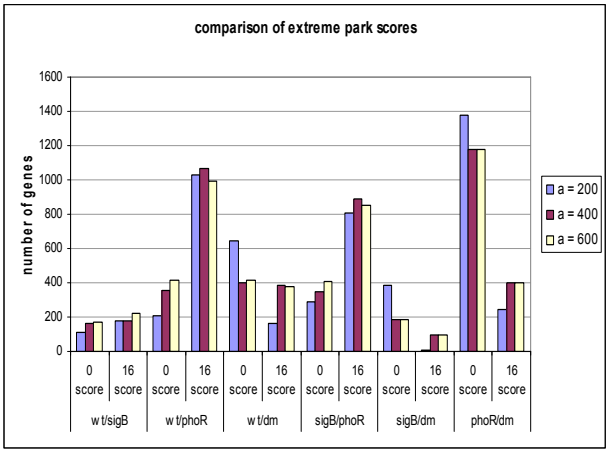

Fig. 4. Park scores for all strain comparisons using data normalised with SCS sets identified with three different settings of parameter a.

The application of the SCS algorithm, using a = 400 and x = 1 %, to the full *B. subtilis* data set resulted in 63 SCS genes (2% of the total number of genes spotted on the array). These genes are shown in Figure 5, grouped into functional categories as defined in SubtiList World-Wide Web Server, Institut Pasteur.

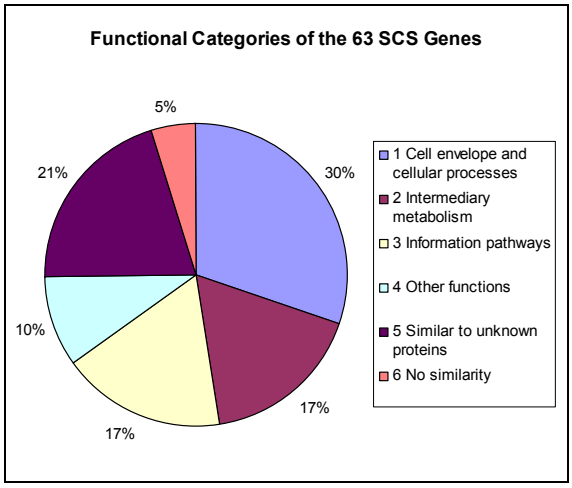

Fig. 5. Functional categories of the 63 SCS genes identified with a = 400 and x = 1%.

From the pie chart in Figure 5 it can be seen that the first category (Cell Envelope and Cellular Processes) is the most represented category, with 30% of the SCS genes belonging to it. A quarter of the SCS genes currently have an unknown function and category 4 (Other Functions) is the least represented with only 10% of the SCS genes being from this category. Category 2 (Intermediary Metabolism) and category 3 (Information Pathways) are evenly represented with 17% of SCS genes belonging to each of them.

This set of genes was compared to scientific literature reports describing the identification of genes that are essential for the survival of *B. subtilis* grown under nutritious conditions (Kobayashi, *et al*, 2003). In their study, following a systematic investigation, 271 genes were found to be essential to the organism. Of these 271 genes a small number are picked out by the SCS algorithm (namely *dnaE*, *ftsZ*, *pgk*, *rplD*, *rplJ* and *yurV*). Further to this, around 30% of the SCS genes are either located in close proximity in the genome or part of the same operon as genes listed as essential to the organism. However, the SCS genes are not expected to match to the list of essential genes too closely as these genes, critical to the organism's survival, were identified by culturing *B. subtilis* in nutritious conditions. Therefore some of these genes may behave differently in the phosphate limiting conditions or the mutant strains used in these experiments. For example, *tagA, B, D* and *F* are listed as essential genes but they are also under the control of *phoR* (Pragai and Harwood, 2002) which is inactivated in the *phoR*-null mutant and therefore would not be expected to fulfil the self-consistent criteria defined in this work.

5.1. Noise reduction by normalisation

The raw expression values for SCS genes vary over the time course of the experiments, however the rank positions of the contributions of these genes are similar for any given time point in each experiment. Therefore the SCS algorithm results in a set of genes which behave similarly in each strain (genotype) over the time course of the experiment. This enables the SCS data set to account for the biological variability between the time points in each strain series as shown in Figure 6.

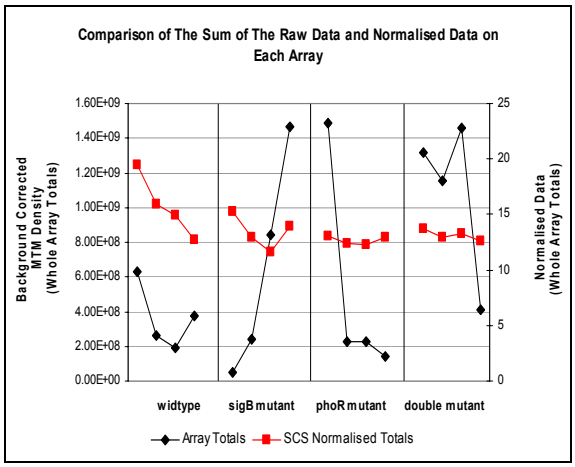

Fig. 6. Totals of gene expression on each array shown for raw data (diamonds) and for SCS

normalised data (squares) using 63 SCS genes described above.

Figure 6 shows that the variability of the normalised data is much lower than that of the raw MTM density data (here MTM stands for Median-based Mean Trimmed Density, an output from ArrayVision™ as a measure of spot intensity). The change in total expression over time is biologically more plausible when the data is normalised using this algorithm.

The three mutants initially have a lower total gene expression than the wildtype strain. The phoR strain and the double mutant show a similar pattern of total expression over time compared to the wildtype and the sigB strain. This can be expected as the organism does not have the mechanism to specifically cope with phosphate stress in the *phoR-null* mutant or the double mutant. The wildtype strain shows decrease in total gene expression over time when the data is normalised. It is expected that under phosphate limited conditions the organism will eventually sporulate and so will down regulate a number of metabolic pathways, hence reducing the overall amount of mRNA in the cells. The sigB strain also shows a decrease in total gene expression, when the data is normalised, up to the last time point, where the total gene expression increases. A biological explanation for this could be related to the hyper-induction of the phoR operon or the onset of sporulation. Upon inspection it transpires that 124 genes are up-regulated by at least 3 fold between the last two time points of the SCS normalised data in the *sigB-null* mutant. Of these genes 32% are either related to sporulation or involved in reaction pathways that result in the release of phosphate. A further 37% of these up-regulated genes currently have an unknown function. The remaining 31% have varying functions but mainly belonging to functional categories 1 and 2 (see SubtiList for functional category classification).

In the three mutant strains, overall gene expression was lower at the outset (as indicated in Figure 6). This is clearly not the case with the non-normalised data, where the total gene expression in the phoR strain and the double mutant is relatively high at the outset (shown by diamonds in Figure 6).

5.2. Differential gene expression

In the absence of technical replicates in the data set studied, it has proved difficult to apply the common methodologies usually employed to identify differentially expressed genes such as the t-statistic or Wilcoxon test. Instead, each gene's Park score has been calculated for every strain-wise comparison for both the nMTM and SCS normalised data sets. There is variation in the Park scores between the two normalisation methods and the results indicate differentially expressed genes are more likely to be correctly identified when the data is normalised with the SCS method rather than the nMTM. To illustrate this, a subset of 33 genes known to be under the control of the Pho regulon is focussed on since the expression of the Pho-regulated genes is expected to be notably lower in the *phoR*-null mutant compared to the wildtype or the *sigB*-null mutant. Therefore the Park scores for the sigB/phoR and the wildtype/phoR comparisons are shown (Figures 7 and 8). These 33 genes (with the exception a small number of genes in the Pho regulon that are repressed by *phoR*) are expected to have a high Park score in the two comparisons, indicating that they are expressed to a greater degree in the wildtype or sigB strain compared to the phoR strain.

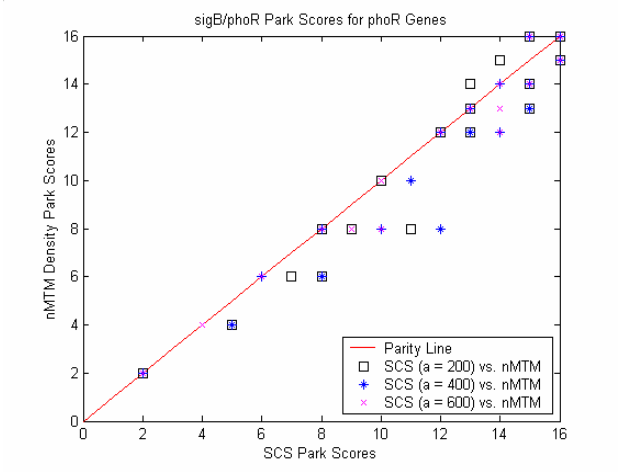

Fig. 7. Parity plot of Park scores of 33 genes in Pho regulon for sigB/phoR comparison using nMTM and SCS normalised gene expression data.

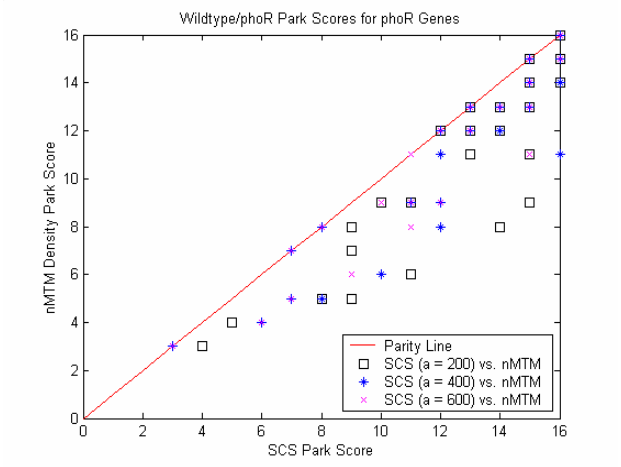

Fig. 8. Parity plot of Park scores of 33 genes in Pho regulon for wildtype/phoR comparison using nMTM and SCS normalised gene expression data.

The parity plots (Figures 7 and 8) show the Park score for the selected genes using the two normalisation techniques. For the SCS normalisation, the Park scores for different values of parameter *a* (200, 400 and 600) are also shown. If the two normalisation techniques (nMTM and SCS) were equal, all the symbols are expected to lie on the parity line. Of the 33 Pho-regulated genes shown in

Figure 7 (the sigB/phoR comparison) 58% have a higher Park score with the SCS normalisation (a = 400), so that the points lie below the parity line. A further 36% have equal Park scores for data using either of the normalisation techniques and only 1 gene has a higher Park score when normalised using the global scaling method. No major improvement is seen when parameter a is set to 200 or 600.

A similar result is seen in Figure 8 (the wildtype/phoR comparison). In this case 42% of the Pho-regulated genes have equal Park scores in both normalisation techniques and the remaining 58% have a higher Park score with the SCS normalisation (a = 400). This indicates that the SCS normalisation allows a clearer discrimination of the genes which are known to be differentially expressed in this experimental system.

6. CONCLUSIONS

A non-parametric normalisation method is proposed for multi-condition time series gene expression data. This method is based on a series of comparisons of ranked gene expression contributions on the individual arrays. If the rank position of a gene contribution to the array total does not change within specified limits across all the arrays, then that gene is included in the self-consistent set (SCS) of genes. The total expression of these genes on each of the arrays is then used to normalise the expression data of the rest of the genes. The algorithm depends upon two user defined parameters, a, the absolute rank difference limit and, to a lesser degree, x, the proportion of genes excluded. Current work concentrates on robustness studies of the SCS normalisation in order to assess the sensitivity of the algorithm to experimental data corrupted by known random and systematic noise. Also the application of this method to other gene expression data containing a number of technical replicates, which exhibits the same structure shown in this manuscript, is being investigated.

We believe that the proposed normalisation method may be useful in other cases of single colour DNA array analysis with a combination of multiple strains, conditions and/or time points. The method provides a way of normalising using all the data simultaneously without having to assign a baseline array or using complex statistics that require replicate data. Using this approach will allow us to apply further data analysis techniques with more confidence in the biological plausibility of the results. Therefore the time, money and effort that have been put into producing this data set in the first place will not be entirely lost due to unavailability of technical and biological replication and therefore some useful knowledge may still be gained from the data.

From the results presented here it appears that the new normalisation technique has successfully decoupled the experimental and biological variations in the array data. Such decoupling would be critical in bioprocess development environment, where rapid learning from a limited number of experiments is required. Deep process understanding, enabled by advanced analytical techniques like DNA array measurements, has the potential to significantly reduce lead times and speed up regulatory approval.

7. ACKNOWLEDGMENTS

This work was supported by the BBSRC. The help of and provision of data by Dr N. Allenby, Prof C. Harwood, Dr Z. Pragai, Prof A. Ward and Dr A. Wipat is gratefully acknowledged.

REFERENCES

Allenby, N.E.E., O'Connor N., Prágai, Z., Ward, A.C., Wipat, A. and Harwood, C.R. (2005). Genome-wide transcriptional analysis of the phosphate starvation stimulon of *Bacillus subtilis*. *J. Bacteriol.*,**187(23)**, 8063-8080

Ballman, K.V., Grill, D.E., Oberg, A.L. and Therneau, T.M. (2004). Faster cyclic loess: normalizing RNA arrays via linear models. *Bioinformatics*, **20(16)**, 2778-2786.

Barash, Y., Elidan G, Kaplan T and Friedman N. (2004). Comparative analysis of algorithms for signal quantitation from oligonucleotide microarrays. *Bioinformatics*, **20(6)**, 839-84

Edwards, D. (2003). Non-linear normalization and background correction in one-channel cDNA microarray studies. *Bioinformatics*, **19(7),** 825-833.

Eymann, C., Homuth G, Scharf C and Hecker M (2002). *Bacillus subtilis* functional genomics: global characterization of the stringent response by proteome and transcriptome analysis. *J. Bacteriol*., **184(9)**, 2500-2520

Kepler, T.B., Crosby, L. and Morgan K.T. (2002). Normalization and analysis of DNA microarray data by self-consistency and local regression. *Genome Biol*, **3(7)**, RESEARCH0037

Kerr, M.K., Martin M., and Churchill G.A. (2000). Analysis of variance for gene expression microarray data. *J Comp Biol*, **7(6)**, 819-837

Kobayashi, K., *et al.*,(2003) Essential *Bacillus subtilis* genes. *PNAS*. **100(8)**, 4678-4683

Park, P.J., Pagano, M. and Bonetti, M. (2001). A nonparametric scoring algorithm for identifying informative genes from microarray data. *Pacific Symp. Biocomputing*, 52-63.

Pragai, Z. and C.R. Harwood (2002). Regulatory interactions between the Pho and σ^B-dependent general stress regulons of *Bacillus subtilis*. *Microbiology*, **148(5)**: 1593-1602

Quackenbush, J., (2002). Microarray data normalization and transformation. *Nature Genetics Supplement*, **32**, 496-501.

Wu, W., Dave N, Tseng GC, Richards T, Xing EP and Kaminski N. (2005). Comparison of normalization methods for Codelink Bioarray data. *BMC Bioinformatics*, **6**, 309