Improved Robustness in Time Series Analysis of Gene Expression Data by Polynomial Model Based Clustering

Michael Hirsch[1,*], Allan Tucker[1], Stephen Swift[1], Nigel Martin[2],
Christine Orengo[3], Paul Kellam[4], and Xiaohui Liu[1]

[1] School of Information Systems Computing and Mathematics, Brunel University,
Uxbridge UB8 3PH, UK

[2] School of Computer Science and Information Systems Birkbeck, University of
London, Malet Street, London, WC1E 7HX, UK

[3] Department of Biochemistry and Molecular Biology, University College London,
Gower Street, London, WC1E 6BT, UK

[4] Department of Infection, University College London, Gower Street, London, WC1E
6BT, UK

Abstract. Microarray experiments produce large data sets that often contain noise and considerable missing data. Typical clustering methods such as hierarchical clustering or partitional algorithms can often be adversely affected by such data. This paper introduces a method to overcome such problems associated with noise and missing data by modelling the time series data with polynomials and using these models to cluster the data. Similarity measures for polynomials are given that comply with commonly used standard measures. The polynomial model based clustering is compared with standard clustering methods under different conditions and applied to a real gene expression data set. It shows significantly better results as noise and missing data are increased.

1 Introduction

Microarray experiments are widely used in medical and life science research [11]. This technology makes it possible to examine the behaviour of thousands of genes simultaneously. Moreover, microarray time series experiments provide an insight into the dynamics of gene activity as an essential part of cell processes.

Despite efforts to produce high quality microarray data, such data is often burdened with a considerable amount of noise. Attempts to reduce the noise are manifold, including intelligent experimental design, multiple repeats of the experiment and noise reduction techniques in the data preprocessing [13]. In addition to the noise problem, parts of the data often can not be retrieved properly so that the dataset contains missing values. For example, a dataset of several experiments with yeast (about 500,000 values) [10] has more than 11% missing values.

[*] This work is in part supported by the BBSRC in UK (BB/C506264/1).

M.R. Berthold, R. Glen, and I. Fischer (Eds.): CompLife 2006, LNBI 4216, pp. 1–10, 2006.

With decreasing quality the direct clustering (DC) of the data with standard methods [5] becomes less reliable. If the data has considerable missing data, the straightforward calculation of the score functions homogeneity and separation [4] for the cluster quality becomes impossible. To overcome these problems this paper suggests the modelling of the data with continuous functions. The model based clustering is done not on the original dataset directly, but on models learnt from it. The models reduce random noise and interpolate missing values, thereby increasing the robustness of clustering.

In this paper the polynomial model based clustering (PMC) is introduced. In contrast to the DC of the data, which calculates the similarity matrix directly from the data, PMC comprised of three steps: the modelling, the calculation of the similarity matrix from the models and the grouping.

2 Methods

The application of continuous functions in time series modelling is motivated by some specific assumptions. Time series result from measurements of a quantity at different time points (TP) over a certain time period. The quantity changes continuously if it could be measured at any time in the presumed time period. Measurement restrictions are due to extrinsic factors such as technical restrictions. Moreover, if a continuous quantity has the value x at TP a and the value y at TP b, then the quantity has any value between x and y at some TP between a and b. Often time series or functions have no sharp edges in the time response, i.e. they are differentiable or smooth.

Any smooth function can be approximated by the Taylor expansion, i.e. by a polynomial. Polynomials are easy to handle since basic operations can be done by simple algebraic manipulations on the parameters. Therefore polynomials are a natural choice in time series modelling. Nevertheless, other classes of functions might be used as well. Previously, polynomials have also been used in other applications of gene expression data modelling [8,12].

2.1 Modelling

Consider series of observations, $y_l(t_i)$ $(l = 1 \ldots N,\ i \in I = \{1, \ldots, T\})$, of N quantities at T TPs. The time elapsed between two measurements at t_i and t_{i+1} might be different through the series. A sub-series of $y_l(t_i)$ in which the missing values are omitted is denoted by $\tilde{y}_l(t_i)$ $i \in J$, where the index set J is the subset of I that contains these time-indices, where a value is available. If J is equal to I, then $\tilde{y}_l(t_i) = y_l(t_i)$.

Polynomials have the general form

$$P(t) = \sum_{i=0}^{n} \alpha_i t^i \ , \tag{1}$$

where n is the degree of the polynomial. To fit a polynomial to the data, the least squares method is used [9]. This method optimises the parameter, α_i, of a function $f(t, \alpha_0, \ldots, \alpha_n)$, $n + 1 < |J|$, $|J|$ is the number of elements in J, such that the function $Q(\alpha_0, \ldots, \alpha_n) = \sum_{i \in J} (f(t_i, \alpha_0, \ldots, \alpha_n) - \tilde{y}_l(t_i))^2$ becomes minimal. Therefore the equations $\partial Q / \partial \alpha_k = 0, \ k = 0 \ldots n$ have to be solved. Applying this equation to polynomials yields

$$\sum_{k=0}^{n} \alpha_k \sum_{j \in J} t_j^{k+i} = \sum_{j \in J} t_j^i \tilde{y}_l(t_j) \quad i = 0, \ldots, n \ . \tag{2}$$

These are $n + 1$ linear equations for the $n + 1$ parameters $\alpha_0, \ldots, \alpha_n$. To solve these equations an inverse matrix of the $(n + 1) \times (n + 1)$ matrix $\sum_{j \in J} t_j^{k+i}$ has to be calculated for each distinct subset J of I that occurs in the data set. To avoid large numbers in the calculation and hence a loss of precision, the time series are scaled to the time interval $[-1, 1]$.

The modelling is done using polynomials with degrees ranging from 2 to 12. Figure 1 shows examples for the degrees 4, 8 and 12. With increasing degree the models fits the data better, but also may over-fit the data.

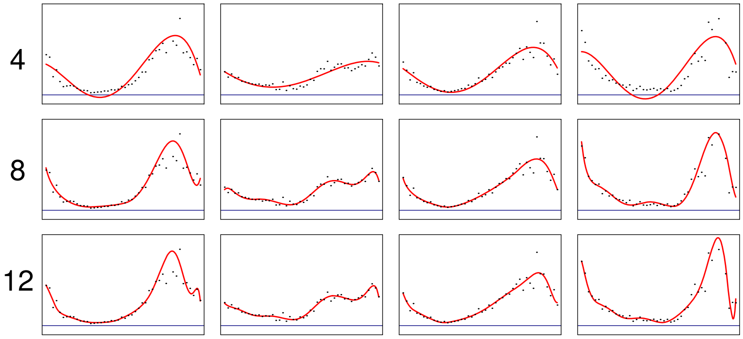

Fig. 1. Modelling of gene expression data with polynomials of different degrees

2.2 Similarity Measures

To calculate the similarity between polynomials, distance measures for functions have to be used. Usually these distance measures involve integration, which replace the sum in the equations for discrete measures. Polynomials are expandable into a Taylor-series, so that a large class of distance measures can be applied. For polynomials it is possible to calculate the anti-derivative, so that numerical integration can be avoided. Each polynomial is represented by the vector of its parameters, $(\alpha_0, \alpha_1, \ldots, \alpha_n)$. Therefore the sum of two polynomials, represented by (α_i) and (β_i), can be written as $(\alpha_0 + \beta_0, \alpha_1 + \beta_1, \ldots, \alpha_n + \beta_n)$ and the anti-derivative of (α_i) is represented by $(0, \alpha_0, 1/2\alpha_1, \ldots, 1/(n+1)\alpha_n)$. The representations for the products and derivatives of polynomials can be found analogously. Therefore, the calculation of the integrals can be reduced to some simple algebraic operations on the $n + 1$ parameters α_i, which keeps the computational complexity for the distance measures low. The calculation of the derivative, the anti-derivative, the sum and the function value of polynomials takes $O(n)$ operations, the calculation of the product of two polynomials takes $O(n^2)$ operations. For the DC the calculation effort depends on the number of TPs T. Because the number of parameters has to be considerably smaller than the number of TPs (otherwise the models would be over-fit), the calculation of the similarity matrix takes less operations for the PMC than for the DC. Two distance measures are considered, the L_p distance and the distance based on a continuous Pearson correlation coefficient.

$\mathbf{L}_p$ Distance. The L_p distance is a standard distance in the space of continuous functions and is equivalent to the p-distance (Minkowski distance) for finite dimensional spaces, such as Euclidean distance,

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum (x_i - y_i)^2} \ , \tag{3}$$

or the Manhattan distance. Let $\boldsymbol{x} = x(t)$ and $\boldsymbol{y} = y(t)$ be continuous functions over the closed interval [a,b], then the L_p distance is given by

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt[p]{\int_a^b |x(t) - y(t)|^p \mathrm{d}t} \ . \tag{4}$$

Usual choices for the exponent are $p = 2$, which is analogous to the Euclidean distance or $p = 1$, which is analogous to the Manhattan distance.

Continuous Correlation Coefficient. Using the mean value theorem of calculus it is possible to formulate the Pearson correlation coefficient of series $\boldsymbol{x} = \{x_i\}$ and $\boldsymbol{y} = \{y_i\}$,

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum x_i y_i - \frac{1}{N} \sum x_i \sum y_i}{\sqrt{\left(\sum x_i^2 - \frac{1}{N}(\sum x_i)^2\right)\left(\sum y_i^2 - \frac{1}{N}(\sum y_i)^2\right)}} \ , \tag{5}$$

for integrable functions. Let $\boldsymbol{x} = x(t)$ and $\boldsymbol{y} = y(t)$ be continuous functions over the closed interval $[a, b]$ and $L = b - a$, then the correlation r can be calculated by

$$r(\boldsymbol{x}, \boldsymbol{y}) = \frac{\int_a^b xy \mathrm{d}t - \frac{1}{L} \int_a^b x \mathrm{d}t \int_a^b y \mathrm{d}t}{\sqrt{\left(\int_a^b x^2 \mathrm{d}t - \frac{1}{L}(\int_a^b x \mathrm{d}t)^2\right)\left(\int_a^b y^2 \mathrm{d}t - \frac{1}{L}(\int_a^b y \mathrm{d}t)^2\right)}} \ . \tag{6}$$

2.3 Grouping

The similarity matrices can be used with any standard clustering technique. To compare the method presented in this paper the Partitioning Around Medoids (PAM) [6] and two variations of hierarchical clustering algorithms were used, the average-linkage cluster analysis and the complete-linkage algorithm [3]. These methods are well-established and have been used for clustering microarray data with some success.

3 Data Set

The PMC is tested with a subset of the gene expression data of the malaria intraerythrocytic developmental cycle [2]. This subset was chosen, because a functional interpretation of the genes is known and can be used to assess the clusterings. It comprises 530 genes in 14 functional groups. The gene expression is measured in 48 TPs with 1 hour time differences. The data set contained 0.32% of missing data and had a low noise level, which has been verified through [2] and by visually plotting many of the functional groups.

4 Experiments

In every experiment the clustering is done with PAM, the average-linkage method and the complete-linkage method. For DC the methods were always applied to both the Euclidean and the correlation based similarity matrix. Polynomials of degrees from 2 to 12 were fitted to each variation of the data set and both the L_2 distance (4) and the correlation (6) were used for clustering. The following experiments were conducted.

1. The data set was clustered without any variations.
2. Normal distributed noise was added to the data. The standard deviation varied between 2% and 66% of the overall mean of the original gene expression values. The experiment was repeated 25 times.
3. The data set was changed by randomly deleting values. The number of missing values varied between 2% and 50%. The experiment was repeated 25 times.

To validate the clustering results, the weighted κ (WK) method [1,7] and quotient of homogeneity and separation (H/S) [4] were used. The WK is a similarity metric between clusters, with possible values between -1 and 1. The larger the WK value the better the agreement between the cluster results. For a clustering $\mathcal{C} = \{C_1, \ldots, C_K\}$ and a distance measure d H/S is given by

$$H(\mathcal{C}) = \sum_{k=1}^{K} H(C_k) = \sum_{k=1}^{K} \sum_{\boldsymbol{x} \in C_k} d(\boldsymbol{x}, \boldsymbol{r}_k)^2 \tag{7}$$

and

$$S(\mathcal{C}) = \sum_{1 \le l < k \le K} d(\boldsymbol{r}_j, \boldsymbol{r}_k)^2 \ , \tag{8}$$

where $\boldsymbol{r}_k = 1/n_k \sum_{\boldsymbol{x} \in C_k} \boldsymbol{x}$ are the cluster centres. A good clustering should have a low homogeneity value and a high separation value, hence a low H/S quotient. Because it is a quotient of sums of distances, H/S is always non-negative.

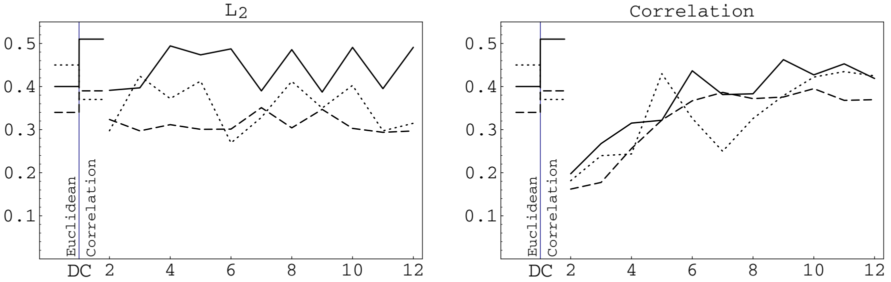

Fig. 2. Weighted κ results of DC and PMC for different similarity measures. PMC is shown depending on the degree of the model. Methods: Average Linkage – normal line, Complete Linkage – dotted line, PAM – dashed line.

5 Results

The results of the clustering of the unchanged data set for WK and H/S are shown in the Figures 2 and 3. The results of DC are marked at the left hand side of each figure, the results for PMC are shown in the main parts, one Figure for each similarity measure. The best DC result yielded the average-linkage clustering of the correlation based similarity matrix, $WK = 0.51$ and $H/S = 0.18$. The PMC showed the best results when applying the average-linkage clustering to L_2 similarity matrix. In particular, the models with even-numbered degrees from 4 to 10 showed constantly good values for WK (0.49) and H/S (0.16–0.19). The correlation based distance showed the best results for models with high degree, but the values for H/S were varied.

The best results in the noise experiment yielded the DC with the correlation based distance and average-linkage clustering and the PMC with L_2 distance and average-linkage clustering. These results are shown in Figure 4. The model of degree 4 is printed with a dotted line and the models of degree 2 and 3 are printed with dashed lines. The other lines show the models of degrees from 5 to 12.

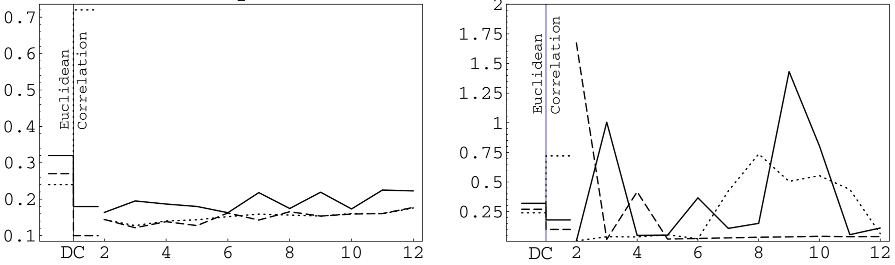

Fig. 3. Homogeneity/Separation results of DC and PMC for different similarity measures. PMC is shown depending on the polynomial degree. Methods: Average Linkage – normal line, Complete Linkage – dotted line, PAM – dashed line.

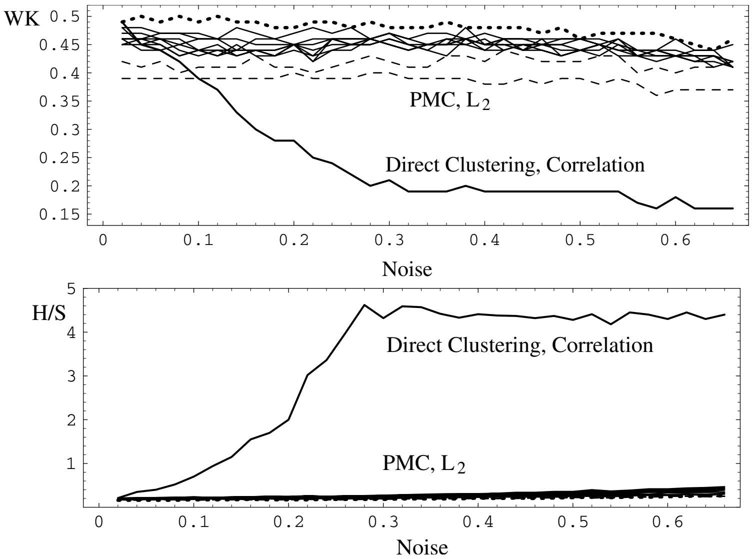

Fig. 4. Weighted κ and Homogeneity/Separation depending on the noise level (in parts of one). Cluster method: Average-linkage. Dotted line – model of degree 4, dashed lines – models of degree 2 and 3.

The PMC displayed excellent robustness, even for a high level of random noise, whereas the DC gave clearly poorer results as the noise level increased. The best choice of the model degree is 4. Models of degree 2 and 3 yielded poorer results for WK than other models, but showed consistently good H/S values. The H/S value for the model of degree 12 changed from 0.21 to 0.47, whereas the H/S value for the degree 2 model only increased from 0.17 to 0.25.

The unchanged data set has 0.32% missing values (not available values – NA). No single time series has more than 3 NAs. The best WK-results in the missing value experiment is yielded by the average linkage method with the correlation based distance for the DC and the L_2 distance for the PMC, see Figure 5 top. Good results were observed using the model with degree 4 (WK 0.49–0.4; dotted line), which worked with up to 46% NA. The DC result is printed with a bold line and aborts at 44% NA. It runs slightly below the degree 4 model (WK 0.48–0.37). The results for the polynomials of degree 2 (WK 0.39–0.38 up to 46% NA) and 3 are very stable as well (dashed lines), but showed poorer WK values. The polynomial models with higher degrees, printed with thin lines, failed sooner.

For the DC methods, the H/S could not be calculated for more than 2% of NAs in the data set. The reason is that every operation that has a NA as any operant returns a NA. For vector operations only these components are used that are not NA for every operant. As it can be seen from the definitions of the cluster centres and separation, eventually every value of a TP is used as an operant. If only one of these values is NA, than this TP is not used in the

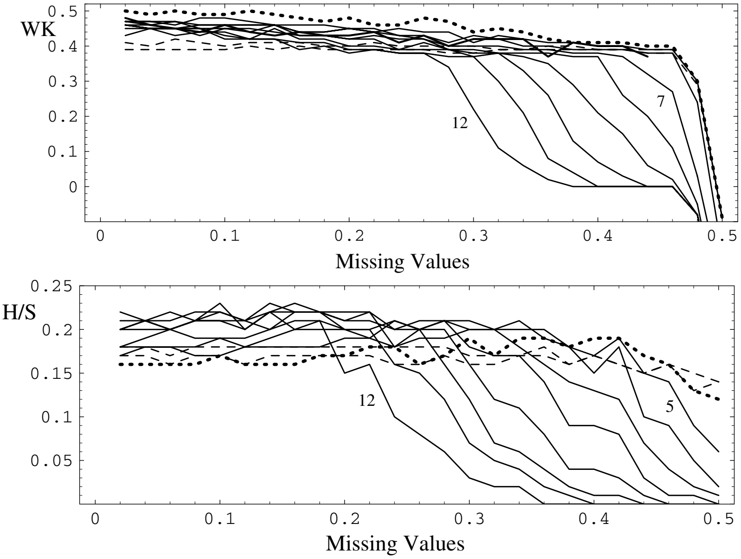

Fig. 5. Weighted κ and Homogeneity/Separation of DC and PMC (degrees 2–12) depending on the rate of missing values. Cluster method: Average Linkage. Bold line – DC, dotted line – model of degree 4, dashed lines – models of degree 2 and 3, models of degrees 5-12 – thin lines. The H/S-value could not be calculated for the DC.

calculation of the separation. In other words, if every TP has at least one NA value the separation will be NA. If p is the ratio of missing values, T the number of TPs and N the number of genes, then the probability that

- a TP contains no NA is $(1-p)^N$,
- a TP contains at least one NA is $1-(1-p)^N$,
- every TP contains at least one NA is $(1-(1-p)^N)^T$,
- at least one TP contains no NA is $1-(1-(1-p)^N)^T$.

The last expression is the probability that the separation is not NA. The likelihood to calculate a non-NA H/S value for the considered data set is 0.9999 for 0.32% NAs, 0.2 for 1% NAs and 0.001 for 2% NAs.

On the other hand, the model based clustering allows H/S to be calculated for any number of NAs in the data, see Figure 5 bottom. Moreover, the H/S value stays stable for low degree models, between 0.16 and 0.19 for the degree 4 model, and for up to 46% missing values in the data set (method: L_2, average linkage).

Summary. The DC and the PMC showed similar results for the good quality data set. Higher degree models showed better results than models of lower degree. With increasing noise the quality of DC dropped significantly, whereas PMC showed excellent robustness. The stability of DC and PMC against missing values is about the same whilst low degree models showed slightly better results.

Homogeneity and separation could not be calculated for DC. The PMC using a model of degree 4 showed the best overall result.

6 Conclusions

This paper has introduced a novel clustering method, polynomial clustering (PMC), for dealing with missing and noisy data. It has demonstrated that PMC yielded consistently good results for different levels of noise and missing values, whereas traditional clustering methods only generated good clusters when applied to high quality data. PMC showed the best robustness when using models with lower degrees, but these were too simple to reflect the features of the data set and hence yielded poorer clustering results. The models with higher degree showed better results but poorer robustness. PMC is easy to calculate. It can be used with a wide range of similarity measures and any standard clustering method.

The optimal degree of the polynomial was calculated empirically in order to balance complexity of the data with over-fitting. Future work will involve looking into ways to determine the degree automatically. The tests will be extended to other real world data sets, a wider range of distance measures and clustering techniques. It is also intended to refine the approximation technique and to analyse the modelling with other classes of functions.

References

1. Altman, D. G.: Practical Statistics for Medical Research. Chapman and Hall (1997)
2. Bozdech, Z., Llinás, M., Pulliam, B. L., Wong, E. D., Zhu, J., DeRisi, J. L.: The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. PLoS Biology **1** (2003) 85–100
3. Eisen, M. B., Spellman, P. T., Brown, P. O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA **95** (1998), 14863–14868.
4. Hand, D. J., Mannila, H., Smyth, P.: Principles of Data Mining. MIT Press (2001)
5. Jain, A. K., Murty, M. N., Flynn, P. J.: Data Clustering: A Review. ACM Computer Surveys **32** No. 3 (1999) 264–323
6. Kaufman, L., Rousseeuw, P. J.: Clustering by means of Medoids, Y. Dodge (ed.), Statistical Data Analysis based on the L1-Norm. North-Holland, Amsterdam (1987) 405–416
7. Kellam, P., Liu, X., Martin, N., Orengo, C., Swift, S., Tucker, A.: Comparing, Contrasting and Combining Clusters in Viral Gene Expression Data. Proceedings of the IDAMAP2001 workshop, London (2001) 56–62
8. Lichtenberg, G., Faisal, S., Werner, H.: Ein Ansatz zur dynamischen Modellierung der Genexpression mit Shegalkin-Polynomen (An Approach to Dynamic Modelling of Gene Expression by Zhegalkin Polynomials). at – Automatisierungstechnik **53** 12 (2005) 589–596

9. Ralston, A.: A First Course in Numerical Analysis. McGraw-Hill (1965)
10. Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D., Futcher, B.: Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization. Molecular Biology of the Cell **9**, 3273–3297, URL http://cellcycle-www.stanford.edu
11. Stekel, D.: Microarray Bioinformatics. Cambridge University Press (2003)
12. Vinciotti, V., Liu, X., Turk, R., de Meijer, E. J., 't Hoen, P. A. C.: Exploiting the full power of temporal gene expression profiling through a new statistical test: Application to the analysis of muscular dystrophy data. BMC Bioinformatics **7** 183 (2006)
13. Wit, E., McClure, J.: Statistics for Microarrays. John Wiley (2004)