Signal Selection in Microarray Data Analysis

Javier Nuñez-Garcia and Olaf Wolkenhauer *
Control Systems Centre
UMIST
Manchester M60 1QD, UK
e-mail: javier@csc.umist.ac.uk, o.wolkenhauer@umist.ac.uk
http://www.umist.ac.uk/csc/people/wolkenhauer.htm

Abstract

The paper describes the use of statistical signal selection in DNA microarray data analysis. The experiments considered produce time-series data. A commonly used method of thresholding is compared to using the 'run test', 'turning point test','Kendall's tau test' and 'autocorrelation function'. We conclude that the use of such techniques, developed for time-series analysis, can help to distinguish between non-informative signals or noise and signals which show either a trend or are correlated over time.

1 DNA Microarray Data

For the data considered in this paper we assume that gene expression profiles were obtained from time course DNA microarray experiments. To illustrate the idea and without loss of generality, we use the yeast data set described by Eisen et. al [2] and choose the 18 measurements obtained for the cell division cycle after synchronization by alpha factor arrest. In general, a gene expression data matrix $\mathbf{X}$ has $i = 1, ..., n$ rows for the genes and $j = 1, ..., r$ columns for samples. Here we have $n = 2467$ yeast genes and $r = 18$ time-points. Each row vector $\mathbf{x}_i = [x_{i1}, ..., x_{ij}, ..., x_{ir}]$ represents a particular gene expression profile. Throughout the paper we refer to a gene by its (row) number in $\mathbf{X}$. x_{ij} is the gene expression level at time j of gene i. $x_{ij} \in \mathbf{x}_i$ is the normalized $\log_2$ ratio E_{ij}/R_{ij} where E_{ij} is the expression level or state at time j of the gene i, and R_{ij} is the reference state of the gene, which is a constant value throughout the experiment [1]:

$$x_{ij} = \frac{\log_2\left(\frac{E_{ij}}{R_{ij}}\right)}{\sqrt{\sum_{k=1}^{18}\left(\log_2\left(\frac{E_{ij}}{R_{ij}}\right)\right)^2}} \quad i = 1, \ldots, 2467 \quad j = 1, \ldots, 18 \tag{1}$$

With the normalization of (1), x_{ij} is positive when $E_{ij} \geq R_{ij}$ and we say the gene is induced or "up-regulated". When $E_{ij} \leq R_{ij}$, x_{ij} is negative and it is said that the gene is repressed or "down-regulated".

In this paper we assume that there are no missing values, i.e., measurements have been obtained for all sampling points of the time course or missing values have been replaced by a suitable method. Assuming that not all genes are involved in the dynamic response to an experiment, the problem considered here is one of signal selection. We investigate test for randomness, stationarity and trend to identify informative signals and to disregard genes from the analysis for which signals appear non-informative or as noise.

2 Signal Selection and Preprocessing Problems

The principal aim of microarray experiments is to study gene activity levels through the measurement of mRNA abundance under varying conditions. Knowledge of the functional role of a

*Author to whom correspondence should be addressed.

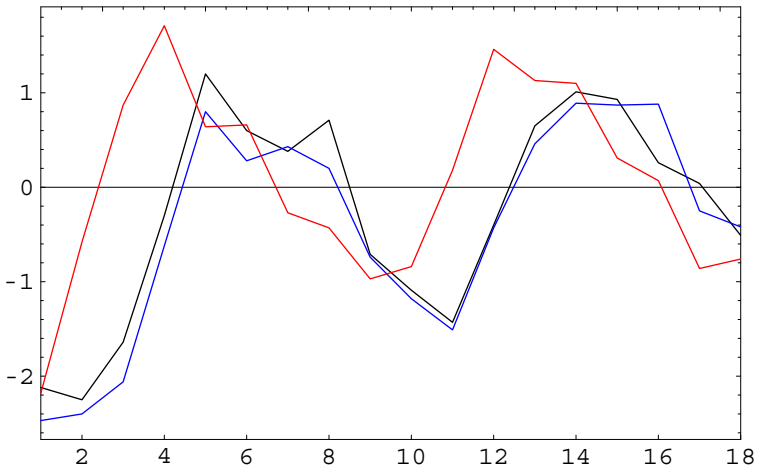

Fig. 1: Three "similar" genes expression profiles (1274, 1281 and 1745).

gene can be used to classify other genes by comparing their expression profiles. Genes involved in the same processes are expected to have "similar" expression profiles in a microarray experiment. Two difficult problems arise from such analysis:

- Conditions of the experiment may not affect all genes. Frequently, a large number of genes are not involved in the process under consideration leading to a "flat" signal with little or no change from the reference state. Those signals can be eliminated from the matrix. This preprocessing can lead to a substantial reduction of the number of genes which makes the analysis of the remaining easier. The criteria for such as selection are not clear as we will discuss later.
- What do we understand by "similar" and how do we decide when two signals are similar? Fig. 1 shows three gene expression profiles. Two of the three profiles are clearly similar to each other. A common approach to discuss similarity is to view the vector $\mathbf{x}_i$ as describing a point in a r-dimensional space where r is the size of the vector. Using the Euclidean distance between the three points, the two signals which are virtually identical would be separated from the third which is time shifted. On the other hand, if we are to view the time course as a dynamic response and fit an autoregressive (AR) model to the data, we find that for all these signals the distance in the parameter space of the AR model is small and all three genes appear related.

In this paper, we will discuss the first problem using statistical test for randomness and trends.

3 Statistical Test for Signal Selection

This section reviews some techniques used to decide whether a signal has been produced by a process with a "deterministic" component or whether it is just "noise". The condition for a sample of data to be regarded as noise is that, if we assume that each data point is a random variable, these have to be independent identically distributed. The distribution that they follow, depends on the process that generates the data. Purely random data such as a lottery, may follow an uniform distribution but very often, processes in life, such as the error of a machine which is taking measurements, follows asymptotically a normal distribution i.e. the random variable which represent the process is a normal distributed when the same experiment is carried out a large number of times. This can be the case for microarray data where noise comes from the apparatus involved in the measurement process. The gene expression ratio x_{ij} in these cases fluctuates around zero if the gene does not responds to the experiment, $E_{ij} \approx R_{ij}$. This generates a "flat" signal on the horizontal axis, $\mathbf{x}_i \approx \mathbf{0}$. This suggest that signals around zero, with little variations may be considered as noise.

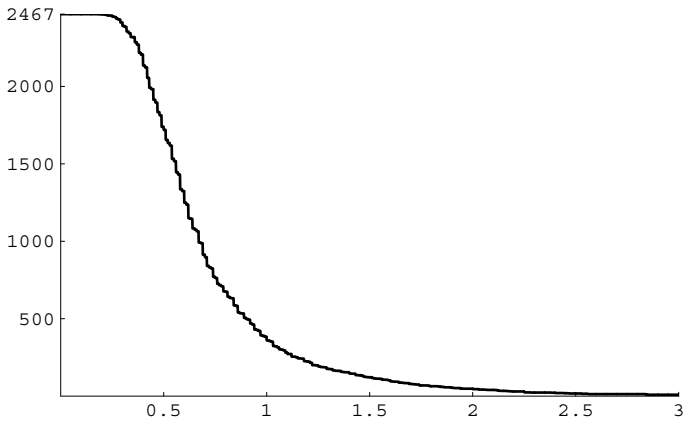

Fig. 2: Length of the radius of the interval around zero against the number of signals considered non-informative.

Threshold Method

The selection criteria used by some authors consist of choosing an interval around zero and all signals falling inside the interval are assumed to be non-significant. Fig. 2 shows the plot of the radius of the interval about zero against the number of significant signals using the threshold method.

This method can also discard signals which, for all j, stay inside the interval but have some trend which could indicate a weak response of the genes to the experiment. In Fig. 3 the signals numbered 256, 260, 268, 330, 452 and 529 illustrate this point. These signal will be rejected using the threshold method but they will be considered to have a "significant" trend if a statistical test is applied. In fact, the responses appear cyclical and suggesting a weak response despite small values. We do not suggest that this would lead to errors in the analysis the specific yeast data set but try to illustrate the problem of using a simple threshold.

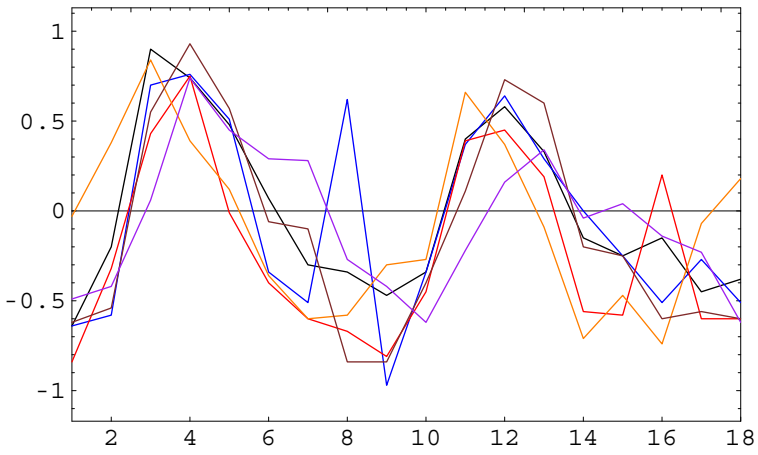

Fig. 3: Genes 256, 260, 268, 330, 452 and 529, all of which would be disregarded using an interval $[-1, 1]$ for signal selection.

Many other signals, for example 13, 25, 31, 33, 49 and 51, plotted in Fig. 4, are inside the interval with radius one and also have no significant trend. For signals verifying these two properties, we make a division between the signals which follow a random pattern and those that do not. To get this division, a statistical test for randomness, such as runs test, discussed in the next section, is applied. In Fig. 5 we can see genes 1 and 129. They do not have a significant trend component and are not considered as a pure random signals by a test for randomness.

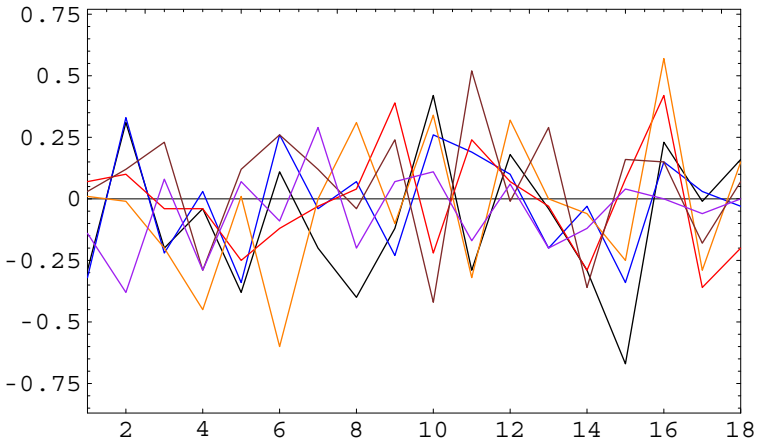

Fig. 4: Genes 13, 25, 31, 33, 49 and 51 fall within the interval $[-1, 1]$ and have not evidence of trend.

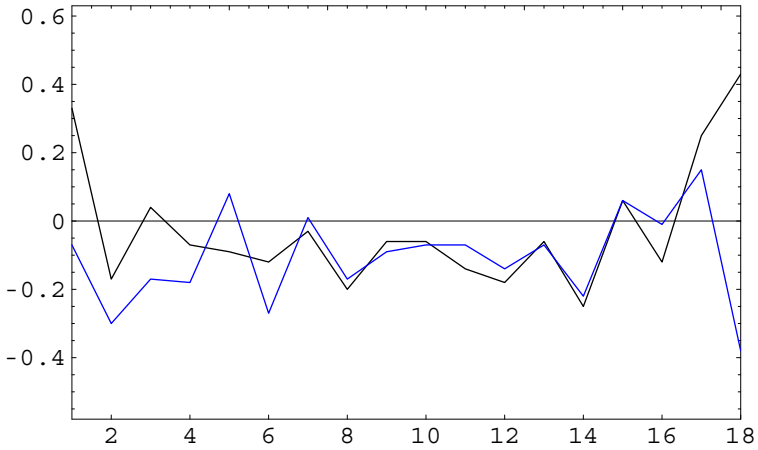

Fig. 5: Genes 1 and 229. They verify the same conditions as the genes in Fig 4 with the difference that they are not considered random signals by the runs test.

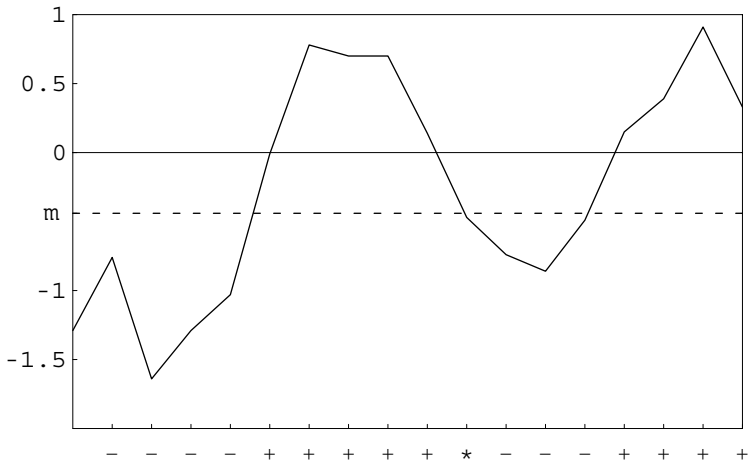

Fig. 6: Runs above "+" and below "-" the median (dashed line).

Runs Test

The runs test is used to check whether the data sample satisfies the following hypotheses:

- H_0: The series has no trend and is independently distributed.
- H_a: The series has a trend and/or has autocorrelated errors.

Under H_0, each data point of the time series is considered as a independently distributed random variable. Note that H_0 is a necessary condition for a sample of data to be regarded as noise. Under H_0 it is supposed that the probability of a point being above the median m of the sample is the same that being below it. Then, if the sample size is even the median under H_0 is $m = r/2$ otherwise, the median is itself an observation and we ignored it, being $m = (r-1)/2$. The statistic used is the number of runs above and below the median, denoted R. If we denote with the symbols "+" and "-" the values above and below the median respectively, R is the number of blocks of pluses and minuses. Fig. 6 shows an example. The median is shown as a dashed line and the "+" and "-" are plotted below the axis. The observation which coincides with the median is denoted by a "*". R is asymptotically normally distributed with

$$\mu_R = m + 1, \quad \sigma_R = \sqrt{\frac{m(m-1)}{2m-1}}$$

Then, normalizing R, $z = \frac{|R-\mu_R|}{\sigma_R}$ we reject H_0 if $|z| > z_{\alpha/2}$. Where α is the confidence level of the test. For $m \leq 20$, R cannot be considered as normally distributed since it is an asymptotic property. However, confidence intervals $[R_L, R_U]$ for small sample sizes can also be calculated. In [3], a table is provided for values of $m \leq 20$ and $\alpha = 0.10$. We have reproduced it in Tab. 1.

m	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
R_L	2	3	3	4	5	6	7	7	8	9	10	11	11	12	13	14
R_U	10	11	13	14	15	16	17	17	19	20	22	23	25	26	27	28

Tab. 1: Critical values for R in a two-tailed runs test for stationarity ($\alpha = 0.10$).

If H_0 is rejected, the test suggests that the series has a trend with $(1-\alpha)\times 100\%$ of confidence or that the values are not independently distributed. The signals in Fig. 7 are considered random by the runs test. Note how this test has to be carefully applied for microarray data, since some signals as the two at the top of the Fig. 7 are not close to zero and may well be informative. In Fig. 8 we can see three trended signals, two inside the unit interval and one outside. The implementation of the runs test is summarized in Fig. 9.

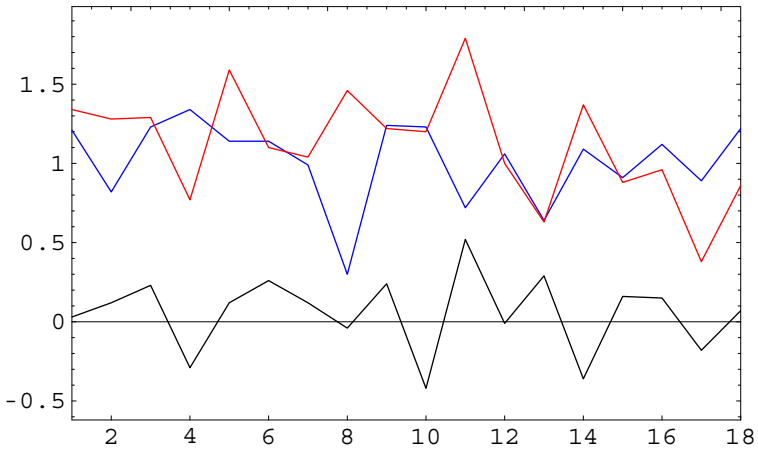

Fig. 7: Signals 49, 561 and 562.

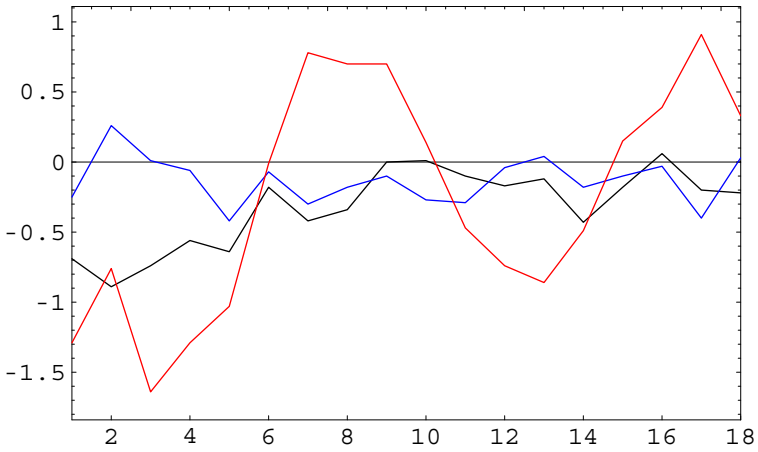

Fig. 8: Signals 4, 8 and 2146.

1. Hypotheses to test for:

   H_0: The series has no trend and is independently distributed.

   H_a: The series has a trend and/or has autocorrelated errors.

2. Statistic: R= the number of runs above and below the median.
3. Decision rule for small sample($m \leq 20, \alpha = 10\%$):

   Reject H_0 if $R \geq R_U$ or $R \leq R_L$. (See Tab. 1 for R_L and R_U.)

   Otherwise do not reject H_0.

4. Decision rule for large sample ($m > 20,\ \alpha$)

   Reject H_0 if $|z| > z_{\alpha/2}$.

   Otherwise do not reject H_0.

   i. $z = \frac{|R-\mu_R|}{\sigma_R}$

   ii. $\mu_R = m + 1, \quad \sigma_R = \sqrt{\frac{m(m-1)}{2m-1}}$ and $m = r/2$ for r even or $m = (r - 1)/2$ for r odd.

   iii. $z_{\alpha/2}$= the upper $(\alpha/2) \times 100\%$ point of the standard normal distribution.

5. Conclusion:

   a) If H_0 is rejected, the test has shown with $(1 - \alpha) \times 100\%$ confidence that a no-trend independent model is inappropriate.

   b) If H_0 is not rejected, the test has shown some support for using a no-trend model.

Fig. 9: Runs test procedure.

Turning Points Test

The Turning points test is used to check whether the time series verify the following hypotheses:

- H_0: The series is a random no-trend series.
- H_a: The series has trend and/or has autocorrelated errors.

A turning point is defined as an inflexion point, i.e., the instant of time at which the series changes from ascending to descending and vice versa. Fig. 10 illustrates the way turning points are identified. Under H_0 the number of turning points U may not be too large and neither too small. The statistic U is asymptotically normally distributed with

$$\mu_U = \frac{2(r-2)}{3}, \quad \sigma_U = \sqrt{\frac{16r-29}{90}}$$

where r is size of the sample. Then, normalizing U, $z = \frac{|U-\mu_U|}{\sigma_U}$, H_0 is rejected with a α confidence level if $|z| > z_{\alpha/2}$. If H_0 is rejected, the test has shown with $(1 - \alpha) \times 100\%$ confidence that the series has trend or the values are not independently distributed. If H_0 is not rejected there is evidence that the time series has no trend. In Fig. 11 the confidence level is plotted against the number of signals rejected by H_0 using the turning points test. The implementation of the turning point test is summarized in Fig. 12.

Kendall's Tau Test

The Kendall's τ-test is based on Kendall's correlation coefficient τ and is used to check whether a time series follows a trend or not. The hypotheses are

- H_0: The series is a random no-trend series.
- H_a: The series has trend either upward or downward.

The procedure is as follows. The $r(r-1)/2$ possible pairs of values from the series are constructed, counting the number of upward pairs N_u (where the first is smaller than the second)

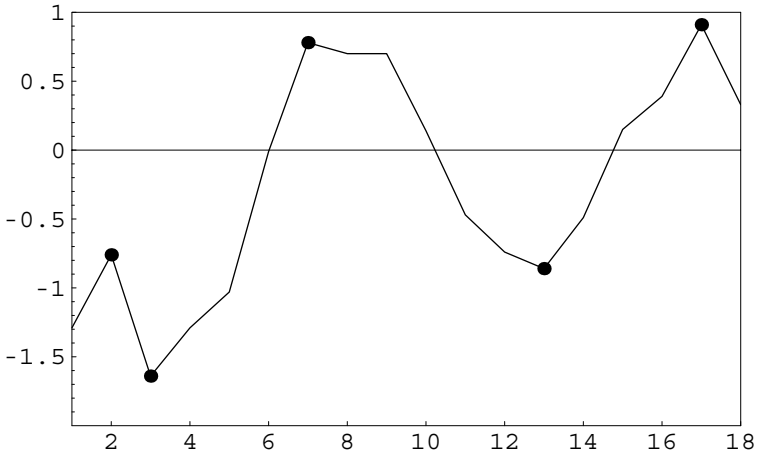

Fig. 10: Example of turning points.

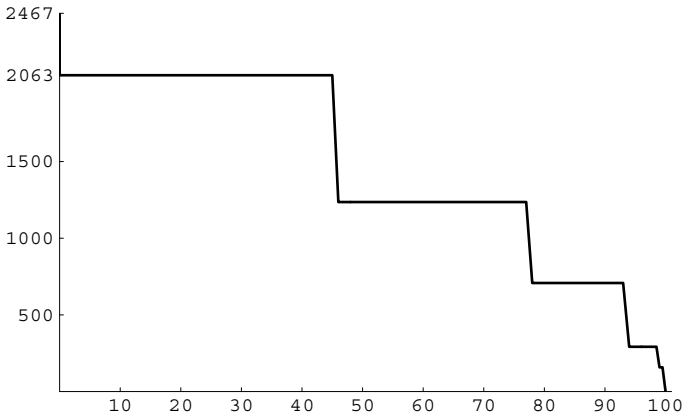

Fig. 11: Confidence of the turning point test against number of signals rejecting H_0.

1. Hypotheses to test for:

   H_0: The series is a random no-trend series.

   H_a: The series is either trended or has autocorrelated errors.

2. Statistic: U= the number of turning points in a series.
3. Decision rule for a moderate or large sample($r \geq 10$):

   Reject H_0 if $|z| > z_{\alpha/2}$.

   Otherwise do not reject H_0.

   i. $z = \frac{|U-\mu_U|}{\sigma_U}$

   ii. $\mu_U = \frac{2(r-2)}{3}$ and $\sigma_U = \sqrt{\frac{16r-29}{90}}$

   iii. $z_{\alpha/2}$= the upper $(\alpha/2) \times 100\%$ point of the standard normal distribution.

4. Conclusion:

   a) If H_0 is rejected, the test has shown with $(1-\alpha) \times 100\%$ confidence that a no-trend independent model is inappropriate.

   b) If H_0 is not rejected, the test has shown some support for a no-trend independent model.

Fig. 12: Turning point test procedure.

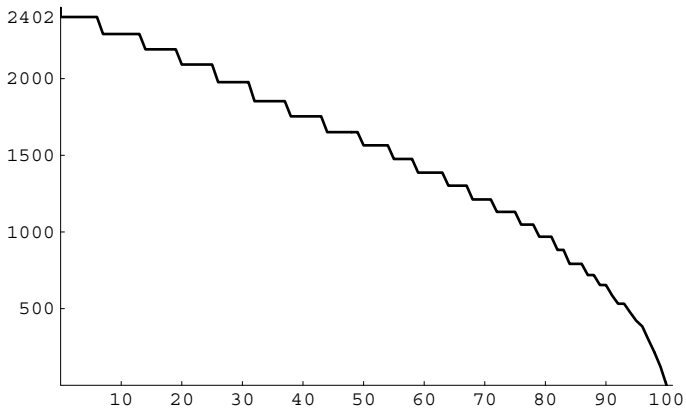

Fig. 13: Confidence of Kendall's tau test against number of signals rejecting H_0.

and the number of downward pairs N_d. The statistic τ is defined as

$$\tau = \frac{N_u - N_d}{r(r-1)/2} \tag{4}$$

or for the normalized version

$$z_\tau = \frac{\tau - \mu_\tau}{\sigma_\tau} \tag{5}$$

where

$$\mu_\tau = 0 \quad \text{and} \quad \sigma_\tau = \sqrt{\frac{2(2n+5)}{9n(n-1)}}. \tag{6}$$

H_0 is rejected if $|z_\tau| > z_{\alpha/2}$ and we conclude with $(1-\alpha) \times 100\%$ confidence that the series is trended. In Fig. 13 the confidence level is plotted against the number of signals rejecting H_0. Note the different graphs in Fig. 11 and 13 despite the fact that the hypotheses tested are similar. The implementation of the Kendall's τ-test is summarized in Fig. 14.

1. Hypotheses to test for:

   H_0: The series is a random no-trend series.

   H_a: The series has trend, either upward or downward.

2. Statistic: $\tau = \frac{N_u - N_d}{r(r-1)/2}$

3. Decision rule for a large sample ($r \geq 10$):

   Reject H_0 if $|z| > z_{\alpha/2}$.

   Otherwise do not reject H_0.

   i. $z = \frac{\tau - \mu_\tau}{\sigma_\tau}$

   ii. $\mu_\tau = 0$ and $\sigma_t au = \sqrt{\frac{2(2r+5)}{9r(r-1)}}$

   iii. $z_{\alpha/2}$= the upper $(\alpha/2) \times 100\%$ point of the standard normal distribution.

4. Conclusion:

   a) If H_0 is rejected, the test has shown with $(1-\alpha) \times 100\%$ confidence that the series is trended.

   b) If H_0 is not rejected, the test has shown some support for a no-trend model.

Fig. 14: Kendall's τ test procedure.

Autocorrelation Function

Another method, commonly used in time series analysis for checking the independence and trend in a time series, is the study of the sample autocorrelation function or ACF of the time series. The autocorrelation of lag k is defined as the correlation between points with a distance of k sampling points, for a $\mathbf{x}_i$, this is

$$r_{ik} = \frac{\sum_{j=1}^{r-k}(x_{ij} - \overline{\mathbf{x}}_i)(x_{i(j+k)} - \overline{\mathbf{x}}_i)}{\sum_{j=1}^{r}(x_{ij} - \overline{\mathbf{x}}_i)^2} \tag{7}$$

where $\overline{\mathbf{x}}_i$ is the mean of x_{ij}, for $j = 1, ..., r$ and $k = 1, ..., m$. m must be smaller than r. Note that for values of m close to r, r_k is constructed from very few data points and the calculated r_k may be not consistent. From (7), it is obvious that $|r_k| \leq 1$. The closer r_k is to zero the more independent are points separated by lag k. Values close to 1 or -1 indicate that there exist a strong relation between values which are separated by k time units. A very simple method to decide whether an autocorrelation is different to zero is based in the asymptotically normally distribution of r_k, with the assumption that r is sufficiently large. The mean is then zero and the variance σ^2 is $1/r$. Since a 95% confidence interval of a normal distribution is approximately $[-2\sigma, 2\sigma]$, we reject the hypothesis $\rho \neq 0$ if r_k is outside of the interval where ρ is the theoretical value of the autocorrelation of lag k.

In Fig. 15 and Fig. 16 we have two examples of the sample autocorrelation functions and their confidence intervals marked by dashed lines. In Fig. 15, there is no significant r_k, and we can conclude that the expression profile has no trend and no dependency for any lag k. In the second figure, r_1 is significant which means a strong dependence between values of distance one. The r_5 is significant too, meaning there exist a dependence between data of distance equal to 5. The reason for this is that the signal is cyclical.

4 Conclusions

In the present paper, we suggest the use of statistical tests for signal selection in DNA microarray time-course experiments. The basic idea is to treat the gene expression profile as a dynamic response and to use methods for time series analysis to identify informative signals.

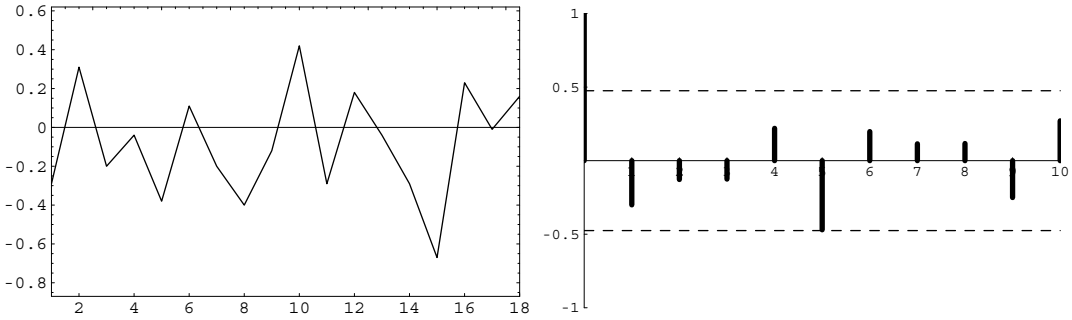

Fig. 15: Gene 13, its ACF and 95% confidence interval.

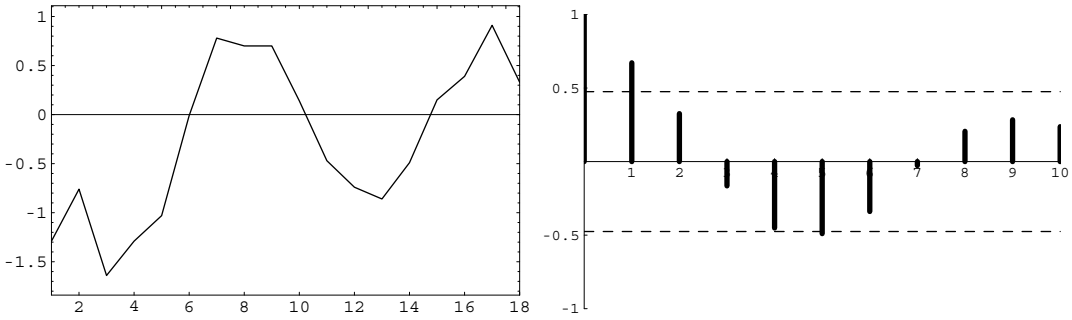

Fig. 16: Gene 2146, its ACF and 95% confidence interval.

Many others statistical tests exist and can be used to check the same or similar hypotheses. Since they are based on the asymptotic distribution of the statistics, these tests work relatively well when the size of the sample is large. For small samples, as is often the case in microarray experiments, different bounds than $z_{\alpha/2}$ for the rejection area may be used. In this case as for any problem were decisions have to be taken from few sample data, the result is less reliable and may require a certain degree of supervision. This may not be feasible for whole-genome data.

An additional difficulty arises when the amount of data is huge as in case of the microarray data where a supervised method for signal selection would increase the effort for pre-processing considerably. On the other hand, we have shown that a simple threshold carries the risk of ignoring informative signals while retaining noisy ones.

References

[1] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S.Furey, M. Ares, and D. Haussler. Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proc. Natl. Acad. Sci. USA (PNAS)*, pages 262–267, 2000.

[2] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA (PNAS)*, pages 14863–14868, 1998.

[3] N.R. Farnum and L.W. Stanton. *Quantitative Forecasting Methods*. PWS-Kent Publishing Company, 1989.