Norsk Regnesentral
NORWEGIAN COMPUTING CENTER

Note

Time development of gene expression

Note no. SAMBA/35/15

Authors Lars Holden

Date 6. nov. 2015

Authors
Lars Holden

Norsk Regnesentral
Norsk Regnesentral (Norwegian Computing Center, NR) is a private, independent, non-profit foundation established in 1952. NR carries out contract research and development projects in information and communication technology and applied statistical-mathematical modelling. The clients include a broad range of industrial, commercial and public service organisations in the national as well as the international market. Our scientific and technical capabilities are further developed in co-operation with The Research Council of Norway and key customers. The results of our projects may take the form of reports, software, prototypes, and short courses. A proof of the confidence and appreciation our clients have in us is given by the fact that most of our new contracts are signed with previous customers.

Title	**Time development of gene expression**
Authors	**Lars Holden**
Date	6. november
Year	2015
Publication number	SAMBA/35/15

Abstract

We test whether there is a slight difference in the time development of the gene expression for a group of genes between two strata relative to time of diagnosis. This group of genes is only identified from all the other genes based on the observed gene expressions. We are able to show that there is a time development in the gene expression the last year before diagnosis. The opposite hypothesis is rejected with a p-value less than 0.001.

It is also possible to analyze in a finer time resolution, but then there is more noise in the results due to the limited number of patients. The time development is most significant up to 6 months before diagnosis. For smaller quantiles (up to 100 genes) there is indication on significance up to 48 months, but the data is so scarce that single patients seem to influence the result.

We estimate that the time development relative to time to diagnosis the first year is present in 600-1.600 genes.

We are also able to make a prognosis on whether a patient has spread or not spread based on the gene expressions the last year before diagnosis. About half the case-control pairs without spread may be classified as a small probability for spread. The classification has a p-value equal 0.01 in a Fisher test.

Keywords	Gene expression, time development
Availability	Open
Project number	220732
Research field	Bioinformatics
Number of pages	13

Table of Content

1 Introduction

This paper is based on a data set for gene expressions for breast cancer. Each measurement is the difference between a case and a control where we for the case know the time before the diagnosis. There are 258 case-control pairs without spread and 101 case-control pairs with spread with 9490 genes. We try to separate between the two strata with and without spread. See xx for further information regarding the data set.

2 Time development in the gene expression

We assume the measured gene expressions after normalization satisfy the equation

$$X_{i,j} = a_i + f_{i,s}(t_j) + \varepsilon_{i,j}$$

for gene i, patient j from stratum s where a_i is a constant, $f_{i,s}(t_j)$ is the time development and $\varepsilon_{i,j} \sim N(0, \sigma_i)$ is noise. The measurement of patient j is performed at time t_j relative to time of diagnosis. We have only one measurement per patient. The expression implies that we expect a smaller variation of $X_{i,j}$ when case-control pairs are from the same strata and with about the time relative to diagnosis. This is measured in the following test: Define $\tau_{s,i,t}$ as the standard deviation for the gene expressions in gene i i, stratum s and an interval around time t relative to diagnosis and $\bar{\tau}_{i,p}$ as the mean of $\tau_{s,i,t}$ for t in a time period p and for the different strata. By estimating $\tau_{s,i,t}$ in a shorter time intervals than the time periods p, we are able to identify time changes also inside each time period. If we don't have $f_{i,s}(t_j) \equiv 0$, then $\bar{\tau}_{i,p}$ and $\tau_{s,i,t}$ are smaller than the corresponding estimates when data is from several time periods. The variance estimate increases when it include the variability in $f_{i,s}(t_j)$. Define $\bar{\tau}_{p,(i)}$ as the i'th smallest of $\bar{\tau}_{i,p}$ and $\bar{\tau}_{(i)}$ is the corresponding value for mean of all the time periods. We test the hypothesis:

H0 $f_{i,s}(t_j) \equiv 0$ for all i and both s

We use $\bar{\tau}_{p,(i)}$ and $\bar{\tau}_{(i)}$ as test statistics. The null model is obtained by randomizing the data between the different case-control pairs. Notice that also a constant value for $f_{i,s}(t_j)$ that is different for each stratum or difference variance in the noise between the two strata, may lead to a rejection of the hypothesis.

The test is formed by randomizing the $X_{i,j}$ between the strata and time periods, i.e. $X_{i,j}$ is replaced with $X_{i,r(j)}$where r(j) is a randomization of the case-control pairs. Hence, we keep the time of each observation, but randomize the data between the case-control pairs. There is not sufficient number of case-control pairs in order to randomize for each strata separately.

The results show that for all time periods, there is no significant change. See Figure 1, left. There are 9.490 genes in the test and 1.000 samples. It is based on data from 258 case-control pairs without spread and 101 case-control pairs with spread. For each gene we have ordered

the genes after increasing $\bar{\tau}_{p,(i)}$ value. From the samples we find a distribution for $\bar{\tau}_{p,(i)}$ that is used in finding the p-value. For the last year before diagnosis, p=1, the variable $\bar{\tau}_{1,(i)}$ is significantly small for (i) less than 8.000. See Figure 1, right. This is based on 53 case-control pairs without spread and 12 case-control pairs with spread. It is not significant results for the other time periods.

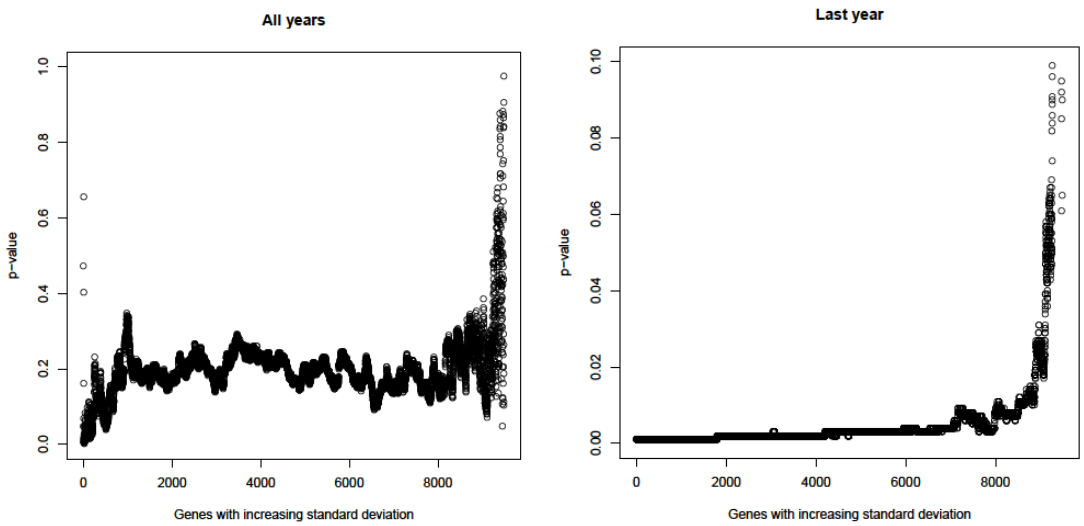

Figure 1. The p-value for the mean sigma value, $\bar{\tau}_{(i)}$ and $\bar{\tau}_{1,(i)}$ respectively. The left figure is for the entire time period and the right figure is for the last year before the diagnosis. Notice the difference in y-axis and that we get very small p-values except for the 1.000 genes with largest standard deviation.

Figure 2 shows the distribution of the sorted $\bar{\tau}_{1,(i)}$ values. The data and the different quantiles for the distribution based on 1.000 randomizations are shown. Notice that the data has smaller values than the minimum of the 1.000 simulations for about 1.000 genes.

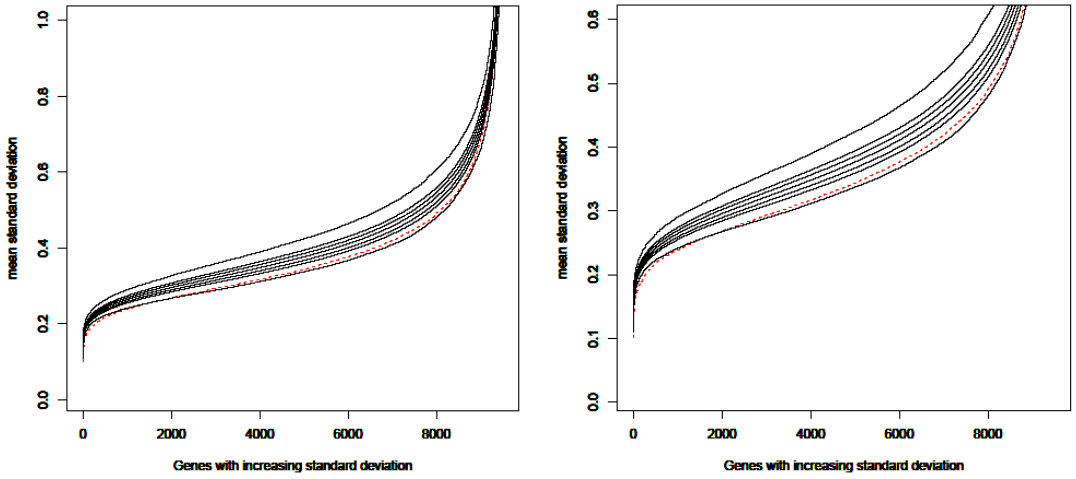

Figure 2. Sorted mean standard deviation $\bar{\tau}_{1,(i)}$ in the last time period (red) and lines for the quantiles min, 0.1, 0.25, mean, 0.75, 0.9 and max based on 1.000 randomizations of the data. Each curve is sorted based on $\bar{\tau}_{1,(i)}$ for this curve which implies that it is not the same gene at each vertical line through the different curves. The only difference between the figures is the scale at the y-axis.

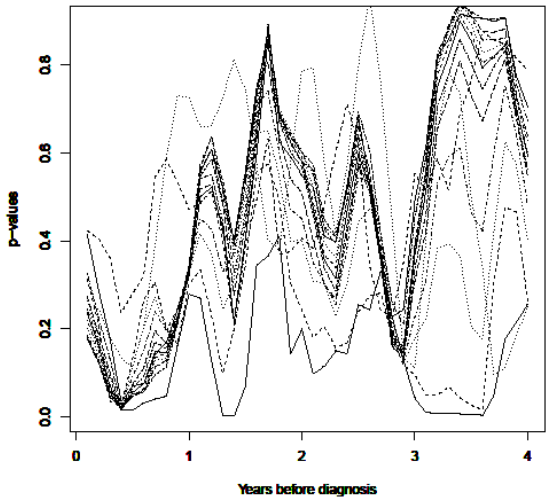

Figure 3. p-values for the 15 quantiles: minimum, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8., 0.9, 0.95, 0.99, maximum shown 40 points in time during the 4 years before diagnosis . Curves are more continuous for small values and more dotted for higher values and higher quantiles gives in most cases larger p-values.

It is also possible to analyze in a finer time resolution, but then there is more noise in the results due to the limited number of patients. See Figure 3. The gene expressions are evaluated at 40 different time points during the four years. For each time point t we use data from the interval (t-50, t+50) days before diagnosis and in most cases we have 10-20 case-control without spread and 4-9 case-control with spread in the interval. For a limited number

of time points it was also necessary to increase the time interval in order to get enough case-controls with spread. We have also smoothed the curves such that each curve shown in Figure 3 is the mean of three original curves. The time development is most significant up to 6 months before diagnosis. For smaller quantiles (up to 100 genes) there is indication on significance in the entire 4 years period, but the data is so scarce that single patients seem to influence the result.

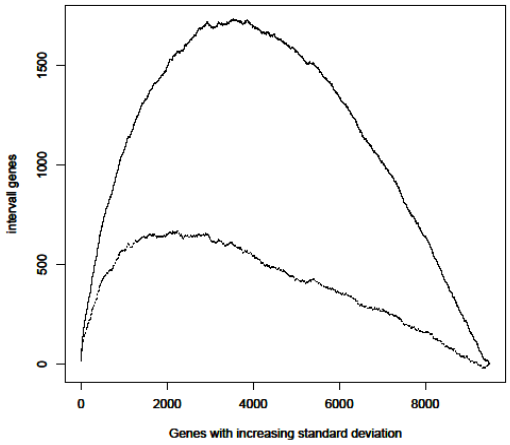

Figure 4. Confidence interval for number of genes with significant time development as a function of quantiles in the distribution of the standard deviation of the gene expressions.

From the results shown in Figure 2, it is also possible to estimate a confidence interval for the additional number of genes with smaller variance than expected, see figure 4. For each value z of the standard deviation $\bar{\tau}_{1,(i)}$ we have $N_{d,z}$ genes with smaller standard deviation than z in the data, and $N_{q,z}$ genes with smaller standard deviation than z for the q-quantile in the randomized data sets. According to the distribution of $\bar{\tau}_{1,(i)}$ we expect with 80% probability that the number of genes with $\bar{\tau}_{1,(i)} < z$ to be in the interval ($N_{0.1,z}$, $N_{0.9,,z}$). Since we observe $N_{q,z}$ genes with $\bar{\tau}_{1,(i)} < z$, we get a confidence interval for the additional number of genes with $\bar{\tau}_{1,(i)} < z$ as ($N_{d,z} - N_{0.9,z}$, $N_{d,,z}$ -$N_{0.1,,z}$). This is a function of z. It is difficult to have an intuition on z, hence we use $N = N_{0.5,z}$, i.e. the mean number of genes $N_{0.5,z}$ satisfies $\bar{\tau}_{1,(i)} < z$. This is shown in Figure 4 where the x-axis is defined based on the x=9.490-$N_{0.5,z}$. The two curves have its maximum for about 7.000 genes, i.e. with about 2.500 genes smaller than this value. The maximum values of the two curves are respectively, 600 and 1.600. This may be interpreted as follows: the standard deviation we would expect for the 2.000-3.000 smallest genes, we have 600-1.600 additional genes with so small standard deviation. The two curves indicated the number of genes with a smaller standard deviation than expected as a function of the size of standard deviation. Notice that we don't have a large number of very small standard deviation, it is only more genes than expected with quite small standard deviation. For other values of z, the additional number of genes with smaller standard deviation is smaller.

3 Prognosis for strata

When there is a time development in the gene expression in the last time period, it is potentially possible to make a prognosis for the strata. We find an estimator that is based on the difference in expected value and variance in the last time period. The variables are defined for all time periods p. For each gene i, we find the relative difference of the expected value in the last time period

$$w_{i,-j,p} = \frac{\mu_{x,i,-j,p,NS} - \mu_{x,i,-j,p,S}}{\sqrt{\sigma^2_{x,i,-j,p,NS} + \sigma^2_{x,i,-j,p,S}}}$$

where $\mu_{x,i,-j,p,s}$ and $\sigma_{x,i,-j,p,s}$ are the expected value and standard deviation in time period p and strata s of the gene expression $X_{i,j}$ where we omit patient j in the estimation. The variable $w_{i,-j,p}$ is made such that the sign depends on whether we expect larger/smaller gene expression for the stratum with spread than without spread and the absolute value of $w_{i,-j,p}$ is large where we expect the absolute value of this difference to be large. When making prognosis for patient j, we use the variable

$$Z_j = \sum_i X_{i,j} w_{i,-j,p}$$

where large values indicate that case control j does not have spread. We don't need a p subscript at the variable Z_j since the case-control j specifies the period p. We predict case-control j to belong stratum without spread if $Z_j > C_p$. We need to set the threshold C.

We set threshold C based on a formula from all the data instead of adjusting it to give an optimal classification for exactly this data set. This variable Z_j is approximately normally distributed in each stratum by the law of large numbers since it is the sum of all the genes. We define $\mu_{z,i,-j,p,s}$ and $\sigma_{z,i,-j,p,s}$ as the expected value and standard deviation in time period p and strata s for Z_j. Hence, we find the probability for belonging to strata NS from the expression

$$P_j = \frac{p_{NS}\varphi(Z_j; \mu_{z,i,-j,p,NS}, \sigma_{z,i,-j,p,NS})}{p_{NS}\varphi(Z_j; \mu_{z,i,-j,p,NS}, \sigma_{z,i,-j,p,NS}) + p_S\varphi(Z_j; \mu_{z,i,-j,p,S}, \sigma_{z,i,-j,p,S})}$$

where p_{NS} is the probability for a patient to belong to strata NS and $\varphi(Z_j; \mu_{z,i,-j,p,NS}, \sigma_{z,i,-j,p,NS})$ is the density of the normal distribution. The threshold $P_j > 0.95$ corresponds to $Z_{j,p} > 80$. This gives the following table of predictions:

	True NS	True S	Sum
Predicted NS	25	1	26
Predicted S	28	11	39
Sum	53	12	65

This shows that we are able to identify about a half the NS case-control pairs as NS with a rather small error probability. 1 out of 26 corresponds to the selected threshold 0.95. This outcome has a 0.01 probability in a Fisher-test.

A standard logit regression gives a p-value equal 0.056 for including the variable Z_j in the regression. A leave-one-out classification based on logit regression does not give as good classification result.