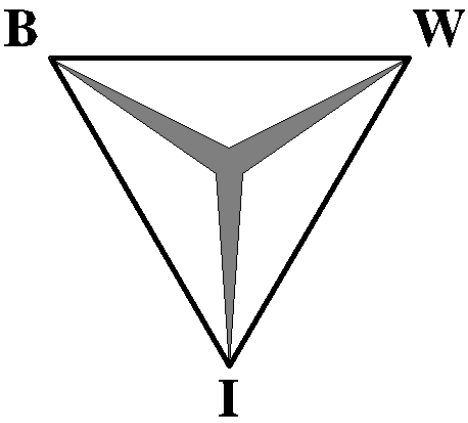
Survival analysis under censoring: Predicting Survival for Cancer patients

Internship paper

August 2008

Combining Genomic (aCGH) and Clinical data



Vu University Amsterdam
Faculty of Sciences
Business Mathematics and Informatics
De Boelelaan 1081a
1081 HV Amsterdam

Name: Christel Nijman
Supervisor: Mark van de Wiel
Second Reader: Mathisca de Gunst

Preface

The last compulsory part of the study Business Mathematics and Informatics is the internship of six months. This internship has to be fulfilled at an external company of the student's choice.

For my internship I carried out an assignment at the Cancer Center Amsterdam (CCA) which is part of the VU medical center (VUmc). The project mainly focused on survival analyses concerning cancer patients. Specifically, combining clinical and genomic data for prediction purposes and comparing standard with alternative prediction models. Also making use of regression analysis, dimension reduction methods and distribution free modeling took up great parts of the assignment

I would like to thank my supervisor Mark van de Wiel for his guidance during this utmost educational period in my study. "Thanks Mark for this great opportunity and your patience while answering my questions and providing guidance when and where needed"

Also thanks to the coordinator of the micro array facility of the VUmc CCA, Bauke Ylstra, for the opportunity to finish my internship at this facility.

Furthermore I want to thank Mathisca de Gunst, my second reader, for her helpful insights and comments.

Last but not least I would like to thank my colleagues, during the past six months, for the nice coffee-breaks, lunchtimes and other great fellowships.

To the reader:

Thanks for taking the time to scroll through and read my work. I hope you will take away some helpful insights into the subject that is discussed.

Amsterdam, August 2008

Christel Nijman

Contents

Preface 5
Contents 7
Introduction 9
Problem definition 9
1. Theoretical part: Survival Analysis 11
1.1. Parametric modeling 11
1.1.1. Hazard functions 11
1.1.2. Exponential Modeling 11
1.1.3. Weibull Modeling 12
1.1.4. Proportional hazards[8] 13
1.1.5. Parameter estimation (Likelihoods)[8] 14
1.2. Nonparametric modeling 15
1.2.1. Cox proportional hazards model 15
1.2.2. Kaplan Meier Curves (Product-Limit Estimator) 16
1.3. Penalizing high dimensional data 17
1.3.1. The Lasso method (L1-penalty) 17
2. Practical part: Data Analysis 19
2.1. Clinical Data 19
2.2. Genomic Data 20
2.2.1. Array Comparative Genomic Hybridization (aCGH) 20
2.2.2. CGHcall[7] 23
2.3. Analysis in R 23
2.3.1. Coxpath[11] 23
2.3.2. Double 10-fold Cross Validation 25
2.3.3. Ranking procedure 25
3. Results 29

Cancer dataset 2 33
Model validating data sets 40
4. Sensitivity analysis 41
Future implementations 49
Conclusion 51
Literature 53
Appendix A: Chromosomal information 55
Appendix B: Figures related to the chapter “Sensitivity analysis” 61

Introduction

In cancer research arrayCGH is a common technique for finding chromosomal aberrations. Linking these chromosomal aberrations with survival through for instance early diagnosis of the disease is high on the researcher's list. In survival analysis statistics is essential for developing robust prediction methods and discovering biomarkers which are associated with survival.

At the CCA, research in the field of cancer development and treatment is done with the aim of providing cancer patients the best possible care, (early) diagnosis and treatment, now and in the future. There is therefore a need for statistical models that can contribute to this highly relevant aim. Such models link relevant genomics data with clinical data and predict survival. Furthermore, the statistics provide a measure for reliability of the prediction.

The goal is to give new insight from a statistical perspective on the relationship between the genomics and clinical data related to cancer. Moreover, we expect that some non-standard procedures, such as stepwise approaches, may have higher prediction accuracy than available standard methods.

Problem definition

- Provide a **statistical model** that **combines clinical and genomic data** and predicts survival
- **IF genomic data adds predictive power**, find **limited selection of genes** that can be used for prospective measurements
- Interpret **alternative model** with **intermediate marker** and compare with standard model

1. Theoretical part: Survival Analysis

In survival analysis we deal with a very important concept called censoring. Censoring occurs when we are unable to observe the response variable of interest. It is a common phenomenon in medical research that some patients can still be in remission when the observation period for remission is terminated. For these patients the exact survival time is unknown because we only know that their true remission time was longer than the observation period. This is called right censoring. In the case of censoring the data is incomplete and special statistical methods have to be applied for analysis[8]. We can either use parametric or nonparametric models to analyze this data. In the usual survival data setup the available data has the following form (y_1, x^1, δ_1)……. (y_N, x^N, δ_N)[9].
The survival time y_i is complete if $\delta_i = 1$ and right censored if $\delta_i = 0$ and x is the usual vector of predictors for the i^{th} individual[9].

1.1. Parametric modeling

In parametric modeling we deal with survival functions and assume that the survival data follows a certain underlying probability density function. Assuming that a random variable Y follows a probability density function f and cumulative distribution function F, then

$$F(y) = P(Y < y) = \int_0^y f(u)du$$

And the survival function or reliability function

$$S(y) = 1 - F(y) = P(Y < y) = \int_y^\infty f(u)du$$

1.1.1. Hazard functions

The hazard function gives the rate of death of an individual given that the individual has survived up to a certain time point. So if we say that individual Y lives up to time y where y>0 and then dies, the hazard function of that individual Y is given by:

$$h(y) = \frac{f(y)}{S(y)}$$

These hazard functions track how the failure rate changes with time[8].

1.1.2. Exponential Modeling

The simplest parametric model is the exponential probability model with probability density function *f*

$$f(y) = \lambda e^{-\lambda y}$$

and cumulative distribution function *F*

$$F(y) = 1 - e^{-\lambda y}$$

The hazard rate belonging to this model is:

$$h(y) = \frac{f(y)}{1 - F(y)} = \frac{\lambda e^{-\lambda y}}{1 - (1 - e^{-\lambda y})} = \lambda$$

The exponential model is the only model with a constant hazard rate. This is also called the memoryless property of the exponential model. This means that if the lifetime of the underlying follows an exponential distribution its future performance is independent of its past performance.

1.1.3. Weibull Modeling

A more extensive parametric model compared to the exponential model is the weibull model. The model has the following probability density function *f*

$$f(y) = \frac{\alpha}{\beta^{\alpha}} y^{\alpha-1} e^{-\left(\frac{y}{\beta}\right)^{\alpha}} \quad , y>0$$

and cumulative distribution function *F*

$$F(y) = 1 - e^{-\left(\frac{y}{\beta}\right)^{\alpha}}$$

The hazard rate belonging to this model is:

$$h(y) = \frac{f(y)}{1 - F(y)} = \frac{\frac{\alpha}{\beta^{\alpha}} y^{\alpha-1} e^{-\left(\frac{y}{\beta}\right)^{\alpha}}}{1 - \left(1 - e^{-\left(\frac{y}{\beta}\right)^{\alpha}}\right)} = \frac{\alpha}{\beta^{\alpha}} y^{\alpha-1}$$

This is a power hazard rate and is used when risk of failure is rapidly increasing with time[8]

1.1.4. Proportional hazards[8]

In survival modeling we are interested in the relationship between the survival time Y and the values of explanatory variables X. So we want to know what the explanatory variables say about the survival time. Now, if the relationship between X and Y is modeled parametrically and Y depends on a vector of the observed values of explanatory variables x, then the hazard rate will also depend on x. If then the hazard rate is given by $h_x(y)$, it is assumed that the covariates, x, have a multiplicative effect on a basic hazard function called the baseline hazard which leads to proportional hazards models.

So the relationship between the hazard function and the baseline hazard is given as follows:

$$h_x(y) = h_0(y)g_1(x)$$

where *g* is a positive function of x and $h_0(y)$ is the baseline hazard, representing the hazard function for an individual having $g_1(x) = 1$.

If two individuals have lifetimes that depend on vectors of covariate values x_1 and x_2 respectively, then the following holds:

$$\frac{h_{x_1}(y)}{h_{x_2}(y)} = \frac{h_0(y)g_1(x_1)}{h_0(y)g_1(x_2)} = \frac{g_1(x_1)}{g_1(x_2)}$$

This shows that the hazard ratio does not depend on *y* because the baseline hazards, which depend on y, cancel from the ratio.

Now we want $g_1(x)$ bigger than or equal to 0 and $g_1(0)$ equal to 1. To accomplish this, a vector β is introduced of *p* parameters by setting $x^T = (x_1, x_2, ..., x_p)$, $\beta^T = (\beta_1, \beta_2, ..., \beta_p)$ and

$$g_1(x) = e^{\beta^T x}$$

The proportional hazards model becomes

$$h_x(y) = h_0(y)e^{\beta^T x}$$

where $\beta^T x = \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p$ is a linear combination of the covariate values from the vector x and the coefficients are taken from the vector of parameters β. Fitting a proportional hazards model will lead to estimation of the p parameters in β using values of the observed responses and covariates

1.1.5. Parameter estimation (Likelihoods)[8]

Parameter estimation is done through the construction of likelihoods and the maximum likelihoods procedure. Let us assume that the parameters in the model are bundled into a variable, θ. Then the likelihood $L(\theta)$ of the observed data is a constant multiple of the joint distribution of the observed data. By maximization of this likelihood function, the estimator $\hat{\theta}$ of θ, is estimated and is the value of θ that maximizes the likelihood function. By searching through the possible values for θ, this maximum is found.

Uncensored data

In the case of uncensored data and observed random variables Y_i for $i = 1.....n$, the likelihood is given as

$$L(\theta) = c\prod_{i=1}^{n} f(y_i;\theta)$$

Solving for which value of θ the derivative with respect to θ of this function equals zero leads to an optimum value. This is called the maximum likelihood procedure. Also checking that the sign of the second derivative is negative gives a maximum. Maximizing the log-likelihood is much easier and preferred. The log-likelihood $l(\theta)$ is defined as $\log_e L(\theta)$ and its maximum is also found through the derivative. The derivative in this case is called the score function, $U(\theta) = l'(\theta)$. The solution to this function is often found through the Newton-Method.

Censored data

For censored data the likelihood function is given as follows:

$$L(\theta) = c\prod_{i=1}^{n} f(z_i)^{\delta_i} S(z_i)^{1-\delta_i}$$

1.2. Nonparametric modeling

In nonparametric modeling there is no assumption on the distribution of the data. So we are dealing with distribution free modeling and there is no parametric form for the baseline. The most important distribution- free regression model used for the analysis of censored data is the Cox model[8]. For visualization purposes the Kaplan Meier curves are widely used. The Cox model is a multivariate analysis model and the Kaplan Meier curve is a univariate model.

1.2.1. Cox proportional hazards model

The Cox proportional hazards model first estimates the values for β and then calculates the baseline through these estimates.

The same proportional hazards model from chapter 1.1.4. is used:

$$h_x(y) = h_0(y)e^{\beta^T x},$$

where $x^T = (x_1, x_2, ..., x_p)$, $\beta^T = (\beta_1, \beta_2, ..., \beta_p)$ and $h_0(y)$ is the baseline hazard that occurs when x=0.
The baseline hazard has no influence on the estimation of β thus at first only the uncensored data points are used.

Analogously to parametric modeling, the likelihood function is constructed and maximized to find the parameter values. A slight difference is that a *conditional* likelihood function is constructed:

$$L_c(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T x_{(j)}}}{\sum_{l \in R_j} e^{\beta^T x_l}}$$

This is done by first specifying a risk set R as the set of individuals that have not died yet. Then the conditional probability, that individual j in the risk set R is the first to die of the members of the risk set is given as follows:

$$= \frac{e^{\beta^T x_{9(j)}}}{\sum_{l \in R_j} e^{\beta^T x_l}}$$

Taking the product of these probabilities leads to the conditional likelihood function.

1.2.2. Kaplan Meier Curves (Product-Limit Estimator)

The Product-Limit estimator estimates the survival function S, using right censored data.

The method acknowledges three variables at a certain time point:

- the number of elements at risk,
- the number of elements that have died,
- the number of elements that are still alive;

By calculating the conditional probability for every individual in the risk set, that the individual will live through a time-span given that it was alive at the beginning of the time-span, the survival function is constructed as the product of these probabilities.
Assume $p_{j,j+1}$ stands for the conditional probability, then the estimate of $p_{j,j+1}$ is equal to

$$p_{j,j+1} = 1 - \frac{\#dying_{j,j+1}}{\#p\_dying_{j,j+1}}$$

Where $\#dying_{j,j+1}$ is the number of individuals that will die in the observed time-span (j,j+1] and $\#p\_dying_{j,j+1}$ is the number of individuals that have the potential of dying within the time-span (j,j+1].

The survival function $\hat{S}(t)$ is the product of these conditional probabilities and is equal to

$$\hat{S}(t) = \prod_{j<t} p_{j,j+1} = \prod_{j<t} \left(\frac{n-j}{n-j+1} \right)^{\delta_{(j)}}$$

where *n* is the number of observations and $\delta_i = \begin{cases} 1 & if\ uncensored \\ 0 & if\ \ censored \end{cases}$

Note that as j increases, the number of elements at risk decreases one at a time.

1.3. Penalizing high dimensional data

With high dimensional data it would be very hard to construct a good prognostic model unless the possibility of dimension reduction was called upon. Conveniently, this is possible and there are penalizing constraints upon which the selection is made. Such dimension reduction models are ridge- and lasso-regression, that ensure L2 and L1 penalties upon the data. The main difference between these two penalties is that the L2-penalty allows all penalized variables to become zero and there is no feature selection possible.

1.3.1. The Lasso method (L1-penalty)

Take the partial likelihood function from the Cox model

$$L_c(\beta) = \prod_{j=1}^{k} \frac{e^{\beta^T x_{(j)}}}{\sum_{l \in R_j} e^{\beta^T x_l}}$$

Then by the lasso method[9] the estimate of β is given by

$$\hat{\beta}(s) = \arg\min l(\beta), \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq s$$

where *s* is a tuning parameter determining how many coefficients and thus the effect of the belonging covariates become zero.

2. Practical part: Data Analysis

2.1. Clinical Data

The raw clinical data consists of observed clinical information concerning 50 patients. Alongside the 50 rows the patients and alongside the columns 58 measured variables are represented.
From these variables the most promising are chosen for the prediction model. This is done through information known beforehand and by input from co-workers. The only condition is that all information concerning this variable is known and present at baseline. The baseline in this case is the date of operation.

For the first prediction model the selected clinical variables are:

1. The age of the patient, (Continuous variable)
2. The sex of the patient, (M or F)
3. The hospital where the patient was treated, (2 values)
4. The differentiation of the tumor, (Poor, Moderate, Well)
5. The state of the tumor, (MSI or MSS)
6. Peritoneal involvement, (Yes or No)
7. The location of the tumor; (2 values)

For “Differentiation” the two values moderate and well were combined because only 3 patients were coded with value “Well”. So now all variables except “Age” are dichotomous.

<Confidential picture not included>

2.2. Genomic Data

The genomic data consists of the calls from the regions with 1120 variables (generated with CGHcall) corresponding to the 50 patients. Every variable can take on 3 values that stand for normal, gain or loss on that part of a chromosome. The 3 values are coded by 0, 1 and -1 respectively. We assume that a gain is the opposite of a loss and this is taken into account when the coefficients are interpreted.

2.2.1. Array Comparative Genomic Hybridization (aCGH)

While reading keep in mind that the goal of aCGH is to find aberrations in the DNA of a patient. These aberrations are of course seen in relation to healthy DNA (read: of a healthy person). One would thus want to compare different DNA samples with each other.

The comparing mechanism is established by hybridization which is the main fundament of DNA microarrays[2]. Two DNA strands can be hybridized if they are complementary to each other reflecting the Watson-Crick rule.[2] A DNA molecule consists of nucleotides and each of these nucleotides consist of a phosphate group, a deoxyribose sugar molecule , and one of the four different nitrogenous bases called: guanine(G), cytosine(C), adenine(A) or thymine(T). By the Watson-Crick rule it is know that G only pairs with C, and A only pairs with T and thus are called complementary.[3] Therefore two strands can only hybridize if they are complementary to each other

$$5' - ATGCCCTGAC - 3'$$
$$3' - TACGGGACTG - 5'$$

Figure 1: Complementary DNA strands[3]

At first only small strings of DNA called oligonucleotides were used to hybridize to complements. Now instead of just using one small string, more of these strings are placed on a surface forming a DNA array.

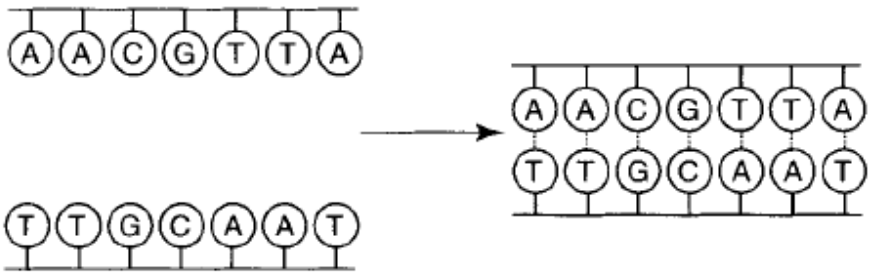


Figure 2: Hybridization of two DNA molecules. Dotted line: hydrogen bonds [2]

A healthy DNA strip has a certain form and by comparing this with an unhealthy DNA sample, aberrations as in copy number changes can be spotted. This is important with cancer because during tumor progression genomic DNA regions are frequently lost or gained.

Comparative genomic hybridization (CGH) was the first efficient approach to scanning the entire genome for variations in DNA copy number[1]. Here a test sample and reference sample are labeled with a fluorescent dye, the reference sample with a red dye and the test sample with a green dye, and then both samples are hybridized to normal human metaphase chromosomes. This then produces images of both fluorescent signals and ratios of these signals are digitally quantified along the length of each chromosome[4]; green means an amplified region with a ratio above one, red means a deleted region with a ratio below one while yellow means a normal region with both signals equally represented. This method is limited due to the poor resolution of the metaphase chromosomes[5]. As follow up to this approach the array CGH method was introduced which produces higher resolution signals because the human metaphase chromosomes are replaced by DNA fragments (100-200 kb) of which the exact chromosomal location is known, identified by its base-pair position on a chromosome. This fragment or array first consisted of spotted genomic sequences inserted into bacterial artificial chromosomes (BACs) but now consists of oligonucleotides (oligo's)[14] .

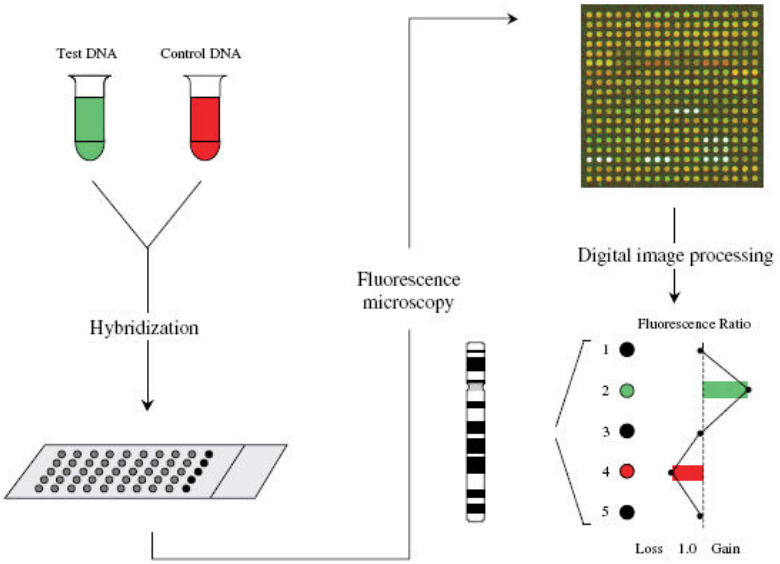


Fig. 1. Schematic overview of the microarray-based comparative genomic hybridization technique. Test and control DNA are labeled with a green and red fluorochrome, respectively. Both DNAs are hybridized to cloned DNA fragments that have been spotted in triplicate on a glass slide (the array). Images of the fluorescent signals are captured and analyzed. Red spots indicate loss of test DNA, green spots indicate gain of test DNA, and yellow spots indicate the presence of equal amounts of test and control DNA. For a precise evaluation, test to control fluorescence signal ratios are measured for each single clone. These results can be translated in a high-resolution overview of chromosomal copy number changes throughout the whole genome.

Figure 3: Schematic overview of aCGH [4]

These hybridization techniques generate profiles in which aberrations can be detected if present. By comparing the generated profiles with normal profiles a clear diagnosis can be established. Due to the fact that the exact chromosomal location of the DNA fragments is known it is possible to link the aberrations to exact parts of chromosomes. The log-ratios are usually converted back to an absolute measure: 3 or 4 underlying discrete states representing loss (<2 copies), normal (2 copies), gain (3-4 copies) and possibly amplification (>4 copies), called calling [7]. Now to translate the profiles into these discrete states certain steps have to be followed. A method that performs this translation is called

CGHcall invented by M.A. van de Wiel[7] which combines strong concepts of previously developed methods.

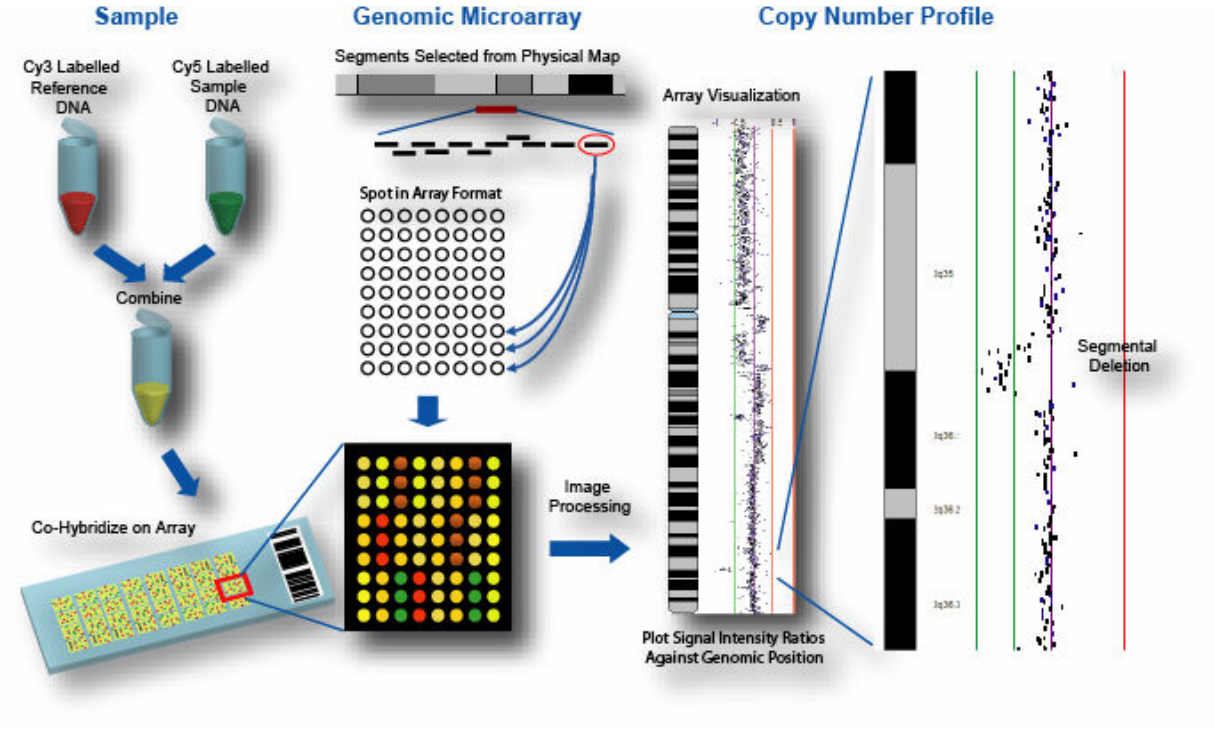


Figure 4: aCGH process (Source unknown)

This method follows 3 main steps: Normalization, Segmentation and Calling and makes a distinction between 6 states instead of 4. After the calling step the dimension of the data is reduced by making Regions. Normalization makes the log2-ratios from different hybridizations comparable (MAD, median normalization). Segmentation divides the genome into contiguous segments and clones that belong to the same segment are used to have the same underlying copy number. Calling is the process of categorizing the different segmentation states into the 6 underlying discrete states: (0) double loss, (1) single loss, (2) normal, (3) single gain, (4) double gain, (≥5) amplification. Regions capture the essential features of the data and are a series of neighboring clones on the chromosome whose aCGH-signature is shared by all clones.[7]

2.2.2. CGHcall[7]

CGHcall is an algorithm that detects copy number changes and classifies them into six states: double deletion, single deletion, normal, gain, double gain and amplification. The algorithm combines strong concepts of other methods that also detect and classify copy number changes. The algorithm is divided into three major parts: Normalization, Segmentation and Calling.

Normalization aims to make $\log_2$-ratios from different hybridizations comparable. Though there are different normalization types, CGHcall uses the median normalization. This normalization type shifts the median of the $\log_2$-ratio to zero.
After the normalization step, the genome is split into contiguous segments. Clones that belong to the same segment are assumed to have the same copy number. The segmentation is needed to reduce the noise in the genomic data and to detect the aberrations and perform breakpoint analysis.
Calling , the last part of the algorithm, is the process of categorizing the different segmentation states as 'loss', 'normal', 'gain', 'amplification'.

2.3. Analysis in R

R is a free software environment for statistical computing and graphics. It is well suited for data analysis purposes and is commonly used when analyzing medically related data.

So now the genomic data is formed and the clinical data is available. And from these data forms the most potential covariates can be analyzed. The genomic data is added to the clinical data and by imposing a penalty constraint on the weights we deal with the unfortunate occurrence of having more predictor variables than observations. Some weights are thus forced to become zero by the Lasso method.

2.3.1. Coxpath[11]

Coxpath is the built-in R function, from the package glmpath, that implements penalization of the cox proportional hazards model by using the predictor-corrector method to determine the entire path of coefficients.

The coefficients are computed with the following criterion

$$\hat{\beta}(\lambda) = \arg\min_{\beta}\left[-\log\{L(y:\beta)\} + \lambda\|\beta\|_1\right]$$

Where L denotes the partial likelihood.

The entire coefficient paths are formulated as $\left\{\hat{\beta}(\lambda) : 0 < \lambda < \lambda_{\max}\right\}$, where $\lambda_{\max}$ is the largest λ that makes $\hat{\beta}(\lambda)$ non-zero, through the predictor-corrector scheme. The active set changes along with λ due to the sparse solution set that is a result of the L1-penalization. The algorithm starts with $\lambda_{\max}$ and estimates the coefficients by reducing λ at each step based on the previous estimate.

The formula for the computation of the coefficients in section 1.3.1 is modified by adding a regularization parameter λ .

2.3.2. Double 10-fold Cross Validation

For the analysis a double cross validation procedure is implemented which ensures minimal fluctuations in the outcome of the prediction model.
One of these is an internal cross-validation method performed in the built-in R function “coxpath” to construct the model and the other is an external cross-validation for prediction purposes.
For the internal cross-validation a standard 10 fold cross validation approach is used which splits the data in 10 equal random sections. These sections are in turn part of the test and training set. At each fold of the external cross-validation the whole data set is split into a training-set and a test-set with a ratio training: test of 45: 5 for the set of 50 patients. Each externally created training set is passed through the internal cross validation method in the function coxpath and therein split into another training and test set. The internal split is used to predict an optimal penalty. This optimal penalty is then used to predict the externally created test set and the coefficients belonging to the selected covariates. The optimal penalty that is used is a median over several values due to instability of the method with respect to different seeds that cause a different split of the dataset to give slightly different optimal penalties. The external 10-fold cross validation is stratified; this means that within the test and training sample the proportion of censored data v.s. uncensored data is approximately the same as the proportion in the original data set. By stratifying the cross validation, the model becomes more stable because it is based on a sample that resembles the test set.

2.3.3. Ranking procedure

The aim of the model is to predict survival of the patients in the data set. The model that is used returns the hazard rates belonging to each patient. These hazards rates are translated into survival by saying that a low hazard rate corresponds to a high survival time using the relationship:

$$S(t) = \exp\left(- \int_0^t \lambda(u)du \right)$$, where $\lambda(t)$ is the hazard function

Within every external cross validation procedure the whole dataset of patients is predicted and saved with a certain order from patients with high hazard rates to patients with low hazard rates. The highest hazard rate gets label 1 and the lowest rate gets label x (# of patients) and everything in between gets a label from 2 to x-1. Now this is done 10 times and every time the label of the initially created test set is taken from the group that it belongs to. When all the test sets have been gathered they are again ordered with respect to their labeling. Now it is clear which patients have a high *predicted* survival rate and which patients have a low survival rate. The patients are then divided into good and bad groups and labeled group 1 and 2 respectively. We use the ranking of the hazard rates rather than the hazard rates themselves to “standardize” between the splits.

Thereafter, Kaplan Meier curves of the *observed* survival times are plotted using the information in the groups. When the prediction is good, the curves lie far apart and the curve of the good group lies above the curve of the bad group.
It is not correct to just rank the hazard rates of the test set within the group of predicted test sets. We are then implicitly assuming that the predictions for the test sets are from the same model. The models that predict each test set are different because coxpath is instable. By using this particular ranking procedure as explained above, the fact that the models within each cross validation are different is well accounted for. It also takes a while before the program is finished running when the genomic data is included in the analysis. For the datasets that were used it took roughly 2 to 3 hours to run the procedure.

Example

There are 10 patients, numbered from 1…10
Suppose that in the first run of the external cross validation, 1/10 of the whole dataset, number 2 is picked as the test set whereas the rest is the training set. So now a prediction is made for this patient based on the remaining 9 patients.

Patient number	Predicted hazard rates
1	?
2	0.8
3	?
4	?
5	?
6	?
7	?
8	?
9	?
10	?

Actually we do not only predict this 1 patient's hazard rate but we predict all 10 patients at once.

Patient number	Predicted hazard rates
1	0.4
2	0.8
3	0.5
4	0.5
5	0.2
6	0.6
7	0.7
8	0.1
9	0.3
10	0.9

Next, the patient numbers are ordered from low to high hazard rates and assigned a label from 1 to 10. The higher the label number, the lower the hazard rate.

Patient number	Predicted hazard rates	label
8	0.1	1
5	0.2	2
9	0.3	3
1	0.4	4
3	0.5	5
4	0.5	6
6	0.6	7
7	0.7	8
2	0.8	9
10	0.9	10

We forget the hazard rates and only store the label and patient number of the test set. Patient no. 2 has label no. 9.

Assume these are the results for the 10 test sample patients when the procedure is repeated 10 times. So every patient has been in the test set once.

Patient number	label
1	8
2	9
3	6
4	1
5	5
6	4
7	7
8	10
9	2
10	3

The patient numbers are again ordered, but now according to the labels from low to high. Also the labels are divided into a good and bad group

Patient number	label	Group
4	1	Good
9	2	Good
10	3	Good
6	4	Good
5	5	Good

3	6	Bad
7	7	Bad
1	8	Bad
2	9	Bad
8	10	Bad

Now we know which patient belongs to which group by selecting the columns “Patient number”and “Group”.

Patient number	Group
4	Good
9	Good
10	Good
6	Good
5	Good
3	Bad
7	Bad
1	Bad
2	Bad
8	Bad

The Kaplan Meier curves are drawn with the information in the group. By ordering the patients in ascending patient number, the column “Group” is in the correct format to place beside the observed survival data.

Patient number	Group
1	Bad
2	Bad
3	Bad
4	Good
5	Good
6	Good
7	Bad
8	Bad
9	Good
10	Good

3. Results

These are the results of the 50 patients when we run the clinical variables combined with the genomic variables.

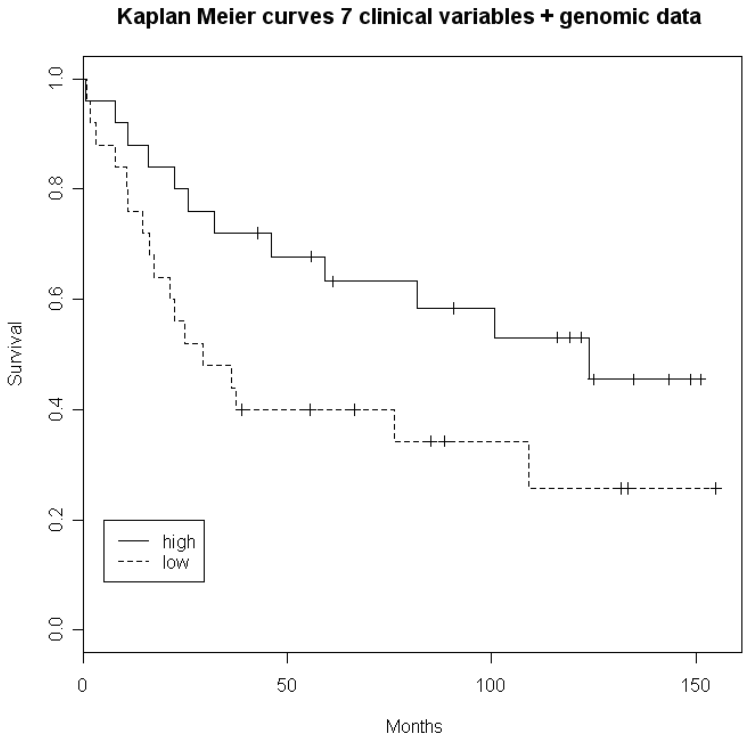


Figure 5: Kaplan Meier curves of survival cancer patients based on both clinical and genomic data.

As we can see in Figure 6, the curves lie apart. Important to know is that to generate these curves the number of genomic covariates added to the 7 clinical variables is only equal to 1. So only 1 region was added to the clinical variables and the model now consists of 8 variables. Actually this is built into the model because the minimum number of genomic covariates that it selects is equal to 1. We can thus say that no regions were selected for prediction purposes. This was already evident from the optimal penalties that were found by the model at each optimization run, the value was always close to 1 which means that all variables should be penalized.

It is not possible to perform a log-rank test on the groups because the data is not independent. The information in the groups is dependent on the observed survival rates; this means that the groups are not prespecified and are only constructed with help from the observed survival rates. Also the point of these Kaplan Meier curves is to measure the performance of the predictions; is there a clear visual difference. Therefore it is not needed to perform any test on the groups.

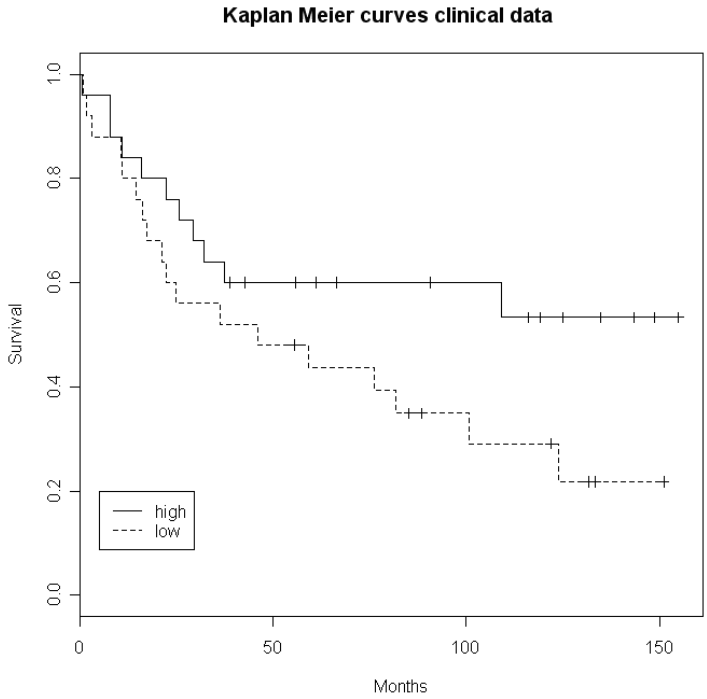


Figure 6: Kaplan Meier curves of 7 clinical variables only

These are the results of only the 7 clinical variables only. From these results we conclude that the genomic data used to generate figure 6 does have some influence on survival prediction.

We now determine which clinical variables are most promising with respect to survival prediction by looking at the Akaike criterion, AIC. We start with a model that includes all clinical variables and drop the variable that has the lowest p-value. Then we look at the AIC of the model to see if it is lower than the previous AIC of the model with that variable in it. The variable combination that gives the lowest AIC is then seen as the best model and thus those variables are seen as the most promising. In this case the most promising variables are the Age of the patient, the differentiation of the tumor and the MSI status.

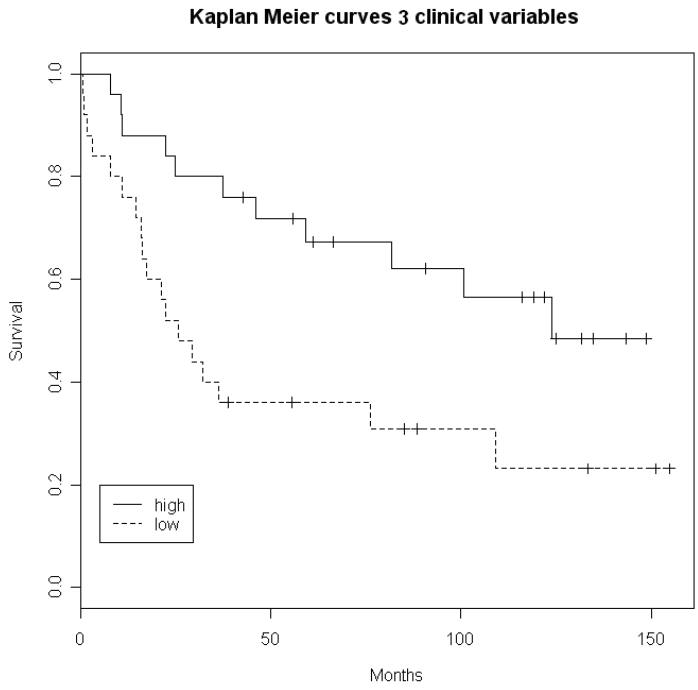


Figure 7: Most promising Clinical variables Age, Differentiation & MSI status

When we now take the most promising clinical variables and repeat the procedure the results are given in Figure 8. These seem to be much better than the previous results. The curves lie much more apart and are more easily separated. So it seems that this model performs much better than the one represented in Figure 6.

Now, if we add the genomic data again to these 3 clinical variables and then consider the results, the findings are gathered in the following figure.

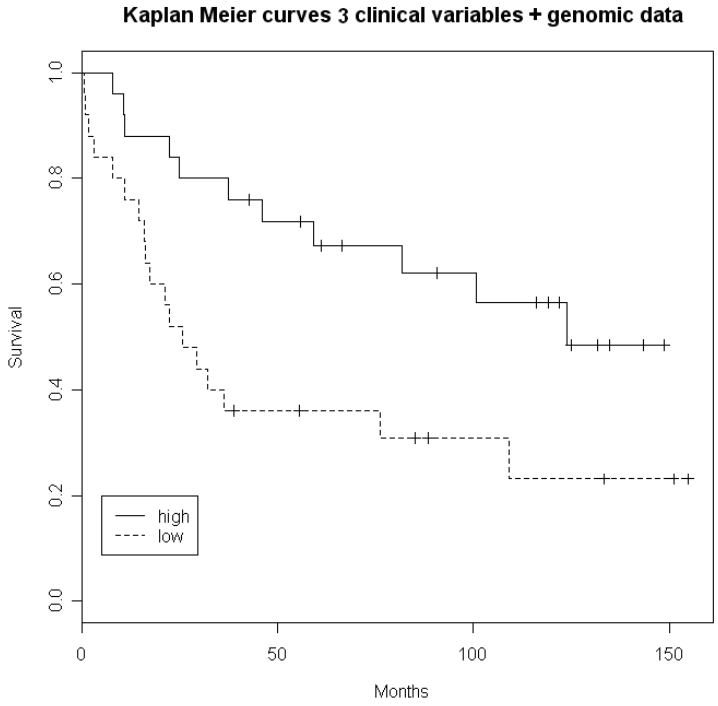


Figure 8: 3 clinical variables combined with genomic data

Even now we see that the genomic data does not add any new information to the predictive power of the clinical variables. The clinical variables are enough for survival prediction with this dataset of 50 patients. The curves are again strangled into each other and do not lie much further apart, on the contrary they even lie a bit closer to each other.

The best model for survival prediction is the model with 3 clinical variables, Age, MSI status and the differentiation of the tumor. It outperforms the other alternative models that include the genomic data for survival prediction. We thus conclude that the genomic data has little added value for survival prediction for this particular data set.

Cancer dataset 2

The next dataset was from 51 patients. There are 6 clinical variables named: Gender, Age, stage of the tumor, PTNMN status, HPV status[a] and silent pattern. All variables except "Age" are binomial variables. For the stage of the tumor, stages 1 and 2 are combined into one group and stages 3 and 4 as well. PTNMN status takes on 3 values: 0, 1 and 2 and has been decoded with dummy variables and is represented by two columns. The genomic data consists of 277 regions and the same procedure as with the other cancer patients was applied. The result of the clinical data combined with genomic data is given in the following figure.

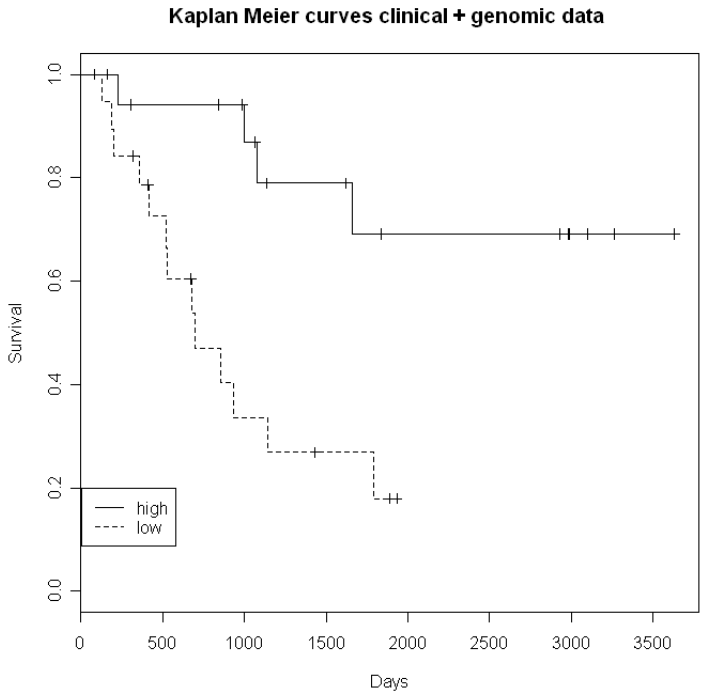


Figure 9: Kaplan Meier curves for the clinical + genomic data

This is a very nice prediction that combines the clinical and genomic data. Now if we only run the clinical variables the figure is as follows.

[a] From a medical point of view, it is worth analyzing the HPV positive tumors only. But the dataset is then too small to construct a good model from a datamining point of view.

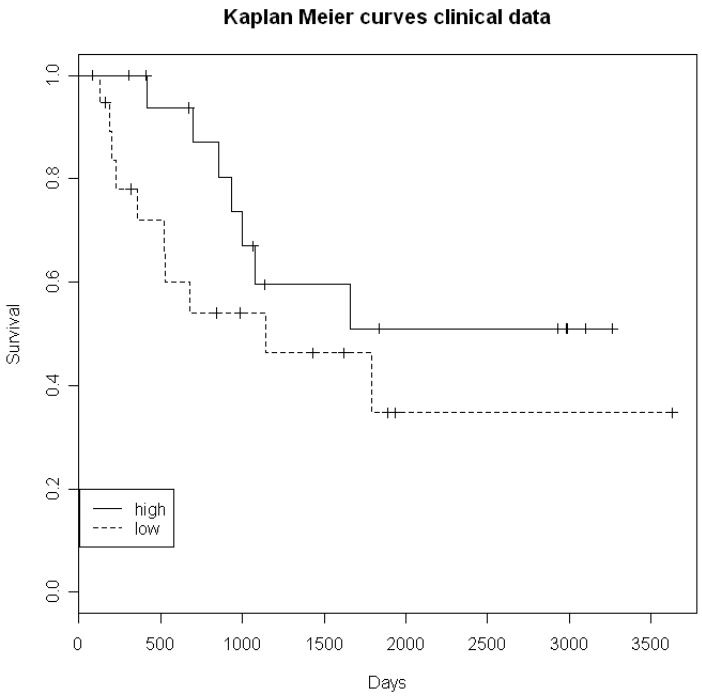


Figure 10: Kaplan Meier curves of the clinical data only

By looking at this figure we conclude that the genomic data adds some value to the clinical data in survival prediction, as figure 10 illustrates a better prediction than figure 11. But just as we saw with the other cancer data set it might be possible that the most promising clinical variables give a much better prediction than the one in figure 10. So now we will run the most promising variables and consider the results.

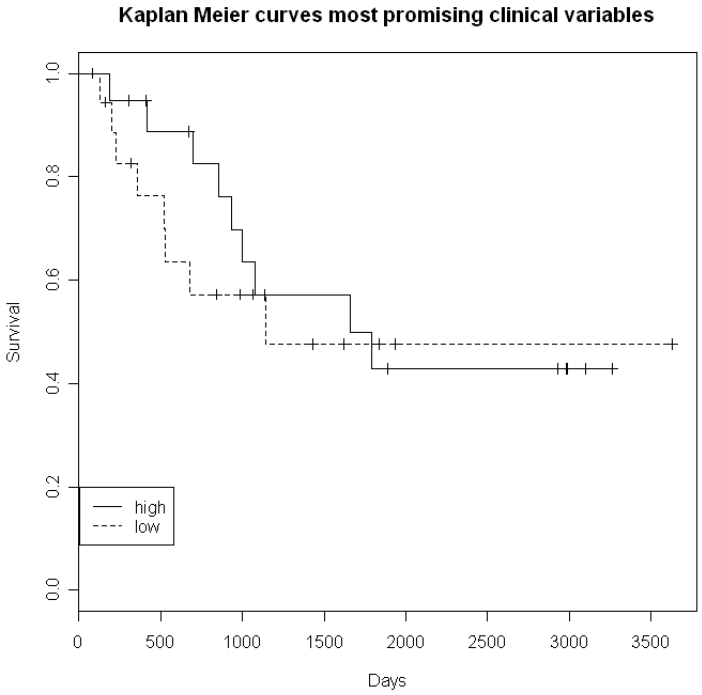


Figure 11: Clinical variable PTNMN

In figure 12 we see that the most promising clinical variable does not predict survival that well. So, based on the previous two figures, for this dataset the genomic data does add some predictive power to the clinical variables. Now we consider which genomic variables cause this nice prediction in figure 10.

When the clinical and genomic data are combined for each cross-validation run, a model is made and some regions are chosen. We consider which regions are chosen the most and then make a graph to see where the "hot-spots" are. But when reading the table you must keep in mind that the variable numbers are NOT region-numbers. To get the region numbers subtract the number of columns that represent the clinical variables, in this case 7 columns. See table 2 for all the information concerning table 1.

Variable no.	Region no.	Occurs(times)
19	12	1
43	34	5
68	61	1
153	146	2
228	221	6
256	249	1
264	257	2

Table 1: Number of times a region occurs in 10-fold cross validation

As we see in table 1, the regions 34 and 221 occur the most, 5 and 6 times resp., followed by region 257 and 146. In the figure below this is visualized. The peaks at the beginning belong to the clinical variables that were not penalized and occur in all 10 models.

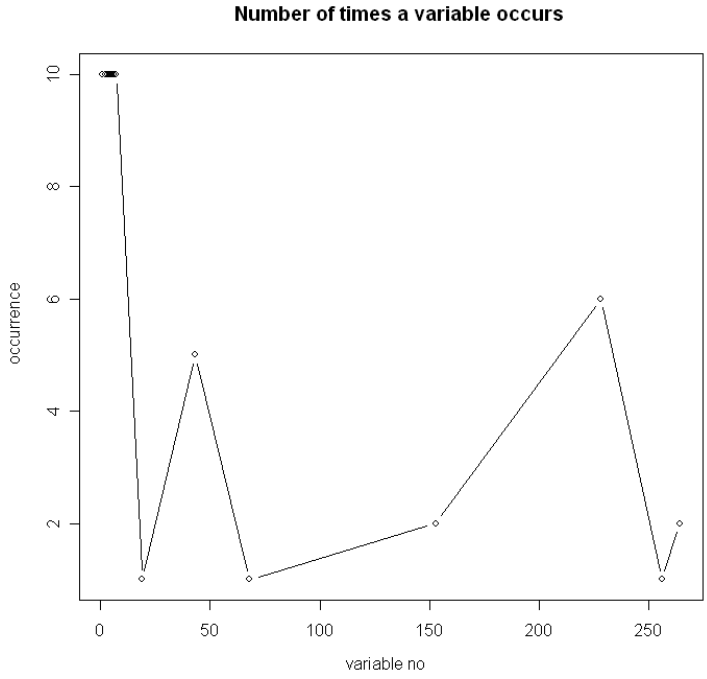


Figure 12: Occurrence of the regions. (Region no. is variable no. -7)

The base pair positions, chromosomes and number of clones corresponding to these region numbers are given in the table below

Region no.	Bp.start	Bp.end	Chromosome	Nclones	Occurs(times)
12	148554987.5	175246568	1	23	1
34	87134141.5	87650067	3	2	5
61	16699355	38253477	5	21	1
146	45156792.5	49333679	10	6	2
221	47082923	47965567	14	2	6
249	33153148.5	38330510	17	12	1
257	17228242	19103014	18	3	2

Table 2: base pair positions and other chromosomal information

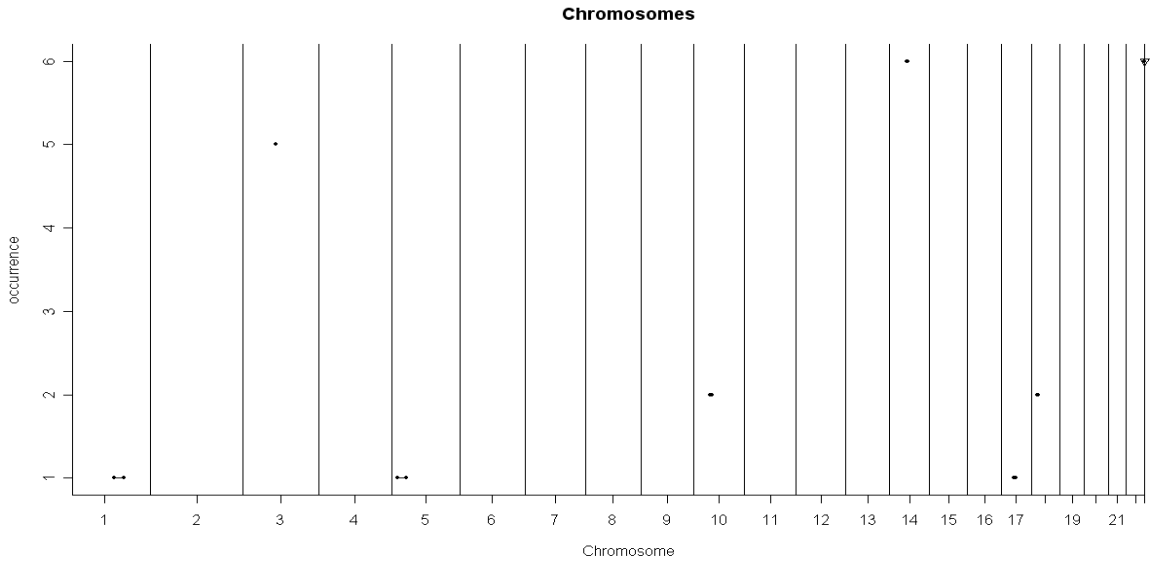


Figure 13: Occurrence on Chromosomes[b]

Now we will only run the genomic data to see which regions are selected.

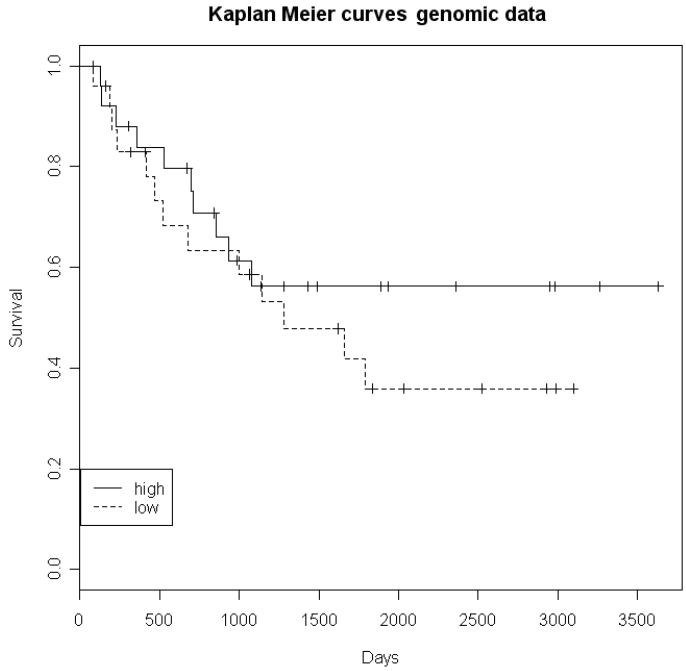


Figure 14: Kaplan Meier curves of genomic data

[b] The up-side-down triangle in the upper right corner has no meaning

The most promising genomic variables that cause figure 14 are given in the table below

Region no.	Occurs(times)
12	3
13	3
33	5
59	1
82	1
83	2
84	1
156	1
193	1
275	1

Table 3: Occurrence of the regions

In table 2, region no. 33 occurs 5 times followed by region no. 12 and 13, 3 times and region 83, 2 times. When we compare these results with the results in table 1, it seems that regions 33 and 34 are important for survival prediction in this dataset. As are region 12 and 13.

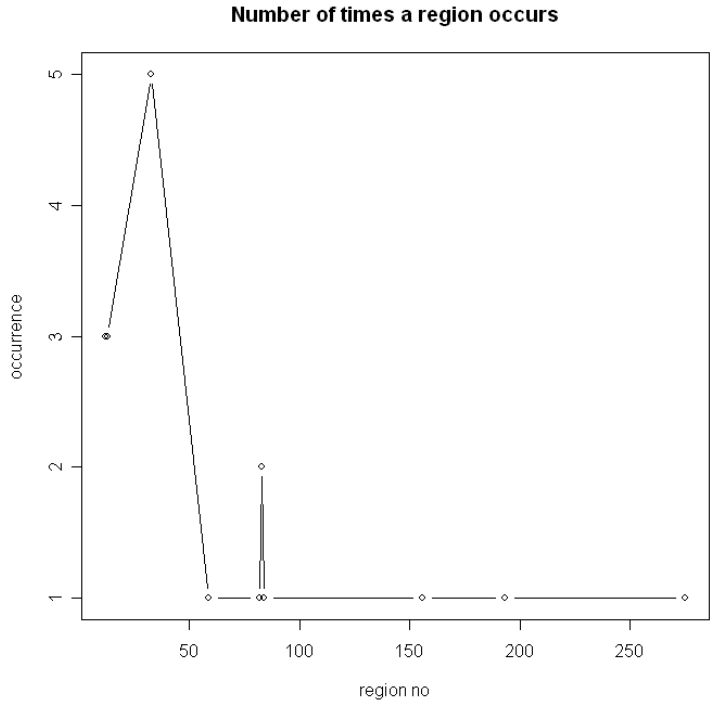


Figure 15: Occurrence of regions

The peak just before region no. 50 belongs to region no 33 as we see in table 2.

The base pair positions, chromosomes and number of clones corresponding to these region numbers are given in the following table

Region no.	Bp.start	Bp.end	Chromosome	Nclones	Occurs(times)
12	148554987.5	175246568	1	23	3
13	177659991.5	180399975	1	5	3
33	83768304	86178494	3	3	5
59	2647827	11407911	5	15	1
82	21493012.5	28338420	7	5	1
83	30097981.5	39348401	7	8	2
84	43018169.5	51470332	7	10	1
156	120459257	135198772	10	19	1
193	46570873.5	53822609	12	9	1
275	14683313	29715891	21	11	1

Table 4: Base pair positions and other chromosomal information

Although in figure 15, the predictive power of these particular regions does not seem that well we say that region 33 and 34 are important for survival prediction of this dataset, also region 12 and 13 as well as region 221.

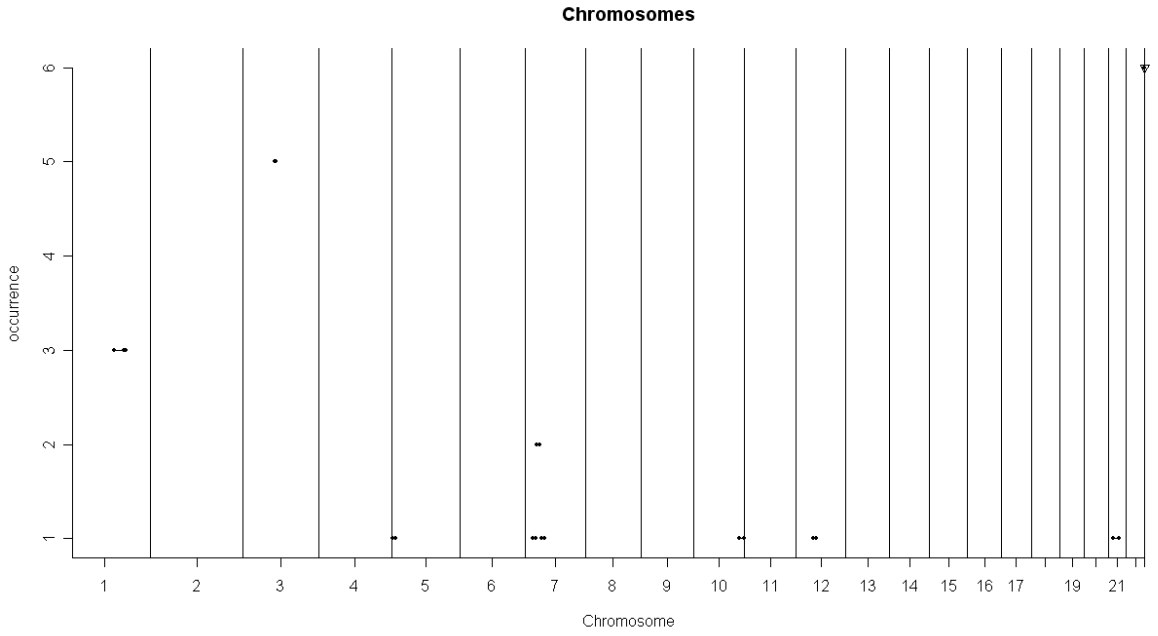


Figure 16: Occurrence on Chromosomes[c]

[c] The up-side-down triangle in the upper right corner has no meaning

Model validating data sets

To check whether this approach to survival prediction really works, the model was tested on 2 non-CCA related datasets. One dataset is the heart transplant dataset provided by R and the other is the breast tumor set from Jane Fridyland[15]. The first set is a low dimensional clinical dataset and the second is a high dimensional aCGH set. For both sets the covariates have predictive power and thus the model has to select the covariates that cause a good survival prediction.

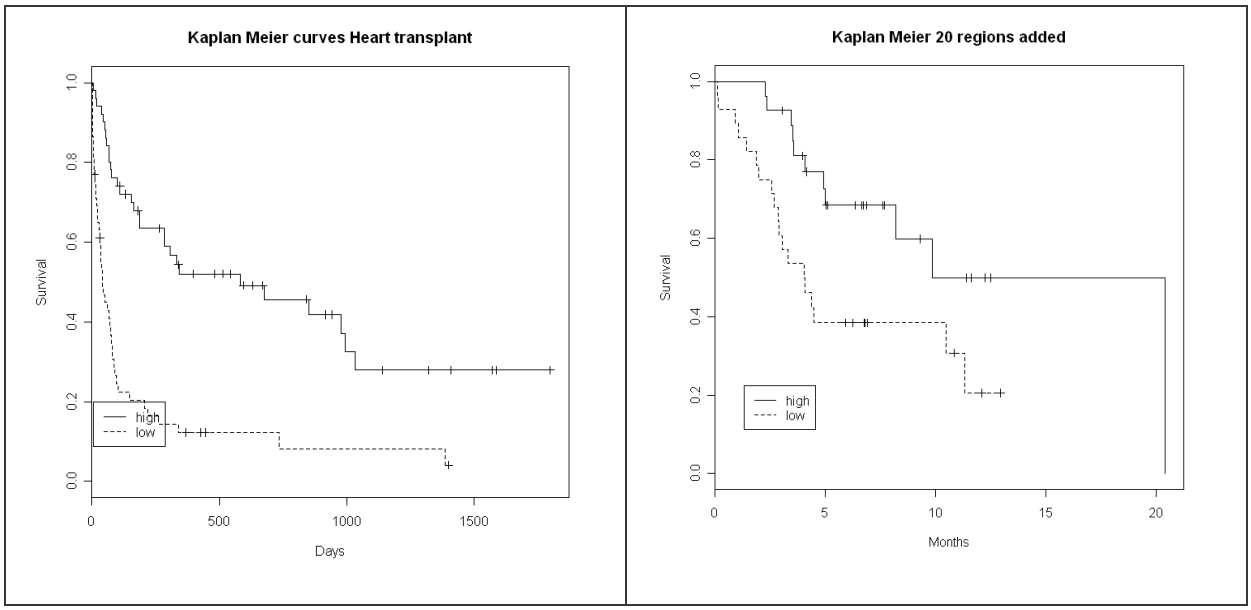


Table 5: Model validating datasets

In these two figures we see that this approach to survival prediction works correctly. As we knew from the beginning that these two datasets consist of predicting variables, the model indeed finds these variables and shows a good prediction.

4. Sensitivity analysis

In this chapter we will take a closer look at the model's performance when some variables are adjusted. This will give us a view on the stability of the model and also help us when interpreting future results. With every good model, the random adjustments should not affect the results drastically. So in this case we would want that every split of the data should give somewhat the same results with respect to survival prediction. The parameter that causes the split is the seed, so different seeds are likely to give a different split and these data-splits might give slightly different results. Now we will consider the prediction results of the datasets for different seeds.

If the model is stable, the results between seeds should not differ that much. The results within one seed are allowed to differ due to the high correlation aspect of aCGH data.

We will compare results of 2 different seeds for the second cancer dataset of the CCA. For the dataset 4 different sets will be examined:

1. Clinical data only,
2. Genomic data only,
3. Clinical with and without penalization + penalized genomic data,
4. Clinical without penalization + wecca clusters + penalized genomic data;

The wecca (weighted clustering of called array CGH data)[16] clusters are added in this part because as it is common to first cluster the data and then predict survival, it is worth checking if this particular order of working adds any value to survival prediction.

The two seeds used are (randomly chosen): `7868 and 81345`
Both the figures and occurrence tables will be given.

Result for clinical data only

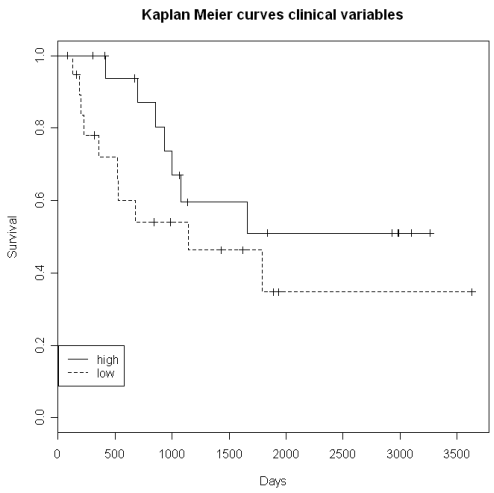
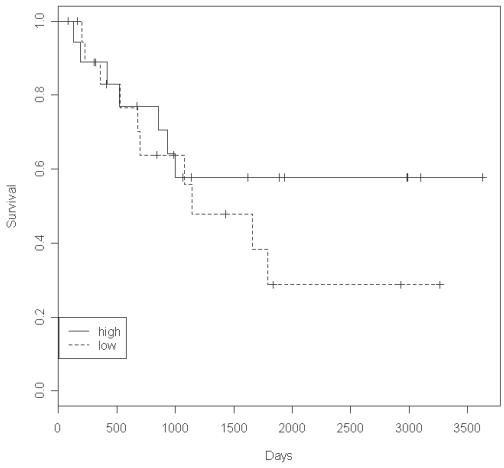
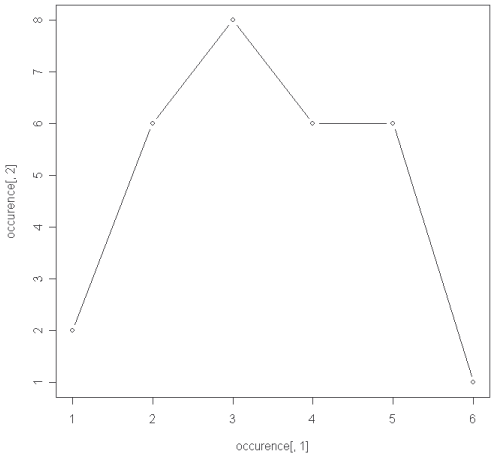
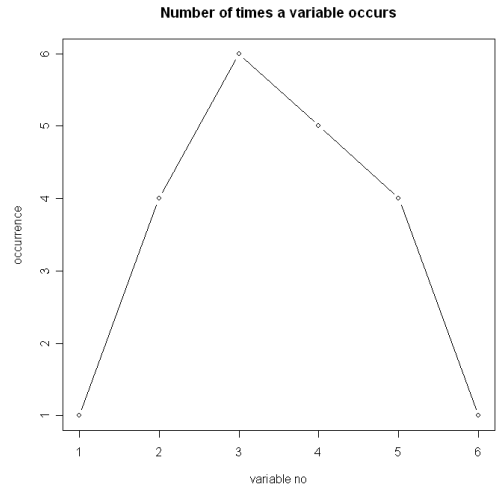
7868	81345
	
	
occur_index occur_count [1,] 1 2 [2,] 2 6 [3,] 3 8 [4,] 4 6 [5,] 5 6 [6,] 6 1	occur_index occur_count [1,] 1 1 [2,] 2 4 [3,] 3 6 [4,] 4 5 [5,] 5 4 [6,] 6 1

Figure 17: Comparison of results of clinical data for different seeds

Results for genomic data only

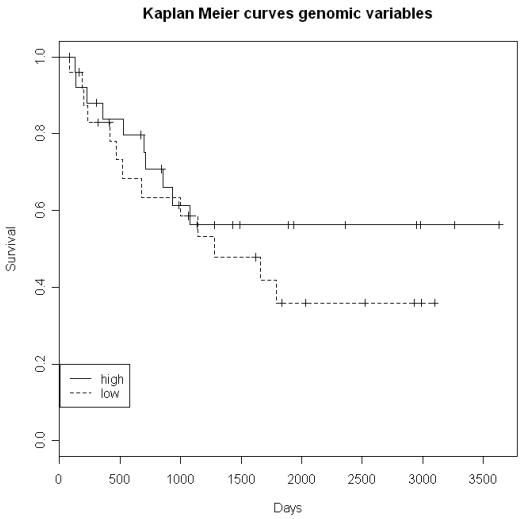
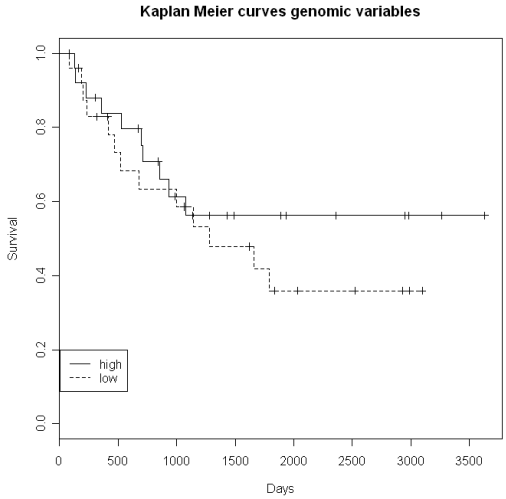
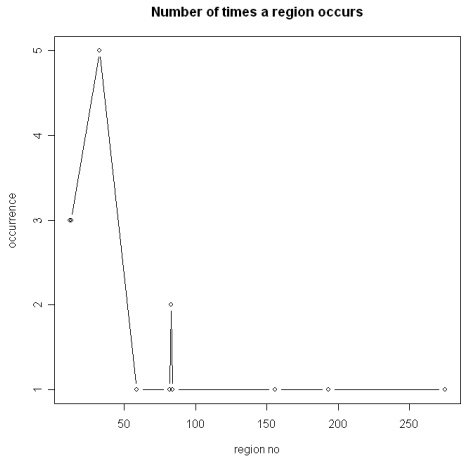
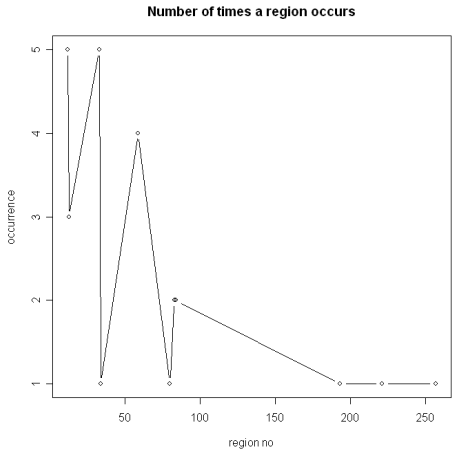
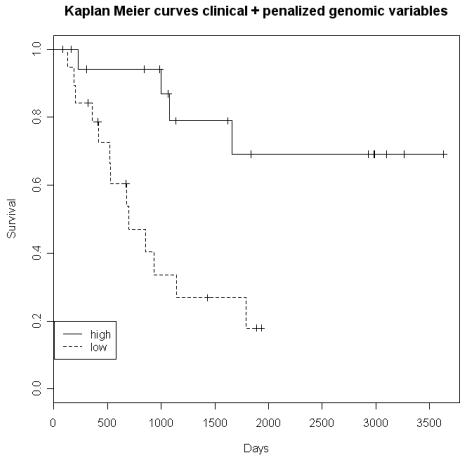
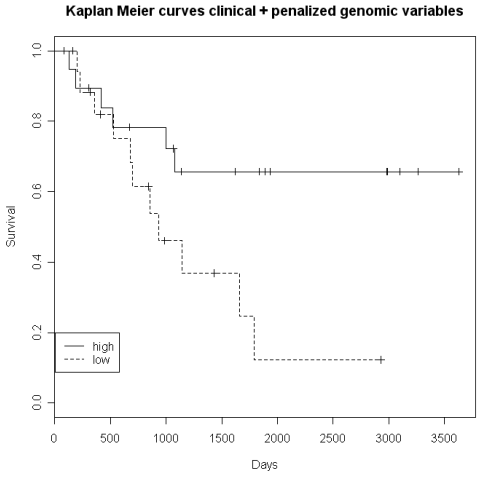
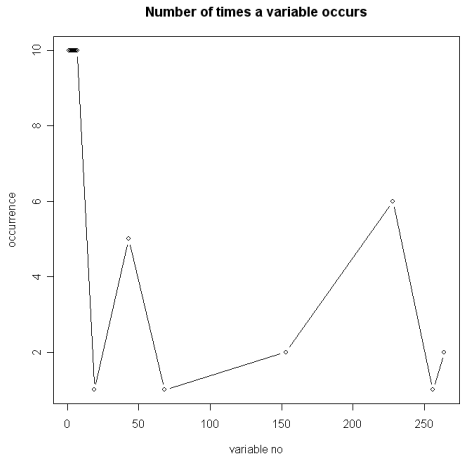
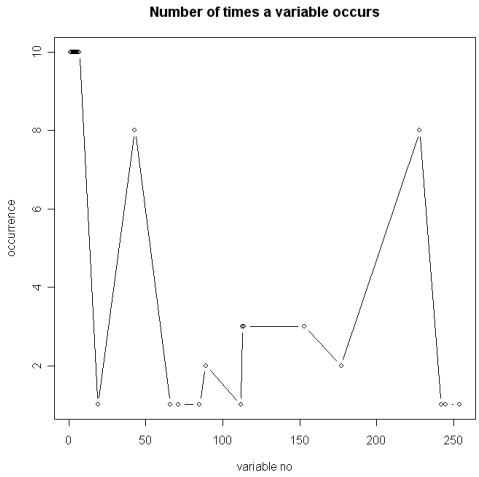
7868	81345
	
	
occur_index occur_count [1,] 12 3 [2,] 13 3 [3,] 33 5 [4,] 59 1 [5,] 82 1 [6,] 83 2 [7,] 84 1 [8,] 156 1 [9,] 193 1 [10,] 275 1	occur_index occur_count [1,] 12 5 [2,] 13 3 [3,] 33 5 [4,] 34 1 [5,] 59 4 [6,] 80 1 [7,] 83 2 [8,] 84 2 [9,] 193 1 [10,] 221 1 [11,] 257 1

Figure 18: Comparison of results of genomic data for different seeds

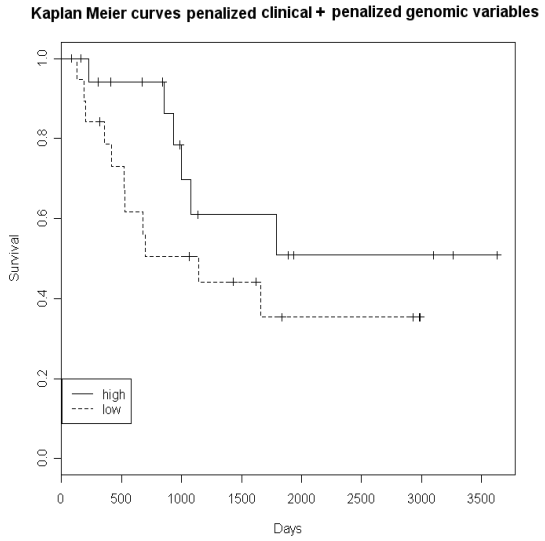
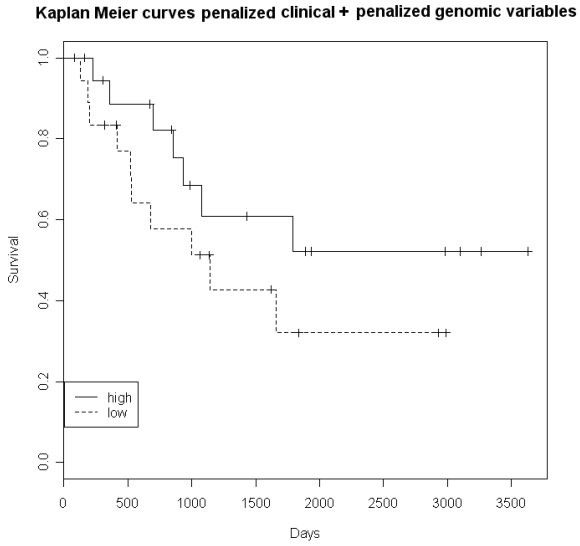
Results for clinical (no penalty) + genomic data with penalty (region no = variable no (occur_ index)-7)

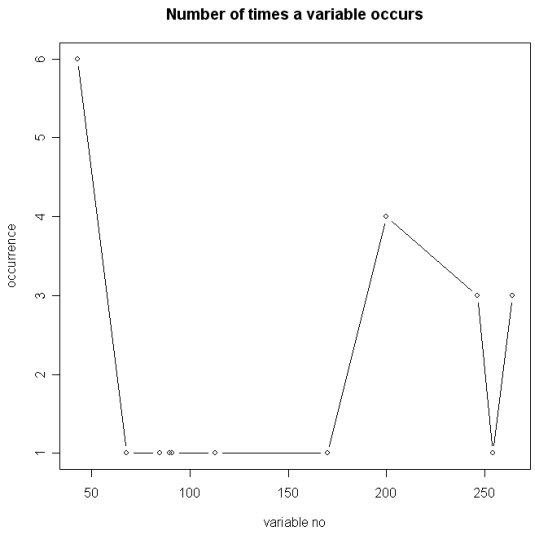
7868	81345
	
	
occur_index occur_count [1,] 1 10 [2,] 2 10 [3,] 3 10 [4,] 4 10 [5,] 5 10 [6,] 6 10 [7,] 7 10 [8,] 19 1 [9,] 43 5 [10,] 68 1 [11,] 153 2 [12,] 228 6 [13,] 256 1 [14,] 264 2	occur_index occur_count [1,] 1 10 [2,] 2 10 [3,] 3 10 [4,] 4 10 [5,] 5 10 [6,] 6 10 [7,] 7 10 [8,] 19 1 [9,] 43 8 [10,] 66 1 [11,] 71 1 [12,] 85 1 [13,] 89 2 [14,] 112 1 [15,] 113 3

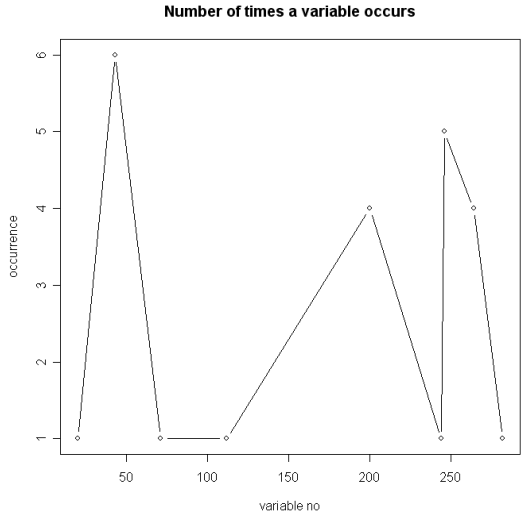
	[16,]	114	3
	[17,]	153	3
	[18,]	177	2
	[19,]	228	8
	[20,]	242	1
	[21,]	245	1
	[22,]	254	1

Figure 19: Comparison of results of clinical(no penalty) +penalized genomic data for different seeds

Results for clinical with penalty + penalized genomic data (region no = variable no (occur_index) -7)

7868	81345
	





	occur_index	occur_count
[1,]	43	6
[2,]	68	1
[3,]	85	1
[4,]	90	1
[5,]	91	1
[6,]	113	1
[7,]	170	1
[8,]	200	4
[9,]	246	3
[10,]	254	1
[11,]	264	3

	occur_index	occur_count
[1,]	20	1
[2,]	43	6
[3,]	71	1
[4,]	112	1
[5,]	200	4
[6,]	244	1
[7,]	246	5
[8,]	264	4
[9,]	282	1

Figure 20: Comparison of results of penalized clinical + penalized genomic data for different seeds(region no = variable no-7)

Results clinical data + wecca clusters

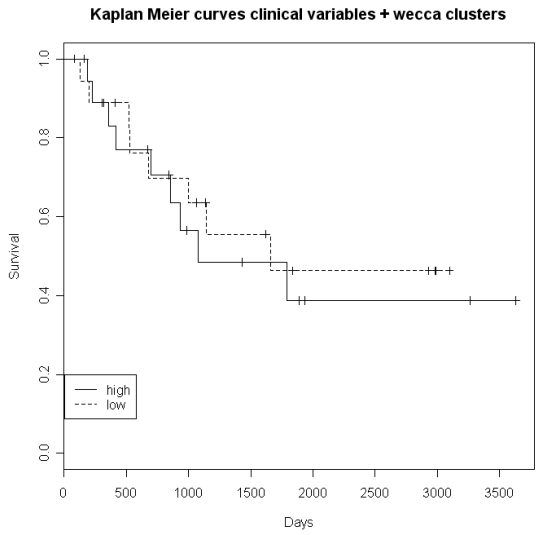
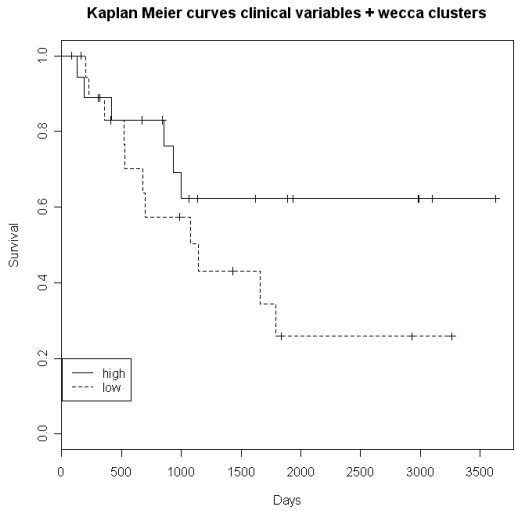
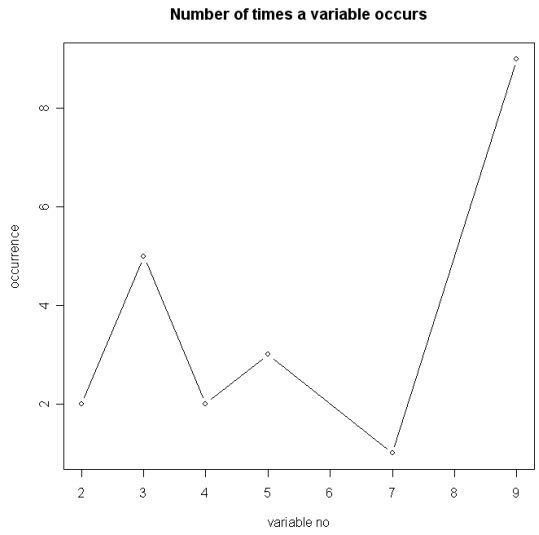
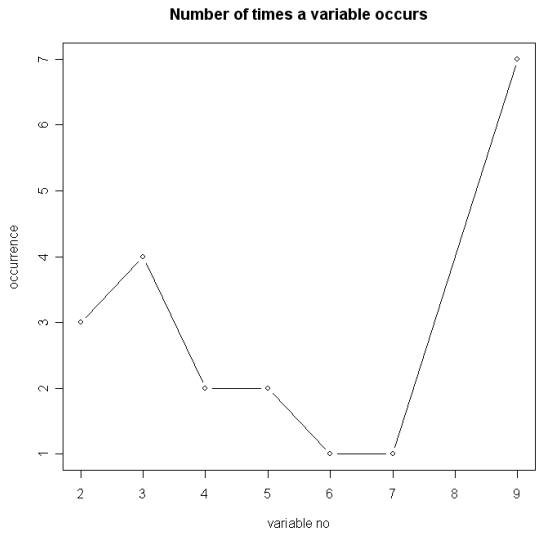
7868	81345
	
	
occur_index occur_count [1,] 2 2 [2,] 3 5 [3,] 4 2 [4,] 5 3 [5,] 7 1 [6,] 9 9	occur_index occur_count [1,] 2 3 [2,] 3 4 [3,] 4 2 [4,] 5 2 [5,] 6 1 [6,] 7 1 [7,] 9 7

Figure 20: Comparison results for clinical data + wecca clusters

Results penalized clinical+ penalized genomic data + clusters (region no=variable no(occur index)-11)

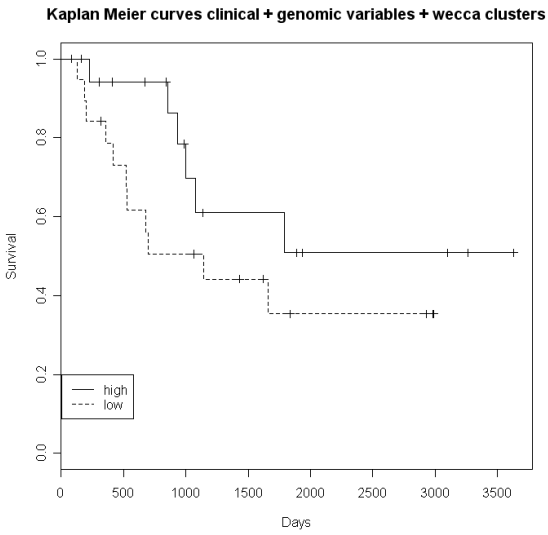
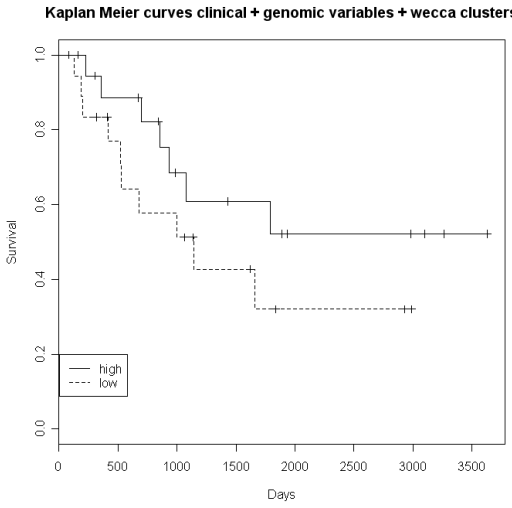
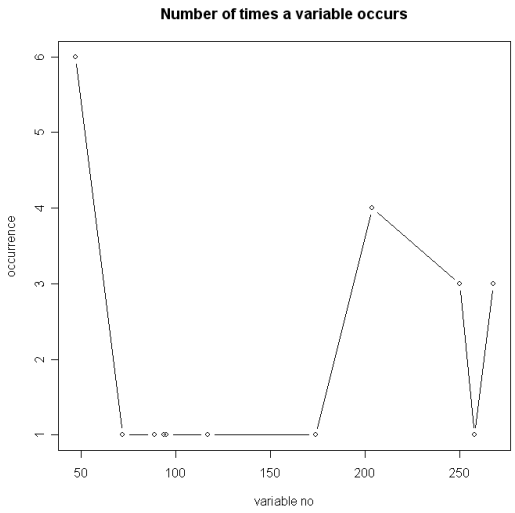
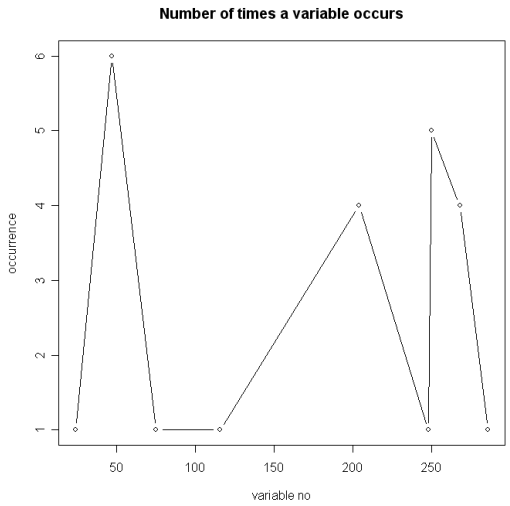
7868	81345
	
	
occur_index occur_count [1,] 47 6 [2,] 72 1 [3,] 89 1 [4,] 94 1 [5,] 95 1 [6,] 117 1 [7,] 174 1 [8,] 204 4 [9,] 250 3 [10,] 258 1 [11,] 268 3	occur_index occur_count [1,] 24 1 [2,] 47 6 [3,] 75 1 [4,] 116 1 [5,] 204 4 [6,] 248 1 [7,] 250 5 [8,] 268 4 [9,] 286 1

Figure 21: Comparison results penalized clinical + penalized genomic data + wecca clusters (region no=variable no-11)

For some of the figures there is a slight difference in the predictions. To get rid of this difference, the mean over several seeds could be taken and worked with. This is not implemented because the whole procedure is very time-consuming when the genomic data is included. For one seed the procedure takes around 2 hours and we would at least want 10 seeds to take the mean over.
Although the differences arise, the model seems to be robust because in all cases we would draw the same conclusion with respect to the most promising variables (regions and/or clinical data). On Chromosome 3 the part between base pair position 17263371 and 134686034, thus between region no. 30 and 39 is chosen very often as adding information for prediction purposes. Also the regions between 230 and 240 are chosen very often.
As mentioned before, aCGH data is highly correlated and this results in appearance of part of the correlated set but not the whole set. So if region 1 and 2 are highly correlated they will unlikely appear together in a predictive model.
By analyzing the results in figure 20 and 21 we conclude that the results are the same as without the clusters. So clustering first does not have an impact on the prediction results. Compare figure 20 with figure 16 and compare figure 21 with figure 19.

For information on the base pair positions and chromosomes corresponding to the region numbers see appendix A. In appendix B the figures of the chromosomes are shown.

Future implementations

For further analysis it would be interesting to consider an alternative model with an intermediate marker and compare it with the standard model. The expectation is then that the alternative model will perform better than the standard model.
Another consideration would be to take the mean over the results of different seeds as the end results. Basically stabilizing the model further would be the main focus in future implementations. The efficiency of the calculations is also improvement worthy.
Concerning the results of this analysis with respect to the second cancer dataset, a biological interpretation of the findings is worth considering.

Conclusion

Assignment

The genomic data has little added value for the first data set that I have worked on. The clinical data gave a much better survival prediction than the clinical and genomic data combined. For this set no limited selection of genes were given for prospective measurements.

The genomic data for the second cancer set had some added value for survival prediction. Multiple regions were selected as promising covariates, switching between different regions that are correlated and separately lead to the same prediction (Appendix C). These regions are translated into base pair positions and chromosomes. It does not seem to matter whether the data is clustered first and then predicted, because the results are the same either way.

Technical

Throughout the analysis of the data some difficulties arose and were dealt with. Apart from the cross validation method in the function coxpath, an external cross validation was implemented to ensure minimal fluctuations in the outcome of the prediction model (section 2.3.2). The issues concerning stability of the function coxpath were solved by working with the median over several optimal penalties that the function produced. With respect to the stability issues of the external cross validation procedure, a stratified approach was implemented. Also ranking the predictions as an evaluation of the predictive performance of the model had its own complications that were taken care of (section 2.3.3).

The built-in R function coxpath, that implements the lasso method through the predictor-corrector scheme, has a hard time finding the most promising covariates. As the search-space (dimension of the data) gets larger, from low to high dimensional, the function needs more time to find the most promising covariates and produce results.

Stratifying the samples in the external 10-fold cross validation leads to much more stable results. Although differences are still detected, the conclusion with respect to the most promising variables stays the same in most cases.

Producing the Kaplan Meier curves needs some special attention. The hazard rates that coxpath produces should not be considered as coming from the same model. Therefore the hazard rates cannot be ranked within only test samples. The ranking should be done over the test and training set for one cross-validation. This gives the correct results.

The last part of the problem definition was not implemented due to unforeseen reasons; the first dataset did not perform as was expected and switching to the next dataset and repeating the whole procedure was very time-consuming.

Literature

1. **Daniel Pinkel & Donna G. Albertson**: Array comparative genomic hybridization and its application in cancer, *Nature Genetics* **37**, S11 - S17 (2005))
2. **S. Knudsen:** A Guide to analysis of DNA microarray data*(Book)*
3. **Mei-Ling Ting Lee**: Analysis of microarray gene expression data*(Book)*
4. **AE Oostlander, GA Meijer & B Ylstra**: Microarray-based comparative genomic hybridization and its applications in human genetics, *Clinical Genetics* 2004 Dec; 66(6):488-95.
5. **Beatriz Carvalho, Marije Weiss, Bauke Ylstra, Gerrit A Meijer**: Microarrays- Detecting DNA Copy-Number changes, *Encyclopedia of Medical Genomics and Proteomics*, 2004 Dec 9
6. **Phillipe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi and Emmanuel Barillot:** Analysis of array CGH data: from signal ratio to gain and loss of DNA regions, *Bioinformatics,* 2004 Dec 12; 20(18):3413-22. Epub 2004 Sep 20.
7. **Mark A. Van de Wiel, Kyung In Kim, Sjoerd Vosse, Wessel N. Van wieringen, Saskia M. Wilting and Bauke Ylstra:** CGHcall: calling aberrations for array CGH tumor profiles, *Bioinformatics* 2007 23(7):892-894
8. **Peter J. Smith :** Analysis of Failure and survival data*(Book)*
9. **Robert Tibshirani :** The Lasso method for the variable selection in the Cox model, *Stat Med.* 1997 Feb 28;16(4):385-95
10. **Venables W., Ripley B.:** Modern applied statistics with S*(Book)*
11. **Mee Young Park and Trevor Hastie:** L1-regularization path algorithm for generalized linear models, *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* Volume 69, Number 4, September 2007 , pp. 659-677(19)
12. **M.C.M. de Gunst:** Statistical Models*(Lecture Notes)*
13. **Ian H. Witten & Eibe Frank:** Datamining, Practical Machine Learning Tools and Techniques*(Book)*
14. **Bauke Ylstra, Paul van den IJssel, Beatriz Carvalho, Ruud H. Brakenhoff and Gerrit A. Meijer:** BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization(array CGH) , *Nucleic Acids Res*. 2006; 34(2): 445–450
15. **Jane Fridyland et al:** Breast tumor copy number aberration phenotypes and genomic instability
16. **Van Wieringen, W.N., Van de Wiel, M.A., Ylstra, B.:** Weighted clustering of called aCGH data", *Biostatistics*, 9(3), 484-500.

Appendix A: Chromosomal information

Region no.	bp.start	bp.end	chromosome	nclone
1	1082137.5	27425550	1	23
2	27602510	36570203	1	11
3	37131978	40791193	1	7
4	42966359	44414952	1	3
5	45332458.5	45421058	1	2
6	46307745	56599427	1	11
7	57482073.5	66680940	1	9
8	67244579	101080679	1	33
9	101256450.5	102773091	1	3
10	105460590.5	107342958	1	4
11	109784674.5	119515493	1	12
12	148554987.5	175246568	1	23
13	177659991.5	180399975	1	5
14	181277242	196309380	1	12
15	198972448	208196008	1	11
16	209802613	245340016	1	29
17	670946	8726800	2	9
18	9713739	33176566	2	23
19	34365393	82693516	2	47
20	95734306	97679803	2	4
21	99279170	113147879	2	14
22	115649114.5	123022408	2	12
23	127171293.5	156612532	2	24
24	158371639	167662287	2	9
25	169112011	180528807	2	17
26	183346163	200147929	2	19
27	201759919.5	205378315	2	5
28	208914039	242568229	2	33
29	282248	15860500	3	22
30	17263371	32650496	3	18
31	35056169	77661857	3	42
32	78480379	82727632	3	6
33	83768304	86178494	3	3
34	87134141.5	87650067	3	2
35	95613019	97099529	3	3
36	101555541.5	115618405	3	14
37	116354803.5	125335575	3	14
38	125993512.5	130084777	3	5
39	131536785	134686034	3	4
40	135928150.5	142172753	3	9
41	143157972	147098138	3	5
42	147780779	157199723	3	16
43	158244141	158898452	3	2
44	161232887	183264103	3	22
45	184223430.5	188090713	3	5

46	188776868	193832313	3	7
47	194509477	199072636	3	6
48	1169331	31275196	4	24
49	31728611.5	33858656	4	3
50	34693806.5	44160499	4	9
51	45840583.5	48750117	4	4
52	53280213	59766010	4	7
53	60441889	72102747	4	11
54	72468569.5	86610272	4	15
55	87696317	91976935	4	7
56	93563909	116534872	4	27
57	116980753	155097151	4	35
58	156601040.5	191117566	4	31
59	2647827	11407911	5	15
60	12206084.5	16441678	5	4
61	16699355	38253477	5	21
62	40197444	42897896	5	4
63	51158140.5	64221566	5	17
64	64691367	87681030	5	28
65	90885304.5	130383778	5	36
66	131471881	150065249	5	22
67	151461554	169688708	5	17
68	171078138	175912462	5	3
69	176798016	180569503	5	3
70	149316.5	11482556	6	13
71	12492992	29408441	6	20
72	32116370.5	44057950	6	14
73	45076252.5	54815485	6	12
74	57372091.5	65840383	6	6
75	66949184.5	77706170	6	11
76	79624339	88867103	6	9
77	90444038	123028985	6	32
78	125465214.5	146165786	6	24
79	148518019.5	170751094	6	26
80	56402.5	5909932	7	6
81	6008802.5	19898273	7	14
82	21493012.5	28338420	7	5
83	30097981.5	39348401	7	8
84	43018169.5	51470332	7	10
85	54816769.5	56527930	7	5
86	62812550	67431714	7	7
87	68141452	73539917	7	6
88	76082487.5	85235255	7	11
89	86787001.5	96970097	7	12
90	97206807.5	100870046	7	6
91	102836447	110712084	7	8
92	115543035.5	119591670	7	7
93	121491979.5	124578081	7	5
94	126994061	131027222	7	6
95	131890264	134570368	7	4

96	136147154.5	158367707	7	25
97	384747.5	1658607	8	4
98	2877062.5	11715250	8	15
99	14369641.5	18545955	8	12
100	20598302	23196929	8	7
101	23750000.5	27369932	8	11
102	28432581	29011164	8	2
103	30903408	32622664	8	4
104	32942029.5	35289479	8	4
105	35686537.5	36439850	8	2
106	37287861	38389035	8	2
107	39911101.5	40575159	8	3
108	42580960.5	43377263	8	2
109	47815257.5	48825348	8	4
110	49214954.5	55586129	8	11
111	57037569.5	61763697	8	9
112	61956207	94506486	8	52
113	94940937	100637597	8	28
114	100838870	104441899	8	20
115	104520750	110524664	8	33
116	110548215.5	110806841	8	3
117	110951150.5	111318355	8	4
118	111571023.5	112599840	8	9
119	112763681	114152064	8	9
120	114329556.5	115837776	8	7
121	116052268	116101802	8	2
122	116779163.5	122798603	8	10
123	123611957	140364554	8	37
124	142286197	146081405	8	5
125	279255.5	1414534	9	3
126	2589860.5	10031363	9	10
127	11256200	14503257	9	6
128	15310900.5	22529658	9	10
129	24167079.5	24973635	9	2
130	25771711	29566162	9	8
131	32083129.5	34971556	9	4
132	68443399	68873034	9	2
133	70602915.5	73710471	9	5
134	76171853	81064934	9	7
135	82108569	87010717	9	9
136	87331683	90313837	9	8
137	91295878	103708844	9	17
138	104661535.5	111043847	9	8
139	111824052.5	114726491	9	4
140	115061420.5	119867044	9	13
141	120187893.5	125359796	9	32
142	126334047.5	129952204	9	7
143	130596817.5	138221187	9	10
144	290395	37758141	10	41
145	38263115	38707135	10	2

146	45156792.5	49333679	10	6
147	52983888	56795996	10	4
148	57631404.5	60183000	10	3
149	60895056.5	68214697	10	9
150	69521424	78888899	10	11
151	79766494	89595067	10	11
152	89613443	95849156	10	12
153	97292594.5	97696741	10	2
154	99700897.5	107774084	10	9
155	108487482.5	119585642	10	13
156	120459257	135198772	10	19
157	222718.5	583612	11	2
158	1685343.5	5840508	11	7
159	7401638.5	9755152	11	4
160	10371720	18642021	11	11
161	20289580.5	29426614	11	6
162	30653078.5	30724721	11	2
163	31838563.5	33978314	11	5
164	34935871.5	35734188	11	4
165	36761414.5	38593022	11	3
166	40282253	42942455	11	4
167	44240252.5	44517873	11	2
168	46299619	54923451	11	4
169	56317876	56960973	11	2
170	59009181	60956645	11	3
171	61395262	66791780	11	7
172	69184373.5	69772711	11	7
173	70306993	70693204	11	3
174	72087733	72475990	11	2
175	73762785.5	73786806	11	2
176	74646515	78716776	11	6
177	79727372.5	81372619	11	3
178	83113166.5	88833802	11	6
179	90989728.5	100222469	11	14
180	101326467	101831922	11	2
181	103884277.5	105021687	11	2
182	107631297.5	110633477	11	8
183	110774445	114126567	11	26
184	114256818	116286509	11	11
185	116559412	118930758	11	28
186	119364093.5	124567036	11	12
187	124673507	133610461	11	9
188	152878.5	7918576	12	12
189	10931834	18529362	12	6
190	19281107.5	30840108	12	12
191	32001761.5	33252021	12	3
192	37284773	46228445	12	11
193	46570873.5	53822609	12	9
194	55820163.5	56391901	12	3
195	57741439.5	58235069	12	2

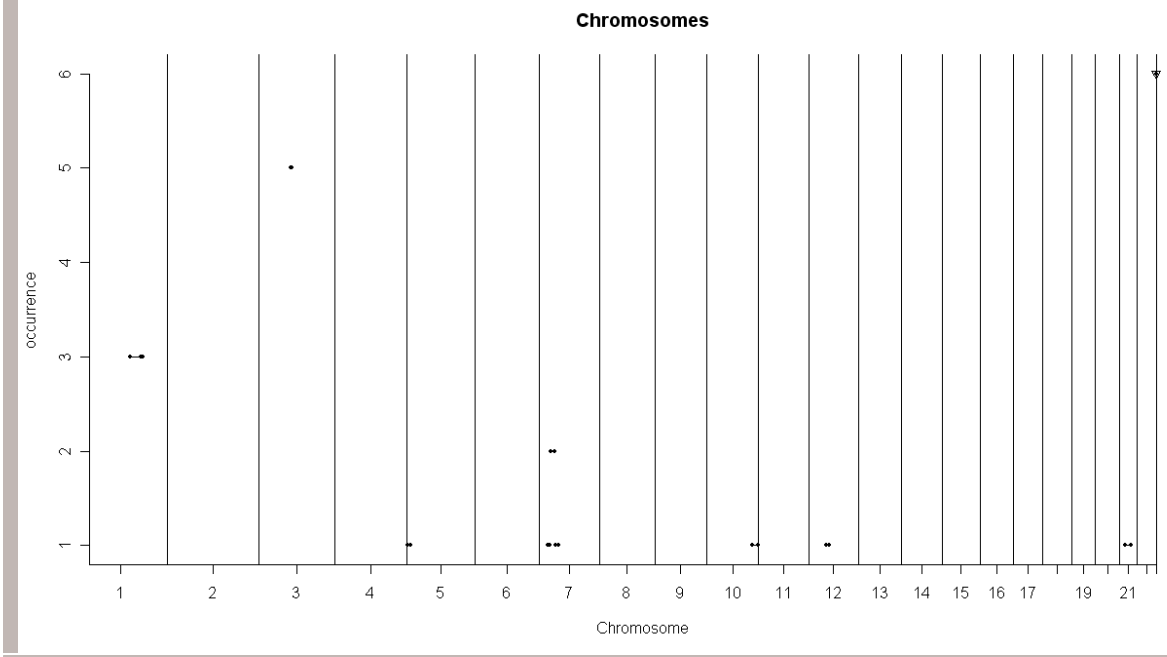
196	59007697.5	60290803	12	2
197	62854620.5	68799206	12	6
198	69815257.5	93104787	12	29
199	94197224	104783981	12	13
200	105048665.5	123052780	12	21
201	124965712	132291094	12	9
202	19072507.5	25211659	13	8
203	25981967.5	32233579	13	10
204	32869061	34995405	13	4
205	37218236	52270268	13	27
206	52889091	54386448	13	8
207	54784267	56939411	13	13
208	57163227	62083043	13	22
209	62134068	67604362	13	24
210	67765227.5	69772427	13	8
211	70387003.5	76289311	13	26
212	76577603	79947363	13	13
213	80285162.5	82118317	13	12
214	82296336	92590531	13	46
215	93033001	94507192	13	5
216	95415986	101413033	13	7
217	104357105	108254496	13	5
218	109284088	113862502	13	7
219	19650648	35696164	14	14
220	37112215.5	46721706	14	15
221	47082923	47965567	14	2
222	49524306	68493100	14	21
223	68610075	77561407	14	12
224	79430708.5	87578121	14	11
225	88902948.5	96220508	14	8
226	97149927.5	106227119	14	8
227	151965	31888948	15	12
228	32974152.5	43513386	15	11
229	45222417	52733937	15	9
230	53412246.5	55012221	15	2
231	56614086.5	58253094	15	3
232	60806644.5	72845394	15	16
233	74573490.5	83784667	15	10
234	85274479.5	99968525	15	13
235	102183	4084406	16	6
236	4520176.5	15492725	16	12
237	16087835.5	31072620	16	14
238	45595453	57045979	16	13
239	58263581	58885009	16	2
240	60015020.5	63539674	16	7
241	64038415	68620150	16	9
242	68722884.5	82465553	16	17
243	82936408.5	84835528	16	5
244	87172104.5	88532308	16	5
245	905054.5	18027536	17	24

246	19189107	23221015	17	6
247	23823873.5	26502533	17	5
248	27506193	30604754	17	6
249	33153148.5	38330510	17	12
250	39009267.5	46276017	17	11
251	46531868.5	69543352	17	27
252	69876869	71155116	17	2
253	72427153	78374443	17	7
254	212950	5352431	18	8
255	6127778	11770828	18	8
256	14921070	17197559	18	2
257	17228242	19103014	18	3
258	22189235.5	22835812	18	2
259	25681738	32803677	18	8
260	33715236	75619995	18	42
261	211053	6435182	19	9
262	6612622.5	8597279	19	4
263	9119442	10609625	19	3
264	12069223.5	18157928	19	10
265	19492271.5	22862445	19	6
266	23643657.5	52418891	19	24
267	53350663	63687254	19	9
268	230000	28168433	20	21
269	29779091.5	34257711	20	14
270	34886227	43401507	20	33
271	43932686.5	50732066	20	31
272	50773048.5	55316098	20	20
273	55918927	59053153	20	12
274	59482816	63542600	20	12
275	14683313	29715891	21	11
276	32934538	46846396	21	14
277	15678161	49416895	22	44

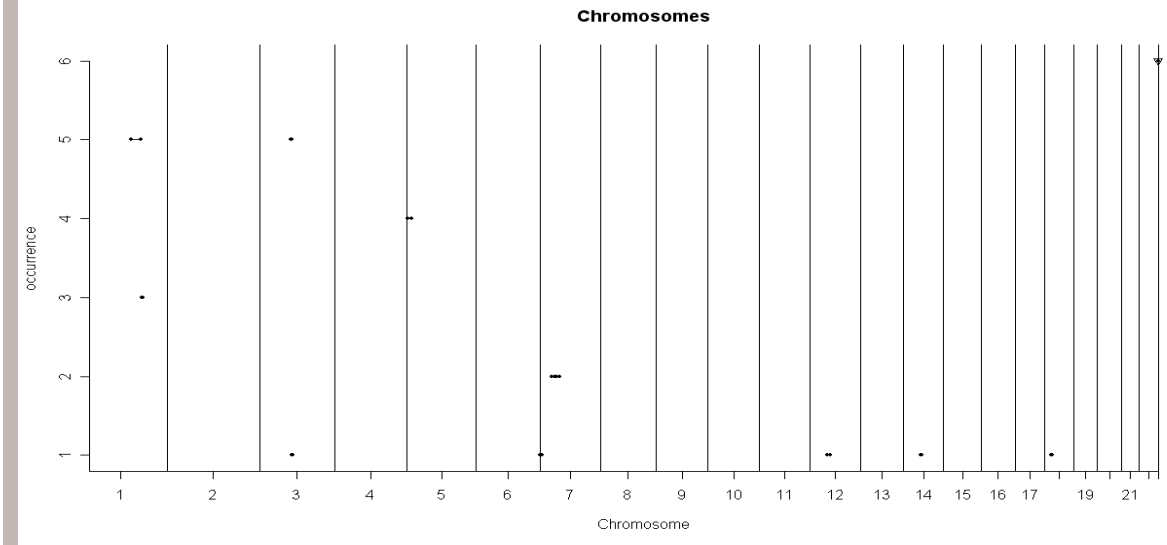
Appendix B: Figures related to the chapter "Sensitivity analysis"[d]

Results for genomic data only

Seed: 7868



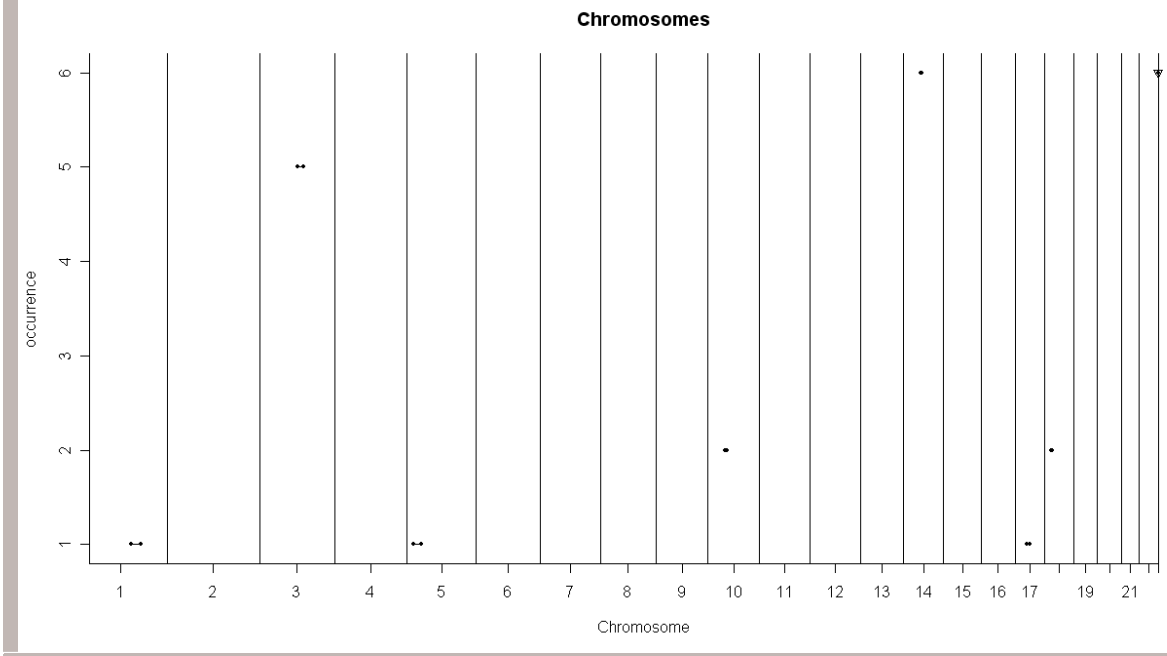
Seed: 81345



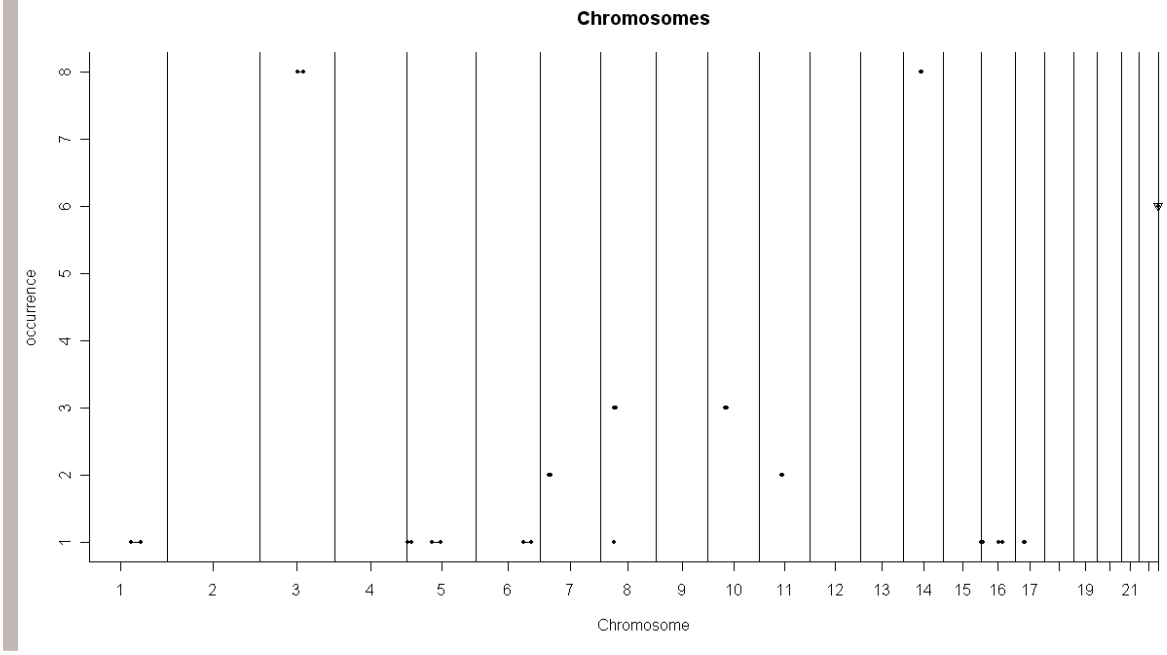
[d] The up-side-down triangles in the plots have no meaning

Results for clinical+ genomic data with penalty (region no = variable no (occur_ index)-7)

Seed: 7868

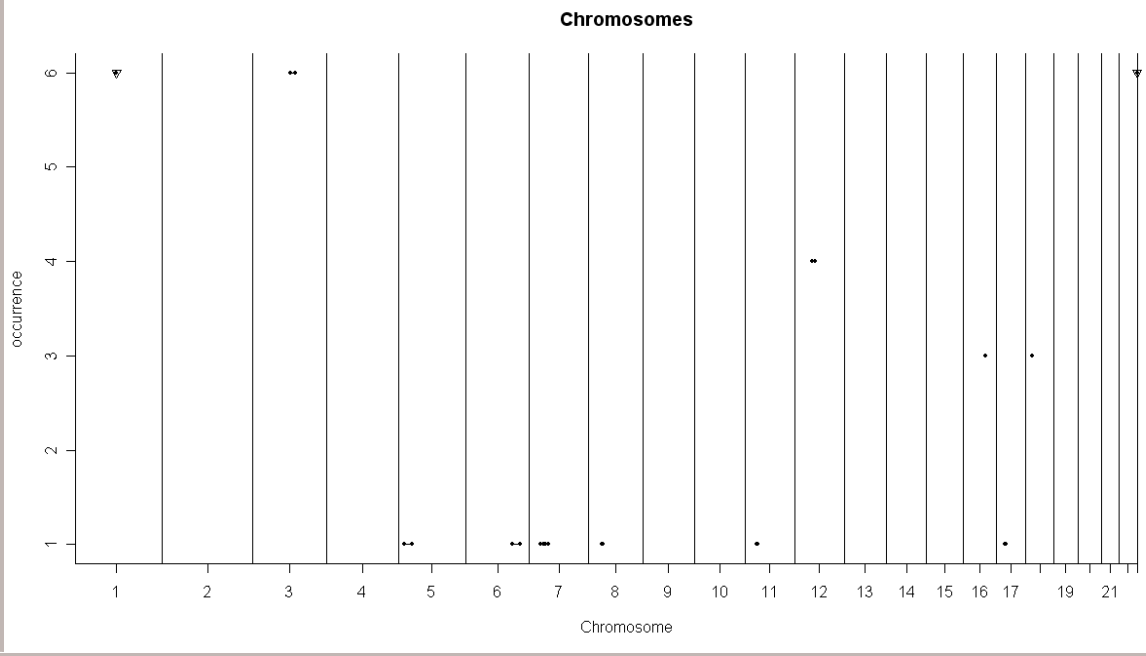


Seed: 81345

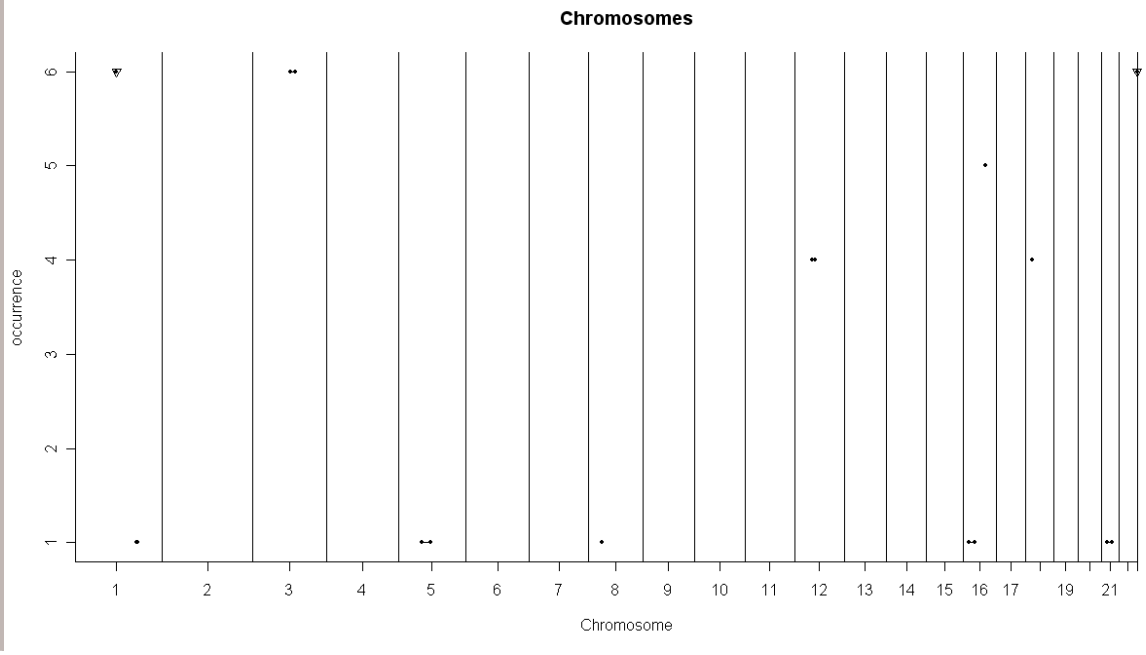


Results for clinical + genomic data without penalty (region no = variable no (occur_index) -7)

Seed: 7868

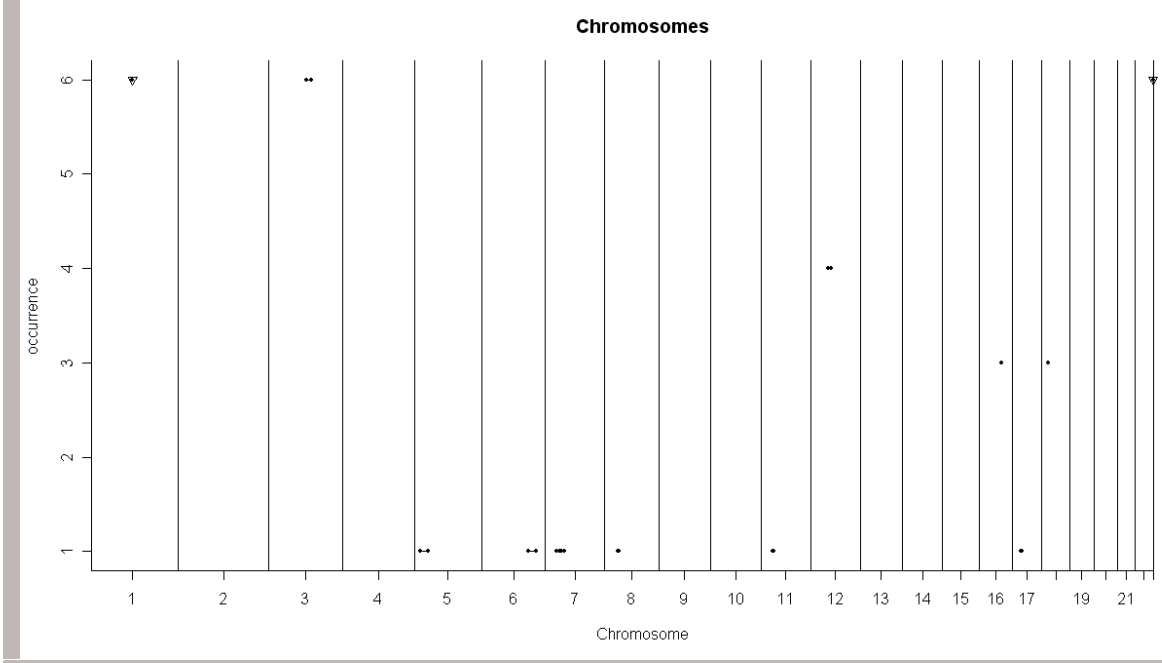


Seed: 81345



Results clinical+(non-penalized)genomic data + clusters(region no=variable no(occur_index)-11)

Seed: 7868



Seed: 81345