MACAT - Microarray Chromosomal Aberration Tool

Joern Toedling, Sebastian Schmeier,Matthias Heinig,
Benjamin Georgi, and Stefan Roepcke

Contents

1 Introduction

This project aims at linking the term *differential gene expression* to the chromosomal localization of genes. MACAT is motivated by the common observation of chromosomal aberrations in tumor cells. The basic idea is to define tumor subtypes by characteristic patterns of aberrations. These patterns might be used for classification of new tumor samples.

In the following we propose a statistical approach for identifying significantly differentially expressed regions on the chromosomes based on a regularized t-statistic (see section 2.2). We address the problem of interpolating the scores between genes by applying kernel functions (see section 2.3). In order to evaluate the significance of scores we conduct permutation experiments (section 2.4). An integral part of the project is the visualization of the results in order to provide convenient access to the statistical analysis without requiring a profound mathematical background (section 2.5).

The package is implemented in the R statistical programming language (*www.r-project.org*). It requires functionalities provided by the Bioconductor package [4], which is a collection of R-libraries dealing with various aspects of the analysis of biological data including normalization, assessing background information on genes, and visualization of data.
We apply our method on a publicly available data set of acute lymphoblastic leukemia (ALL), described in [2]. This data set consists of 327 tumor samples subdivided into 10 classes. In order to investigate chromosomal aberrations within one tumor class we consider the expression levels of this class versus all the other subclasses.

2 Methods

2.1 Data Preprocessing

We assume normalized expression data, which can be provided already as a matrix in R or in form of a delimited text file. In the preprocessing step the expression data is integrated with gene location data for the given microarray into one common data format. Herefore *macat* provides the `preprocessedLoader` or its convenience-wrapper `buildMACAT`, which employ various Bioconductor [4] functions.
The resulting data format is a list containing:

- Gene identifier
- Gene location (chromosome, strand, coordinate)
- Sample labels, denoting for instance tumor (sub)types
- Expression levels as a matrix

- An identifier for the type of microarray, used in the experiments

Currently *macat* is limited to commercial Affymetrix$^{®}$ oligonucleotide microarrays.

2.2 Scoring of Differential Gene Expression

For each gene we compute a statistic denoting the degree of differential expression between two given groups of samples.In the context of the leukemia data set the two groups of samples are given by one tumor class in the first group and the remaining nine classes forming the second. The statistic is the regularized t-score introduced in [9]. This so-called *relative difference* is defined as

$$d(i) = \frac{\overline{x}_A(i) - \overline{x}_B(i)}{s(i) + s_0}$$

where $\overline{x}_A(i)$, $\overline{x}_B(i)$ are the mean expression levels of gene i in group A and B respectively, $s(i)$ is the pooled variance of the expression values of gene i, and s_0 is constant for all genes. In essence, $d(i)$ is Student's t-statistic augmented by a fudge factor s_0 in the denominator, which is supposed to prevent a high statistic for genes with a very low variance $s(i)$. We chose s_0 to be the median over all gene variances $s(i)$, analogous to [3].
Since the null-distribution of these regularized t-scores is conceptually different from a known t-distribution, we have to approximate it by random permutations (for details see section 2.4).
However, *macat* also allows to compute Student's classical t-statistic for each gene and assess significance based on the underlying t-distribution.

2.3 Smoothing Kernels

The distribution of genes measured on the array over the whole length of the chromosome is not uniform. Since we want to compute differential expression statistics for larger chromosomal regions we need a method to interpolate between the measured values. Several different approaches will be discussed here:

- k-Nearest Neighbor: For every chromosomal coordinate compute the average of the k nearest genes.
- Radial basis function (rbf): For every chromosomal coordinate compute the average over all genes weighted by distance as explained in detail below.
- Base-Pair-Distance Kernel: Similar to the k-Nearest-Neighbors, but using this kernel the average is taken over all genes within a certain radius of the position, whose value has to be determined.

All of the approaches presented above can be expressed as weighted sums of expression values. This allows us to use a matrix multiplication as basic framework for the computations where we multiply our expression matrix with the so called kernel matrix.

Let E be the expression data with *ngenes* genes in the rows and *nsamples* samples in the columns and K the kernel matrix. Further assume that we want to interpolate the smoothed expression values at *nsteps* genomic locations that are stored in a vector we call *steps*. So the kernel matrix has the dimensions (*ngenes* x *nsteps*). One entry of $K(gene, step)$ represents the weight that the gene *gene* has at the location *step*. So the product

$$E^T K = S$$

gives the smoothed matrix S of dimension (*nsamples* x *nsteps*), where one entry of $S(sample, step)$ represents the smoothed interpolation of all expression values of the *sample* at the location $steps_{step}$.

The only thing that needs to be done is the computation of the kernel matrix K. The weights depend only on the location of the genes relative to the steps, for which interpolated values are to be computed.
For the three kernels listed above the kernel functions are:

- k-nearest neighbors

$$K(gene, step) = \begin{cases} 1 & \text{if gene is one of the } k \text{ nearest neighbors of step} \\ 0 & \text{else} \end{cases}$$

- radial basis function

$$K(gene, step) = \exp(-\gamma \| geneLocation_{gene} - steps_{step} \|^2)$$

- base-pair distance

$$K(gene, step) = \begin{cases} 1 & \text{if } geneLocation_{gene} \text{ is within } distance \text{ of step} \\ 0 & \text{else} \end{cases}$$

The free parameters of the kernels determine the degree of smoothing. Take for instance the kNN kernel: the smaller k gets the fewer genes are averaged and in the extreme case interpolations take only the value of the next gene ($k = 1$). In the case of the base-pair distance kernel with a very small distance, you will see spikes at the locations, where genes are located, and zero elsewhere. For very large distances the smoothing will remove all spikes and the value for each position is roughly the same. This shows that a good choice of the kernel parameters is very important. For the three kernel functions presented here we fit the parameters from the data:

- kNN: the number of genes on the chromosome is determined and k is set to cover approximately 10% of the genes.

- rbf: the width of the kernel is controlled by the parameter gamma. We demand that for each position where we interpolate there should be at least two neighboring genes that both have a weight of 1/2 because this yields a simple average between these. To guarantee weights of one half the maximum gene distance (*max*) between the genes on the chip has to be determined. Then gamma can be computed as:
$$\gamma = \frac{\ln 2}{(max/2)^2}$$
- base-pair distance: as with the rbf kernel we demand that there are at least two genes within the radius that form the average. So the distance is set to *max*.

2.4 Statistical Evaluation

To judge the significance of differential gene expression, we propose to investigate random permutations of the class labels. To obtain a reliable simulation of the empirical distribution, we suggest to choose at least $B \geq 1000$ permutations, preferably more. For each of these permutations the (regularized) t-statistic is recomputed for each gene. Thus, for each gene we obtain B permutation statistics and consequently an *empirical p-value*, denoting the proportion of the B permutation statistics being greater or equal than the actual statistic that is based on the true class labels.

Optionally, to judge the significance of differential expression over whole chromosomal regions, one could instead investigate permutations of the ordering of genes on chromosomes. For each of these permutations the smoothing of gene-specific scores is recalculated. This way, one obtains a null-distribution for scores over chromosomal regions. Given this null-distribution, it is possible to define confidence intervals for scores from random sample groupings and assess significance of scores observed in a relevant sample grouping.

However, the standard procedure in *macat* is to permute the class labels, which is considerably faster and yields results that are easier to validate and interpret statistically.

This approach is implemented in the *macat*-function `evalScoring`, which can be seen as the core function of *macat*. See appendix A for an examplary use of the function and its arguments.

One important issue is that, when analyzing many chromosomes and classes, one might obtain statistically significant results by chance (*multiple- testing problem*). In the given setting, classical procedures correcting for multiple testing, such as the *Bonferroni* correction, cannot be applied. We advise the user to be aware of the problem and to validate results by alternative methods.

2.5 Visualization

In order to facilitate a better understanding of both the data and the statistical analysis, it is helpful to employ meaningful and concise visualizations. *macat* includes plotting functionality for several questions of interest.

- Plotting raw and kernelized expression levels versus coordinates of genes on one chromosome.
- Visualize raw and kernelized statistical scores versus coordinates of genes on one chromosome.
- Emphasis of interesting chromosomal regions with listing of relevant genes.

Chromosomal regions showing significant differential expression can be visualized by plotting the result of the `evalScoring` function (see figure 1). Hereby scores for genes (black dots), the sliding average of the 0.025 and 0.975 quantiles of the permuted scores (grey lines), the sliding average permuted scores (red line), and highlighted regions (yellow dotted), where the score exceeds the quantiles, are plotted along one chromosome. The yellow regions are deemed interesting, showing significant over- or under-expression according to the underlying statistic. The plot region ranges from zero to the length of the respective chromosome.

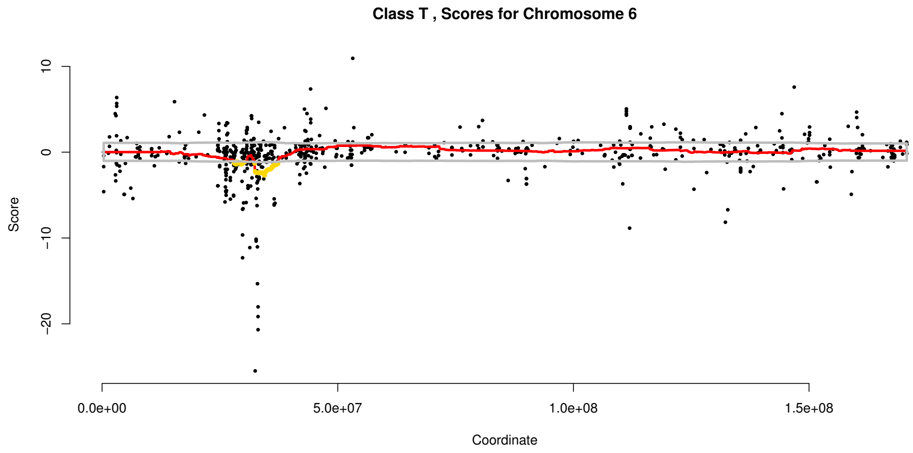
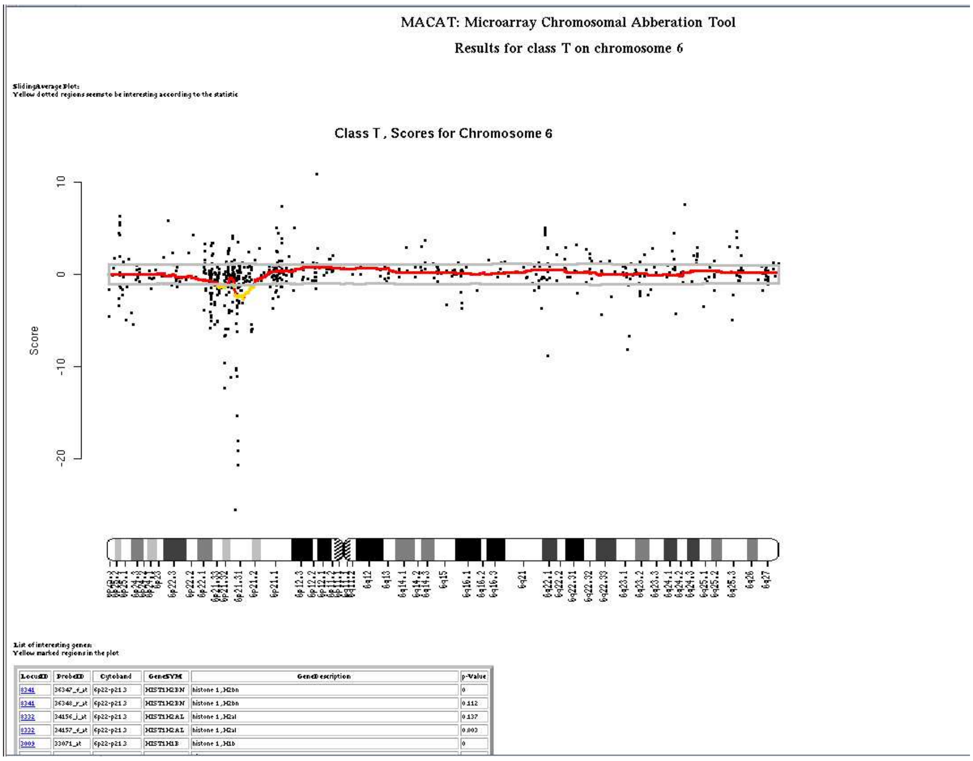


Figure 1: Example plot of plot.MACATevalScoring

One can generate an HTML-page (see figure 2) on-the-fly by setting the argument `output` in the function `plot.MACATevalScoring` to “html”. The generated HTML-page provides information about genes located within the highlighted chromosomal regions. For each gene some annotation, a click-able LocusID linked to the NCBI web site, and the empirical p-value is provided.

3 Results

This section describes the results, we obtain from an exemplary analysis on “T- vs. B-lymphocyte ALL”.
First the regularized t-score [3] is computed as described in section 2.2. To include



MACAT: Microarray Chromosomal Abberation Tool

Results for class T on chromosome 6

SlidingAverage Plot:
Yellow dotted regions seems to be interesting according to the statistic

List of interesting genes:
Yellow marked regions in the plot

LocusID	ProbeID	Cytoband	GeneSYM	GeneDescription	p-Value
8341	36347_f_at	6p22-p21.3	HIST1H2BN	histone 1, H2bn	0
8341	36348_r_at	6p22-p21.3	HIST1H2BN	histone 1, H2bn	0.112
8332	34156_i_at	6p22-p21.3	HIST1H2AL	histone 1, H2al	0.137
8332	34157_f_at	6p22-p21.3	HIST1H2AL	histone 1, H2al	0.003
3009	33071_at	6p22-p21.3	HIST1H1B	histone 1, H1b	0

Figure 2: Example for a generated HTML-page

information about the distances between the genes, we used the radial basis kernel for smoothing with the default parameters (see section 2.3). We have investigated regions of chromosome six for significant differential expression. Figure 2 shows that there is a region on the p-arm of chromosome six that is significantly under-expressed. The genes within that region comprise also the well known MHC class II genes that are known to be expressed by B-lymphocytes, but not by T-lymhocytes. Other genes in this region remain to be investigated in more detail.
Apart from this region the data reveals no other significant differences in gene expression between T- and B-lymphocyte ALL on chromosome six.

In table 1, we show a list of genes found to be located in significantly differentially expressed regions on different chromosomes when analyzing different subtypes of leukemia versus all other subtypes.

4 Discussion and Outlook

As described in the previous section, applying our method, we detected a chromosomal stretch as significantly differentially expressed, which is in agreement with biological knowledge of the two sample classes involved. This gives an indication that the chromosomal regions annotated as being significant by the method are indeed biologically meaningful. Many of the genes found by MACAT (see table 1) are known oncogenes or are at least associated with oncogenesis.

This fits well with our expectation, since it appears reasonable that different tumor subtypes can be characterized on the molecular level by different expression levels for genes relevant to oncogenesis.

One point of interest for future research would be the application of MACAT on other publicly available data sets and contrasting the results with relevant biological expert knowledge to evaluate performance. This way, one could get an clearer impression of the extend of possible applications.

Another point would be the exploration of different approaches for obtaining relative gene expression values for groups of samples. For instance one could use another data set as a reference for samples with ”normal” gene expression levels (“normal” denoting samples from patients with a tumor). Of course, due to the noisy nature of microarray data, this approach would have to overcome some structural and technical difficulties to make the measured expression levels comparable. Given that a suitable data set could be found and the compatibility issues resolved, one could examine the characteristic patterns of chromosomal aberrations in tumor subclasses as opposed to normal tissue which is a more general question than the question of differences between subclasses of the same tumor type, which we have considered in the analysis above.

The method, which we have described, can detect significant differential expression for chromosomal regions. However, the reason for the differential expression, be it

Class	**Chrom**	**Cytoband**	**LocusID**	**OMIM Annotation**
MLL	8	8q24.12	4609	”Alitalo et al. [1] found that the MYC gene, which is involved by *translocation* in the generation of *Burkitt lymphoma*, is amplified, resulting in homogeneously staining chromosomal regions (HSRs) in a human neuroendocrine tumor cell line derived from a *colon cancer*. The HSR resided on a distorted X chromosome; amplification of MYC had been accompanied by translocation of the gene from its normal position on 8q24.”
MLL	9	9p22.3	6595	”Mammalian SWI/SNF complexes are ATP-dependent chromatin remodeling enzymes that have been implicated in the regulation of gene expression, cell cycle control, and *oncogenesis*.”
MLL	9	9p24.3	23189	” The suggested role for this protein is in tumorogenesis of renal cell *carcinoma*.”
MLL	11	11p13	7490	”Mutations in this gene can be associated with the development of *Wilms tumors* in the kidney or with abnormalities of the genitourinary tract.”
E2A	1	1q23	5087	”Wiemels et al.[10] sequenced the genomic fusion between the PBX1 and E2A genes in 22 pre-B acute lymphoblastic leukemias and 2 cell lines. Kamps et al. [5] discussed the chimeric genes created by the human t(1;19) translocation in pre-B-cell acute lymphoblastic leukemias. The authors cloned 2 different E2A-PBX1 fusion transcripts and showed that NIH-3T3 cells transfected with cDNAs encoding the fusion proteins were able to cause malignant tumors in nude mice.”
E2A	9	9p24.3	23189	”By RT-PCR and Western blot analysis, Sarkar et al. [7] demonstrated that KANK expression was suppressed in most renal tumors and in kidney tumor cell lines due to methylation at CpG sites in the gene.”

Table 1: Example genes with OMIM annotation [6] detected by MACAT to be located in differentially expressed regions. ’Class’ denotes the analyzed tumor subtype.

a deletion, amplification,translocation or other chromosomal event, remains to be investigated.

Thus, results obtained by our method should be verified by suitable experiments. One could think of a customized cDNA or oligo chip that contains all known genes, including fusion-genes, stemming from translocation events, in the regions that were found to be differentially expressed. To validate the borders of found regions, it would be useful to also incorporate genes that neighbor found regions.

Other possibly useful experiments include Real-time PCR to measure the amount of mRNA for one (or few) specific genes, comparing the results to measured expression levels. The nature of the chromosomal event leading to the differential expression can be investigated by methylation array experiments or by fluorescence in situ hybridization.

Nevertheless, we have shown that our method provides a solid baseline for gene-wise experiments.

A Example Session

A.1 The Beginning

In this section, we show an example MACAT session. First of all, one has to include the library and to load the data set provided in the package.

```
> library(macat)
```

```
Loading required package: Biobase
Welcome to Bioconductor
         Vignettes contain introductory material.  To view,
         simply type: openVignette()
         For details on reading vignettes, see
         the openVignette help page.
Loading required package: annotate
Loading required package: XML
Loading required package: reposTools
Loading MicroArray Chromosomal Aberration Tools...
Loading required packages...
Type 'demo(macatdemo)' for a quick tour...
```

```
> loaddatapkg("stjudem")
```

```
Loading data set 'stjude' ...
Type 'summary(stjude)' for an overview of the data structure.
See package 'macat' for examplary analyses.
```

We now have a data object called `stjude`.

A.2 First Investigation

Let us have a closer look on the data. The data is already in the right format. It has been pre-processed by the function `preprocessedLoader`.

```
> summary(stjude)
```

```
             Length  Class  Mode
geneName       12637 -none- character
geneLocation   12637 -none- numeric
chromosome     12637 -none- character
expr         4128375 -none- numeric
labels           327 -none- character
chip               1 -none- character
```

We can for example access the first 10 gene names by typing

```
> stjude$geneName[1:10]
```

```
 [1] "34916_s_at" "34917_at"   "34462_at"   "163_at"     "163_at"
 [6] "35219_at"   "31641_s_at" "33300_at"   "33301_g_at" "38950_r_at"
```

The different labels in the data set can be accessed by

```
> unique(stjude$labels)
```

```
 [1] "BCR"         "E2A"         "Hyperdip"    "Hyperdip47" "Hypodip"
 [6] "MLL"         "Normal"      "Pseudodip"   "T"          "TEL"
```

```
> table(stjude$labels)
```

```
       BCR        E2A   Hyperdip Hyperdip47    Hypodip        MLL     Normal
        15         27         64         23          9         20         18
 Pseudodip          T        TEL
        29         43         79
```

There are ten different classes of tumor patients. The next question is how many probeIDs lie on chromosome 1.

```
> length(stjude$geneName[stjude$chromosome == 1])
```

```
[1] 1255
```

Now for some visualization. We want to plot the sliding average of the expression values from sample 3 along chromosome 6 with the default rbf-kernel (for details see section 2.5).

```
> plotSliding(stjude, 6, sample = 3, kernel = rbf)
```

See the result in figure 3.

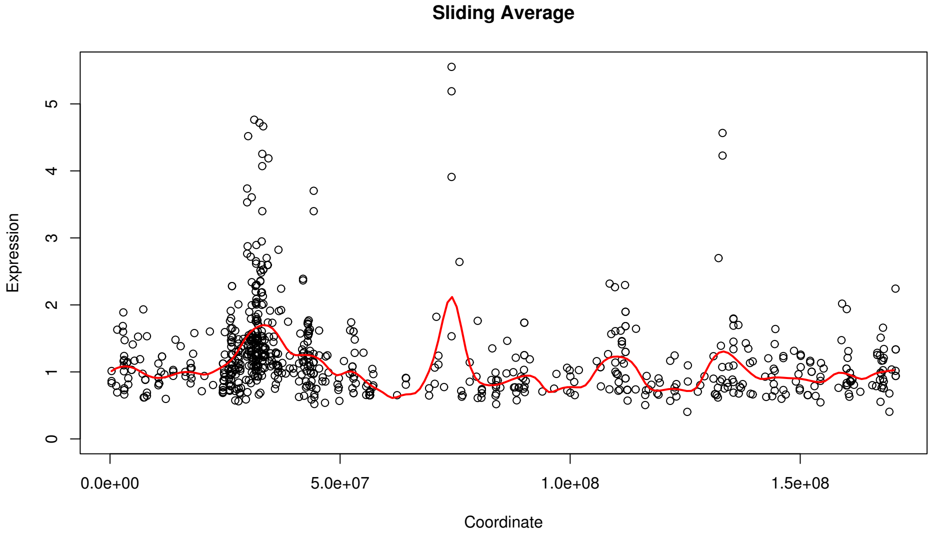


Figure 3: Sliding Average of the expression values of chromosome 6 with rbf-kernel.

A.3 Deeper Investigation

Next we investigate the data for chromosomal regions showing differential expression. We again look at chromosome 6 for some specific differences between T-lymphocyte ALL (class ”T”) and B-lymphocyte ALL (all other classes). Take a look on section 2.2 (Scoring differential expression) for details on the score. First we use the `evalScoring` function to build a *MACATevalScoring* object. This may take some time due to the number of permutations. We first use the default Gaussian(rbf) kernel for smoothing the scores.

```
> e1 = evalScoring(stjude, class="T", chromosome=6, nperms=1000)

Investigating 43 samples of class T ...
Compute observed test statistics...
Building permutation matrix...
Compute 1000 permutation test statistics...
250 ...500 ...750 ...1000 ...
Compute empirical p-values...
Compute quantiles of empirical distributions...
Computing sliding values for scores...
Compute sliding values for permutations...
All done.
```

Next, the same analysis with a k-nearest neighbor kernel.

```
> e2 = evalScoring(stjude, class="T", chromosome=6, nperms=1000,kernel=kNN)
```

We can compare the two results by using the plot function `plot.MACATevalScoring` (see figure 4).

```
> x11(width=12,height=12)
> par(mfrow=c(2,1))
> plot(e1, output="x11", new.device=F)
> plot(e2, output="x11", new.device=F)
```

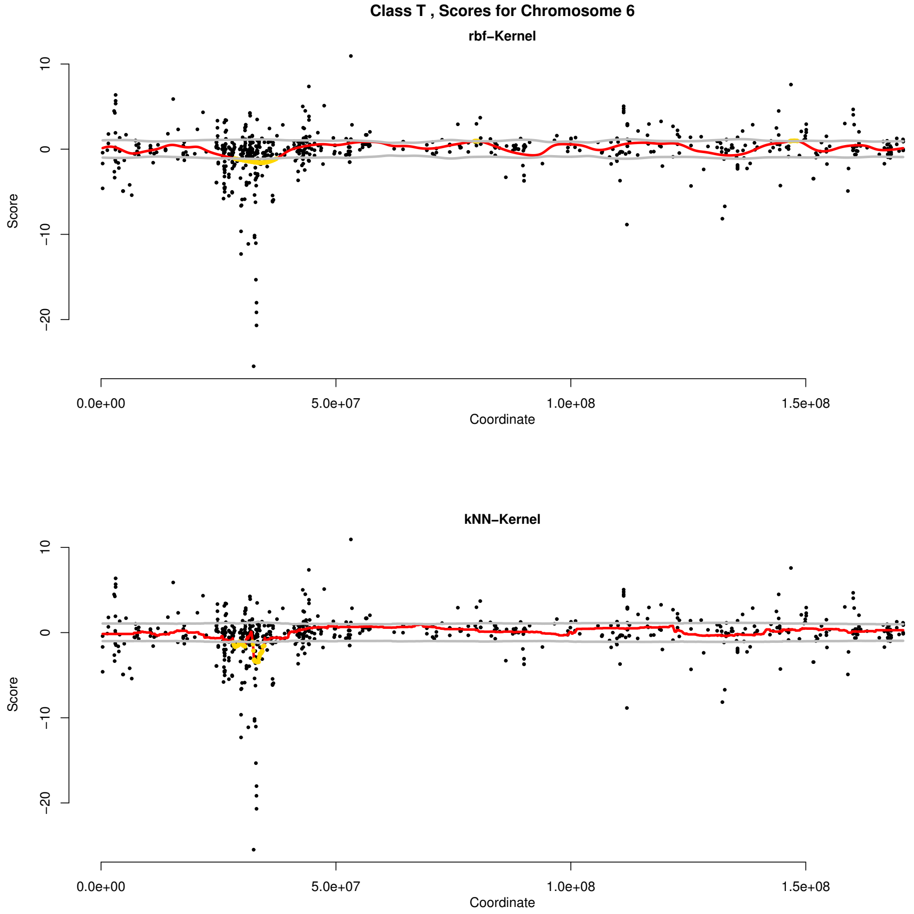


Figure 4: Comparison of two kernel functions for smoothing scores. Top: Radial basis function, Bottom: k-nearest neighbor

A.4 Collecting Results

Next, we use the plot function `plot.MACATevalScoring` to generate an HTML-page on-the-fly. Therefore, we set the parameter `output` to ”html” (see section 2.5).

```
> plot(e1, output="html")
```

See the result in Figure 2. We get some useful information about genes, which lie in the highlighted regions.

This ends our short example session. Have fun using *macat*!

References

[1] K. Alitalo, M. Schwab, C. Lin, H.E. Varmus, and J.M. Bishop. Homogeneously staining chromosomal regions contain amplified copies of an abundantly expressed cellular oncogene (c-myc) in malignant neuroendocrine cells from a human colon carcinoma. *Proc. Nat. Acad. Sci.*, 80:1707–1711, 1983.

[2] E.J. Yeoh et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1(2):133–143, March 2002.

[3] R. Tibshirani et al. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Nat. Acad. Sci.*, 99(10):6567–6572, 2002.

[4] Robert C. Gentleman, Vincent J. Carey, Douglas J. Bates, Benjamin M. Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Guenther Sawitzki, Colin Smith, Gordon K. Smyth, Luke Tierney, Yee Hwa Yang, , and Jianhua Zhang. Bioconductor: Open software development for computational biology and bioinformatics. *Bioconductor Project Working Papers. Working Paper 1.*, 2004.

[5] M. P. Kamps, A. T. Look, and D. Baltimore. The human t(1:19) translocation in pre-b all produces multiple nuclear e2a-pbx1 fusion proteins with differing transforming potentials. *Genes Dev.*, 5:358–368, 1991.

[6] V.A. McKusick. *Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders.* Johns Hopkins University Press, Baltimore, 12 edition, 1982.

[7] S. Sarkar, B. C. Roy, N. Hatano, T. Aoyagi, K. Gohji, and R. Kiyama. A novel ankyrin repeat-containing gene (kank) located at 9p24 is a growth suppressor of renal cell carcinoma. *J. Biol. Chem.*, 277:36585–36591, 2002.

[8] A. Schliep, A. Schonhuth, and C. Steinhoff. Using hidden markov models to analyze gene expression time course data. *Bioinformatics*, 19 Suppl 1:i255–63, 2003.

[9] V.G. Tusher, R. Tisbhirani, and G. Chu. Significance analysis of microarrays applied to ionizing radiation response. *Proc. Nat. Acad. Sci.*, 98(9):5116–5121, April 2001.

[10] J. L. Wiemels, B. C. Leonard, Y. Wang, M. R. Segal, S. P. Hunger, M. T. Smith, V. Crouse, X. Ma, P. A. Buffler, and S. R. Pine. Site-specific translocation and evidence of postnatal origin of the t(1;19) e2a-pbx1 fusion in childhood acute lymphoblastic leukemia. *Proc. Nat. Acad. Sci.*, 99:15101–15106, 2002.