Ludwig-Maximilians-Universtät München
Institut für Informatik
Chair Practical Informatics and Bioinformatics
Prof. Dr. Ralf Zimmer
Constantin Ammar (1)
Evi Berchtold (3)
Alexander Grün (2)

Summer term 2017
pearls@bio.ifi.lmu.de

Exercises for the Lecture Bioinformatics Pearls: ENCODE and The Regulatory Genome

Assignment 6
Hand-in: Tuesday, 8.06.2017, 10 h

Save your solution to

```
${S6} = /home/proj/biocluster/praktikum/neap_pearl/${stud_account}/Solution6.
```

Also provide your implementation in this directory that allows to reproduce your results. Your program should print a usage info if invoked without parameters. If you need additional input files, please store them in your solution folder. Please explain your solutions with appropriate descriptions.

Task 1: MS2 Spectrum Analysis (10P)

Consider the MaxQuant **msms.txt* files that are deposited in

```
/home/proj/biosoft/praktikum/neap_pearl/assignments/a6/proteomics
```

The first part of the filename indicates the mass spec instrument used to measure the data. These files list all MS2 spectra that were assigned to a peptide sequence by MaxQuant, also called the *peptide spectrum matches* (PSMs). Each line in the file corresponds to one PSM and contains a variety of parameters extracted by MaxQuant. If you want to read about the individual parameters you can consider the *tables.pdf* file, which is also in the folder. For all plots considered, make one plot for each instrument type.

(a) **PSM Statistics:** When looking into the files, you will notice that the same peptides (same sequence!) are often identified multiple times. Plot a histogram of the number of PSMs for all peptides.

(b) **Andromeda Scores:** Plot (b1) the distribution of Andromeda scores (*Score* header in the msms.txt file) for peptides with $>$ 10 PSMs. Also plot (box-plots) the Andromeda score against peptide length (b2) and against number of PSMs (b3).

(c) **Spectrum Intensities:** Consider the *Matches* and *Intensities* columns. For each ion match (e.g. b- or y- ions), the corresponding intensity is listed. Extract the (ion, intensity) pairs from these two columns and plot the summed intensity for each PSM against the Andromeda score.

(d) **Mass Spectra:** Choose your favourite PSM from each msms.txt file. Give a quick reason why it's your favourite. Make a plot of this spectrum.

(e) **Ion intensities:** Choose your favourite ion. Give a quick reason why it's your favourite. Display the distribution of intensities and variances (box plots) against the number of PSMs with that ion.

Task 2: 3D DNA Interactions (10P)

Download the Hi-C dataset from the 4DGenome Database

`https://4dgenome.research.chop.edu/Download.html`.

Keep only contacts that are found using Hi-C and originate from the human genome (hg19).

(a) show distributions of

- contact counts for all chromosomes
- distance between the interactors for all contacts

(b) create a tool that generates contact-maps (heatmaps) for a certain region on a specified chromosome. The heatmap should be created by binning the genome and counting contacts between and within the respective bins. Your tool should take the following input parameters:

- input file
- number of bins
- region: 'chr:START-END' (e.g: chr2:10000-50000)

(c) create a GUI or webpage for the tool you created in (b). The user should be able to specify the parameters and the tool must display the resulting heatmap.

Task 3: Disease classification (10P)

Use the 'genefu' R package and load the NKI dataset (available from bioconductor). Make predictions for all patients in the NKI cohort using the following risk scores/subtype classifiers:

- PAM50
- SCMGENE
- EndoPredict
- OncotypeDX

Show one Kaplan-Meier plot for each classifier/risk score considering overall survival. For the risk scores, use the predefined cutoffs and limit the patient cohort to ER positive patients. Also provide an overview of the hazard ratios and log rank test p-values for all classifiers/risk scores.

Additionally, calculate contingency matrices (remember the tool you made the other day (Task 2)!) between the two classifiers and the two risk scores, as well as between the classifiers and the ER/PR status.