A introduction on NuCMap R package

Liqun Xi, Qingyang Zhang, Kristin Brogaard,
Bruce Lindsay, Jonathan Widom, Ji-Ping Wang

September 16, 2012

1 About NuCMap

NuCMap is an R package for analyzing **Nu**cleosome mapping data from the newly developed **C**hemical **Map**ping technology. *NuCMap* depends on the experiment data package *nucmapData*, which provides the published data in Brogaard et al. (2012) for illustration. This package is built upon a locally convoluted Poisson model proposed in Brogaard et al. (2012) and a generalized EM algorithm by Xi et al. (2012). The core of the package was written in Fortran and C++. *NuCMap* integrates eleven functions including `plotCUTS`, `peakDIST`, `trainTEMP1`, `estNCP1`, `callUNIQUE`, `callRED`, `trainTEMP4`, `estNCP4`, `plotAATT`, `estNCPcall`, `calOccup` to fulfill a complete analysis of chemical mapping data from data visualization, model training and diagnostic, parameter estimation to nucleosome map definition.

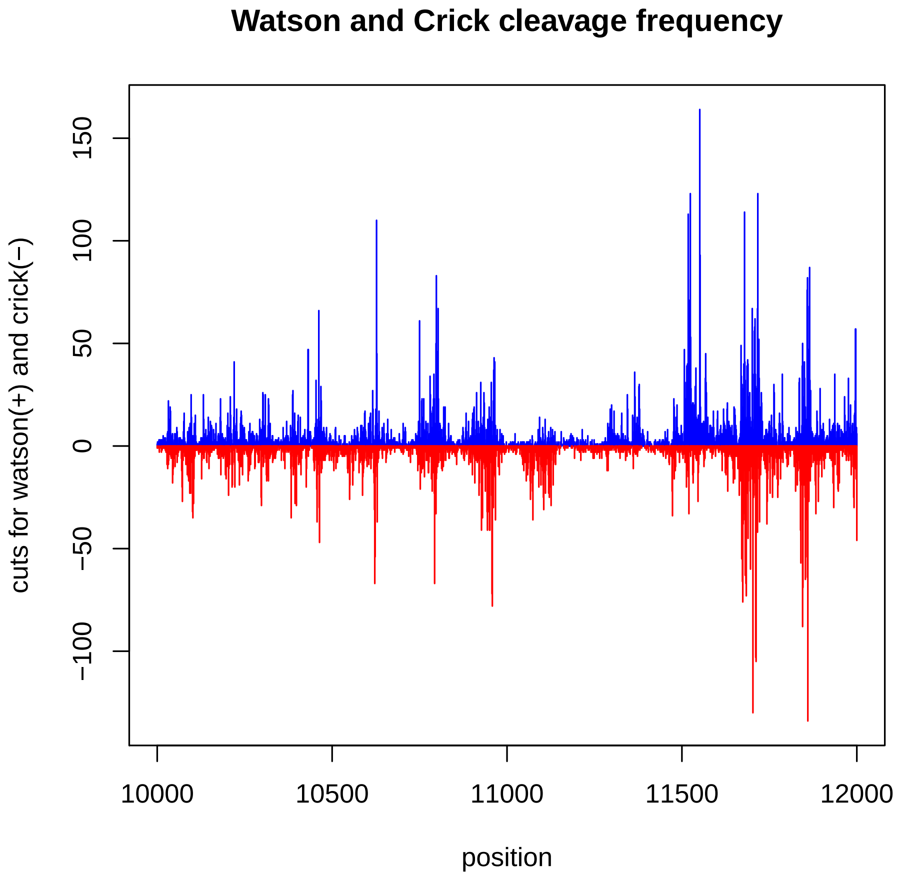
2 NuCMap functions

2.1 Raw data visualization

```
> library(NuCMap)
> library(nucmapData)
```

Function `plotCUTS` visualizes the cleavage frequency in any specified region on Watson and Crick strands simultaneously.

```
> wfile=system.file("extdata", "watson12.txt",package="nucmapData")
> cfile=system.file("extdata", "crick12.txt",package="nucmapData")
> plotCUTS(seqname="chrI",watsonfile=wfile,crickfile=cfile,startpos=10000,endpos=12000)
```



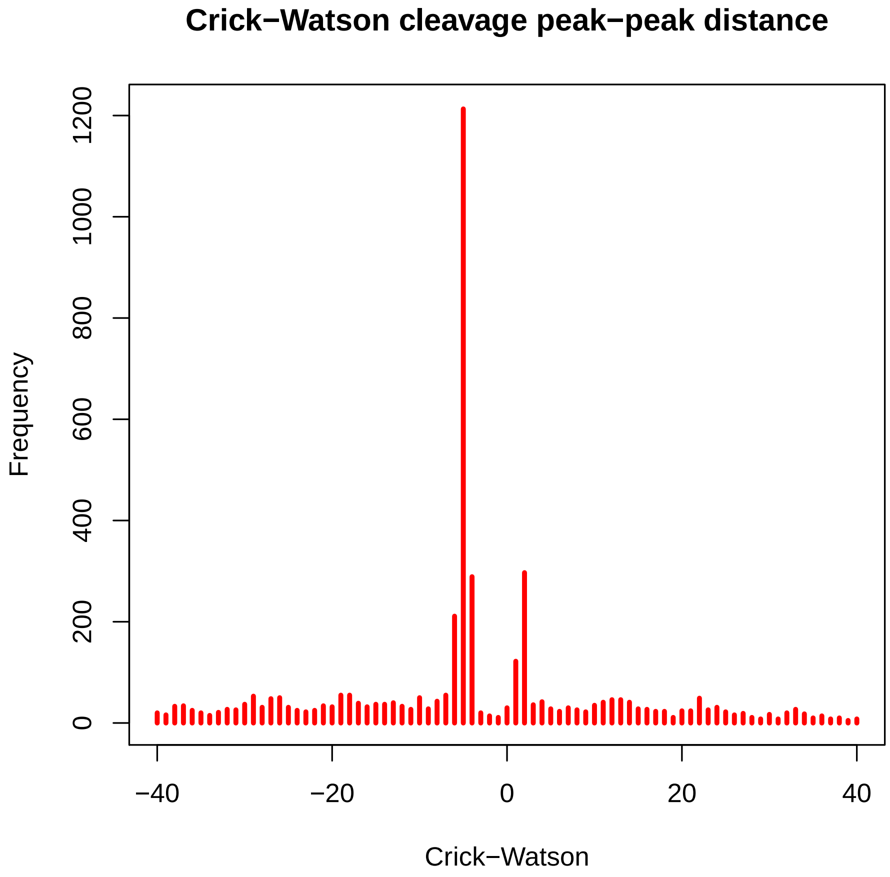
Here we have two raw files named "*watson*12*.txt*" and "*crick*12*.txt*" (system data in the experiment data package), which contain the cleavage frequency on chromosomes I and II from the published data in Brogaard et al. (2012). *NuCMap* requires that each raw data file contain three columns, the first of which is the chromosome name, the second is the chromosome coordinate and the last is the number of cleavages observed at each coordinate on the Watson strand or the Crick strand. The argument *seqname* must be the name of one particular chromosome, e.g. "*chrI*", which should be consistent with the chromosome's name in the raw cleavage files as specified in *watsonfile* and *crickfile*. Note that the arguments *watsonfile* and *crickfile* must be specified as the file names of the raw data attached with the complete path. If only the file names are specified without the path, R will search the working directory for the specified raw data files. The same requirement applies to all functions below that involve these two arguments. The region for visualization is specified by the starting position argument *startpos* and end position argument *endpos*. For example, if "*watson.txt*" and "*crick.txt*" are located in the directory "*C* : /*Users*/*jon*/*DNA*", we can plot the cleavages on chromosome I from coordinate 10000 to 12000 as follows:

```
>wfile="C:/Users/jon/DNA/watson.txt"

>cfile="C:/Users/jon/DNA/crick.txt"

>plotCUTS(seqname="chrI",watsonfile=wfile,crickfile=cfile,startpos=10000,endpos=12000)
```

2.2 Cross-strand cleavage peak-peak distance

As a check on the primary-secondary sites configuration (see Brogaard et al. (2012)), function `peakDIST` identifies the local maximum cleavage peaks on both strands, and calculates the frequency of cross-strand peak-peak distance. Typically the input data is the cleavages frequency from the entire chromosomes (though user can also specify individual chromosomes for this calculation).

```
> wfile=system.file("extdata","watson12.txt",package="nucmapData")
> cfile=system.file("extdata","crick12.txt",package="nucmapData")
> peakDIST(seqname=c("chrI","chrII"),watsonfile=wfile,crickfile=cfile)
```



The cleavage file may contain more than two chromosomes. The default value for the argument *seqname* in `peakDIST` is ”all”, which results in a plot using all chromosomes listed in the raw data files. One can also use *seqname* to specify one or more individual chromosomes for this plot. Those rules apply to all functions that carries the *seqname* argument except function `plotCUTS` (in which *seqname* must be specified as one individual chromosome name). For example, the codes above generate a plot only based on chromosome I and II, showing the two peaks at +2 and −5 nt distance, confirming the primary + primary and primary + secondary cleavage site combinations (Brogaard et al. (2012)).

2.3 Template training and NCP score calculation

Function `trainTEMP1` iteratively trains a single-template model to characterize the average cleavage frequency at eight positions including (-2, -1, 0, 1, 4, 5, 6, 7) around nucleosome center. It carries the identical arguments as function `peakDIST`. It outputs the template results into a file named "*trainTEMP1result.txt*" under the current working directory.

```
> wfile=system.file("extdata","watson12.txt",package="nucmapData")
> cfile=system.file("extdata","crick12.txt",package="nucmapData")
> trainTEMP1(seqname=c("chrI","chrII"),watsonfile=wfile,crickfile=cfile)
```

Function `estNCP1` invokes Fortran codes to estimate nucleosome center positioning (NCP) score and NCP score/noise ratio using one-template model. Call `estNCP1` as follows:

```
> chrI=system.file("extdata", "chrI.fa",package="nucmapData")
> wfile=system.file("extdata", "watson12.txt",package="nucmapData")
> cfile=system.file("extdata", "crick12.txt",package="nucmapData")
> estNCP1(seqname="chrI",genfile=chrI,watsonfile=wfile,crickfile=cfile,temp1="default")
```

Function `estNCP1` requires the input of chromosome sequences, the name and path of which are sepcified by the argument *genfile*. If the path is not specified, the R will search the working directory for the chromosome sequences. The chromosome files must be in standard FASTA format with line length $\leq$ 400. The names of the chromosomes specified by *genfile* must be consistent to as specified by argument *seqname*. The argument *temp1* specifies the template to be used to fit the Poisson cluster model (Xi et al. (2012)). The default choice (by setting *temp1* = "*default*") is the template trained based on the yeast chemical mapping data from function `trainTEMP1` as input by setting *temp1* = "*trainTEMP1result.txt*" (assuming this file resides under the working directory) for this calculation.

The output file, named "*NCPscore.ratio_1temp.txt*", is automatically saved under the working directory. It contains five columns:

1. `chr.`: chromosome name;
2. `Position`: chromosome coordinate;
3. `NCPscore`: estimated NCP score;
4. `Ratio`: NCP score/noise ratio;
5. `cNCPscore`: NCP score after correction for strand asymmetry of cleavages. It is used for nucleosome occupancy calculation

```
> result=read.table(system.file("extdata", "NCPscore.ratio_1temp.txt",package="nucmapData"),header=T
> result[1:10,]
```

```
   chro Position NCPscore   Ratio cNCPscore
1  chrI        4    0.002 0.00114      0.00
2  chrI        5    0.539 0.30366      0.54
3  chrI        6    0.400 0.22534      0.40
4  chrI        9    0.103 0.05783      0.10
5  chrI       10    0.094 0.05315      0.09
```

```
6  chrI       13    1.319 0.74366      1.32
7  chrI       27    0.064 0.03632      0.06
8  chrI       28    0.156 0.08795      0.16
9  chrI       31    0.321 0.18105      0.32
10 chrI       36    0.071 0.03998      0.07
```

Function `callUNIQUE` defines a unique nucleosome map based on the estimated NCP score from function `estNCP1` or `estNCP4` (see Brogaard et al. (2012)):

```
> NCP1=system.file("extdata","NCPscore.ratio_1temp.txt",package="nucmapData")
> callUNIQUE(estresults=NCP1,seqname=c("chrI","chrII"))
```

The argument *estresults* specifies the file name of the output file from function `estNCP1` or `estNCP4`. The defined unique map will be saved into a file named "*UNIQUEcenters.txt*".

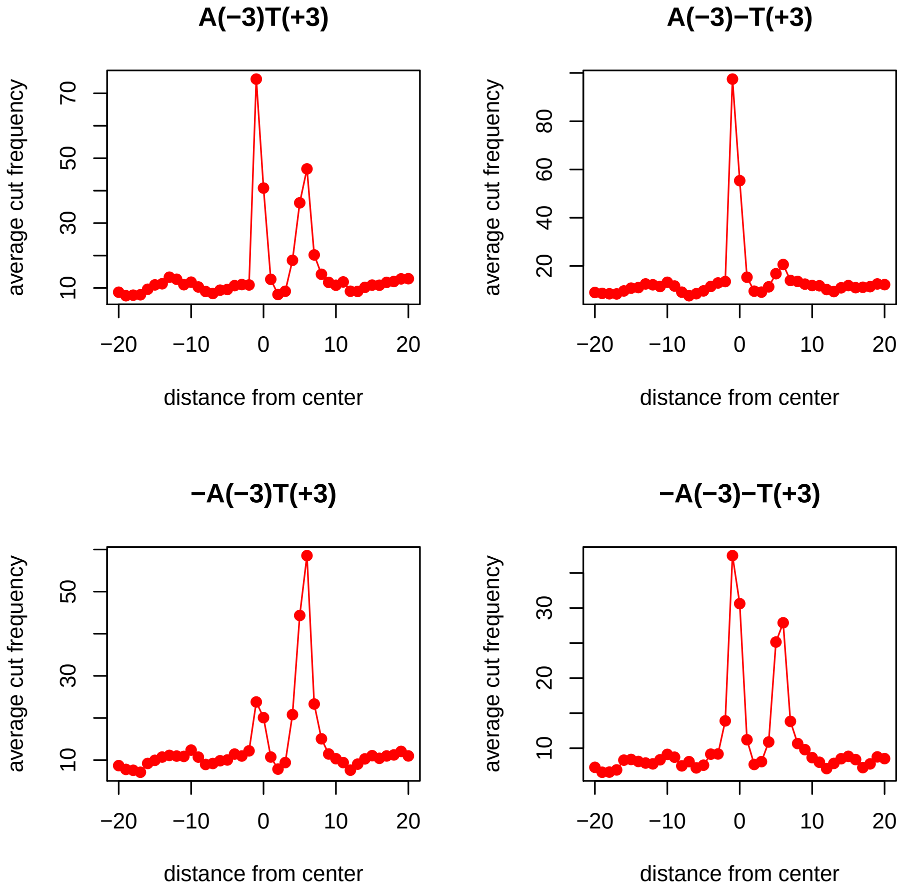
Function `callRED` defines a redundant nucleosome map based on the estimated NCP scores from function `estNCP1` or `estNCP4` (see Brogaard et al. (2012)) by thresholding the NCP score/noise ratio. All the positions where the NCP score/noise ratio exceeds the cutoff specified by *threshold* are defined as nucleosome centers. The default threshold value is the minimum NCP score/noise ratio from unique nucleosome map used in Brogaard et al. (2012).

```
> NCP1=system.file("extdata","NCPscore.ratio_1temp.txt",package="nucmapData")
> callRED(estresults=NCP1,seqname=c("chrI","chrII"),threshold="default")
```

The defined redundant map will be saved into a file named "*REDcenters.txt*" under the working directory.

Function `trainTEMP4` identifies possible cleavage bias related to the base composition around primary and secondary sites. In particular, the presence or absence of a nucleotide A or T at position -3 and $+3$ respectively results in four distinct cleavage patterns (Brogaard et al. (2012)). Based on the unique maps from the single-template model, this function plots the average cleavage frequency at positions around nucleosome centers from the following four categories of nucleosomes: $+A(-3)+T(+3)$, $+A(-3)-T(+3)$, $-A(-3)+T(+3)$, and $-A(-3)-T(+3)$, where a $+/-$ in front of the A/T nucleotide strands for presence/absence of the letter at $-3/+3$ position respectively relative to the nucleosome center. The average cleavage frequencies at clustered positions including $(-2,-1,0,1,4,5,6,7)$ for the four categories, referred to as four templates, are outputed into a file named $trainTEMP4result.txt$, with each row representing a template.

```
> wfile=system.file("extdata","watson12.txt",package="nucmapData")
> cfile=system.file("extdata","crick12.txt",package="nucmapData")
> chrI=system.file("extdata","chrI.fa",package="nucmapData")
> chrII=system.file("extdata","chrII.fa",package="nucmapData")
> umap=system.file("extdata","UNIQUEcenters.txt",package="nucmapData")
> trainTEMP4(seqname=c("chrI","chrII"),genfile=c(chrI,chrII),watsonfile=wfile,crickfile=cfile,center
```



Function `estNCP4` calculates the NCP score and NCP score/noise ratio using four-template model.

```
> chrI=system.file("extdata", "chrI.fa",package="nucmapData")
> wfile=system.file("extdata", "watson12.txt",package="nucmapData")
> cfile=system.file("extdata", "crick12.txt",package="nucmapData")
> temp4_file=system.file("extdata", "trainTEMP4result.txt",package="nucmapData")
> estNCP4(seqname="chrI",genfile=chrI,watsonfile=wfile,crickfile=cfile,temp4=temp4_file)
```

The argument *temp4* specifies the four-template model to be used in estimation. The default (by setting *temp4* = ”*default*”) is the model trained from yeast chemical mapping data (Brogaard et al. (2012) and Xi et al. (2012)). One can also specify their own template model trained based on the current data (e.g., output from function `trainTEMP4`) as exemplified by the sample codes. This function generates an output file named ”*NCPscore.ratio_4temp.txt*”, containing same five columns as in the `estNCP1` function.

To simplify the steps of calculation, *NuCMap* also provides a function `estNCPcall`, which realizes the four-template model training, NCP score estimation and unique map definition.

```
> chrI=system.file("extdata", "chrI.fa",package="nucmapData")
> wfile=system.file("extdata", "watson12.txt",package="nucmapData")
> cfile=system.file("extdata", "crick12.txt",package="nucmapData")
> estNCPcall(seqname="chrI",genfile=chrI,watsonfile=wfile,crickfile=cfile)
```

3 Other functions

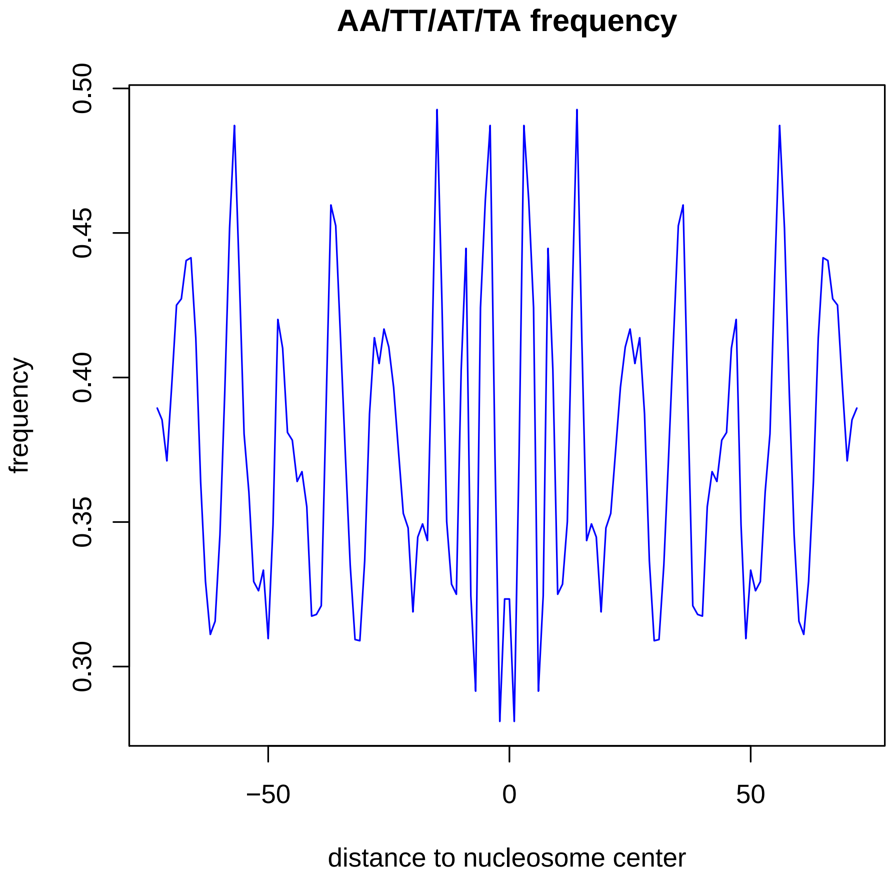
NuCMap provides two additional functions for post-estimation analysis. Function calOccup calculates nucleosome occupancy score genome-wide.

```
> NCP4=system.file("extdata","NCPscore.ratio_4temp.txt",package="nucmapData")
> chrI=system.file("extdata","chrI.fa",package="nucmapData")
> rmap=system.file("extdata","REDcenters.txt",package="nucmapData")
> calOccup(estresults=NCP4,seqname="chrI",genfile=chrI,rednufile=rmap)
```

This function requires a redundant map that defines all the possible nucleosomes to construct the occupancy (specified by argument *rednufile*). The redundant map generated from function callRED is defined by the threshold value, which controls the noise. Thus the user can define the occupancy curve based on his/her own choice of threshold value in the redundant map. This function automatically saves the output into a file named "*NuOccupancy.txt*".

Funtion plotAATT plots the AA/TT/AT/TA dinucleotide frequency as a function of position within the nucleosome region. Nucleosomal DNA is known for the enrichment of AA/TT/AT/TA signal at periodical locations. This plot provides a check of the validity of the defined nucleosome map. The argument *center* is the unique or redundant nucleosome map from function callUNIQUE or callRED.

```
> chrI=system.file("extdata", "chrI.fa",package="nucmapData")
> chrII=system.file("extdata", "chrII.fa",package="nucmapData")
> umap=system.file("extdata", "UNIQUEcenters.txt",package="nucmapData")
> plotAATT(seqname=c("chrI","chrII"),genfile=c(chrI,chrII),center=umap)
```



References

Brogaard, K., Xi, L., Wang, J.-P. and Widom, J. (2012). A base pair resolution map of nucleosome positions in yeast. *Nature 486:496-501*

Xi, L., Brogaard, K., Zhang, Q., Lindsay, B., Widom, J., Wang, J.-P. (2012). A locally convoluted Poisson cluster model for nucleosome positioning signals in chemical map. Submitted for publication.