Getting Started with Centipede

There is a help document associated with most of the functions in the package (you can install the package from centipede.uchicago.edu). Here are some rough instructions though.

The simplest input to centipede is a matrix of read counts around motif matches. For getting to this, we use an in-house set of tools for

1) finding the genomic locations of motif matches we want to use as input
2) filtering for locations with mappability issues
3) extracting all the DNase cut-sites for + and - strands around the remaining motif matches (we use 100 bp on either edge of the motif).

I think the tomtom package has a tool for doing #1 http://meme.sdsc.edu/meme4_5_0/cgi-bin/fimo.cgi). The tools we use to do this would need some work to bring to a state where someone could use it without some time investment.

Once you have candidate motif matches, you need to build a matrix where each row is a candidate motif match, and each column is a position relative to the center of the motif match. The values in this matrix will be number of read start-sites that occur at that position. We simply concatenate the forward and reverse strands together for the purpose of model fitting.

An example of this matrix is supplied as part of the centipede package. To have a look at the format, you can try this in R:

```
library(CENTIPEDE)
data(NRSFcuts, package='CENTIPEDE')
head(NRSFcuts)
```

There is a help document associated with the fitCentipede function. To see this do:

```
?fitCentipede
```

At the bottom, you will see an example of sequence of commands to run to fit centipede on the NRSFcuts object.

Namely,

```
#GETS EXAMPLE DATA FOR NRSF
   data(NRSFcuts, package='CENTIPEDE')
   data(NRSF_Anno, package='CENTIPEDE')

   #FITS THE CENTIPEDE MODEL
   centFit <- fitCentipede(Xlist = list(DNase=as.matrix(NRSFcuts)
```

```
), Y=cbind(rep(1, dim(NRSF_Anno)[1]), NRSF_Anno[,5], NRSF_Anno[,6]))

    #PLOTS IMAGE OF CUTSITES RANKED BY CENTIPEDE POSTERIORS
    imageCutSites(NRSFcuts[order(centFit$PostPr),][c(1:100, (dim(NRSFcuts)[1]-
100):(dim(NRSFcuts)[1])),])

    #PLOT ESTIMATED FOOTPRINT
    plotProfile(centFit$LambdaParList[[1]],Mlen=21)
```

Now, all the output of the centipede run (e.g. Posterior probability, log likelihood ratios, etc.) will be in exactly the same order as your original input matrix (equivalent to NRSFcuts). So, to look at the posterior probabilities of every candidate motif match in your original file, you would just use the object:

```
centFit$PostPr
```

These could be written to a text file which you could merge with your original bed file defining the locations of candidate motif matches.