A introduction on NuPoP R package

Ji-Ping Wang*, Liqun Xi
Department of Statistics, Northwestern University

October 29, 2019

1 About NuPoP

NuPoP is an R package for **Nu**cleosome **Po**sitioning **P**rediction. *NuPoP* is built upon a duration hidden Markov model, in which the linker DNA length is explicitly modeled. The nucleosome or linker DNA state model can be chosen as either a fourth order or first order Markov chain. *NuPoP* outputs the Viterbi prediction of optimal nucleosome position map, nucleosome occupancy score (from backward and forward algorithms) and nucleosome affinity score (Xi et al. (2010),Wang et al. (2008)).

In addition to the R package, we also developed a web server prediction engine and a stand-alone Fortran program available at http://nucleosome.stats.northwestern.edu. The *NuPoP* R package and the Fortran program can predict nucleosome positioning for a DNA sequence of any length.

For comments, please contact: jzwang@northwestern.edu. For technical details of *NuPoP* and the methods, please refer to Xi et al. (2010) and Wang et al. (2008).

2 NuPoP functions

NuPoP does not depend on any other R packages. It has three major functions, `predNuPoP`, `readNuPoP`, and `plotNuPoP`. The `predNuPoP` function predicts the nucleosome positioning and nucleosome occupancy, the `readNuPoP` reads in the prediction results, and the function `plotNuPoP` visualizes the predictions.

```
> library(NuPoP)
```

The `predNuPoP` calls a Fortran subroutine to process the DNA sequence and make predictions, and outputs the predictions into a text file into the current working directory. This method is based on a duration Hidden Markov model consisting of two states, one for the nucleosome and the other for the linker state. For each state, a first order Markov chain and a fourth order Markov chain models are built in. For example, a sample DNA sequence is `test.seq` located in `inst/extdata` subdirectory of the package. Call the `predNuPoP` function as follows:

```
> predNuPoP(system.file("extdata", "test.seq", package="NuPoP"),species=7,model=4)
```

*jzwang@northwestern.edu

```
Prediction output: '/tmp/RtmpkLHjaj/Rbuild27dc2b5e4552/NuPoP/vignettes/test.seq_Prediction4.txt'
```

Note that the argument *file* must be specified as the complete path and file name of the DNA sequence in FASTA format in any directory. For example, if the "test.seq" file is located in "/Users/jon/DNA", the function can be called by *>predNuPoP(file="/Users/jon/DNA/test.seq",species=7,model=4).* The user should not use *file="~/DNA/test.seq"* to speficify the path to avoid error messages. The argument *species* can be specified as follows: 1 = Human; 2 = Mouse; 3 = Rat; 4 = Zebrafish; 5 = D. melanogaster; 6 = C. elegans; 7 = S. cerevisiae; 8 = C. albicans; 9 = S. pombe; 10 = A. thaliana; 11 = Maize; 0 = other. If species = 0 is specified, the algorithm will identify a species from 1-11 that has most similar base composition to the input sequence, and use the models from the selected species for prediction. The default value is 7. The argument *model* can be either 1 or 4, standing for the order of Markov chain models used for the nucleosome and linker states.

The output file, generated in the current working directory, will be named after the sequence file name, with an added extension as *_Prediction1.txt* or *_Prediction4.txt*. For the above codes, the output file will be *test.seq_Prediction4.txt*. The four columns in the output file are

1. `Position`: position in the input DNA sequence.
2. `P-start`: probability that the current position is the start of a nucleosome.
3. `Occup`: nucleosome occupancy score. The nucleosome occupancy score is defined as the probability that the given position is covered by a nucleosome.
4. `N/L`: 1 indicates the given position is covered by nucleosome and 0 for linker linker based on Viterbi prediction.
5. `Affinity`: nucleosome binding affinity score. This affinity score is defined for every 147 bp of DNA sequence centered at the given position. Therefore for the first and last 73 bp of the DNA sequence, the affinity score is not defined (indicated as NA).

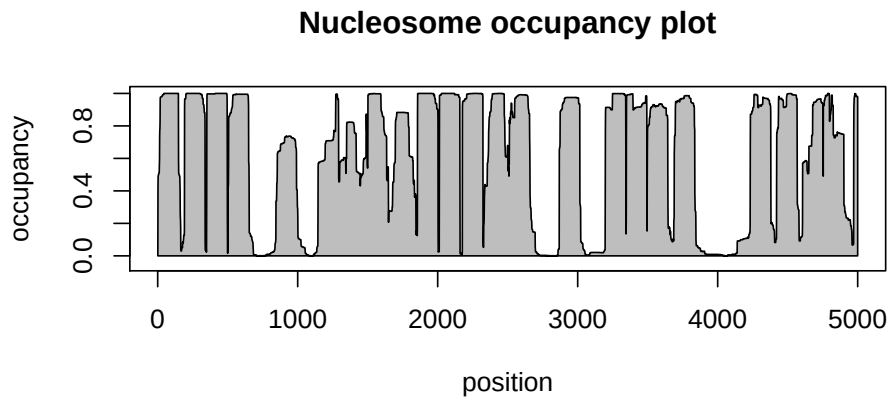
The output file can be imported into R by `readNuPoP` function:

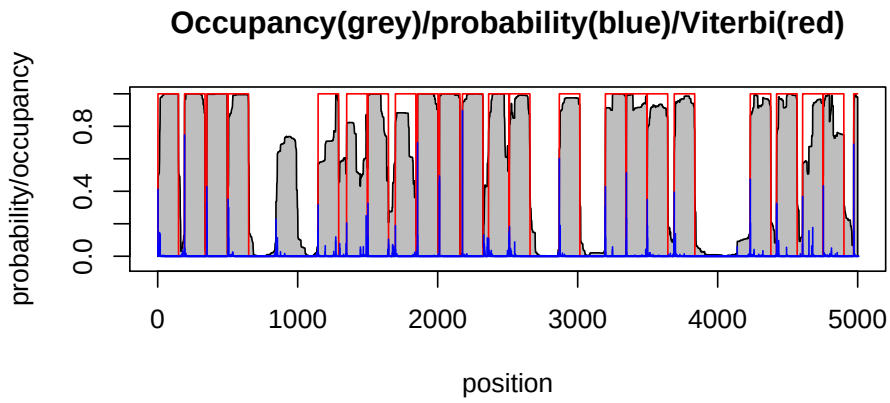
```
> results=readNuPoP("test.seq_Prediction4.txt",startPos=1,endPos=5000)
> results[1:5,]
```

```
  Position P.start Occup N/L Affinity
1        1   0.000 0.000   0       NA
2        2   0.412 0.412   1       NA
3        3   0.075 0.487   1       NA
4        4   0.005 0.492   1       NA
5        5   0.005 0.497   1       NA
```

The genomic sequence can be extremely long. The user can import a part of the predictions by specifying the start position (*startPos*) and the end position (*endPos*) in the `readNuPoP` function. For example, to visualize prediction results from *startPos=1* to *endPos=5000*,

```
> plotNuPoP(results)
```





References

Wang, J.-P., Fondufe-Mittendorf, Y., Xi, L., Tsai, G.-F., Segal, E., and Widom, J. (2008). Preferentially quantized linker DNA lengths in *Saccharomyces cerevisiae*. *PLoS Computational Biology*, 4(9):e1000175.

Xi, L., Fondufe-Mittendorf, Y., Xia, L., Flatow, J., Widom, J., and Wang, J.-P. (2010). Predicting nucleosome positioning using a duration hidden markov model. *BMC Bioinformatics*, pages doi:10.1186/1471–2105–11–346.