A MATLAB toolbox to identify RNA-protein binding sites in HITS-CLIP

November 21, 2013

`The HITS-CLIP analysis`

MATLAB functions for analysing HITS-CLIP

Description

High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation (HITS-CLIP) has been widely used to investigate genome-wide maps of RNA-protein interactions. Inspired by the underlying spatial association of the read coverage and the positional bias of mutations, we devised a two-stage model to identify RNA-protein binding sites. The model is established on all the sequencing reads (including non-clustered read sequences) to investigate binding sites at single base pair resolution.

The first round of the model employs a hidden Markov model (HMM) framework with extended nodes to identify highly read-enriched locations. The second round of the model employs an analogue of mixture models to assess the reliability of mutations and to determine binding sites on the highly enriched location. More details about the model is presented in Yun et al. (2013) This toolbox provides essential MATLAB functions to implement our model for the identification of binding sites using heterogeneous logit models via semi-supervised learning.

Details

For the computational efficiency, the HITS-CLIP data are grouped by the region length, which facilitates parallel computing in the inference of the HMM. Some functions in `http://perso.telecom-paristech.fr/cappe/Code/H2m/` are used in the toolbox. Users can specify initial values of the HMM and the mixture model as well as a cutoff for the posterior probability of being binding sites.

Author(s)

Jonghyun Yun <`jonghyun yun@utsouthwestern.edu`>

Examples

Please look at `example.m` for an example usage of the toolbox.

`an example dataset`	*Aligned read sequences obtained from the HITS-CLIP data generated by Chi et al. (2009)*

Description

The HITS-CLIP data generated by Chi et al. (2009) are included in `Chi2009` directory. The data are pre-processed to obtain aligned read sequences, and clustered and non-clustered reads are separated in two files. `clustered.txt` contains clustered read sequences, while `non-clustered.txt` contains non-clustered (isolated) reads. Seven columns in each file are tab-separated, and mutations appeared in a single read sequence are separated by white-space.

Value

`1st column`	Region ID (integer)
`2nd column`	Name of a read sequence (not to be used in the analysis)
`3rd column`	Chromosome where a read sequence resides (string)
`4th column`	Starting location of a read sequence (integer)
`5th column`	Ending location of a read sequence (integer)
`6th column`	Strand of a read sequence (- or +)
`7th column`	Genomic location of mutations found in a read sequence (integer)

References

Chi, S. W., Zang, J. B., Mele, A., et al. (2009). Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature*, 460(7254):479–86.

Yun, J., Wang, T., Wnag, X, and Xiao, G. (2013). Identification of RNA-protein binding sites in HITS-CLIP using heterogeneous logit models via semi-supervised learning.