PHM Users' Guide

Rongrong Zhang (zhan1602@purdue.edu)

April 12, 2017

```
  @article{rzhang2017phm,
    title={Inferring Spatial Organization of Individual Topologically
    Associated Domains via Piecewise Helical Model
},
    author={Zhang, Rongrong and Hu, Ming and Zhu, Yu,
      and Qin, Zhaohui and Deng, Ke and Liu, Jun S.},
    journal={IEEE/ACM Transactions on Computational
      Biology and Bioinformatics},
    note={submitted}
   }
```

1 Introduction

Piecewise helical model (PHM) is a parsimonious, easy to interpret, and robust model for inferring three-dimensional (3D) chromosomal structure from Hi-C data. PHM takes the Hi-C contact matrix and local genomic features (restriction enzyme cutting frequencies, GC content and sequence uniqueness) as input and produces, via MCMC computation, the posterior distribution of three-dimensional (3D) chromosomal structure.

In Piecewise Helical Model, we assume that chromatin within a topologically associating domain exhibits a consensus spatial organization among the cell population. Additionally, it is known from geometry that any 3D curve can be uniquely determined by its local curvature and torsion. As a special case, a constant curvature and constant torsion lead to a helical curve. Analogous to the fact that any continuous function can be approximated by a constant at each point, the curvature and torsion of an arbitrary 3D curve can be approximated by piecewise constant functions. Therefore, any continuous 3D curve can be approximated by several well-connected helixes, which we refer to as a piecewise helical curve. Based on this, we further assume that chromatin folds like a piecewise helical curve in three dimension space.

2 How to Run PHM

The full command is:
`./phm -i heatmap_filename -v covariates_filename -NP no_of_helixpieces -NG no_of_iteration -NT tune_interval -SEED seed`

`heatmap_filename`: string of characters, file name of the input Hi-C contact matrix.
`covariates_filename`: string of characters, file name of the input local genomic features.
`no_of_helixpieces`: integers, number of helixes within the piecewise helical curve. Default value is 2.
`no_of_iteration`: integers, number of Gibbs sampler iterations. Default value is 5,000.
`tune_interval`: integers, length of tune interval in HMC. Default value is 50.
`seed`: integers, seed for gsl random number generator. Default value is 1.
`output_directory`: string of characters, output directory

Example:
`./phm -i heatmap.txt -v cov.txt -NP 2 -NG 5000 -NT 50 -SEED 1`

3 Input Files

3.1 Format of input Hi-C contact matrix

Assume the genomic region of interest contains N loci. The input file of Hi-C contact matrix is a $N \times N$ symmetric matrix separated by the tab delimiter. All off-diagonal numbers should be non-negative integers. All diagonal numbers should be zero. The number in the (i, j) th cell is the total number of Hi-C reads spanning the i th locus and the j th locus.

Example:

0	197	175	154	140	147	102	122	...
197	0	210	138	124	98	84	102	...
175	210	0	348	143	110	115	130	...
154	138	348	0	176	171	202	167	...
140	124	143	176	0	448	248	153	....
147	98	110	171	448	0	303	180	...
102	84	115	202	248	303	0	243	...
122	102	130	167	153	180	243	0	...
...	...	...	...	...	...	...	...	...

3.2 Format of input local genomic features

The input file of local genomic features is a $N \times 6$ matrix separated by the table delimiter. For the i th row $(i - 1, \cdots, N)$:

Column 1: chromosome name for the i th locus.

Column 2: start position for the i th locus.

Column 3: end position for the i th locus.

Column 4: number of restriction enzyme cut fragment ends in the i th locus (positive integer).

Column 5: mean GC content in the i th locus (positive real number).

Column 6: mean mappability score in the i th locus (positive real number).

Example:

22	1.6e+07	1.7e+07	389	0.4511	0.8831
22	1.7e+07	1.8e+07	272	0.4534	0.8951
22	1.8e+07	1.9e+07	218	0.5365	0.9065
22	1.9e+07	2.0e+07	230	0.4726	0.8704
22	2.0e+07	2.1e+07	235	0.4366	0.9304
22	2.1e+07	2.2e+07	400	0.4562	0.9118
22	2.2e+07	2.3e+07	246	0.4928	0.8788
22	2.3e+07	2.4e+07	336	0.4595	0.9067
...	...	...	...	...	...

4 Output Files

`time.txt`: start, end and duration of the PHM running time.

`mode_loglike.txt`: the posterior model of the log likelihood.

`record_loglike.txt`: the log likelihood in each iteration of the Gibbs sampler (a vector with size no_of_iteration). It can be used to check the convergence of MCMC chain.

`mode_p.txt`: the posterior mode of the 3D coordinates ($N \times 3$ matrix).

`record_p.txt`: the posterior samplers of the 3D coordinates (no_of_iteration$\times 3N$ matrix). In each row, the 3D coordinates are in the order of: $x_1, y_1, z_1, x_2, y_2, z_2, \cdots, x_N, y_N, z_N$.

`mode_kappa_tau.txt`: the posterior mode of kappas and taus of helixes in the piecewise helical curve (κ, τ).

Column 1: number of helix.

Column 2: posterior mode of the curvature of the helix(κ).

Column 3: posterior mode of the torsion of the helix(τ).

`mode_tnb.txt`: the posterior mode of t, n, b vectors in Frenet framework at the change points.

`mode_nui.txt`: the posterior mode of the nuisance parameters (5×3 matrix).

Row 1: posterior mode of the scaling parameter (β_0).

Row 2: posterior mode of the association between # of Hi-C reads and spatial distance (β_1).

Row 3: posterior mode of the restriction enzyme effect (β_{enz}).

Row 4: posterior mode of the GC content effect (β_{gcc}).

Row 5: posterior mode of the mappability effect (β_{map}).

Row 6: posterior mode of the overdispersion parameter (φ).

`record_nui.txt`: the posterior samplers of the nuisance parameters (no_of_iteration$\times$6 matrix).

Column 1: posterior samples of the scaling parameter (β_0).

Column 2: posterior samples of the association between # of Hi-C reads and spatial distance (β_1).

Column 3: posterior samples of the restriction enzyme effect (β_{enz}).

Column 4: posterior samples of the GC content effect (β_{gcc}).

Column 5: posterior samples of the mappability effect (β_{map}).

Column 6: posterior samples of the overdispersion parameter (φ).

5 Contact

Comments, suggestions, questions are welcomed, and should be directed to Rongrong Zhang.

Email: zhan1602@purdue.edu