HSA User Manual

*Publications that use results obtained from this software please include a citation of the paper:

Zou, C., Zhang, Y., Ouyang, Z. (2016) HSA: integrating multi-track Hi-C data for genome-scale reconstruction of 3D chromatin structure. Genome Biology, 17:40

1. Prerequisites

1.1 HSA is written and tested under R 3.1.1. R Package 'MASS' is required. If you want a real-time visualization, install also R package 'rgl' and uncomment the line 885 and 886 in the R script 'cstruct1.R'.

1.2 Required inputs: contact maps' files, output filename and Indicator of the Markov term.

Contact map files:
Each contact map file contains one contact map. Suppose there are n loci, contact map file should contain a matrix with n rows and n+2 column, in which:
Column 1: start position of each locus
Column 2: end position of each locus
Column 3 to n+2: the n*n contact map in matrix form with all entries nonnegative.

Output file name: the entire path and name of the output file. PLEASE NOTE THIS TERM CANNOT BE OMITTED.

Indicator of the Markov term: 1 for using Markov modeling and 0 not using. We recommend it be set to 1 if the percentage of nonzero entries in contact map is lower than 10%.

1.3 Optional inputs: covariate files, initial 3D structure file

Covariate files:
Each file contains the values of the covariates for one contact map. The number of rows should be the same as its corresponding contact map file. The first two columns are the same as those in its corresponding contact map file. Columns after 2 are bias correction factors, including enzyme-cutting fragment length, GC content, and mappability score. The number of covariates can be more or less than three.
If the input contact map is already normalized, covariate file is not needed.

Initial 3D structure file: An initial 3D structure file should contain the start and end positions, as well as the 3D coordinates of its loci, thus altogether 5 columns. The resolution of the initial 3D structure can be lower or equal to the contact map to be fitted. For contact maps at high resolution (e.g., higher than 200kb or more than 1000 loci), we recommend using the reconstructed 3D structure from its lower resolution (e.g. 1Mb) counterpart as an initial structure. If the lower resolution contact map is not available, a quick substitute is a sub

contact map derived from extracting equally spaced columns and rows from the high-resolution map.

2. Run HSA

2.1 Run by Rscript command

If the inputs contain covariate files, the command line is:
Rscript myR.R contactmap_file1 … contactmap_fileK covariate_file1 … covariate_fileK outputfile Markov_Indicator Initial_structure_file (if available)

If the inputs do not contain covariate files:
Rscript myR.R contactmap_file1 … ,contactmap_fileK 0 outputfile Markov_Indicator Initial_structure_file (if available)

Parameters can be tuned in script myR.R.

2.2 Run in R console:

In R console, after importing related contact maps, covariate files, and output file's directory and name, type:
>source("yourpath/HSA/cstruct1.R")
>fmain(lsmap0,lscov0,outputfile,Maxiter,submaxiter,lamda,Leapfrog,epslon,mkfix,rho,mk,initialS)

Input annotation:

lsmap0: list(map1,…, mapK). A list containing contact maps to be modeled. Suppose there are n loci in a contact map1, then map format1 is a matrix with n rows and n+2 column, in which:
Column 1: start position of each locus
Column 2: end position of each locus
Column 3 to n+2: the n*n contact map in matrix form with all entries nonnegative.

lscov0:list(listcov1,…,listcovK) or 0 if no bias-correction covariate is need.
list(listcov1,…,listcovK), a list containing bias covariates for each contact map. e.g. Suppose there are 3 covariates, listcovk is as list(covmatk_1,..., covmat1_3),
covmatk_1~covmatk_3 are n*n matrices in which n is the number of rows in mapk.

output: output filename.

Maxiter (50~1000), submaxiter (100~150), lamda (submaxiter/4 ~ submaxiter/3), Leapfrog (20~50), epslon (0.0003~0.02): tuning parameters for maximum iteration numbers, HMC sampled length, Leapfrog step length/step size, etc., that can be set by users. The range in each bracket is the value range we use empirically for reference. The larger the Maxiter, submaxiter, and Leapfrog's values are, the longer the time it takes to run. The larger the epslon

is, the more radical the sampling process will be, and the less smooth the structure is updated. You might need to try a few epslon values to see which one fits your data. Do NOT use too large epslon. You can also use real-time visualization as mentioned above to tune these parameters and decide when to terminate.

mkfix: 0 if use default Markov coefficients and 1 if estimating from data. Default: 0

rho: ranges in [0,1) for tuning the kinetic and momentum in Hamilton dynamics. Default: 0

mk: use Markov distance control (1) or not (0).We recommend it be set to 1 if the percentage of nonzero entries in contact map is low than 10%.

initialS: Initial 3D structure if the user has one he or she wants to input. Null by default. If the input structure has the same number of loci as contact maps, this initalS can be just a 3-column coordinates. If the loci in this structure are not the same as contact map, this variable should have five columns, with the first two columns specifying the start and end locations of each locus and the last three as 3D coordinates.

Other options:

coarsefit: use a larger step size in Leapfrog at the beginning, Default=True. Note that large step size is more likely to generate outliers and NA values that cause error in computation. If that happens, try coarsefit=F.

rmoutlier: remove outliers in the output structure (T) or not (F). Default=F.

fitmode: 0 (Defualt) for a radical fitting pattern that achieves higher log-likelihood faster but is more prone to outliers and computation overflow, and 1 for a gentler yet also slower mode.

3. Interpret the results
Output includes:

outputfile.txt:
column 1: start position of each locus
column 2: end position of each locus
column 3 to 5: 3D coordinates of structure with maximum likelihood ever occurred in iteration. This optimal structure is stored and updated at each iteration thus can be checked any time during the fitting process. If you think the output is already good enough at some point, you can terminate the algorithm and do not have to wait until the end of fitting process.

outputbeta.txt:
outputbeta.txt: the estimated coefficients in the model. The coefficients are sorted as one vector, including the intercept for contact map 1, the coefficients of bias-correction covariates for contact map 1, the coefficient of power-law

distance conversion for contact map 1, ... , the intercept for contact map C, the coefficients of bias-correction covariates for contact map K, the coefficient of power-law distance conversion for contact map K.

outputtemp.txt:
column 1: start position of each locus
column 2: end position of each locus
column 3 to 5: 3D coordinates of structure sampled at latest iteration.

----------Developed and written by Chenchen Zou (chenchen.zou@jax.org)