HiCEnterprise

Documentation

Hania Kranas, Irina Tuszyńska, Bartek Wilczyński
https://github.com/hansiu/HiCEnterprise

August 21, 2019

Contents

Chapter 1

Introduction

HiCEnterprise predicts long-range interactions between regions/domains based on Hi-C maps. Establish contact!

1.1 What is in the package?

HiCEnterprise takes a list of small (1-3 bins) regions or Topological Associated Domains and searches for their significant interactions on given Hi-C maps. Package consists of two parts that perform different types of analysis: prediction of interactions between regions and domains.

1.1.1 Types of analyses

1.1.1.1 Regions

This part of the package is an implementation of the method for identification of long-range interacting regions and creating interaction profiles based on Hi-C data as introduced by (Won et al. 2016).
More in 2.

1.1.1.2 Domains

In the domain part of the package we implemented the method for prediction of long-range interactions between TADs (Topologically Associating Domains) as presented in (Niskanen et al. n.d.).
More in 3.

1.2 Dependencies

- Python version 2.7 or 3.5/6
- required python packages: numpy, scipy, statsmodels, matplotlib.
- if you want to plot through R, you need have R installed with needed packages.

1.3 Basic usage

```
HiCEnterprise type options
```

'type' can be either regions or domains. Other options are type dependent.
If you want to see the list of available arguments with short explanations type either HiCEnterprise regions −−help or HiCEnterprise domains −−help. Full specifications for each argument can be found in the documentation (2 and 3)

1.3.1 Installation

```
python setup.py install
```

You may need administrative privileges to do this.

1.3.1.1 Testing the installation

To test the scripts, please install pytest package, go to the main HiCEnterprise directory and run:

```
python setup.py test
```

1.3.2 Examples

We provided some example files (used for testing too) with which you can learn how to use the program. In order to easy download package here are available small example files that are useful for testing if the program works properly. In order to test program on the real Hi-C maps, please download example files from the HiCEnterprise web page: `http://regulomics.mimuw.edu.pl/wp/hicenterprise/` Here you can see two example runs:

1.3.2.1 Regions

```
HiCEnterprise regions regions_options
```

1.3.2.2 Domains

```
HiCEnterprise domains domains_options
```

Additional examples and specifications for each argument can be found in the documentation for a specific type (2 and 3).

Chapter 2

HiCEnterprise regions

Regions part of package can be used to identify long-range interacting regions and creating interaction profiles based on Hi-C data. It is based on the method described by (Won, 2016). The significance of the individual pairwise interaction between bins is calculated as the probability of observing a stronger contact under the fitted Weibull distribution matched by chromosome and distance. FDR (False Discovery Rate) is used for correcting for multiple testing.

2.1 Input files

To run the point interactions analysis you need:

- at least one Hi-C map in numpy format for an N chromosome (named mtx-N-N.npy)
- BED file with coordinates of regions to extract interaction profiles for (or Enhancer Atlas FASTA file - it has positions of the enhancer region in fasta header of each sequence).

All coordinates should be provided in bp. Analysis results will be more reliable if you provide multiple Hi-C maps that are biological replicates.

2.1.0.1 Example files

Example files are provided in HiCEnterprise/HiCEnterprise/tests/test_files directory in maps and enhseq.

2.2 Usage

2.2.1 Basics

Overall, the usage looks like this:

```
HiCEnterprise regions regions_options
```

To see the available options:

```
HiCEnterprise regions -h
```

2.2.1.1 Example runs

There are a few required arguments for the analysis.

- Path to folders with Hi-C maps in a numpy format:
  −−hic_folders ./map40kb_replicate1 ./map40kb_replicate2 \ldots
- File with coordinates of regions to extract interaction profiles for:
  −−region_file regions.bed
- Name of the chromosome to run the analysis for: −−chr 22
- Resolution (bin size) of Hi-C maps in bp (all maps provided should be of the same resolution): −−bin_res 40000

For information on formats of files see 2.1.
Run with only required arguments:

```
HiCEnterprise regions --hic_folders ./map40kb_replicate1
./map40kb_replicate2 --region_file ./regions.bed --chr 22
--bin_res 40000
```

2.2.1.2 Visualisation

If you want to plot the results, add –plotting argument to the command. It is possible to choose if you want to plot with matplotlib (mpl) or rpy2 and R (rpy2). Resulting figures can be found in directory provided by –figures_folder option.
Example plotting with matplotlib:

```
HiCEnterprise regions --hic_folders ./map40kb_replicate1
./map40kb_replicate2 --region_file ./regions.bed --chr 22 --bin_res
40000 --plotting mpl --figures_folder ./my_folder_for_figures
```

If you do not provide the –figures_folder argument, ./figures directory will be created and the plots will be put there.

2.2.1.2.1 Example usage files There are also some files with example usage provided in HiCEnterprise/example_usage directory i.e.:

- TSS enrichment analysis for found predictions of interactions
- generating random bins
- extracting unique bins and adjusting statistics for comparison between two analyses

Those additional scripts may generate new directories and results. For now, as they are not an official part of the package they are not discussed in this documentation.

2.3 Output files

2.3.1 stats folder

This folder contains the results from the analysis. Result filenames consist of regions file name, map folders names, either 'stats' or 'significant', chromosome name, number of bins considered, resolution and statistical cutoff threshold. Example: regions-map40kb_replicate1-map40kb_replicate2-stats22-200x40000bp-0_01.txt.

2.3.1.1 'stats' files

Stats files are text files that contain statistics derived from analysis. They provide the number of regions (bin regions) in the analysis, how many had predictions and some simple statistics.

2.3.1.2 'significant' files

Files with 'significant' in name contain the statistically significant (with given FDR threshold) interactions derived from the analysis. They can be created in three formats (see 2.4):

2.3.1.2.1 custom txt format Txt contains bin and regions coordinates and corresponding predictions of interactions with q-values.

2.3.1.2.2 modified BED format 3 first columns are coordinates of the predicted interaction, 4th column encodes coordinates of region that interacts, 5th column is score (-log10 q-value).

2.3.1.2.3 modified GFF format Modified GFF contains columns as follows:

- chromosome name
- program name
- type: group (group of regions and their predictions), region, prediction
- start coordinate
- end coordinate
- score (-log10 q-value) for predictions
- strand (none, '.')
- frame (none, '.')
- ID of Parent (group of regions and their predictions) with chromosome name and bin from original assembly

Results in GFF files can be remapped from the original assembly to another one with argument –remap (see 2.4). GFF files are suitable for a representation in Jbrowse with modified CSS (thanks to Karolina Sienkiewicz).

2.3.2 figures folder

In this folder you will find the plotted figures if the –plotting argument was used. Interaction profiles are extracted to PDF files. Files are named by the regions file name and additional plotting arguments (–num_regs, –section). Each interaction profile is composited of weighted log Intensity, -log10 p-values and -log10 q-values.

2.3.3 pickles folder

Pickles folder contains pickle files (generated by pickle python library) with saved parts of results that may be useful for running new analyses on the same data (i.e. same Hi-C maps but other regions). Weibull fitting takes a while, so saving the parameters of the distribution for a given map, chromosome and distance can save a lot of time in future calculations.

2.4 All arguments listed alphabetically

- Python version 2.7 or 3.5/6
- required python packages are provided in requirements.txt file.
- if you want to plot through R, you need have R installed with needed packages.
- -a, –section : Section of bp for the plot
- -b, –bin_res : REQUIRED Resolution (size of the bins on the Hi-C maps) in bp i.e. 10000 for 10kb resolution. All maps should be of same resolution.
- -c, –chr : REQUIRED Name of the chromosome for which to extract domain interactions.
- -f, –figures_folder : Folder to save the figures (plots) in. Default is '../figures/'.
- -m, –hic_folders : REQUIRED Folder/folders in which the Hi-C data is stored with file names in format mtx-N-N.npy, where N = chromosome name
- -n, –num_bins : Number of bins left and right to be considered when extracting the interaction profile.
- –num_regs : Number of regions to plot
- -p, –pickled_folder : Folder with pickled files (to load from/save in). Default is '../pickles/'.
- –plotting : If there should be plotting, and if so should it be with rpy2 or matplotlib. Options: mpl, rpy2. Default is mpl.
- -r, –region_file : REQUIRED Any tab-delimited BED file or FASTA File from Enhancer-Atlas with regions and their positions.
- –regs_algorithm : Algorithm for sorting regions to bins. Options: all, one. Default is all. 'all' assigns the region to all bins that if fits into; 'one' chooses the bin in which most of the region is.
- –remap : Assembly that maps and regions are currently in, and assembly in which to save the stats in. Format: assembly:new_assembly i.e. hg_19:hg_38. Currently only available for GFF files.

- -s, –stats_folder : Folder to save the statistics and significant points in. Default is '../stats/'.
- –single_sig : If the single map significant points should be saved or not.
- –stat_formats : In which file formats to save statistics. It is possible to provide multiple formats. Available: txt, bed, gff. Default is only txt.
- -t, –threshold : Statistical cutoff threshold. Default is 0.01.

Chapter 3

HiCEnterprise domains

In the domain part of the package the significance of interaction between domains A and B is calculated as the probability of observing a stronger contact under the fitted distribution. There are three distributions available: Hypergeometric (default), Negative Binomial and Poisson.

3.1 Available distributions

3.1.1 Poisson

Another approach is to assume that the contacts are distributed according to the Poisson distribution with the λ parameter depending on the distance from the diagonal. In this case, for every domain pair, we estimate the average contact frequency λ_d in the diagonals corresponding to the distance between them and score the given domain contact against the Poisson distribution with the mean and variance λ_d. Poisson distribution is used as default by the program.

3.1.2 Hypergeometric

$$Pvalue = 1 - \sum_{i=0}^{k-1} \frac{\binom{a}{i}\binom{N-a}{b-i}}{\binom{N}{b}} \tag{3.1}$$

Here N (population size) is the sum of all frequencies in the Hi-C map, a is the sum from the column of domain A and b is the sum from the row of domain B (number of success states in the population and number of draws). k is the number of contacts between domains A and B. This idea and original code was developed by Irina Tuszyńska and presented in (Niskanen et al. n.d.).

3.1.3 Negative Binomial

The negative binomial distribution is slightly more general than the Poisson distribution discussed in the previous section. In particular it allows for overdispersion of the data. In this case, for each pair of domains at the distance d we estimate the mean contact frequency μ_d and variance $\sigma_d^2 \geq \mu$ from all values in the respective diagonals and then score the contact significance of this domain pair against the negative binomial distribution with the estimated mean and variance.

3.2 Input files

To run the domain-domain interactions analysis you need:

- one Hi-C map in numpy format for one chromosome
- file with domains definition in the format: domain_id chromosome_name start_coordinate end_coordinate (space delimited). Optionally fifth column may contain sherpa_level.

All coordinates should be provided in bp.

3.2.0.1 Example files

Example files are provided in HiCEnterprise/HiCEnterprise/tests/test_files directory in maps and doms.
Overall, the usage looks like this:

```
HiCEnterprise domains domains_options
```

To see the available arguments:

```
HiCEnterprise domains -h
```

3.2.0.2 Example runs

There are a few required arguments for the analysis.

- Path to the Hi-C map in a numpy format: --hic_map ./map40kb.npy
- File with domain definition: --domain_file domains.txt
- Name of the chromosome to run the analysis for: --chr 22
- Resolution (bin size) of Hi-C maps in bp (all maps provided should be of the same resolution): --bin_res 40000

For information on formats of files see 3.2.
Run with only required arguments:

```
HiCEnterprise domains --hic_map ./map40kb.npy --domain_file
./domains.txt --chr 22 --bin_res 40000
```

3.2.0.3 Visualisation

If you want to plot the results with matplotlib, add –plotting argument to the command. Resulting figures can be found in director provided by –figures_folder option.
In example:

```
HiCEnterprise domains --hic_map ./map40kb.npy --domain_file
./domains.txt --chr 22 --bin_res 40000 --plotting
--figures_folder ./my_folder_for_figures
```

If you don not provide the –figures_folder argument, ./figures directory will be created and the plots will be put there. There are additional options that you can use to change your visualization, i.e. if you want to change colors that represent the Hi-C map and domain-domain interactions, you can use –hic_color and –interactions_color with any available palette from https://matplotlib.org/api/pyplot_summary.html.

3.3 Output files

3.3.1 stats folder

This folder contains the results from the analysis. Result filenames consist of domains file name, map name, 'stats' or 'corr_stats', chromosome name, statistical cutoff threshold, and distribution name. Example: domains-map40kb-stats22-0_01-hypergeom.txt.

3.3.2 figures folder

In this folder you will find the plotted figures if the –plotting argument was used. Interaction maps are extracted to PNG files. Files are named by the map name, "corr" if the FDR corrected data is on the plot, domains file name, and distribution name. Example: map40kb-corr_domains-hypergeom.png
More interesting examples with real HiC maps are available on the HiCEnterprise site http://regulomics.mimuw.ed

3.4 All arguments listed alphabetically

- –all_domains : Stop remove pericentromeric domains and domains with rare interdomains contacts, where a mean number of contacts in one row is less than n*numer_of_domains. n = 1 and can be changed by –interact_indomain parameter
- -b, –bin_res : REQUIRED Resolution (size of the bins on the Hi-C maps) in bp i.e. 10000 for 10kb resolution.
- -c, –chr : REQUIRED Name of the chromosome for which to extract domain interactions.
- –distribution : The distribution on which you would like to base the identification of domain-domain interactions. Available: hypergeom, negbinom, poisson. Default is hypergeom.
- -d, –domain_file : REQUIRED Text file with domain definition in the format: dom_id(int) chromosome(int) dom_start(bp) dom_end(bp) sherpa-lvl(optional).
- -e, –ticks_separation : Frequency of ticks on the plot.
- -f, –figures_folder : Folder to save the figures (plots) in. Default is './figures/'.
- -g –interact_indomain : Multiplier of domains_number, that is a threeshold for neglecting pericentromeric domains. Mutualy exclusive with –all_domains option. Default = 1, higher number - more domains will be removed.
- -l, –plot_title : The title of the plot. If it contains spaces, use quotation marks. Default is "Interactions".
- -m, –hic_map : Path to the single-chromosome Hi-C map in numpy format.
- -n, –hic_name : Name to use for Hi-C map. Default is the name of the file.
- -o, –hic_color : The color of HiC map. You can choose the palette from Colormaps available here: https://matplotlib.org/api/pyplot_summary.html. Recommended are Reds, Blues, YlOrBr, PuBu. Default is Greens.
- –plotting : If results should be plotted. matplotlib library is required.

- -r, –interactions_color : The color of interactions. You can choose the palette from Colormaps available here: https://matplotlib.org/api/pyplot_summary.html. Recommended are Reds, Blues, YlOrBr, PuBu. Default is YlOrBr.
- -s, –stats_folder : Folder to save the statistics and significant points in. Default is './stats/'.
- –sherpa_lvl : If there are sherpa levels in the file and which one to use.
- -t, –threshold : Statistical cutoff threshold. Default is 0.01.