DIVERSITY

Authors: Sneha Mitra, Anushua Biswas, and Leelavati Narlikar

# Documentation

DIVERSITY is an unsupervised learning method that identifies different modes of protein-DNA binding from high-throughput chromatin immunoprecipitation (ChIP) data. The aim is to report the sequence component that is likely to be the cause of a specific region being reported in the ChIP experiment. This component may be a direct binding site of the profiled protein or a contact through intermediaries.

DIVERSITY is available as a web-server `http://diversity.ncl.res.in/` as well as a standalone software downloadable from `https://github.com/NarlikarLab/DIVERSITY`.

DIVERSITY can be run using the command `diversity` after installation. A successful run produces an HTML file `all_models.html` where all models can be viewed at a glance. It contains the link to the best model, the one with the optimal number of modes. Data of all learned models are stored in individual directories of the form `modelWith_`$\langle m \rangle$`_modes`. Here, $m$ denotes the number of modes of the model. Within each such directory, DIVERSITY saves the output of all trials, corresponding to different Gibbs sampling runs, for the given mode. These trial directories contain the sequence logos of the discovered motifs along with `info.txt` that lists the mode of binding for every sequence of the input file. A copy of the output of the best trial is saved outside of the trial directories.

A binary file `models.bin.p` is generated that consists of information regarding the models learned and the number of free parameters found in every model. This binary file helps in deciding the best model using Bayesian model selection.

## Prerequisites

The following packages need to be installed in order to run DIVERSITY:

- Python 2.7+
- python-numpy
- python-ctypes
- python-multiprocessing
- python-re

## Installation

DIVERSITY is freely available at `https://github.com/NarlikarLab/DIVERSITY`. Execute the following commands to download and install DIVERSITY:

```
$ wget https://github.com/NarlikarLab/DIVERSITY/archive/v1.0.0.tar.gz
$ tar -xvf v1.0.0.tar.gz
$ cd DIVERSITY-1.0.0/
$ make
```

To execute DIVERSITY from anywhere export the path to DIVERSITY to the PATH variable:
`$ export PATH=/path/to/DIVERSITY:$PATH`

## Usage

To run DIVERSITY:

`$ /path/to/DIVERSITY/diversity [options]`

## Options

The various options for running DIVERSITY are as follows:

- -f filename

  Mandatory input. Data file for which motifs are to be identified. File must be in fasta format.

- -o directory

  Valid directory name. If it exists then a new one is create with given name along with an extension number. Default directory: diversityOut in the current working directory.

- -maskReps $b$ (binary)

  1 masks lower case nucleotides by replacing them with 'N' (default). 0 will not mask.

- -r $b$ (binary)

  1 to allow motifs to be present in both strands (default), 0 to allow only on given strand

- -zoops $z$ (real number between 0 and 1)

  0 insists on all sequences to have a motif (default). 1 to let the sequence length decide prior on presence of motifs. Value strictly between 0 and 1 is the probability of a sequence not having a motif.

- -minWidth $w$ (non negative integer)

  Minimum width of motifs while training models. Default value: 6

- -initialWidth $w$ (non negative integer)

  Starting width of motifs while training models. Default value: 12

- -minMode $m$ (non negative integer)

  Minimum number of modes to allow. Default value: 3

- -maxMode $M$ (non negative integer)

  Maximum number of modes to allow in the model. Default value: 10

- -lcount $c$ (non negative integer)

  Number of random initializations for Gibbs sampling per model. Default value: 5

- -proc $np$ (non negative integer)

  Maximum number of processors to be used for computation. Default value: number of processors present in the system

- -fast $f$ (non negative number)

  Speed up/slow down execution by $f$ times. $f > 1$ speeds up by n times. $0 < f < 1$ slows down by a factor of $f$. $f = 0$ runs till last $n$ likelihood values has a linear fit (where $n$ is number of sequences: very slow). Default value: 1

## Examples

The following examples illustrate the usage of the options.

- To run with all the default options:

  $ `diversity -f example.fa`

- To run for minimum 6 and maximum 12 modes without masking repeats and initial motif width of 10:

  $ `diversity -f example.fa -minMode 6 -maxMode 12 -maskReps 0 -initialWidth 10`

- To find list of options:

  $ `diversity`