

DICE: Cell lineage reconstruction from single-cell CNA data

Description

DICE (short for Distance-based Inference of Copy-number Evolution) is a collection of fast and accurate methods for reconstructing cell lineage trees from single-cell copy number aberration data. Most notable among these methods are DICE-star and DICE-bar, which use standard-root and breakpoint-root distances, respectively, and reconstruct the phylogeny using a balanced minimum evolution criteria. DICE-star and DICE-bar have both been found to be generally more accurate and far more scalable than other, more complex, model-based approaches for reconstructing cell lineage trees from single-cell somatic copy number alteration data. Both approaches, and many variants, are implemented in a single Python file and can be easily run using a python interpreter. DICE can be cited as follows:

DICE: Fast and Accurate Distance-Based Reconstruction of Single-Cell Copy Number Phylogenies

Samson Weiner and Mukul S. Bansal

Under review.

Installation

If on github, you can clone the this repository with

```
git clone https://github.com/samsonweiner/DICE.git
```

Otherwise, the source code can be downloaded at:

<https://compbio.engr.uconn.edu/software/dice/>.

Dependencies

The following python packages are required to run the software: * Numpy * Pandas

Additionally, DICE requires the fastme package. The easiest and recommended approach to install fastme is with `conda` (see here). Otherwise, users can download existing executables from the website. In this case, it is recommended that the executable be added to the user's `$PATH` variable.

The commands used for a full install with `conda` following best practices are shown below.

```
conda create -n DICE python=3
conda activate DICE
```

```
conda install numpy
conda install pandas
conda install -c bioconda fastme
```

Usage

Running DICE under default parameter settings will use the DICE-star method (standard root distance) and save the distance matrix to a file. The only required input is the path to the file containing the copy number profiles. To run DICE-star with the balanced ME phylogenetic reconstruction algorithm, use the command

```
python3 DICE.py -i inputProfiles.tsv -o outputDir -m balME
```

To run DICE-bar with balanced ME, use the same command with the `-b` flag:

```
python3 DICE.py -i inputProfiles.tsv -o outputDir -m balME -b
```

Input File Format: DICE** takes as input a single file (specified using the `-i` command line option) containing a tab-separated values (TSV) file describing the copy number profiles of all cells. The following headers are required for each file and should be placed on the first line: CELL (the cell id of the current row), chrom (the chromosome **X** of the current row in the form of **chrX**), start (the starting location in bp of the copy number bin), end (the ending location in bp of the copy number bin), CN states (the actual copy number of the bin in the current row). If total copy numbers are used, the value of CN states should be a single numerical value. If allele-specific copy numbers are used, the value of CN states should be **a,b** where **a** is the copy number for haplotype A, and **b** is the copy number for haplotype B.

E.g.

CELL	chrom	start	end	CN states
leaf1	chr1	0	10000	1,1
leaf2	chr1	0	10000	1,2
leaf5	chr3	50000	60000	3,4

See the attached sample input file for a full example.

Available command line options for DICE

`-i, --input` Path to input file (Required).

`-o, --output` Path to output directory (Default: current directory).

`-p, --prefix` Prefix to add to output files (Default: DICE variant).

`-s, --save-dm` Toggle to save distance matrix to file in PHYLIP format (Default: False).

`-b, --breakpoint` Toggle to use breakpoint profiles. Otherwise, uses standard profiles (Default: False).

`-t, --total-cn` Toggle to use total copy numbers. Otherwise, assumes allele-specific copy numbers (Default: False).

`-d, --dist-type` Choice of distance function. Options are *root*, *log*, *euclidean*, and *manhattan*. (Default: root).

`-m, --rec-method` Choice of phylogenetic reconstruction algorithm. Options are *balME*, *olsME*, *NJ*, and *uNJ*. If not specified, computes pairwise distances and saves to a file. (Default: None).

`-n, --use-NNI` Toggle to use NNI tree search. Otherwise, uses an SPR tree search. (Default: root).

`-f, --fastme-path` Path to the fastme executable. (Default: fastme).

`-z, --seed` RNG seed used in fastme. (Default: None).

Available distance measures

Currently there are 4 distance measures available: manhattan, euclidean, log, and root. The breakpoint variants of these measures can be achieved by using the `-b` toggle. The root and log distances are defined as taking the square root and logarithm of each term in the manhattan distance, respectively. DICE-star and DICE-bar both use the root distance, as it has been shown to perform the best among all four.

Available Reconstruction algorithms

DICE makes available the four phylogenetic reconstruction algorithms implemented in fastme for computing a tree from a distance matrix. These are: balanced minimum evolution (balME), ordinary least-squares minimum evolution (olsME), neighbor-joining (NJ), and unweighted neighbor-joining (uNJ). For more information on these algorithms, see the fastme documentation.

Contact

If there are any questions related to DICE, please contact Samson Weiner (samson.weiner@uconn.edu) or Mukul Bansal (mukul.bansal@uconn.edu), or visit <https://compbio.engr.uconn.edu/> for a complete list of available software and datasets.