destiny – diffusion maps for large-scale single-cell data in R

Philipp Angerer [1], Laleh Haghverdi [1], Maren Büttner [1], Fabian J. Theis [1,2], Carsten Marr [1,*], and Florian Buettner [1,†,*]

[1]Helmholtz Zentrum München - German Research Center for Environmental Health, Institute of Computational Biology, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
[2]Technische Universität München, Center for Mathematics, Chair of Mathematical Modeling of Biological Systems, Boltzmannstr. 3, 85748 Garching, Germany

ABSTRACT

Summary: Diffusion maps are a spectral method for non-linear dimension reduction and have recently been adapted for the visualization of single cell expression data. Here we present *destiny*, an efficient R implementation of the diffusion map algorithm. Our package includes a single-cell specific noise model allowing for missing and censored values. In contrast to previous implementations, we further present an efficient nearest-neighbour approximation that allows for the processing of hundreds of thousands of cells and a functionality for projecting new data on existing diffusion maps. We exemplarily apply *destiny* to a recent time-resolved mass cytometry dataset of cellular reprogramming.

Availability and implementation: *destiny* is an open-source R/Bioconductor package `http://bioconductor.org/packages/destiny` also available at `https://www.helmholtz-muenchen.de/icb/destiny`. A detailed vignette describing functions and workflows is provided with the package.

Contact: `carsten.marr@helmholtz-muenchen.de`, `f.buettner@helmholtz-muenchen.de`

1 INTRODUCTION

Recent technological advances allow for the profiling of individual cells, using methods such as single-cell RNA-seq, single-cell RT qPCRqPCR or cyTOF (Vargas Roditi and Claassen, 2015) These techniques have been used successfully to study stem cell differentiation with time-resolved single-cell experiments, where individual cells are collected at different absolute times within the differentiation process and profiled. While differentiation is a smooth but nonlinear process (Buettner and Theis, 2012; Haghverdi, Buettner, and Theis, 2015) involving continuous changes of the overall transcriptional state, standard methods for visualizing such data are either based on linear methods such as Principal Component Analysis and Independent Components Analysis or they use clustering techniques not accounting for the smooth nature of the data.

In contrast, diffusion maps – initially designed by Coifman, Lafon, et al. (2005) for dimensionality reduction in image processing – recover a distance measure between each pair of data points (cells) in a low dimensional space that is based on the transition probability from one cell to the other through several paths of a random walk. Diffusion maps are especially suited for analyzing single-cell gene expression data from differentiation experiments (such as time-course experiments) for three reasons. First, they preserve the global relations between data points. This feature makes it possible to reconstruct developmental traces by re-ordering the asynchronously differentiating cells according to their internal differentiation state. Second, the notion of diffusion distance is robust to noise, which is ubiquitous in single-cell data. Third, by normalizing for sampling density, diffusion maps become insensitive to the distribution of the data points (i. e. sampling density), which aids the detection of rare cell populations.

Here, we present a user friendly R implementation of diffusion maps including previously proposed adaptations to single cell data (Haghverdi, Buettner, and Theis, 2015) as well as novel functionality. The latter includes approximations allowing for the visualisation of large data sets and the projection of new data on existing maps.

2 DESCRIPTION: THE *DESTINY* PACKAGE

2.1 Algorithm

As input, *destiny* accepts an expression matrix or data structure extended with annotation columns. Gene expression data should be pre-processed and normalized using standard workflows (see Supplementary text S1) before generating the diffusion map. *destiny* calculates cell-to-cell transition probabilities based on a Gaussian kernel with width σ to create a sparse transition probability matrix M. If the user does not specify σ, *destiny* employs an estimation heuristic to derive this parameter (see Supplementary Text S2). In contrast to other implementations, *destiny* allows for the visualisation of hundreds of thousands of cells by only using distances to the k nearest neighbors of each cell for the estimation of M (see Supplementary Text S2). Optionally *destiny* uses an application-specific noise model for censored and missing values in the dataset (see Figure S1). An eigendecomposition is performed on M after density normalization, considering only transition probabilities between different cells. By rotating M, a symmetric adjoint matrix can be used for a faster and more robust eigendecomposition (Coifman, Kevrekidis, et al., 2008) The resulting data-structure contains the eigenvectors with decreasing eigenvalues as numbered diffusion components, the input parameters and a reference to the data.

† *Current address:* European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Cambridge, CB10 1SD, UK

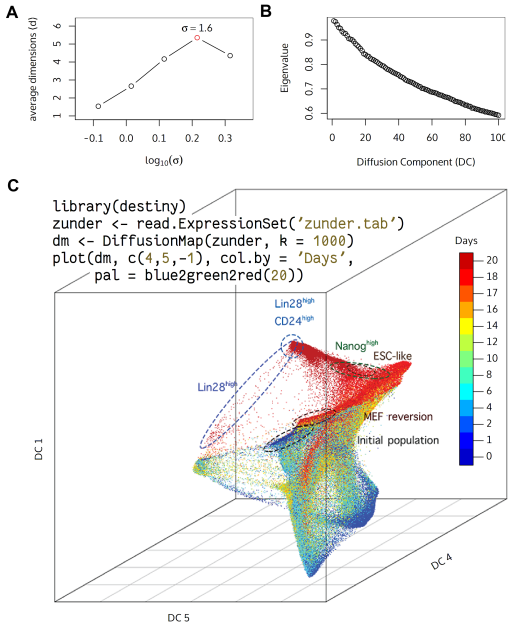

Figure 1. *destiny* applied to the mass cytometry reprogramming dataset of Zunder et al. (2015) with 36 markers and 256,000 cells. A) The optimal Gaussian kernel width σ. B) The Eigenvalues of the first 100 diffusion components decrease smoothly, indicating a large intrinsic dimensionality of the data. C) The initial population of mouse embryonic fibroblasts (MEFs, blue) is reprogrammed and profiled over 20 days. While a final cell population expressing stem cell markers is clearly separated, cells that revert to the MEF state are found proximal to the initial population in the diffusion map. (inset:) *destiny* code to generate the diffusion map

2.2 Visualization and projection of new data

This data-structure can be easily plotted and colored using the parameters of provided plot methods. An automatic color legend integrated into R's palette system facilitates the generation of publication-quality plots. A further new feature in *destiny* is the ability to integrate new experimental data in an already computed diffusion map. *destiny* provides a projection function to generate the coordinates for the new data without recalculating the diffusion map by computing the transition probabilities from new data points to the existing data points (see Supplementary Text S3).

3 APPLICATION

We applied *destiny* to four single-cell datasets of different size (hundreds to hundreds of thousands of cells) and characteristics (qRT-PCR, RNA-Seq and mass cytometry, see Supplementary Table S1). We first estimate the optimal σ that matches the intrinsic dimensionality of the data (Fig. 1A and Supplementary Figs. S2A and S3A). Using a scree plot (Fig. 1B and Supplementary Figs. S2B, S3B, and S4A), the relevant diffusion components can be identified. However, for big datasets as the mass cytometry data from Zunder et al. (2015) with 256,000 cells and 36 markers, corresponding Eigenvalues decrease smoothly. Although only a part of the intrinsic dimensionality can be represented in a 3D plot, the diffusion map reveals interesting properties of the reprogramming dynamics (Fig. 1C and Supplementary Fig. S5). We compared *destiny*'s performance to other implementations, including our own in MATLAB (based on Maggioni code[1], published with Haghverdi, Buettner, and Theis, 2015) and the diffusionMap R package (Richards, 2014) *destiny* performs similarly well for small datasets, while outperforming other implementations for large datasets (see Supplementary Table S1).

4 DISCUSSION AND CONCLUSION

We present a user-friendly R package of the diffusion map algorithm adapted to single-cell gene expression data and include new features for efficient handling of large datasets and a projection functionality for new data. We illustrate the capabilities of our package by visualizing gene expression data of 250,000 cells and show that our package is able to reveal continuous state transitions. Together with an easy to use interface this facilitates the application of diffusion map as new analysis tool for single-cell gene expression data.

ACKNOWLEDGEMENT

We thank Chris McGinnis (Seattle, USA) and Vicki Moignard (Cambridge, UK) for helpful comments on *destiny*.

Funding: Supported by the UK Medical Research Council (Career Development Award to FB) and the ERC (starting grant Latent-Causes to FJT). MB is supported by a DFG Fellowship through the Graduate School of Quantitative Biosciences Munich (QBM).

REFERENCES

Buettner, F. and F. J. Theis (2012) A novel approach for resolving differences in single-cell gene expression patterns from zygote to blastocyst. *Bioinformatics* 28.18, i626–i632.

Coifman, R. R., I. G. Kevrekidis, et al. (2008) Diffusion Maps, Reduction Coordinates, and Low Dimensional Representation of Stochastic Systems. *Multiscale Modeling & Simulation* 7.2, 842–864.

Coifman, R. R., S. Lafon, et al. (2005) Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* 102.21, 7426–7431.

Haghverdi, L., F. Buettner, and F. J. Theis (2015) Diffusion maps for high-dimensional single-cell analysis of differentiation data. *Bioinformatics*.

Richards, Joseph (2014) *diffusionMap: Diffusion map*. CRAN

Vargas Roditi, Laura de and Manfred Claassen (2015) Computational and experimental single cell biology techniques for the definition of cell type heterogeneity, interplay and intracellular dynamics. *Current Opinion in Biotechnology* 34, 9–15.

Zunder, Eli R. et al. (2015) A Continuous Molecular Roadmap to iPSC Reprogramming through Progression Analysis of Single-Cell Mass Cytometry. *Cell Stem Cell* 16.3, 323–337.

[1] `http://www.math.duke.edu/~mauro/code.html`