SLICER: Inferring Branched, Nonlinear Cellular Trajectories from Single Cell RNA-seq Data

Joshua D. Welch[1], Ziqing Liu[2], Li Wang[2], Junjie Lu[3], Paul Lerou[3], Jeremy Purvis[4], Li Qian[2], Alexander Hartemink[5], and Jan F. Prins[1]

[1] Department of Computer Science,
The University of North Carolina at Chapel Hill, Chapel Hill, USA
{jwelch,prins}@cs.unc.edu

[2] Department of Pathology, The University of North Carolina at Chapel Hill,
Chapel Hill, USA

[3] Department of Pediatric Newborn Medicine, Harvard Medical School, Boston, USA

[4] Department of Genetics, The University of North Carolina at Chapel Hill,
Chapel Hill, USA

[5] Department of Computer Science, Duke University, Durham, USA

1 Abstract

Understanding the dynamic regulation of gene expression in cells requires the study of important temporal processes, such as differentiation, the cell division cycle, or tumorigenesis. However, in such cases, the precise sequence of changes is generally not known, few if any marker genes are available, and individual cells may proceed through the process at different rates. These factors make it very difficult to judge a given cell's position within the process. Additionally, bulk RNA-seq data may blur aspects of the process because cells at sampled at a given wallclock time may be at differing points along the process. The advent of single cell RNA-seq enables study of sequential gene expression changes by providing a set of time slices or "snapshots" from individual moments in the process. To combine these snapshots into a coherent picture, we need to infer an "internal clock" that tells, for each cell, where it is in the process.

Several techniques, most notably Monocle and Wanderlust, have recently been developed to address this problem. Monocle and Wanderlust have both been successfully applied to reveal biological insights about cells moving through a biological process. However, a number of aspects of the trajectory construction problem remain unexplored. For example, both Monocle and Wanderlust assume that the set of expression values they receive as input have been curated in some way using biological prior knowledge. Wanderlust was designed to work on data from protein marker expression, a situation in which the number of markers is relatively small (dozens, not hundreds of markers) and the markers are hand-picked based on prior knowledge of their involvement in the process. In the initial application of Monocle, genes were selected based on differential expression analysis of bulk RNA-seq data collected at initial and final timepoints. In addition, Monocle uses ICA, which assumes that the trajectory lies along a linear projection of the data. In general, this linearity assumption may

M. Singh (Ed.): RECOMB 2016, LNBI 9649, pp. 239–240, 2016.
DOI: 10.1007/978-3-319-31957-5

not hold in biological systems. In contrast, Wanderlust can capture nonlinear trajectories, but works in the original high-dimensional space, which may make it more susceptible to noise, particularly when given thousands of genes, many of which are unrelated to the process being studied. Another challenging aspect of trajectory construction is the detection of branches. For example, a developmental process may give rise to multiple cell fates, leading to a bifurcation in the manifold describing the process. Wanderlust assumes that the process is non-branching when constructing a trajectory. Monocle provides the capability of dividing a trajectory into a branches, but requires the user to specify the number of branches.

In this paper, we present SLICER (Selective Locally linear Inference of Cellular Expression Relationships), a new approach that uses locally linear embedding (LLE) to reconstruct cellular trajectories. SLICER provides four significant advantages over existing methods for inferring cellular trajectories: (1) the ability to automatically select genes to use in building a cellular trajectory with no need for biological prior knowledge; (2) use of locally linear embedding, a nonlinear dimensionality reduction algorithm, for capturing highly nonlinear relationships between gene expression levels and progression through a process; (3) automatic detection of the number and location of branches in a cellular trajectory using a novel metric called geodesic entropy; and (4) the capability to detect types of features in a trajectory such as “bubbles” that no existing method can detect. Comparisons using synthetic data show that SLICER outperforms existing methods, particularly when given input that includes genes unrelated to the trajectory. We demonstrate the effectiveness of SLICER on newly generated single cell RNA-seq data from human embryonic stem cells and murine induced cardiomyocytes.