CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2020-0001

Vol. **2**, No. 1, pp. 1-55
March 2021

On the Mathematics of RNA Velocity I: Theoretical Analysis

Tiejun Li[1,*], Jifan Shi[2,*], Yichong Wu[1,*] and Peijie Zhou[3,*]

[1] *LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, China.*
[2] *International Research Center for Neurointelligence, The University of Tokyo Institutes for Advanced Study, The University of Tokyo, Tokyo 113-0033, Japan.*
[3] *Department of Mathematics, University of California, Irvine, Irvine, CA 92697, USA.*

Received 21 September 2020; Accepted 29 October 2020

Abstract. The RNA velocity provides a new avenue to study the stemness and lineage of cells in the development in scRNA-seq data analysis. Some promising extensions of it are proposed and the community is experiencing a fast developing period. However, in this stage, it is of prime importance to revisit the whole process of RNA velocity analysis from the mathematical point of view, which will help to understand the rationale and drawbacks of different proposals. The current paper is devoted to this purpose. We present a thorough mathematical study on the RNA velocity model from dynamics to downstream data analysis. We derived the analytical solution of the RNA velocity model from both deterministic and stochastic point of view. We presented the parameter inference framework based on the maximum likelihood estimate. We also derived the continuum limit of different downstream analysis methods, which provides insights on the construction of transition probability matrix, root and ending-cells identification, and the development routes finding. The overall analysis aims at providing a mathematical basis for more advanced design and development of RNA velocity type methods in the future.

AMS subject classifications: 60J28, 62P10, 92B15

Key words: RNA velocity, stochastic model, continuum limit, kNN density estimate.

1 Introduction

Single-cell RNA sequencing (scRNA-seq) is a rapid maturing technique, which makes the elaborate study of biological processes in the single cell resolution possible [50, 58].

*Corresponding author. *Email addresses:* tieli@pku.edu.cn (T. Li), shijifan@ircn.jp (J. Shi), wuyichongyt@pku.edu.cn (Y. Wu), peijiez1@uci.edu (P. Zhou)

The rich and diverse scRNA-seq datasets are revealing to us the mysteries of stem cell differentiation [54], heterogeneity in multicellular organisms [25], cancer cell dissection [8,37,64], drug discovery [22,59], etc. Every year, a swarm of analysis tools are produced by researchers all over the world [42, 62]. Some popular choices include the clustering tools [6, 28], trajectory inference tools [21, 40, 42, 47, 51, 53], and energy landscape tools [26,44,46,63], etc.

The characterization of stemness and lineage of the cells is a fundamental question in developmental biology. Although some practical indices, such as the signalling entropy and Markov chain entropy [47, 51], etc., are proposed to quantify the stemness of different cells in the scRNA-seq data analysis, they are more or less heuristic in nature. Recently, another promising method, the RNA velocity [29], was proposed to address this issue based upon the fact that the nascent (unspliced) and mature (spliced) mRNA can be distinguished in common single-cell RNA-seq protocols, such as SMART-seq2 [38], Drop-seq [32] and 10X genomics [66]. Thus, the relative abundance of unspliced and spliced mRNA are utilized to infer the velocity of each cell in the spliced mRNA abundance space, and predict the tendency of transition from one cell to another according to the RNA velocity model [29]. Improved methods in kinetic modeling, parameter inference and downstream analysis have been subsequently proposed [3, 41], showing the potential of RNA velocity to quantify the stemness of cells in a rational way.

Despite the fruitful results and promising applications of RNA velocity, it is of prime importance to understand the rationale underlying the algorithm design, as well as the subtle differences between different proposals from mathematical point of view. For instance, when constructing the cell-cell stochastic transition probability matrix from RNA velocity, La Mano et al. [29] and Qiu et al. [41] used the correlation scheme in the velocity kernel, while the cosine scheme was proposed in [3]. In the recent version of dynamo package [39], a scheme with local kernels [4] of diffusion was also utilized. In spite of their intuitive plausibility, the theoretical implications of different kernels demands further investigation. In addition, a tracking strategy of root and ending cells has been applied based on forward and backward diffusions [3, 29], whose theoretical basis remains to be established. Resolution of these puzzles based on a formal mathematical study will not only shed light on these theoretical problems, but also lead to a deeper comprehension of the RNA velocity and inspire further rational design of more delicate RNA velocity models. The current paper is devoted to this purpose.

In this work, we will present a thorough mathematical study on the whole process of RNA velocity model from kinetic model derivation, parameter inference algorithm to the downstream dynamical analysis. Our analysis will contribute insights toward several fundamental questions regarding RNA velocity and relevant downstream analysis, including:

- How to derive the deterministic and stochastic kinetic models of RNA velocity, and find analytical solutions?
- How to build the maximum likelihood estimator (MLE) of the parameters, built on