DOI:10.11721/cqnuj20240105

单细胞转录组学细胞动态分化数据模拟算法的比较与评估*

朵泓睿[1], 李映红[2], 李 勃[1]
(1. 重庆师范大学 生命科学学院, 重庆 401331; 2. 重庆邮电大学 生物信息学院, 重庆 400065)

摘要:对 9 种单细胞转录组学(single cell RNA sequencing,scRNA-seq)细胞动态分化数据模拟算法进行系统性地比较和评估,为算法的开发者和用户提供可靠的参考和帮助。利用多种评价指标对上述算法在准确性、可拓展性、适用性等 3 个方面进行全面评估,并对运行时间和内存消耗情况进行建模预测。结果显示:9 种被评估的算法均不能在数据特征与拓扑结构这 2 个方面同时有完美的表现;算法 Dyngen 虽然能模拟与参考数据拓扑结构和细胞分化轨迹相似度较高的数据,但是它的运行时间较长且内存消耗过大;将近一半的算法还需要及时更新版本信息并维护相关功能。用户在使用 scRNA-seq 细胞动态分化数据模拟算法时应综合考虑不同的使用场景和特点来选择最适合的模拟算法。

关键词:单细胞转录组学;动态分化;数据模拟;算法评估
中图分类号:Q31 **文献标志码**:A **文章编号**:1672-6693(2024)01-0125-08

单细胞转录组学(single cell RNA sequencing,scRNA-seq)的蓬勃发展和广泛应用积累了大量的相关数据,而对这些数据的整合分析和深度挖掘为阐明复杂生物学现象或过程提供了一条新的重要途径。在 scRNA-seq 数据挖掘中,轨迹推断(trajectory inference)是极具代表性的一种分析策略:它能按照一定的轨迹路线对细胞进行排列和分配,从而显化细胞分化发育阶段的整个过程[1]。通常研究者认为基因表达越相似的细胞,它们之间的距离越近。基于该假设,通过计算细胞与细胞之间基因表达的相似性,这些细胞会被有序地排列成一条代表着细胞状态动态变化过程的轨迹,并被赋予伪时间(pseudotime)的数值以衡量该细胞与祖细胞之间的距离[2]。轨迹推断不仅能更直观地反映和描绘细胞分化轨迹与谱系演化,且有利于揭示调控细胞命运选择的关键基因和基因表达值的变化过程[3-5]。

自从 Monocle 和 Wanderlust 开辟了轨迹推断分析的先河[6-7],各种相关算法层出不穷,建立的轨迹模型也越来越复杂,有线性、二叉、多分叉、树形、环形、无规则图等[8]。截至目前,大约共有 70 余种方法[9]专门被用于 scRNA-seq 数据的轨迹分析。在此类方法和软件开发的过程中,与其他类似软件进行比较和评估更能反映和突出新方法的优势。通常来讲,除了收集一些真实的细胞分化或时间序列的 scRNA-seq 数据之外,模拟数据也能作为另一种重要的参考数据被纳入评估研究中。为了实现对细胞动态分化过程中 scRNA-seq 的数据模拟,近些年来有一些模拟算法被陆续地开发出来以满足研究者在各种评估研究中对此类模拟数据的需求[10-12]。其中出现和应用最早的是基于分布模型的模拟算法,例如 Splatter[13]、MFA[14]、PROSSTT[15]等。此外,为了在模拟过程中体现真实的生物学过程,提高算法的可解释性与实用性,研究者开始利用基因表达调控的原理来模拟单细胞动态表达数据,例如 SymSim[16]。最近,以 RNA 速率(RNA velocity)[17]为基本原理的算法也相继出现,例如 VeloSim[18],表明数据模拟研究始终紧跟生物学理论研究的热点领域而不断创新发展。

现有的 scRNA-seq 细胞动态分化数据模拟算法在原理上差别较大,输入数据的格式与要求也不尽相同,且目前没有统一的标准和流程来评价某种算法模拟效果的优劣,仅从原理上无法确定哪种算法具有良好的模拟效果。于是,对此类算法的评估比较就显得十分关键。有鉴于此,本研究拟用参考数据集对目前多种流行的 scRNA-seq 细胞动态分化数据模拟算法从准确性(accuracy)、可拓展性(scalability)、适用性(usability)等 3 个方

* **收稿日期**:2022-11-23 **修回日期**:2023-06-20 **网络出版时间**:2023-06-25T10:34
资助项目:国家自然科学基金青年科学基金项目(No.31871274);重庆市教育委员会科学技术研究项目(No.KJQN202100538);重庆市中小学创新人才培养工程项目计划(No.CY220506)
第一作者简介:朵泓睿,男,研究方向为生物信息学,E-mail:duohongrui@cqnu.edu.cn;**通信作者**:李勃,男,副教授,博士,E-mail:libcell@cqnu.edu.cn
网络出版地址:https://link.cnki.net/urlid/50.1165.N.20230621.1951.010

面进行系统评估,并构建了一系列完整的数据模拟和结果评价的流程。此外,在对评估结果深入分析的基础上,本研究还依据不同模拟算法的评价表现与适用场景总结了一套实用指南,以期为用户选择合适的算法提供可靠参考,并为数据模拟算法的开发提供更紧贴实际的建议。

1 研究方法

1.1 模拟算法来源

通过在scRNA-seq数据分析算法网站数据库scRNA-tools(https://www.scrna-tools.org/)和PubMed文献数据库中查找和筛选,共有9种数据模拟算法被纳入评估队列(表1)。其中一些算法只需要通过参数设置就能实现数据模拟,而另外一些算法则需要相应的细胞分化轨迹信息(例如由细胞分化轨迹抽象得到的网络结构或者是标准的Newick tree格式的文件)作为额外输入。

表1 纳入评估的算法

Tab.1 Algorithms involved in the benchmarking

名称	编程语言	发表年份	模型算法	输入数据	理论上模拟的分化轨迹	来源文献
Splatter	R	2017	γ-泊松分布	参数	树形	文献[13]
MFA	R	2017	正态分布	参数	二叉树	文献[14]
PROSSTT	Python	2019	负二项分布	tree文件、参数	树形	文献[15]
Dyngen	R	2021	Gillespie随机模拟	参数	图	文献[19]
Dyntoy	R	2019	零膨胀负二项分布	网络、参数	树形	文献[20]
SymSim	R	2019	动力学模型	tree文件、参数	网络	文献[16]
TedSim	R	2021	动力学模型	tree文件、参数	树形	文献[21]
ESCO	R	2021	γ-泊松分布	tree文件、参数	图	文献[22]
VeloSim	R	2021	动力学模型	tree文件、参数	图	文献[17]

1.2 参考数据集的获取

本研究从开源社区zenodo(https://doi.org/10.5281/zenodo.1443566)中选取23组数据作为标准参考数据集。每个数据集均包含原始的基因表达矩阵、标准化后的基因表达矩阵、细胞类群注释信息、分化轨迹网络、标准的分化轨迹类型等信息。为最大限度保留数据的原始特征,所有数据均不进行预处理。

1.3 评估指标

为全面详尽地评估和展示各个算法的优劣,本研究使用了3类评估指标即准确性、可拓展性和适用性。

1.3.1 准确性评价

准确性评价分为数据特征分析和拓扑结构分析,分别用于衡量原始数据与模拟数据的特征和分化轨迹拓扑结构的相似程度(图1a):1) 针对数据特征分析,本研究使用基因平均表达值(gene mean expression)、基因表达值的标准差(standard deviation)、基因表达值的变异系数(coefficient of variability)、文库大小(library size)以及基因和细胞的零值比例(zero proportion per gene/cell)来概括和总结原始数据和模拟数据特征。随后,每组来自参考和模拟数据的特征队列(array)由小到大重排后,将根据表2左半部分列出的评价指标的有关公式计算出数据特征结果。2) 对于参考数据和模拟数据的拓扑结构比较,本研究首先使用slingshot[23]对模拟数据进行轨迹推断(trajectory inference),并将得到的结果与参考数据集作为输入,计算表2右半部分列出的4种指标[9]。

1.3.2 可拓展性评价

可拓展性分析(图1a)是在保证基因或者细胞数量一方不变(本研究中设置为1 000个)的前提下,设置细胞或者基因的数量梯度(以1 000个为步长,取1 000个到10 000个之间的数值梯度),检测每1次运行的时间和内存消耗,并利用结果建立单变量多项式回归模型(univariate polynomial regression model)[9]

$$y \approx \ln x + x^{1/2} + x + x^2 + x^3$$

以判断随着细胞(或基因)个数(x)的增长和运行时间(或内存消耗)数值(y)的趋势走向。

此外，本研究利用训练数据（利用梯度的细胞或基因数量模拟的监测数据）建立双变量多项式回归模型（bivariate polynomial regression model）：

$$y \approx \ln m + m^{1/2} + m + m^2 + m^3 + \ln n + n^{1/2} + n + n^2 + n^3,$$

其中：m 和 n 分别表示细胞个数和基因个数。随后，本研究利用测试数据集（每次模拟过程中监测的数据）对运行时间（或内存消耗）数值进行预测，计算真实值与预测值之间的皮尔逊相关系数（r），并对结果进行可视化展示。

1.3.3　适用性评价

本研究设计了一套详尽透明的计分表以量化算法的适用性（图 1a），即从可用性、代码、说明文档、方法维护、文章等方面进行人为主观评价和打分，并将打分结果作为适用性得分。

1.4　评估流程及过程

本研究整体评估流程如图 1b 所示。首先，一系列程序会被用于学习和提取参考数据中的重要参数。以得到的参数作为输入，程序将启动数据模拟步骤，同时对运行时间和内存消耗进行监测。数据模拟完成后，slingshot 会进行轨迹推断，并将结果与参考数据作准确性方面的比较。在评估的同时，按照计分准则对所有算法的适用性进行评分。完成以上步骤后，根据运行时间和消耗内存的数据计算算法的可拓展性得分。将准确性、可拓展性、适用性等 3 个方面总得分的平均值作为该模拟算法的最终得分。

本研究所有计算和评估过程均在 R 语言（版本 4.1.1）环境下完成，且每次运行时仅分配 1 个 CPU 核心（Intel(R) Xeon(R)W-2255 CPU@3.70 GHz）。为方便在 R 语言中直接调用和执行 Python 命令，本研究使用 reticulate 软件包（版本 1.22）中的 source_python 函数直接读取并执行“.py”格式文件中的 python 命令，以保证所有评估过程均在同一编程环境下进行。测试时，每组数据在同一算法下重复执行 10 次，以减小其他因素对实验的影响。本研究中使用 R 软件包 peakRAM（版本 1.0.2）中的 peakRAM 函数对运行步骤的时间和计算机的内存消耗进行检测。在数据整合与处理过程中，未在 0～1 范围之内的数值首先将被缩放到均值为 0、标准差为 1 的范围内，再利用正态密度函数将数据调整到 0～1 范围内。数据结果处理和可视化主要使用 dplyr（版本 1.0.8）和 ggplot2 软件包（版本 3.3.5）实现。

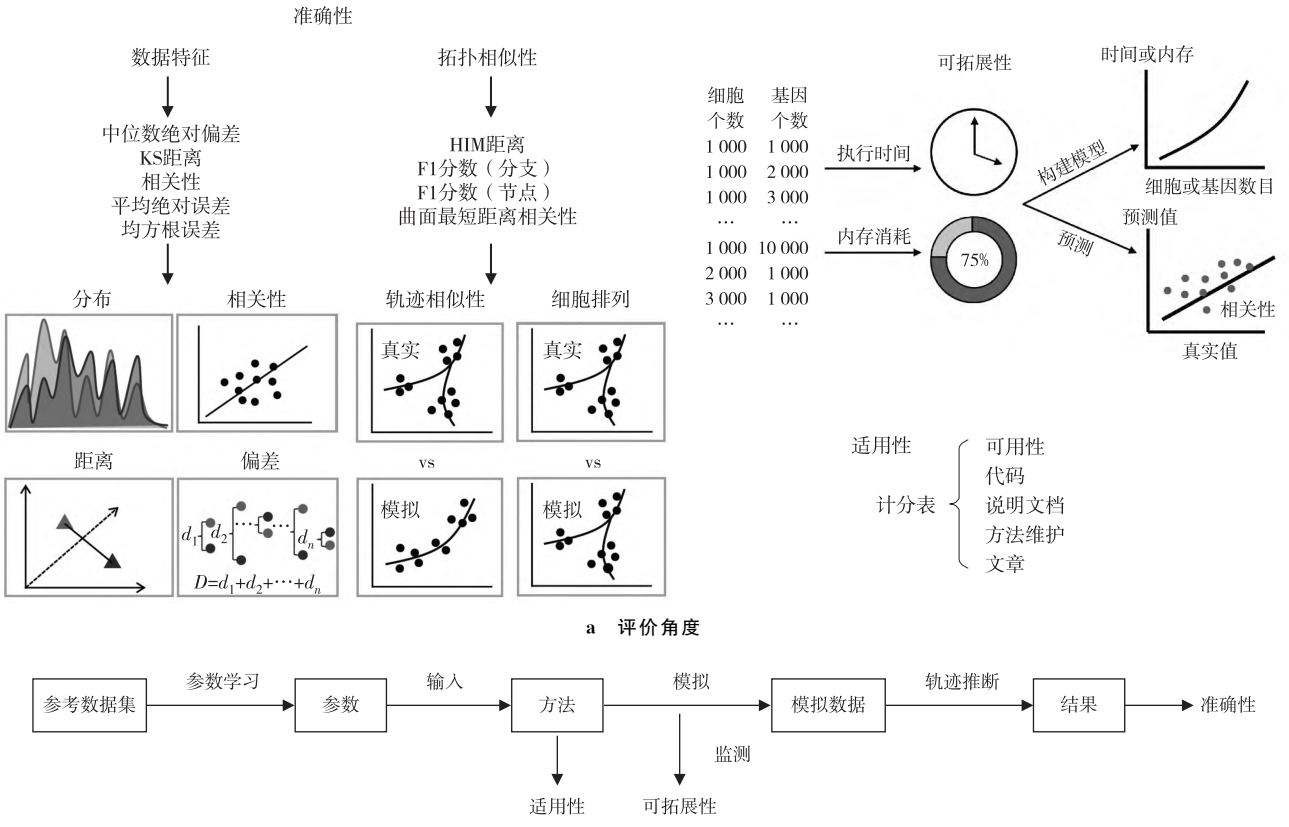


图 1　评价角度及评估流程示意图

Fig.1　An overview of the benchmark criteria and pipeline

表 2 准确性评估采用的评价指标

Tab.2 The evaluation metrics for accuracy assessment

指标名称	分析角度	指标名称	分析角度
中位数绝对偏差(median absolute deviation,MAD)	数据特征	HIM 距离(Hamming-Ipsen-Mikhailov distance,HIM)	拓扑结构
KS 距离(Kolmogorov-Smirnov distance,KS)	数据特征	F1 分数(分支)(F1 score for branches,$F1_{branches}$)	拓扑结构
相关性(correlation,Cor)	数据特征	F1 分数(节点)(F1 score for milestones,$F1_{milestones}$)	拓扑结构
平均绝对误差(mean absolute error,MAE)	数据特征	曲面最短距离相关性(correlation between geodesic distances,Cor_{dist})	拓扑结构
均方根误差(root mean squared error,RMSE)	数据特征		

2 结果与分析

各算法的总体评估结果及在准确性、可拓展性、适用性等方面的详细结果如图 2、图 3 和表 3 所示。

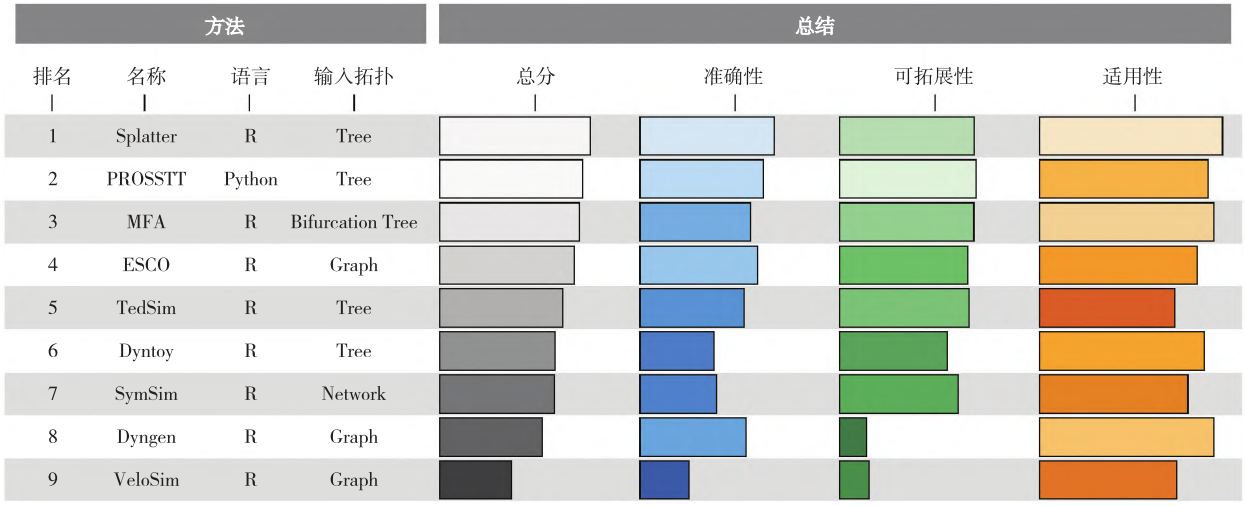


图 2 模拟算法基本特征及总体评估结果

Fig.2 Basic characteristics of the simulation algorithms and an overview of the evaluation results

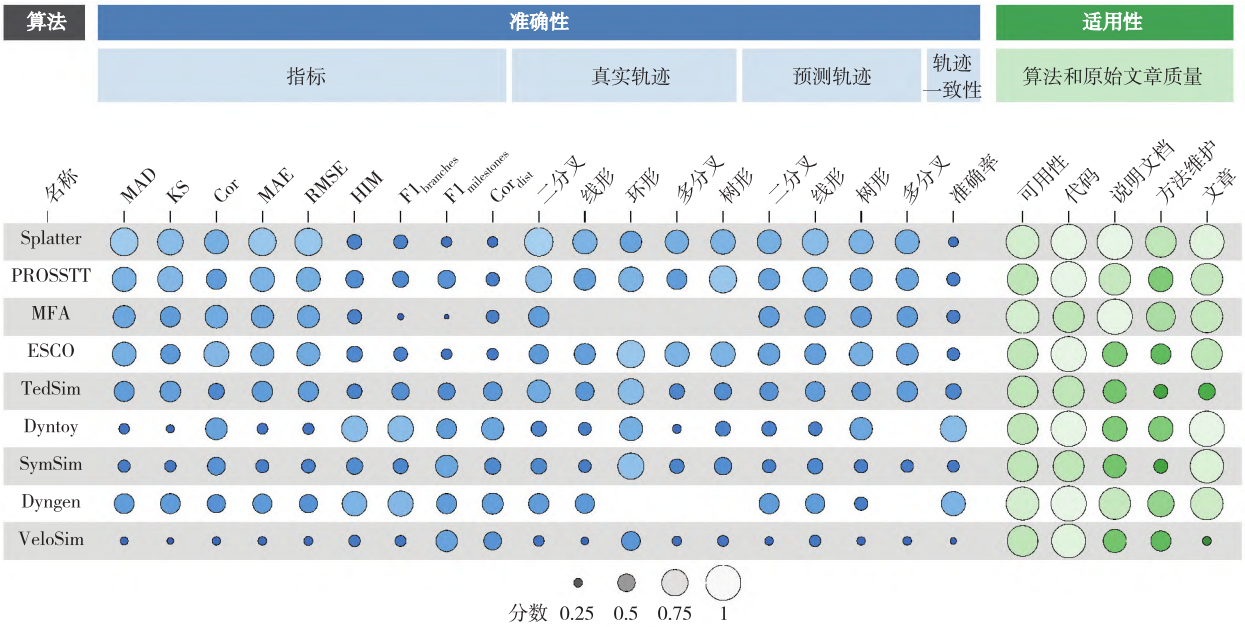


图 3 模拟算法在准确性和适用性方面的详细结果

Fig.3 Detailed results of simulation algorithms in terms of accuracy and usability

表 3　模拟算法在可拓展性方面的详细结果

Tab.3　Detailed results of simulation algorithms in scalability

算法	1 000 个细胞、5 000 个基因		5 000 个细胞、1 000 个基因		1 000 个细胞、10 000 个基因		10 000 个细胞、1 000 个基因	
	运行时间	内存消耗	运行时间	内存消耗	运行时间	内存消耗	运行时间	内存消耗
Splatter	3.64 s	625 MB	5.72 s	363 MB	6.8 s	840 MB	7.09 s	＞1 GB
PROSSTT	2.73 s	76 MB	9.31 s	76 MB	8.5 s	153 MB	12.1 s	153 MB
MFA	1.23 s	759 MB	1.23 s	763 MB	2.26 s	＞1 GB	2.33 s	＞1 GB
ESCO	9.5 s	＞1 GB	7.53 s	＞2 GB	18.17 s	＞2 GB	14.03 s	＞5 GB
TedSim	48.86 s	928 MB			1 min40 s	＞2 GB		
Dyntoy	1 min16 s	625 MB	5 min6 s	＞1 GB	2 min28 s	797 MB	22 min38 s	＞6 GB
SymSim	44.05 s	＞5 GB	46.03 s	＞4 GB	1 min35 s	＞8 GB	1 min37 s	＞4 GB
Dyngen	12 min1 s	＞50 GB	3 min48 s	＞8 GB	22 min41 s	＞84 GB	3 min54 s	＞9 GB
VeloSim	12 min52 s	＞10 GB	22 min11 s	＞9 GB	25 min33 s	＞16 GB	＞1 h	＞23 GB

2.1　准确性分析

2.1.1　各算法在数据特征和轨迹拓扑结构方面的表现

本研究将每个算法在 9 个准确性指标上的得分进行可视化（图 3）。显而易见的是，总评分前 5 名的算法 Splatter、PROSSTT、MFA、ESCO 和 TedSim 虽然在 MAD 等 5 个描述数据特征的指标上获得了远高于其他 4 种算法的得分，但是在 HIM 等 4 个评价分化轨迹的指标上得分较低（图 3）。这一结果表明上述 5 个算法虽然能模拟出与参考数据相似的数据特征，但是在模拟分化轨迹方面效果较差（图 2）。

SymSim 和 VeloSim 仅在少数几个指标中得分较高，但是从整体来看，它们模拟的数据与参考数据的相似性较低。虽然 Dyntoy 在 HIM 等 4 个评价指标的得分较高，但是仅在相关性这一项数据特征方面表现良好（图 3），因此它在数据特征模拟方面优势不大。Dyngen 与 Dyntoy 的情况类似，但 Dyngen 在数据特征方面的模拟效果要优于 Dyntoy，因此它在准确性方面的排名处于中等（图 2）。

显然，没有一种算法能同时在数据特征和拓扑结构的模拟上与参考数据保持高相似性，而是只能在其中某些方面具有良好效果。因此，现有的算法都存在一定缺陷，还无法为用户提供完美的选择。

2.1.2　各算法在不同拓扑结构数据中的表现

按照不同数据结构对每种算法进行评分（图 3 中真实轨迹部分），结果显示 Splatter 与 PROSSTT 以任何拓扑结构的数据为参考都能得到良好的模拟效果。ESCO 则是在呈环形、多分叉和树形拓扑结构数据上表现良好。此外，TedSim、Dyntoy 和 SymSim 的表现非常类似，都在环形拓扑结构数据的表现最好。除了 MFA、Dyngen 和 VeloSim，其他所有的算法都能在环形拓扑结构数据上得到较高的结构数据评分。值得注意的是，MFA 由于自身的特殊性而只能模拟二叉型分化轨迹数据，但是得到的结果却并不理想。VeloSim 则基本上在所有拓扑结构数据模拟中的表现都很逊色。

2.1.3　Dyntoy 和 Dyngen 在模拟分化轨迹方面的表现

总体上看，大多数算法在模拟拓扑结构相似性方面的结果不太理想，特别是总分排名靠前的几种算法（图 3）。相比之下 Dyntoy 和 Dyngen 在模拟具有相似分化轨迹数据方面的优势更加明显。具体来看，2 种算法在 HIM 指标上的得分远高于其他算法，表明模拟数据的拓扑结构与参考数据的结构相似性最高。此外，它们还在 $F1_{branches}$、$F1_{milestones}$、Cor_{dist} 等 3 种指标上的表现较好，说明被分配在轨迹上的细胞及它们的排列、分布和相互距离的相关性较高。另一方面，图 3 轨迹一致性部分显示 Dyntoy 和 Dyngen 模拟的数据具有较高的拓扑结构一致性准确率，这表明模拟数据拓扑结构与原数据的拓扑结构类型高度一致。因此，上述 2 种算法能更准确、更有效地模拟相似分化轨迹的数据。

2.2　可拓展性分析

从分析结果来看，Dyngen 和 VeloSim 在可拓展性方面的表现较差，即模拟数据消耗的时间长且需要较大的运行内存。例如 Dyngen 模拟1 000 个细胞、10 000 个基因的数据需要占用 84 GB 以上的内存；而 VeloSim 模拟 10 000 个细胞、1 000 个基因的数据则需要消耗 1 h 以上的时间（表 3）。这极大地影响了算法使用的灵活性，提

高了设备和时间的硬性要求,不利于用户快速地获得大量 scRNA-seq 模拟数据。

根据单变量回归模型的结果(图 4),可以发现大多数模拟算法都遵循以下规律:随着自变量细胞(或基因)个数的增长,因变量运行时间(或内存消耗)的增长速率逐渐减慢,即自变量与因变量之间呈自然对数或者 1/2 次幂的关系。另外只有 VeloSim 运行时基因个数与内存消耗之间呈线性关系。

双变量回归模型结果显示:该模型预测的运行时间与真实时间之间具有较强的相关性,相关系数值为 0.77(图 5a),且在内存消耗的预测方面具有良好效果(相关系数值为 0.68)(图 5b)。

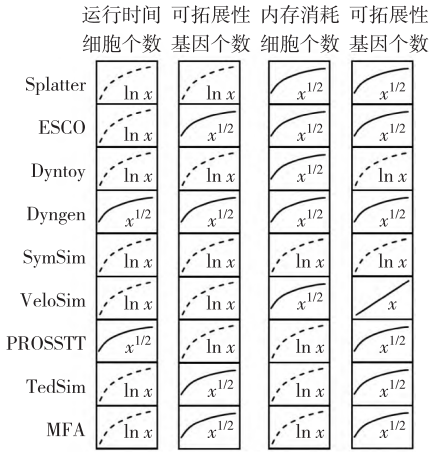
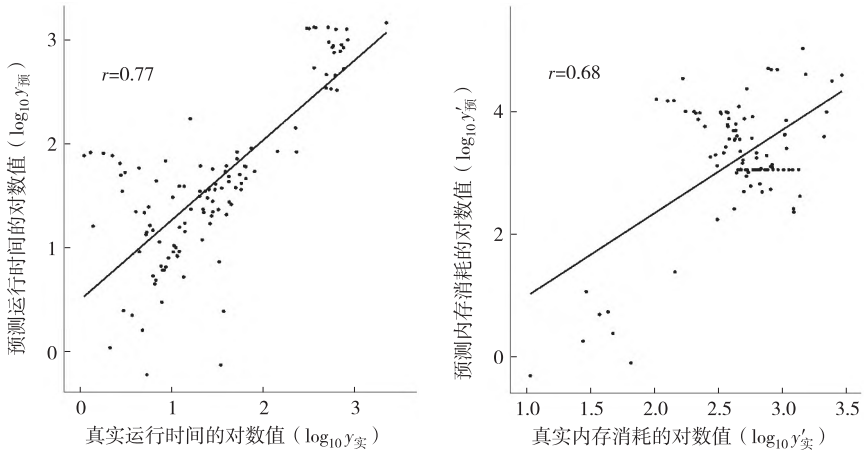


图 4 单变量回归分析预测时间和内存消耗的增长趋势

Fig.4 Univariate regression analysis for predicting the increasing trend in time consuming and memory usage



a 运行时间预测 b 内存占用预测

图 5 双变量回归模型对时间和内存的预测值与真实值之间的相关性分析

Fig.5 The bivariate regression model for correlation analysis between the predicting and true values of time and memory

2.3 适用性分析

适用性分析结果显示:所有算法都满足一些基本的要求,比如说能免费获取且包含简单的使用说明、帮助文档、开源的代码等;Splatter、PROSSTT、MFA 和 Dyngen 在适用性方面的表现近乎完美。但是仍有一些算法与上述 4 种算法存在较大差距,主要体现在算法维护和原始文章方面(图 3):ESCO、TedSim、SymSim 和 VeloSim 在算法维护方面得分较差,主要是因为它们从开发至今仍然没有升级过版本,也没有进行过内容和功能的修改;发布 VeloSim 的原始文章未经过同行评议(peer-review),故该算法的原理、论证过程和结果的科学性至今还不能被完全认同。因此,算法开发者应该以 Splatter 等几个算法为参照,借鉴和学习其中值得参考的地方,这将有利于增强用户学习并使用有关算法时的体验感。

3 总结与展望

本研究从准确性、可拓展性和适用性等 3 个方面系统地评估了 9 种 scRNA-seq 细胞动态分化数据模拟算法,结果表明:目前的模拟算法均不能在数据特征与拓扑结构这 2 个方面同时有完美的表现;Dyngen 虽然能模拟与参考数据拓扑结构和细胞分化轨迹相似度较高的数据,但是它的运行时间较长且内存消耗过大,难以满足用户对设备和等待时间的基本要求;此外将近一半的算法还需要及时更新版本信息并维护相关功能。

基于上述结果,本研究提出了一个关于算法选择问题的实用指南(图 6)。在选择相关算法时,用户需要考虑输入文件是否容易获取、编程语言是否熟悉以及算法使用体验感等因素。因为目前的算法都存在一定的不足,因此用户只能在数据特征或者拓扑结构相似性中的一方做出选择,以便最大限度地提高模拟质量并满足自身的使用需求。

在非必要情况下,用户更倾向于使用操作步骤简单、用时较短及对硬件设施要求较低的算法。为尽可能改善和提高用户体验感,开发者在开发有关算法时可以考虑以下措施:1) 针对算法中可能出现的流程繁杂、步骤冗余等情况,除了对流程、步骤等进行精简之外,还可以考虑用 snakemake[24] 将数据模拟步骤搭建为一系列流程,或者封装为 Docker[25] 镜像。2) 在算法使用手册中,应尽可能详细地描述输入文件格式、可能出现的错误、参数

的意义及注意事项，从而帮助用户准确地了解算法的使用。3）需要定时检查各种依赖包的版本，并重复检验算法流程的完整性与可重复性，及时回复用户的反馈并做好改善工作；在必要时，开发者还可对原来的算法进行更深层地改进与优化，推出新一代版本[26]。

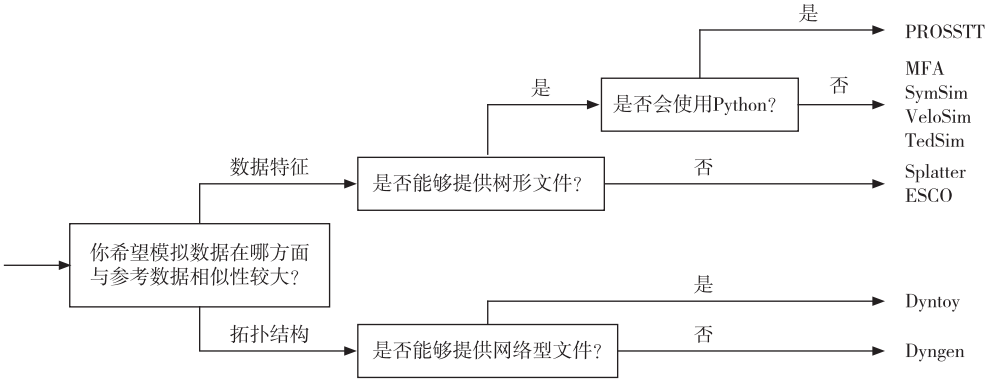


图 6　算法选择的参考指南

Fig.6　A reference guideline for selecting algorithms

scRNA-seq 细胞动态分化数据模拟算法的原理及背景趋向复杂。最早的模拟算法从真实数据服从的分布中抽取基因表达数据[13]，但 scRNA-seq 数据究竟遵循哪种分布却众说纷纭[27-28]。后来研究者将基因转录过程中外部和内部因素带来的误差考虑在内[16]，最近又根据 RNA 速率原理[17]进行模拟，使得数据模拟过程越来越接近真实的生物学过程，提高了数据模拟原理的可解释性，但模拟效果却并未呈现逐渐向好的趋势。因此，scRNA-seq 细胞动态分化数据模拟算法的开发工作不能只重视该算法针对特定任务的表现，还必须详细考察验证模拟数据与真实数据之间在数据特征等基本评价角度上的相似性，以确保算法能在一些基本的评价体系中发挥出色，使自身具有更大的优势与竞争力。

参考文献：

[1] HIE B，PETERS J，NYQUIST S K，et al.Computational methods for single-cell RNA sequencing[J].Annual Review of Biomedical Data Science，2020，3：339-364.

[2] TRITSCHLER S，BUTTNER M，FISCHER D S，et al.Concepts and limitations for learning developmental trajectories from single cell genomics[J].Development，2019，146(12)：dev170506.

[3] LIANG Y，KANEKO K，XIN B，et al. Temporal analyses of postnatal liver development and maturation by single-cell transcriptomics[J].Developmental Cell，2022，57(3)：398-414.

[4] ROSZAK P，HEO J O，BLOB B，et al. Cell-by-cell dissection of phloem development links a maturation gradient to cell specialization[J].Science，2021，374(6575)：eaba5531.

[5] WU F G，BARD J E，KANN J，et al.Single cell transcriptomics reveals lineage trajectory of retinal ganglion cells in wild-type and Atoh7-null retinas[J].Nature Communications，2021，12(1)：1465.

[6] TRAPNELL C，CACCHIARELLI D，GRIMSBY J，et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells[J].Nature Biotechnology，2014，32(4)：381-386.

[7] BENDALL S C，DAVIS K L，AMIR EL A D，et al. Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development[J].Cell，2014，157(3)：714-725.

[8] LUECKEN M D，THEIS F J.Current best practices in single-cell RNA-seq analysis：a tutorial[J].Molecular Systems Biology，2019，15(6)：e8746.

[9] SAELENS W，CANNOODT R，TODOROV H，et al. A comparison of single-cell trajectory inference methods[J]. Nature Biotechnology，2019，37(5)：547-554.

[10] NGUYEN N D，HUANG J W，WANG D F.A deep manifold-regularized learning model for improving phenotype prediction from multi-modal data[J].Nature Computational Science，2022，2(1)：38-46.

[11] STARK S G，FICEK J，LOCATELLO F，et al. SCIM：universal single-cell matching with unpaired feature sets[J]. Bioinformatics，2020，36(Suppl 2)：i919-i927.

[12] YU H S，WELCH J D.MichiGAN：sampling from disentangled representations of single-cell data using generative adversarial

networks[J].Genome Biology,2021,22(1):158.
[13] ZAPPIA L,PHIPSON B,OSHLACK A.Splatter:simulation of single-cell RNA sequencing data[J].Genome Biology,2017,18(1):174.
[14] CAMPBELL K R,YAU C.Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers[J].Wellcome Open Research,2017,2:19.
[15] PAPADOPOULOS N, GONZALO P R, SODING J. PROSSTT: probabilistic simulation of single-cell RNA-seq data for complex differentiation processes[J].Bioinformatics,2019,35(18):3517-3519.
[16] ZHANG X W,XU C L,YOSEF N.Simulating multiple faceted variability in single cell RNA sequencing[J].Nature Communications,2019,10(1):2611.
[17] La MANNO G,SOLDATOV R,ZEISEL A,et al.RNA velocity of single cells[J].Nature,2018,560(7719):494-498.
[18] ZHANG Z Q, ZHANG X W. VeloSim: Simulating single cell gene-expression and RNA velocity[EB/OL].[2022-11-23]. https://www.biorxiv.org/content/10.1101/2021.01.11.426277v2.
[19] CANNOODT R, SAELENS W, DECONINCK L, et al. Spearheading future omics analyses using dyngen, a multi-modal simulator of single cells[J].Nature Communications,2021,12(1):3942.
[20] SONG D Y,LI J J.PseudotimeDE:inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data[J].Genome Biology,2021,22(1):124.
[21] PAN X H,LI H C,ZHANG X W.TedSim:temporal dynamics simulation of single-cell RNA sequencing data and cell division history[J].Nucleic Acids Research,2022,50(8):4272-4288.
[22] TIAN J J,WANG J B,ROEDER K.ESCO:single cell expression simulation incorporating gene co-expression[J].Bioinformatics,2021,37(16):2374-2381.
[23] STREET K,RISSO D,FLETCHER R B,et al.Slingshot:cell lineage and pseudotime inference for single-cell transcriptomics[J].BMC Genomics,2018,19(1):477.
[24] MÖLDER F,JABLONSKI K P,LETCHER B,et al.Sustainable data analysis with snakemake[J].F1000Research,2021,10:33.
[25] MERKEL D J L J.Docker:lightweight linux containers for consistent development and deployment[J].Linux Journal,2014,2014(239):2.
[26] SUN T Y,SONG D Y,LI W V,et al.scDesign2:a transparent simulator that generates high-fidelity single-cell gene expression count data with gene correlations captured[J].Genome Biology,2021,22(1):163.
[27] SARKAR A,STEPHENS M.Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis[J].Nature Genetics,2021,53(6):770-777.
[28] VIETH B,ZIEGENHAIN C,PAREKH S,et al.powsimR:power analysis for bulk and single cell RNA-seq experiments[J].Bioinformatics,2017,33(21):3486-3488.

A Comparison and Benchmarking for Data Simulation Algorithms of Cellular Dynamic Differentiation in Single-Cell Transcriptomics

DUO Hongrui[1], LI Yinghong[2], LI Bo[1]

(1. College of Life Sciences, Chongqing Normal University, Chongqing 401331;
2. School of Biological Information, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: A systematic comparison and benchmark for 9 simulation algorithms of cellular dynamic differentiation in single-cell transcriptomics was conducted, and reliable guideline and reference for developers and users were provided. Various metrics were used for the comprehensive evaluation of 9 algorithms in terms of accuracy, scalability and usability, and a model was established for predicting the time consuming and memory usage. Results showed that 9 evaluated algorithms are not capable of performing well both in data property and cellular differentiation trajectory simulation. Dyngen can simulate data which is more similar to the reference data in topology and cellular differentiation trajectory, but it consumed more time and used more memory. Almost half of the algorithms needed updating versions and maintaining relevant functions. When using simulation methods designed for cellular dynamic differentiation, users are supposed to take different applying situations and features of the tasks into consideration in order to select the most suitable simulation algorithm.

Keywords: single-cell transcriptomics; dynamic differentiation; data simulation; algorithm evaluation

(责任编辑 方 兴)