Journal of Health and Biomedical Informatics *2024; 11(1): 72-82*
doi: 10.34172/jhbmi.2024.14
Medical Informatics Research Center, Kerman University of Medical Sciences

Original article

Providing a Multi-Scale Self-Encoding Method to Improve Clustering and Analysis of Single-Cell Data

Amin Einipour[1]*

1. Assistant Professor, Department of Computer Engineering, Andimeshk Branch, Islamic Azad University, Andimeshk, Iran

ARTICLE INFO:

Article History:

Received: 30 Apr 2024

Accepted: 31 May 2024

Published: 20 Jun 2024

Corresponding Author:

Amin Einipour

Email:
Amin.Einipour@iau.ac.ir

Citation: Einipour A. Providing a Multi-Scale Self-Encoding Method to Improve Clustering and Analysis of Single-Cell Data. Journal of Health and Biomedical Informatics 2024; 11(1): 72-82. [In Persian]

Abstract

Introduction: In bioinformatics, analyzing single-cell data is crucial for understanding cellular functions' complexities. However, this analysis faces challenges like inefficient dimensionality reduction and suboptimal clustering. This study aimed to present a method that enhances the clustering of single-cell data, improves reconstruction quality, and reduces data dimensions.

Method: This paper introduces SAMS (Single-cell Analysis using Multi-Scale Autoencoder), which uses a multi-scale autoencoder model to improve the challenges in single-cell data analysis. The SAMS method involves three primary steps: (1) data preprocessing and normalization, (2) employing a deep neural network model to reconstruct and reduce data dimensions with the help of a multiscale autoencoder, and (3) clustering the reduced data using the K-means algorithm to assess the method's performance.

Results: The SAMS method was implemented using Python on single-cell datasets. The results demonstrate that SAMS can effectively visualize cells in a two-dimensional space with an average Nearest Neighbor Error (NNE) rate of 89%, indicating a strong preservation of data structure. Additionally, the Silhouette index and Davis-Bouldin index, which measure clustering accuracy, show significant improvement with averages of 0.66 and 0.50, respectively.

Conclusion: The proposed SAMS method by combining the multiscale self-encoder model and the K-means algorithm could obtain better results than the previous methods. Its application in single-cell data analysis can aid researchers in gaining deeper insights into cellular functions and discovering new patterns.

Keywords: Single-cell analysis, Dimensionality reduction, Clustering analysis

[DOI: 10.34172/jhbmi.2024.14]

مجله انفورماتیک سلامت و زیست پزشکی دوره یازدهم، شماره اول، سال ۱۴۰۳، ص ۸۲-۷۲
doi: 10.34172/jhbmi.2024.14
مرکز تحقیقات انفورماتیک پزشکی، دانشگاه علوم پزشکی کرمان

مقاله پژوهشی

ارائه یک روش خودرمزگذار چند مقیاسی جهت بهبود خوشه‌بندی و تحلیل داده‌های سلول–منفرد

امین عینی‌پور[1]*

۱. استادیار، گروه مهندسی کامپیوتر، واحد اندیمشک، دانشگاه آزاد اسلامی، اندیمشک، ایران

اطلاعات مقاله:

سابقه مقاله:

دریافت: ۱۴۰۳/۲/۱۱

پذیرش: ۱۴۰۳/۳/۱۱

انتشار برخط: ۱۴۰۳/۳/۳۱

نویسنده مسئول:

امین عینی‌پور

ایمیل:

Amin.Einipour@iau.ac.ir

ارجاع: عینی‌پور امین. ارائه یک روش خودرمزگذار چند مقیاسی جهت بهبود خوشه‌بندی و تحلیل داده¬های سلول–منفرد. مجله انفورماتیک سلامت و زیست پزشکی ۱۴۰۳؛ ۱۱(۱): ۸۲-۷۲.

چکیده

مقدمه: تحلیل داده‌های سلول–منفرد نقش بسزایی در فهم پیچیدگی‌های عملکرد سلول‌ها ایفا می‌کند. تحلیل این داده‌ها با چالش‌هایی مانند کاهش ابعاد ناکارآمد و خوشه‌بندی نامطلوب مواجه هستند. هدف این مقاله ارائه روشی است که ضمن افزایش کیفیت بازسازی و کاهش ابعاد داده‌ها، خوشه‌بندی داده‌های سلول–منفرد را بهبود بخشد.

روش کار: در این مطالعه یک روش جدید به نام (Single-cell Analysis using Multi-Scale autoencoder) SAMS ارائه می‌شود که از یک مدل خود رمزگذار چندمقیاسی برای بهبود چالش‌های موجود در تحلیل داده‌های سلول–منفرد بهره می‌برد. روش پیشنهادی SAMS شامل سه مرحله اصلی است: (۱) پیش‌پردازش و نرمال‌سازی داده‌ها، (۲) استفاده از مدل شبکه عصبی عمیق برای بازسازی و کاهش ابعاد داده‌ها به کمک خودرمزگذار چندمقیاسی و (۳) خوشه‌بندی داده‌های کاهش‌یافته با استفاده از الگوریتم K-means برای ارزیابی عملکرد روش پیشنهادی.

یافته‌ها: روش پیشنهادی SAMS با استفاده از زبان پایتون پیاده‌سازی شده و نتایج به دست آمده بر روی مجموعه داده‌های سلول–منفرد نشان می‌دهد که SAMS می‌تواند سلول‌ها را با کیفیت بالایی در یک فضای دو بُعدی با میانگین نرخ ۸۹٪ =NNE نمایش دهد که نشان‌دهنده حفظ مناسب ساختار داده‌ها است. همچنین، شاخص‌های سیلوئت و دیویس–بولدین در ارزیابی دقت خوشه‌بندی، نشان می‌دهد که روش پیشنهادی به ترتیب با میانگین ۰/۶۶ و ۰/۵۰ توانسته بهبود خوبی را در خوشه‌بندی سلول‌ها ایجاد کند.

نتیجه‌گیری: روش پیشنهادی SAMS با ترکیب مدل خودرمزگذار چندمقیاسی و الگوریتم K-means توانسته نتایج بهتری نسبت به روش‌های پیشین به دست آورد. به‌کارگیری SAMS در تحلیل داده‌های سلول–منفرد می‌تواند به پژوهشگران در درک بهتر عملکرد سلول‌ها و کشف الگوهای جدید کمک کند.

کلیدواژه‌ها: تحلیل سلول–منفرد، کاهش ابعاد، تحلیل خوشه‌بندی

[DOI: 10.34172/jhbmi.2024.14]

مقدمه

تحلیل داده‌های سلول-منفرد (Single-Cell)، یکی از جدیدترین و قدرتمندترین روش‌ها در زیست‌شناسی سلولی و بیوانفورماتیک است که به محققان امکان می‌دهد تغییرات ژنتیکی و پروتئینی را در هر سلول به‌صورت مجزا بررسی کنند. این تکنیک‌ها اطلاعات دقیق‌تری درباره ناهمگنی سلولی درون بافت‌ها و پاسخ‌های سلولی به محیط‌های مختلف ارائه می‌دهند. تحلیل سلول-منفرد نقش مهمی در پیشرفت تحقیقات زیست‌پزشکی و به ویژه در درک بهتر بیماری‌های پیچیده مانند سرطان و توسعه روش‌های درمانی دقیق‌تر دارد. روش‌های تحلیل سلول-منفرد از تکنیک‌های مختلفی مانند RNA-Seq، پروتئومیکس و توالی‌یابی ژنومی سلول منفرد بهره می‌برند [۱،۲].

RNA-Seq سلول-منفرد یکی از ابزارهای مهم در تحلیل این نوع داده‌ها است. این تکنیک امکان بررسی بیان ژن در سطح هر سلول را فراهم می‌کند و به شناسایی سلول‌های نادر و بررسی دینامیک سلولی کمک می‌کند. تکنولوژی‌هایی مانند 10x-Genomics و Fluidigm به توسعه این حوزه کمک شایانی کرده‌اند [۲،۳]. تحقیقات نشان داده‌اند که RNA-Seq سلول-منفرد می‌تواند تفاوت‌های بیان ژن بین سلول‌ها را با دقت بالایی شناسایی کند.

پروتئومیکس سلول-منفرد یکی دیگر از حوزه‌های مهم در این حوزه است که به بررسی پروتئین‌های بیان‌شده در سطح هر سلول می‌پردازد. این روش‌ها اطلاعات دقیقی درباره فعالیت‌های زیستی سلول‌ها فراهم می‌کنند و به شناسایی پروتئین‌های کلیدی در فرآیندهای زیستی مختلف کمک می‌کنند [۴،۵].

توالی‌یابی ژنومی سلول-منفرد نیز به محققان امکان می‌دهد تا تغییرات ژنتیکی را در سطح هر سلول به‌صورت مجزا بررسی کنند. این تکنیک‌ها برای شناسایی جهش‌های نادر و بررسی تنوع ژنتیکی بین سلول‌ها بسیار مفید هستند [۶،۷].

با وجود ارزش بسیار بالای این نوع داده‌ها، اما چالش‌های این حوزه همچنان وجود دارد. یکی از چالش‌های اصلی در تحلیل داده‌های سلول-منفرد، حجم بالای داده‌ها و پیچیدگی‌های زیستی مرتبط با این مجموعه از داده‌ها است. روش‌های مختلفی برای کاهش ابعاد داده‌ها و تجسم الگوهای سلولی ارائه شده‌اند که شامل تکنیک‌های استانداردی مانند تحلیل مؤلفه‌های اصلی (Principal Component PCA(Analysis و توزیع تصادفی همسایگی ترکیبی (t-SNE) می‌باشند [۸،۹]. PCA با استفاده از تبدیل خطی، ابعاد داده‌ها را کاهش می‌دهد، اما ممکن است در تجسم ناهمگنی‌های پیچیده ناکارآمد باشد. t-SNE، یک تکنیک غیرخطی است که برای تجسم داده‌های تک‌سلولی استفاده می‌شود، اما به دلیل پیچیدگی محاسباتی، زمان پردازش بالایی دارد. با توجه به محدودیت‌های دقت و سرعتِ روش‌های کلاسیک کاهش ابعاد، در سال‌های اخیر روش‌های ترکیبی و جدید برای کاهش ابعاد و تحلیل این داده‌ها ارائه شده است [۱۰]. توسعه ابزارهای بیوانفورماتیکی مختلف می‌تواند به تحلیل داده‌های سلول-منفرد کمک کند. ابزارهایی مانند Seurat، SC3، Monocle و SIMLR از جمله ابزارهای محبوب در این حوزه هستند [۱۱-۱۳]. این ابزارها امکانات مختلفی برای پیش‌پردازش، کاهش ابعاد، خوشه‌بندی و تحلیل مسیرهای زیستی ارائه می‌دهند.

در حوزه بالینی نیز تحلیل داده‌های سلول-منفرد کاربردهای گسترده‌ای در تحقیقات بالینی دارد. این روش‌ها به شناسایی بیومارکرهای جدید، بررسی مکانیسم‌های بیماری و توسعه درمان‌های شخصی‌سازی‌شده کمک می‌کنند [۱۴،۱۵].

با توجه به موارد مذکور به نظر می‌رسد که تحلیل داده‌های سلول-منفرد به عنوان یک فناوری نوین، فرصت‌های بی‌نظیری برای درک بهتر از ناهمگنی سلولی و تحلیل دقیق‌تر تغییرات زیستی در سطح سلول‌های منفرد فراهم می‌کند. با استفاده از روش‌های پیشرفته تحلیل داده و تکنیک‌های جدید کاهش ابعاد، این حوزه به سرعت در حال پیشرفت است و پتانسیل زیادی برای کاربردهای بالینی و تحقیقات بنیادی دارد. روش معرفی شده در این مطالعه که SAMS نامگذاری شده از یک مدل خود رمزگذار چندمقیاسی به همراه الگوریتم K-means استفاده می‌کند. روش پیشنهادی ضمن افزایش کیفیت بازسازی و کاهش ابعاد داده‌ها، خوشه‌بندی داده‌های سلول-منفرد را بهبود می‌بخشد که می‌تواند به عنوان یک ابزار قدرتمند در این زمینه مورد استفاده قرار گیرد.

روش کار

در تحلیل داده‌های سلول-منفرد، هدف اصلی کاهش ابعاد داده‌ها و خوشه‌بندی سلول‌ها به گونه‌ای است که سلول‌های با ویژگی‌های مشابه در یک خوشه قرار گیرند. روش‌های سنتی مانند PCA و t-SNE به خوبی کار می‌کنند؛ اما ممکن است در مواجهه با داده‌های پیچیده

[DOI: 10.34172/jhbmi.2024.14]

مجله انفورماتیک سلامت و زیست پزشکی دوره یازدهم، شماره اول، سال ۱۴۰۳، ص ۸۲-۷۲
doi: 10.34172/jhbmi.2024.14
مرکز تحقیقات انفورماتیک پزشکی، دانشگاه علوم پزشکی کرمان

عملکرد ضعیفی داشته باشند. روش پیشنهادی، ترکیبی از الگوریتم‌های یادگیری عمیق و تکنیک‌های آماری است که برای بهبود دقت و کارایی تحلیل داده‌های سلول-منفرد طراحی شده است.

روش پیشنهادی که در این مطالعه به اختصار SAMS(Single-cell Analysis using Multi-Scale autoencoder) نامگذاری شده است یک روش خودرمزگذار چند مقیاسی تحلیل داده‌های سلول-منفرد و شامل سه مرحله اصلی است: پیش‌پردازش، بازسازی و کاهش ابعاد با استفاده از خودرمزگذار چندمقیاسی، و خوشه‌بندی داده‌ها.

مرحله ۱: جمع‌آوری و پیش‌پردازش داده‌ها

داده‌های سلول-منفرد از طریق روش‌های توالی‌یابی RNA سلول-منفرد (scRNA-seq) جمع‌آوری می‌شوند. در این روش، هر سلول به‌طور مجزا ایزوله شده و RNA آن استخراج و توالی‌یابی می‌شود. از روش‌های متداول در این زمینه می‌توان به Drop-seq و Smart-seq2 اشاره کرد. هر یک از این روش‌ها دارای مزایا و معایب خاص خود هستند. به عنوان مثال، روش Drop-seq امکان پروفایلینگ سریع و ارزان تعداد زیادی سلول را فراهم می‌کند، در حالی که Smart-seq2 دقت بالاتری در شناسایی ژن‌ها دارد.

پیش‌پردازش داده‌ها:

داده‌های خام به دلیل نویز و مشکلات فنی نیاز به پیش‌پردازش دارند. مراحل پیش‌پردازش به‌طور کلی شامل موارد زیر است:

۱- فیلتر کردن داده‌های بی‌کیفیت: حذف سلول‌هایی که کمتر از تعداد معینی ژن بیان شده دارند یا ژن‌هایی که در تعداد کمی از سلول‌ها بیان شده‌اند.

۲- نرمال‌سازی

نرمال‌سازی کتابخانه‌ای(Library Size Normalization): یکی از روش‌های معمول برای نرمال‌سازی، استفاده از فاکتور مقیاس (size factor) است که برای هر سلول محاسبه می‌شود. فاکتور مقیاس برای سلول i به صورت زیر و طبق رابطه (۱) محاسبه می‌شود:

$$SF_i = \frac{Total\ reads\ in\ cell\ i}{Median\ of\ total\ reads\ across\ all\ cells} \quad (۱)$$

سپس، مقادیر بیان هر ژن در سلول i با تقسیم بر این فاکتور، نرمال می‌شوند.

نرمال‌سازی لگاریتمی(Log-normalization): در این مرحله با استفاده از نرمال‌سازی لگاریتمی و طبق رابطه (۲) اقدام به متعادل کردن تفاوت‌های توالی‌یابی شدیم.

$$Log-normalized\ value = log_2\left(\frac{Read\ count + 1}{Size\ Factor}\right) \quad (۲)$$

این نرمال‌سازی تفاوت‌های بزرگ را کاهش می‌دهد و داده‌ها را برای تحلیل‌های آماری آماده می‌کند.

۳- تبدیل داده‌ها: در این مرحله، داده‌ها به فرم مناسبی برای ورود به مدل و در قالب یک ماتریس بیان ژنی آماده می‌شوند. در این ماتریس، ردیف‌ها نمایانگر سلول‌ها و ستون‌ها نمایانگر ژن‌ها هستند. هر مقدار در ماتریس بیانگر سطح بیان یک ژن خاص در یک سلول مشخص است.

[DOI: 10.34172/jhbmi.2024.14]

مرحله ۲: کاهش ابعاد مسئله با استفاده از یک روش خودرمزگذار چندمقیاسی

خودرمزگذار چندمقیاسی (Multi-Scale Autoencoder) یک نوع شبکه عصبی عمیق است که شامل لایه‌های متعددی است که به‌طور متوالی داده‌ها را به فضای با ابعاد کمتر کدگذاری می‌کنند و به این ترتیب داده‌ها را بازسازی می‌کنند. این ساختار به مدل اجازه می‌دهد تا ویژگی‌های پیچیده و چندمقیاسی داده‌ها را به خوبی یاد بگیرد. در این مرحله عملیات زیر انجام می‌شود:

۱- طراحی معماری خودرمزگذار: خودرمزگذار چندمقیاسی شامل یک شبکه عصبی است که داده‌های ورودی را به فضای با ابعاد کمتر کدگذاری می‌کند و سپس تلاش می‌کند که این داده‌ها را بازسازی کند. در این مرحله موارد زیر در نظر گرفته می‌شوند:

- شبکه عصبی کدگذار (Encoder): کدگذار شامل چندین لایه عصبی است که به صورت متوالی داده‌های ورودی را به ابعاد کمتری نگاشت می‌کنند. هر لایه از توابع فعال‌سازی غیرخطی مانند ReLU یا Leaky ReLU استفاده می‌کند تا توانایی مدل در یادگیری الگوهای پیچیده افزایش یابد. این شبکه وظیفه دارد که ویژگی‌های مهم داده‌ها را استخراج و به فضای ابعاد کمتر نگاشت کند. اگر ورودی داده‌ها را X در نظر بگیریم، فرآیند کدگذاری طبق رابطه (۳) به صورت زیر انجام می‌شود:

$$Z = f_{encoder}(X) = \sigma(W_{encoder}X) + b_{encoder} \quad (۳)$$

که در آن Z فضای کدگذاری شده با ابعاد کمتر، W وزن‌های کدگذار، b بایاس‌ها و σ تابع فعال‌سازی می‌باشد.

شبکه عصبی دیکدگذار (Decoder): دیکدگذار شامل چندین لایه عصبی است که داده‌های کدگذاری‌شده را به فضای اولیه بازسازی می‌کند. در این لایه‌ها نیز از توابع فعال‌سازی مانند ReLU یا sigmoid استفاده می‌شود. این فرآیند طبق رابطه (۴) و به‌صورت زیر انجام می‌شود:

$$X' = f_{decoder}(Z) = \sigma(W_{decoder}Z) + b_{decoder} \quad (۴)$$

که در آن X' داده‌های بازسازی شده، W وزن‌های کدگذار، b بایاس‌ها و σ تابع فعالسازی می‌باشد.

۲- آموزش شبکه: هدف از آموزش شبکه به حداقل رساندن خطای بازسازی با استفاده از الگوریتم‌های بهینه‌سازی مانند Adam است. الگوریتم Adam یک روش بهینه‌سازی مبتنی بر گرادیان است که از تخمین‌های مرتبه اول و دوم طبق رابطه (۵) استفاده می‌کند:

$$\theta_{t+1} = \theta_t - \frac{\alpha}{\epsilon + \sqrt{\hat{t}^{\upsilon}}} m_{\hat{t}} \quad (۵)$$

که در آن θ_t پارامترهای مدل در گام t، α نرخ یادگیری، $m_{\hat{t}}$ تخمین میانگین گرادیان، $\hat{t}^{\upsilon}$ تخمین واریانس گرادیان و ϵ یک مقدار کوچک برای جلوگیری از تقسیم بر صفر است.

مرحله ۳: خوشه‌بندی داده‌ها

در این مرحله با استفاده از ویژگی‌های استخراج شده حاصل از مرحله قبل، و با استفاده از الگوریتم K-means اقدام به خوشه‌بندی داده‌ها شد. این الگوریتم داده‌ها را به K خوشه تقسیم می‌کند به‌طوری‌که هر داده به خوشه‌ای اختصاص داده می‌شود که میانگین داده‌های آن خوشه به آن نزدیک‌تر باشد. در نهایت برای ارزیابی کیفیت خوشه‌بندی ایجاد شده از معیارهایی مانند شاخص سیلوئت و شاخص دیویس-بولدین استفاده می‌شود.

[DOI: 10.34172/jhbmi.2024.14]

روش پیشنهادی SAMS با استفاده از زبان پایتون و بر روی سه مجموعه داده سلول-منفرد به نام‌های Kolod، Buettner و Usoskin پیاده‌سازی شد.

مجموعه داده Usoskin شامل ۶۲۲ نمونه از سلول‌های عصبی بخشی از مغز موش است. تعداد ژن‌های این مجموعه داده ۱۷۷۷۲ و تعداد انواع سلول‌های واقعی نیز ۴ نوع است. مجموعه داده Buettner شامل سلول‌های بنیادی جنینی در مراحل مختلف چرخه سلولی هستند. این داده‌ها مجموعه‌ای از یک مطالعه کنترل شده است که اثر چرخه سلولی را بر سطح بیان ژن تعیین می‌کرد و شامل ۱۸۲ نمونه سلول، ۹۵۷۳ ژن و تعداد ۳ نوع سلول است. مجموعه داده Kolod نیز داده‌های سلولی بنیادی پرتوان در شرایط محیطی مختلف را نشان می‌دهد و شامل ۷۰۴ نمونه، ۱۳۴۷۳ ژن و ۳ خوشه سلولی واقعی است.

برای ارزیابی روش پیشنهادی ابتدا با استفاده از توابعی مانند UMAP اقدام به مصورسازی داده‌های اصلی و داده‌های بازسازی شده توسط روش پیشنهادی بر روی یک فضای دوبُعدی شد و با استفاده از معیار کمی به نام NNE کیفیت مصورسازی خوشه‌بندی محاسبه شد.

UMAP یک روش مصورسازی و کاهش ابعاد است که بر مبنای نظریه توپولوژی و هندسه دیفرانسیل عمل می‌کند و هدف آن حفظ ساختار محلی داده‌ها در فضای با ابعاد کمتر است. UMAP با ساخت یک گراف از همسایگی محلی داده‌ها و سپس بهینه‌سازی آن در فضای با ابعاد پایین، توزیع داده‌ها را به شکلی فشرده و قابل تجسم در می‌آورد. این روش به خصوص در تصورسازی داده‌های پیچیده با ابعاد بالا مانند داده‌های ژنومی و داده‌های تصویر موفق بوده است [۱۶]. همچنین برای بررسی کیفیت نهایی خوشه‌بندی از شاخص‌های سیلوئت و دیویس-بولدین استفاده شد.

معیار NNE: معیار NNE یا همان Nearest Neighbor Embedding یک معیار ارزیابی است که اغلب برای ارزیابی کیفیت نمایش داده‌ها در روش‌های کاهش ابعاد استفاده می‌شود. این معیار به بررسی حفظ روابط نزدیک‌ترین همسایگان در فضای کاهش‌یافته نسبت به فضای اصلی می‌پردازد. هدف NNE این است که بررسی کند که آیا ساختار همسایگی در فضای کاهش‌یافته به درستی حفظ شده است یا خیر. به طور خاص، NNE بررسی می‌کند که چند درصد از نزدیک‌ترین همسایگان هر نقطه در فضای اصلی، در فضای کاهش‌یافته نیز نزدیک‌ترین همسایه باقی مانده‌اند. مقدار حاصل از محاسبه NNE عددی در بازه [0,1] است به‌طوری‌که عدد نزدیک به ۱ به این به معنی است که ساختار همسایگی در فضای کاهش‌یافته به خوبی حفظ شده است و عدد نزدیک به ۰ به معنی این است که ساختار همسایگی به خوبی حفظ نشده است. این معیار به عنوان یک ابزار ارزشمند برای ارزیابی روش‌های کاهش ابعاد و حفظ ساختار داده‌ها در فضای کاهش‌یافته استفاده می‌شود.

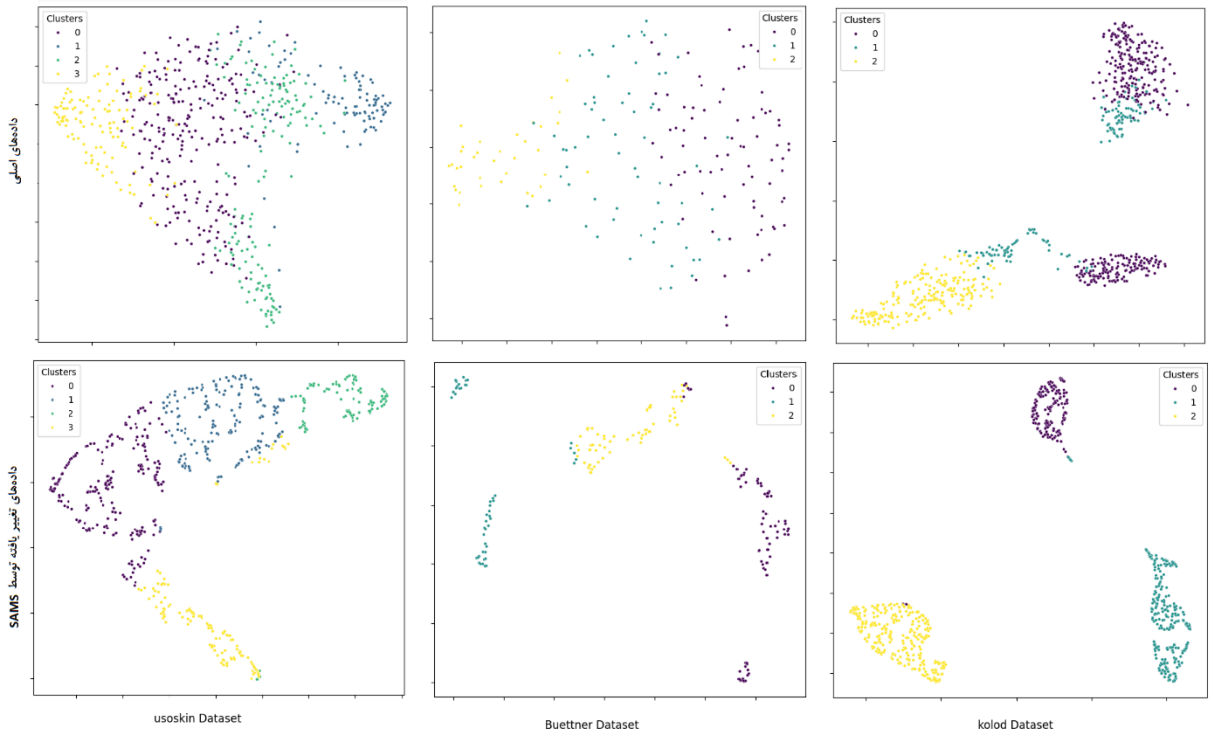
شاخص سیلوئت (Silhouette Index): این شاخص کیفیت خوشه‌بندی را با توجه به نزدیکی سلول‌ها به مراکز خوشه‌ها و فاصله بین خوشه‌ها ارزیابی می‌کند. مقدار سیلوئت معیار میزان شباهت یک شی به خوشه خود در مقایسه با خوشه‌های دیگر است. محدوده سیلوئت از ۱+ تا ۱ - است، که در آن مقادیر نزدیک ۱+ نشان دهنده نزدیکی نمونه به خوشه خود (خوشه‌بندی خوب)، مقدار نزدیک ۰ نشان دهنده نمونه‌های مرزی و مقادیر نزدیک ۱- نشان‌دهنده عدم نزدیکی نمونه‌ها به خوشه مربوطه (خوشه‌بندی نا مناسب) است. اگر بیشتر اشیاء از مقدار بالایی برخوردار باشند، ساختار خوشه بندی مناسب است. اگر بسیاری از نقاط دارای مقدار کم یا منفی باشند، در این صورت ممکن است ساختار خوشه‌بندی دارای خوشه‌های بسیار زیاد یا بسیار کم باشد. برای ارزیابی کلی خوشه‌بندی، میانگین شاخص سیلوئت برای تمام نمونه‌ها محاسبه شد.

شاخص دیویس-بولدین (Davies-Bouldin): این شاخص میزان جدایی بین خوشه‌ها و فشردگی هر خوشه را ارزیابی می‌کند به این ترتیب که برای محاسبه شاخص دیویس-بولدین برای یک روش خوشه‌بندی کافی است ابتدا بیشینه فاصله هر خوشه را نسبت به خوشه‌های دیگر به دست آورد سپس میانگین بیشینه فاصله‌های محاسبه شده برای همه خوشه‌های ایجاد شده توسط الگوریتم را محاسبه کرد. در حقیقت این شاخص، میانگین حداکثر نسبت پراکندگی درون به پراکندگی بین خوشه‌ها را محاسبه می‌کند. هر چه مقدار شاخص کمتر باشد، عمل خوشه‌بندی بهتر صورت گرفته است.

[DOI: 10.34172/jhbmi.2024.14]

نتایج

بررسی نتایج به دست آمده بر روی سه مجموعه داده سلول-منفرد Kolod، Buettner و Usoskin نشان می‌دهد که روش پیشنهادی می‌تواند سلول‌ها را با کیفیت بالایی در یک فضای دو بُعدی با میانگین نرخ NNE= ۰/۸۹ نمایش دهد که نشان دهنده حفظ مناسب ساختار داده‌ها در فضای جدید است. این نگاشت را با استفاده از روش مصورسازی UMAP در حالت‌های مختلف بر روی سه مجموعه داده مورد نظر انجام داده و نتایج مصورسازی مبتنی بر روش پیشنهادی با داده‌های اصلی مقایسه شد. نتایج این مصورسازی در شکل ۱ نمایش داده شده است. در این شکل، هر نقطه بیانگر یک سلول و هر رنگ نشان دهنده یک نوع سلولی می‌باشد.



شکل ۱: مصورسازی مجموعه داده‌های سلول-منفرد با نگاشت آن‌ها در یک فضای دو بُعدی

نتایج ارزیابی کمّی مربوط به کیفیت خوشه‌بندی (معیار NNE) نیز در جدول ۱ نمایش داده شده است. نتایج حاصل نشان می‌دهد که روش پیشنهادی SAMS نه تنها کیفیت خوشه‌بندی سلول‌ها را افزایش می‌دهد بلکه می‌تواند سلول‌ها را با دقت بسیار بالاتری توسط روش‌های مصورسازی نمایش دهد.

جدول ۱: مقادیر مربوط به معیار NNE روش پیشنهادی

مجموعه داده	NNE
Usoskin	۰/۸۷
Buettner	۰/۸۴
Kolod	۰/۹۷

همچنین، نتایج شاخص سیلوئت و شاخص دیویس-بولدین در ارزیابی دقت نهایی خوشه‌بندی، نشان می‌دهد که روش پیشنهادی به ترتیب با میانگین ۰/۶۶ و ۰/۵۰ نسبت به میانگین ۰/۴۸ و ۰/۶۹ داده‌های اصلی، بهبود قابل توجهی در خوشه‌بندی سلول‌ها ایجاد کرده است. نتایج کاملِ مربوط به این شاخص‌ها بر روی مجموعه داده‌های مذکور در قالب جدول ۲ نمایش داده شده است. نتـایج به دست آمده در

[DOI: 10.34172/jhbmi.2024.14]

مجله انفورماتیک سلامت و زیست پزشکی دوره یازدهم، شماره اول، سال ۱۴۰۳، ص ۸۲-۷۲
doi: 10.34172/jhbmi.2024.14
مرکز تحقیقات انفورماتیک پزشکی، دانشگاه علوم پزشکی کرمان

مقایسه با داده‌های اصلی (Original) نشان می‌دهد که روش پیشنهادی به خوبی با کاهش ابعاد داده‌ها توانسته دقت نهایی خوشه‌بندی را بهبود بخشد.

جدول ۲: شاخص سیلوئت و شاخص دیویس-بولدین

مجموعه داده	شاخص سیلوئت (SAMS)	شاخص سیلوئت (Original)	شاخص دیویس-بولدین (SAMS)	شاخص دیویس-بولدین (Original)
Usoskin	۰/۶۲	۰/۴۲	۰/۵۶	۰/۸۱
Buettner	۰/۶۵	۰/۴۱	۰/۵۶	۰/۸۳
Kolod	۰/۷۳	۰/۶۳	۰/۳۸	۰/۴۵

بحث و نتیجه‌گیری

همان‌طور که اشاره شد هدف اصلی روش پیشنهادی در این مطالعه که SAMS نامگذاری شد، کاهش ابعاد داده‌ها و خوشه‌بندی دقیق سلول‌ها به گونه‌ای است که سلول‌های با ویژگی‌های مشابه در یک خوشه قرار گیرند. روش پیشنهادی، ترکیبی از الگوریتم‌های یادگیری عمیق و تکنیک‌های آماری است که برای بهبود دقت و کارایی تحلیل داده‌های سلول-منفرد طراحی شده است.

نتایج نشان داد که روش پیشنهادی SAMS توانسته است سلول‌ها را با کیفیت بالایی در فضای دو بُعدی به تصویر کشیده و خوشه‌بندی کند که نشان دهنده حفظ مناسب ساختار داده‌ها در فضای جدید است [۲۰-۱۷]. همچنین، جهت ارزیابی کمّی مربوط به کیفیت خوشه‌بندی، از معیار NNE استفاده شد که نتایج حاصل باز کیفیت خوشه‌بندی داده‌ها توسط روش پیشنهادی در فضای جدید را تأیید می‌کند [۲۱].

نتایج مربوط به شاخص‌های سیلوئت و دیویس-بولدین برای ارزیابی دقت نهایی خوشه‌بندی نیز نشان دهنده بهبود قابل توجه اِعمال روش پیشنهادی نسبت به داده‌های اصلی است. همان‌طور که اشاره شد شاخص سیلوئت بالاتر نشان‌دهنده خوشه‌بندی بهتر و شاخص دیویس-بولدین پایین‌تر نشان‌دهنده جدایی بهتر بین خوشه‌ها است [۲۳، ۲۲]. نتایج حاصل نشان داد که روش پیشنهادی با بازسازی و کاهش ابعاد مناسب داده‌ها، هم از نظر شاخص سیلوئت و هم از نظر شاخص دیویس-بولدین عملکرد خوبی را ارائه می‌دهد.

کیفیت خوشه‌بندی داده‌های اصلی با داده‌های بازسازی شده توسط روش پیشنهادی مورد ارزیابی قرار گرفت که نشان می‌دهد روش پیشنهادی با بازسازی و کاهش ابعاد مناسب داده‌ها می‌تواند بهبود قابل توجهی را در خوشه‌بندی داده‌ها ایجاد کند. در ادامه روش پیشنهادی SAMS با روش‌های معروفی مثل Seurat، SC3 و SIMLR مورد مقایسه می‌گیرد.

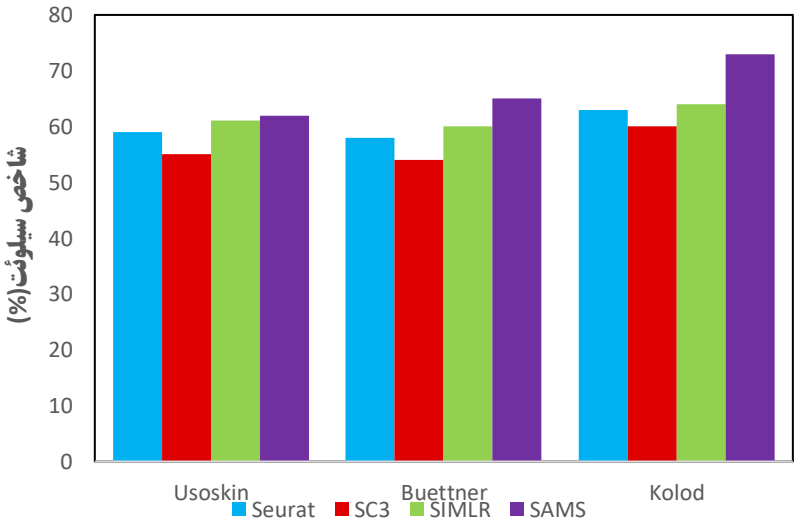
Seurat یک ابزار جامع برای تحلیل داده‌های سلول-منفرد است که به کاربران امکان می‌دهد داده‌ها را پیش‌پردازش، نرمال‌سازی، کاهش ابعاد و خوشه‌بندی کنند. از ویژگی‌های کلیدی Seurat می‌توان به تشخیص ژن‌های با تغییرات بالا، استفاده از روش‌های کاهش ابعاد مانند PCA و UMAP، و الگوریتم‌های خوشه‌بندی مانند Louvain اشاره کرد. این ابزار به طور گسترده‌ای برای شناسایی و تجسم انواع مختلف سلول‌ها در نمونه‌های بیولوژیکی استفاده می‌شود [۲۴].

SC3 یک الگوریتم خوشه‌بندی برای داده‌های سلول-منفرد است که با ترکیب چندین روش خوشه‌بندی به اجماع می‌رسد. این روش شامل نرمال‌سازی، کاهش ابعاد با استفاده از PCA یا t-SNE، و ترکیب نتایج چندین خوشه‌بندی مختلف برای به دست آوردن یک خوشه‌بندی نهایی است. SC3 به دلیل دقت بالا و استفاده از روش‌های متنوع در خوشه‌بندی داده‌ها، یکی از ابزارهای محبوب در تحلیل داده‌های سلول-منفرد است [۱۱].

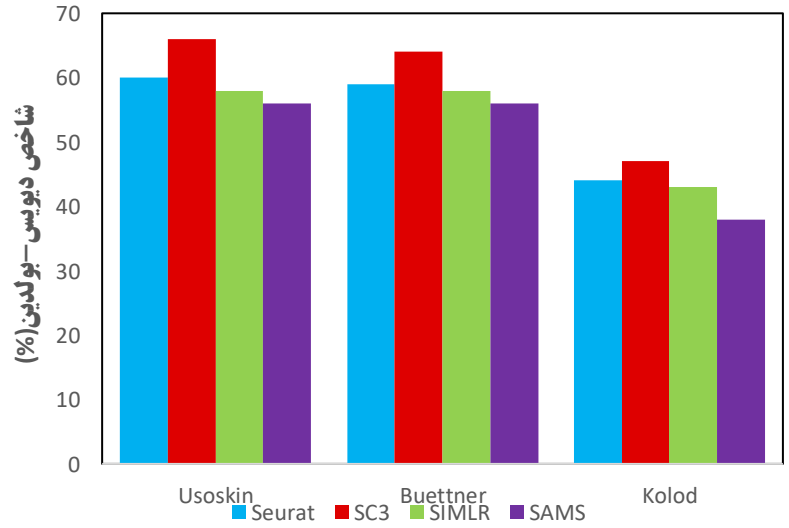
SIMLR نیز یک الگوریتم یادگیری ماشین برای شناسایی و تجسم ساختارهای درونی داده‌های سلول-منفرد است. این روش از یادگیری چند هسته‌ای (multi-kernel learning) برای یادگیری شباهت‌های بین سلول‌ها استفاده می‌کند و به کاهش ابعاد و خوشه‌بندی داده‌ها کمک می‌کند. SIMLR به خاطر توانایی بالا در مدیریت داده‌های پیچیده و نویزی، به ویژه در کاربردهای تجسم و خوشه‌بندی داده‌های سلول-منفرد، مورد توجه قرار گرفته است [۲۵].

برای ارزیابی نهایی و مقایسه دقت خوشه‌بندی حاصل از روش پیشنهادی با سایر روش‌های مورد اشاره نیز از شاخص‌های سیلوئت و دیویس-بولدین استفاده شده است. نتایج مربوطه در قالب نمودارهایی در شکل‌های ۲ و ۳ نمایش داده شده است. همان‌طور که نمودارها

نشان می‌دهد، روش پیشنهادی SAMS نسبت به سایر روش‌ها همواره از دقت بالاتری برخوردار است. به‌عنوان مثال شاخص سیلوئت روش پیشنهادی با میانگین ۰/۶۶ نسبت به بالاترین مقدار شاخص سیلوئت سایر روش‌ها یعنی SIMLR با میانگین ۰/۶۱ با بهبود حدود ۸ درصدی مواجه شده است. از نظر شاخص دیویس-بولدین نیز روش پیشنهادی با میانگین ۰/۵۰ نسبت به بالاترین مقدار شاخص دیویس-بولدین سایر روش‌ها یعنی SIMLR با میانگین ۰/۵۳، بهبود حدود ۶ درصدی حاصل شده است.



شکل ۲: مقایسه شاخص سیلوئت روش‌های مختلف با روش پیشنهادی SAMS



شکل ۳: مقایسه شاخص دیویس-بولدین روش‌های مختلف با روش پیشنهادی SAMS

به طور کلی نتایج به دست آمده نشان دهنده همسو بودن روش SAMS با نتایج پژوهش‌های پیشین است. به طور مثال، Seurat و SIMLR نیز بهبودهایی در خوشه‌بندی سلول‌ها نشان داده‌اند، اما روش SAMS با ترکیب الگوریتم‌های یادگیری عمیق و تکنیک‌های آماری، عملکرد بهتری نسبت به آن‌ها داشته است. در واقع روش SAMS توانسته است ضمن کاهش ابعاد داده‌های اصلی با حفظ ساختار آن‌ها، خوشه‌بندی دقیق‌تری را ارائه دهد. این نتایج همگی نشان از توانایی روش پیشنهادی SAMS در شناسایی ناهمگنی سلولی و ارائه

[DOI: 10.34172/jhbmi.2024.14]

یک تحلیل جامع و دقیق از داده‌های سلول-منفرد حکایت دارد که می‌تواند به محققان در شناسایی بیومارکرهای جدید و درک بهتر دینامیک سلولی کمک کند.

از جمله محدودیت‌های این مطالعه می‌توان نیاز به پردازش بالا و پیچیدگی محاسباتی بالای آن اشاره کرد. همچنین، وابستگی به کیفیت داده‌های اولیه و دقت پیش‌پردازش نیز از دیگر محدودیت‌های این روش می‌باشد که می‌تواند دست‌مایه‌ای برای پژوهش‌ها و کارهای آتی باشد. با این حال، نقاط قوت اصلی روش پیشنهادی در این مطالعه در مقایسه با روش‌های پیشین شامل دقت بالای خوشه‌بندی و همچنین کیفیت بالای نمایش سلول‌ها در فضای دو بُعدی است که می‌تواند به عنوان ابزاری مناسب به محققان در حوزه تحلیل داده‌های سلول-منفرد کمک کند.

تشکر و قدردانی

نویسنده پژوهش بر خود لازم می‌داند از همکاری و مساعدت دانشگاه آزاد اسلامی واحد اندیمشک سپاسگزاری کند.

حمایت مالی

این پژوهش حاصل تحقیق مستقل بدون حمایت مالی می‌باشد.

تعارض منافع

بنابر اظهار نویسنده این مقاله تعارض منافع ندارد.

سهم مشارکت نویسندگان

مقاله یک نویسنده دارد.

References

[1]. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Molecular Cell 2017;65(4):631-43. doi: 10.1016/j.molcel.2017.01.023
[2]. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, et al. Exponential scaling of single-cell RNA-Seq in the past decade. Nat Protoc 2018;13(4):599-604. doi: 10.1038/nprot.2017.149
[3]. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun 2017:8:14049. doi: 10.1038/ncomms14049
[4]. Budnik B, Levy E, Harmange G, Slavov N. Mass-spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. Genome Biol 2018;19(1):161. doi: 10.1186/s13059-018-1547-5
[5]. Specht H, Emmott E, Petelski AA, Huffman RG, Perlman DH, Serra M, et al. Single-cell proteomic and transcriptomic analysis reveals protein-protein interactions and cell heterogeneity. Genome Biol 2021;22: 50. doi: 10.1186/s13059-021-02267-5.
[6]. Baslan T, Hicks J. Genome-wide copy number analysis of single cells. Nat Protoc 2012;7(6):1024-41. doi: 10.1038/nprot.2012.039
[7]. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, McIndoo J, et al. Tumor evolution inferred by single-cell sequencing. Nature 2011;472(7341):90-4. doi: 10.1038/nature09807
[8]. van der Maaten L, Hinton G. Visualizing data using t-SNE. Journal of Machine Learning Research 2008;9(86):2579-605. doi: jmlr.org/papers/v9/vandermaaten08a.
[9]. Abdi H, Williams LJ. Principal component analysis. Wiley Interdisciplinary Reviews: Computational Statistics. 2010;2(4):433-59. https://doi.org/10.1002/wics.101.
[10]. Einipour A, Mosleh M, Ansari-Asl K, A graph-based clustering approach to identify cell populations in single-cell RNA sequencing data. Journal of Health and Biomedical Informatics 2020;7(1):60-72. [In Persian]
[11]. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods 2017;14(5):483-6. doi: 10.1038/nmeth.4236
[12]. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888-902. doi: 10.1016/j.cell.2019.05.031

[13]. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. Pseudo-temporal ordering of individual cells reveals dynamics and regulators of cell fate decisions. Nat Biotechnol 2014; 32(4): 381–6. doi: 10.1038/nbt.2859
[14]. Tirosh I, Venteicher AS, Hebert C, Escalante LE, Patel AP, Yizhak K, et al. Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma. Nature 2016;539(7628):309-13. doi: 10.1038/nature20123
[15]. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science. 2014;344(6190):139-401. doi: 10.1126/science.1254257.
[16]. Vermeulen M, Smith K, Eremin K, Rayner G, Walton M. Application of Uniform Manifold Approximation and Projection (UMAP) in spectral imaging of artworks. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy 2021;252:119547. doi: 10.1016/j.saa.2021.119547.
[17]. Liu S, Maljovec D, Wang B, Bremer P-T, Pascucci V. Visualizing High-Dimensional Data: Advances in the Past Decade. IEEE Trans Vis Comput Graph. 2017;23(3):1249-1268. doi: 10.1109/TVCG.2016.2640960.
[18]. McInnes L, Healy J, Saul N, Groberger L. UMAP: Uniform Manifold Approximation and Projection. J Open Source Softw 2018;3(29):861. doi:10.21105/joss.00861.
[19]. Arbelaitz O, Gurrutxaga I, Muguerza J, Pérez JM, Perona I. An extensive comparative study of cluster validity indices. Pattern Recognit 2013;46(1): 243-56. doi: 10.1016/j.patcog.2012.07.021.
[20]. Vahldiek K, Klawonn F. Cluster-centered visualization techniques for fuzzy clustering results to judge single clusters. Appl Sci 2024;14(3):1102. doi: 10.3390/app14031102.
[21]. Yang J, Lin CT. Multi-View Adjacency-Constrained Nearest Neighbor Clustering (Student Abstract). In Proceedings of the AAAI Conference on Artificial Intelligence 2022; 36(11):13097-8. doi: 10.1609/aaai.v36i11.21685.
[22]. Shutaywl M, Kachoule NN. Silhouette analysis for performance evaluation in machine learning with applications to clustering. Entropy 2021;23(6):759. doi: 10.3390e23060759.
[23]. Xiao J, Lu J, Li X. Davies-Bouldin index based hierarchical initialization K-Means. Intell Data Anal 2017; 21(3):13271338. doi: 10.3233/IDA-163129.
[24]. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol 2018;36(5):411-420. doi: 10.1038/nbt.4096.
[25]. Wang B, Zhu J, Pierson E, Ramazzotti D, Batzoglou S. Visualization and analysis of single-cell RNA-seq data by kernel-based similarity learning. Nat Methods 2017;14(5):414-6. doi: 10.1038/nmeth.4207.

[DOI: 10.34172/jhbmi.2024.14]