A graph-based practice of evaluating collective idendities of cell clusters

Yuji Okano[1,2], Yoshitaka Kase[2], and Hideyuki Okano[2]

[1] Department of Extended Intelligence for Medicine, The Ishii-Ishibashi Laboratory, Keio University School of Medicine
[2] Division of CNS Regeneration and Drug Discovery, International Center for Brain Science, Fujita Health University

February 28, 2024

Abstract

Random sentences

Introduction

It has been more than 10 years since the birth of single-cell RNA-sequencing (scRNA-seq)[1], and the technology now is recognized as a prominent game changer of the molecular biology of this decade. Likewise the pioneering technologies, DNA micro array and bulk RNA-seq, scRNA-seq can observe multidimensional gene expression profiles, while it also can provide such information in single-cell-level. Although this informative assay have contributed to reveal detailed biology of various cell types, the excessive resolution blurred the conceptual boundary between static cell types and transient cell status[2]. Consequently, clusters, chunks of sampels that shares similar geometrical properties in the data space overwrote the classical notion of cell types. As the cell clusters are dependent on sampling stochasticity of the dataset, and their biological properties might sway from the original doctrine of cell types[3]. Hence, a theoretical backbone and a effective method to glue the theory and real data are essential to identify universal characters of specific samples from piles of extrinsic noises.

In our previous research, we proposed a gene regulatory network (GRN)-based representation of cell clusters while edges of GRNs explain statistical dependencies between two genes, and demonstrated that similarity of two clusters can be defined as a quasi-pseudo-metric function d^*[3]. To discuss mathematical properties of the space of cell clusters, we defined novel terms, cell class and eigen-cascades, and step-by-step introduced their algebraic structures. Eigen-cascades refer to a set of marker genes and pairs of genes that are statistically dependent (i.e., isomorphic to the direct sum of the direct sums of the vertex set and the edge set of a GRN). A cell class refers to a cell cluster characterized with the corresponding eigen-cascades. Note that the nuances of cell clusters and cell classes are slightly different even though we might use those terms interchangeably in this article (See Appendix for more descriptions). When two cell classes ${}^{\forall}[x], [y]$ are represented by the GRNs regarding a set of genes G, and the two GRNs (eigen-cascades) respectively denoted as $C_{[x]}(G)$ and $C_{[y]}(G)$, a bivariate function d^* that maps a pair of cell classes to real numbers are defined as follows[3]:

$$d^*([x],[y]) := 1 - \frac{|C_{[x]}(G) \cap C_{[y]}(G)|}{|C_{[x]}(G)|}. \tag{1}$$

Eq. (1) is derived from the Hamming distance function (a metric function that measures the difference of two character strings) and modeified to embrace the tendency of the Peter and Clark (PC) algorithm, one of the most simple bayesian network algorithms[3, 4, 5]. With those concepts, we also proposed frameworks to compare the similarities of given two cell clusters. Our scheme comprises two fundamental steps: formation of GRNs and evaluation of their similarity (Figure 1A). As the concept of cell classes are independent from the choice of data analysis methods, this frameowrk itself can be applied into various cases regardless of any feature engineering (such as the data preprocessing) and clustering methods.

The framework can be applied into the annotation of scRNA-seq data when a referential dataset is available (Figure 1B). As the annotation is the act of tagging clusters with descriptions in natural languages, the biological features of annotated clusters often treated as general and preserved properties of the cell types which the clusters are named after. Accordingly, it is better to have a large enough referential dataset which seem to reflect canonical states of specific cell types which are shared with the query dataset. Using GRN-based characterization, cell classes are annotated with the name of the most similar cell class, however, comparisons of cellular identities can be bidirectional due to asymmetry of d^*. We named the similarity of cell classes from the perspectives of the query data as estimation, and the one from the point of view of the referential data as labeling. Those GRN-based annotations can be visualized with planet plots, where the subjective cell class (here we denote it $[x]$) is located in the center and the radii of the circles reflect $d^*([x], \cdot)$ values for all cell classes placed on the circumferences.

The performance of those frameworks build around GRNs have its bottoleneck in the step to create GRNs, and the process can be broken down into the configuration of the vertex sets and the choice of the network algorithm (Figure 1C). As well as the methods of the feature engineering and the clustering, each step of the GRN formation also has a variety of options. In the last paper, we introduced a combined method of manual curration referring review articles and a machine learning (ML)-based feature selection using a gradient boosting decision tree (GBDT) model with the L1 and the L2 regularizations[3]. For the manual supervision, gene ontology (GO) terms can be another information source. Nevertheless, a priori identification of the sample components are essential to create meaningful GRNs by injecting the domain-specific information. The differentially expressed gene (DEG)-based method can be a more heuristic and a less interactive option because the differential expression analysis (DEA) semi-automatically scoops DEGs. Regarding the network algorithms, we mentioned that there are several possible options.e implemented our codes based on the numerical PC algorithms

⋮

In addition to the theoretical proposal, we applied our logic to the annotation of scRNA-seq data, and showed that the GRN-based annotation can visualize the similarities and the difference of cell clusters, which the conventional DEG-based manual annotation could not address due to the objectives of DEA to summerize the biological semantics of the clusters.

Furthermore, we intended to make our framework dependent on researchers' expertises on the sample domains so that the metrics of cellular identities are tailor-made for the research scopes providing necessary and sufficient resolutions. Constrary, this design made our algorithm unfriendly to users. As the legitimacy our theory needs to be validated in various cases, A semi-automated system to help users select key marker genes is desired.

⋮

Leveraging the backbone theory of GRN-based comparisons of cluster-wise cellular identities (i.e., cell classes), we implemented

⋮

To simplify the contents of this study to highlight our foci, we would not discuss any practices of designing data spaces or clustering in depth.

Results

Challenges of the framework of GRN-based methods

In this reserch, we revisited the workflow of the GRN-based annotation

⋮

To demonstrate those difficulties, we analyzed an open source scRNA-seq data of peripheral blood mononuclear cells from 10X Genomics (for short, we named the dataset PMBC3k in this research).

1. Difficulty of effective gene selection

In our previous study, we proposed a method to combine supervised curration reffering a review paper and a

⋮

The standard workflow consists the quality control (QC) of the data, normalization such as reads per million (RPM) transformation and logarithmic transformation, highly variable gene (HVG) extraction, demensionality reduction such as principal component analysis (PCA), truncated singular value decomposition (TSVD), Uniform

Manifold Approximation and Projection (UMAP)[6], etc., clustering, DEA, annotation, and other downstream analysis (Figure S1).

Considering the fact that our method's primary application is the annotation of scRNA-seq data, we need an alternative method to find marker genes to reduce computational costs and streamline the overall time required to initiate main analyses.

2. Statistical issue: independence v.s. uncorrelation

The statistical independence of two events A and B is defined as a situation where the following equation holds:

$$P(A \cap B) = P(A)P(B), \tag{2}$$

where $P(\cdot)$ is the probability of an event. On the other hand, the correlation coefficient $Corr(X, Y)$ of stochastic variables X and Y is defined as follows:

$$Corr(X, Y) := \frac{Cov(X, Y)}{\sqrt{Var[X]Var[Y]}} = \frac{E[(X - E[X])(Y - E[Y])]}{\sqrt{Var[X]Var[Y]}}, \quad \text{if } Var[X]Var[Y] > 0, \tag{3}$$

where $E[\cdot]$ is the expected value, $Var[\cdot]$ is the variance, and $Cov(\cdot, \cdot)$ is the covariance. Independent variables exhibit a correlation coefficient of zero, the converse is false (e.g., when $X \sim U(-1, 1)$, where $U(-1, 1)$ refers to the uniform distribution over the interval from -1 to 1, $Corr(X, X^2) = 0$ although X and X^2 are dependent). Therefore, strictly speaking, it is not appropriate to substitute the chi squared test or the exact test with the t test of correlation.

However, we introduced a correlation-based algorithm to get GRNs compromising rigor in order to adjust to continuous gene expression values. To address this issue, we need to implement an effective method to binarize the gene expression values so that the new algorithm would rely on the statistical tests of independence. This update would make our algorithm align better to the original concept of our theory.

3. Irresponsibility to gene expression values

To demonstrate this difficulty, we analyzed an open source scRNA-seq data of peripheral blood mononuclear cells from 10X Genomics (for short, we named the dataset PMBC3k in this research).

Automated marker-gene suggestion

Although we intended to require experimenters to curate marker genes to use in GRNs, overly recurrsive trials to find

Dropout-based binarization

With a map $q : \Gamma \times X \to \mathbb{N}$ which returns a raw gene counts of gene ${}^{\forall}g \in \Gamma$ for sample ${}^{\forall}x \in X$ where Γ is the whole set of genes and X is the whole set of samples, here we introduce the coverage function $Coverage_{[x]} : \Gamma \to \mathbb{Q}$ of cell class $[x]$ as follows:

$$Coverage_{[x]}(g) := \frac{|\{x \mid x \in [x] \; s.t. \; q(g, x) \neq 0\}|}{|[x]|} \tag{4}$$

Note that $Coverage_{[x]}$ relies on q only for identifying zeros in raw counts, therefore, any kind of values converted from raw counts by a transformation $\psi : \mathbb{N} \to \mathbb{R}$ such that $\psi^{-1}[\{0\}] = \{0\}$ can be used in lieu of $q(g, x)$. For instance, RPM values and $\log_2(RPM + 1)$ are accepted (see Appendix for more detailed explanations).

Weighted evaluation function

$$d^*([x], [y]) := 1 - \frac{|C_{[x]}(G) \cap C_{[y]}(G)|}{|C_{[x]}(G)|} = 1 - \frac{|E_{[x]}(G) \cap E_{[y]}(G)| + |G|}{|E_{[x]}(G)| + |G|} \tag{5}$$

$$Whqpm([x], [y]) := 1 - \frac{|E_{[x]}(G) \cap E_{[y]}(G)| + \sum_{g \in G} Coverage_{[x]}(g)}{|E_{[x]}(G)| + \sum_{g \in G} Coverage_{[x]}(g)}. \tag{6}$$

Weighted Hamming quasi-pseudo-metric

Discussion

I have no idea.

Methods

GRNet Impletemtations

GO term-assisted gene selection referring Jaccard Index

$$J(A,B) := \frac{A \cap B}{A \cup B} \tag{7}$$

Jaccard Index of two sets A, B is defined as Eq. (7). We expanded this definition to pairwise comparisons of multiple elements by forming a matrix where each element is the corresponding Jaccard Index, and we named the matrix Jaccard index matrix (JIM). For example, the element in i-th row and j-th column (where $i, j, k \in \mathbb{N}$ and $i \leq k, j \leq k$), $JIM_{i,j}$, can be defined as follows when a JIM of sets $X_1, X_2, \cdots, X_k$ are considered:

$$JIM_{i,j} := J(X_i, X_j) \tag{8}$$

Especially for seed markers, sets of subscribed GO terms (let $G_1, \cdots, G_k$) and their JIM are calulated in order to set $min_{i,j}(J(G_i, G_j))$ as a threshold of biological correspondence.

$\vdots$

For detailed method of implementation, we calculated the JIM of the related GO terms of given seed markers. We used mygene.py[7] to query the GO database, and Numpy[8] to calculate JIM.

GRNs and the evaluation function

Following our previous report[3], we computed GRNs by calculating correlations of continuous gene expression values (e.g., $log_2(RPM + 1)$) using Pgmpy[9]. In this study, we introduced

scRNA-seq data analysis

Dataset List

The scRNA-seq data we used in this research were publicly available as online resources as follows:
M1C10X: https://portal.brain-map.org/atlases-and-data/rnaseq/human-m1-10x
hFB: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE165388
PBMC3k: https://support.10xgenomics.com/single-cell-gene-expression/datasets/1.1.0/pbmc3k
aHSPC: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE137864
BCA: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149938

Preprocessing, dimensionality reduction, and visualization

We performed data preprocessing, dimensionality reduction, data visualization of the scRNA-seq datasets using Python packages (including Scanpy[10], Polars, Pandas[11], Numpy, Matplotlib[12], Seaborn[13]) and Julia packages.

Clustering and DE analysis

We performed leiden clustering, DE analysis using Scanpy.

Resource availability

Data availability

Not applicable

Code availability

GRNet and the analysis codes are available on GitHub (https://github.com/yo-aka-gene/grnet). Online documentation for GRNet is also provived (https://grnet.readthedocs.io).

Author contributions

YO designed the project, implemented the algorithms, excecuted the analyses, and wrote the manuscript. YK contributed as the senior author, and edited the manuscript. HO edited the manuscript, and supervised the project.

Acknowledgements

We thank hogehoge for thorough support.

Abbreviations

DEA differential expression analysis

DEG differentially expressed gene

GBDT gradient boosting decision tree

GO gene ontology

GRN gene regulatory network

HVG highly variable gene

JIM Jaccard index matrix

ML machine learning

PC Peter and Clark

PCA principal component analysis

QC quality control

RPM reads per million

scRNA-seq single-cell RNA-sequencing

TSVD truncated singular value decomposition

UMAP Uniform Manifold Approximation and Projection

References

[1] F. Tang, C. Barbacioru, Y. Wang, E. Nordman, C. Lee, N. Xu, X. Wang, J. Bodeau, B. B. Tuch, A. Siddiqui, *et al.*, "mrna-seq whole-transcriptome analysis of a single cell," *Nature methods*, vol. 6, no. 5, pp. 377–382, 2009.

[2] A. Regev, S. A. Teichmann, E. S. Lander, I. Amit, C. Benoist, E. Birney, B. Bodenmiller, P. Campbell, P. Carninci, M. Clatworthy, *et al.*, "The human cell atlas," *elife*, vol. 6, p. e27041, 2017.

[3] Y. Okano, Y. Kase, and H. Okano, "A set-theoretic definition of cell types with an algebraic structure on gene regulatory networks and application in annotation of rna-seq data," *Stem cell reports*, vol. 18, no. 1, pp. 113–130, 2023.

[4] A. Bookstein, V. A. Kulyukin, and T. Raita, "Generalized hamming distance," *Information Retrieval*, vol. 5, pp. 353–375, 2002.

[5] P. Spirtes, C. N. Glymour, and R. Scheines, *Causation, prediction, and search*. MIT press, 2000.

[6] L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[7] C. Wu, A. Mark, and A. I. Su, "Mygene. info: gene annotation query as a service," *bioRxiv*, p. 009332, 2014.

[8] C. R. Harris, K. J. Millman, S. J. Van Der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, *et al.*, "Array programming with numpy," *Nature*, vol. 585, no. 7825, pp. 357–362, 2020.

[9] A. Ankan and A. Panda, "pgmpy: Probabilistic graphical models using python," in *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*, Citeseer, 2015.

[10] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, pp. 1–5, 2018.

[11] W. McKinney *et al.*, "Data structures for statistical computing in python," in *Proceedings of the 9th Python in Science Conference*, vol. 445, pp. 51–56, Austin, TX, 2010.

[12] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[13] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021.

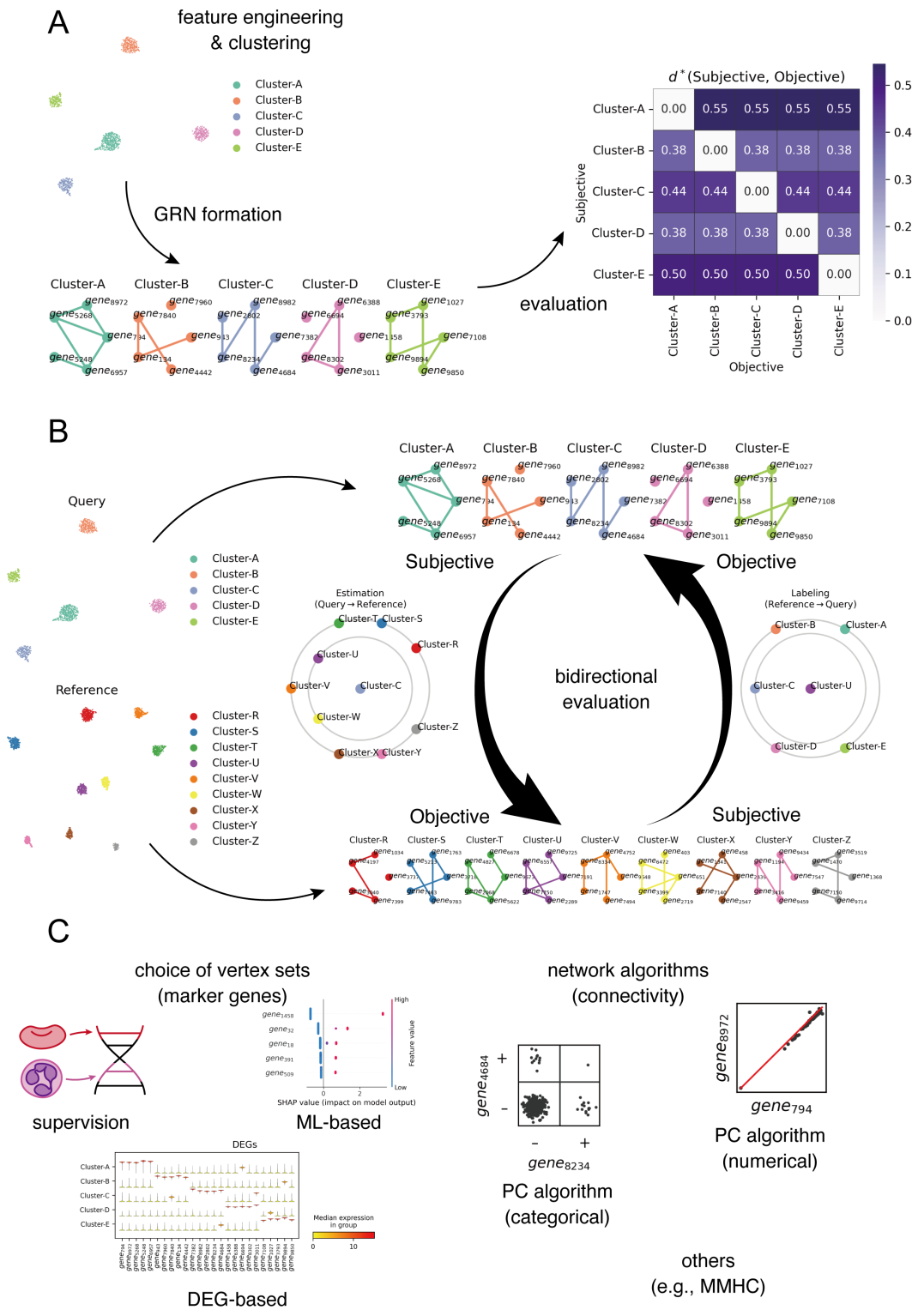
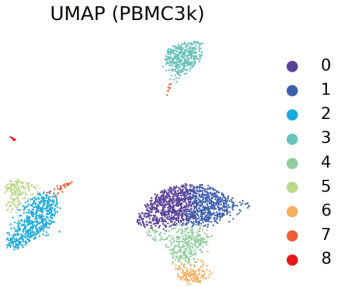


Figure 1: The framework of the GRN-based characterization and annotation of cell classes

A: The foundation of the GRN-based characterization of cell classes. After clustering in designed data space by arbitrary methods, cell classes (the clusters) can be represented by GRNs of corresponding genes of choice. The similarity of two GRNs of the same vertex (marker gene) are evaluated with the assymetrical function d^*, where the return values reflects the similarity from the viewpoint of the subjective clusters. **B**: Schematic of the GRN-based scRNA-seq data annotation. Expecting the referential data to reflect canonical states of target sample domains, the evaluation of the similarity among cell classes can be performed bidirectionally. **C**: Methodological variations of the selection of vertex sets (marker genes) and the algorithms to compute the network structures of GRNs.

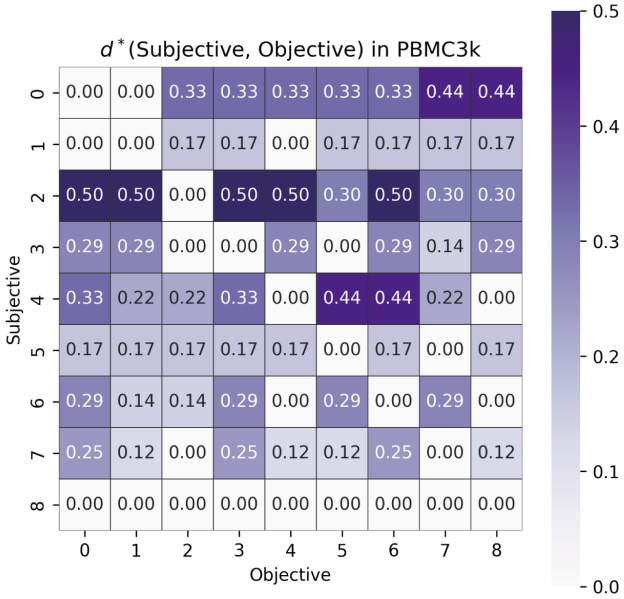
Figure 2

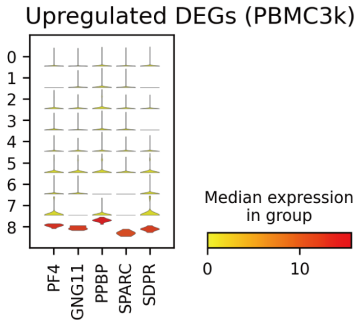
Gene symbols and GO terms



SHAP

time comparison





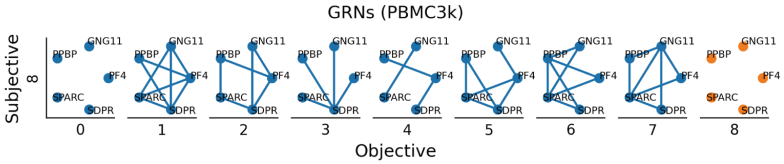


Figure 2: Gene expression patterns of clusters in PBMC3k