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

Yuji Okano[1,2]

[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 (ICBS), Fujita Health University

February 22, 2024

Abstract

Random sentences

Introduction

It has been more than 10 years since the birth of single-cell RNA-sequencing (scRNA-seq), and the technology now is recognized as a prominent game changer of modern molecular biology. Likewise the pioneering technology, bulk RNA-seq, scRNA-seq can observe multidimensional gene expression profiles, while it also can provide such information in single-cell-level.

$\vdots$

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[1].

Results

Quite interesting, while being misterious.

Discussion

I have no idea.

Methods

GO term-assisted gene selection

Although we intended to require experimenters to curate marker genes to use in GRN

Jaccard Index

$$J(A, B) := \frac{A \cap B}{A \cup B} \tag{1}$$

Jaccard Index of two sets A, B is defined as Eq. (1). 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:

$$JIM_{i,j} := J(X_i, X_j) \tag{2}$$

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.

Impletemtation

Querying the GO database by using mygene.py[2], we calculated the JIM of the related GO terms of given seed markers with Numpy[3].

scRNA-seq data analysis

The scRNA-seq data we used in this research were publicly available as online resources.

Preprocessing, Dimensionality Reduction, and Visualization

We performed $\cdots$ using Scanpy[4], Polars.

Abbreviations

GRN gene regulatory network

JIM Jaccard index matrix

scRNA-seq single-cell RNA-sequencing

References

[1] 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.

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

[3] 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.

[4] 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.