Vol.2018-BIO-53 No.11
2018/3/10

スペクトラルグラフ理論を応用した 1細胞RNA-seqデータのトップダウン型クラスタリング

蒲原 純[1]　加藤 有己[1,2,a)]　河原 行郎[1,2]

概要：近年，1細胞単位でのRNAシークエンシング技術（RNA-seq）が確立され，細胞の遺伝子発現レベルがトランスクリプトーム網羅的に説明できるようになった．古典的な教師なし学習により，与えられた細胞集団を発現データに基づき既知の範疇である程度クラスタリングできるものの，1細胞発現データにはノイズやドロップアウトが存在するため，高精度の細胞分類を行うためには高度な数理的手法が必要である．本稿では，1細胞RNA-seqデータに対し，トップダウン型の階層的クラスタリングを行うアルゴリズムを提案する．ここで，スペクトラルグラフ理論を応用し，細胞集団からなる重み付きグラフを2つの同質な部分集団に分割するための最小カットの計算を再帰的に実行する．公開されている1細胞RNA-seqデータセットを用いて提案手法を評価したところ，良いクラスタリング精度を示した．

1. はじめに

現在我々が共有している細胞の種類に関する知識は完全ではない．例えば，ヒトの脳には約1,000億個の神経細胞があるといわれているが，実のところそれらが何種類に分類され，各々がどのような機能を持つかは解明されていない．この問題を解決する方法の1つとして細胞のカタログを作成することが考えられ，これにより脳疾患の治療などに有益なデータが得られると期待されている．

近年，1細胞単位でのRNAシークエンシング技術 (RNA-seq) が確立され，細胞の遺伝子発現レベルがトランスクリプトーム網羅的に説明できるようになった．群平均法などの階層的クラスタリングや主成分分析などの非階層的クラスタリングを含む古典的な教師なし学習で，与えられた細胞集団を遺伝子発現データに基づき，既知の範疇である程度分類できるものの，新規細胞サブタイプの発見という観点では現在黎明期にあたる [2], [5], [7]．これまで，発現データに存在する揺らぎなどのノイズを考慮し，高精度の細胞分類に向けた高度な数理的手法が開発されてきた [3], [8], [10] が，クラスター数をあらかじめ設定するなど，アドホックな手法が多いのも事実である．

本稿では，クラスター数を決定せずに実行可能な，1細胞RNA-seqデータのトップダウン型クラスタリング手法を提案する．

[1] 大阪大学 医学部
Faculty of Medicine, Osaka University, 2–2 Yamadaoka, Suita, Osaka 565–0871, Japan
[2] 大阪大学 大学院医学系研究科
Graduate School of Medicine, Osaka University, 2–2 Yamadaoka, Suita, Osaka 565–0871, Japan
[a)] ykato@rna.med.osaka-u.ac.jp

2. 手法

2.1 スペクトラルグラフ理論による最小カットの計算

与えられた細胞集団を，細胞を頂点に，細胞間の類似度を重みとして辺に持つ重み付き無向グラフによってモデル化することを考える（図1）．このとき，細胞集団を2つの部分集団に分割することは，グラフの最小カットを与える分割を求めることに相当する．最小カットにはいくつかのバリエーションがあるが，ここでは次に定義する正規化カットの概念を用いる．まず，グラフ $G = (V, E)$ の頂点集合 V の2つの部分集合 A, B の間のカットを，A と B を接続する辺の重みの総和として定める．具体的に，頂点 $i, j \in V$ 間の辺の重みを w_{ij} として，カットを

$$Cut(A, B) = \sum_{i \in A, j \in B} w_{ij}$$

によって定義する．この値は A, B の間の辺の重み（2つの集団の類似度）が小さい時に小さくなるが，単純に A または B のいずれかが小さい集合の時にも小さくなり，V の分割としてふさわしい指標とは言えない．そこで，$A, B \in V$ の正規化カットを，

$$Ncut(A, B) = \frac{Cut(A, B)}{Cut(A, V)} + \frac{Cut(A, B)}{Cut(B, V)}$$

として定める．この定義は，$Ncut$ が小さいような $A, B \in V$ はそれら部分集合間での辺の重みは小さく，それぞれの部

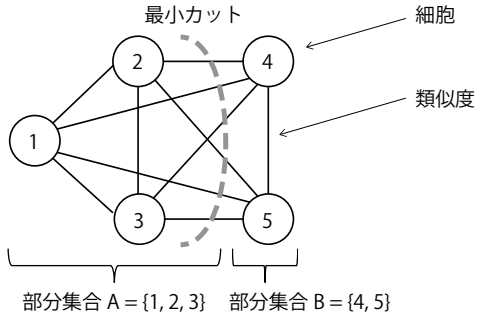

図 1 細胞分類問題のグラフによるモデル化

分集合内部では辺の重みが大きくなることを意味している．

このように考えると，正規化カットが小さくなるような V の分割 $\{A, B\}$ は，モデルの元となった細胞集団の部分集団への分割を表現していると考えられる．したがって，解くべき問題は次のように与えられる:

$$\begin{aligned} &\text{minimize} && Ncut(A, B) \\ &\text{subject to} && V = A \cup B,\ A \cap B = \emptyset. \end{aligned}$$

最小正規化カットを厳密に求めるのは困難と考えられているが，スペクトルグラフ理論を用いると，グラフラプラシアンの固有値問題に帰着させて，近似的に最小正規化カットを求めることが可能である [4], [6], [9]．遺伝子発現カウントとサンプルからなる 2 次元の RNA-seq データに適用する場合，それを重み付き無向グラフに変換してから，最小正規化カットを与える分割を求める．以上を階層的クラスタリングの分岐で用いる 2 分割のアルゴリズムとする．

2.2 2 分割の反復によるトップダウン型クラスタリング

本アルゴリズムでは 2 分割を繰り返すことで，事前にクラスター数を指定することなく細胞集団のクラスタリングを行う．具体的に，2 分割によって得られた部分細胞集団のそれぞれに，あらかじめ設定した最小正規化カットの閾値を超えない限り，再帰的に 2 分割を適用する．これにより，細胞集団の階層的クラスタリングを得ることが可能になる．

2.3 ベンチマークデータ

クラスタリングの性能を評価するベンチマークとして，公開された 1 細胞 RNA-seq 由来の 6 つの FPKM データセットを SCPortalen [1] から取得した．

3. 結果

上述のデータセットを用いて，細胞ラベルとクラスタリング結果を比較した．クラスタリング精度を評価するため，再現率と適合率の調和平均である F スコアを計算し，表 1 のような結果を得た．

表 1 提案手法によるクラスタリング精度評価

データ ID	1	2	3	4	5	6	平均
サンプル数	124	288	340	480	869	1205	
F スコア	0.745	0.533	0.568	0.776	0.752	0.829	0.700

4. おわりに

本稿では 1 細胞 RNA-seq データのトップダウン型クラスタリング手法を提案し，実際の細胞分類データと比較して良いクラスタリング精度が達成されることを示した．今後，より規模の大きなデータやラベル未知のデータに対して提案手法を適用し，適宜改良を行っていく予定である．

参考文献

[1] Abugessaisa, I., Noguchi, S., Böttcher, M., Hasegawa, A., Kouno, T. *et al.* SCPortalen: human and mouse single-cell centric database. *Nucleic Acids Res.* **46,** D781–D787 (2018).
[2] Kolodziejczyk, A.A., Kim, J.K., Svensson, V., Marioni, J.C. & Teichmann, S.A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58,** 610–620 (2015).
[3] Pierson, E. & Yau, C. ZIFA: dimensionality reduction for zero-inflated single-cell gene expression analysis. *Genome Biol.* **16,** 241 (2015).
[4] Shi, J. & Malik, J. Normalized cuts and image segmentation. *IEEE T. Pattern Anal.* **22,** 888–905 (2000).
[5] Stegle, O., Teichmann, S.A. & Marioni, J.C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16,** 133–145 (2015).
[6] Teran Hidalgo, S.J., Wu, M. & Ma, S. Assisted clustering of gene expression data using ANCut. *BMC Genomics* **18,** 623 (2017).
[7] Villani, A.C., Satija, R., Reynolds, G., Sarkizova, S., Shekhar, K. *et al.* Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356,** eaah4573 (2017).
[8] 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* **14,** 414–416 (2017).
[9] Xing, E.P. & Karp, R.M. CLIFF: clustering of high-dimensional microarray data via iterative feature filtering using normalized cuts. *Bioinformatics* **17,** S306–S315 (2001).
[10] Xu, C. & Su, Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics* **31,** 1974–1980 (2015).