Vol.2018-MPS-117 No.1
2018/3/1

Pachinko Allocation Model を用いたクラスタリングによるシングルセル発現解析手法

広田 航[1,a)]　江藤 充宏[2]　瀬尾 茂人[2,b)]　松田 秀雄[2]

概要：近年のシングルセル解析技術の発展により，遺伝子発現プロファイルを用いた細胞のクラスタリングや機能解析が一細胞単位で行われるようになった．このシングルセル遺伝子発現プロファイルは行が遺伝子，列が細胞となる行列であり，行・例ともに数万にものぼるという特徴がある．既存研究では Latent Dirichlet Allocation (LDA) を用いたクラスタリング手法が提案されているが，LDA にはトピック数を増やすとクラスタリングの精度が下がり，またトピックに基づく細胞の機能解析が十分に行えないという問題があった．そこで本研究では，トピック数の増加に対してトピックの推定精度が頑健なトピックモデルである Pachinko Allocation Model を遺伝子発現プロファイルに適用し，細胞のクラスタリングを行った．その結果，大きいトピック数においても高い精度のクラスタリングを実現した．

キーワード：トピックモデル，シングルセル発現解析，クラスタリング

A method for cluster analysis of single-cell gene expression data using the Pachinko Allocation Model

Wataru Hirota[1,a)]　Mitsuhiro Eto[2]　Shigeto Seno[2,b)]　Hideo Matsuda[2]

Abstract: Improving single-cell analysis technologies satisfies our desire to observe the gene expressions of individual cells. We now have single-cell gene expression profiles, large-scaled matrices whose numbers of both rows and columns are often more than 10,000. Many existing studies have attempted to manage such large matrices, conducting analyses such as clustering and functional analysis of cells. A successful approach by these researchers has been to use the Latent Dirichlet Allocation (LDA). However, the more the number of topics of LDA grows, the lower the precision of clustering becomes. Therefore, we propose a clustering method using the Pachinko Allocation Model (PAM), which is a robust model against increasing topics. Here we demonstrate the high performance of the method.

Keywords: topic model, single-cell expression analysis, clusteirng

1. はじめに

生物を構成する基本単位は細胞である．その細胞は同一個体内では全て同じ遺伝情報を持つ．しかし我々の体内を見てもわかるように，これらの細胞は実に多様な働きを見せる．これは細胞の設計図（遺伝情報）が同じであっても，どの部分の設計図（遺伝子）が利用されるかが違うことに起因する．したがって細胞の性質を理解するためには遺伝情報のみならず遺伝子の働き方を調べることが不可欠で

[1] 大阪大学基礎工学部情報科学科
Department of Information and Computer Sciences, School of Engineering Science, Osaka University, Toyonaka, Osaka 560–8531, Japan
[2] 大阪大学大学院情報科学研究科バイオ情報工学専攻
Department of Bioinformatic Engineering, Graduate Schoolof Information Science and Technology, Osaka University, Suita, Osaka 565–0871, Japan
[a)] w-hirota@ist.osaka-u.ac.jp
[b)] senoo@ist.osaka-u.ac.jp

ある．

一細胞単位で遺伝子の発現量を測定できるシングルセル解析技術はこれに大きな進歩をもたらし，近年は細胞の機能解析や擬時間推定，遺伝子制御ネットワークの推定などが盛んに行われるようになった．これらの多くはあらかじめ細胞のクラスタリングを行うため，クラスタリングはこれらの手法の基礎となる．

これまでも細胞のクラスタリング手法は数多く提案されてきた．その 1 つが CellTree[1] である．CellTree はトピックモデルの 1 つである Latent Dirichlet Allocation[2]（LDA）を用いたクラスタリング手法である．CellTree はこの LDA を用いて，クラスタリングを高い精度で実現した上に得られたクラスタの機能解析も可能にした．

しかしその LDA にも，トピック数の増大とともにクラスタリング精度が低下し，クラスタの詳細な機能解析は十分には行えないという問題があった．そこで本研究は，トピック数の増大に対しトピックの推定精度が頑健なトピックモデルである Pachinko Allocation Model[3] を採用し，トピックモデルの利点を活かしつつより詳細な機能解析を可能にする手法を提案する．

本論文は次のような構成を取る．2 章 ではシングルセル解析に着目して，その発現およびデータの解析について述べる．3 章 では提案手法について説明する．4 章 では提案手法の有用性を確かめるための実験とその結果について述べる．5 章ではまとめと今後の課題について述べる．

2. シングルセル発現解析

本章ではまず遺伝子の発現について述べ，次に遺伝子の発現データを用いて行われているクラスタリングについて述べる．

2.1 遺伝子の発現

DNA（DeoxyriboNucleic Acid; デオキシリボ核酸）は生物の遺伝情報を保持する物質である．DNA は A（アデニン），G（グアニン），C（シトシン），T（チミン） の 4 種類のヌクレオチドがなす配列によってその情報を保持する．この配列を塩基配列，または単に配列と呼ぶ．

DNA 配列はその一部が生体内で動的に転写され，タンパク質の合成など様々な用途に用いられる．この転写される DNA の配列の種類のことを遺伝子と呼ぶ．DNA から転写された遺伝子は RNA（RiboNucleic Acid; リボ核酸）として生体内に存在する．遺伝子が DNA から転写され RNA が生成される過程を遺伝子の発現（expression）という．一般的に同じ DNA からでも異なる量や種類の遺伝子が生じる．これが同じ DNA を持つ同一個体内の細胞に多様性が生じる一因である．

遺伝子がどの程度発現したか，定量化したものが遺伝子発現量である．遺伝子発現量は単に発現量とも呼ばれる．また，細胞 j の遺伝子 i の発現量を (i, j) 要素に持つ行列を遺伝子発現プロファイル，または発現プロファイルと呼ぶ．

提案手法では発現量として Unique Molecular Identifier[4] count（UMI count）を採用したものを対象とする．UMI は 1 つ 1 つが異なる（unique な）配列を持った分子のことで，これを RNA の末尾に付加することで RNA の識別子として利用できる．発現した全ての RNA に UMI を付与することで，ある細胞から出た UMI を RNA ごとに分類してそれぞれ数を数えたものを，その遺伝子の発現量と考えることができる．これを UMI count と呼ぶ．UMI count は，その定義から非負整数となる．

2.2 クラスタリング

シングルセル発現プロファイルを用いた細胞のクラスタリング手法はこれまで多く提案されている．以下ではまずその従来手法の多くが行っている次元削減に関して，その必要性や問題点を述べる．次にクラスタリングの従来手法である k-means 法や pcaReduce[5]，DIMM-SC[6]，SC3[7] について述べる．

2.2.1 クラスタリングと次元削減

シングルセル発現プロファイルには，行数・列数がともに大きいという特徴や疎行列であるという特徴がある．したがって，その解析は高次元の遺伝子次元空間から効果的に低次元の特徴空間へと次元削減を行う方法と組み合わせて行われることが多い．

次元削減で広く用いられている手法は主成分分析である．しかし発現プロファイルに主成分分析を適用することにはいくつかの問題がある．まず，発現量の分布は正規分布に従わないという問題がある．主成分分析は標本分散を最大化する次元を選択するため，データが正規分布に従うことを暗に仮定している．しかし UMI count によって発現量を定義した場合，発現量は 0 の多い非負整数となるためこの仮定が成立しない．また主成分分析は負の因子負荷量の生物学的な解釈が難しいという問題もある．

2.2.2 クラスタリングの従来手法

2.2.2.1 pcaReduce

pcaReduce は，主成分分析による次元削減と k-means 法によるクラスタリングを交互に行うクラスタリング手法である．pcaReduce は $q+1$ 次元にまで削減した次元で q 個のクラスタへ分割する．q は任意である．次に q を 1 つ減らして，最も近接するクラスタ同士を結合する．これを q（クラスタ数）が指定した数になるまでくり返す．

2.2.2.2 DIMM-SC

Dirichlet mixture model for clustering droplet-based single cell（DIMM-SC）は，混合ディリクレ分布に基づいたクラスタリング手法である．DIMM-SC では，ある細胞から発現する遺伝子の種類やその発現量は，細胞 1 つ 1

つが持つ潜在変数である「クラス」から確率的に生成されたものとみなす．そのクラスの分布から発現プロファイルが生成される確率の算出と，それを最大化するクラス分布の更新を EM アルゴリズムを用いてくり返すことでクラスタリングを行う．

2.2.2.3 SC3

Single-Cell Consensus Clustering（SC3）は，複数のクラスタリング方法を組み合わせたパイプライン形式のクラスタリング手法である．SC3 は 3 種類の距離尺度と 2 種類の次元削減手法を組み合わせた計 6 通りの方法で細胞間の距離を計算し，それぞれクラスタリングを行う．その後，それらのクラスタリング結果を統合して最終的なクラスタリングを行う．

2.3 トピックモデルを用いたクラスタリングの従来手法

2.2.1 で述べた主成分分析の問題を解決する 1 つの手法として，LDA を用いたクラスタリングの手法である CellTree が提案されている．以下では LDA と CellTree について述べたあと，LDA の問題点について述べる．

2.3.1 Latent Dirichlet Allocation

LDA は代表的なトピックモデルの 1 つである．トピックモデルは主に自然言語処理の分野で用いられる文章の生成モデルである．トピックモデルは，文章を単語の重複集合（Bag of Words）とみなし，それぞれの単語は文章がそれぞれ持つ「トピック」の分布にもとづいて生成されるとみなす．トピックは単語の共起関係にもとづいて推定されるもので，それぞれ単語の出現確率を持つ．例えばある文章は「スポーツ」というトピックが多くを占める場合，その文章はスポーツトピックからの出現確率が高い「サッカー」や「野球」といった単語が多く出現する[*1]．

LDA がある細胞の遺伝子を発現させる過程を **図 1** に示す．LDA は，文章を細胞，単語の出現頻度を遺伝子の発現量と置き換えて考えることで発現プロファイルにも用いることができる．この場合，細胞は遺伝子の重複集合（Bag of Genes）とみなされ，それぞれの細胞はトピックの分布からトピックを生成し，生成されたトピックはそれぞれ遺伝子を発現させる．

本研究の文脈におけるトピックとは，細胞周期に関わる遺伝子や分化の制御を行う遺伝子など，機能的に関連のある遺伝子の集合であることが期待される．なぜなら各トピックは遺伝子発現の共起関係を抽出したものだからである．したがって，トピックに関係のある遺伝子は高い発生確率を持ち，関係しない遺伝子は発生確率が 0 となるため，2.2.1 で述べた主成分分析の問題点である解釈の困難さを解決する．

LDA のグラフィカルモデルを **図 2** に示す．図 2 におけ

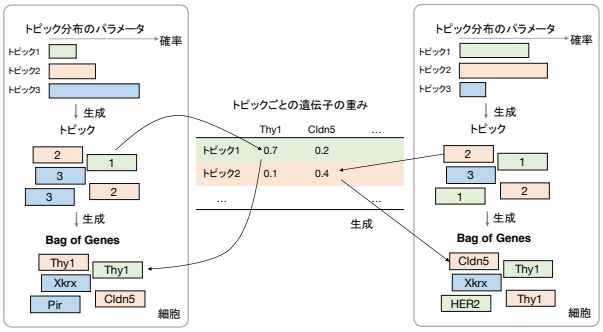

図 1 LDA が発現量を生成する過程
Fig. 1 Process of generating gene expressions in LDA

る各パラメータは以下に示す分布から生成される．細胞 j のトピック分布 $\boldsymbol{\theta}_j$ とトピック k の遺伝子の出現確率 $\boldsymbol{\phi_k}$ はそれぞれディリクレ分布 $Dir(\boldsymbol{\alpha})$，$Dir(\boldsymbol{\beta})$ に従う（$\boldsymbol{\alpha}$，$\boldsymbol{\beta}$ はハイパーパラメータ）．細胞 j の各トピックの遺伝子の出現確率は多項分布 $Mult(\boldsymbol{\theta}_j)$ に従う．

2.3.2 CellTree

CellTree は LDA を次元削減手法に用いたクラスタリング手法である．LDA によって得られた細胞のトピック次元はカイ二乗距離[*2] を距離尺度とした階層型クラスタリングに用いられる．

2.3.3 LDA の問題点

数万種類の遺伝子からなる発現プロファイルの特徴次元をいくつのトピックまで削減するかという問題は，データにも依存するため難しい．ただしトピックを数個にまで削減することは，仮に細胞のクラスタリングは十分に行うことができたとしても，遺伝子集団を「モレなくムダなく」機能解析するためには少なすぎると考えられる．なぜなら，トピック数が小さい場合は「ムダ」は少なくとも「モレ」が多くなってしまうからである．つまり，トピック数が小さい場合は全てのトピックが何らかの機能を持つことはあっても，トピックに関連付けられる機能数が少なくなるからである．

しかし，LDA では推定するトピック数を増やした場合にトピックの推定精度が低下するという問題が知られている [8]．そのため，LDA は詳細な遺伝子機能解析のためにトピック数を増やすことは困難である．

3. 提案手法

本章ではまず提案手法の概要を述べたあと，提案手法で用いる Pachinko Allocation Model（PAM）と階層型クラスタリングについて述べる．

3.1 提案手法の概要

提案手法では PAM を利用して次元削減を行う．PAM

[*1] ただし，トピックモデルでは「スポーツ」のようなトピックの意味は推定できないことに注意されたい．

[*2] $d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}}$ で定義される．

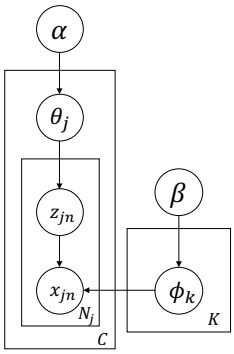

図 2 LDA のグラフィカルモデル

Fig. 2 Graphical model representation of LDA

図中のパラメータの意味は次のとおりである．$\boldsymbol{\alpha}$ は $\boldsymbol{\theta}_j$ のハイパーパラメータ．$\boldsymbol{\theta}_j$ は 細胞 j のトピック分布のパラメータ．z_{jn} は 細胞 j から n 番目に生成された遺伝子のトピック番号．x_{jn} は z_{jn} から生成された遺伝子．$\boldsymbol{\beta}$ は $\boldsymbol{\phi}_k$ のハイパーパラメータ．$\boldsymbol{\phi}_k$ は トピック k の遺伝子の出現確率．C は 細胞数．N_j は 細胞 j の遺伝子数．K は トピック数（事前に指定する）．

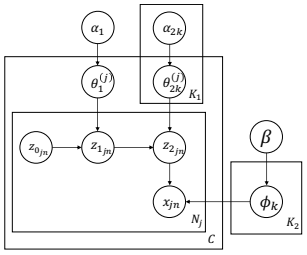

図 3 PAM のグラフィカルモデル

Fig. 3 Graphical model representation of PAM

図中のパラメータの意味は次のとおりである．$\boldsymbol{\alpha}_1$ は $\boldsymbol{\theta}_1^{(j)}$ のハイパーパラメータ．$\boldsymbol{\alpha}_{2k}$ は $\boldsymbol{\theta}_{2k}^{(j)}$ のハイパーパラメータ．$\boldsymbol{\theta}_1^{(j)}$ は 細胞 j の上位トピック分布．$\boldsymbol{\theta}_{2k}^{(j)}$ は 細胞 j の上位トピックが k である場合の下位トピック分布．$z_{0_{jn}}$ は 細胞 j の n 番目の遺伝子のルートトピック．（便宜上配置されるもので，実際には 1 つの値しか取らない）．$z_{1_{jn}}$ は $z_{0_{jn}}$ から生成された上位トピック．$z_{2_{jn}}$ は $z_{1_{jn}}$ から生成された下位トピック．x_{jn} は z_{jn} から生成された遺伝子．$\boldsymbol{\beta}$ は $\boldsymbol{\phi}_k$ のハイパーパラメータ．$\boldsymbol{\phi}_k$ は 下位トピック k の遺伝子の出現確率．C は 細胞数．N_j は 細胞 j の遺伝子数．K_1 は 上位トピック数（事前に指定）．K_2 は 下位トピック数（事前に指定）．

はトピック間の関連を考慮するため，2.3.3 で述べたトピック数を増やした時にトピックの推定精度が下がる問題を解決し，機能解析における網羅性の向上が見込まれる．PAM で得られた次元は CellTree における LDA と同様に細胞の特徴量として階層型クラスタリングに用いられる．階層型クラスタリングの距離尺度にはユークリッド法を，結合法には Ward 法をそれぞれ用いる．

3.2 Pachinko Allocation Model

PAM はトピックと遺伝子を有向グラフで表現するトピックモデルである．本来 PAM は任意の DAG で表現されたトピックモデルを指すが，提案手法では特にルートトピック層・上位トピック層・下位トピック層・遺伝子層の 4 層からなる 4-level PAM を用いる．ただしルートトピックは便宜上配置しているものであり，実際は全ての細胞の全ての遺伝子が同一の値を取る．以降，とくに断りがなければこのモデルを PAM と呼ぶ．PAM のグラフィカルモデルを **図 3** に示す．

3.3 クラスタリング

提案手法では，細胞の特徴量はその細胞の下位トピックの出現確率とする．細胞 j の下位トピックの出現確率 $\boldsymbol{R}_j$ を式 (1) で定義する．これは K_2 次元上のベクトルとなる．

$$\begin{aligned} R_{jl} &= \sum_{k=1}^{K_1} \theta_{2kl}^{(j)} \\ \boldsymbol{R}_j &= (R_{j1}, \ldots, R_{jK_2}) \end{aligned} \tag{1}$$

その上で，細胞間の距離はユークリッド距離で定義する．したがって細胞 j と j' の距離は (2) に示す式で定義される．

$$d(j, j') = \sqrt{\sum_{l=1}^{K_2} (R_{jl} - R_{j'l})^2} \tag{2}$$

クラスタの結合法には Ward 法を用いる．Ward 法では，全てのクラスタ A, B 間の距離 $d(A, B)$ を 式 (3) で定義し，最も距離が小さいクラスタ同士を結合する．これを与えられたクラスタ数になるまでくり返す．

$$\begin{aligned} d(A, B) &= E(A \cup B) - E(A) - E(B) \\ \text{where } E(X) &= \frac{1}{|X|} \sum_{\boldsymbol{x} \in X} ||\boldsymbol{x} - \boldsymbol{g}_X||^2 \end{aligned} \tag{3}$$

ただし，$|X|$ は集合 X の要素数，$||\cdot||$ は L^2 ノルム，$\boldsymbol{g}_X$ は X の重心である．

4. 実験と考察

本章では提案手法を検証する実験について述べる．実験 1 では，提案手法のトピック数の評価を行った．そのあと，実験 2 で提案手法のクラスタリング性能を従来手法と比較した．

クラスタリングの性能評価においては Klein[9], Zeisel[10], Zheng[11] の 3 種類のデータセットを用いた．ただし Zheng は，生物学的に分類が易しい Zheng-simple と，Zheng-simple より分類が難しい Zheng-challenging の 2 種類のデータセットを使用した．Klein，Zeisel，Zheng-simple，Zheng-challenging は生物学的な意味からそれぞれ 4 クラス，7 クラス，3 クラス，3 クラスに分類される．

4.1 実験 1. トピック数の評価

LDA や PAM においてトピック数の推定は重要かつ難しい問題である．本実験では上位トピック数の推定には平

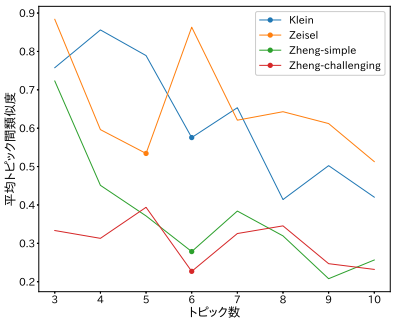

図 4 トピック数ごとの平均トピック間類似度の推移
Fig. 4 The average of topic similarities of LDA models

均トピック間類似度を用いた [12]．この手法においては平均トピック間類似度が小さいほど各トピックが独立している，すなわち各トピックの意味が明瞭であると考えられる．本実験では，平均トピック間類似度が最初に極小値を取るようなトピック数を PAM の上位トピック数として採用する．最小値ではなく最初の極小値を取る理由は，上位トピック数が大きくなると PAM のパラメータ推定にかかる計算時間が大きくなるという問題があるためである．

トピック $\mathrm{T_i}$，$\mathrm{T_j}$ 間の類似度の定義を式 (4) に示す．

$$\mathrm{sim}(\mathrm{T_i}, \mathrm{T_j}) = \frac{\sum_{g=1}^{G} \phi_{ig}\phi_{jg}}{\sqrt{\sum_{g=1}^{G}(\phi_{ig})^2 \sum_{g=1}^{G}(\phi_{jg})^2}} \qquad (4)$$

ただし，G は遺伝子数である．モデルの平均トピック間類似度は 式 (4) で算出した値の平均値である．

各トピック数で算出した平均トピック間類似度を **図 4** に示す．この結果より，PAM での上位トピック数は Klein で 6，Zeisel で 5，Zheng-simple で 6，Zheng-challenging で 6 と推定された．このトピック数を次に示す実験 1.2 で用いる．

また，PAM の下位トピック数を 30 から 60 の範囲で 10 ずつ変動させ，同じ数のトピック数の CellTree と 性能を比較した結果を **図 5** に示す．性能の評価指標には Adjusted Rand Index[13]（ARI）を用いた．ARI は 2 つの分布の類似度をはかる指標で，ARI が大きいほど 2 つの分布は類似しているとみなせる．ARI の最大値は 1 である．本実験では先に述べた正解付きのデータセットを用いて，この正解の分布とクラスタリング結果の分布を ARI によって比べた．すなわち，ARI の値が高いほどクラスタリングの性能が高いとみなせる．

図 5 の結果より次の 2 点がいえる．1 つ目は，いずれのデータセット・トピック数においても、提案手法の ARI は CellTree を上回っていることである．これはシングルセル解析においても CellTree で大きなトピック数を推定することは困難であることを示している．2 つ目は，提案手法は下位トピック数を変化させてもクラスタリング性能に大

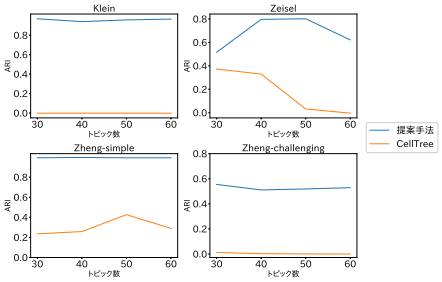

図 5 CellTree と提案手法の ARI の比較
Fig. 5 ARI between CellTree and the proposed method

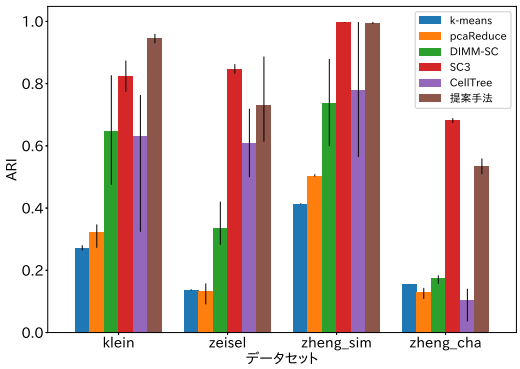

図 6 従来手法と提案手法の ARI
Fig. 6 ARI among existing methods and the proposed method

きな違いが見られないことである．したがって実験 1.2 で用いる提案手法では，全てのデータセットで PAM の下位トピック数を 50 と定めた．

4.2 実験 2. 提案手法と従来手法のクラスタリング性能の比較

本実験では従来手法と提案手法のクラスタリング性能を実験 1.1 と同じく ARI によって評価した．従来手法および提案手法の ARI を **図 6** に示す．ただし，従来手法としては 2.2.2 で挙げた 4 手法と k-means 法，CellTree を用いた．

本実験の実験条件を以下に示す．本実験ではクラスタリングを各手法・データセットごとに 4 回ずつ実行して，それらの ARI の平均値を棒グラフの値に採用した．またこれら 4 回の ARI の最大値・最小値を棒グラフ上の黒の実線で示す．各手法の設定を以下に示す．

k-means 重心の計算方法には算術平均を採用した．クラスタの割り付け回数は 300 回とした．

pcaReduce 次元数の初期値を 100 と定めた．近接するクラスタの結合法はデフォルト設定のメソッド S[5] を用いた．

DIMM-SC 設定はデフォルトの設定を用いた．

SC3 各設定はデフォルトの設定を用いた．遺伝子フィル

タリング，および SVM の実行は行っていない．

CellTree パラメータの推定には MAP 推定 [14] を用いた．トピック数は実験 1.1 で推定した値を用いた．

提案手法 PAM のギブスサンプリングのくり返し回数を 10,000 回とした．上位トピック数，下位トピック数はいずれも実験 1.1 で推定した値を用いた．

実験 1.2 において注目すべき結果は次の 2 点である．1 つ目は，この結果より，いずれのデータセットに対しても CellTree より提案手法のほうが平均 ARI が高いことである．したがってクラスタリング手法としては CellTree よりも提案手法の方が適していると考えられる．2 つ目は，Klein に対しては提案手法の ARI が最も高かったのに対し，SC3 は Zeisel，Zheng-simple，Zheng-challenging の 3 つのデータセットに対して最も ARI が高かったことである．ただし提案手法は SC3 に比べて，機能解析が行いやすいという利点がある．また，提案手法はクラスタリング手法としてユークリッド距離による階層型クラスタリングという単純な手法を用いているが，これをより最適な手法に変更することで分類精度が高まる可能性がある．

5. おわりに

本論文ではシングルセル発現解析のための PAM を用いた細胞のクラスタリング手法を提案した．本論文で示した実験により，提案手法は高い分類精度でクラスタリングが可能であること，また機能解析の精度も既存手法より優れていることが確認された．また，本論文で示した実験はすべて性質が既知であるデータセットに対して行われたが，既知のクラスや機能を推定できることは未知のサブクラスや機能を解析する際にも有用であることを意味すると考えられる．

本研究の今後の課題としてはトピック数の推定方法が挙げられる．本論文の実験では上位トピック数の推定方法としてトピック間類似度を用いたが，これは LDA のトピック数推定の方法として提案されたものであり，必ずしも PAM で最適であるとは限らない．これについては，Hierarchical Dirichlet Process のような方法を用いて PAM のトピックを自動推定するという改善策が考えられる．

謝辞 本研究は JST CREST JPMJCR15G1 の支援と，JSPS 科研費 JP15K00403，JP16K12525 の助成を受けたものです．

参考文献

[1] DuVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H. and Tsuda, K.: CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-seq data, *BMC Bioinformatics*, Vol. 17, No. 1, p. 363 (2016).

[2] Blei, D. M., Edu, B. B., Ng, A. Y., Edu, A. S., Jordan, M. I. and Edu, J. B.: Latent Dirichlet Allocation, *Journal of Machine Learning Research*, Vol. 3, pp. 993–1022 (2003).

[3] ACM: *Pachinko allocation: DAG-structured mixture models of topic correlations* (2006).

[4] Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S. and Taipale, J.: Counting absolute numbers of molecules using unique molecular identifiers, *Nature Methods*, Vol. 9, No. 1, pp. 72–74 (2011).

[5] Žurauskienė, J. and Yau, C.: pcaReduce: hierarchical clustering of single cell transcriptional profiles, *BMC Bioinformatics*, Vol. 17, No. 1, p. 140 (2016).

[6] Sun, Z., Wang, T., Deng, K., Wang, X.-F., Lafyatis, R., Ding, Y., Hu, M. and Chen, W.: DIMM-SC: a Dirichlet mixture model for clustering droplet-based single cell transcriptomic data, *Bioinformatics* (2017).

[7] Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R. and Hemberg, M.: SC3: consensus clustering of single-cell RNA-seq data, *Nature Methods*, Vol. 14, No. 5, pp. 483–486 (2017).

[8] Blei, D. M. and Lafferty, J. D.: Correlated topic models, *Proceedings of the 18th International Conference on Neural Information Processing Systems*, MIT Press, pp. 147–154 (2005).

[9] Klein, A. M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D. A. and Kirschner, M. W.: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells., *Cell*, Vol. 161, No. 5, pp. 1187–1201 (2015).

[10] Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C. et al.: Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq, *Science*, Vol. 347, No. 6226, pp. 1138–1142 (2015).

[11] Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J. et al.: Massively parallel digital transcriptional profiling of single cells, *Nature communications*, Vol. 8, p. 14049 (2017).

[12] Cao, J., Xia, T., Li, J., Zhang, Y. and Tang, S.: A density-based method for adaptive LDA model selection, *Neurocomputing*, Vol. 72, No. 7-9, pp. 1775–1781 (2009).

[13] Hubert, L. and Arabie, P.: Comparing partitions, *Journal of classification*, Vol. 2, No. 1, pp. 193–218 (1985).

[14] Taddy, M.: On estimation and selection for topic models, *International Conference on Artificial Intelligence and Statistics*, pp. 1184–1193 (2012).