Vol.2021-MPS-133 No.25
Vol.2021-BIO-66 No.25
2021/6/30

シングルセル遺伝子発現プロファイル解析へのテンソル分解を用いた教師なし学習による変数選択法の応用

田口 善弘[1,a)]　ターキー　ターキー[2]

概要：シングルセルの遺伝子発現プロファイル解析は分子生物学の重要な研究手段となっているが、その研究を行うことはなかなか難しい状況になっている。特に細胞ごとのラベル情報が不足しているため、従来の教師あり学習をそのまま適用することが難しい。本研究ではテンソル分解を用いた教師なし学習による変数選択法を用いてヒトとマウスの中脳の発生過程の一細胞遺伝子発現プロファイルの統合解析を目指す。

Application of tensor decomposition based unsupervised feature extraction to single cell gene expression profile analysis

1. はじめに

一細胞遺伝子発現プロファイル解析 (single cell RNA-seq, 以下 scRNA-seq) は分子生物学における重要な研究手段となっている。しかし、細胞ごとのラベリングが不足しているため、効果的な研究が難しい状態にある。本研究ではテンソル分解を用いた教師なし学習による変数選択法 [2] を用いて、ヒトとマウスの中脳の発生過程の scRNA-seq データを解析した結果 [1] を報告する。本報は第 57 回 BIO 研究発表会で発表した「主成分分析を用いた教師無し学習による変数選択の一細胞 RNA-seq への応用」 [3] で解析したデータを再度「テンソル分解を用いた教師なし学習による変数選択法」で解析し直した研究である。

2. 方法と材料

2.1 scRNA-seq データ

Gene Expression Omnibus (GEO) の ID GSE76381 から “GSE76381_EmbryoMoleculeCounts.cef.txt.gz”(ヒト) と “GSE76381_MouseEmbryoMoleculeCounts.cef.txt.gz”(マウス) の 2 つのファイルをダウンロードした。これらは各々中脳の発生過程のデータであり、ヒトの胚の腹側中脳細胞（妊娠後 6 週間が２８７細胞、７週間が１３１細胞、８週間が３３１細胞、９週間が３２２細胞、１０週間が５０９細胞、１１週間が３９７細胞、計１９７７細胞）とマウスの E11.5 から E18.5 までの期間の 6 時点での腹側中脳細胞（E11.5 が３４９細胞、E12.5 が３５０細胞、E13.5 が３４５細胞、E14.5 が３０８細胞、E15.5 が３５６細胞、E18.5 が１４２細胞、時期不明が５７細胞、の計１９０７細胞）からなる。

[1] 中央大学
Chuo University, Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
[2] キング・アブドゥルアズィーズ大学
King Abdulaziz University
[a)] tag@granular.com
本研究は原著論文として刊行済みである [1]。

2.2 テンソル分解を用いた教師なし学習による変数選択法

遺伝子 i を共有する 2 つの遺伝子発現プロファイル $x_{ij} \in \mathbb{R}^{N\times M}$ と $x_{ik} \in \mathbb{R}^{N\times K}$ があるとする（j,k はそれぞれ、サンプルのインデックス)。これらを統合解析するために

$$x_{jk} = \sum_i x_{ij}x_{ik} \in \mathbb{R}^{M\times K} \tag{1}$$

を計算する。これに対して特異値分解 (singular value decomposition, SVD) を適用し

$$x_{jk} = \sum_\ell \lambda_\ell u_{\ell j} v_{\ell k} \tag{2}$$

を得る。ここで $u_{\ell j} \in \mathbb{R}^{M\times M}, u_{\ell k} \in \mathbb{R}^{K\times K}$ は特異値ベクトルである。

これからサンプルに付与される特異値ベクトル

$$u^{(j)}_{\ell i} = \sum_j x_{ij}u_{\ell j} \tag{3}$$

ⓒ 2021 Information Processing Society of Japan

Vol.2021-MPS-133 No.25
Vol.2021-BIO-66 No.25
2021/6/30

$$u_{\ell i}^{(k)} = \sum_k x_{ik} v_{\ell k} \tag{4}$$

を計算する。

次に遺伝子選択に用いる $u_{\ell i}^{(j)}$ と $u_{\ell i}^{(k)}$ を選択する。このためには $u_{\ell j}, v_{\ell k}$ に戻って、興味がある j,k 依存性のある $u_{\ell j}, v_{\ell k}$ を選択し、対応する $u_{\ell i}^{(j)}, u_{\ell i}^{(k)}$ を用いて、遺伝子選択を行う。ここでは $u_{\ell i}^{(j)}, u_{\ell i}^{(k)}$ がガウス分布に従っているという帰無仮説を用いて、

$$\langle u_{\ell i}^{(j)} \rangle = \frac{1}{N} \sum_i u_{\ell i}^{(j)} \tag{5}$$

$$\langle u_{\ell i}^{(k)} \rangle = \frac{1}{N} \sum_i u_{\ell i}^{(k)} \tag{6}$$

$$\sigma_\ell^{(j)} = \sqrt{\frac{1}{N} \sum_i \left(u_{\ell i}^{(j)} - \langle u_{\ell i}^{(j)} \rangle \right)^2} \tag{7}$$

$$\sigma_\ell^{(k)} = \sqrt{\frac{1}{N} \sum_i \left(u_{\ell i}^{(k)} - \langle u_{\ell i}^{(k)} \rangle \right)^2} \tag{8}$$

$$P_i^{(j)} = P_{\chi^2} \left[> \sum_\ell \left(\frac{u_{\ell i}^{(j)} - \langle u_{\ell i}^{(j)} \rangle}{\sigma_\ell^{(j)}} \right)^2 \right] \tag{9}$$

$$P_i^{(k)} = P_{\chi^2} \left[> \sum_\ell \left(\frac{u_{\ell i}^{(k)} - \langle u_{\ell i}^{(k)} \rangle}{\sigma_\ell^{(k)}} \right)^2 \right] \tag{10}$$

のように P 値を遺伝子 i に付与する。$P_{\chi^2}[> x]$ は累積カイ二乗分布で値が x 以上の確率、$\sigma_\ell^{(j)}, \sigma_\ell^{(k)}$ は標準偏差、和は興味がある j,k 依存性のある $u_{\ell j}, v_{\ell k}$ を選択した時に選んだ ℓ についてのみ取る。P 値は BH 法 [2] で多重比較補正し、補正 P 値が 0.01 以下の遺伝子を選択する。

2.3 結果

ヒトとマウスの共通遺伝子として、gene symbol を共有する１３８８９遺伝子を選んだ（つまり $N = 13889$）。方法と材料で記述した方法で j 番目のヒトの細胞に付与された特異値ベクトル $u_{\ell j}$ と k 番目のマウスの細胞に付与された特異値ベクトル $u_{\ell k}$ を得た。ここでは時間依存性がある遺伝子発現プロファイルに興味があるので、時間依存性がある特異値ベクトルを選択するために

$$u_{\ell j} = a_\ell + \sum_{t=1}^{T} a_{\ell t} \delta_{jt} \tag{11}$$

$$u_{\ell k} = b_\ell + \sum_{t=1}^{T} b_{\ell t} \delta_{kt} \tag{12}$$

というカテゴリ回帰を行った。$a_\ell, a_{\ell j}, b_\ell, b_{\ell k}$ は回帰係数、δ_{tj}, δ_{tk} は j,k が t 番目の時刻での観測である場合に１，それ以外は０の δ 関数である。カテゴリ回帰で P 値を計算し、BH 法で多重比較補正して、補正 P 値が０．０１以下の ℓ を選択した。**図 1** と**表 1** は選択された特異値ベクトルのヒトとマウスの間の比較である。ヒトに対して５５個、マウスに対しては４４個の特異値ベクトルが選択されているが、うち３２個が共通である。ヒトとマウスの実験は全く独立であることを考えるとこれはかなり驚くべき一致度であると言えるだろう。

表 1 ヒト細胞に付与された１９７７個の特異値ベクトルのうち、選択された５５個の特異値ベクトルと、マウス細胞に付与された１９０７個の特異値ベクトルのうち、選択された４４個の特異値ベクトルとの混同行列

Table 1 Confusion matrix of coincidence between selected 55 singular value vectors selected among all 1977 singular value vectors, $u_{\ell j}$, attributed to human cells and 44 singular value vectors selected among all 1907 singular value vectors, $v_{\ell k}$, attributed to mouse cells.

		human	
		not selected	selected
mouse	not selected	1833	23
	selected	12	32

表 2 共通の１３３８４遺伝子のうち、選択されたヒト４５６遺伝子とマウス５０５遺伝子の混同行列。オッズ比は１３３で、フィッシャーの正確確率検定で計算された P 値は数値誤差の範囲内で０。

Table 2 Confusion matrix of coincidence between selected 456 genes for human and selected 505 genes for mouse among all 13384 common genes. Selected: corrected P-values, computed with χ^2 distribution eqs. (9) and (10), are less than 0.01, not selected: otherwise. Odds ratio is as many as 133 and P-values computed by Fisher's exact test is 0 (i.e. less than numerical accuracy).

		human	
		not selected	selected
mouse	not selected	13233	151
	selected	200	305

次に、(9) 式と (10) 式を用いて付与した P 値を多重比較補正して 0.01 以下の遺伝子を選んだ結果を**表 2** に示す。共通の１３３８４遺伝子のうち、マウス５０５遺伝子、ヒト４５６遺伝子が選ばれたが、うち３０５遺伝子が共通に選ばれていた。全く別々の実験の統合解析であることを考えるとこの一致度は非常に大きいと言えるだろう。

この様に、２つの独立なデータセットから共通の（発現プロファイルが時間依存性を持っていると思われる）遺伝子を選択することに成功した。これ自体は非常に望ましいことだが、これらの選択された遺伝子に生物学的な意味が無い場合には意味がない。まったく独立のデータセットであるから、生物学的な意味以外で共通の遺伝子が選択されるとは考えにくいが、しかし、確認しないわけにはいかない。そこで、ここでは Enrichr [4] に遺伝子をアップロードしてエンリッチメント解析を行い、選択された遺伝子の生物学的な意味を解析した。**表 3** は “Allen Brain Atlas up”

ⓒ 2021 Information Processing Society of Japan

Vol.2021-MPS-133 No.25
Vol.2021-BIO-66 No.25
2021/6/30

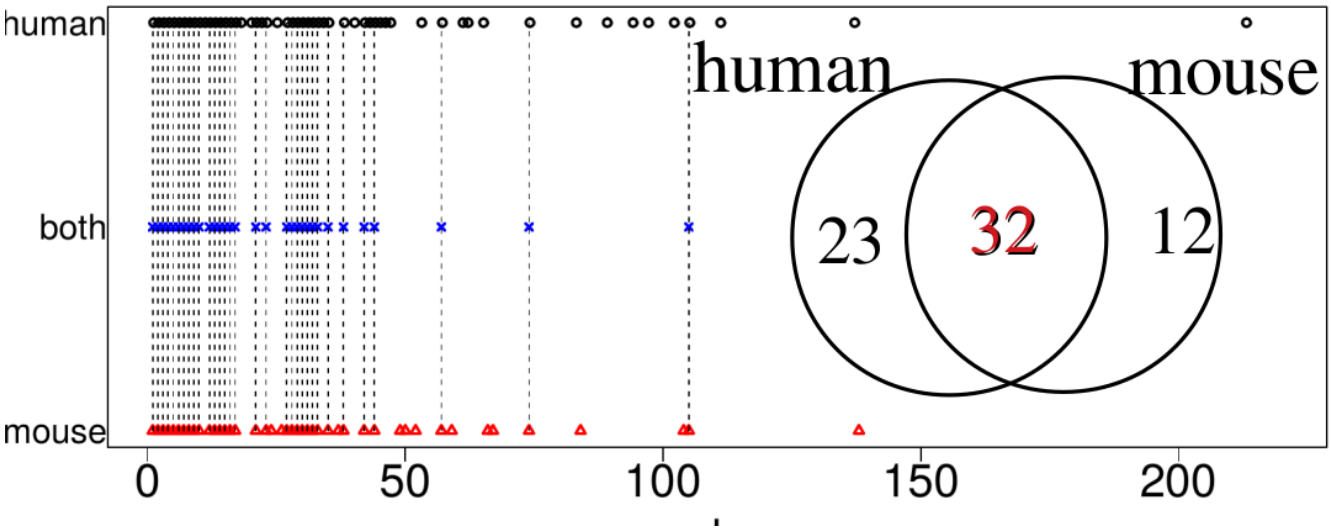

図 1　多重比較補正された０．０１以下の P 値が付与されて選択されたヒトとマウスのサンプル特異値ベクトル $u_{\ell j}, v_{\ell k}$ の比較。

Fig. 1　Comparison between singular value vectors, $u_{\ell j}$ and $v_{\ell k}$, to which adjusted P-values less than 0.01 are attributed.

表 3　選択されたヒト４５６遺伝子とマウス５０５遺伝子を Enrichr にアップロードした時の “Allen Brain Atlas up” カテゴリのエンリッチメント結果

Table 3　Five top ranked terms from “Allen Brain Atlas up” by Enrichr for selected 456 human genes and 505 mouse genes.

human			
Term	Overlap	P-value	Adjusted P-value
Paraventricular hypothalamic nucleus, magnocellular division, medial magnocellular part	31/301	2.68×10^{-12}	2.91×10^{-9}
Paraventricular hypothalamic nucleus, magnocellular division	31/301	2.68×10^{-12}	2.91×10^{-9}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part	28/301	3.39×10^{-10}	1.47×10^{-7}
Paraventricular hypothalamic nucleus	29/301	7.02×10^{-11}	5.08×10^{-8}
paraventricular nucleus, dorsal part	27/301	1.57×10^{-9}	4.88×10^{-7}
mouse			
Paraventricular hypothalamic nucleus, magnocellular division, medial magnocellular part	31/301	4.03×10^{-11}	2.19×10^{-8}
Paraventricular hypothalamic nucleus, magnocellular division	31/301	4.03×10^{-11}	2.19×10^{-8}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part	31/301	4.03×10^{-11}	2.19×10^{-8}
lower dorsal lateral hypothalamic area	29/301	8.40×10^{-10}	3.65×10^{-7}
Paraventricular hypothalamic nucleus, magnocellular division, posterior magnocellular part, lateral zone	31/301	4.03×10^{-11}	2.19×10^{-8}

の結果である。

ここにあるのはトップ５だけだが（フルデータは原著論文 [1] の補遺を参照）、ヒト、マウス共に脳に関係した遺伝子の発現が上昇していることが解る。これは “Allen Brain Atlas down”（つまり発現が下ってい遺伝子）でしか有意にエンリッチメントされた遺伝子しか見つからなかった、前回の主成分分析を使った結果 [3] に比べて大きく改善していることが解るだろう。**表 4** は “JENSEN TISSUES” の胎児脳の結果である。P 値が 10^{-10} 程度だった前回の

表 4　Enrichr の “JENSEN TISSUES” カテゴリの胎児脳のエンリッチメント結果

Table 4　Enrichment of Embryonic brain by “JENSEN TISSUES” in Enrichr

Term	Overlap	P-value	adjusted P-value
human			
Embryonic_brain	330/4936	3.36×10^{-104}	4.30×10^{-102}
mouse			
Embryonic_brain	366/4936	3.59×10^{-115}	4.59×10^{-113}

ⓒ 2021 Information Processing Society of Japan

結果 [3] よりこちらも大きく改善している。**表 5** は同じく “ARCHS4 Tissues” カテゴリの結果だが中脳そのもの (MIDBRAIN) が非常に高い有意性をもって検出されている。これも前回 [3] は不可能だった。

表 5 Enrichr の “ARCHS4 Tissues” カテゴリの中脳のエンリッチメント結果

Table 5 Enrichment of Embryonic brain by “ARCHS4 Tissues” in Enrichr

Term	Overlap	P-value	adjusted P-value
human			
MIDBRAIN	248/2316	1.02×10^{-129}	1.11×10^{-127}
mouse			
MIDBRAIN	248/2316	1.44×10^{-99}	1.56×10^{-97}

また、これらの遺伝子をタンパク質タンパク質相互作用 (Protein protein interaction, PPI) データベースである STRING [5] にもアップロードしてみた。その結果、ヒト４５６遺伝子の間には 7488 個の PPI がありこれは期待値である 3524 個を大きく上回っており 1×10^{-16} という非常に小さな P が付与されることが解った。同様に、マウスの５０５遺伝子に対しては、6788 個の PPI があり、これは期待値である 3290 個をやはり大きく上回っており、1×10^{-16} という非常に小さな P が付与されることが解った。この結果は PPI という生物学的な観点から見てもヒトとマウスで選択された遺伝子は生物学的に大きな意味があるものが選ばれていると思われる。

最後にこれらの遺伝子を制御する転写因子があるかどうかを調べてみた。**表 6** は Enrichr の “ENCODE and ChEA Consensus TFs from ChIP-X” カテゴリの結果である。ヒトとマウスそれぞれに多数の転写因子が見つかっているだけではなく、マウスとヒトの間で大きく共通した転写因子群が見つかっている。このことから、結果はこの観点から見ても大いに生物学的に意味があると思われる。また、regnetworkweb [6] を参照する限りではこれらの転写因子は互いに制御関係にあり、意味あるネットワークを組んでいるらしいことも判明した（**図 2**）。これらの転写因子に関する結果を前回の主成分分析を用いた結果 [3] と比較すると、ヒトに対しては前回のほうがよかったものの、マウスでは大きく改善し、特に regnetworkweb との整合性では今回の方が圧倒的によくなっている。

以上、全ての結果が、テンソル分解を用いた教師なし学習による変数選択法が RNA-seq での遺伝子選択に有効であることを示している。

3. おわりに

本報告ではテンソル分解を用いた教師なし学習による変数選択法が RNA-seq での遺伝子選択に有効であることをヒトとマウスの遺伝子発現プロファイルの統合解析で示した。RNA-seq はただでさえ、ラベル情報が不足しているので、条件間の発現差で遺伝子を選択することは難しい。しかも、今回の場合、ヒトとマウスという異種間の比較であるだけではなく、ヒトの場合は中絶が許されてる３ヶ月までのデータしか無いためにマウスとヒトの計測時刻の対応がとれないという条件の中で両方に共通に生物学的な遺伝子を選択することに成功している。

このようなことからテンソル分解を用いた教師なし学習による変数選択法は RNA-seq においてもデファクトスタンダードの方法として広く普及することが望まれている。

謝辞 この研究は科研費の 17K00417, 19H05270, 及び、大川財団のグランド番号 17-10 の補助を受けて実行された。

参考文献

[1] Taguchi, Y.-h. and Turki, T.: Tensor Decomposition-Based Unsupervised Feature Extraction Applied to Single-Cell Gene Expression Analysis, *Frontiers in Genetics*, Vol. 10, p. 864 (online), DOI: 10.3389/fgene.2019.00864 (2019).

[2] Taguchi, Y.-H.: *Unsupervised Feature Extraction Applied to Bioinformatics*, Springer International Publishing (2020).

[3] 田口善弘：主成分分析を用いた教師無し学習による変数選択の一細胞 RNA-seq への応用，情報処理学会研究報告，Vol. 2019-BIO-57, No. 6, pp. 1–6 (2019).

[4] Kuleshov, M. V., Jones, M. R., Rouillard, A. D., Fernandez, N. F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S. L., Jagodnik, K. M., Lachmann, A., McDermott, M. G., Monteiro, C. D., Gundersen, G. W. and Ma’ayan, A.: Enrichr: a comprehensive gene set enrichment analysis web server 2016 update, *Nucleic Acids Research*, Vol. 44, No. W1, pp. W90–W97 (online), DOI: 10.1093/nar/gkw377 (2016).

[5] Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. and Mering, C. v.: STRING v11: protein-Protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets, *Nucleic Acids Research*, Vol. 47, No. D1, pp. D607–D613 (online), DOI: 10.1093/nar/gky1131 (2018).

[6] Liu, Z.-P., Wu, C., Miao, H. and Wu, H.: RegNetwork: an integrated database of transcriptional and post-transcriptional regulatory networks in human and mouse, *Database*, Vol. 2015 (online), DOI: 10.1093/database/bav095 (2015).

ⓒ 2021 Information Processing Society of Japan

Vol.2021-MPS-133 No.25
Vol.2021-BIO-66 No.25
2021/6/30

表 6 Enrichr の “ENCODE and ChEA Consensus TFs from ChIP-X” カテゴリでエンリッチメントされた転写因子。ヒトとマウスで共通のものは太字表示。

Table 6 TFs enriched in “ENCODE and ChEA Consensus TFs from ChIP-X” by Enrichr for human and mouse. Bold TFs are common.

human	BCL3, **BHLHE40**, **EGR1**, **GABPA**, **IRF3**, **PPARG**, **REST**, **RFX5**, SP1, SP2, SRF, **STAT3**, **TCF7L2**, TRIM28, TRIM28, **ZBTB33**,
mouse	**BHLHE40**, CTCF, E2F4, E2F6, **EGR1**, ESR1, ETS1, FLI1, **GABPA**, **IRF3**, NFIC, NRF1, **PPARG**, RCOR1, **REST**, **RFX5**, SPI1, **STAT3**, **TCF7L2**, USF1, USF2, YY1, **ZBTB33**, ZNF384,

(A) (B)

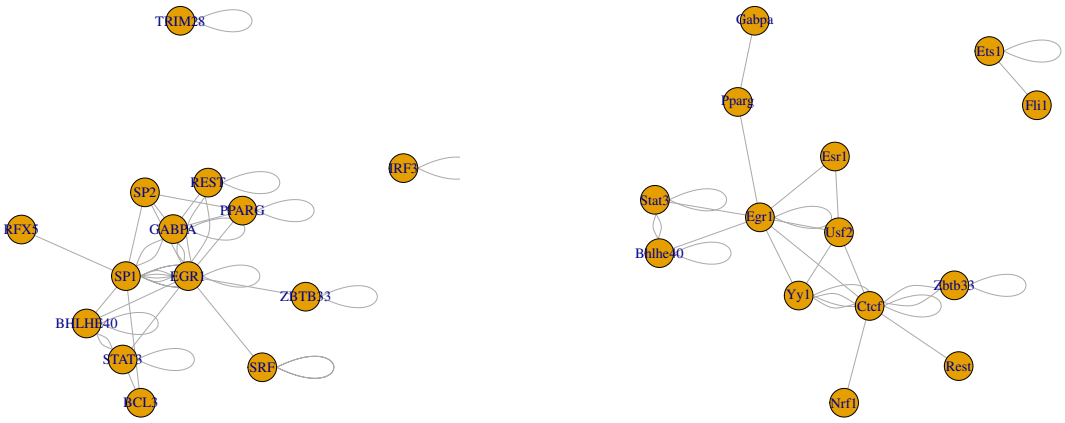

図 2 TF network identified by regnetworkweb for TFs in Table 6. (A) human, (B) mouse.

ⓒ 2021 Information Processing Society of Japan