Vol.2022-MPS-138 No.55
Vol.2022-BIO-70 No.55
2022/6/29

テンソル分解を用いた赤血球遺伝子発現データの解析

相川　隼人[1,a)]　田口　善弘[2]

概要：m6A は哺乳類に広くみられる一般的な遺伝子修飾であり様々な機能があると考えられている. この修飾は赤血球の生成にも関与していることが分かっており,m6A mRNA メチルトランスフェラーゼ複合体をコードする遺伝子は赤血球マーカー CD235a の発現に影響を与えることが実験により明らかとなっている. 本研究はこの遺伝子発現データをテンソル分解によって解析することで発現量に有意に差の生じた遺伝子を選択することを目的として行った. その結果, 候補遺伝子の選択に成功した. これらの遺伝子をエンリッチメント解析したところダイアモンド・ブラックファン貧血といった遺伝性の造血不全症と関連していることが分かり, 候補遺伝子が赤血球生成に関連していることを強く示唆している.

1. はじめに

N6-methyladenosine(m6A) は哺乳類をはじめとした高等真核生物の RNA に最もよく見られる遺伝子修飾であり, 人体では癌の発生や造血など様々な生理機能に影響を及ぼしていると考えられている [1].m6A は METTL3 や WTAP といったライターによってコードされる m6A メチルトランスフェラーゼによって RNA に付加される. Kuppers らによる実験 [2] により METTL3,WTAP を除去した Human erythroleukemia (HEL) 細胞では赤血球に主要なシアロ糖タンパク質である CD235a の発現が有意に減少したことが報告されている. 本研究ではこのデータをテンソル分解による変数選択法により解析し発現量に有意に差の生じた遺伝子を選択することを目的に行った.

2. 方法

2.1 扱うデータと形式

本解析では Gene Expression Omnibus (GEO) で公開されている GSE106124 というデータセットを用いた. このデータセットは Kuppers らによる HEL 細胞の遺伝子発現プロファイリング [2] である. GSE106124 は 5 つのサブシリーズから構成されており, そのうち GSE95372,GSE105782 の 2 つについて解析を行った. これらのデータをテンソル分解で解析するため, それぞれをテンソルの形式に成形した. また成形する過程で平均 0, 分散 1 となるように規格化した.

2.1.1 GSE95372

GSE95372 は GATA1, GYPA, LMO2, METTL3, WTAP をそれぞれ KO した場合と Non targeting control の遺伝子発現プロファイリングデータである. このデータを以下の 2 つのテンソル x_{i_1,j_1,j_2},x_{i_2,j_3,j_4,j_5} として扱う. ここで x_{i_1,j_1,j_2} は i_1 番目の遺伝子について j_1 番目の KO 遺伝子 ($j_1 = 1$: GATA1, $j_1 = 2$: GYPA, $j_1 = 3$: LMO2, $j_1 = 4$: NTC),j_2 番目の複製の発現プロファイルであり, $24819 \times 4 \times 3$ の 3 階のテンソルである. 同様に x_{i_2,j_3,j_4,j_5} は i_2 番目の遺伝子における j_3 番目の KO 遺伝子 ($j_3 = 1$: METTL3, $j_3 = 2$: WTAP),j_4 番目の sgRNA,j_5 番目の複製についての発現プロファイルであり,$25370 \times 2 \times 2 \times 2$ の 4 階のテンソルである.

便宜上 3 階のテンソルを GSE95372A,4 階のテンソルを GSE95372B とする.

2.1.2 GSE105782

GSE105782 は WTAP を KO した場合のリボソームプロファイリングデータである. これを x_{i_3,j_6,j_7,j_8} とテンソルの形に成形した. ここで x_{i_3,j_6,j_7,j_8} は i_3 番目の遺伝子の j_6 番目の KO 遺伝子 ($j_6 = 1$: NTC, $j_6 = 2$: WTAP),j_7 番目のライブラリー ($j_7 = 1$: ribosome, $j_7 = 2$: total),j_8 番目の複製における発現プロファイルであり,$19224 \times 2 \times 2 \times 3$ の 4 階のテンソルである.

2.2 テンソル分解による教師なし変数選択法

テンソル分解による解析は Taguchi によるテンソル分解を用いた教師なし学習による変数選択法 [3] を基に行った. 分解には R パッケージである rTensor の HOSVD 関数を

[1] 中央大学大学院理工学研究科物理学専攻
Major in Physics, Graduate School of Science and Engineering, Chuo University
[2] 中央大学理工学部物理学科
Major in Physics, Faculty of Science and Engineering, Chuo University
[a)] aikawa.rane@gmail.com

用いた.

3. 結果

3.1 テンソル分解による遺伝子選択

それぞれのテンソルについてテンソル分解を行い各モードに沿った特異値行列とコアテンソルを得た. 図 1 は GSE95372A の特異値行列の値を列ごとに示したものである. この図より KO 遺伝子特異値行列については第 4 特異値ベクトルが GATA1,LMO2 と NTC の間の差を反映していることがわかる. また複製特異値行列については複製間で同じ傾向を示すと考えられるため第 1 特異値ベクトルが選ばれた. これらの特異値ベクトルをもとにコアテンソルの絶対値が大きくなるように遺伝子特異値行列を選び, 第 6 特異値ベクトルを選択した.

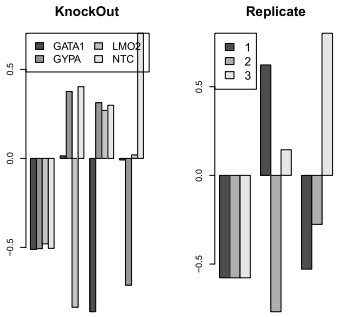

図 1 GSE95372A の各特異値行列

GSE95372B についても同様に解析を行った. 図 2 は GSE95372B の特異値行列の値である. METTL3 と WTAP の KO, sgRNA, 複製のそれぞれについて発現プロファイルは似た傾向になると考えられるため図より各特異値行列について第 1 特異値ベクトルが選ばれる. 選ばれた特異値ベクトルから遺伝子特異値行列については第 1 特異値ベクトルを選択した.

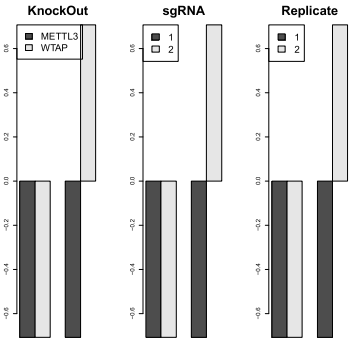

図 2 GSE95372B の各特異値行列

GSE105782 の特異値行列の値を図 3 に示す.KO 遺伝子,ribosome と total ライブラリーについては異なる傾向, 複製については似た傾向を示すと考えられる. そのため KO 遺伝子特異値行列とライブラリー特異値行列は第 2 特異値ベクトル, 複製特異値行列は第 1 特異値ベクトルが選ばれた. これらの特異値ベクトルとコアテンソルの値から遺伝子特異値ベクトルは第 1 特異値ベクトルを選択した.

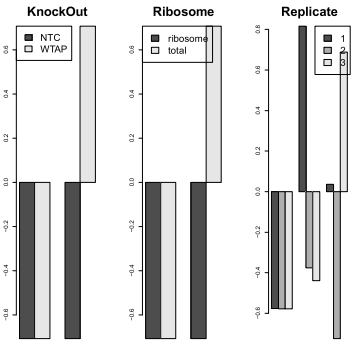

図 3 GSE105782 の各特異値行列

その後, 選択した遺伝子特異値ベクトルが多重ガウス分布に従っていると仮定し各遺伝子に棄却確率を付与.Benjamini-Hochberg 法 [4] で多重比較補正を行い補正 P 値が 0.05 以下となるような遺伝子を選択した.GSE95372 の 1 つ目のテンソルからは 266,2 つ目のテンソルからは 189 の遺伝子が選ばれた.GSE105779 のテンソルからは 208 の遺伝子が選択された.

3.2 エンリッチメント解析

選択された遺伝子について Enrichr でエンリッチメント解析を行った. 表 1 にエンリッチメント解析の結果上位 5 種を示す.Orphanet Augmented 2021 は難病のデータベースである. 表を見ると Blackfan-Diamond anemia という先天性の造血不全症が濃縮されていることがわかる. また造血機能とは直接関係がないものの小頭症や異形成, 成長遅延, 知能障害などの関連遺伝子が濃縮されていることがわかる. これらは遺伝性の貧血と合併して発症することが多く知られており関連が示唆されている.

4. おわりに

本研究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 m6A 転移酵素をコードする遺伝子を除去した細胞の発現プロファイルから発現に有意な差の生じた遺伝子を選択することを目的として行った. その結果として各データについてそれぞれ約 200 の遺伝子を選択することに成功した. またエンリッチメント解析の結果から選択遺伝子が造血機能に深く関連していることがわかり, テンソル分解による遺伝子選択が適切にできたことを示している.

表 1 各テンソルで選択された遺伝子のエンリッチメント解析結果上位 5 種

GSE95372A	P 値	補正 P 値
X-linked intellectual disability-cerebellar hypoplasia-spondylo-epiphyseal dysplasia syndrome ORPHA:459070	2.1251E-54	2.29086E-51
X-linked microcephaly-growth retardation-prognathism-cryptorchidism syndrome ORPHA:435938	2.1251E-54	2.29086E-51
Blackfan-Diamond anemia ORPHA:124	6.17031E-54	4.4344E-51
Myelodysplastic syndrome associated with isolated del(5q) chromosome abnormality ORPHA:86841	4.88993E-49	2.63567E-46
Familial colorectal cancer Type X ORPHA:440437	6.97583E-47	3.00798E-44
GSE95372B		
Blackfan-Diamond anemia ORPHA:124	1.99067E-88	3.15322E-85
Familial colorectal cancer Type X ORPHA:440437	5.6425E-84	4.46886E-81
Myelodysplastic syndrome associated with isolated del(5q) chromosome abnormality ORPHA:86841	2.09493E-82	6.63675E-80
X-linked intellectual disability-cerebellar hypoplasia-spondylo-epiphyseal dysplasia syndrome ORPHA:459070	2.09493E-82	6.63675E-80
X-linked microcephaly-growth retardation-prognathism-cryptorchidism syndrome ORPHA:435938	2.09493E-82	6.63675E-80
GSE105782		
Familial colorectal cancer Type X ORPHA:440437	6.96384E-39	1.25279E-35
Blackfan-Diamond anemia ORPHA:124	3.45303E-37	3.106E-34
Myelodysplastic syndrome associated with isolated del(5q) chromosome abnormality ORPHA:86841	3.80671E-34	1.36966E-31
X-linked intellectual disability-cerebellar hypoplasia-spondylo-epiphyseal dysplasia syndrome ORPHA:459070	3.80671E-34	1.36966E-31
X-linked microcephaly-growth retardation-prognathism-cryptorchidism syndrome ORPHA:435938	3.80671E-34	1.36966E-31

参考文献

[1] Xiulin Jiang et al. *The role of m6A modification in the biological functions and diseases.* Signal Transduction and Targeted Therapy, volume 6, Article number: 74. doi: 10.1038/s41392-020-00450-x. (2021).

[2] Daniel A Kuppers et al. *N 6-methyladenosine mRNA marking promotes selective translation of regulons required for human erythropoiesis.* Nature Communications, 10(1):4596. doi: 10.1038/s41467-019-12518-6. (2019).

[3] Taguchi, Y. *Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases.* BMC Med. Genomics, 10(S4):67. doi: 10.1186/s12920-017-0302-1. (2017).

[4] Benjamini, Y. and Hochberg, Y. *Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.* Journal of the Royal Statistical Society: Series B (Methodological), Volume57, Issue1, pp. 289-300. (1995).