Vol.2022-BIO-69 No.1
2022/3/10

腎臓明細胞癌の miRNA 指標のテンソル分解による解析

田口 善弘[1,a)]　呉 家樂[2,b)]

概要：腎臓 (淡) 明細胞癌（ kidney renal clear cell carcinoma,KIRC）は腎細胞癌の8 割ほどを占めるがんであり、その原因を解明することは重要である。今回は我々は mRNA と miRNA の発現量をがんと正常細胞で比較することで病因に関係すると目される遺伝子と miRNA を複数個選ぶことに成功した。これらの遺伝子は先行研究でがんで有意に発現が変化していることが知られているだけではなく、KIRC 患者の生存率に関係している遺伝子が多く含まれていた。また miRNA については、その標的遺伝子ががんに関係していることが解った。さらに選択された遺伝子や miRNA は独立な2 つの研究（ それぞれが共通した患者について mRNA と miRNA の発現を計測している）において有意に重なりがある形で選択されることがわかったが、t 検定、SAM、limma などの既存手法ではこのようなデータセットに依らないロバストな遺伝子や miRNA の選択は出来なかった。

1. はじめに

いわゆるマルチオミックスデータの計測は広く行われるようになってきたが、標準的な解析方法は確立していない。多くの場合、せっかく共通サンプルに対して複数種類の計測を行っても個別に解析するにとどまることになる。本研究では従来から提唱している「 テンソル分解を用いた教師なし学習による変数選択法」を肝臓（ 淡）明腎細胞がんの mRNA と miRNA の統合解析に応用し、疾患原因となる mRNA や miRNA の特定を試みた例 [1] を報告する。

2. 材料と方法

2.1 mRNA 及び miRNA の発現プロファイル

解析に用いた mRNA 及び miRNA の発現プロファイルはいずれも公的なデータベースのものを用いた。1 つめのデータセットは TCGA からダウンロードしたものであり、がん２５３ サンプル、正常な腎臓7 1 サンプルに対して mRNA と miRNA を計測したデータセットであり、２つめのデータセットは GEO から得たもので、がんと正常な腎臓がそれぞれ1 7 サンプルずつのデータセットである。詳しいことについては原論文 [1] を参照されたい。

[1] 中央大学
Chuo University, Kasuga, Bunkyo-ku, Tokyo 112–8551, Japan
[2] 亜州大学
Asia Uniersity, 500, Lioufeng Rd., Wufeng, Taichung 41354, Taiwan
[a)] tag@granular.com
[b)] ppiddi@gmail.com
本研究は原著論文として刊行済みである [1]

2.2 解析手法

解析フローの概要は図 **1** にあるとおりである。以下では個々の解析手法について説明する。詳細は近著 [2] を参照されたい。

2.2.1 テンソル分解を用いた教師なし学習による変数選択法

$x_{ij}^{(\mathrm{mRNA})} \in \mathbb{R}^{N\times M}$ を j 番目のサンプルに対する i 番目の mRNA の発現量、$x_{kj}^{(\mathrm{miRNA})} \in \mathbb{R}^{K\times M}$ を j 番目のサンプルに対する k 番目の miRNA の発現量、とする。ここでこの2 つを統合解析するため以下のような3 階のテンソルを構成する。

$$x_{ijk} = x_{ij}^{(\mathrm{mRNA})} \times x_{kj}^{(\mathrm{miRNA})} \in \mathbb{R}^{N\times M\times K} \tag{1}$$

これをテンソル分解したいのだが、いまの場合、$N \sim 10^4, K \sim 10^3, M \sim 10^2$ であり、必要なメモリーが $\sim 10^9$ のオーダーになってしまい、通常の計算機ではテンソル分解を行うことが出来ないサイズになってしまった。そこでこれを縮約し

$$x_{ik} = \sum_{j=1}^{M} x_{ijk} \in \mathbb{R}^{N\times K} \tag{2}$$

とし、これを特異値分解

$$x_{ik} = \sum_{\ell=1}^{\min(N,K)} \lambda_\ell u_{\ell i} u_{\ell k} \tag{3}$$

を実行し（$u_{\ell i} \in \mathbb{R}^{N\times N}, u_{\ell k} \in \mathbb{R}^{K\times K}$）、サンプルに付与される特異値は

$$u_{\ell j}^{\mathrm{mRNA}} = \sum_{i=1}^{N} x_{ij}^{(\mathrm{mRNA})} u_{\ell i} \in \mathbb{R}^{M\times M} \tag{4}$$

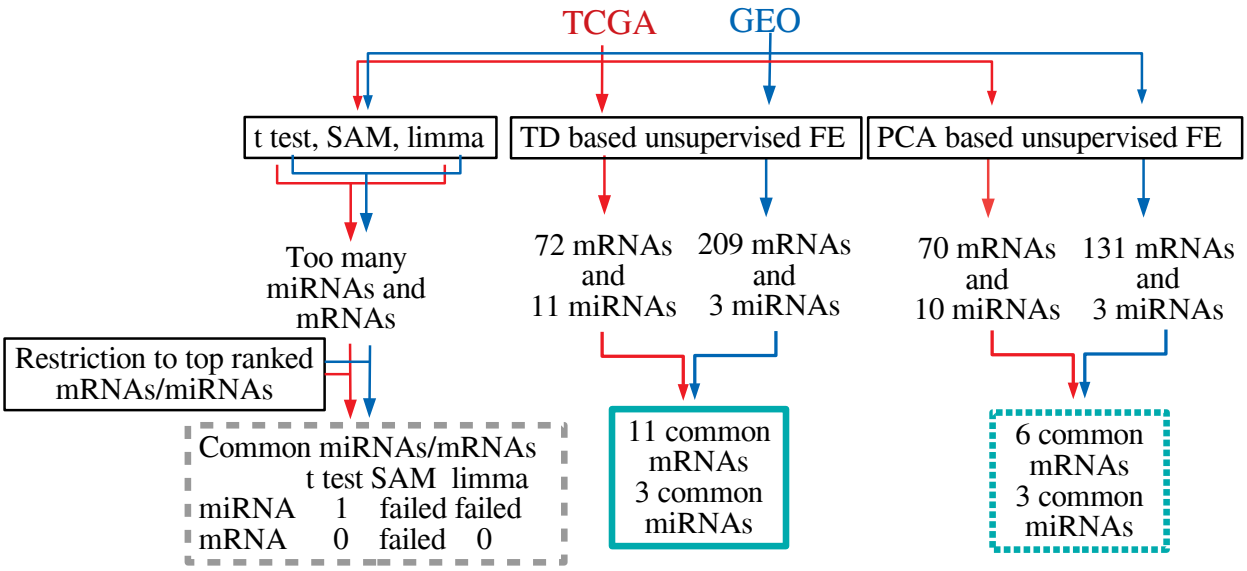

図 1　解析フローの説明
Fig. 1　Analysis flow

$$u_{\ell j}^{\mathrm{miRNA}} = \sum_{k=1}^{K} x_{kj}^{(\mathrm{miRNA})} u_{\ell k} \in \mathbb{R}^{M\times M} \tag{5}$$

という計算で求めることにする。

次に

- $u_{\ell j}^{\mathrm{mRNA}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ が (j 依存について) 相関している。
- $u_{\ell j}^{\mathrm{mRNA}}$ ががんと正常腎臓で有意差がある。
- $u_{\ell j}^{\mathrm{miRNA}}$ ががんと正常腎臓で有意差がある。

の3 条件を満たす ℓ を特定し、対応する $u_{\ell i}, u_{\ell k}$ を用いて

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell i}}{\sigma_\ell}\right)\right] \tag{6}$$

$$P_k = P_{\chi^2}\left[> \left(\frac{u_{\ell k}}{\sigma'_\ell}\right)\right] \tag{7}$$

の様に i, k に P 値を付与する。但し、$P_{\chi^2}[> x]$ は引数が x 以上である場合の累積 χ 二乗分布であり、$\sigma_\ell, \sigma'_\ell$ は標準偏差である。P_i, P_k は BH 基準 [2] で多重比較補正し、補正 P 値が 0.01 以下の mRNA,miRNA を選択する。

2.2.2　主成分分析を用いた教師なし学習による変数選択法

$x_{ij}^{(\mathrm{mRNA})}$ と $x_{kj}^{(\mathrm{miRNA})}$ を

$$\sum_{i=1}^{N} x_{ij}^{(\mathrm{mRNA})} = 0 \tag{8}$$

$$\sum_{i=1}^{N} \left(x_{ij}^{(\mathrm{mRNA})}\right)^2 = N \tag{9}$$

$$\sum_{k=1}^{K} x_{kj}^{(\mathrm{miRNA})} = 0 \tag{10}$$

$$\sum_{k=1}^{K} \left(x_{kj}^{(\mathrm{miRNA})}\right)^2 = K \tag{11}$$

となるように規格化しておき

$$\sum_{i'=1}^{N}\left(\sum_{j=1}^{M} x_{ij}^{(\mathrm{mRNA})} x_{i'j}^{(\mathrm{mRNA})}\right) u_{\ell i'} = \lambda_\ell u_{\ell i} \tag{12}$$

$$\sum_{k'=1}^{K}\left(\sum_{j=1}^{M} x_{kj}^{(\mathrm{miRNA})} x_{k'j}^{(\mathrm{miRNA})}\right) u_{\ell k'} = \lambda'_\ell u_{\ell k} \tag{13}$$

という対角化問題を解くことで $u_{\ell i} \in \mathbb{R}^{N\times N}, u_{\ell k} \in \mathbb{R}^{K\times K}$ を求める。(4) 式以下の計算は前節と全く同じである。

3. 結果

3.1 テンソル分解を用いた教師なし学習による変数選択法

$\ell = 2$ について

- $u_{\ell j}^{\mathrm{mRNA}}$ に対して、がんと正常な腎臓で二群の t 検定を行った結果 $P = 7.10 \times 10^{-39}$
- $u_{\ell j}^{\mathrm{miRNA}}$ に対して、がんと正常な腎臓で二群の t 検定を行った結果 $P = 2.13 \times 10^{-71}$
- $u_{\ell j}^{\mathrm{mRNA}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ の相関係数は $0.905 (P = 1.63 \times 10^{-121}$、 図 **2**)

が成り立っていたので $\ell = 2$ を用いることにした。この結果、表 **1** にあるような 11 個の miRNA と 72 個の mRNA が選択された。これらについて様々な検証を行った。

まず、miRNA と mRNA の相関を調べた。この結果 $11 \times 72 = 792$ ペアの内、353 ペアが正に、358 ペアが負に有意に相関していた(ペアごとに P 値を計算してから多重比較補正し、BH 基準で多重比較補正し、補正 P 値が 0.01 以下のものを有意であるとした)。従って、９０％のペアが相関していたことになる。

次にエンリッチメント解析を用いて生物学的な評価を

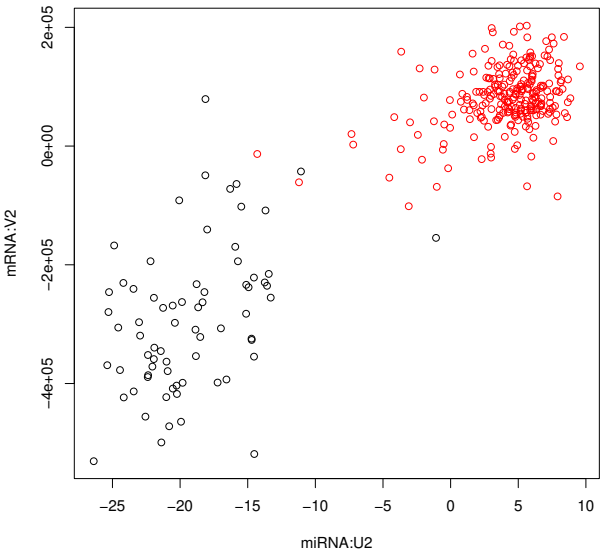

図 2 $u_{\ell j}^{\mathrm{mRNA}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ の散布図。赤はがん。

Fig. 2 Scatter plot between $u_{\ell j}^{\mathrm{mRNA}}$ and $u_{\ell j}^{\mathrm{miRNA}}$. Tumors are red colored.

表 1 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 TCGA で選択された miRNA と mRNA

Table 1 miRNA and mRNAs selected by TD based unsupervised FE for TCGA

miRNA ID
hsa-mir-210 hsa-mir-891a hsa-mir-155 hsa-mir-200c hsa-mir-141 hsa-mir-508 hsa-mir-122 hsa-mir-514-3 hsa-mir-514-1 hsa-mir-514-2 hsa-mir-184
Gene symbol
ACTG1 ADAM6 AIF1L ALDOA ALDOB ANGPTL4 APLP2 APP AQP1 AQP2 ASS1 ATP1A1 ATP1B1 ATP5A1 ATP5B B2M C3 C4A C7 CA12 CCND1 CD74 CDH16 COL4A1 COL4A2 CP CYFIP2 ENO1 FN1 FTL GAPDH GATM GNB2L1 GPX3 HLA-A HLA-B HLA-C HLA-DRA HSD11B2 HSP90AA1 HSPA8 IGFBP3 IGFBP5 ITM2B KNG1 LDHA LDHB LOC96610 NDRG1 NDUFA4L2 NNMT P4HB PCK1 PEBP1 PLIN2 PLVAP PODXL RGS5 SERPINA1 SLC12A1 SLC12A3 SOD2 SPARC SPP1 TGFBI TMBIM6 TMSB10 UBC UMOD VEGFA VIM VWF

行った。MSigDB [3] を用いた結果、図 **3** にあるように多数のがん関連の遺伝子セットと有意に被っていることが判明した。同じ MSigDB の REACTOME カテゴリでも多くのヒットが観測された (図 **4**)。

次に OncoLnc [4] を用いた生存解析を行った。図 **5** にあるように、２４個の遺伝子が生存確率と関係していることが解った。

次に DIANA-miRPath [5] を用いて選択された miRNA の評価を行った(図 **6**)。miRNA もがんに関連したものが選ばれていることがこの結果から解った。

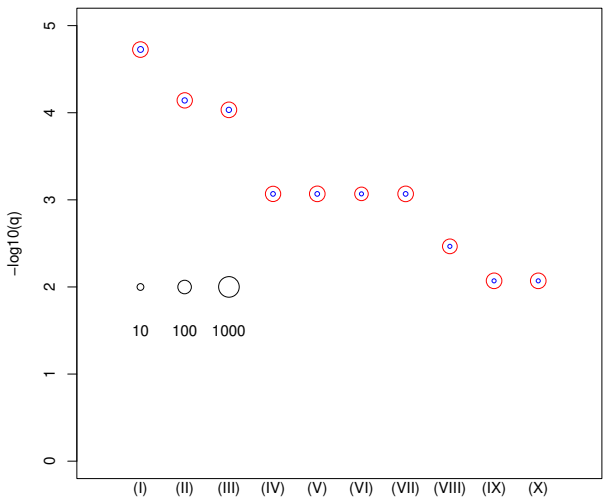

図 3 MSigDB の oncology category でのエンリッチメント解析。ローマ数字との対応は英文のキャプションを参照。縦軸は補正された P 値の自然対数の絶対値。赤円の半径はそれぞれのカテゴリに含まれる遺伝子の総数で青円の半径はそのうち、選択された遺伝子と被っていた数。

Fig. 3 Enrichment analysis of oncogenic category in MSigDB. (I) CAMP_UP.V1_UP (II) SNF5_DN.V1_DN (III) ESC_V6.5_UP_LATE.V1_UP (IV) ESC_V6.5_UP_EARLY.V1_DN (V) ESC_J1_UP_LATE.V1_UP (VI) SIRNA_EIF4GI_UP(VII) P53_DN.V1_DN (VIII) MEL18_DN.V1_UP (IX) LTE2_UP.V1_UP (X) RPS14_DN.V1_UP. Vertical axis is the negative normal logarithmic-adjusted P-values. The radii of open red and blue circles show the normal logarithmic values of the number of genes in each category and those of genes included in both the category and the selected genes shown in Table 1.

3.2 GEO との比較

TCGA のデータセットにテンソル分解を用いた教師なし学習による変数選択を適用した結果、mRNA や miRNA を選択することに成功し、生物学的にも妥当なものが選ばれていることが解った。ここでこの選択がどれくらいロバストかを確認するため、GEO からのデータセットで同じことを行った。$\ell = 2$ について

- $u_{\ell j}^{\mathrm{mRNA}}$ に対して、がんと正常な腎臓で二群の t 検定を行った結果 $P = 6.74 \times 10^{-22}$
- $u_{\ell j}^{\mathrm{miRNA}}$ に対して、がんと正常な腎臓で二群の t 検定を行った結果 $P = 2.54 \times 10^{-18}$
- $u_{\ell j}^{\mathrm{mRNA}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ の相関係数は $0.931 (P = 1.58 \times 10^{-15})$

が成り立っていた。サンプル数が1 桁少ないので有意差は弱くなってしまったが $\ell = 2$ が上記3 つの性格を満たして

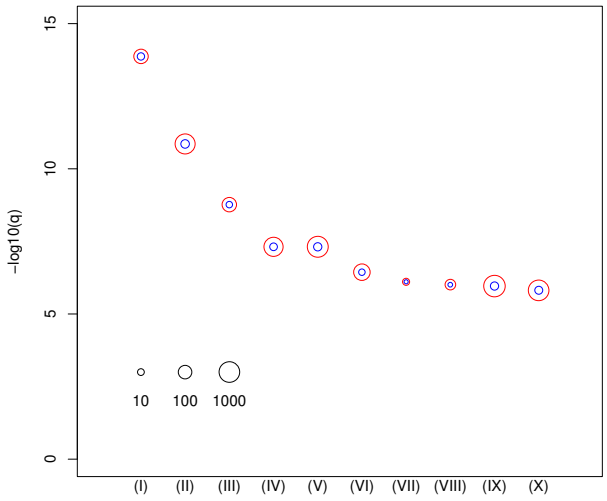

図 4 MSigDB の REACTOME category でのエンリッチメント解析。ローマ数字との対応は英文のキャプションを参照。縦軸は補正された P 値の常用対数の絶対値。赤円の半径はそれぞれのカテゴリに含まれる遺伝子の総数で青円の半径はそのうち、選択された遺伝子と被っていた数。

Fig. 4 Enrichment analysis of REACTOME category in MSigDB. (I) REACTOME_regulation of insulin-like growth factor (IGF) transport and uptake by IGF binding proteins IGFBPS (II) REACTOME_cytokine signalling in immune system (III) REACTOME response to elevated platelet cytosolic CA2+ (IV) REACTOME_signalling by interleukins (V) REACTOM_innate immune system (VI) REACTOME_platelet activation, signalling, and aggregation (VII) REACTOME_endosomal vacuolar pathway (VIII) REACTOME_gloconeogenesis (IX) REACTOME_post-translational protein modification (X) REACTOME_disease. The radii of open red and blue circles show the normal logarithmic values of the number of genes in each category and those of genes included in both the categories and the selected genes shown in Table 1.

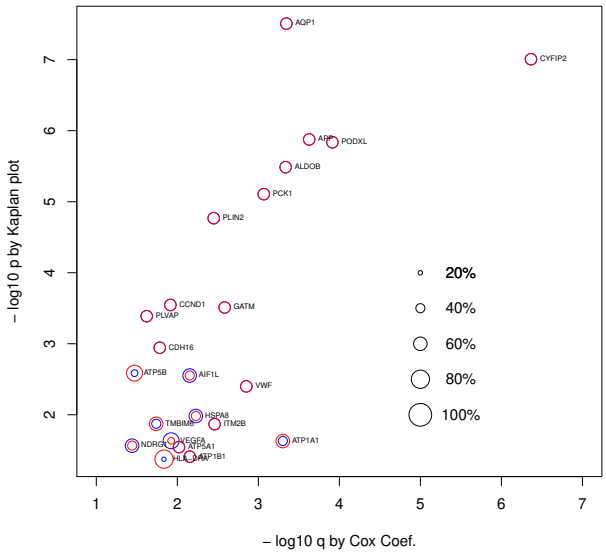

図 5 表 1 のうちの生存解析で有意な結果が得られた24 遺伝子の結果。縦軸はカプランプロット得られた補正された P 値の常用対数の絶対値。横軸は同じ量のコックス解析の結果。赤丸は発現量の下位何%で患者を二分したかの大きさを表す。50 %で2 分しなかった場合に限り上位%の大きさを青い円で表現した

Fig. 5 Survival analysis of 24 genes from Table 1 that significantly contribute to patients' survival. Vertical axis: negative normal logarithmic values of P-values computed by Kaplan plot. Horizontal axis: negative normal logarithmic values of adjusted P-values computed by Cox analysis. Red open circles indicate lower expression percentile of patient groups. Only when they are not 50%, upper expression percentiles are displayed with blue circles.

いる点は変わらないので、この結果はロバストだとみなすことが出来る。

次に選択された mRNA と miRNA の比較を行った(表 **2**)。オッズ比は 19.7 でフィッシャーの正確確率検定による P 値は 8.97×10^{-11} であり、高い有意度で一致していることがわかり、得られた結果のロバストさが確認された。miRNA については3 個の miRNA が共通していた。

表 **2** TCGA と GEO で選択された遺伝子間の混同行列

Table 2 Confusion matrix between genes selected in TCGA and GEO dataset.

		GEO	
		Not selected	Selected
TCGA	Not selected	17,209	160
	Selected	60	11

3.3 既存手法との比較

非常によい結果が得られたが、もっと簡単な既存手法で同じような結果が得られるなら意味がない。そこで t 検定、SAM [6]、limma [7] といった今回解析されたデータの測定に使われたマイクロアレイ技術でも適用可能な方法と比較した。結果から言うとこれらの手法で選ばれる mRNA や miRNA は数が多すぎることが判明した。つまり、P 値だけでは十分な絞り込みができないという結果になることが解った(表 **3**)。しかし、選択数が多いのは単に、検出力が高すぎるためであるかもしれず、上位に限ればテンソル分解を用いた教師なし学習による変数選択法と遜色ない性能を出す可能性はある。そこで、この3 つの手法について、テンソル分解を用いた教師なし学習による変数選択

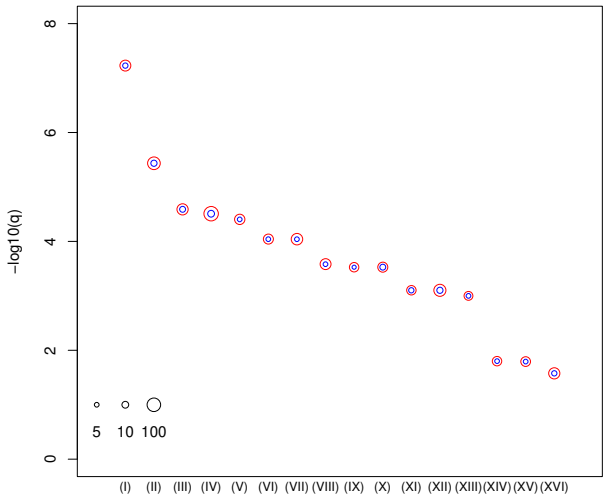

図 6 DIANA-miRPath の KEGG でのエンリッチメント解析。ローマ数字との対応は英文のキャプションを参照。縦軸は補正された P 値の常用対数の絶対値。赤円の半径はそれぞれのカテゴリに含まれる遺伝子の総数で青円の半径はそのうち、選択された遺伝子と被っていた数。

Fig. 6 Enrichment analysis of KEGG pathway provided by DIANA-mirpath to which miRNAs in Table 1 were uploaded. Vertical axis: negative normal logarithmic values of adjusted P-values. (I) Chronic myeloid leukemia (II) Proteoglycans in cancer (III) Prostate cancer (IV) Pathways in cancer (V) Pancreatic cancer (VI) Glioma (VII) Hepatitis B (VIII) Small cell lung cancer (IX) Non-small cell lung cancer (X) Colorectal cancer (XI) Endometrial cancer (XII) Viral carcinogenesis (XIII) Bladder cancer (XIV) Melanoma (XV) Renal cell carcinoma (XVI) Hepatitis C. The radii of red open circles indicate the normal logarithmic values of the number of genes in each category targeted by miRNAs in Table 1 whose normal logarithmic numbers are proportional to the radii of blue open circles. See Table S4 for numerical values and full description.

表 3 TCGA,GEO において比較手法で選択された mRNA や miRNA の数

Table 3 The number of mRNAs and miRNAs selected by other conventional methods applied to TCGA anf GEO data set

	TCGA		GEO	
	mRNA	miRNA	mRNA	miRNA
t test	13,895	399	12,152	78
SAM	14,485	441	16,336	108
limma	18,225	662	28,524	319

法と同じ数だけの上位の mRNA,miRNA を選択し、GEO と TCGA の間の一致具合をしらべた。t 検定については mRNA の共通数は0，miRNA の共通数は1 個であり、テンソル分解を用いた教師なし学習による変数選択法に大きく劣る結果となった。SAM についてはそもそも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選ばれた数以上の mRNA に $P = 0$ が割り当てられてしまっていたので、同じ数の上位 mRNA を絞り込めず、比較が不可能だった。limma については、共通する mRNA,miRNA が一個も無かった。このことから、これら3 つの比較手法は、単に検出力が高すぎて多くの mRNA や miRNA を選んでしまっているわけではなく、有意差の選定基準が適当ではないので有意差のあるものとないものがうまく区別できないために、多くの候補を選んでしまっているのだと思われることが解った。

最後に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原型である主成分分析を用いた教師なし学習による変数選択法と性能の比較を TCGA を用いて行った。結果は $\ell = 2$ について

- $u_{\ell j}^{\mathrm{mRNA}}$ に対して、がんと正常な腎臓で二群の t 検定を行った結果 $P = 2.33 \times 10^{-36}$
- $u_{\ell j}^{\mathrm{miRNA}}$ に対して、がんと正常な腎臓で二群の t 検定を行った結果 $P = 2.39 \times 10^{-77}$
- $u_{\ell j}^{\mathrm{mRNA}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ の相関係数は $0.839 (P = 2.74 \times 10^{-87})$

となった。miRNA はテンソル分解よりもよい結果を得られたが mRNA と相関係数についてはテンソル分解に劣っているので総合的には主成分分析の方が劣っていた。また TCGA と GEO との比較では、mRNA については 70 個が TCGA で、131 個が GEO で選ばれ、共通のものは6 個しか無く、テンソル分解より劣る結果となった (miRNA の共通数は3 で同じ)。従って、この点からもテンソル分解は mRNA と miRNA で別々の解析を行う主成分分析より良い結果を出すことができることが解った。

4. おわりに

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 mRNA と miRNA の発現量の統合解析に用いることで、t 検定、SAM、limma の様な既存手法では不可能な、十分に数が絞られた選択を行うことが出来た。また、このような統合解析を行うことで TCGA と GEO で有意に一致するロバストネスが高い結果を得ることが出来た。これは既存手法で上位の mRNA や miRNA を（ P 値による有意差の検出を越えて）さらに絞り込むことでは達成できないロバストネスであり、mRNA と miRNA を別々に解析する主成分分析を用いた方法でもテンソル分解と同じだけのロバストネス（ TCGA と GEO での一致度）を得ることは出来ず、統合解析の有効性が改めて示された。

参考文献

[1] Ng, K.-L. and Taguchi, Y.-H.: Identification of miRNA signatures for kidney renal clear cell carcinoma using the tensor-decomposition method, *Scientific Reports*, Vol. 10, No. 1 (online), DOI: 10.1038/s41598-020-71997-6 (2020).

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

[3] Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J. P. and Tamayo, P.: The Molecular Signatures Database Hallmark Gene Set Collection, *Cell Systems*, Vol. 1, No. 6, pp. 417–425 (online), DOI: 10.1016/j.cels.2015.12.004 (2015).

[4] Anaya, J.: OncoLnc: linking TCGA survival data to mRNAs, miRNAs, and lncRNAs, *PeerJ Computer Science*, Vol. 2, p. e67 (online), DOI: 10.7717/peerj-cs.67 (2016).

[5] Vlachos, I. S., Zagganas, K., Paraskevopoulou, M. D., Georgakilas, G., Karagkouni, D., Vergoulis, T., Dalamagas, T. and Hatzigeorgiou, A. G.: DIANA-miRPath v3.0: deciphering microRNA function with experimental support, *Nucleic Acids Research*, Vol. 43, No. W1, pp. W460–W466 (online), DOI: 10.1093/nar/gkv403 (2015).

[6] Tusher, V. G., Tibshirani, R. and Chu, G.: Significance analysis of microarrays applied to the ionizing radiation response, *Proceedings of the National Academy of Sciences*, Vol. 98, No. 9, pp. 5116–5121 (online), DOI: 10.1073/pnas.091062498 (2001).

[7] Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. and Smyth, G. K.: limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Research*, Vol. 43, No. 7, pp. e47–e47 (online), DOI: 10.1093/nar/gkv007 (2015).