Vol.2019-BIO-60 No.2
2019/12/2

テンソル分解を用いた教師なし学習による変数選択法のＴＣＧＡデータベースにおける卵巣がんデータのmicroRNA発現プロファイルとメチル化プロファイルへの適用

田口 善弘[1,a)]　呉 家樂[2,b)]

概要：一般に、microRNA の発現量とタンパク質コード遺伝子のプロモーターメチル化を同じサンプルについて計測することは可能であるので、相関関数を計算することも数学的には可能である。但し、生物学的にこの２つの観測量が直接関係しているとは限らないので、生物学的に意味がある microRNA とタンパク質コード遺伝子のペアを探すのは難しいように思われる。本研究では、最近、我々が提案している「テンソル分解を用いた教師なし学習による変数選択法」を用いるとこの様なことが可能であることを示す。

1. はじめに

いわゆるマルチオミックスデータ解析の発展に伴い、多くのオミックスデータが並行して観測されることも増えてきた。しかし、それらの解析を行う場合にはある程度生物学的なバックグラウンドを仮定した解析が行われることが多い。例えば、タンパク質コード遺伝子の発現量とその遺伝子のプロモーターのメチル化の統合解析、とか、microRNA(miRNA) とその標的遺伝子の統合解析、などである。これにはいろいろな理由があるが、１つには生物学的な背景を元にある程度関係性の絞り込みを行わないと解析が困難であるということがある。

例えば、遺伝子の数が N 個、あった場合、ある遺伝子のプロモーターメチル化と別の遺伝子の発現量が関係しているかを調べたい、とすると調べるべきペアの数は N^2 個に昇る。N は通常数万個あるため、ペアの数は数億ペアになってしまう。この場合、例えば、相関係数が有意に非ゼロのペアを調べたいとするといろいろな問題が起きる。仮にサンプル数が少ないとなると、かなり大きな相関係数でないかぎり、多重比較補正を考えると有意でないことになり、実際には関係があっても無くなってしまう。一方、サンプル数が多すぎると、非常に小さな相関係数でも有意に非ゼロになってしまう。実際、間接的には任意の遺伝子のプロモーターメチル化と別の任意の遺伝子の発現量は、なんらかの関係のがあるであろうから、これは統計的には正しいが、そうなると実際に直接関係しているペアを選択したいという目的から外れている。

本研究では、最近提案した「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、卵巣がんの microRNA の発現量とタンパク質コード遺伝子のプロモーターメチル化のデータを TCGA からダウンロードして解析し、生物学的に意味があるペアを選択したことを報告する。

2. 方法

2.1 TCGA

データは TCGA からダウンロードした。8 個の正常卵巣、569 個の卵巣がんプロファイルからなるデータで、計 577 個である。プロモーターメチル化は 24906 遺伝子に対して、miRNA 発現量は 723miRNA に対して計測された。

2.2 テンソル分解

まず、テンソル分解について述べる。テンソル分解にはいろいろな種類があるが [3]、本稿ではタッカー分解と呼ば

[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]（プレプリは https://doi.org/10.1101/380071）。また、刊行済みの書籍 [2] の §7.9 の内容にも基づいている。

れる分解を用いる。簡単のため三相のテンソルを用いて説明するが、四相以上の場合への拡張は自明であろう。要素が $x_{ijk} \in \mathbb{R}^{N\times M\times K}$ であるような三相のテンソル $\mathcal{X}$ を考える。この時、タッカー分解は

$$x_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{M}\sum_{\ell_3=1}^{K} G(\ell_1,\ell_2,\ell_3)u_{\ell_1 i}u_{\ell_2 j}u_{\ell_3 k}$$

で定義される。$G \in \mathbb{R}^{N\times M\times K}$ はコアテンソル、$u_{\ell_1 i} \in \mathbb{R}^{N\times N}, u_{\ell_2 j} \in \mathbb{R}^{M\times M}, u_{\ell_3 k} \in \mathbb{R}^{K\times K}$ は特異値行列である。特異値行列は直交行列である。まず、これは明らかに過完備であり、ユニークな答えは存在しない。したがってタッカー分解の結果はタッカー分解を実行する具体的なアルゴリズムに依って変わってしまう。ここでは高次特異値分解 [3] と呼ばれるアルゴリズムを採用する。高次特異値分解で得られたタッカー分解では一般にコアテンソルは対角テンソルではない。従って、特異値ベクトル $\boldsymbol{u}_{\ell_1} \in \mathbb{R}^N, \boldsymbol{u}_{\ell_2} \in \mathbb{R}^M, \boldsymbol{u}_{\ell_3} \in \mathbb{R}^K$ は様々な組み合わせで複数回掛けあわされた上で和が取られる。

2.3 行列からのテンソル作成

本研究で実際にテンソル分解されるテンソルは、メチル化を表現する行列 $x_{ij} \in \mathbb{R}^{N\times M}$ (i はプロモーターメチル化を計測した遺伝子を、j はサンプルを表す) と、miRNA の発現量を表す行列 $x_{ik} \in \mathbb{R}^{N\times K}$ (k は miRNA を、j はサンプルを表す) から作られた三相のテンソル $\tilde{x}_{ijk} = x_{ij}x_{ik} \in \mathbb{R}^{N\times M\times K}$ を作ってこれをテンソル分解することにする。しかし、ここで $N = 24906, M = 577, K = 723$ であることを考えるとこのテンソルをこのままテンソル分解して

$$\tilde{x}_{ijk} = \sum_{\ell_1=1}^{N}\sum_{\ell_2=1}^{M}\sum_{\ell_3=1}^{K} G(\ell_1,\ell_2,\ell_3)u_{\ell_1 i}u_{\ell_2 j}u_{\ell_3 k} \qquad (1)$$

を求めるのは容易ではない。通常の数百 GB 程度のメモリーしか積んでいない計算機ではメモリーに入り切らない。

そこで以下のような近似を行う。$x_{ik} = \sum_{j=1}^{M}\tilde{x}_{ijk} \in \mathbb{R}^{N\times K}$ を計算し、x_{ik} に対して、テンソル分解ならぬ特異値分解

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

を実行することで、(プロモーターメチル化を計算した) 遺伝子、及び、miRNA に対する特異値ベクトル $u_{\ell i}, u_{\ell k}$ をそれぞれ計算するのである。ここでサンプル j に対する和を取ってしまったため、サンプルに対する特異値ベクトル $u_{\ell j}$ を計算することができない。これは以下の近似式で計算する。

$$u^{\mathrm{methyl}}_{\ell j} = \sum_{i=1}^{N} u_{\ell i}x_{ij} \qquad (2)$$

$$u^{\mathrm{miRNA}}_{\ell j} = \sum_{k=1}^{K} u_{\ell k}x_{ik} \qquad (3)$$

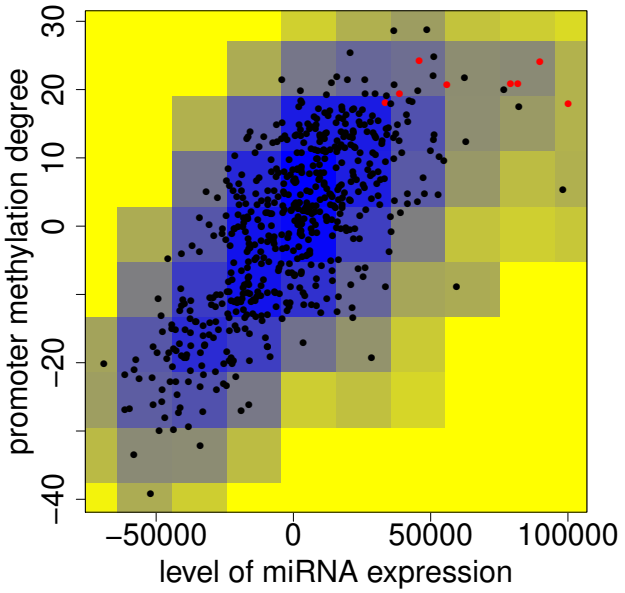

図 1 サンプルに付与された特異値ベクトル $u^{\mathrm{methyl}}_{\ell j}$ (縦軸) と $u^{\mathrm{miRNA}}_{\ell j}$ (横軸)。黒丸：腫瘍、赤丸：正常細胞

Fig. 1 Singular value vectors, $u^{\mathrm{methyl}}_{\ell j}$(vertical axis) and $u^{\mathrm{miRNA}}_{\ell j}$(horizontal axis) that are attributed to samples. Black circile : tumors, red circles : normal tissue

2.4 変数選択

今回の解析の場合は、プロモーターメチル化、miRNA 発現量共に、正常臓器と腫瘍の間の有意差を持つようなものを選ぶ必要がある。そこで、正常臓器と腫瘍の間の有意差を持つ特異値ベクトル $u^{\mathrm{methyl}}_{\ell j}, u^{\mathrm{miRNA}}_{\ell j}$ が見つかったとしよう。これらと相補的な遺伝子特異値ベクトル $u_{\ell i}$ と miRNA ベクトル $u_{\ell k}$ を用いて、正常臓器と腫瘍の間の差に寄与する (プロモーターメチル化を計測した) 遺伝子と miRNA の組み合わせが選択できる。

具体的には $u_{\ell i}$ と $u_{\ell k}$ がガウス分布していることを仮定し、χ 二乗分布を用いて

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell i}}{\sigma^{\mathrm{methyl}}_\ell}\right)^2\right], P_k = P_{\chi^2}\left[> \left(\frac{u_{\ell k}}{\sigma^{\mathrm{miRNA}}_\ell}\right)^2\right]$$

の様にして遺伝子 i 及び miRNAk に与えられた P 値、P_i 及び P_k を、Benjamin-Hochberg で多重補正して、補正 P 値が 0.01 以下である遺伝子と miRNA を選択する。ここで $P_{\chi^2}[> x]$ は引数が x 以上である χ 二乗分布の累積確率分布、$\sigma^{\mathrm{methyl}}_\ell, \sigma^{\mathrm{miRNA}}_\ell$ は標準偏差である。

3. 結果

図 1 は $\ell = 2$ の時の特異値ベクトル $u^{\mathrm{methyl}}_{\ell j}, u^{\mathrm{miRNA}}_{\ell j}$ の散布図である。この 2 つのベクトルは両方共、サンプルに対して付与されたものであり、(1) 式で計算をするならば本来は $u_{\ell_2 j}$ という単一の特異値ベクトルが求まらないといけないところが近似的に x_{ik} の特異値分解を経て (2) 式と (3) 式から再計算したために二重に計算されてしまったも

のである。従って、$u_{\ell j}^{\mathrm{methyl}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ が全く別のものだったらそもそも近似は破綻している（し遺伝子のプロモーターメチル化と、miRNA の発現量には相関はない）ことになる。

幸いにも図 1 を見る限りではこの 2 つはよく相関している。実際、相関係数を計算すると０．７２であり、サンプル数が５７７個もあることを考慮すると $P = 2.0 \times 10^{-92}$ という非常に小さな P 値を得ることができる。そういう意味では目論見通り一致度の高い $u_{\ell j}^{\mathrm{methyl}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ が得られた。

次に $u_{\ell j}^{\mathrm{methyl}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ が正常細胞と腫瘍の間で有意差があるかが重要である。我々が知りたいのはどの遺伝子のプロモーターメチル化と miRNA 発現量が協働的に正常細胞と腫瘍の差に貢献しているかを知りたいからだ。TCGA からのデータは正常例が非常に少ない偏ったデータであることが多い。今回の卵巣がんの場合にも偏りは非常に大きい。５７７サンプルのうち、正常例は 8 例しか無い。この様な場合、有意差があることを統計的に示すことができない場合が多い。だが、幸いにもこの場合は、実際に t 検定で二群の差を検定してみると $u_{\ell j}^{\mathrm{methyl}}$ に対しては $P = 1.2 \times 10^{-11}$、$u_{\ell j}^{\mathrm{miRNA}}$ に対しては $P = 1.3 \times 10^{-4}$ と十分に小さい P 値をえることができた。このことから、得られたサンプル特異値ベクトル $u_{\ell j}^{\mathrm{methyl}}$ と $u_{\ell j}^{\mathrm{miRNA}}$ は、お互いの相関が大きいというだけではなく、正常細胞と腫瘍の間で差があるという意味でも非常に意味があるものが得られたと言えよう。

次に遺伝子特異値ベクトル $u_{\ell i}$ と miRNA ベクトル $u_{\ell k}$ に立ち戻り、第 2.4 節方法に書かれた方法で遺伝子と miRNA を選択したところ、それぞれ、２４１遺伝子（一覧は原著論文 [1]）と 7 miRNA（具体的な miRNA 名は後述）を選ぶことができた。

次に確認すべきは、これらの遺伝子のプロモーターメチル化 x_{ij} と miRNA の発現量 x_{kj} が正常細胞と腫瘍で本当に差があるかをであるが、検定を行ったところ、確かに差があることが解った。

最後に２４１遺伝子のプロモーターメチル化 x_{ij} と 7 miRNA の発現量 x_{kj} が相関しているかどうかを確認した（**表 1**）。幸いにも全 $1687 = 7 \times 241$ 個の遺伝子–miRNA ペアの内、94%にあたる 1592 ペアが有意に相関していた。この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は、腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択することができていることが解るだろう。

最後にこの「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」がいかに難しいかを他の伝統的な方法で試みることで実際に考えてみよう。まずは単純に個々の遺伝子とプロモーターメチル化と

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で選択された遺伝子と miRNA の内、遺伝子のプロモーターメチル化と miRNA の発現量が有意な相関（補正 P 値が０．０１以下）を示す miRNA–遺伝子ペアの数

Table 1 The number of miRNA–gene pairs showing a significant correlation (adjusted P-vales less than 0.01) when they are identified with TD-based unsupervised FE.

		negative correlation	
		T	F
positive	T	0	985
correlation	F	607	95

個々の miRNA の発現量に独立に t 検定を適用して P 値を計算し、BH 基準で多重比較補正した後、P 値が 0.01 以下の遺伝子と miRNA を選択してみた。その結果、19395 遺伝子と 214miRNA という非常に多くの遺伝子と miRNA が選ばれてしまった。t 検定による P 値は基本的にサンプル数依存である。理論的にはサンプル数無限大の場合には全ての遺伝子のプロモーターメチル化と全ての miRNA の発現量が正常細胞と腫瘍で有意差があるという結果になってしまうことが避けられない。これは勿論、我々が知りたい「生物学的には腫瘍と正常細胞の差に関係しているのはどの遺伝子のメチル化とどの miRNA の発現量なのか？」という問の答えとしては妥当ではない。今回はサンプル数が 577 個と非常に多いため、それが悪い方に働いて、十分な性能のスクリーニングが行えなかった、とみなすべきだろう。

通常であればここで更にフォールドチェンジ（FC）などを考慮して絞り込みを行うべきであるが、どのくらいの大きさの FC が適当なのかなどの任意性があり、また、テンソル分解を用いた教師なし学習による変数選択は P 値だけで十分に絞り込みが行えているところ、また、ここでの目的なあくまでテンソル分解を用いた教師なし学習による変数選択の既存手法に対する有意性を示すことだけであることをかんがえると、今回はこれ以上先に踏み込まないことが妥当であろう。

次に t 検定で選択された 19395 遺伝子のプロモーターメチル化と 214miRNA の発現量が有意に相関しているかを考える（**表 2**）。残念ながら全ペアのうち１３％しか有意の相関をしていないことが解る。従って、この方法では「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」ことは難しいことが解る。

それではバカ正直に t 検定で腫瘍と正常細胞で有意差がある全ての遺伝子と miRNA を考慮するのではなく、有意差が上位の遺伝子と miRNA にだけ絞り込んだらもうちょっと相関が高くなるのではないだろうか？そこでテン

表 2　t 検定においてプロモーターメチル化が腫瘍と正常細胞で有意差があるとされた 19395 遺伝子と発現量が同じく腫瘍と正常細胞で有意差があるとされた 214miRNA のペアの内、相関が有意である（補正 P 値で 0.01）のペアの数。

Table 2　The number of miRNA–gene pairs showing a significant correlation (adjusted P-vales less than 0.01) when these are identified by Student's t test.

		negative correlation	
		T	F
positive	T	0	329896
correlation	F	225495	3595139

表 3　t 検定で選択されたもののうち、上位 7miRNA と 241 遺伝子に限定した場合の、有意に相関している（相関係数で計算した補正 P 値が 0.0 以下）miRNA-遺伝子ペアの数

Table 3　The number of miRNA–gene pairs showing a significant intrapair correlation (adjusted P-vales less than 0.01) when only seven top-ranked miRNAs and 241 top-ranked genes selected by Student's t test are considered.

		negative correlation	
		T	F
positive	T	0	13
correlation	F	28	1646

ソル分解を用いた教師なし学習による変数選択で選択された数と同じ数の 7miRNA と 241 遺伝子を「より P 値の小さい（＝有意差が大きい)」という基準で意図的に選択して、相関がより強くなるかを調べた（**表 3**)。残念ながら上位の遺伝子と miRNA に絞り込みを行っても有意に相関しているペアの割合は増えず、むしろ３％に減ってしまった。そもそも、t 検定には正常細胞と腫瘍の間で差があるものを見つける能力はあっても、まったく無関係（だとされる）遺伝子のプロモーターメチル化と miRNA の発現量の間の相関を見つける能力は無いだろう。このように従来の方法では「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」が難しいからこそ、ずっとそのような関係が見逃されてきたのだとも言える。

最後に、そもそも t 検定から出発するからいけないので、まず遺伝子のプロモーターメチル化と miRNA の発現量を見て、相関があるペアを絞り込めば、自動的に miRNA と遺伝子のペアも絞り込め、そのペアに含まれている miRNA と遺伝子だけを対象にして正常細胞と腫瘍の間の差を t 検定すれば「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」のではないか、という考えはどうだろう（**表 4**）？ 全遺伝子のプロモーターメチル化と全 miRNA の発現量の間の相関係数を計算したところ、全体のわずか７％のペアしか有意な相関をもっていないことが解った。一見、この戦略はうまく行ったように思える。しかし、実際に選ばれたペアをよく見てみるとこの印象は覆される。なんと全ての遺伝子、全ての miRNA がたった７％しか選ばれていないはずのペアのどれかには含まれてしまっているのである。これでは相関係数の有意性から遺伝子や miRNA を全く絞り込めないことになる。またも５７７個と言うサンプル数の多さが裏目にでてしまった。これでは最初から全ての遺伝子と miRNA を対象に t 検定で正常細胞と腫瘍で有意差のある遺伝子や miRNA を選ぶとの何もかわならくなってしまう。

表 4　遺伝子のプロモーターメチル化と miRNA の発現量が有意に相関（補正 P 値が 0.01 以下）している miRNA–遺伝子ペアの数

Table 4　The number of miRNA–gene pairs (an expression level and degree of promoter methylation, respectively) showing a significant intrapair correlation (adjusted P-vales less than 0.01).

		negative correlation	
		T	F
positive	T	0	608989
correlation	F	588783	16809266

この様に「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」ことは案外難しく、難しいことを考えなくてもさっくりそれが実行でき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は非常に優れた方法であると言うことができるだろう。

次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で選択された 7miRNA と２４１遺伝子の生物学的な意味を精査しよう。あまりそういうことは考えにくいとはいうもの、「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」ことに成功したとしてもそれはなんらかのアーティファクトで生物学的に無意味だ、という可能性もあるからだ。そこで選択された hsa-miR-142-3p, hsa-miR-142-5p, hsa-miR-150,hsa-miR-21*, hsa-miR-22, hsa-miR-224, hsa-miR-96 の７つの miRNA を DIANA-mirpath にアップトードして KEGG パスウェイエンリッチメント解析を行った（**表 5**)。その結果、６６個ものパスウェイのエンリッチメントが確認された（補正 P 値で０．０５以下)。特にそのうちの１５パスウェイはがん関連のものである。このことから、少なくとも選ばれた 7miRNA が生物学的に無意味だとは言えないと解った。

次に２４１個の遺伝子を MSigDB [4] にアップロードし

表 5 7miRNA を DIANA-miRPath にアップロードして得られた KEGG パスウェイエンリッチメントの一覧。g #:遺伝子数,m #:miRNA 数、P 値は補正 P 値である。

Table 5 Enriched KEGG pathways detected by DIANA-miRPath among seven miRNAs. Bold ones are cancer related. g #: the number of genes, m #: the number of related miRNAs; *p*-values are adjusted.

KEGG pathway	p-value	g #	m #
Viral carcinogenesis	4.17E-11	91	7
Proteoglycans in cancer	4.17E-11	86	7
Prion diseases	4.87E-09	10	6
Adherens junction	4.87E-09	41	7
Renal cell carcinoma	4.87E-09	38	7
Bacterial invasion of epithelial cells	2.79E-08	41	7
Central carbon metabolism in cancer	4.84E-08	37	7
Hippo signaling pathway	5.90E-08	57	7
Cell cycle	7.20E-08	62	7
TGF-beta signaling pathway	8.55E-08	37	7
Fatty acid biosynthesis	1.61E-07	4	4
Glycosaminoglycan biosynthesis - keratan sulfate	2.71E-07	8	6
Hepatitis B	1.69E-06	60	7
Prostate cancer	3.75E-06	46	7
Shigellosis	5.12E-06	33	6
Pathogenic Escherichia coli infection	8.67E-06	33	7
Pancreatic cancer	1.51E-05	34	7
Fatty acid metabolism	2.08E-05	14	5
FoxO signaling pathway	3.02E-05	59	7
Protein processing in endoplasmic reticulum	5.58E-05	74	7
Regulation of actin cytoskeleton	5.58E-05	83	7
p53 signaling pathway	5.72E-05	36	7
HIF-1 signaling pathway	6.18E-05	50	7
2-Oxocarboxylic acid metabolism	2.20E-04	9	5
Lysine degradation	2.20E-04	19	7
Oocyte meiosis	2.31E-04	45	7
Ubiquitin mediated proteolysis	3.30E-04	55	7
Endocytosis	4.41E-04	81	7
SNARE interactions in vesicular transport	4.44E-04	17	7
Colorectal cancer	6.24E-04	31	7
Endometrial cancer	9.95E-04	25	6

た（**表 6**）。MSigDB は発現量の差に特化したエンリッチメント解析なので今回の結果の解析に向いている。その結果、がん関連の３３の遺伝子セットが有意に（$FDR < 0.05$）重なっていたが（ここには上位２０位まで）、このことはちゃんとがん関連の遺伝子が選ばれており、「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」が単なるアーティファクトではない生物学的に意味がある遺伝子を選択していることを示している。

4. おわりに

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 [2] を用いて、卵巣がんにおけるタンパクコード遺伝子のプロモーターメチル化と miRNA の発現量の相関について調べた。通常の方法ではこのような「腫瘍と正常細胞で有意差があり、かつ、互いにプロモーターのメチル化と発現量がそれぞれ相関している、遺伝子と miRNA のペアを選択する」ペアを見つけるのは難しく、また、選ばれたペアはがんに関連したものがちゃんと選ばれていることを確認した。

謝辞 データは TCGA Research Network: https://www.cancer.gov/tcga からダウンロードした。

参考文献

[1] Taguchi, Y.-H. and Ng, K.-L.: Tensor Decomposition-Based Unsupervised Feature Extraction for Integrated Analysis of TCGA Data on MicroRNA Expression and

表 6 MSigDB の C6: oncogenic signatures と選択された２４１遺伝子との重なり。#G1 (K):各カテゴリの遺伝子数、#G2 (k): ２４１遺伝子との重なり

Table 6 Overlaps between C6: oncogenic signatures in MSigDB and gene symbols associated with genes identified by TD-based unsupervised FE as showing differential promoter methylation between normal ovarian tissues and tumor tissues. #G1 (K): the number of genes in each overexpressed genes set. #G2 (k): overlaps with genes selected by TD-based unsupervised FE.

Gene Set Name	#G1 (K)	Description	# G2 (k)	k/K	p-value	FDR q-value
WNT_UP.V1_DN	170	Genes downregulated in C57MG cells (mammary epithelium) by overexpression of WNT1 [Gene ID=7471] gene.	9	0.0529	3.63E-07	4.58E-05
KRAS.600.LUNG.BREAST_UP.V1_UP	288	Genes upregulated in epithelial lung and breast cancer cell lines overexpressing an oncogenic form of KRAS [Gene ID=3845] gene.	11	0.0382	4.85E-07	4.58E-05
RPS14_DN.V1_UP	192	Genes upregulated in CD34+ hematopoietic progenitor cells after a knockdown of RPS14 [Gene ID=6208] by RNA interference (RNAi).	9	0.0469	1.01E-06	5.64E-05
VEGF_A_UP.V1_UP	196	Genes upregulated in HUVEC cells (endothelium) by treatment with VEGFA [Gene ID=7422].	9	0.0459	1.19E-06	5.64E-05
MEL18_DN.V1_UP	141	Genes upregulated in DAOY cells (medulloblastoma) upon a knockdown of PCGF2 [Gene ID=7703] by RNAi.	7	0.0496	1.13E-05	4.26E-04
KRAS.LUNG.BREAST_UP.V1_UP	145	Genes upregulated in epithelial lung and breast cancer cell lines overexpressing an oncogenic form of KRAS [Gene ID=3845] gene.	6	0.0414	1.32E-04	3.37E-03
KRAS.300_UP.V1_UP	146	Genes upregulated in four lineages of epithelial cell lines overexpressing an oncogenic form of KRAS [Gene ID=3845] gene.	6	0.0411	1.37E-04	3.37E-03
BMI1_DN.V1_UP	147	Genes upregulated in DAOY cells (medulloblastoma) upon a knockdown of BMI1 [Gene ID=648] by RNAi.	6	0.0408	1.43E-04	3.37E-03
KRAS.600_UP.V1_UP	287	Genes upregulated in four lineages of epithelial cell lines overexpressing an oncogenic form of KRAS [Gene ID=3845] gene.	8	0.0279	1.64E-04	3.44E-03
CAHOY_ASTROGLIAL	100	Genes upregulated in astroglia cells.	5	0.05	2.02E-04	3.82E-03
ATF2_UP.V1_DN	187	Genes downregulated in myometrial cells overexpressing ATF2 [Gene ID=1386] gene.	6	0.0321	5.19E-04	7.71E-03
CYCLIN_D1_UP.V1_UP	188	Genes upregulated in MCF-7 cells (breast cancer) overexpressing CCND1 [Gene ID=595] gene.	6	0.0319	5.33E-04	7.71E-03
SRC_UP.V1_UP	188	Genes upregulated in primary epithelial breast cancer cell culture overexpressing SRC [Gene ID=6714] gene.	6	0.0319	5.33E-04	7.71E-03
PTEN_DN.V1_UP	191	Genes upregulated upon a knockdown of PTEN [Gene ID=5728] by RNAi.	6	0.0314	5.80E-04	7.71E-03
MTOR_UP.N4.V1_DN	193	Genes downregulated in CEM-C1 cells (T-CLL) by rapamycin (sirolimus) [PubChem = 6610346], an mTOR pathway inhibitor.	6	0.0311	6.12E-04	7.71E-03
KRAS.LUNG_UP.V1_UP	141	Genes upregulated in epithelial lung cancer cell lines overexpressing an oncogenic form of KRAS [Gene ID=3845] gene.	5	0.0355	9.74E-04	1.15E-02
ALK_DN.V1_UP	145	Genes upregulated in DAOY cells (medulloblastoma) after a knockdown of ALK [Gene ID=238] by RNAi.	5	0.0345	1.10E-03	1.16E-02
BMI1_DN_MEL18_DN.V1_UP	145	Genes upregulated in DAOY cells (medulloblastoma) upon a knockdown of BMI1 and PCGF2 [Gene ID=648, 7703] by RNAi.	5	0.0345	1.10E-03	1.16E-02
P53_DN.V2_UP	148	Genes upregulated in HEK293 cells (kidney fibroblasts) upon a knockdown of TP53 [Gene ID=7157] by RNAi.	5	0.0338	1.21E-03	1.20E-02
KRAS.50_UP.V1_UP	48	Genes upregulated in four lineages of epithelial cell lines overexpressing an oncogenic form of KRAS [Gene ID=3845] gene.	3	0.0625	2.14E-03	2.03E-02

Promoter Methylation of Genes in Ovarian Cancer, *2018 IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 195–200 (online), DOI: 10.1109/BIBE.2018.00045 (2018).

[2] Taguchi, Y.-H.: *Unsupervised Feature Extraction Applied to Bioinformatics, A PCA Based and TD Based Approach*, Springer Nature International (2020).

[3] 石黒勝彦，林　浩平：関係データ学習 (機械学習プロフェッショナルシリーズ)，講談社 (2016).

[4] : MSigDB, The GSEA/MSigDB Team (online), available from ⟨http://software.broadinstitute.org/gsea/msigdb/annotate.jsp⟩ (accessed 2019-11-04).