Vol.2018-BIO-56 No.2
2018/12/14

筋萎縮性側索硬化症のための マイクロRNAバイオマーカーの探索

田口善弘[1,a)]　王秀瑛[2,b)]

概要：筋萎縮性側索硬化症（ ALS）は、広く利用可能な有効な治療法がない重度の神経変性疾患の 1 つである。病気が進行するにつれて、患者は自発的な筋肉の制御を失う。ニューロンの変性がこの疾患の原因であるにもかかわらず、機能不全のメカニズムは未だ不明である。 ALS を開始および進行させる遺伝子機構を捜すために、この疾患とのマイクロ RNA（ miRNA）発現の関連性が考慮された。健常者、孤発性 ALS（ sALS)、家族性 ALS（ fALS）および ALS 変異キャリアからの血清 miRNA を調べた。これらの血清 miRNA プロファイルには、主成分分析を用いた教師無し学習による変数選択を適用した。その結果、我々は、高精度で健常者と患者を区別することができる miRNA を予測した。したがって、これらの miRNA は、ALS の潜在的な予後の miRNA バイオマーカーであり得る。

1. はじめに

ALS(Amyotrophic lateral sclerosis, 筋萎縮性側索硬化症) [1] は治癒のための有効な治療法が確立されていない神経変異性疾患である。治療法が確立しない原因は多岐に渡るが、大きな問題として再生しない運動ニューロンが病変部位であるため、病変部位の組織標本をとることが難しく、そもそも、実験的に病気の原因を確定するのが難しいという問題がある。このため、近年の大規模なマルチオミックスデータの取得技術の発展にも関わらず、病変部位のマルチオミックスデータの取得さえままならないというのが現状となっている。このような事情のため、マルチオミックスデータの取得は、組織標本の取得が可能な、患者血液などが主となっているが、それでも、どのような成分を分析すればそもそもバイオマーカーとして有効なのかさえ、コンセンサスがないのが現状である。

本研究では、過去に取得された家族型 ALS, 孤発型 ALS, 遺伝子変異保持者（ 未発症)、健常者、の4 グループの血中 microRNA のデータ [3] を用いてバイオマーカーを構成することを目指す。治療法の確立されていない同疾患に対して、取得が容易な血中 miRNA を用いたバイオマーカーが構成できれば、実用的な意味はあるし、また、血中であっても発現量の変化する miRNA が特定できれば、その標的遺伝子の解析などから病因を同定できる可能性もある。

解析手法としては著者がかねてから提案している「 主成分分析を用いた教師なし学習による変数選択法」とカテゴリ回帰を用いた（ 原著論文 [2] では前者だけを用いて医学･生物学的な観点から主に検証を行ったがここでは読者の興味に沿って内容をやや変更した)。

2. データと方法

全体的なデータ解析フローを図 **1** に示す。

2.1 miRNA 発現プロファイルデータ

解析対象の miRNA 発現プロファイルデータ [3] は GEO の GSE52917 からダウンロードした。具体的には GSE52917_series_matrix.txt.gz という名前のファイルが “Series matrix” に存在し、これを用いた。このファイルには5３個の血清 miRNA プロファイルが含まれている。その内訳は、家族性 ALS が6，遺伝子変異を保持しているが未発症のもの1２、孤発性 ALS が1８，健常者が1７となっている。またこのデータには 3391 種類の miRNA が含まれている。

2.2 カテゴリ回帰による変数選択

$x_{ij} \in \mathbb{R}^{N\times M}$ を i 番目の miRNA の j 番目のサンプルにおける miRNA 発現プロファイルとする。カテゴリ回帰ではこれを

$$x_{ij} = a_i + \sum_k b_{ik}\delta_{kj}$$

[1] 中央大学理工学部物理学科
[2] 國立交通大學統計學研究所
[a)] tag@granular.com
[b)] wang@stat.nctu.edu.tw

本研究は原著論文として出版済みである。[2]

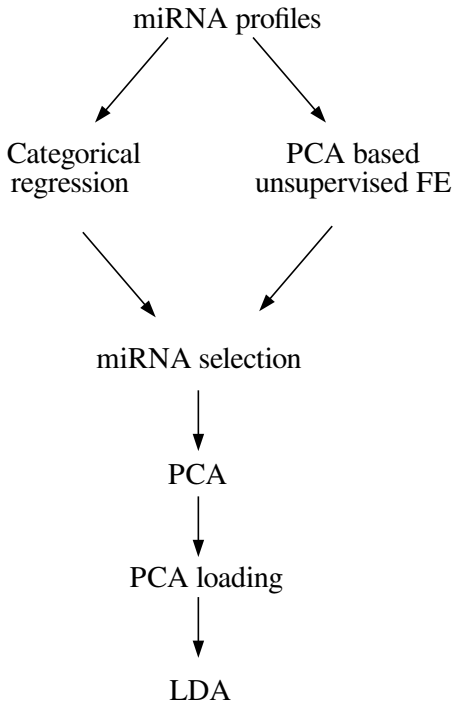

図 1　本研究のデータ解析フロー。LDA：線形判別、PCA：主成分分析

Fig. 1　Data analysis flow in this study. PCA: principal component analysis, LDA: liner discriminat analysis.

という式で回帰する。a_i, b_{ik} は i 番目の miRNA 固有の回帰係数であり δ_{jk} は j サンプルが k 番目のカテゴリに入ったときだけ1 、それ以外は0 になる関数である。R に実装されている回帰計算の関数である lm を用いて各 miRNA ごとに P 値を計算、求まった P 値を BH 基準 [4] で多重比較補正して、補正 P 値が 0.01 以下である miRNA を選択した。

2.3　主成分分析を用いた教師なし学習による変数選択

まず x_{ij} を $\sum_i x_{ij} = 0, \sum_i x_{ij}^2 = N$ になるように標準化する。次に $S_{ii'} = \sum_j x_{ij} x_{i'j}$ で定義される行列を対角化し、固有ベクトル $\boldsymbol{u}_\ell \in \mathbb{R}^N$ を計算することで主成分得点を計算し、主成分分析を実行する。主成分負荷量 $\boldsymbol{v}_\ell \in \mathbb{R}^M$ は $v_{\ell j} = \sum_i u_{\ell i} x_{ij}$ で計算する。次に今度は主成分負荷量に対するカテゴリ回帰

$$v_{\ell j} = a_\ell + \sum_k b_{\ell k} \delta_{kj}$$

を実行する。$a_\ell, b_{\ell k}$ は第 ℓ 主成分固有の回帰係数である。計算した P 値を多重比較補正し、補正 P 値が 0.05 以下の ℓ を選ぶ。この ℓ を用いて、$u_{\ell i}$ がガウス分布であるという帰無仮説のもとに、χ 二乗分布を使って i 番目の miRNA に P 値 P_i を

$$P_i = P_{\chi^2}\left[> \left(\frac{u_{\ell i}}{\sigma_\ell} \right)^2 \right] \tag{1}$$

という式で付与する。ここで σ_ℓ は標準偏差、$P_{\chi^2}[> x]$ は引数が x より大きい時の χ 二乗分布の累積確率である。多重比較補正した P 値が 0.01 以下の miRNA を選択する。

2.4　PCA を用いた線形判別

選択された miRNA だけを用いて再度 PCA を行い、主成分負荷量 $\boldsymbol{v}_{\ell j}$ を再計算する。次に主成分負荷量に対するカテゴリ回帰

$$v_{\ell j} = a_\ell + \sum_k b_{\ell k} \delta_{kj}$$

を再度、実行する。計算した P 値を多重比較補正し、補正 P 値が 0.05 以下の ℓ を選ぶ。選ばれた $\boldsymbol{v}_\ell$ を用いて、R に実装された lda 関数を用いて線形判別を実行する。prior = rep(1/4,4) と CV=T のオプションを指定し、4 群の重みが同じであることと、leave one out cross validation を要請し、判別能を測定する。

3.　結果

3.1　カテゴリ回帰による変数選択

カテゴリ回帰を用いた変数選択の計算結果について述べる。カテゴリ回帰によって選択された miRNA はは全部で９０ miRNA であった(表 **1**)。この miRNA がバイオマーカーになっているかどうかを確認したいがサンプルが５３個しかないため、このままでは必ず4 群に分けることができてしまい、意味がない。そこで自由度を低減するため、選択された 90miRNA のみを用いて再度主成分分析を行い、主成分負荷量 $\boldsymbol{v}_\ell$ を計算してカテゴリ回帰を当てはめたところ、$\ell = 1, 2, 5, 8$ が 0.05 以下の補正 P 値を持っていることが解った (図 **2**)。この4 つの主成分得点を用いて５３サンプルを四群に判別分析した結果が表 **2** である。孤発性 ALS 患者以外はよく判別されている。

3.2　主成分分析を用いた教師なし学習による変数選択

主成分分析を行った結果、第二主成分得点がカテゴリ回帰で 0.05 以下の補正 P 値 ($P = 2.45 \times 10^{-5}$) を持っていることが解った (図 **3**)。第二主成分得点を用いて各 miRNA に付与される P 値を χ 二乗分布, (1) 式, で計算し、補正 P 値が 0.01 以下の miRNA を数えたところ、全部で 67miRNA がこの条件を満たしていた (表 1, 図 **4**)。この miRNA がバイオマーカーになっているかどうかを確認したいがサンプルが５３個しかないため、このままでは必ず4 群に分けることができてしまい、意味がない。そこで自由度低減のため、選択された 67miRNA のみを用いて再度主成分分析を行い、主成分負荷量 $\boldsymbol{v}_\ell$ を計算してカテゴリ回帰を当てはめたところ、$\ell = 1, 2, 3, 8$ が 0.05 以下の補正 P 値を持っ

表 1　選択された miRNA のリスト。太字は Liguori ら [5] が同定した孤発性 ALS で低下する３８個の miRNA とのオーバーラップ。

Table 1　List of selected miRNAs. Bold letters indicate ovelaps with 38 miRNAs downrefulated in sALS [5].

カテゴリ回帰を用いた変数選択固有 (68 miRNA)
mir-1271 mir-185 mir-1910 mir-297 mir-3611 mir-3689b mir-3689f mir-3937 mir-3960 mir-423 mir-4525 mir-4539 mir-4669 mir-466 **mir-550a-1** mir-550a-2 mir-550a-3 **miR-106b** miR-1268b miR-1273d miR-1285 miR-1290 miR-1296 miR-1322 miR-1538 miR-188-5p miR-1909-star miR-194-star miR-195-star miR-2278 miR-2392 miR-296-3p miR-3120-5p miR-32-star miR-320a miR-320b miR-320c miR-320d miR-338-5p miR-3655 miR-3663-5p miR-3907 miR-3960 miR-4271 miR-4429 miR-4440 miR-4462 miR-4493 miR-4521 miR-4530 miR-4647 miR-4667-5p miR-4701-3p miR-4710 miR-4728-5p miR-4746-3p miR-4769-5p miR-4783-3p miR-4793-3p miR-574-5p **miR-584** miR-610 miR-629 miR-642b miR-661 miR-760 miR-769-3p miR-99b-star
両法に共通 (22 miRNA)
miR-1180 miR-1275 miR-1306 **miR-130b** miR-134 miR-1469 miR-185 miR-1915 miR-2861 miR-297 miR-3064-5p miR-3665 miR-4306 miR-4455 miR-4497 miR-4656 miR-4745-5p miR-4787-5p miR-483-5p miR-595 miR-638 miR-665
主成分分析を用いた教師なし学習による変数選択 固有 (45 miRNA)
let-7a let-7b let-7c **let-7d miR-103a** miR-106a miR-107 miR-122 miR-1246 miR-1280 miR-1281 miR-140-3p miR-146a miR-151-3p **miR-151-5p miR-16** miR-1825 miR-191 miR-19b miR-2110 **miR-221 miR-22 miR-23a** miR-24 miR-25 miR-30d miR-3135b miR-3175 miR-3185 miR-320e miR-3613-5p **miR-425** miR-4454 miR-4466 miR-4485 miR-4488 miR-4532 miR-455-3p miR-4707-5p miR-4734 miR-652 miR-663 miR-92a **miR-93** miR-940

表 2　図 2 に示された4 つの主成分負荷量を用いた線形判別の混同行列。fALS:家族性 ALS, sALS:孤発性 ALS

Table 2　Confusion matrix of liner regression analysis using four PC loading shown in Fig. 2.

	正解			
予測	変異保持者	fALS 患者	健常者	sALS 患者
変異保持者	8	1	0	5
fALS 患者	2	5	0	0
健常者	1	0	16	8
sALS 患者	1	0	1	5

ていることが解った (図 5)。この4 つの主成分得点を用いて５３サンプルを四群に判別分析した結果が表 3 である。表 2 に比べると孤発性 ALS 患者の判別が大きく向上している。表 2 では、１８人の孤発性 ALS 患者のうち、５人しか孤発性 ALS 患者として予測でき無かった。４群の判別であることを考えると、これはほぼランダムと変わらない。しかし、表 3 では1 ８人中8 人と約半数が孤発性 ALS 患者であると予測できていることがわかる。

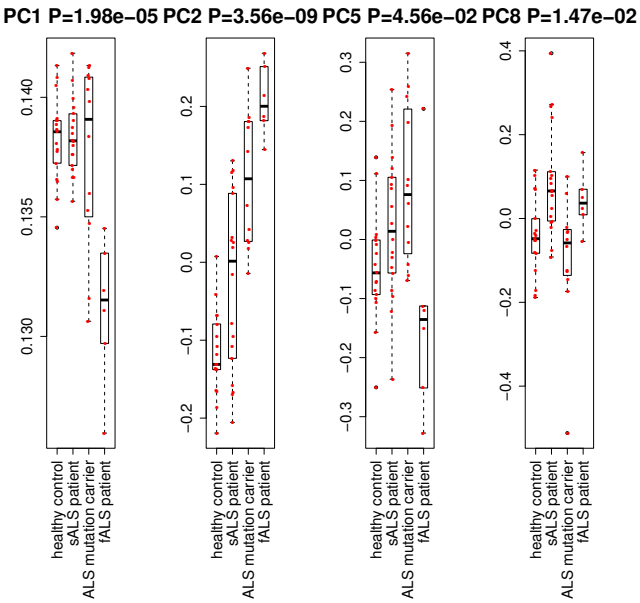

図 2　カテゴリ回帰で選ばれた 90miRNA(表 1) のみを用いて計算した結果で得られる主成分負荷量のうち、カテゴリ回帰で計算した補正 P 値が、0.05 以下であった第1，２，５，８主成分負荷量の箱ひげ図。

Fig. 2　Boxplots using four PC loading computed by PCA applied to 90 miRNAs selected by categorical regression (Table 1).

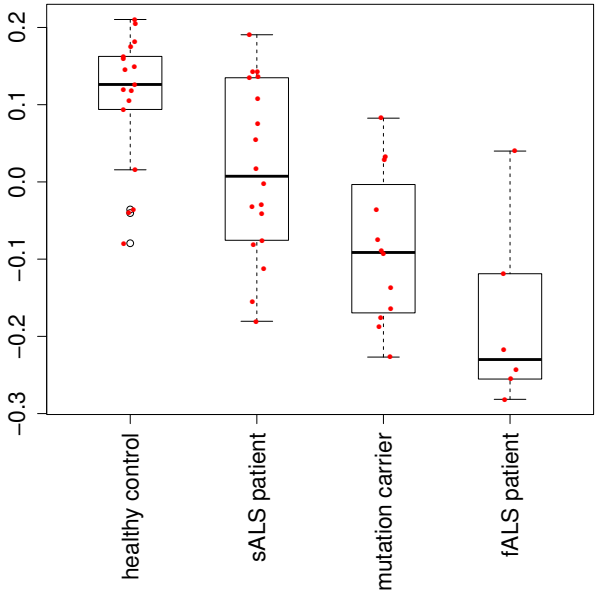

図 3　主成分分析を用いて計算した主成分負荷量のうち、カテゴリ回帰で計算した補正 P 値が、0.05 以下 ($P = 2.45 \times 10^{-5}$) であった第2 主成分負荷量の箱ひげ図。

Fig. 3　Boxplot using the second PC loading computed by PCA applied to miRNA profiles ($P = 2.45 \times 10^{-5}$).

4. 議論

本研究では３０００個以上の多数の miRNA の中から少

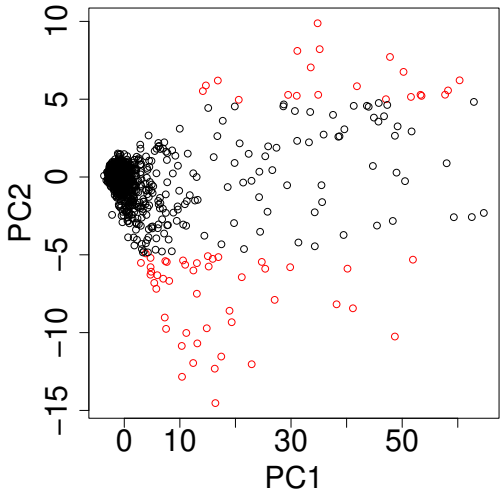

図 4　第一、第二主成分得点散布図。赤丸が選択された 67miRNA (表 1)。

Fig. 4　Scatter plot between the first and second PC scores. Red open circles correponds to selected 67 miRNAs (Table 1)

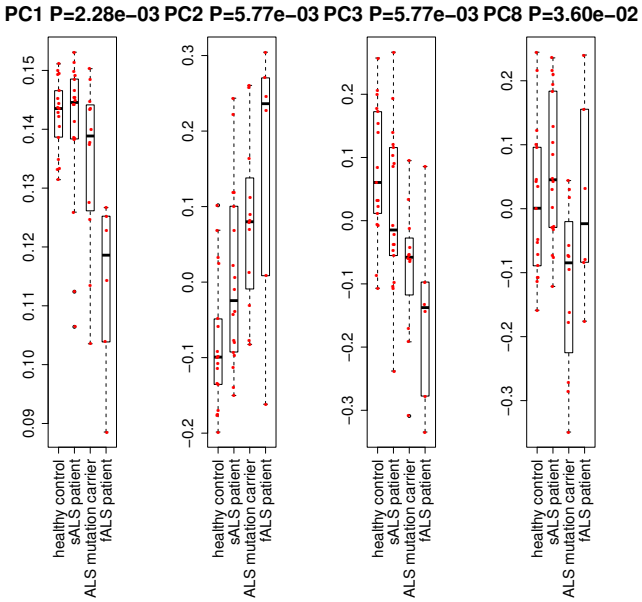

図 5　主成分分析で選ばれた 67miRNA(表 1) のみを用いて計算した結果で得られる主成分負荷量のうち、カテゴリ回帰で計算した補正 P 値が、0.05 以下であった第1，2，3，8 主成分負荷量の箱ひげ図。

Fig. 5　Boxplots using four PC loading computed by PCA applied to 67 miRNAs selected by PCA (Table 1).

数個（百個以下）の miRNA を選択し、選択した miRNA から主成分分析を用いて更に少数個の合成変数を生成・選択することで ALS のバイオマーカーを構成することを目指した。miRNA の選択にはカテゴリ回帰と主成分分析を用いた教師なし学習による変数選択を用いた。両法とも、補正 P 値が 0.01 以下という基準で、miRNA を選択した。カテゴリ回帰は 90 miRNA、主成分分析は 67 miRNA とほぼ同じくらいの数を選択し、また、かなりの数の miRNA が両法で共通に選択された（表 1, フィッシャーの正確確率検定で、$P = 1.5 \times 10^{-19}$、オッズ比で２３倍）。このことから、両法はある程度共通した基準で miRNA を選択していることが伺われる。大きな違いは、カテゴリ回帰では4群に対するどのような依存性であっても、miRNA の選択基準になるのに対して、主成分分析を用いた教師なし学習による変数選択の場合には図 3 に示されたような依存性以外は考慮されないということである。にも関わらず、これだけ統計的に有意度の高い、共通性のある選択がなされるということは、図 3 で示されるような四群への依存性が非常に支配的だということを示しているだろう。

また、選択された９０および６７ miRNA だけを用いて PCA した場合の主成分負荷量のうち、四群に有意に相関している負荷量同士の類似性も大きい（図 2, 図 5）。そもそも両法とも同数の4 つの主成分負荷量が、「カテゴリ回帰で計算した補正 P 値が 0.05」という同じ基準で選ばれており、選ばれた主成分も、カテゴリ回帰が第1，2，5，8 主成分、主成分分析を用いた教師なし学習による変数選択のほうが第1，2，3，8 主成分で4 主成分中、第1，2，8 主成分の3 つが共通して選ばれている。実際、単に第何主成分かということだけではなく、共通に選ばれている第1，2，8 成分については相関係数も大きい (表 4 の第1，2，4 番目の対角成分が該当する。P 値はそれぞれ、$P = 1.11 \times 10^{-7}, 6.14 \times 10^{-6}, 5.41 \times 10^{-5}$)。実際に共通に

表 3　図 5 に示された4 つの主成分負荷量を用いた線形判別の混同行列。fALS:家族性 ALS, sALS:孤発性 ALS

Table 3　Confusion matrix of liner regression analysis using four PC loading shown in Fig. 5.

	正解			
予測	変異保持者	fALS 患者	健常者	sALS 患者
変異保持者	8	1	0	2
fALS 患者	2	5	0	1
健常者	0	0	14	7
sALS 患者	2	0	3	8

表 4　図 2 と図 5 に示された4 つの主成分負荷量の間のピアソン相関係数の値。行：図 2、列：図 5。

Table 4　Person's correlation coefficients between PC loading show in Fig. 2 (rows) and Fig. 5 (columnes).

	PC1	PC2	PC3	PC8
PC1	0.65	-0.24	0.41	0.10
PC2	-0.50	0.58	-0.79	-0.08
PC5	0.32	-0.04	0.04	-0.39
PC8	0.26	-0.05	-0.06	0.53

選ばれている miRNA の数はカテゴリ回帰を用いた選択の場合は９０ miRNA 中２２ miRNA、主成分分析を用いた教師なし学習による変数選択の場合は、６７ miRNA 中２２ miRNA に過ぎない（ 表 1）ことを考えると、この一致は目覚ましい。変数選択を行ったあと、それを更にもう一回主成分分析にかけることで、よりロバスト性のある主成分負荷量が得られるのは非常に興味深い。

それでは一致していない図 2 の第三主成分得点と図 5 の第八主成分得点の差は何を反映しているだろうか。それは表 2 と表 3 の違い、すなわち、孤発性 ALS 患者の判別能に関係していると思われる。実際、健常者、家族性 ALS 患者、変異保持者の間の判別性能は、表 2 と表 3 とではほぼ同じだが、孤発性 ALS 患者の判別能だけは表 3 で大きく向上した。

このことは、変数選択に教師なし学習を用いることの重要性を表現していると思われる。教師あり学習であるカテゴリ回帰ではカテゴリ内の分散が少なく、カテゴリ間の分散が大きい変数を選択するという基準が適用される。このことは一見、いいように見えるが、問題はペナルティとされるカテゴリ内分散の扱いである。あるカテゴリ内の分散が大きい変数は一見、よくない（ カテゴリ間の判別には使えない）変数の様に思える。だが、一方で、人間には認識できていない複数のサブカテゴリからなっているために大きな分散をもっている可能性もある。かりに孤発性 ALS が実際はいくつかのサブカテゴリに別れているが人間には認識できないとしよう。そうなると、カテゴリ内分散が小さい変数はサブカテゴリが認識できない、より劣った変数ということになってしまい、サブカテゴリを正しく反映している変数は、生物学的には正しいことをしているにも関わらず、不適当な変数として排除されてしまうことになる。もし、サブカテゴリをグループとして反映できる変数がなければ、選ばれるのは「 ゆらぎ」でたまたまサブカテゴリ間の差が小さかった変数になってしまい、このような変数を用いた判別モデルは過学習を起こし、性能が低下するだろう。表 2 を見る限りでは、まさにこれが起きてしまっている様に見える。実際、図 3 を見る限りでは、孤発性 ALS のカテゴリ内分散が他のカテゴリに比べて明らかに大きい。これはひょっとすると人間には未知のサブカテゴリが孤発性 ALS に含まれていることを意味しているのかもしれない。

Liguori ら [5] は本研究の原著論文 [2] 出版「 後」に３８個の血中 miRNA が孤発性 ALS の患者で低下していることを報告した。うち１４個が表 1 にも含まれているが大部分（ １４個中１０個）は主成分分析を用いた教師なし学習でのみ選択されている。これではカテゴリ回帰を用いた変数選択に基づいて、孤発性 ALS の判別がうまく行くはずもなかっただろう。カテゴリ回帰の限界を越えるには、Liguori ら [5] がやったように、孤発性 ALS と健常者の間の二群間比較を行うしかない。しかし、多群（ 今回の場合は四群）がある場合の、ペアワイズな比較の繰り返しは一般に統計的な手法としては誤りとされており、今回の様に四群のデータから始めた場合、孤発性 ALS と健常者の二群間比較だけを特別に行うことを正当化するにはなんからの生物学的、あるいは、医学的な議論が必要だろう。それは決して、科学的な立場としては、悪いことではないが、主成分分析を用いた教師なし学習による変数選択ではその様な議論なしに、データ駆動的に、孤発性 ALS 特異的に変動している miRNA をきちっと選択できている。どちらを使うべきかは議論を待たないのではないだろうか。そして発生頻度から言えば家族性 ALS は ALS 全体の1 割程度を占めるにすぎず、大部分の ALS は孤発性 ALS に属するのだからなおさらである。

このような議論に対してはこのようなフィルター型の変数選択ではなく、ランダムフォレスト [6] や LASSO [7] の様なラッパー型の変数選択を行えばこのような誤った変数選択（ より良いパフォーマンスを達成できるのに見逃す）を行わないで済むのではないかという反論があるだろう。だが、その場合、全サンプルを学習セットとテストセットに分けて学習セットで変数選択を行い、性能はテストセットで確認するという方法を取らないと、フィルター型の変数選択以上に過学習をする可能性が上がってしまう。しかし、今回、サンプル数は4 群で５３サンプルと必ずしも多くはなく、一番数が少ない家族型 ALS 患者はわずか6 サンプルしかない。この状態で、全体を学習セットとテストセットに分けて変数選択を行うことは現実的であるか大いに疑問である。

5. 結論

機械学習的な手法の中では一般に教師あり学習の方が有効であるとされ、特に変数選択の場合に教師なし学習が使われることは稀である。しかし、この例で見るように、人間が勝手に決めた基準（ この場合は四群間の差異が大きく、群内の分散は小さくあるべきだ、という基準）を「 教師」とみなして、それに合うようにデータを加工することは危険を伴う。データ駆動科学の基本に戻ってより我々は謙虚であるべきだと思うのは私だけだろうか。

謝辞　本研究の原著論文は台湾科学技術省の助成番号 105-2118-M-009-001-MY2 のグラントの助成、及び科研費基盤研究 (C) 17K00417 の助成を受けて行われた。

参考文献

[1] Zarei, S., Carr, K., Reiley, L., Diaz, K., Guerra, O., Altamirano, P., Pagani, W., Lodin, D., Orozco, G. and Chinea, A.: A comprehensive review of amyotrophic lateral sclerosis, *Surgical Neurology International*, Vol. 6, No. 1, p. 171 (online), DOI: 10.4103/2152-7806.169561 (2015).

[2] Taguchi, Y.-H. and Wang, H.: Exploring microRNA Biomarker for Amyotrophic Lateral Sclerosis, *International Journal of Molecular Sciences*, Vol. 19, No. 5, p. 1318 (online), DOI: 10.3390/ijms19051318 (2018).

[3] Freischmidt, A., Mller, K., Zondler, L., Weydt, P., Volk, A. E., Božič, A. L., Walter, M., Bonin, M., Mayer, B., von Arnim, C. A. F., Otto, M., Dieterich, C., Holzmann, K., Andersen, P. M., Ludolph, A. C., Danzer, K. M. and Weishaupt, J. H.: Serum microRNAs in patients with genetic amyotrophic lateral sclerosis and pre-manifest mutation carriers, *Brain*, Vol. 137, No. 11, pp. 2938–2950 (online), DOI: 10.1093/brain/awu249 (2014).

[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)*, Vol. 57, No. 1, pp. 289–300 (online), available from ⟨http://www.jstor.org/stable/2346101⟩ (1995).

[5] Liguori, M., Nuzziello, N., Introna, A., Consiglio, A., Licciulli, F., EustachioD' Errico, Scarafino, A., Distaso, E., Simone, I. L.: Dysregulation of MicroRNAs and Target Genes Networks in Peripheral Blood of Patients With Sporadic Amyotrophic Lateral Sclerosis, *Frontiers in Molecular Neuroscience*, Vol. 11, p. 288 (online), DOI: 10.3389/fnmol.2018.00288 (2018).

[6] Breiman, L.: Random Forests, *Machine Learning*, Vol. 45, No. 1, pp. 5–32 (online), DOI: 10.1023/A:1010933404324 (2001).

[7] Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 73, No. 3, pp. 273–282 (online), DOI: 10.1111/j.1467-9868.2011.00771.x (2011).