Vol.2017-BIO-51 No.1
2017/9/26

テンソル分解を用いた教師なし学習による心的外傷後ストレス障害由来の心臓病原因遺伝子の同定

田口　善弘[1,a)]

概要：心的外傷ストレス障害はテロや戦争等生命の危機に瀕するような強いストレスに晒された場合、無事に危機を切り抜けた後も精神的な不調に悩まされる疾患である。近年、このような疾患が肉体的な不調も誘発する場合があることがわかってきた。例えば、心的外傷ストレス障害をもつ患者は心臓疾患を発症する割合が有意に高い。しかし、心的外傷ストレス障害が生じているのはあくまで脳であり、空間的に遠く隔たった心臓に心的外傷ストレス障害がどのようにして疾患を引き起こすのかよくわかっていない。本研究では、心臓と脳で同じような遺伝子の発現異常を共有することが心的外傷後ストレス障害由来の心臓病を発生させるのではないかという仮説に基づいて、心的外傷後ストレス障害由来の心臓病のマウスモデルに於ける多種臓器の遺伝子発現プロファイルの解析をテンソル分解を用いた教師無し学習による変数選択を用いて解析し、同時以上発言している遺伝子を特定したのでこれについて報告する。

Y-H. Taguchi[1,a)]

1. はじめに

心的外傷ストレス障害の患者は心臓疾患を引き起こす可能性が有意に高いことが疫学的に知られている [1]。しかし、それはあくまで疫学的な観点からの研究であり、その背後にある分子生物学的な機構は解明されていない。双子においては心的外傷ストレス障害と心臓疾患の関係が有意に高いことから [2]、心的外傷ストレス障害と心臓疾患の関係にはなんらかの遺伝子学的な機構が関係していることが期待される。そこで我々は、最近刊行された心的外傷ストレス障害のマウスモデルにおける、脳と心臓を含む多臓器の遺伝子発現プロファイル (GEO [3] ID GSE68077) を、近年提案された主成分分析を用いた教師なし学習による変数選択 [4–24] をテンソル分解に拡張した、テンソル分解を用いた教師無し学習による変数選択を用いて解析し、心的外傷後ストレス障害由来の心臓病原因遺伝子の同定を試みた。本レポートはその結果の報告である。

[1] 中央大学理工学部物理学科
東京都, 112-8551, 日本
[a)] tag@granular.com
本研究内容は既に原著論文として InCob2017 に受理済みであり BMC Medical Genomics の Supplement として刊行予定の内容である [25]。

2. 手法

2.1 mRNA プロファイル

GSE68077 は心的外傷ストレス障害のマウスモデルの多臓器遺伝子発現プロファイルである。具体的には、扁桃体、海馬、内側前頭前皮質、中隔核、線条体、腹側線条体、血液、心臓、半脳、脾臓の１０部位について、１０日間と５日間の2 種類のストレス期間、１．５週、２４時間、６週の3 種類の休息期間（ストレスを加え終えてからの経過時間）のそれぞれについて参照群と操作群の2 種を行った大規模な実験である（各実験条件について、３～4 個の生物学的な反復がなされている。一部のストレス期間と休暇期間の組み合わせは実行されていない）。これらは遺伝子数 × 10 × 2 × 3 × 2 の5 階のテンソルとして表現できる（存在しないストレス期間と休息期間の組み合わせについては０を導入する）。

2.2 人工データ

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性能試験のための模擬遺伝子発現プロファイルである。３０、０００×１０×１０の3 階のテンソルであり、それぞれ、遺伝子、臓器、実験条件を表現しているとする。各実験条件では異った１００遺伝子ずつが発現するとする。した

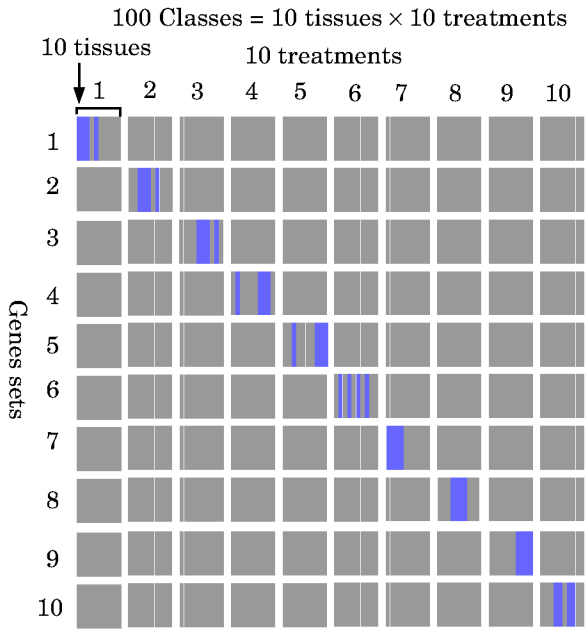

図 1　人工データの発現条件。各実験条件で異った１００遺伝子が、１０実験条件に対して、異った4 臓器(青線) で発現する。

Fig. 1　Expression conditions of synthetic data. Distinct 100 genes are expressive in each of 10 treatments for distinct sets of four tissues (blue lines).

がって、なんらかの実験条件で発現するのは合計で１０００遺伝子だけであり、残りの２９、０００遺伝子は発現しない。また、各実験条件で発現するのは１０臓器中4 臓器だけであり、どの実験条件でどの4 臓器が発現するかという条件は１０実験条件で異なっているものとする (図 **1**)。テンソルの値は分散1 のガウス分布から生成される。このうち、発現している場合には平均値として4 を付加する。

2.3　テンソル分解を用いた教師なし学習による変数選択法

任意の $m+1$ 階のテンソルは

$$x_{i,j_1,\ldots,j_m} = \sum_{\ell_1,\ldots,\ell_{m+1}} G(\ell_1,\ldots,\ell_{m+1}) \cdot x_{\ell_{m+1},i} \prod_{k=1}^{m} x_{\ell_k,j_k}$$

の様に展開することができる [26](HOSVD)。ここで G はコアテンソル、$x_{\ell_{m+1},i}$ と x_{ℓ_k,j_k} は特異値ベクトルである。ここで絶対値が大きな G をもつ $\ell_1,\ldots,\ell_{m+1}$ の組み合わせが、元のテンソル $x_{i,j_1,\ldots,j_m}$ に対して大きな寄与を持つので重要である。この展開を用いた変数選択は次のようにして行う。ここで選択したい変数は i(今回の例では遺伝子) であるとしよう。$j_1,\ldots,j_m$ は実験条件や臓器を表現するとする。まず、x_{ℓ_k,j_k} を観察し、目的とする現象に関係する特異値ベクトルを選択する(例えば、患者と健常者で差がある、など)。次にこれらの特異値ベクトルと絶対値の大きなコアテンソルを共有する ℓ_{m+1} のセット Ω を選択する。最後に $x_{\ell_{m+1},i}$ に対して多重ガウス分布を仮定し、χ 二乗分布を使い i に P 値 P_i を付与する。

$$P_i = P_{\chi^2}\left[> \sum_{\ell_{m+1}\in\Omega} \left(\frac{x_{\ell_{m+1},i}}{\sigma_{\ell_{m+1}}} \right)^2 \right]$$

ここで $P_{\chi^2}[>x]$ はカイ二乗分布の因数が x より大きい場合の累積積分確率、$\sigma_{\ell_{m+1}}$ は $x_{\ell_{m+1},i}$ の標準偏差である。得られた P_i は Benjamini-Hochberg(BH) 基準と R の fdrtool 関数とで多重比較補正され、この補正 P 値が十分小さい(具体的には人工データの場合は 0.1,0.05,0.01 の3 通り、mRNA プロファイルの場合には 0.01) i を選択する。また、$j_1,\ldots,j_m$ が多カテゴリにより構成されている(例えば、多臓器や多実験条件) 場合には同じように P_{j_k} を計算することで臓器や実験条件の選択を行うことも可能である。

3.　結果

3.1　人工データの解析結果

目的は2 つあり

(1) ３０，０００個中、なんらかの条件で発現している１、０００個(全体の 3.3%) の遺伝子を判別できるか。

(2) １、０００個の遺伝子が１００個ずつの１０クラスターに分かれていることが認識できるか。

である。テンソル分解

$$x_{i_1,i_2,i_3} = \sum_{\ell_1,\ell_2,\ell_3} G(\ell_1,\ell_2,\ell_3) x_{\ell_1,i_1} x_{\ell_2,i_2} x_{\ell_3,i_3}$$

を適用する。i_1 が遺伝子、i_2 が臓器、i_3 が実験条件に対応する。図 **2** は $x_{\ell_1,i_1}, 2 \le \ell_1 \le 5, 1 \le i \le 1000$ を4 次元ベクトルとみなして、表記した結果である。きれいに１０個のクラスターに分かれていることが判る。次に、１，０００個の遺伝子を選択できるかに挑戦した。x_{ℓ_1,i_1} を用いて前節、「方法」に書かれた方法で、しきい値に設定する補正された P 値を変えながら、第何番目までの特異値ベクトルを用いるか(つまり $1 \le \ell_1 \le k$ の k をいくつにとるか) で真陰性率、真陽性率、AUC がどう変わるかを図 **3** に表記した。まず、しきい値の P 値に関わらず、誤認識はほぼ無いことが判る。１、０００個の遺伝子のうち、どれくらいを認識できるかはしきい値の P 値で変化しているが、最大値で２０%から６０%の間である。但し、AUC は最低でも０．７，最大で０．９に達しており、正例が３．３%しか無いという厳しい条件を考えると十分な性能が出ていると言える。更に AUC と真陽性率は $1 \le \ell_1 \le 10$ の時に最大値になっている。一般に K 個のクラスターをユークリッド空間で表現するには $K-1$ 次元が必要であるが、今回のデータでは発現している１０遺伝子グループとそれ以外であわせて１１遺伝子グループになっているのでこの結果は妥当だと言えるだろう。目的(１) は達成されているといってよい。

更に実際に１０個の遺伝子グループの分離を定量的に再現できるかを確認するために、選択された遺伝子にユーク

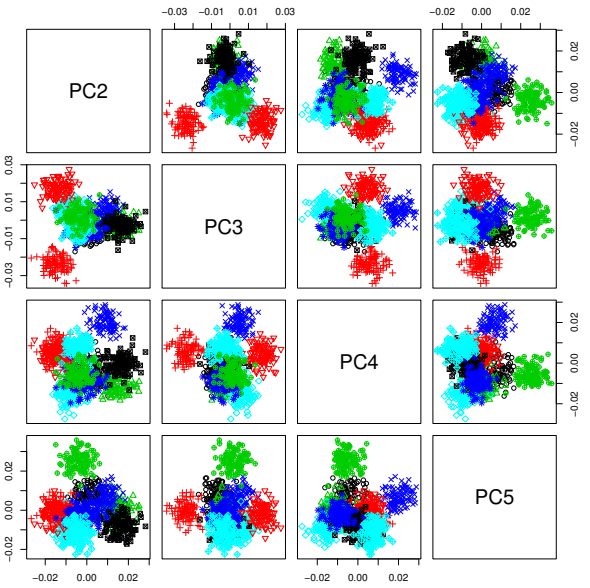

図 2　元の発現する１０遺伝子グループに属する１，０００遺伝子の $x_{\ell_1,i_1}, 2 \leq \ell_1 \leq 5$, of 1,000 genes $(1 \leq i_1 \leq 1,000)$ の散布図。色と記号の組み合わせを変えて表記してある。見やすさのため、残る２９、０００遺伝子のベクトルの表記は省略している

Fig. 2　Scatter plots involving the second gene's through fifth gene's singular value vectors. $x_{\ell_1,i_1}, 2 \leq \ell_1 \leq 5$, of 1,000 genes $(1 \leq i_1 \leq 1,000)$ that belong to one of the 10 gene sets. These 10 gene sets are represented by distinct combinations of colors and symbols. The 29,000 genes not included in any of the 10 gene sets are omitted for clarity.

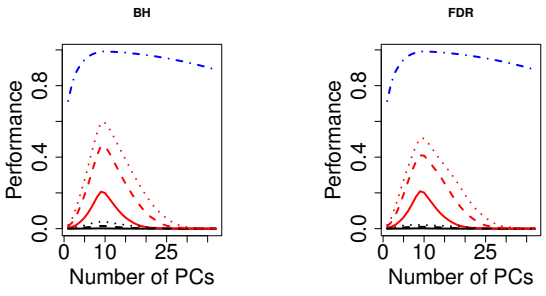

図 3　１００個の独立な人工データの平均で求めた判別性能。１，０００個の発現遺伝子を正確に判別できたかを表記している。赤が真陽性率、黒が真陰性率、実線が $P = 0.01$、破線が $P = 0.05$、点線が $P = 0.01$ の場合である。青の一点鎖線は AUC である。BH は Benjamini-Hochberg 法で FDR は false discovery rate で補正した場合である。

Fig. 3　Performance of synthetic data (averaged over 100 trials). BH: Benjamini-Hochberg, FDR: false discovery rate. Red curves: true positive rates, black curves: false positive rates, solid curves: $P = 0.01$, dashed curves: $P = 0.05$, dotted curves: $P = 0.1$, blue dash-and-dot curves: area under the curve (AUC).

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（補正された P 値が 0.1 以下）が元の遺伝子グループ（行）とクラスター結果（列）の組み合わせでどのように分布しているかを見た。２９，０００個の発現していない遺伝子は遺伝子グループ１１であり、クラスター１１に多く分類されている。

Table 1　Clustering of genes identified by TD-based unsupervised FE for synthetic data (adjusted P-values less than 0.1). Rows: gene sets (the first to the tenth are the gene sets to which the first 1,000 genes are likely to belong, and the 11th is the gene set to which the remaining 29,000 genes are likely to belong), columns: clustering

Mclust

	1	2	3	4	5	6	7	8	9	10	11
1	80	0	0	0	0	0	0	0	0	0	5
2	0	71	0	0	0	0	0	0	0	0	5
3	0	0	90	0	0	0	0	0	0	0	1
4	0	0	0	65	0	0	0	0	0	0	1
5	0	0	0	0	69	0	0	0	0	0	2
6	0	0	0	0	0	00	0	0	0	0	16
7	0	0	0	0	0	66	0	0	0	0	2
8	0	0	0	0	0	0	66	0	0	0	5
9	0	0	0	0	0	0	0	77	0	0	4
10	0	0	0	0	0	0	0	0	74	0	3
11	0	0	0	0	0	0	0	0	0	81	11

Ward

	1	2	3	4	5	6	7	8	9	10	11
1	81	0	0	0	0	0	0	0	0	0	5
2	0	71	0	0	0	0	0	0	0	0	4
3	0	0	90	0	0	0	0	0	0	0	3
4	0	0	0	65	0	0	0	0	0	0	1
5	0	0	0	0	69	0	0	0	0	0	8
6	0	0	0	0	0	66	0	0	0	0	4
7	0	0	0	0	0	0	66	0	0	0	8
8	0	0	0	0	0	0	0	77	0	0	5
9	0	0	0	0	0	0	0	0	64	0	4
10	0	0	0	0	0	0	0	0	0	35	3
11	0	0	0	0	0	0	0	0	0	46	10

リッド距離と ward 法を用いた階層的クラスタリングを行い全体を１１個のクラスターに分けた場合と、混合ガウス分布で１１個のクラスターに分けた場合の、元の１０遺伝子グループとの関係を表 **1** に示す。得られた１１個のクラスターのうちの１０個に元の１０遺伝子グループがほぼ綺麗に分かれていることが判る。したがって、目的（２）も達成されていると言える。

テンソル分解を用いた教師なし学習による変数選択が上記目的（１）と（２）を達成できることは確認できたが、他の方法はどうであろうか。こんなややこしいことをしなくても既存手法で可能ならこんなことをする必要はない。ここで既存手法を試す前にまず指摘したいことは、群間の差異、例えば、平均値の差異などを遺伝子の選択基準とす

表 2　カテゴリカル回帰での変数選択の混同行列。行が正解で列が予測。太字が真陽性例の数である。

Table 2　Confusion matrix of synthetic data using categorical regression. Bold numbers are true positives.

１０実験条件×１０サンプルとみなした場合, AUC=0.94

adjusted P-values	$P < 0.01$	$P > 0.01$	$P < 0.05$	$P > 0.05$	$P < 0.1$	$P > 0.1$
$i \leq 1000$	**2**	998	**16**	984	**77**	923
$i > 1000$	0	29000	0	29000	7	28993

１０臓器×１０サンプルとみなした場合 , AUC=0.58

adjusted P-values	$P < 0.01$	$P > 0.01$	$P < 0.05$	$P > 0.05$	$P < 0.1$	$P > 0.1$
$i \leq 1000$	**0**	1000	**0**	1000	**0**	1000
$i > 1000$	0	29000	0	29000	0	29000

るような方法は一切使うことができないということである。なぜならば、今回のデータでは１０臓器×１０実験条件で１００サンプルしかなく、個々の組み合わせに対して一個のサンプルしか無いのでそもそも群間比較は原理的に不可能である。したがって、厳密なことを言うのであれば、テンソル分解を用いた教師なし学習による変数選択でしかこの人工データは解析できず、性能比較以前の問題であると言うことができる。しかし、一方、一実験条件に一サンプルというのはあまり現実的な条件とは言えず、これでテンソル分解を用いた教師なし学習による変数選択が他手法より優れているというのはやや問題があるので、仮想的に実験条件、または、臓器の区別を無視して、１０サンプルずつが計測された１０実験条件、または、１０サンプルずつが計測された１０臓器の１００サンプルデータとして解析し、１，０００遺伝子が選択できるかを確かめてみた。最初に試みたのはカテゴリカル回帰での変数選択（ 別名 ANOVA）である（ 表 **2**)。まず気づくことは真陽性率の低さである。テンソル分解を用いた教師なし学習による変数選択（ 図 3）の場合は $P = 0.1$ の場合には６０％近い真陽性率になっているがカテゴリカル回帰での変数選択では１０％にも達していない。それでも AUC 自体は１０実験条件とみなした場合には高くなってはいる。ただ、１０臓器とみなした場合は AUC が０．５しかなく事実上判別はできていない。これは図 1 に見るように、実験条件では比較的発現している遺伝子がまとまっているが臓器単位で考えると発現しているサンプルの数も大きくばらついており判別が困難だったのが伺われる（ P 値は BH 基準で補正した)。そもそもこの問題は正例が３．３％しかない難しい問題なのでこのような結果になっても仕方はない。

次に、SAM [27] を試した（ 表 **3**)。SAM は遺伝子発現プロファイルの解析で分類群間に差がある遺伝子の選択で広く用いられる手法である。結果はカテゴリカル回帰の場合（ 表 2）と大差ない。１０実験条件とみなした場合、AUC こそ０．９４になっているが、FDR が０．１９で１９６遺伝子，０．０７で３４遺伝子しか選択できておらず、FDR を０．１にした場合には総数で１，０００個ある正例の１０％ていどしか検出できていないし、１０臓器と見た場合にはそもそも、AUC が０．５８とほぼ判別ができていない状態である。SAM を使ってもテンソル分解を用いた教師なし学習による変数選択に比べると極めて劣った結果しか出せていない。最後に limma [28] を試した（ 表 **4**, 補正された P 値としては limma が提供する q 値を採用した)。limma は最新のベイズ統計を考慮したパッケージであり、マイクロアレイの解析で発現差のある遺伝子を選択する場合にはデファクトスタンダードになっているような方法である。本来は対数比に対して用いるべきであるが、人工データは負値をたくさん含んでいるのでデータ自体が対数比をとったあとの値であるとみなして limma を用いた。さすがに AUC がどちらでも０．９９になっており SAM やカテゴリカル回帰による変数選択よりはこの点において改善している。ただ、補正 P 値として妥当な 0.1 程度をしきい値にした場合には真陽性率が１０％以下なのは変わっていない。

以上より、この人工データは１０臓器、または、１０実験条件とみなす、という妥協をした上でも、３０，０００個の遺伝子の中から１，０００個の遺伝子を選ぶという問題に関する限り、カテゴリカル回帰による変数選択、SAM、limma のいずれも、テンソル分解を用いた教師なし学習による変数選択に大きく劣っているということがわかった。次に現実のデータに本手法を適用した場合、どのような結果が得られるかを見てみよう。

3.2　mRNA プロファイルの解析結果

今回解析するデータは表 **5** のとおりであり、これを遺伝子が i、臓器が j_2、ストレス期間が j_3、ストレス後休憩期間が j_4、そして参照群か操作群かを j_1 とする５階のテンソル x_{i,j_1,j_2,j_3,j_4} として扱う。このテンソルに HOSVD を用いた結果、図 **4** にあるように $x_{\ell_1=2,j_1}$ と $x_{\ell_2=4,j_2}$ がそれぞれ参照群と操作群が逆符号で、扁桃体、海馬、心臓が同符号で大きな寄与をもつことがわかった。したがって、これらと対になって絶対値の大きな G をもつ遺伝子、ストレス時間、休憩時間の特異値ベクトルをみつければ、どの

表 3　SAM の結果。p0 は帰無仮説割合、FDR は補正 P 値に相当。Called は帰無仮設に該当しない遺伝子の数。真陽性例の数の期待値は False × FDR × p0。

Table 3　Results by SAM. p0 is the ratio of the null hypothesis, and FDR corresponds to the adjusted P-values. Called is the number of genes that break the null hypothesis. Expected number of false positives is False × FDR × p0.

１０実験条件×１０サンプルとみなした場合 , AUC=0.94

	Delta	p0	False	Called	FDR
1	0.1	0.974	365.47	799	0.44560
2	0.2	0.974	38.59	196	0.19180
3	0.3	0.974	2.59	34	0.07421
4	0.4	0.974	0.02	3	0.00649
5	0.5	0.974	0.02	3	0.00649
6	0.6	0.974	0	2	0
7	0.7	0.974	0	0	0
8	0.8	0.974	0	0	0
9	0.9	0.974	0	0	0
10	1.0	0.974	0	0	0

１０臓器×１０サンプルとみなした場合、 AUC=0.58

	Delta	p0	False	Called	FDR
1	0.1	1	0	0	0
2	0.1	1	0	0	0
3	0.1	1	0	0	0
4	0.1	1	0	0	0
5	0.1	1	0	0	0
6	0.1	1	0	0	0
7	0.1	1	0	0	0
8	0.1	1	0	0	0
9	0.1	1	0	0	0
10	0.1	1	0	0	0

表 4　limma を使った場合の混同行列。太字が真陽性例の数

Table 4　Confusion matrix of synthetic data using limma. Bold numbers are true positives.

１０実験条件×１０サンプル, AUC=0.99

adjusted P-values	$P<0.01$	$P>0.01$	$P<0.05$	$P>0.05$	$P<0.1$	$P>0.1$
$i \leq 1000$	**2**	998	**10**	984	**82**	923
$i > 1000$	0	29000	0	29000	0	29000

１０臓器×１０サンプルとみなした場合, AUC=0.99

adjusted P-values	$P<0.01$	$P>0.01$	$P<0.05$	$P>0.05$	$P<0.1$	$P>0.1$
$i \leq 1000$	**0**	1000	**0**	1000	**0**	1000
$i > 1000$	0	29000	0	29000	0	29000

表 6　絶対値の大きい上位の $G(\ell_1 = 2, \ell_2 = 4, \ell_3, \ell_4, \ell_5)$ 。1

Table 6　Top-ranked $G(\ell_1 = 2, \ell_2 = 4, \ell_3, \ell_4, \ell_5)$ with greater absolute values.

ℓ_3	ℓ_4	ℓ_5	$G(2,4,\ell_3,\ell_4,\ell_5)$
1	1	11	-35.0
1	1	1	-30.8
2	2	1	-30.3
2	3	4	-30.0
2	3	1	28.7
2	2	4	28.5

実験条件でどの遺伝子が扁桃体、海馬、心臓で同時に異常発現しているかを知ることができる。 表 **6** にあるように、ℓ_3(ストレス時間) と ℓ_4(休憩時間) については、全部の特異値ベクトルが上位に現れて居て、特に選択の必要はないが、ℓ_5(遺伝子) については $\ell_5 = 1, 4, 11$ の3 つが特にたくさん出現していることが判る。そこでこれらを用いて P 値を計算し、BH 基準で多重比較補正して 0.01 以下のものを選んだところ、全部で８０１ 個のプローブが選択された。そこでこれら８０１ プローブが実際に参照群と操作群において有意に発現差があるかをt 検定したところ、表 **7** にあるように４０ 種類の臓器と実験条件の組み合わせのうち、１３ の組み合わせで有意差があるということがわかった。このことから、テンソル分解を用いた教師なし学習に

表 7　８０１ プローブの発現が参照群と操作群で有意に差がある１３ 種類の臓器と 実験条件の組わせ。

Table 7　Thirteen combinations of tissues and experimental conditions where the selected 801 probes are differentially expressed between stress-exposed and control samples.

ストレス期間	10 日		5 日	
休憩期間	24 時間	6 週	２４時間	１．５週
扁桃体		○		○
海馬		○	○	○
内側前頭前皮質		○		
心臓	○			○
半脳			○	○
脾臓		○	○	○

よる変数選択はこの様な複雑な実験条件と臓器の組み合わせの中から参照群と 操作群で差がある遺伝子のセットを抽出するという 操作をほぼ自動的に実行できる能力があることがわかった。

最後にカテゴリカル回帰による変数選択、SAM、limma を用いて mRNA プロファイルで遺伝子選択を行った結果を示す(表 **8**、表 **9**、表 **10**)。細かい説明は省略するが、全般的に非常に多数個の遺伝子が検出されてしまっている。この理由は単純で、全体で１０臓器×４実験条件×２(参

表 5　本研究で使われたサンプル一覧。カンマの前後の数字は参照群と匝瑳郡のサンプル数。h は時間、w は週。

Table 5　Samples used in this study. Numbers before/after comma are control/treated samples. h: hours, w: weeks.

ストレス期間（単位：日）	5		10			5		10	
休憩期間	24h	1.5 w	24h	6w		24h	1.5 w	24h	6w
扁桃体	3,2	5,4	3,4	3,4	海馬	3,5	4,5	5,4	4,5
内側前頭前皮質	4,5	5,5	3,4	4,4	中隔核	3,2	2,3	3,3	3,3
線条体	5,5	5,5	5,4	4,4	腹側線条体	5,5	5,5	3,4	5,4
血液	5,5	5,5	4,5	4,5	心臓	5,5	4,5	5,5	5,5
半脳	5,5	4,5	5,5	5,5	脾臓	5,5	5,5	5,4	5,5

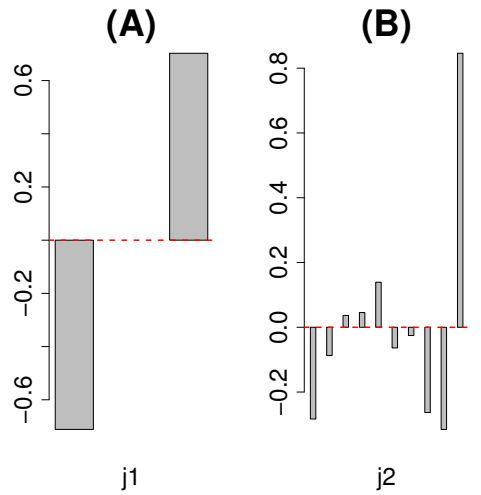

図 4　変数選択に使用された特異値ベクトル。（A）参照群／操作群の第2 特異値ベクトル、$x_{\ell_1=2,j_1}$。j_1 は1 と2 が各々、参照群と操作群に対応。（B）第4 臓器特異値ベクトル、$x_{\ell_2=4,j_2}$。j_2 が1，2，8，9，10 が各々、扁桃体、海馬、心臓、半脳、そして脾臓に対応。

Fig. 4　Singular value vectors employed. (A) The second control-related or treatment-related singular value vector, $x_{\ell_1=2,j_1}$. Control: $j_1 = 1$, and treatment (stress): $j_1 = 2$. (B) The fourth tissue singular value vector, $x_{\ell_2=4,j_2}$, AY: $j_2 = 1$, HC: $j_2 = 2$, heart: $j_2 = 8$, hemibrain: $j_2 = 9$, and spleen: $j_2 = 10$.

照群対操作群）＝８０カテゴリーになってしまい、SAM やカテゴリカル回帰でそれらのうちのどれかが一個でも他と差があれば有意であるとみなされてしまうからである。参照群と操作群の対関係を操作に盛り込めない。limma の場合はペアワイズに差をとることで４０カテゴリに減らすことが可能だが、それでもこの４０カテゴリーのどれか一個でも有意に値を持ってしまうと検出されてしまうために数は減るものの多数個が検出される問題は解消できない。それはやり方が悪いだけで４０対を別々に扱って発現差のある遺伝子を選択すればいいだろう、と思うかもしれないが、このやり方には問題が多い。まず、表 5 にあるように４０個の臓器と実験条件の組み合わせの中で、参照群と操作軍はそれぞれ数個ずつしかない。このような少数個で統計検定して P 値を求めても、多重比較の補正数がプローブ数×４０という膨大な数になってしまう。プローブ数は4 万個以上あるので全部で１６０万個に対して多重比較補正しなくてはならず、例えば、しきい値の補正 P 値を０．０１とした場合、これは 10^{-9} とかいうとんでもなく小さな値の生の P 値がなくてはならないことになる。参照群5 個、操作群5 個の少数サンプルでこれだけの P 値を得るのはほぼ絶望的だろう。さらに別の問題として、仮に遺伝子が４０個の条件で検出できたとしても、その遺伝子は海馬、扁桃体、心臓では選択されるが、他では選択されない、という都合のよい条件を満たさなくては心的外傷ストレス障害と心臓疾患を結びつける要因にはならないだろう。テンソル分解を用いた教師なし学習による変数選択の場合は図 4 で臓器特異値ベクトルを選んだことで（完全ではないが）ある程度は満たされていて、それが表 7 の様な結果につながっている。

この様な議論から現実のデータの場合にもテンソル分解を用いた教師なし学習による変数選択は非常に有効な手段であると結論付けられるだろう。遺伝子について生物学的な議論も行ったがそれはいずれ刊行予定の原著論文 [25] を参照されたい。

4.　終わりに

本稿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を提案し、それを人工データと mRNA プロファイルに適応し有効な結果を得た。本稿で見たように同手法は非常に多数の実験条件がある場合、それを統合的に解析する能力に優れている。この方法が今後広く用いられることを念じて本稿の結びに換える。

参考文献

[1]　Deng, L. X., Khan, A. M., Drajpuch, D., Fuller, S., Ludmir, J., Mascio, C. E., Partington, S. L., Qadeer, A., Tobin, L., Kovacs, A. H. and Kim, Y. Y.: Prevalence and Correlates of Post-traumatic Stress Disorder in Adults With Congenital Heart Disease, *Am. J. Cardiol.*, Vol. 117, No. 5, pp. 853–857 (2016).
[2]　Vaccarino, V., Goldberg, J., Rooks, C., Shah, A. J.,

表 8　カテゴリ回帰による変数選択（ mRNA プロファイル）
Table 8　Results of gene selection based on CR.

adjusted P-values	$P > 0.01$	$P < 0.01$	$P > 0.05$	$P < 0.05$	$P > 0.1$	$P < 0.1$
	2222	41157	1986	41713	1839	41860

表 10　limma の結果（ mRNA プロファイル）
Table 10　Results of gene selection based on limma.

adjusted	ケース A: 参照群と操作群の差を考慮しない					
P-values	$P > 0.01$	$P < 0.01$	$P > 0.05$	$P < 0.05$	$P > 0.1$	$P < 0.1$
	0	43379	0	43379	0	43379
adjusted	ケース B: 参照群と操作群の差を考慮					
P-values	$P > 0.01$	$P < 0.01$	$P > 0.05$	$P < 0.05$	$P > 0.1$	$P < 0.1$
	25992	17387	17745	25634	13542	29837

表 9　SAM の結果（ mRNA プロファイル）。p0 は帰無仮説割合、FDR は補正 P 値に相当。Called は帰無仮設に該当しない遺伝子の数。真陽性例の数の期待値は False × FDR × p0。
Table 9　Results by SAM. p0 is the ratio of the null hypothesis, FDR corresponds to the adjusted P-values. Called is the number of genes that break the null hypothesis. Expected number of false positives is False × FDR × p0.

	Delta	p0	False	Called	FDR
1	0.1	0.011	38538.08	43379	0.0094
2	11.4	0.011	0.02	5424	3.9e-08
3	22.7	0.011	0	323	0
4	34.0	0.011	0	40	0
5	45.2	0.011	0	7	0
6	56.5	0.011	0	4	0
7	67.8	0.011	0	2	0
8	79.1	0.011	0	1	0
9	90.3	0.011	0	1	0
10	101.6	0.011	0	1	0

Veledar, E., Faber, T. L., Votaw, J. R., Forsberg, C. W. and Bremner, J. D.: Post-traumatic stress disorder and incidence of coronary heart disease: a twin study, *J. Am. Coll. Cardiol.*, Vol. 62, No. 11, pp. 970–978 (2013).

[3] Barrett, T. and Edgar, R.: [19] Gene Expression Omnibus: Microarray Data Storage, Submission, Retrieval, and Analysis, *Methods in Enzymology*, Elsevier, pp. 352–369 (online), DOI: 10.1016/s0076-6879(06)11019-8 (2006).

[4] Taguchi, Y.-H., Iwadate, M., Umeyama, H. and Murakami, Y.: Principal component analysis based unsupervised feature extraction applied to bioinformatics analysis, *Computational Methods with Applications in Bioinformatics Analysis* (Tsai, J. J. P. and Ng, K.-L., eds.), World Scientific, Singapore, chapter 8, pp. 153–182 (online), DOI: 10.1142/9789813207981_0008 (2017).

[5] Taguchi, Y. H.: microRNA-mRNA Interaction Identification in Wilms Tumor Using Principal Component Analysis Based Unsupervised Feature Extraction, *2016 IEEE 16th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 71–78 (online), DOI: 10.1109/BIBE.2016.14 (2016).

[6] Taguchi, Y. H.: Principal Components Analysis Based Unsupervised Feature Extraction Applied to Gene Expression Analysis of Blood from Dengue Haemorrhagic Fever Patients, *Sci Rep*, Vol. 7, p. 44016 (2017).

[7] Taguchi, Y.-H.: Principal component analysis based unsupervised feature extraction applied to publicly available gene expression profiles provides new insights into the mechanisms of action of histone deacetylase inhibitors, *Neuroepigenetics*, Vol. 8, pp. 1 – 18 (online), DOI: http://dx.doi.org/10.1016/j.nepig.2016.10.001 (2016).

[8] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for post-traumatic stress disorder-mediated heart disease, *BMC Bioinformatics*, Vol. 16, p. 139 (2015).

[9] Taguchi, Y. H. and Okamoto, A.: Principal Component Analysis for Bacterial Proteomic Analysis, *Pattern Recognition in Bioinformatics* (Shibuya, T., Kashima, H., Sese, J. and Ahmad, S., eds.), LNCS, Vol. 7632, Springer International Publishing, Heidelberg, pp. 141–152 (2012).

[10] Ishida, S., Umeyama, H., Iwadate, M. and Taguchi, Y. H.: Bioinformatic Screening of Autoimmune Disease Genes and Protein Structure Prediction with FAMS for Drug Discovery, *Protein Pept. Lett.*, Vol. 21, No. 8, pp. 828–39 (2014).

[11] Kinoshita, R., Iwadate, M., Umeyama, H. and Taguchi, Y. H.: Genes associated with genotype-specific DNA methylation in squamous cell carcinoma as candidate drug targets, *BMC Syst Biol*, Vol. 8 Suppl 1, p. S4 (2014).

[12] Taguchi, Y. H. and Murakami, Y.: Principal component analysis based feature extraction approach to identify circulating microRNA biomarkers, *PLoS ONE*, Vol. 8, No. 6, p. e66714 (2013).

[13] Taguchi, Y. H. and Murakami, Y.: Universal disease biomarker: can a fixed set of blood microRNAs diagnose multiple diseases?, *BMC Res Notes*, Vol. 7, p. 581 (2014).

[14] Murakami, Y., Toyoda, H., Tanahashi, T., Tanaka, J., Kumada, T., Yoshioka, Y., Kosaka, N., Ochiya, T. and Taguchi, Y. H.: Comprehensive miRNA expression analysis in peripheral blood can diagnose liver disease, *PLoS ONE*, Vol. 7, No. 10, p. e48366 (2012).

[15] Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., Tamori, A., Kawada, N., Taguchi, Y. H. and Azuma, T.: Comparison of Hepato-

cellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray, *PLoS ONE*, Vol. 9, No. 9, p. e106314 (2014).

[16] Murakami, Y., Kubo, S., Tamori, A., Itami, S., Kawamura, E., Iwaisako, K., Ikeda, K., Kawada, N., Ochiya, T. and Taguchi, Y. H.: Comprehensive analysis of transcriptome and metabolome analysis in Intrahepatic Cholangiocarcinoma and Hepatocellular Carcinoma, *Sci Rep*, Vol. 5, p. 16294 (2015).

[17] Umeyama, H., Iwadate, M. and Taguchi, Y. H.: TINAGL1 and B3GALNT1 are potential therapy target genes to suppress metastasis in non-small cell lung cancer, *BMC Genomics*, Vol. 15 Suppl 9, p. S2 (2014).

[18] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: Heuristic principal component analysis-based unsupervised feature extraction and its application to gene expression analysis of amyotrophic lateral sclerosis data sets, *Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2015 IEEE Conference on*, pp. 1–10 (online), DOI: 10.1109/CIBCB.2015.7300274 (2015).

[19] Taguchi, Y. H., Iwadate, M., Umeyama, H., Murakami, Y. and Okamoto, A.: Heuristic principal component analysis-aased unsupervised feature extraction and its application to bioinformatics, *Big Data Analytics in Bioinformatics and Healthcare* (Wang, B., Li, R. and Perrizo, W., eds.), pp. 138–162 (2015).

[20] Taguchi, Y. H.: Integrative Analysis of Gene Expression and Promoter Methylation during Reprogramming of a Non-Small-Cell Lung Cancer Cell Line Using Principal Component Analysis-Based Unsupervised Feature Extraction, *Intelligent Computing in Bioinformatics* (Huang, D.-S., Han, K. and Gromiha, M., eds.), LNCS, Vol. 8590, Springer International Publishing, Heidelberg, pp. 445–455 (2014).

[21] Taguchi, Y. H.: Identification of aberrant gene expression associated with aberrant promoter methylation in primordial germ cells between E13 and E16 rat F3 generation vinclozolin lineage, *BMC Bioinformatics*, Vol. 16 Suppl 18, p. S16 (2015).

[22] Taguchi, Y. H.: Identification of More Feasible MicroRNA-mRNA Interactions within Multiple Cancers Using Principal Component Analysis Based Unsupervised Feature Extraction, *Int J Mol Sci*, Vol. 17, No. 5, p. E696 (2016).

[23] Taguchi, Y. H.: Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression, *BioData Min*, Vol. 9, p. 22 (2016).

[24] Taguchi, Y. H., Iwadate, M. and Umeyama, H.: SFRP1 is a possible candidate for epigenetic therapy in non-small cell lung cancer, *BMC Med Genomics*, Vol. 9 Suppl 1, p. 28 (2016).

[25] Taguchi, Y.-H.: Tensor decomposition-based unsupervised feature extraction identifies candidate genes that induce post-traumatic stress disorder-mediated heart diseases, *BMC Med. Genomics* (2017). in press.

[26] Lathauwer, L. D., Moor, B. D. and Vandewalle, J.: A Multilinear Singular Value Decomposition, *SIAM Journal on Matrix Analysis and Applications*, Vol. 21, No. 4, pp. 1253–1278 (online), DOI: 10.1137/S0895479896305696 (2000).

[27] 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).

[28] 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).