

乳がんデータセットを用いた 主成分分析を用いた教師なし学習による変数選択法と 部分的最小二乗回帰との比較

藤澤 孝太^{1,a)} 中村 晃輔¹ 宮田 龍太^{†1,b)}

概要: 本研究では、主成分分析を用いた教師なし学習による変数選択法 (PCAUF) の実データに対する有用性を評価する。具体的には、乳がん患者の遺伝子発現プロファイルに PCAUF を適用し、生理学的に裏付けされた乳がん関連遺伝子をどれだけ選択できるか検証する。その結果、約 7000 個ある候補から 67 個に限定でき、うち 6 個が既知の乳がん関連遺伝子であった。教師あり学習を用いた従来手法では 5 個しか拾えなかったので、PCAUF の方が検出力が高いと考えられる。

キーワード: 主成分分析を用いた教師なし学習による変数選択法, 部分的最小二乗回帰, 乳がん

Comparison between the principal components analysis based unsupervised feature extraction and the partial least squares regression using the breast cancer dataset

FUJISAWA KOTA^{1,a)} NAKAMURA KOSUKE¹ MIYATA RYOTA^{†1,b)}

Abstract: In this study, we evaluated the performance of the principal components analysis based unsupervised feature (PCAUF) extraction with a real data set. Concretely, the PCAUF was applied to the gene expression profiles and verified how many genes physiologically related to breast cancer could be selected. As a result, approximately 7,000 candidates could be limited to 67 genes. Moreover, 6 of them were known as the genes related to breast cancer, while the conventional method using supervised learning i.e., partial least squares regression (PLSR) could pick up only the 5 genes. Therefore, we concluded that the PCAUF could be more powerful method of feature selection.

Keywords: principal components analysis based unsupervised feature, partial least squares regression, breast cancer

1. はじめに

近年、機械学習的手法を用いて、ある疾患に関与する遺伝子群を数万ある候補から特定する研究がいくつか行われており、教師あり学習に基づいた変数選択法 (例えば

[1], [2]) の利用が主流となっている。しかし、重篤な症状を引き起こす疾患は事例が少ない場合が多く、お手本となるデータを大量に必要とする教師あり学習では有意な遺伝子選択は困難であった。

データドリブンで選択基準を与えるべく、Taguchi (例えば [3], [4], [5], [6]) は教師なし学習の代表格である主成分分析 (principal component analysis, PCA) に基づいた変数選択法 (PCA-based unsupervised feature extraction, PCAUF) を提案した。文献 [3] で Taguchi は、人工データにより PCAUF の有用性を証明した後、 Deng 熱に関

¹ 琉球大学大学院理工学研究科
IPJS, Nakagami, Okinawa, 903-0213, Japan

^{†1} 現在, 琉球大学工学部
Presently with University of the Ryukyus

^{a)} k198422@eve.u-ryukyu.ac.jp

^{b)} miyata26@tec/u-ryukyu.ac.jp

する実データに適用し、疾患関連遺伝子の特定を試みた。しかし、選択された遺伝子群に対する生物学的解釈はなされていたものの、まだ生理学的な裏付けがないため、実データに対する PCAUFE の有用性については議論の余地がある。

そこで本研究では、デング熱よりも生理学的な解明が進んでいる乳がんに関するデータセットに PCAUFE を適用し、実データに対する有用性を検証する。ここで用いるデータセットには生理学的に乳がんとの関連が指摘されている遺伝子が7つあり、PCAUFE を適用することでいくつ検出できるか、またそのデータセットに部分的最小二乗回帰に基づく変数選択法 (partial least squares regression, PLSR) を適用した先行研究 [7] の結果と比較することで、PCAUFE の性能を評価する。

先行研究より一つでも多く乳がんに関連する遺伝子が選択されれば、PCAUFE の有用性を実データで証明できた一例となる。

2. データセットと方法

2.1 乳がんデータセット [7], [8]

本研究で用いた mRNA プロファイルはすべて NCBI GEO [9] からダウンロードした。ここでは解析対象のデータセットとして、GSE16443 [7] を使用する。このデータセットには各被験者から採取した血液中の各 mRNA の発現量が記載されており、被験者ごとに標準化が施されている。乳がん患者 (cancer) 67 人、健常者 (healthy) 60 人、そして 11,217 個のプロープで構成されている。ただし、プロープの重複を排除した遺伝子数は 6,968 個である。また、このデータセットの乳がん患者は、発症したがんの種類によって①非浸潤性乳管がん (ductal carcinoma in situ, DCIS: 10 人)、②浸潤乳管がん (invasive ductal carcinoma, IDC: 49 人)、③浸潤性小葉がん (invasive lobular carcinoma, ILC: 4 人)、④その他の浸潤性乳がん (4 人) に細分化されている (詳細は文献 [7] を参照)。そして文献 [7] によると、生理学的に裏付けされている乳がん関連遺伝子として以下の7つを挙げている: CAMP, CSTA, DEFA3, LCN2, LTF, PPBP, S100A12。

2.2 PCAUFE [3]

本節では、上述の乳がんデータセット [7] に PCAUFE を適用した手順の概要について述べる。PCAUFE の詳細については、例えば文献 [3], [4], [5], [6] を参照されたい。

まず、GSE16433 [7] のサンプルを列、候補となる遺伝子の mRNA プロープを行としたデータセットに対して PCA を適用し、主成分負荷量と主成分得点をそれぞれ算出する。この操作の結果、サンプルに主成分得点が、プロープに負荷量が割り当てられる通常の PCA とは逆に、PCAUFE では主成分負荷量が各サンプルとの関連を表し、各プロープ (遺伝子) に主成分得点が付与される。

次に主成分負荷量を参考に、遺伝子選択に用いる主成分 (軸) を選択する。具体的には、各サンプルにあらかじめラベル付けされている healthy と cancer の 2 群間で、 t 検定により主成分負荷量に統計的有意差が確認できる主成分を 2 軸選択する。

そして選択された主成分負荷量に対応する主成分得点を標準化したものを 2 乗し、プロープごとに足し合わせる。その値に自由度 2 のカイ二乗分布を適用し、求めた P 値を Benjamini & Hochberg 法により調整し、 P 値が 0.01 未満のプロープのみを選択する。

最後に、選択されたプロープに対応する遺伝子を乳がん関連遺伝子として列挙する。

2.3 PLSR に基づいた変数選択法 (従来手法) [7]

本節では、前述の PCAUFE の比較対象として、先行研究 [7] で Aaroe et al. が採用した PLSR に基づく変数選択法の概要について述べる。

PLSR は、目的変数との共分散が大きい主成分のみで線形回帰を行う手法である。多重共線性の問題を回避でき、サンプル数が少なくても安定したパラメータ推定値が得られるため、バイオインフォマティクスでも広く使われている。手法の詳細については、文献 [10] を参照されたい。

先行研究 [7] では、まず GSE16443 [7] データセットの cancer (= 1), healthy (= -1) を目的変数とし、各プロープの発現量を次元圧縮前の説明変数とした PLSR モデルで leave-one-out 交差検証を行う。

次に、leave-one-out 交差検証で得られた各回帰係数にジャックナイフ検定を行い、 P 値が 0.05 未満となる回帰係数をもつ主成分に関連するプロープを選択する。この操作は逆説的に言えば、0 に近い回帰係数をもつ、すなわち統計的に疾患発症とは無関連なプロープを除外することに相当する。

最後に、選ばれたプロープに the database for annotation, visualization and integrated discovery (DAVID) [11] で gene ontology (GO) 解析を適用しプロープをアノテーションした上で、誤発見率 20% 未満の遺伝子を乳がん関連遺伝子として列挙する。

詳細は文献 [7] を参照されたい。

3. 結果

3.1 乳がんデータセットへの PCAUFE の適用

まず、GSE16433 [7] のサンプルを列、項目 (候補となる遺伝子の mRNA) を行としたデータセットに対して PCA を適用し、サンプルに関する主成分負荷量を求めた。主成分負荷量において t 検定で乳がん患者と非患者とで統計的有意差が確認できた主成分は次の 5 つであった: PCs 4, 18, 21, 32, 105。その中でも P 値が小さかった第 18, 21 主成分について、ここでは着目する。図 1 にそれらの主成分負

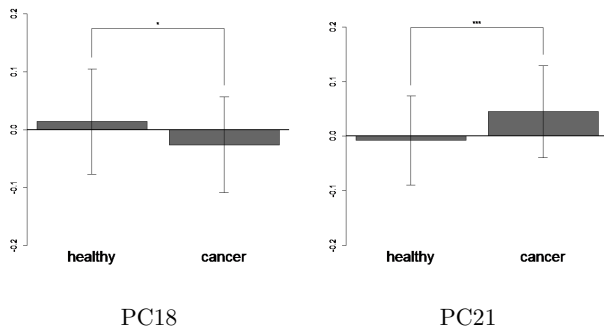


図 1 第 18, 21 主成分負荷量 (t 検定, $* p < 0.05$, $*** p < 0.001$).
Fig. 1: PC18 and PC21 loadings ($*** : p < 0.001$ and $* : p < 0.05$ in t test).

荷量をそれぞれ示す。図 1 の PC18 より、第 18 主成分負荷量の平均は healthy で正、cancer で負の値であった。一方図 1 の PC21 より、第 21 主成分負荷量の平均は healthy で負、cancer で正の値であった。

次に、表 1 の 3 列目に PCAUFE で選択できたプローブ数、遺伝子数、生理学的根拠を持つ遺伝子の数を示す。PCAUFE によって選択されたプローブ数は 6968 個から 111 個まで限定できた。PCAUFE で選択されたプローブに関連した遺伝子を抽出したところ、選ばれた遺伝子は 67 種類であった。それらの遺伝子を表 2 に記す。そのうち文献 [7] で生理学的根拠を持つと言われていたものが 6 個あった：CAMP, DEFA3, LCN2, LTF, PPBP, S100A12。

3.2 PCAUFE と従来手法 [7] との比較

表 1 の 4 列目に Aarøe et al. [7] の結果を示す。選択されたプローブの数は 738 個であった。PLSR で選ばれたプローブに DAVID で GO 解析を適用し、アノテーションした結果、選ばれた遺伝子は 67 種類であった。表 3 にそれらの遺伝子を示す。そのうち生理学的根拠を持つものは次の 5 個であった：CAMP, DEFA3, LTF, PPBP, S100A12。

また、表 1 の 4 列目に PCAUFE と PLSR で選ばれた遺伝子の論理積（アンサンブル）で選択した結果を示す。PCAUFE と PLSR の論理積（アンサンブル）で選択された遺伝子は 7 個であった。表 2, 3 ではその 7 個に * をつけた。そのうち生理学的根拠を持つものは 5 個あった：CAMP, DEFA3, LTF, PPBP, S100A12。よって、LCN2 は PCAUFE のみで選択できた。しかし、CSTA はどちらでも検出できなかった。これらの結果を踏まえた散布図を図 2 に示す。この図より、PCAUFE では特徴空間の原点からより距離があるプローブが選ばれていることがわかる。

4. 考察

図 1 より、本解析の場合、患者と非患者とで平均的に主成分負荷量が正負で分かれており、PC18 の値が減少または PC21 の値が増大するほど、乳がん患者の傾向が高まる

表 1 PCAUFE と PLSR の比較。

Table 1 Comparison between the PCAUFE and the PLSR.

	データセット	PCAUFE	PLSR	論理積
プローブ数	11217	111	738	7
遺伝子数	6968	67	67	7
乳がん遺伝子数	7	6	5	5

ことが示唆された。

ここで、図 2 より、PCAUFE で選択できた生理学的根拠が取れる 6 種の遺伝子（CAMP, DEFA3, LTF, PPBP, S100A12：丸とアスタリスクがついているもの、LCN2 丸印と濃い十字がついているもの）はすべて特徴空間上では、第 18 主成分の値が負でかつ第 21 主成分の値が正という cancer（乳がん患者）傾向の高い領域に位置していることがわかる。よって、この乳がんデータセットの場合、PCAUFE で選択された遺伝子は生理学的知見からも、整合性が取れる結果が得られたといえる。

また表 1 に示す通り、PCAUFE と PLSR とのアンサンブル（論理積）で選択できた遺伝子 7 つのうち、CAMP, DEFA3, LTF, PPBP, S100A12 の 5 種類は細菌の防御応答に関連する遺伝子、すなわち乳がんのバイオマーカーであった [7]。これらの遺伝子は好中球の活性化において高度に発現する。好中球は白血球の一種で 50~70% を占め、病原体を認識し殺す役割を持っている。また、腫瘍細胞にも反応する [13]。好中球の増加は急性の細菌感染症の合図であるが、がん患者にもリンパ球の減少とともに報告されていて、好中球/リンパ球比が上昇すると言われていた [14][15]。

PCAUFE と PLSR を用いた先行研究 [7] との比較を行う。表 1 より、プローブ数に関しては、PLSR よりも PCAUFE を用いた方が圧倒的に限定できることがわかる。一方、遺伝子数に関しては、両手法ともに同じ個数に関連遺伝子を限定できることがわかる。これは、Aarøe ら [7] が PLSR 適用後に DAVID で GO 解析を施すことで、PLSR で選択されたプローブに生物学的な解釈を行うことによる部分が大きい。PCAUFE は完全なるデータドリブン型アプローチで、プローブの重複を除外する以外は遺伝子数を限定する操作を行わないため、生命科学の専門知識が乏しくても解析できるという観点からも、PLSR より有用である。

さらに、限定できた個数は両者でたまたま同数でも、選択された遺伝子群の種類はまったく異なるため、PCAUFE と PLSR で論理積をとるアンサンブルアプローチを採用すると、候補をたった 7 種類までに限定できる。PLSR はラベルとプローブとの相関が大きくなるように少数の変数で表現し、その変数に回帰を行うのでラベルとプローブとの相関があるものを選択し、DAVID で GO 解析を施すことで上流の遺伝子だけを選択する。一方、PCAUFE はサン

表 2 PCAUFE によって選ばれた遺伝子. 遺伝子数は 67 個である. * がついている遺伝子は PLSR と PCAUFE 両方で選ばれた遺伝子である.

Table 2 The genes selected by the PCAUFE, of which the number was 67. The genes marked with * are the genes selected both with the PCAUFE and the PLSR.

ADAM28	ALPL	ANKRD22	BANK1	BLK	BLNK	CAMP ○ *
CCR3	CD160	CD19	CD22	CD72	CD79A	CH13L1
CLC	CPA3	CST7	CTSG	CTTN	DEFA1	DEFA3 ○ *
DHRS9	FCER1A	FCER2	FCGR1A	FER1L3	FLJ10665	FLJ39378
FOLR3	FREB	G0S2	GBP1	GBP5	GNG11	GP1BA
GPR44	GZMH	HDC *	HLA-DQA2	HLA-DRB3	HSU79303	IFIT1
IFITM3	INDO	KCNH2	KRTAP19-2	LCN2 ○	LOC440607	LTF ○ *
MGC33414	MPL	MX1	MYOM2	OAS2	P2RX5	P2RY14
PF4 *	POU2AF1	PPBP ○ *	PSTPIP2	RBMS3	RPS26	S100A12 ○ *
S100P	STAC	SYTL2	VPREB3			

表 3 PLSR によって選ばれた遺伝子 [7]. 遺伝子数は 67 種類である. * がついている遺伝子は PCAUFE と PLSR 両方で選ばれた遺伝子である.

Table 3 The genes selected by the PLSR, the number was 67. The genes marked with * are the genes selected both with the PCAUFE and the PLSR.

AARSD1	ACAA1	ACAT2	AIF1	ANXA1	ATP5E	ATP5L
ATP6V0B	C10orf33	C8B	CAMP *	CDKN2D	CLU	CMAS
CMTM5	CXCR3	CYP2J2	DEFA3 *	EEF1G	ETF1	FDPS
GP1BB	GRN	GUK1	HDC *	HEXA	IDI1	INSIG1
LASS6	LOC440587	LTF *	MBTPS1	MTHFS	NMI	OSBPL7
PECI	PEMT	PF4 *	PLAA	PMVK	PPBP *	PRODH
RALBP1	RNPEPL1	RPL11	RPL12	RPL14	RPL15	RPL17
RPL24	RPL26L1	RPL36A	RPL37A	RPL4	RPL6	RPS25
RPS29	RPS3A	S100A12 *	STAT3	SULT1A2	SULT1A2	SULT1A4
SYK	TIRAP	TRSPAP1	UBA52			

プルに 2 群の差が表れるような軸を選択し, カイ二乗分布によって p 値 0.01 以下のプローブを選択することで, 病気との相関が「かなり高い」ものだけを上流/下流にかかわらず選択するので, 共通して選択される遺伝子が少ないと考えられる. 生命科学の分野では, これらの機械学習のアプローチで選択された遺伝子群が本当に生理学的にも病気と関連していると言えるのか, 実際には wet 実験で証明しなければならない. 実験では候補となる遺伝子が多くてはコストがかかってしまうので, 候補となる遺伝子は少ないほうがよい. よってコストパフォーマンスを重視するなら, PCAUFE と PLSR との論理積が, 選択された遺伝子 7 個のうち 5 個がバイオマーカーだったため, 最良の疾患関連遺伝子選択方法と言えるだろう. もしかしたら, このアンサンブル法で選択された残り 2 種の遺伝子 (図 2 * だけついているもの) も, 現時点でバイオマーカー認定されていないだけであって, wet 実験でさらなる確証が得られれば, 将来有用なバイオマーカーになり得るかもしれない.

次に, PCAUFE で生理学的に乳がんとの関連が実際に

示唆されている遺伝子がいくつ検出できるか, について言及する. Aarøe ら [7] は乳がんに関連する遺伝子として 7 個の遺伝子に着目していた. それぞれの手法で 7 個中何個検出できたかを表 1 にも記載した. 表 1 より, PCAUFE で検出できた乳がんのバイオマーカー遺伝子は 6 個であり, Aarøe ら [7] の PLSR を用いて検出できた数より一つ多い. 表 4 に, PCAUFE で検出できた / できなかったバイオマーカーを示す. この表より, PCAUFE は PLSR で検出できなかった LCN2 を検出できたことがわかる. 侵襲性乳がんの組織や尿サンプルから高い濃度の LCN2 が検出されたという報告がされていて [12], 乳がんのバイオマーカーとして提案されている [7].

さらに, PCAUFE でも PLSR でも検出できなかったバイオマーカー CSTA について考察する. この遺伝子はバイオマーカーの中でも予後マーカーと呼ばれるもので, 特定の治療によらない病気の経過を予測するための指標である [16][17]. 本研究では病気の経過を考慮せずに解析を行ったので, がんが進行していない患者が多かった場合は

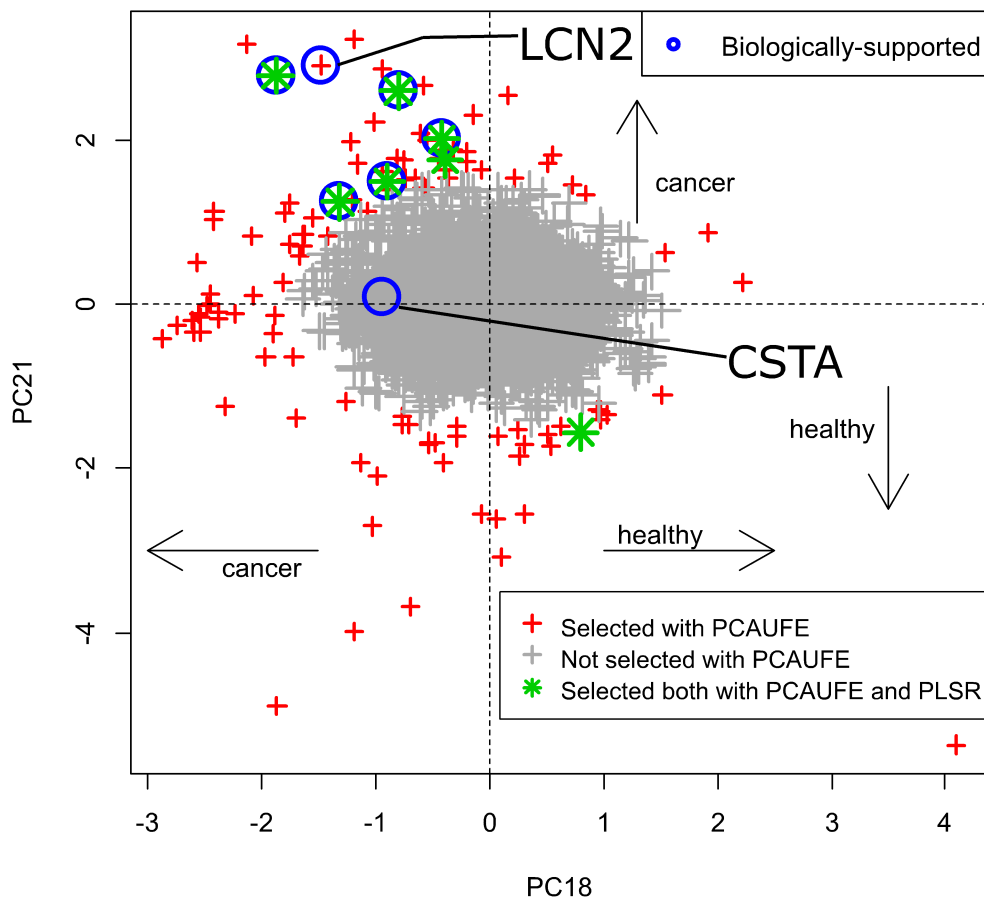


図 2 PCAUFÉ における主成分得点の散布図. 濃い十字は PCAUFÉ で選ばれたプローブ, 薄い十字は選ばれなかったプローブである. 図中のアスタリスクは PCAUFÉ と PLSR とのアンサンブル (論理積) によって選ばれた 7 種類の遺伝子に関するプローブを表す. 図中の○は Aarøe et al. [7] で言及されていた乳がんに関連する遺伝子である.

Fig. 2 Biplots of PC score (PCAUFÉ results). The light color crosses represent the probes which aren't selected. The dark color crosses correspond to the selected probes. The circles in the figure are genes mentioned in Aarøe et al. [7] and related to breast cancer. The asterisks in the figure represent probes attributed to 7 genes the gene selected both with PCAUFÉ and PLSR.

CSTA が選択できない可能性が高い. また, 図 2 ではひし形印が CSTA に対応しており, PC18 軸上では cancer より (< 0) の傾向が見えるものの, PC21 軸上では cancer と healthy とのほぼ境界 (≈ 0) に位置しており, 軸の選択次第では選択できた可能性も否定できない. よって, PCAUFÉ の検出力を向上させるための今後の課題として, 既知のバイオマーカーを可能な限り検出できる最適な軸の探索方法が挙げられる.

最後に確認として, PCAUFÉ で選択された 67 種の遺伝子のみを列として持つデータセットに PCA を適用し, サンプルに関する主成分得点で患者と非患者とで統計的に有意差が出ることを確認した. さらに, がんの種類で DCIS, IDC, ILS, その他の乳がん患者で主成分得点の重点を求めたところ正負で分かれることを確認した.

5. おわりに

本研究では, 生理学的に乳がんとの関連が示唆されている遺伝子 7 つを含むデータセットに対して PCAUFÉ を適用し, 7 個中何個選択できたか, 同じデータセットに PLSR を適用した先行研究 [7] と比較した.

PCAUFÉ を用いると, 乳がんのバイオマーカーとなる遺伝子 7 個中 6 個が選択でき, 先行研究 [7] よりも一つ多く検出できた. どちらの手法も 6,968 個あった候補から最終的に 67 個まで遺伝子を限定できたため, PCAUFÉ の方が PLSR よりも検出力は高いと言える.

表 4 各手法で検出できたバイオマーカー.

Table 4 Biomarkers detected by each method.

PCAUFÉ & PLSR					PCAUFÉ	unselected
CAMP	DEFA3	LTF	PPBP	S100A12	LCN2	CSTA

また、各手法で検出された遺伝子の個数は偶然一致していたものの、選択された遺伝子の種類の間にはオーバーラップはほとんど見られなかったため、PCAUFとPLSRとで論理積を取るアンサンブル的アプローチを採用すると、候補となる遺伝子は7つまで絞ることができた。その中で乳がんとの関連に生理学的根拠を持つ遺伝子は5個あり、すべて細菌の防御応答に関連する遺伝子であった。

これらのバイオマーカー遺伝子と上述のアンサンブル的アプローチで選択した遺伝子9種に関するプローブは、いずれもPCAUFの特徴空間上で患者/非患者の区別の整合性が取れる軸上に位置しており、軸の選択次第でさらに検出力を向上できる可能性があることが示唆された。

参考文献

- [1] Schwender H. (2012) *siggenes: Multiple Testing using SAM and Efron's Empirical Bayes Approaches*. (<https://rdrr.io/bioc/siggenes/>)
- [2] Ritchie M.E., Phipson B., Wu D., Hu Y., Law C.W., Shi W, Smyth G.K. (2015) limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research* vol.43, no.7, pp:e47.
- [3] Taguchi Y (2017) Principal components analysis based unsupervised feature extraction applied to gene expression analysis of blood from dengue haemorrhagic fever patients. *Scientific Reports*. 7:44016-44029.
- [4] Taguchi Y.H. (2016) 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, pp:e696.
- [5] Taguchi Y.H. (2016) 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* 8:1-18.
- [6] Taguchi Y. H., Iwadate M., Umeyama H. (2015) Principal component analysis-based unsupervised feature extraction applied to in silico drug discovery for post-traumatic stress disorder-mediated heart disease. *BMC Bioinformatics* 16:139.
- [7] Jørgen Aarøe, Torbjørn Lindahl, Vanessa Dumeaux, Solve Sæbø, Derek Tobin, Nina Hagen, Per Skaane, Anders Lønneborg, Praveen Sharma, and Anne-Lise Børresen-Dale (2010) Gene expression profiling of peripheral blood cells for early detection of breast cancer. *Breast Cancer Research*. 12:R7. doi: 10.1186/bcr2472.
- [8] <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE16443>
- [9] NCBI: Gene Expression Omnibus, NCBI (online), available from (<https://www.ncbi.nlm.nih.gov/geo/>) (accessed 2019-5-16)
- [10] Paul Geladi and Bruce R. Kowalski (1986) Partial least-squares regression: a tutorial. *Analytica Chimica Acta*. 185:1-17.
- [11] <https://david.ncifcrf.gov/>
- [12] Yang J, Bielenberg DR, Rodig SJ, Doiron R, Clifton MC, Kung AL, Strong RK, Zurakowski D, Moses MA (2009) Lipocalin 2 promotes breast cancer progression. *Proc Natl Acad Sci USA*. 106:3913-3918. doi: 10.1073/pnas.0810617106.
- [13] Koga Y, Matsuzaki A, Suminoe A, Hattori H, Hara T (2004) Neutrophil-derived TNF-related apoptosis-inducing ligand (TRAIL): a novel mechanism of antitumor effect by neutrophils. *Cancer Research*. 64:1037-1043. doi: 10.1158/0008-5472.CAN-03-1808.
- [14] Cho H, Hur HW, Kim SW, Kim SH, Kim JH, Kim YT, Lee K (2008) Pre-treatment neutrophil to lymphocyte ratio is elevated in epithelial ovarian cancer and predicts survival after treatment. *Cancer Immunol Immunother*. 58:15-23. doi: 10.1007/s00262-008-0516-3.
- [15] Walsh SR, Cook EJ, Goulder F, Justin TA, Keeling NJ (2005) Neutrophil-lymphocyte ratio as a prognostic factor in colorectal cancer. *J Surg Oncol*. 91:181-184. doi: 10.1002/jso.20329.
- [16] Kuopio T, Kankaanranta A, Jalava P, Kronqvist P, Kotkansalo T, Weber E, Collan Y (1998) Cysteine proteinase inhibitor cystatin A in breast cancer. *Cancer Research*. 58:432-436
- [17] Levicar N, Kos J, Blejec A, Golouh R, Vrhovec I, Frkovic-Grazio S, Lah TT. (2002) Comparison of potential biological markers cathepsin B, cathepsin L, stefin A and stefin B with urokinase and plasminogen activator inhibitor-1 and clinicopathological data of breast carcinoma patients. *Cancer Detection and Prevention*. 26:42-49. doi: 10.1016/S0361-090X(02)00015-6.