

# 主成分分析を用いた教師なし学習による出芽酵母の時間周期 遺伝子発現プロファイルの解析

田口 善弘<sup>1,a)</sup>

**概要:** 遺伝子発現プロファイルから「重要な」遺伝子を選択することは広く行われているが依然として「統計的に有意に大きな発現差があるもの」という古典的な基準が大半の研究で行われているのが現状である。しかし、この戦略は多クラスがある場合の実験には適応が難しい。ペアワイズに比較するとペアの数だけ遺伝子セットが選ばれてしまい、解釈が難しくなる。著者は近年、この問題を解決するため、主成分分析を用いた教師なし学習による変数選択を提案し、様々な場合の遺伝子選択に用いてきた。本研究では出芽酵母の細胞分裂周期遺伝子の選択というやり尽くされた感がある問題にあえてこの手法を用いることで一般的な正弦回帰による変数選択を凌ぐ性能をあげることを見出した [1] のでここに報告する。

## Principal component analysis based unsupervised feature extraction applied to budding yeast temporally periodic gene expression

Y-H. TAGUCHI<sup>1,a)</sup>

**Abstract:** Although identifying “critical genes” based upon gene expression profiles are widely employed, significant differential expression is an only criterion widely employed; this strategy is hardly applied to multi class problems. The series of pairwise comparisons results in identification of sets of genes as many as the number of pairwise comparisons, which makes the interpretation harder. Principal component analysis (PCA) based unsupervised feature extraction (FE) was recently proposed and applied to many biological problems by the author in order to overcome this difficulty. In this study, I intentionally applied this methodology to the extensively studied problem, budding yeast cell division cycle genes identification. PCA based unsupervised FE turned out to outperform traditional sinusoidal fittings.

### 1. はじめに

遺伝子発現プロファイルから「重要な遺伝子」を同定することは広く行われている。例えば、患者と健常者の遺伝子発現プロファイルを比較して、健常者との比較で、患者側で有意に発現が上昇・下降している遺伝子を同定することで疾患原因遺伝子の候補を同定する、などである。これはデータ科学的には変数選択問題であり、これ自体は特に新しい問題とは言えない。しかし、遺伝子発現プロファイルに基づく遺伝子選択には、通常の変数選択問題とは本質的に異った部分が多々ある。まず、変数の数がサンプルの

数に比べて圧倒的に多いということである。サンプルの数は実験動物の個体数や患者の数であり、数万サンプルを集めるのは容易ではない。一方、遺伝子は数万個あるので、サンプル数が遺伝子の数より圧倒的に少ない。この様な場合、なんらかの目的関数を設定しても、サンプル数以上の変数を選んでしまうと、必ず「解」がある（条件式の数と変数の数が同じ）ことになってしまい選択ができない（どの変数を選んで解になるのだから）。この様な場合にはいわゆる交差検定を用いて過学習を防ぐのが一般的であるが、交差検定は解析的に実行することができず、変数選択を試行錯誤で行うしか方法がない。この場合、多数の遺伝子から少数個（≒サンプル数程度）の遺伝子を選択する組み合わせは非常に多数になってしまうため、収束に困難を生じることが多い（例えば、多数の局所解が存在すること

<sup>1</sup> 中央大学理工学部物理学科 Tokyo 112-8551, Japan

<sup>a)</sup> tag@granular.com

本研究内容はすでに原著論文として刊行済み [1] である

でユニークな解に辿りつかない、選択肢が多すぎてそもそも局所解にさえ辿りつかない、など)。この様な場合の解決法として正則化項をつけて変数選択の範囲を狭める試みもなされている [2] が、正則化項を用いた変数選択には、その当初の設計思想上、変数選択の安定性はまったく考慮されておらず(サンプルごとに異った遺伝子が選択される)、どの遺伝子が生物学的に重要かという問に答えなくてはならない遺伝子選択には向いているとは言えない。

我々はこの様な問題を克服することを目的として近年、「主成分分析を用いた教師なし学習による変数選択」を提案し、様々な問題に適用してきた [3-18]。本研究では、出芽酵母の細胞分裂周期遺伝子探索というある意味やり尽くされた問題に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ることで、本手法の性能を確認することを試みた。本手法を出芽酵母の細胞分裂周期遺伝子探索に用いる理由は2つある。第一に、たった今述べたようにかなりやり尽くされた研究なので結果の生物学的な妥当性が判断しやすい。第二に、通常の遺伝子選択のようにある条件で発現が大きい/小さいという単純な条件では出芽酵母の細胞分裂周期遺伝子を探索できないことである。単純に細胞が分裂する瞬間に発現が変化している遺伝子を選べばいいというわけではない(実際、DNAの複製は細胞分裂がおきるはるか以前に終了している)ために、特定の時刻での単純な遺伝子発現量変化は重要ではない。出芽酵母の細胞分裂周期遺伝子探索問題に一番良く使われる方法は周期的に振動している遺伝子の同定だが、周期性という漠然とした基準を一般的に変数選択に直接適用するのは難しく(与えられた周期の周期関数というだけでは無限の選択肢があり、周期性という基準だけで優劣をつけるのは難しい)、遺伝子発現プロファイルの形状が正弦波であることを仮定する正弦回帰が用いられることが多い。しかし、冷静に考えれば、細胞分裂の周期関数が正弦波であることにはなんの生物学的な理由も無いことは明らかであり、正弦波であることに優先順位を与えることが生物学的に妥当かどうかは甚だ疑わしい。以下では、この問題に主成分分析を用いた教師なし学習による変数選択が非常に有効であることを示す。

## 2. 主成分分析を用いた教師なし学習による変数選択

本手法についての詳細は原著論文 [1] を見ていただくとして概要を述べる。まず、第一に主成分分析を用いて遺伝子を低次元空間に埋め込む。この結果、サンプルに主成分得点が、遺伝子に主成分負荷量が割り当てられる通常のサンプル埋め込みとは逆に、遺伝子に主成分得点が、サンプルに主成分負荷量が付与される。通常のサンプル埋め込みとの違いは、サンプル埋め込みの場合は共分散行列、または、相関係数行列が対角化されるために遺伝子ごとに発現プロファイルの平均がゼロになるのに対して、遺伝子埋め

込みの場合はグラム行列が対角化されるためにサンプルごとの平均プロファイルがゼロになる点である。また、一般に、サンプル平均、遺伝子ごと平均のそれぞれをゼロにした場合の効果は自明ではなく、この2つの主成分分析は一般には異った結果を与えることになり、どの様に異なるかを事前に計算することもできない点は注意を要する。また、サンプル平均と遺伝子平均を同時にゼロにすることはできないので、どちらか一方を選択することになる。主成分分析を用いた教師なし学習による変数選択において、サンプル平均をゼロにする遺伝子埋め込みの方を選択する理由は、一般に遺伝子発現プロファイルの計測では絶対値の計算は不可能であり、なんらかの正規化が必要なためである。遺伝子平均をゼロにするサンプル埋め込みの場合は、前処理として遺伝子発現プロファイルがサンプルごとに何らかの仮定に基づいてあらかじめ正規化されていることが多い。しかし、そこに任意性が入ってしまったのでは、せっかくの教師なし学習の意味がない。そこで、通常、主成分分析を用いた教師なし学習による変数選択ではサンプル平均をゼロにする方が選択される。

次に、遺伝子選択に用いる主成分を主成分負荷量を参考に選択する。今の場合、主成分負荷量はサンプルに紐付けられているために、主成分負荷量を観測することでどの主成分に生物学的に意味がある特徴が表れているかを知ることができる。本研究の場合には、2つの主成分が張る空間で軌跡がリミットサイクルを構成しているかどうかで選択を行った(それが周期関数であるということであるから)。

最後に、選択された主成分得点に多重ガウス分布を仮定し、 $x$ 二乗分布を適用して  $P$  値を計算し、外れ値となっている遺伝子を選択する(多重比較補正された  $P$  値が 0.01 以下を外れ値とする)。なぜなら、このような遺伝子は注目する主成分に対して、偶然では説明できない大きな寄与を持っている、つまり、周期性に特異的に寄与している遺伝子であるとみなすことができるからである。主成分得点がガウス分布に従うという仮定の是非には議論があるものと思われるが、この様な仮定は確率主成分分析と呼ばれる、主成分分析の統計学習的フレームでの再解釈でも採用されており [19]、それほど現実から離れたものではないと思われる。

## 3. 結果

### 3.1 人工データ

本節では主成分分析を用いた教師なし学習による変数選択の有効性をデモンストレーションするため、人工データを用いて同手法のパフォーマンスを確認する。図1は人工データを用いた結果である。この人工データは1万個の遺伝子の長さ100の時系列データからなっている。1万個の遺伝子のうち100個だけが長さ25の周期関数(つまり長さ100で4周期)になっておりそれ以外はノイズで

表 1 人工データの混同行列。P 値は BH 基準で補正されている。

A	1		2		3	
主成分分析	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$
$i > 100$	9900	0	9900	0	9900	0
$i \leq 100$	0	100	0	100	0	100
正弦回帰	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$
$i > 100$	9900	0	9900	0	9900	0
$i \leq 100$	0	100	17	83	48	52

A	4		5		6	
主成分分析	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$
$i > 100$	9900	0	9900	0	9900	0
$i \leq 100$	0	100	0	100	0	100
正弦回帰	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$	$P \geq 0.01$	$P < 0.01$
$i > 100$	9900	0	9900	0	9900	0
$i \leq 100$	64	36	72	28	85	15

ある。周期データは図 1 の (C) と (D) に青で描かれている正弦/余弦関数に振幅の  $A$  倍のノイズを加えて作られている (図 1(C) と (D) の黒線)。これらのノイズは、乱数だが 4 周期 (1 周期の長さは 25) は完全に満たしており、周期性を満たすノイズになっている。周期性を満たす 100 個の遺伝子の発現プロファイルは、この 2 つの周期関数を任意の比率で線形結合させて作られた、やはり周期 4 を満たすランダムな時系列になっている。これら周期的な遺伝子 100 個とそれ以外の単なるノイズからなる総計 1 万個の遺伝子を主成分分析で 2 次元に埋め込みした結果が図 1(A) である。赤い丸が 100 個の周期遺伝子である。周期性のある遺伝子がそれ以外の遺伝子ときれいに分離されていることがわかる。図 1(C) と (D) の赤線が第 1、2 主成分負荷量である。もとの「ランダム」周期関数をほぼ完全に再現していることがわかるだろう。このように、主成分分析を用いた教師なし学習による変数選択は、周期関数を全く仮定せずに、周期関数を単なるノイズからきれいに分離できることがわかる。図 1(B) は第 1、2 主成分得点で構成されるリミットサイクルである。4 色はそれぞれ、4 周期を表現している。乱れていて見難いがよく見ればきっちり 4 周期のリミットサイクルになっていることがわかるだろう。これをリミットサイクルであると視認することは困難だが、主成分分析を用いた教師なし学習による変数選択ではこれを容易に認識できるわけだ。

表 1 は 100 回の独立試行の平均で求められたパフォーマンスの結果である。主成分分析を用いた教師なし学習による変数選択、正弦回帰、ともに多重比較補正された  $P$  値が 0.01 以下のものを周期遺伝子として選択した (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リミットサイクルが構成されている平面において外れ値であること、正弦回帰の場合には正弦関数に近いこと、を意味する  $P$  値なのでその意味は全く異なる)。ノイズの正弦波に対する振幅比  $A$  が 1 の時はどちらも 100% の精度で選択できているが、 $A$  が増えるに従って正弦回帰の精度は低下

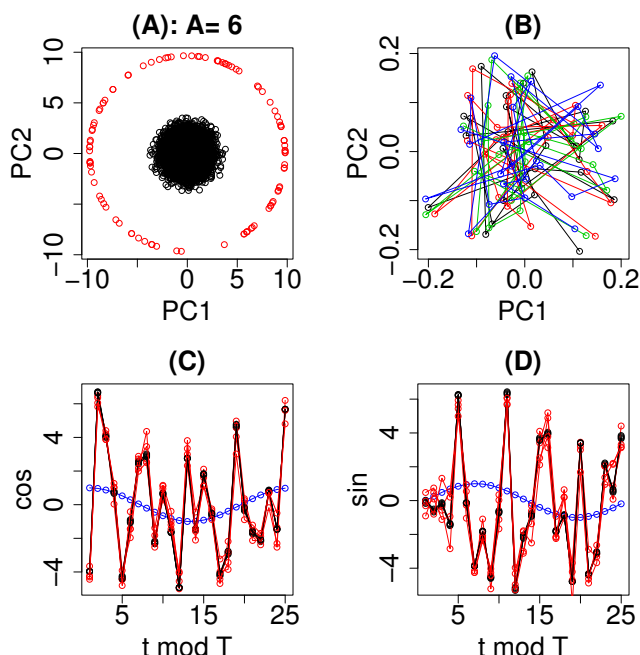


図 1 人工データへの主成分分析を用いた教師なし学習による変数選択の適用例。(A) 1 万個の遺伝子の第 1、2 主成分得点の散布図。赤が 100 個の周期遺伝子。(B) 第 1、2 主成分得点によって構成されるリミットサイクル。4 周期ごとに色が変えてある。(C)(D) 青: 正弦、余弦関数。黒: ノイズを加えた周期関数。赤: 第 1、2 主成分得点。

する。一方、主成分分析を用いた教師なし学習による変数選択の方は、全く影響を受けず、100% の精度を維持している。これはあくまで人工データではあるが、少なくとも、主成分分析を用いた教師なし学習による変数選択は、周期性を全く仮定しなくても周期関数を選択する能力があることがわかる。問題は、現実のデータの場合にこんなに都合よく行っているか、という問題である。以下の二小節では、出芽酵母の細胞分裂周期遺伝子の同定という「現実の」問題に主成分分析を用いた教師なし学習による変数選択を適用することで本小節で見たような主成分分析を用い

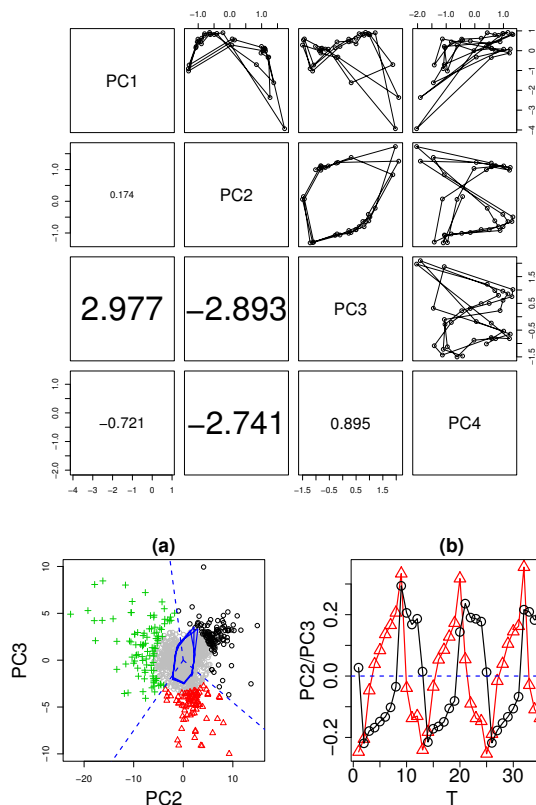


図 2 出芽酵母のメタボリックサイクルの結果。上: 第1 から4 主成分負荷量の散布図。下半分は周期性の基準となる回転数。下 (a) 第2、3 主成分得点の散布図。赤、黒、緑が選択された遺伝子で色は K-means のクラスタリングの結果。青線が主成分負荷量。(b) 第2 (黒)、3 (赤) 主成分負荷量。

た教師なし学習による変数選択の性能が現実のデータ解析にも見いだせるかを検証していく。

### 3.2 出芽酵母: メタボリックサイクル

酵母の細胞分裂サイクルは当たり前だが、同期しているわけではない。個々の出芽酵母が勝手な位相で細胞分裂を起こしていたのでは、当然のごとく周期性は見いだせない。何らかの方法で全ての酵母が同時に細胞分裂を起こすように同期させる必要がある。同期させるにはいろいろな方法があるが、その1 つに、増殖に必要なエネルギーを制限する方法がある。エネルギーを断れば増殖は一度停止する。そこで再度エネルギーを与えれば、全ての酵母が一斉に増殖を開始するのしばらくの間、細胞分裂周期は同期したままになる。これをメタボリックサイクルと呼ぶ。図 2 はメタボリックサイクルで同期させた出芽酵母の遺伝子発現プロファイルの解析に、主成分分析を用いた教師なし学習による変数選択を適用した場合の結果である。第1 から4 主成分負荷量にリミットサイクルが見える。周期性を示す回転数も3 周期程度を示している。3 周期を示している第2、3 主成分を用いて変数選択を行い、3 つのクラスター

に分けた結果が図 2 下 (a) である。この3 つの遺伝子群に GO タームのエンリッチメント解析を行うと、それぞれ、リボソーム、ミトコンドリア、細胞分裂に関係していた。これは元データを計測したグループ [20] が詳細な解析で同定した3 遺伝子グループの生物学的な機能とよく一致していた。これは完全に教師なし学習でも、専門家が詳細な生物学的な考察に基づいて同定した遺伝子群を容易に同定できることを示しており、教師なし学習の有用性をよく示している。

更に図 2 下 (b) に示されている周期関数は正弦波から大きくずれている。実際、正弦回帰による遺伝子の選択は生物学的な観点からみて、主成分分析を用いた教師なし学習による変数選択に劣っていることが示せる [1]。

これらのことから、主成分分析を用いた教師なし学習による変数選択は周期的な振動をする発現プロファイルをもつ遺伝子を、実際の周期の長さや関数形を仮定せずに同定出来るだけでなく、正弦回帰を用いた伝統的な方法より、より生物学的に妥当な遺伝子選択ができることを示している。

### 3.3 出芽酵母: 細胞分裂周期

出芽酵母の細胞分裂周期を揃えるより直接的な方法として、温度感受性変異体を用いる方法がある。この変異株は、ある特定の温度下では細胞分裂が停止してしまう。この性質を利用して出芽酵母の細胞分裂を一度停止、位相が揃ったところで一斉に細胞分裂を開始させることが可能になる。この方法は変異株を使ったより人工的な方法ではあるが、メタボリックサイクルを用いた同期と異なって、代謝パスウェイが周期的に振動して細胞分裂周期遺伝子とまざって検出されることが無い分、より直接的な実験法ではある。

この実験方法はメタボリックサイクルよりメジャーであり、すでに複数回の実験がなされている。これらは CycleBase [21] というデータベースに蓄積されている。今回、我々は CycleBase に登録されている複数の出芽酵母の細胞分裂時の遺伝子発現プロファイルへの主成分分析を用いた教師なし学習による変数選択の適用を試みた。その結果、登録されている8 つのプロファイルのうち、7 つについて、細胞周期に一致するリミットサイクルを構成する主成分負荷量が同定でき、対応する主成分得点で遺伝子選択を行うことに成功した。使えなかった1 つのプロファイルについてはあらかじめ周期性の強い一部の遺伝子しか登録されていず、外れ値を同定するために対照として必要な非周期遺伝子がなかったため、うまく行かなかったと思われる。

更に、これらの遺伝子の選択数は、各プロファイルについて200 個以下ずつだったにも関わらず、実に37 個の遺伝子が、7 つ中6 つ以上のプロファイルで同時に同定されていることがわかった。出芽酵母の遺伝子の総数は4000 個程度であり、そのうちの200 個というわずかに

5%程度にすぎない。この選択比率で37個、つまり各プロファイルで選択された遺伝子のうち20%以上が、7つ中6つのプロファイルで選択されるということは、主成分分析を用いた教師なし学習による変数選択による遺伝子選択が高度にロバストでサンプル依存性が非常に低いことを示している。

37個の遺伝子についてエンリッチメント解析を行ったところ、CycleBaseにおいて正弦回帰を用いてランキングされた上位の同数の遺伝子と同等以上の生物学的な妥当性をもつ遺伝子セットであることがわかった。これらの結果は、メタボリックサイクルにおける主成分分析を用いた教師なし学習による変数選択の成功が単なる偶然ではなく、非常にロバストであり、主成分分析を用いた教師なし学習による変数選択は周期性を仮定せずに周期的な振動をする遺伝子を選択するという一見矛盾した作業を効率よくできるものであることがわかる。

#### 4. 考察

主成分分析を用いた教師なし学習による変数選択は分裂酵母の細胞分裂周期遺伝子の同定において、広く使われている正弦回帰と同等以上のパフォーマンスを持つことが確認された。しかし、最初から周期関数を仮定した場合よりも何も仮定しない方がよりよい、というのは一見、矛盾であるように思われる。これにはいくつかの理由がある。まず、第一には正弦回帰は正弦関数という特定の周期関数を仮定していることである。図1を見るまでもなく、よい周期関数であることと正弦関数であることには本来はなんの関係もない。実際、図2下(b)に見るように、出芽酵母の周期的な遺伝子発現プロファイルは、正弦関数から大きくずれることもありうる。しかし、事前にこの関数形を知ることにはできない。正弦回帰を仮定する以上、この問題からは自由になれない。これは周期関数の同定に正弦回帰を使うことに対してペナルティとして働くだろう(関数形を仮定しない主成分分析を用いた教師なし学習による変数選択には当然のごとくこのような問題は存在しない)。一方、主成分分析はリミットサイクルの基底となる周期関数をその形のいかに関わらず同定できる可能性がある。主成分負荷量に、細胞分裂周期に相当するプロファイルが必ず現れるという保証は全く無いが、そもそも、細胞分裂を研究するための実験設定なのだから、主成分負荷量に細胞分裂周期に相当するプロファイルが現れてもなんの不思議もない。このことが主成分分析を用いた教師なし学習による変数選択が正弦回帰に勝る第一の理由であろう。

しかし、これだけが主成分分析を用いた教師なし学習による変数選択が正弦回帰に勝る唯一の理由ではない。正弦回帰では遺伝子発現プロファイルと正弦関数の相関係数の大小で遺伝子をランク付けしているが、主成分分析を用いた教師なし学習による変数選択では同定された主成分(こ

の場合は、リミットサイクルを構成するような主成分)に寄与が大きい物、つまりは射影が大きい物を選んでいる。射影と相関係数の違いは、前者が注目する成分以外の寄与が考慮されないのに対し、後者は注目する成分以外の寄与は全てペナルティになることである。後者の考えは果たして正しいだろうか? 相関係数を基準とする場合、つまり、正弦関数からのずれをペナルティとみなす場合には、細胞分裂周期以外に寄与がある遺伝子は、ランクが下がってしまう。しかし、遺伝子は単体で作用するものではなく、複数の遺伝子が協力して機能を発揮する場合も多い。その場合、細胞分裂周期に重要な役目を担っている遺伝子が別の重要な役割を果たしていることもありうるだろう。実際、図2上を見てもわかるように、第1から4主成分負荷量には8の字を構成している場合がある。この場合、8の字を構成しないもの、するものの、両方の負荷量が共に細胞周期と一致することは原理的にありえず、どちらかはペナルティとして扱われてしまう。しかし、この8の字が単なるノイズではなく、なんらかの生物学的な機能を担っている可能性は捨てきれない。その場合、どちらかをペナルティに扱うのは正しくないだろう。実際、図2下(b)にある2つの主成分負荷量への射影ではなく相関係数を基準にして遺伝子選択を行うとむしろ生物学的な妥当性が下がることがわかった(詳しくは原著論文 [1] を参照)。

#### 5. 終わりに

主成分分析を用いた教師なし学習による変数選択を出芽酵母の細胞分裂周期遺伝子の同定という古典的な問題に応用した結果、伝統的な正弦回帰を用いた方法と同等以上のパフォーマンスが得られた。教師なし学習はデータ駆動的な方法であり、遺伝子発現プロファイルの形状を正弦関数に近いと仮定する正弦回帰より前提が少なく、同等以上の性能が得られるなら、教師なし学習的な手法の方がロバストネスが高いことが期待される。主成分分析を用いた教師なし学習による変数選択は生物学的に意味がある成分は主成分負荷量に現れるはずだという仮定と、対応する主成分得点が高い遺伝子が、主成分負荷量が表現する生物学的な意味を担っているはずだという、ただそれだけの仮定に基づいた手法である。データ駆動型の手法であるために、人間の前提知識に関係なく生物学的に意味がある成分を発見的に見いだせる利点がある。世はすべからく、深層学習やベイズ統計のようなモデルベースの統計手法が全盛であるが、モデルベースの手法を用いるということは、逆にいえば、我々がモデル化できる程度には対象をよく理解していると仮定している事にもなる。それは本当に正しい仮定なのだろうか。教師なし学習と同等程度のパフォーマンスしか得られないのではその仮定は大いに疑わしいのではないだろうか。

参考文献

- [1] 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).
- [2] 川野秀一, 廣瀬慧, 立石正平 and 小西貞則: 回帰モデリングと  $L_1$  型正則化法の最近の展開, *日本統計学会誌*, Vol. 39, No. 2, pp. 211–242 (2010).
- [3] 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).
- [4] 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 (2016).
- [5] 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).
- [6] 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).
- [7] Taguchi, Y.-h., Iwadate, M., Umeyama, H., Murakami, Y. and Okamoto, A.: Heuristic principal component analysis-based 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).
- [8] 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).
- [9] 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).
- [10] 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).
- [11] 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).
- [12] Murakami, Y., Tanahashi, T., Okada, R., Toyoda, H., Kumada, T., Enomoto, M., Tamori, A., Kawada, N., Taguchi, Y. H. and Azuma, T.: Comparison of Hepatocellular Carcinoma miRNA Expression Profiling as Evaluated by Next Generation Sequencing and Microarray, *PLoS ONE*, Vol. 9, No. 9, p. e106314 (2014).
- [13] 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).
- [14] 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).
- [15] 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).
- [16] 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).
- [17] 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).
- [18] 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).
- [19] ビンヨップ, C.: パターン認識と機械学習 上, 丸善出版 (2012).
- [20] Tu, B. P., Kudlicki, A., Rowicka, M. and McKnight, S. L.: Logic of the yeast metabolic cycle: temporal compartmentalization of cellular processes, *Science*, Vol. 310, No. 5751, pp. 1152–1158 (2005).
- [21] Santos, A., Wernersson, R. and Jensen, L. J.: Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes, *Nucleic Acids Res.*, Vol. 43, No. Database issue, pp. D1140–1144 (2015).