半教師あり学習を用いた遺伝子発現データの解析

景山彩璃　(指導教員：瀬々潤)

1　はじめに

近年の技術進歩により，細胞で利用されている遺伝子の量を網羅的に観測するマイクロアレイが広く利用されるようになった．マイクロアレイで得られる遺伝子発現量情報は，観測遺伝子数分の次元，つまり 1 万を超す次元のデータであり，状況把握が容易ではない．そこでこのデータを俯瞰する方法として主成分分析 (PCA) による次元削減手法が用いられている [1]．しかし，PCA は各サンプル (実験) に対して事前知識を仮定しない教師無し手法であり，必ずしも状況に適した次元が削減できないことがある．一方，フィッシャー判別分析 (FDA) を利用した教師有り次元削減法もあるが，遺伝子発現量実験では時系列情報（時間に応じて取得したデータ）や濃度系列情報（濃度に応じて取得したデータ）も多く，この場合クラスを設定することが容易ではない．そこで，本研究では一部のデータにのみ教師ラベルを設定し，残りのデータはラベルを持たない状態で学習する半教師あり手法を遺伝子発現量解析に適用することを提案する．実験では，細胞に対して分化 (異なった種の細胞に変質する事) を誘導する刺激と増殖 (同一の細胞に分裂を繰り返す事) を誘導する刺激を様々な濃度で与え，時間を追って発現量を調査したデータ [2] 解析に適用し，細胞状態の変化が追えることを確認する．

2　方法

本章では，教師無し手法である PCA による次元削減と，教師あり手法である FDA の改良版である局所フィッシャー判別分析による次元削減の間を取る半教師あり次元削減手法である半教師あり局所フィッシャー判別分析法 (SELF) の導入を行い，次章においてこれらの手法を遺伝子発現量データに対して適用する．

2.1　主成分分析 (PCA)

PCA は，最も分散の大きくなる軸から順に削減する次元を求めていく方法である．

サンプル (実験)i の発現量ベクトルを $\boldsymbol{x}_i$ で表す．$\boldsymbol{x}_i$ は遺伝子数分の次元を含むベクトルである．サンプルは n 個あるとする．この n 個のベクトルに対する分散共分散行列を $\boldsymbol{S}^{(t)}$ とすると，

$$\begin{aligned}\boldsymbol{S}^{(t)} &= \sum_{i=1}^{n}(\boldsymbol{x}_i-\boldsymbol{\mu})(\boldsymbol{x}_i-\boldsymbol{\mu})^{\mathsf{T}},\ \text{但し}\ \boldsymbol{\mu}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i \\ &= \frac{1}{2}\sum_{i,j=1}^{n}W_{i,j}^{(t)}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathsf{T}}\end{aligned}$$

但し $\boldsymbol{W}^{(t)}$ は $W_{ij}^{(t)}=1/n$ なる要素の $n\times n$ 行列である．この共分散行列に対し，PCA で射影する次元は，以下の固有値，固有ベクトル問題の固有値が大きい方から優先的に選択する．$\boldsymbol{S}^{(t)}\boldsymbol{\varphi}=\lambda\boldsymbol{\varphi}$ （ここで $\boldsymbol{\varphi}$ は固有ベクトル).

2.2　局所フィッシャー判別分析 (LFDA)

本節ではフィッシャー判別分析の改良手法 LFDA [3] の導入を行う．フィッシャー判別分析と異なるところは，同一クラス内の点を全て均一に見るのではなく，距離が近い点同士に重みを付けて解析する事である．

局所クラス間散布行列 $\boldsymbol{S}^{(lb)}$ 及び局所クラス内散布行列 $\boldsymbol{S}^{(lw)}$ を n' をラベル付きのサンプル数として，以下の通り定義する．

$$\begin{aligned}\boldsymbol{S}^{(lb)} &= \frac{1}{2}\sum_{i,j=1}^{n'}W_{i,j}^{(lb)}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathsf{T}} \\ \boldsymbol{S}^{(lw)} &= \frac{1}{2}\sum_{i,j=1}^{n'}W_{i,j}^{(lw)}(\boldsymbol{x}_i-\boldsymbol{x}_j)(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathsf{T}}\end{aligned}$$

但し $y_i=y_j$の時, $\boldsymbol{W}_{i,j}^{(lb)}=A_{i,j}(1/n'-1/n'_{y_i}), \boldsymbol{W}_{i,j}^{(lw)}=A_{i,j}/n'_{y_i}$．$y_i\neq y_j$の時, $\boldsymbol{W}_{i,j}^{(lb)}=1/n', \boldsymbol{W}_{i,j}^{(lw)}=0$ である．この時，LFDA で射影する次元は次の固有値固有ベクトル問題を解いた解の内，固有値が大きい方から優先的に選択する．$\boldsymbol{S}^{(lb)}\boldsymbol{\varphi}=\lambda\boldsymbol{S}^{(lw)}\boldsymbol{\varphi}$.

2.3　半教師つきフィッシャー分析 (SELF)

PCA と LFDA がいずれも類似の固有値固有ベクトル問題を解いている事に着目し，半教師つき問題として定式化したものが SELF [4] である．SELF では，変数 $\beta\in[0,1]$ を用意し以下の様に正則化局所クラス間散布行列 $\boldsymbol{S}^{(rlb)}$, 正則化局所クラス内散布行列 $\boldsymbol{S}^{(rlw)}$ を定義する．

$$\begin{aligned}\boldsymbol{S}^{(rlb)} &= (1-\beta)\boldsymbol{S}^{(lb)}+\beta\boldsymbol{S}^{(t)} \\ \boldsymbol{S}^{(rlw)} &= (1-\beta)\boldsymbol{S}^{(lw)}+\beta\boldsymbol{I}_d\end{aligned}$$

ここで I_d は d 次元の単位行列を示す．この式を利用し，SELF は以下の固有値固有ベクトル問題を解く．

$$\boldsymbol{S}^{(rlb)}\boldsymbol{\varphi}=\lambda\boldsymbol{S}^{(rlw)}\boldsymbol{\varphi}$$

大きい固有値に対応する固有ベクトルが，射影する次元となる．式より，$\beta=1$ の時 PCA と同等に，$\beta=0$ の時 LFDA と同等になる．

3　実データによる実験

3.1　問題設定と前処理

前章で導入した SELF を用いて，細胞に分化誘導刺激である HRG 刺激及び増殖誘導刺激である EGF 刺激を与えて観測された遺伝子発現量データ [2] の解析を行う．実験は各刺激，濃度 4 種類 (0.1nM, 0.5nM, 1.0nM, 10nM) 及び刺激後の時間 (5, 10, 15, 30, 45, 60, 90 分) の計 28 点について採取を行っているが，EGF 刺激の 10nM，60 分のみ結果が欠けているため，合計 55 サンプルのデータとなっている．各サンプル 10,124 個の遺伝子発現量が含まれている．

ここでは，実験の仮定として HRG と EGF が類似の ErbB 受容体を刺激していることから，濃度が弱く時間の短い 6 実験 (0.1nM 及び 0.5nM における 5 分, 10 分，15 分) では類似した細胞状態にあると考え同一のクラス”1” を与える．また，濃度が高く，時間も経った 1.0nM 及び 10nM における 45 分，60 分，90 分では HRG, EGF の種類によって状態が十分変化していると考え，HRG ではクラス”2” を，EGF ではクラス”3” を与える．併せると，クラス”1” が 12 点，クラ

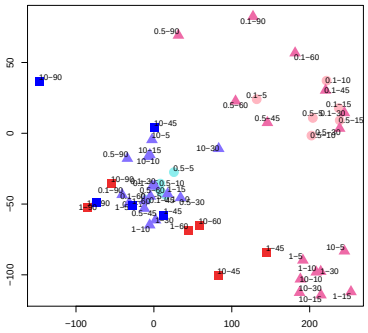

図 1: PCA

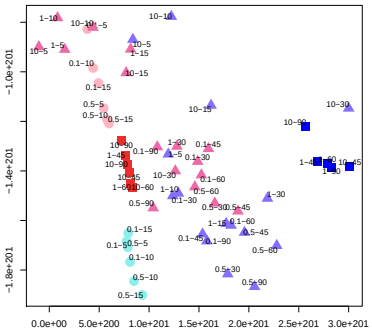

図 2: LFDA

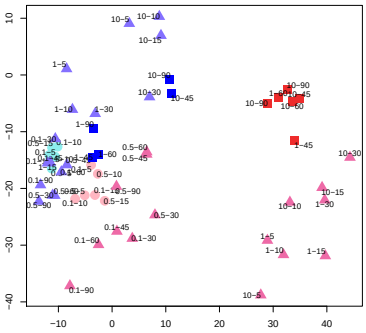

図 3: SELF

ス”2” が 6 点，クラス”3” が 5 点，残りの 32 点がクラスラベル無しのデータとして扱う．

生命に関するデータには常にノイズが入る．遺伝子発現量も同様にノイズが入るため，クラス”1” における発現と”2”，“3” における発現に変動がみられる遺伝子のみを予め選択した．この選択には t 検定を用い，p 値が 0.01 以下と成る遺伝子のみを選択した．結果，435 個の遺伝子が残った．

3.2 実験結果

PCA, LFDA, SELF による結果を図 1 から 3 に示す．赤色が HRG, 青色が EGF による結果であり, 色が濃くなるほど, 試薬の濃度が濃いもしくは刺激から時間が経っている事を示している．点の形はクラスの種類を表し, 丸がクラス “1”，四角が “2” もしくは “3”，三角はクラス無しを示す．点のラベルは (濃度-時間) となっている．また，利用しているデータが高次元の少数サンプルであるため，特に PCA で 1 次元目がノイズを表している事が多い．そこで 2, 3 番目に固有値の大きい値に対応する次元のみを表示した．また，SELF のパラメータ β は，0.99 を用いている．

図 1 では，クラス毎に分布がまとまらず刺激の変化を観測する事が難しい．これは PCA が教師無しの次元削減方法であり，その弱点が現れている．図 2 では, クラス毎に分布はまとまっているが，三角で示したクラス無しの点が HRG と EGF で混ざっており，刺激に対する変動を見ることが難しい．また，クラス”1” の HRG,EGF で分離している事も刺激による反応の差異の比較を難しくしている．図 3 ではクラス”1” の点が全て集約し,EGF は Y 軸の正方向へ，HRG は X 軸の正方向へ伸びている．また，HRG と EGF で濃度や時間に対し異なる動きをしており，HRG 刺激は，時間の経過と濃度上昇に伴って，ほぼ単調に X 軸の大きい方にサンプルが移動していくが，EGF に対しては，刺激の濃度が強いと一旦発現が大きく動くが，時間が経つにつれて濃度が弱い時の値，つまり刺激前の状態に戻っていく動きが観察される．これは，HRG により分化という異なる細胞に変質するのに対し，EGF が増殖という同一の細胞状態に戻る刺激であることに合致した反応である．

3.3 射影軸の解析

SELF で求めた射影軸に関して，関連する遺伝子の抽出を行った．射影軸に対応する固有ベクトルの内，係数の絶対値が大きな遺伝子が，その軸に大きく寄与している．

固有ベクトルの絶対値が大きい方から上位 10 遺伝子を表 1,2 に示す．2 つの表に共通する遺伝子が存在しない．また，表 1 中，MCL1 は分化誘導，SKAP2 は増殖抑制因子として知られており，また細胞骨格に関連した遺伝子が多いため分化に関連する遺伝子が多い．更に表 2 中，EGR1 と NAB2 は協調して働く増殖因子，分化関連因子の ING1, TNFRSF12A, 細胞制御全般に関わる転写因子 JUNB が含まれており，生物学的に信頼性の高い特徴量を抽出できたと考えられる．

表 1: 図 3 の X 軸に寄与する遺伝子

順位 2D	Gene.Symbol	係数
1	ACTR2	1.64
2	PSME4	1.40
3	MCL1	-1.39
4	SKAP2	1.31
5	EGR3	1.15
6	SKAP2	1.14
7	SKAP2	1.05
8	CYR61	1.03
9	ZNF91	0.98
10	PDLIM5	0.95

表 2: 図 3 の Y 軸に寄与する遺伝子

順位 3D	Gene.Symbol	係数
1	HMGCS1	1.46
2	ING1	-1.09
3	EGR1	-1.09
4	NAB2	-1.04
5	—	0.99
6	ID1	-0.96
7	PTPN12	0.92
8	JUNB	-0.90
9	TNFRSF12A	-0.89
10	MAP3K8	0.87

4 考察と今後の課題

次元削減方法として PCA，LFDA，SELF を行った結果，特に SELF では刺激 HRG,EGF の実験状況における細胞状態の変化を可視化でき, それぞれの細胞に対する応答も分離する事が出来た．また, 高次元の遺伝子発現量データの中から分化，増殖に関与する遺伝子を抽出出来た．今回の実験では t 検定により遺伝子数を絞ったが，情報の損失に繋がっている恐れもあり，より多くの遺伝子を利用した次元削減を行う事で精度の高い解析を目指したい．

参考文献

[1] Principal component analysis for clustering gene expression data. K. Y. Yeung and W. L. Ruzzo. *Bioinformatics*, pp. 763–774, vol. 17, no 9, 2001.

[2] Quantitative transcriptional control of ErbB receptor signaling undergoes graded to biphasic response for cell differentiation. T. Nagashima, *et al. Journal of Biological Chemistry*, pp. 4045-4056, vol. 282, no. 6, 2007.

[3] Dimensionality reduction of multimodal labeled data by local Fisher discriminant analysis. M. Sugiyama. *Journal of Machine Learning Research*, vol.8 (May), pp.1027-1061, 2007.

[4] Semi-supervised local fisher discriminant analysis for dimensionality reduction. M. Sugiyama, T. Ide, S. Nakajima, and J. Sese. *Machine Learning*, vol. 78, no. 1-2, pp. 35-62, 2010.