# マイクロアレイを用いた糖尿病患者の遺伝子データの解析

リスク工学グループ演習 6 班
居城 秀明　鈴木 昭平　松原 史浩
201120659　201120626　201120639

佐藤 (イリチュ) 美佳
アドバイザー教員

2011 年 9 月 30 日

## 1　はじめに

現代の生活習慣病と言われる糖尿病は合併症を起こす可能性が問題視されており，早急な対策が要求されている．糖尿病対策の中で特に「笑い」によって糖尿病の症状である血糖値の上昇を抑えるような結果を得た実験を取り上げ，その結果と遺伝子との関わりについて，バイオインフォマティクスの技術を用いこれを解析することで確かめる．

本稿は糖尿病の概要及び「笑い」と糖尿病との関わりについての実験について述べたのち，遺伝子発現データの概要及び解析手法，結果を述べ考察し，最後にむすびを述べる．

## 2　糖尿病進展を抑制する「笑い」と遺伝子発現

この章では生活習慣病として広く知られる糖尿病についてその概要及び「笑い」によりその進展を抑制するような結果を得られた実験について詳説する．

### 2.1　糖尿病とその影響

糖尿病には幾つかの種類があり，インスリンを生成するための膵臓の $\beta$ 細胞が破壊され発症する (1 型糖尿病) 場合や，遺伝子の異常や他の病気が原因となって発症することもあるが，大多数が 2 型糖尿病とよばれる，過食や運動不足など生活習慣が原因となっている．

糖尿病が及ぼす影響の 1 つとして合併症が挙げられる．中でも糖尿病神経障害，糖尿病網膜症，糖尿病腎症は 3 大合併症と呼ばれる．これらの症状は日常生活に大きく支障をきたすものである．

### 2.2　「笑い」と血糖値変化に関する実験

2 型糖尿病患者 25 名を対象とした 2 日にわたる実験が行われた．1 日目の昼食後，医学部の助教授による講義を行い，2 日目の昼食後は漫才の鑑賞を行い，それぞれ食前と講義後，食前と漫才鑑賞後との血糖値の差を調べたところ，1 日目は平均 123mg 上昇したのに対し 2 日目は 77mg の上昇にとどまった．

講義と漫才による「笑い」とで糖尿病患者の血糖値の上昇に変化が見られた実験であるといえる．[2]

### 2.3　遺伝子と血糖値変化の関連性

その後遺伝子の発現解析を行い，「笑い」によってスイッチの状態がオンになる遺伝子の存在が明らかとなった．その中には，糖尿病の合併症を防ぐ重要な遺伝子が含まれていることがわかった．

## 3　研究目的および解析対象の概要

本研究の目的及び対象とするデータについて記す．

### 3.1　研究目的

本研究の目的は，糖尿病患者に「笑い」を与えた場合とそうでない場合に検出した遺伝子発現の計測データ (マイクロアレイデータ) を解析する．この時，糖尿病患者への「笑い」への影響がどの遺伝子によって識別できるかについて調査を行うものである．

また，マーカー遺伝子を同定し，識別関数を推定を行い，この結果から糖尿病患者の「笑い」による治療・効果を知るにあたり必要とされる遺伝子のデータの特定，効果の有無を判定する．

これらを行うことにより「笑い」によって変化した血糖値が，どの遺伝子の変化によるものなのかについて明らかとなり，糖尿病に対する新たな治療方法としての可能性が期待できる．

### 3.2 解析対象データについて

解析の対象データとして NBCI (National Center for Biotechnology Information) のマイクロアレイによる糖尿病患者の遺伝子発現データを用いた [4]．以下にデータに関する用語説明及びデータについての説明を記す．

#### 3.2.1 遺伝子発現

遺伝子発現 (単に発現ともいう) は，遺伝子の情報が細胞における構造および機能に変換される過程をいう．つまり遺伝情報に基づいてタンパク質が合成されることを指す．遺伝情報に基づいて生成されたタンパク質は体内に作用し様々な効果を及ぼす．

ヒトに何らかの変化が生じたとき，その前後に遺伝子発現の変化が見受けられれば，そのような変化を及ぼす遺伝子の存在を確認することができる．

#### 3.2.2 マイクロアレイ

細胞内の遺伝子発現量を測定するために, 多数の DNA 断片をプラスチックやガラス等の基板上に高密度に配置した分析器具で，DNA・RNA を 検出することができる．今回のデータでは, 食前，及び食後の血糖値の差の比較を反映させるため二色法という方法を用いている．これは，遺伝子発現の変化を調べる際，細胞と対照条件の細胞に対し，それぞれ異なる蛍光色素 (Cy3, Cy5) で標識し，1 枚のガラス板上のプローブに競合的にハイブリダイゼーションを行う方法である．

マイクロアレイの出力結果は蛍光色素の信号強度となり，数値が高いほど発現量が多く見られることを表す．

#### 3.2.3 マイクロアレイによる遺伝子発現データ

解析対象データの一例を以下の表に記す

表 1　遺伝子発現データ (例)

| data table_1 | | | | |
|---|---|---|---|---|
| ID_REF | VALUE | CH1_MEAN | CH2_MEAN | PRE_VALUE |
| 1 | -0.76 | 18.5 | 31.4 | 0.59 |
| 2 | | | | |
| ⋮ | | | | |

各々の列の示す意味は次のとおりである．

ID_REF　遺伝子の番号.

VALUE　PRE_VALUE に対数をとって正規化したもの.

CH1_MEAN　食前の遺伝子発現量

CH2_MEAN　講義，または漫才鑑賞後の遺伝子発現量

PRE_VALUE　CH1_MEAN，および CH2_MEAN の比

これが被験者 1 人分のデータである．本研究では被験者 21 人のデータを用いるものとする．

### 3.3 解析方法

遺伝子発現データに対しクラスタリングを行い，退屈な講義を聞いた被験者のデータと漫才を鑑賞した被験者のデータとの分解を試み，その結果について考察を行うものとする．具体的な解析手法については次章で述べる．

## 4 解析手法

遺伝子データを解析ソフトである R によって処理を行った．使用した分析手法はヒートマップ，階層的クラスタリング，非階層的クラスタリング，サポートベクターマシン (SVM) となっている．以下にそれぞれの分析手法を簡単に説明する．

### 4.1 階層的クラスタリング

階層的クラスター分析とは，個体間の類似度あるいは非類似度 (距離) に基づいて，最も似ている個体から順次に集めてクラスターを作っていく方法である．クラスターが作られていく様子を樹形図で示すことができる．以下にその手順を示す.

アルゴリズム

1. 距離 (あるいは類似度) を求める方法を選択し，個体間の距離 (類似度) を計算する.
2. クラスター分析の方法（最近隣法，最遠隣法など）を選択する.
3. 最もこの距離の近い二つのクラスタを逐次的に併合する.
4. この併合を，全ての対象が一つのクラスタに併合されるまで繰り返す.

### 4.2 ヒートマップ

個体と変数それぞれに対し階層的クラスタリングを適用し，類似度を色の濃淡で表現している．個体と変数を同時に可視化できるため，全体像をつかみやすい.

### 4.3 非階層的クラスタリング

階層的クラスター法は，個体数が多いと計算量が膨大になり，大量のデータ解析には向いていない．大規模のデータセットのクラスター分析には，非階層的クラスター法が多用されている．非階層的クラスター法の代表的な方法として $k$ 平均 ($k$-means) 法がある．以下にその手順を示す.

アルゴリズム

1. $k$ 個の初期クラスターの中心を与える．
2. すべてのデータと $k$ 個のクラスター中心との距離を求め最も近いクラスターに分類する．
3. 新たに形成されたクラスターの中心を求める．
4. クラスターの中心がすべて前の段階の結果と同じになるあるいは事前に指定している．繰り返しの回数に達するまで 2，3 を繰り返す．

### 4.4 SVM(Support Vector Machine) による識別

SVM はその識別精度が比較的良いという性質がある。しかし，計算量が多く速度が遅いという特徴がある．SVM の精度が良いのは，① 線形分離不可能な場合はカーネルを使って高次元で線形分離，② マージン最大になるように線形分離を行っているからである．

まず，① カーネルを用いた高次元での線形分離の仕組みについて説明する．例えば，図 1 に示す 2 次元平面座標系 $(x, y)$ 上の 4 つの点 $(1, 1)$，$(1, -1)$，$(-1, -1)$，$(-1, 1)$ を考える．仮に点 $(1, 1)$，$(-1, 1)$ がひとつのクラス，点 $(1, -1)$，$(-1, -1)$ がひとつのクラスであるとする．このとき，直線 $y = 0$ で 2 つのクラスを分けることができる．

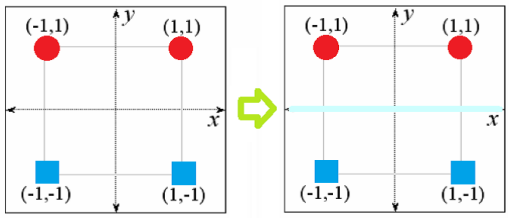


図 1　二次元の点の線形分離

しかし，図 2 に示すように仮に点 $(1, 1)$，$(-1, -1)$ がひとつのクラス，点 $(-1, 1)$，$(1, -1)$ がひとつのクラスであるとすると，平面上でクラスの境界線を一本の直線で引くことはできない．そこで，新しい変数 $z = xy$ を導入し，2 次元平面 $(x, y)$ 上の 4 つの点を 3 次元空間 $(x, y, z)$ に写像すると $(1, 1, 1)$，$(1, -1, -1)$，$(-1, -1, 1)$，$(-1, 1, -1)$ となり，両クラスは平面 $z = 0$ で切り分けることができる．これが，カーネルを用いた高次元での線形分離の仕組みである．

次は，② マージン最大になるように線形分離を行う仕組みについて説明する．例えば，図 3 に示す 2 次元の特徴空間に以下のような 2 つのクラス A と B に属するデータがあったとする．

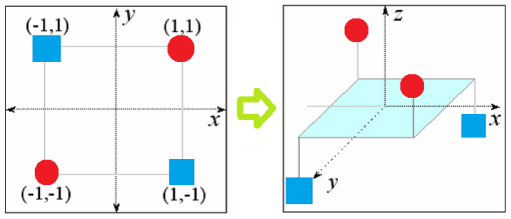


図 2　高次元での線形分離

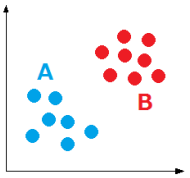


図 3　A,B に属するデータ

これをうまく分離したいと思ったときは，大半の人は図 4 のように線を引くはずである．

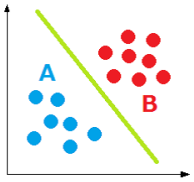


図 4　クラス分離例 1

では，なぜ図 5 のように線を引かないのかについて考えてみると，

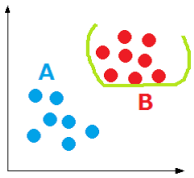


図 5　クラス分離例 2

このような引き方をすればクラス A のデータの周りの領域がクラス B のデータの周りの領域に対し大きくなりすぎるためである．そこで SVM では，「マージン最大化」と呼ばれる手法でこの識別境界を決定している．学習データの中で最も他クラスと近い位置にいるもの (これをサポートベクトルと呼ぶ) を基準とし，そのユークリッド距離 (図 6 の黒い矢印の長さ) が最も大きくなるような位置に識別境界を設定する．つまり，クラスの最端から他クラスまでのマージンを最大にするようにするのだ．

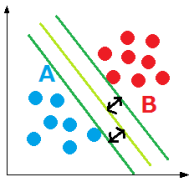


図 6　SVM によるクラス分離

これが，マージン最大になるように線形分離を行う仕組みである．

## 5　解析結果

遺伝子発現データに対し，実際に行った操作について以下に記す．

### 5.1　データのクリーンアップ

遺伝子発現データの各項目 (VALUE, CH1_MEANS など) ごとに被験者をまとめたデータ配列を作成した．また，対象データには数値のない項が含まれていたのでクラスタリング時に適切な計算ができるよう処理を施した．

### 5.2　SVM を用いた遺伝子配列の識別

CH1_MEAN, CH2_MEAN, PRE_VALUE, VALUE の 4 項目に対してそれぞれ SVM 分析を行った．CH1 のデータを図 7 に示す．1 列目の前半が漫才を鑑賞した糖尿病患者 14 名のデータ，後半が退屈な講義を聞いた糖尿病患者 7 名のデータである．2 列目以降はそれぞれの患者の遺伝子情報 (約 4 万個) である．

| | A | B | C | D | E | F | G | H | I |
|---|---|---|---|---|---|---|---|---|---|
| 1 | type | ID307 | ID514 | ID1204 | ID1557 | ID1608 | ID2354 | ID4150 | ID4249 |
| 2 | laugh | 35898.45 | 38738.25 | 56732.45 | 65098.95 | 64637.65 | 65128.55 | 30450.02 | 50106.55 |
| 3 | laugh | 55225.69 | 60089.87 | 74117 | 100940.7 | 97533.51 | 113297.2 | 38740.9 | 58470.1 |
| 4 | laugh | 40501.23 | 53042.41 | 83849.24 | 92539.66 | 95016.08 | 99486.83 | 53536.5 | 43849.63 |
| 5 | laugh | 54040.01 | 63958.15 | 77363.22 | 75111.65 | 85971.09 | 78474.93 | 67578.94 | 67279.35 |
| 6 | laugh | 40427.89 | 43535.18 | 65058.75 | 60084.13 | 72432.76 | 74822.95 | 42927.88 | 44171.44 |
| 7 | laugh | 48737.76 | 57619.69 | 66761.56 | 77054.41 | 79612.87 | 90784.63 | 52114.21 | 38278.09 |
| 8 | laugh | 50373.57 | 52893.49 | 53705.29 | 65132.69 | 71942.41 | 79192.65 | 56226.97 | 45932.04 |
| 9 | laugh | 45497.44 | 61727.52 | 81609.46 | 88670.25 | 95084.85 | 105847.4 | 56132.46 | 41973.34 |
| 10 | laugh | 61119.78 | 79958.59 | 6472.01 | 65119.2 | 65119.4 | 65119.6 | 53298.72 | 70633.86 |
| 11 | laugh | 81667.15 | 87547.88 | 93742.39 | 97169.97 | 97703.35 | 102618.1 | 66104.95 | 63814.41 |
| 12 | laugh | 86919.58 | 103152.3 | 11769 | 121313.1 | 128848.6 | 129234.2 | 100363.3 | 67272.26 |
| 13 | laugh | 65165.56 | 76746.42 | 87884.88 | 104722.3 | 113483.5 | 118331.2 | 71671.52 | 50270.27 |
| 14 | laugh | 119085.2 | 98328.11 | 116088.4 | 118948.6 | 128923.1 | 129644.6 | 86646.57 | 82989.07 |
| 15 | laugh | 50853.71 | 61313.21 | 65119.21 | 65143.81 | 65143.91 | 65143.91 | 59196.41 | 58888.71 |
| 16 | boring | 69388.81 | 81010.16 | 92689.28 | 94936.16 | 104101.7 | 112918.9 | 82926.13 | 57929.57 |
| 17 | boring | 56912.77 | 64816.37 | 61121.47 | 69010.72 | 70790.86 | 74906.39 | 58387.67 | 61161.17 |
| 18 | boring | 66321.16 | 69315.96 | 72095.48 | 83630.93 | 87350.45 | 99103.05 | 75786.75 | 61394.93 |
| 19 | boring | 49330.28 | 68971.34 | 78895.57 | 88979.4 | 95051.51 | 115806.4 | 66373.08 | 45720.99 |
| 20 | boring | 58429.55 | 65130.95 | 65131.05 | 68014.87 | 72431.83 | 82986.81 | 66248.71 | 66993.1 |
| 21 | boring | 58397.49 | 58896.4 | 72002.4 | 79450.46 | 82665.56 | 96624.49 | 63479.76 | 47810.69 |
| 22 | boring | 67618.42 | 67676.47 | 80944.45 | 102062.5 | 122844.5 | 165340.3 | 56650.58 | 44702.11 |

図 7　クリーンアップ後の各患者の CH1_MEAN のデータ配列の一部

### 5.3　階層的クラスタリングによるヒートマップ

取得した遺伝子データに対し，ヒートマップを適用し結果を出力した．　これによって遺伝子データ全体の見通しを立てることを試みた．

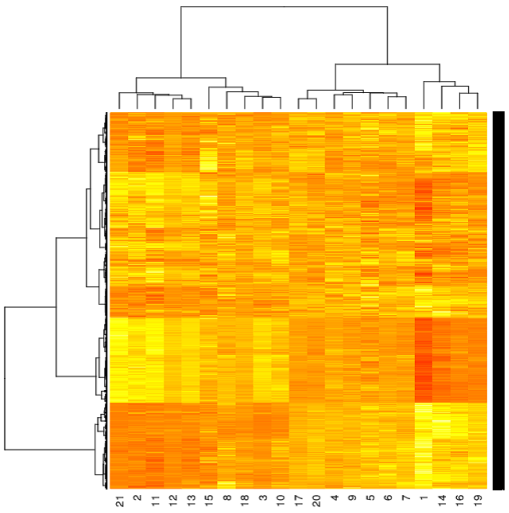


図 8　ヒートマップの結果

図 8 の結果から，個体に関して笑った個体と，笑っていない個体にうまくわけることができなかった．また，変数に対しての結果からは，変数の数が 2 万と膨大なため潰れてしまい，うまく特徴抽出ができなかった．しかし，個体の分類構造が明確に抽出できている．この事に着目して，分類された個体ごとにヒートマップを出力した．図 8 において二つに分類されたサンプルの一方を用いて出力した結果を図 9 に示す．

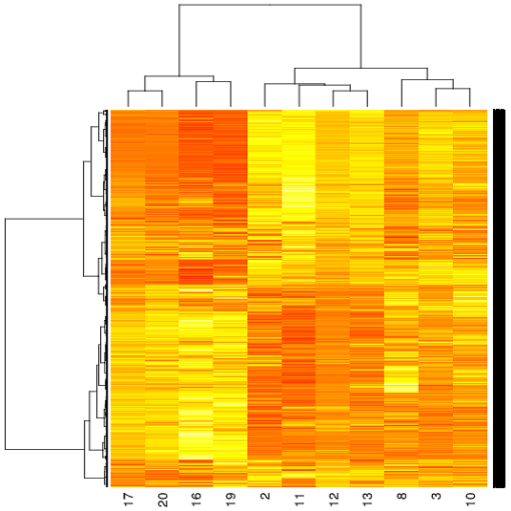


図 9　特定の個体を省いたヒートマップの結果

図 9 の結果から，笑った個体とそうでない個体の 2 つにうまく分割することができていることがわかる．このように望ましい分割を行うことによって，特徴抽出を行えることが分かった．

### 5.4　非階層クラスタリングによる遺伝子配列の分割

階層的クラスタリングでは変数の様子が考察できず，視覚的にもわからなかったため，変数に対して非階層的

クラスタリングを適用してみる．分割数を 4 とし出力した結果を以下に示す．

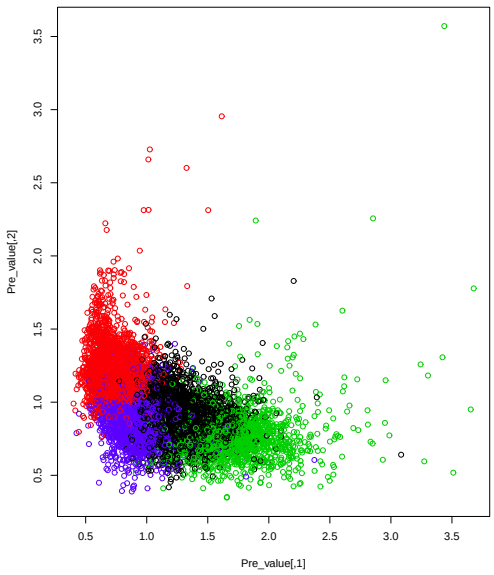


図 10　非階層的クラスタリングの結果

この $k$-means による非階層クラスタリングを CH1_MEAN, CH2_MEAN, PRE_VALUE, VALUE の 4 項目それぞれ適用して各々 4 つのクラスタにデータを分割し，計 16 個のデータとなった．

このように分割したデータに対し各々 SVM 分析を行った．以下, 分解した要素を $\mathrm{CH1\_MEAN}_i (i = 1, 2, 3, 4)$ などと記述する．

### 5.5　分割した各クラスタに対しての SVM 分析

分割したクラスタを区別するために各クラスタに名前を付けた．VALUE データを分割した際のクラスタを $\mathrm{VALUE}_1$ とし，同様にデータの種類とクラスターラベルを対応させている．

図 10 のクラスタリング結果を基にして, $\mathrm{VALUE}_1$ に SVM 分析を適用した結果を以下の表 2 から表 5 までに示す．まず，$\mathrm{VALUE}_1$ から $\mathrm{VALUE}_4$ までのそれぞれに対し，各々のデータすべてを学習用データとして評価し，次に各々のデータすべてをテスト用データとして評価した．

表 2　$\mathrm{VALUE}_1$ の SVM 分析適応結果

| pred | boring | laugh |
|---|---|---|
| boring | 7 | 0 |
| laugh | 0 | 14 |

表 3　$\mathrm{VALUE}_2$ の SVM 分析適応結果

| pred | boring | laugh |
|---|---|---|
| boring | 7 | 0 |
| laugh | 0 | 14 |

表 4　$\mathrm{VALUE}_3$ の SVM 分析適応結果

| pred | boring | laugh |
|---|---|---|
| boring | 4 | 0 |
| laugh | 3 | 14 |

表 5　$\mathrm{VALUE}_4$ の SVM 分析適応結果

| pred | boring | laugh |
|---|---|---|
| boring | 6 | 0 |
| laugh | 1 | 14 |

この表の列成分は予め与えたラベル，行成分は SVM による分離結果である．例えば表の 2 行 1 列成分は退屈な講義を受けたとされる被験者の遺伝子配列が，漫才を鑑賞したとされるグループに含まれると判断されてしまった被験者数である．分解結果からクラスタによっては誤分解を起こしているがほぼ正確に識別されているといえる．

また，以下の図 11 に SVM 出力結果図を示す．退屈な講義を聞いた糖尿病患者を黒色で，漫才を鑑賞した糖尿病患者を赤色で表している．

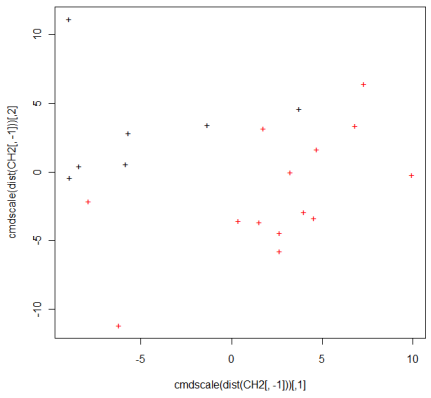


図 11　SVM 出力結果 (2 次元写像)

## 6　考察

階層的クラスタリングを用いたヒートマップの出力によって遺伝子とサンプルの全体像を見ることができた．

また，サンプルに対する階層的クラスタリングに関して特定のサンプルを省くことで漫才の講演を聞いたサンプルと講義を聞いたサンプルとを分割することができた．これは，すべてのサンプルで笑うという生理的な要素が数値として明確に表れる保証はないため，うまく分割されるような処理を行うことで特徴抽出をしていることを意味する．

ヒートマップからは分割した各々の被験者に対して何らかの構造が見られ，これは笑いによって遺伝子が発現したものである可能性が考えられ．マーカー遺伝子の範囲をある程度特定する手がかりとなりうる．

また今回 SVM による全領域解析を行うことが出来なかったため，前処理として非階層的クラスタリングにより分割し特定の領域に SVM をかけることを行ったが，これによりうまく判別できたという結果が得られた．この結果から「笑い」が遺伝子発言に対し何らかの関与をしているのではないかという可能性が伺える．

また，今回は被験者が 21 人とサンプル数が少ないため，全てを訓練データ，テストデータとして SVM を行っている．高次元少標本データの分析は一般に “ 次元の呪い ” により，通常の多変量解析による分析が出来ないが，今回，本グループ演習で実施した解析法により適切な識別が可能となった．

## 7 今後の展望

被験者が 21 人とサンプル数が少ないため，より多くのサンプルに対して実行し本提案法の妥当性を検証する必要がある．

今回，使用したデータの遺伝子の分類では，４つのクラスターに分類し，明確なクラスターを得ることが出来た．この結果については，同じクラスターに分類された遺伝子の性質等の専門的な知識により，より一層の詳細な特徴を見出せる可能性があると考えられる．また，分類構造が明らかとなった被験者について，何らかの要因を見出せる可能性が考えられる．さらに，「笑い」のような生理的な現象には個人差が考えられるため，この個人差をどのように適切に分析に反映させるかについては議論の余地がある．

## 8 むすび

本研究では「笑い」と遺伝子発現の関連性に関する既往研究を基に，遺伝子発現データに対し解析を行った．ヒートマップによって遺伝子とサンプルの全体像をつかむことができ，サンプルに明確な分類構造があることがわかった．そのため，特定のサンプルを抽出することで，識別子の同定を行えることがわかった．

また，高次元小標本データの問題を解決するため，次元縮約を図り，非階層クラスタリングによって遺伝子（次元）を分類することで，次元を縮約した．さらに，縮約した遺伝子群に対するサンプルに SVM を適用することで，識別する方法を提案した．SVM によって明確にサンプルの判別が行われる結果が得られ，糖尿病患者治療に対する笑いの効用を示唆する結果が得られた．また，本研究で得られた識別関数と遺伝子群により，識別に効用のある遺伝子を抽出し，かつ，識別関数を同定することに成功した．今後，同種のデータに適用することにより，本研究で提案した方法の妥当性を検証することが必要と考えられる．

## 参考文献


[1] Trupti Joshi, Jinrong Wan, Curis J. Palm, Kara Juneau, Ron Davis, Audrey Southwich, Katrina M. Ramonell, Gary Stacey, Dong Xu, Jiexun Li, Hua Su, Hsinchun Chen, Huiqing Liu, Limsoon Wong, Ying Xu “KNOWLEDGE DISCOVERY IN BIOINFORMATICS”, WILLEY-INTERSCIENCE, p.p.57-111, 2007

[2] Takashi Hayashi, Osamu Urayama, Miyo Hori, Shigeko Sakamoto, Uddin Mohammad Nasir, Shizuko Iwanaga, Keiko Hayashi, Fumiaki Suzuki, Koichi Kawai, Kazuo Murakami, “Laughter modulates prorenin receptor gene expression in patients with type 2 diabetes”, Jounal of Statistical Software, Volume 62, Issue 6, June 2007, p.p.703-706

[3] Alexandros Karatzoglou, David Mayer, Kurt Hornik, “Support Vector Machines in ***R***”, Jounal of Statistical Software, Volume 15, Issue 9, April 2006, p.p.1-28
`http://www.jstatsoft.org/v15/i09/paper`

[4] NCBI, マイクロアレイによる糖尿病患者の遺伝子発現データ
`URL:http://www.ncbi.nlm.nih.gov/`