Vol.2015-MPS-103 No.23
Vol.2015-BIO-42 No.23
2015/6/23

# COG 分類を考慮した 遺伝子発現プロファイルデータのクラスタリング手法

大古田みのり[†1]　石川由羽[†1]　高田雅美[†1]　佐伯和彦[†1]　城和貴[†1]

遺伝子はタンパク質の設計図であり，生成されるタンパク質の機能によって遺伝子を分類することが可能である．遺伝子の機能を明らかにすることは生物学的実験の短縮や，医療，農業の発展に役立つ．機械学習を用いてデータをクラスタリングすることで，機能未知の遺伝子の機能を推定できると考えられる．本稿では，機能が明らかになっている遺伝子発現プロファイルデータを学習データとし，正しく学習するための手法を提案する．提案手法は，学習データのクラスを複数のグループに分割することによって精度の向上を図っている．クラスタリング手法は SVM を用いる．遺伝子発現プロファイルデータは，土壌微生物であるミヤコグサ根粒菌の時系列データを用いる．また，データのクラスにはかずさ DNA 研究所が提供する遺伝子の機能データベースである COG の分類を用いる．

キーワード：遺伝子発現プロファイルデータ，ミヤコグサ根粒菌，COG，SVM

# A Clustering Method for Gene Expression Data Based or COG Classification

MINORI OHGODA[†1]　YU ISHIKAWA[†1]　MASAMI TAKATA[†1]
KAZUHIKO SAEKI[†1]　KAZUKI JOE[†1]

Genes are a blueprint of the protein. Gene can be classified by using the function of the protein. It helps that shorting if the biological experiment time, medical care and agricultural development to clarify the function of genes. In this study, we propose a learning method for correct machine learning that treats gene expression profile data as learning data. The proposed method improves the accuracy by divides classes of learning data to multiple. It is considered that the function of unknown genes can be estimated with machine learning. We adopt SVM to clustering method. We used time series gene expression profile data of Mesorhizobium loti which is a soil microbe and database called COG which Kazusa DNA Res. Inst. provides as classes of data.

## 1. はじめに

近年，ヒトをはじめとする様々な生物種で展開されたゲノムプロジェクトの成果により，多くのゲノムの DNA 塩基配列が解読されている．このような遺伝子のデータの整備や蓄積にともない，大量の遺伝子のデータが計算機を用いて解析されつつある．この試みの中には，個々の遺伝子の機能解析[1]から，複数の遺伝子の相互関係の解析[2]，患者の疾患の有無の判別[3]など様々なものがある．

遺伝子の機能や相互関係を解明するために用いられるデータとして，遺伝子発現プロファイルデータが挙げられる[1][2][3]．遺伝子発現プロファイルデータとは，生物の細胞を時系列で観測し，その細胞内の遺伝子の発現量を計測した行列データである．このデータは，主にバイオインフォマティクスの分野で利用され，ある現象に関わる遺伝子群の候補を絞り込みたい場合や，遺伝子の発現量から特定の組織の挙動を調べたい場合，疾病を発見したい場合に利用される．計算機を用いて遺伝子の機能や相互関係を推定することは，実際に生物学的実験を行って推定するときの指標として役立つ．このことより，生物学的実験の所要時間短縮に繋がると考えられる．また，計算機を用いることで生物学者とは違う視点で解析が行え，新たな機能を発見できる可能性もある．

遺伝子発現プロファイルデータを用いて，遺伝子の機能や複数の遺伝子の相互関係を推定するために様々な手法が提案されている．機能未知な遺伝子の機能を推定する手法として，生物学的な実験により既に機能が明らかとなっている遺伝子の遺伝子発現プロファイルデータを用いる手法がある[1]．機能未知遺伝子とは，遺伝子の機能が未だ明らかになっていない遺伝子のことを指す．遺伝子の中には 2 つ以上の機能を持つものもあり，複数の機能のうちの 1 つしか明らかになっていないものも存在する．そのため，この手法は既に機能が明らかとなっている遺伝子に対しても，新しい視点の解析となり有効であると考えられる．また，複数の遺伝子の相互関係を推定するための手法として，数理モデルを用いるものがある[2]．しかし，これらの研究の問題点は，対象となるデータに対してどの手法が適しているのかは明らかになっておらず，対象のデータに特定の手法が適している，と仮定して推定が行われている点である．つまり，既存研究の手法が各自の解析したいデータに向いている確証がないという問題がある．そのため，例えば機

†1 奈良女子大学
Nara Women's University

Vol.2015-MPS-103 No.23
Vol.2015-BIO-42 No.23
2015/6/23

械学習を用いて遺伝子発現プロファイルデータを解析する場合,使用するデータを正しく学習する手法が必要となる．データを正しく学習することで，機能未知の遺伝子の機能を推定できると考えられる．加えて，異なる条件で計測されたデータの学習結果を比較することで，複数の機能を持つ遺伝子を発見し，機能を推定できる可能性がある．

本稿では 6 つの条件下で計測されたミヤコグサ根粒菌の遺伝子発現プロファイルデータに機械学習を適用し，クラスタリングを行うための学習方法を提案する．ミヤコグサ根粒菌は，マメ科の植物であるミヤコグサと共生関係にあり，窒素固定産物を生成する．ミヤコグサ根粒菌の持つ共生窒素固定に関わる遺伝子はまだ完全に解明されていない．遺伝子の機能を解明することで，共生窒素固定のメカニズムがさらに明らかになり，他の植物でも共生窒素固定が可能になれば，農業の生産性の向上や，環境問題の解決に大いに役立つ．本稿で扱うミヤコグサ根粒菌の遺伝子の機能については COG（Clusters of Orthologous Groups）[4][5]分類を用いるものとする. COG とは遺伝子の機能データベースである．遺伝子の機能データベースには COG の他に京都大学が提供する KEGG（Kyoto Encyclopedia of Genes and Genomes）[6]や MultiFun[7]などが存在する．今回，ミヤコグサ根粒菌の遺伝子プロファイルデータの提供元であるかずさ DNA 研究所[8]が公開しているデータベースが COG であるため，本稿では COG 分類を用いる．

2 章では，SVM を用いた遺伝子発現プロファイルデータのクラスタリング手法の既存研究について述べる．3 章ではクラスタリング手法と，本稿で利用する遺伝子発現プロファイルデータ，その遺伝子の機能について説明する．4 章では 3 章の提案手法を用いて条件ごとに実験を行い，比較する．

## 2. 既存研究

SVM を用いて遺伝子データをクラスタリングした既存研究を 2 つ挙げる．

1 つ目の研究[1]では酵母菌の遺伝子発現プロファイルデータに SVM と，カーネル密度推定，Fisher 線形判別分析，C4.5，MOC1 の計 5 種類のアルゴリズムを適用してその正答率を混合行列を用いて比較している. SVM のカーネルは多項式カーネル，ガウシアンカーネルを利用している．クラスタリングした遺伝子発現プロファイルデータのデータ数は 2467 個，データの次元数は条件数の 76，学習データのクラス数は 5 である．クラス分けに用いた遺伝子の機能は，MIPS（Munich Information Center for Protein Sequences Yeast Genome Database）[9]の分類をもとにしている．結果は SVM を適用したものがすべてのクラスにおいて最も良い正答率を出している．カーネルはガウシアンカーネルの正答率が最も高いクラスと，高次元の多項式カーネルの正答率が最も高いクラスの２通りあった．また，この研究では機能未知な遺伝子をテストデータとして扱うことで，その機能の推定も行われている．

表 1　遺伝子発現プロファイルデータの計測条件

| 条件番号 | フラボノイド | 遺伝子の様子 |
|---|---|---|
| 1 | + | 野生型 |
| 2 | - | 野生型 |
| 3 | + | nod 遺伝子なし |
| 4 | - | nod 遺伝子なし |
| 5 | + | tts1 遺伝子なし |
| 6 | - | tts1 遺伝子なし |

2 つ目の研究[3]では，大腸菌の遺伝子発現プロファイルデータに SVM と，提案手法である k-nearest-neighbor アルゴリズムを適用してその正答率を比較している. SVM のカーネルは多項式カーネルを利用している. この研究で SVM を適用したデータは遺伝子発現プロファイルデータだけでなく，各遺伝子間の相関係数など他の情報も合わせることで正答率の向上を実現している．結果は SVM を遺伝子発現プロファイルデータと他のデータを組み合わせた例が最も良い精度を出している．

## 3. 遺伝子発現プロファイルデータのクラスタリング手法

### 3.1 ミヤコグサ根粒菌

根粒菌とはマメ科の植物の根に感染し，大気中の窒素を窒素固定して宿主植物に提供する土壌微生物である．その中でもミヤコグサ根粒菌とは，宿主であるミヤコグサと共生し，窒素固定を行う．2000 年に全ゲノム塩基配列が，かずさ DNA 研究所で解読され，現在は遺伝子の機能の解析が進められている．本稿では，かずさ DNA 研究所から提供を受けた遺伝子発現プロファイルデータを利用する．

遺伝子発現データとは様々な実験条件化における遺伝子発現量を測定した，遺伝子×実験条件で表される行列データ，または実験条件を固定し，遺伝子×単位時間で表される行列データである．遺伝子×単位時間のデータを特に，遺伝子発現プロファイルデータという．遺伝子発現プロファイルデータを含む，遺伝子データのデータベースは様々な機関によって管理，公開されており，研究者が自由に使えるものも多い．本稿のデータは国立生物工学情報センター（National Center for Biotechnology Information, NCBI）[10]の Web ページで取得可能である．

本稿で扱う遺伝子発現プロファイルデータは Nod 因子に関わるデータである．Nod 因子とは根粒菌がマメ科植物と共生を始める際に生成する物質である．データの条件数は 6 であり，次元数は 4 である．次元数は遺伝子発現量を計測した回数と等しく，実験開始時，1 時間後，6 時間後，24

Vol.2015-MPS-103 No.23
Vol.2015-BIO-42 No.23
2015/6/23

表 2　COG 分類

| | クラス名 | 機能 |
|---|---|---|
| 1 | A | Amino acid biosynthesis |
| 2 | B | Biosynthesis of cofactors, prosthetic groups, and carriers |
| 3 | C | Cell envelope |
| 4 | D | Cellular processes |
| 5 | E | Central intermediary metabolism |
| 6 | F | Energy metabolism |
| 7 | G | Fatty acid, phospholipid and sterol metabolism |
| 8 | I | Purines, pyrimidines, nucleosides, and nucleotides |
| 9 | J | Regulatory functions |
| 10 | K | DNA replication, recombination, and repair |
| 11 | L | Transcription |
| 12 | O | Translation |
| 13 | P | Transport and binding proteins |
| 14 | U | Other categories |
| 15 | Z | Hypothetical |

時間後に 1 回ずつ発現量を計測したことを表す．条件についての説明を表 1 にまとめる．

表 1 中の用語について説明する．野生型とは，自然界に最も多く存在する状態の遺伝子を指す．つまり本稿の場合は，計測時に条件を何も与えず，nod 遺伝子と tts1 遺伝子の両方を残した状態で計測したデータのことを指す．nod 遺伝子とは，Nod 因子を合成する遺伝子であり，tts1 遺伝は，nod 遺伝子の制御下にあることが知られている遺伝子である．フラボノイドとは，根粒菌と共生関係にあるミヤコグサが分泌する nod 遺伝子を活性化させる物質である．表中の+はフラボノイドを添加した状態でデータを計測したことを，−は無添加の状態で計測したことを表す．

### 3.2 COG（Clusters of Orthologous Groups）

COG とは，NCBI などが提供している遺伝子の機能分類データベース[4][5]である．各ゲノムのタンパク質コードに記号を割り当てることで，遺伝子を分類している．遺伝子によっては，最上位の分類の下に更に細かい分類が続くものもある．遺伝子の機能データベースには COG の他に KEGG や MultiFun が存在する．ミヤコグサ根粒菌の遺伝子プロファイルデータの提供元であるかずさ DNA 研究所が公開しているデータベースが COG であるため，本稿では COG の提供する分類を用いる．

遺伝子の機能について説明する．生物の細胞中には遺伝子が存在し，その遺伝子はタンパク質の設計図となることが知られている．遺伝子を元に生成されたタンパク質には，代謝に関わるもの，細胞の機能に関わるもの等さまざまな役割を持つものがある．COG はそのタンパク質の機能を遺伝子の機能と読み替え，分類クラスとしている．分類クラスは，調べたい生物の DNA 配列を機能既知の生物の DNA 配列と比較し，類似性が認めれば同じクラスに分類する，という手法で決定される．

本稿では遺伝子発現プロファイルデータの提供元であるかずさ DNA 研究所が提供する RhizoBase[11]内の最上位の機能クラスを利用する．RhizoBase の機能クラスは表 2 のようになっている．クラスは遺伝子の機能ごとに割り振られており，例えば，クラス L は遺伝子の転写に関わる機能，クラス O は遺伝子の翻訳に関わる機能を持った遺伝子群であることを表している．クラス Z は機能未知の遺伝子，クラス U はさまざまな機能を持つ遺伝子を便宜上ひとまとめにしたものである．

### 3.3 クラスタリング手法

本稿ではクラスタリング手法として SVM を用いる．カーネルはガウシアンカーネルを用いる．遺伝子Xの発現量を$\vec{X}$，遺伝子Yの発現量を$\vec{Y}$，αをハイパーパラメータとすると，ガウシアンカーネルKは，

$$\mathrm{K}(X,Y) = exp\left(\frac{\left\|\vec{X}-\vec{Y}\right\|^2}{2\alpha^2}\right)$$

と表せる．既存研究[1]で最も良い結果が得られているためこのカーネルを用いる．また，COG 分類でクラス U と Z に属するデータは解析から省く．これは，クラス Z と U に属する遺伝子は機能が 1 つに絞られていない遺伝子のクラスであり，学習データに向かないと考えられるためである．よって，実際には A，B，C，D，E，F，G，I，J，K，L，O，P の合計 13 クラスを使用する．

## 4. 提案手法

本稿の目的は，既存研究ではうまく分類できない遺伝子発現プロファイルデータに合った学習方法を提案することである．そのために，提案手法では学習データをいくつかのグループに分類して，学習を複数回行うことで学習率を向上させる．

提案手法は 2 つのフェーズで構成される．手法の概略図を図 1 に示す．フェーズ 1 では，13 のクラスからグループを作る．グループは総当たりの組み合わせで作るため，$\sum_{i=1}^{13}{}_{13}C_i = 8178$グループできる(1-1)．次にそのグループごとに SVM を適用し，正答率を求める(1-2)．次に正答率が上位のグループを抽出し降順にソートする(1-3)．ソートが終了するとフェーズ 1 は終了する．

フェーズ 2 では，フェーズ 1 で抽出したグループを基準とし，そこへ新たなグループを付け加えることで最適な組み合わせを決定する．はじめに，フェーズ 1 の抽出グルー

Vol.2015-MPS-103 No.23
Vol.2015-BIO-42 No.23
2015/6/23

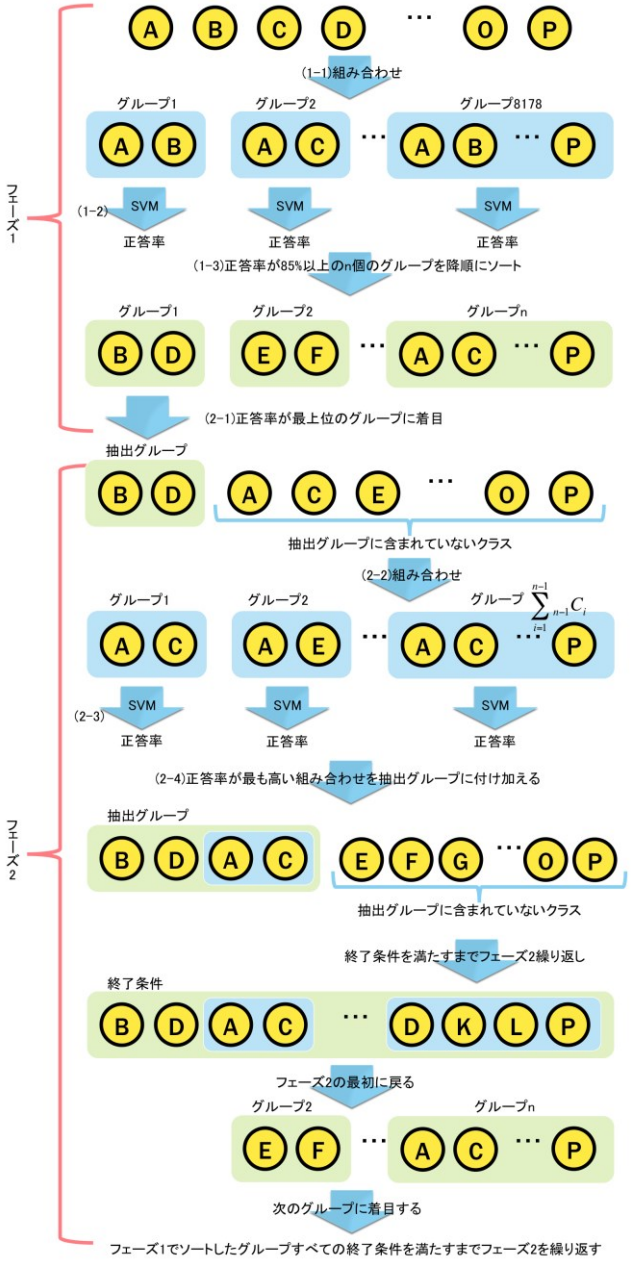


図 1　提案手法の概略図

プの中で最も正答率の高いグループを基準とする(2-1)．次に，13 のクラスの中でこの基準に含まれていないクラスだけを使い，更に組み合わせでグループ 2 を作る(2-2)．例えば，抽出グループの 1 つ目にクラス A と B が含まれていた場合，グループ 2 は残りの 11 クラスから，C と D，D と E のような組み合わせの複数グループから構成される．次にそのグループ 2 を構成する組み合わせごとに SVM を適用し，正答率を求める(2-3)．その中で最も正答率の良かった組み合わせを，はじめの抽出グループに付け加える(2-4)．例えば，先ほどの 11 クラスの中でクラス C と D を組み合わせたグループ 2 の正答率が最も高かった場合は，抽出グループは A と B に，新しい組み合わせである C と D を付け加えることで再構成される．フェーズ 2 の終了条件 1 は，

表 3　予備実験 1 の結果

| データ番号 | 正答率(%) |
|---|---|
| 1 | 26.4 |
| 2 | 26.3 |
| 3 | 25.7 |
| 4 | 26.3 |
| 5 | 26.0 |
| 6 | 25.9 |
| 7 | 26.5 |

抽出グループの構成要素が 13 全てのクラスとなったとき，もしくは 12 となったときである．構成要素が 12 の場合，抽出グループに残りの 1 つのクラスを付け加えることで抽出グループは完成する．次にフェーズ 1 へ戻り，初めにソートした際の，2 組目に正答率の高かったグループを次の抽出グループの基準とする． フェーズ 2 の終了条件 2 は，フェーズ 1 で正答率が閾値以上であったグループそれぞれに対してフェーズ 2 を適用することである．

フェーズ 1，2 を満たすと実行を終了し，データごとに最適な組み合わせを出力して終了する．

## 5. 実験

4 章の提案手法を遺伝子発現プロファイルデータに適用する実験を行う．はじめに予備実験を 2 つ行い，既存手法の精度を確認し，その後予備実験の結果を受けて本実験を行う． 実験は 6 つの条件ごとのデータセットと，6 つの条件を全て合わせたデータセットの合計 7 つのデータを用いた．以降，表 1 の条件番号の 1 から 6 をそれぞれデータ 1 からデータ 6 とする．また，6 つの条件を全て合わせたデータセットをデータ 7 とする．つまりデータ 7 の次元数は 24 となる．データは 7 つとも同じスケールで正規化した．ハイパーパラメータは，コストパラメータを$C = 1$とする．

### 5.1 予備実験 1

予備実験 1 として，既存研究の 1 つ目の手法を適用し，13 クラスのデータに対して交差検定を行う．交差検定で用いるデータの分割数は 3 とする．実験結果はデータごとに 3 回ずつ交差検定を行い，それぞれの正答率の平均を予備実験 1 の正答率とする．結果は表 3 に示すとおりである．

表 3 より，交差検定の結果は 7 つの正答率の平均で 26.2%となり，非常に正答率が低くなることを確認した．このことにより，本稿で使用する遺伝子発現プロファイルデータには，既存研究の手法は適切でないことが明らかになった．

予備実験 1 により，13 のクラスに対して同時に SVM を適用すると正答率が低くなることが分かった．その原因を探るため，さらに予備実験 2 を行う．

Vol.2015-MPS-103 No.23
Vol.2015-BIO-42 No.23
2015/6/23

### 5.2 予備実験 2

予備実験 2 として，学習データとテストデータを同様のものにして SVM を実行する．この実験も予備実験 1 と同様，データごとに 3 回ずつ実験を行い，それぞれの正答率の平均を予備実験 2 の結果とした. この実験を行うことで，13 クラスを同時にクラスタリングした際の学習率を調べることができる．この実験においても学習率が低ければ，学習の方法が使用したデータに適当でないことが明らかとなる．予備実験 2 では，SVM との比較として同じ教師あり学習のニューラルネットワークをデータ 7 に適用した結果も示す．この実験結果は表 4 に示す．また，13 クラスのデータがそれぞれどのクラスにいくつ属すると判別されたかについて，データ 1 の判別結果を表 5 に示す．表 5 は行が学習データ，列がテストデータを表す．例えば 1 列目について説明すると，13 クラスを同時に学習した後，クラス A に属するテストデータを判別したところ，クラス A に属する 189 のデータのうち，50 が正しく判別され，残りの 139 のデータは誤って別のクラスである，と判別され，その結果クラス A の正答率は 26.5%であることが示されている．

結果 1 より，13 クラス全てを一度にクラスタリングすると正答率は平均 53.5%となった．予備実験 2 は，学習データとテストデータに同じデータを使用しているので，これは非常に低い正答率であるといえる．このことより，本稿で使用するデータは既存の手法では学習が上手くいかないことが分かる．また，ニューラルネットワークを使った判別の正答率も 39.9%と低いことからも，データの正しい判別には何らかの工夫が必要であることが分かる．したがって，学習の方法が使用したデータに適していないことが明らかとなったため，提案手法を適用する必要があると考えられる．また，データ 7 は 24 の次元数を持ち，他の 6 つのデータの 6 倍の特徴量を持つため，正答率が高くなったと考えられる．

結果 2 より，学習の正答率を向上させるための方針を考察する．ここでは，結果 2 のクラス J の行とクラス P の行に着目する．テストデータの列をそれぞれ見ると，どのクラスでも正しいクラスに判別されたデータ数と同程度，もしくは 2 倍以上のデータがクラス J, P に属するデータである，と誤判別されている．つまり，クラス J，P の存在が全体の正答率の低下を招いている可能性がある．クラス J と P とうまく分離できるクラスを見つけ，学習の早期に学習できれば正答率の向上を見込めると考えられる．

### 5.3 実験

5.1 節と 5.2 節の予備実験の結果を受け，4 章の提案手法を遺伝子発現プロファイルデータに適用する．結果を表 6 に示す．予備実験 2 と同様に学習データとテストデータは同じものを使用した．フェーズ 1 で抽出するグループは本実験では正答率が 85.0%以上のものとする．表 6 は，それぞれのデータごとに, 下段に最適なクラスの組み合わせを，上段にそのクラスを判別したときの正答率を示している．提案手法では抽出グループを決める際，グループの数を固定せず, 正答率が 85.0%以上のものを抽出する. そのため，条件によって抽出グループの数は異なる．抽出グループの数はデータ 1 が 51，データ 2 が 45，データ 3 が 47，データ 4 が 38，データ 5 が 49，データ 6 が 47，データ 4 が 114 であった．表 6 はその中で最も結果の良かったもののみを表す．結果とは，抽出グループ中ではじめに着目した組の正答率である. また, データごとに正答率の平均を示した．

この結果により，データごとに最適なクラスの分類が明らかとなった．例えば，データ 1 ならば，クラス L とクラス P の判別方法を学習した後，クラス I とクラス J の判別方法を学習し, クラス C とクラス F, クラス D とクラス O, クラス G とクラス K，クラス A とクラス B の順番で学習すれば良い，ということが分かる．

表 4　予備実験 2 の結果 1

| データ番号，手法 | 正答率(%) |
|---|---|
| 1 | 52.7 |
| 2 | 52.4 |
| 3 | 52.7 |
| 4 | 50.6 |
| 5 | 52.8 |
| 6 | 51.1 |
| 7 | 61.9 |
| ニューラルネットワーク | 39.9 |

表 5　予備実験 2 の結果 2

| | A | B | C | D | E | F | G | I | J | K | L | O | P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 50 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| B | 1 | 40 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 |
| C | 0 | 0 | 53 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| D | 1 | 1 | 0 | 101 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| E | 0 | 0 | 0 | 0 | 26 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| F | 3 | 7 | 4 | 1 | 1 | 144 | 5 | 1 | 3 | 2 | 2 | 2 | 2 |
| G | 0 | 0 | 1 | 0 | 0 | 0 | 28 | 1 | 0 | 0 | 1 | 0 | 0 |
| I | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 30 | 0 | 0 | 0 | 0 | 0 |
| J | 14 | 13 | 12 | 8 | 15 | 28 | 23 | 7 | 232 | 16 | 4 | 10 | 34 |
| K | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 27 | 0 | 0 | 1 |
| L | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 16 | 0 | 0 |
| O | 1 | 0 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 125 | 0 |
| P | 119 | 98 | 44 | 95 | 78 | 163 | 107 | 43 | 301 | 57 | 29 | 59 | 727 |
| 要素数 | 189 | 160 | 114 | 206 | 121 | 339 | 168 | 82 | 538 | 103 | 54 | 196 | 764 |
| 正答率(%) | 26.5 | 25.0 | 46.5 | 49.0 | 21.5 | 42.5 | 16.7 | 36.6 | 43.1 | 26.2 | 29.6 | 63.8 | 95.2 |

表 6　実験の結果

| | 1 組目 | 2 組目 | 3 組目 | 4 組目 | 5 組目 | 6 組目 | 7 組目 | 平均正答率（%） |
|---|---|---|---|---|---|---|---|---|
| データ 1 | 95.5<br>(L,P) | 93.7<br>(I,J) | 87.0<br>(C,F) | 84.6<br>(D,O) | 81.9<br>(G,K) | 77.9<br>(A,B) | 100.0<br>(E) | 88.7 |
| データ 2 | 95.6<br>(L,P) | 93.2<br>(I,J) | 85.7<br>(C,F) | 85.3<br>(D,O) | 79.7<br>(G,K) | 77.4<br>(A,E) | 100.0<br>(B) | 88.1 |
| データ 3 | 95.5<br>(L,P) | 93.2<br>(I,J) | 86.5<br>(C,F) | 84.3<br>(D,O) | 82.3<br>(A,E) | 77.5<br>(G,K) | 100.0<br>(B) | 88.5 |
| データ 4 | 95.5<br>(L,P) | 92.0<br>(C,J) | 87.7<br>(F,I) | 82.8<br>(D,O) | 79.3<br>(G,K) | 79.0<br>(A,E) | 100.0<br>(B) | 88.1 |
| データ 5 | 95.8<br>(L,P) | 93.2<br>(I,J) | 86.9<br>(C,K) | 85.8<br>(F,O) | 82.0<br>(D,E) | 77.4<br>(A,B) | 100.0<br>(G) | 88.6 |
| データ 6 | 95.7<br>(L,P) | 93.6<br>(I,J) | 87.6<br>(C,F) | 82.2<br>(B,D) | 81.7<br>(E,O) | 77.9<br>(A,G) | 100.0<br>(K) | 88.4 |
| データ 7 | 96.2<br>(L,P) | 94.4<br>(I,J) | 92.1<br>(C,F) | 89.6<br>(D,O) | 85.1<br>(B,G) | 82.5<br>(A,K) | 100.0<br>(E) | 91.4 |

また，データ 7 以外のデータは最適な組み合わせが異なりながら，正答率が平均 88.4%であり，データが変わっても同程度の正答率が期待できることが分かる．データ 7 に関しては，予備実験と同様，他の 6 つのデータよりも特徴量が多いため，正答率が高くなったと考えられる．また，予備実験 2 の正答率と本実験の正答率を比較すると正答率は向上しており，よりデータに最適な学習方法である，と言える．また，学習に失敗した遺伝子を抽出し，アミノ酸配列上，あるいは，ゲノム上の位置を他の機能クラスのものと比較することで，新たな機能を発見できる可能性も得られる．

## 6. まとめ

本稿ではミヤコグサ根粒菌の機能が明らかになっている遺伝子の遺伝子発現プロファイルデータを学習データとし，遺伝子の属するクラスを適切に組み合わせることで，SVM を用いてデータを正しく学習するための手法を提案した．学習の精度を調べるために，学習データとテストデータに同様のデータを採用し，実験を行った．その結果，7 つのミヤコグサ根粒菌のデータに関して既存の手法より学習の精度が向上し，平均 88.4%という正答率を得ることができた．既存手法を用いると正答率は平均で 53.5%であった．このことより，既存の手法と比較して提案手法はよりデータに最適な学習方法であると言える．また，特定の手法を用いて学習を行い，COG 分類とほぼ一致したクラスタリグが可能であることは，手法が生物学的な意味を持つと予想できる．したがって提案手法は生物学的な意味を持つといえる．

今後の展望としては，本手法を用いて学習を行い，テストデータにミヤコグサ根粒菌の機能が不明な遺伝子のデータを使用することで機能を推定する．その際，学習に失敗した遺伝子の情報が活用できると考えらえる．また，データ別に求めた最適な学習方法を組み合わせることで計測条件が異なるデータを用いて正答率の高いクラスタリング手法を提案できると考えられる．

## 参考文献

1) Brown, M. P., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., ... & Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. Proceedings of the National Academy of Sciences, 97(1), 262-267.
2) Watanabe, Y., Seno, S., Takenaka, Y., & Matsuda, H. (2012). An estimation method for inference of gene regulatory net-work using Bayesian network with uniting of partial problems. BMC genomics, P(Suppl 1), S12.
3) Lee, Y., & Lee, C. K. (2003). Classification of multiple cancer types by multicategory support vector machines using gene expression data.Bioinformatics, 19(9), 1P2-1P9.
4) COGs Phylogenetic classification of proteins encoded in complete genomes, http://www.ncbi.nlm.nih.gov/COG/ (2015-5-21 参照)
5) Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A genomic perspective on protein families. Science, 278(5338), 631-637.
6) KEGG: Kyoto Encyclopedia of Genes and Genomes, http://www.genome.jp/kegg/ （2015-5-21 参照）
7) MultiFun, a cellfunction assignment schema, http://genprotec.mbl.edu/files/MultiFun.html（2015 -5-21 参照）
8) かずさ DNA 研究所, http://www.kazusa.or.jp/ (2015-5-21 参照)
9) Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., ... & Weil, B. (2002). MIPS: a database for genomes and protein sequences. Nucleic acids research, 30(1), 31-34.
10) NCBI http://www.ncbi.nlm.nih.gov/ (2015-5-21 参照)
11) RhizoBase http://genome.microbedb.jp/rhizobase/Mesorhizobium（2015-5-21 参照）