Vol.2011-MPS-86 No.4
Vol.2011-BIO-27 No.4
2011/12/1

Two-way AIC: マイクロアレイデータに基づく発現量変動遺伝子検出の新手法

露﨑 弘毅[†] 富永 大介[††] 權 娟大[†††] 宮崎 智[†††]

DNA マイクロアレイの遺伝子発現データから、真に生物学的に繋がりがある遺伝子群を選択的に検出するのは未だ難しい問題であり、今後も更なる発展が求められている。近年では大量の遺伝子発現データが個人レベルで取得可能となってきた。そこで本研究では、複数の比較実験データを統合したメタデータセットから得られる実験側で見た変動と遺伝子側で見た変動を利用し、両方向で発現変動を判定する、two-way AIC という手法を開発した。two-way AIC と他の統計手法を比較したところ、どの手法よりも two-way AIC は特異度が高く、また超幾何分布の*p*値が低かった。そのため two-way AIC は機能的に繋がりがある遺伝子群を選択的に検出する能力に優れているということが示された。

Two-way AIC: A Novel Method for Detection of Differentially Expressed Genes from Microarray Data Sets

Koki Tsuyuzaki[†], Daisuke Tominaga[††], Yeondae Kwon[†††] and Satoru Miyazaki[†††]

Detection of gene clusters which truly have biological relationship from DNA microarray data sets is still a difficult problem. Recently, many different microarray data are publicly available. In this paper, we built the integrated microarray data meta-data sets composed by multiple comparative experiment data and developed a method, called two-way AIC, which makes use of experiment's differential and gene's differential to detect differentially expressed genes. Compared to other methods, two-way AIC has high specificity and low hypergeometric *p*-value. In conclusion, two-way AIC is superior to any other methods in terms of selective detection of gene clusters which have biological relationship.

1. 背景・目的

DNA マイクロアレイの遺伝子発現データから発現変動遺伝子を検出する方法は、現在数多く開発されている。共に発現変動を起こした共発現遺伝子群は、実験系で与えられた刺激に対して、同じタイミングで変動を起こしているため、生物学的な繋がりが示唆される。しかし、内在的に様々なノイズを含むこれらのデータから、真に生物学的に繋がりがある遺伝子群を選択的に検出するのは未だ難しい問題であり、今後も更なる発展が求められている。

近年では遺伝子発現データのデータベースへの登録件数の増加や、次世代シーケンサの登場により、大量の遺伝子発現データが個人レベルで取得可能となってきた。そのため、研究室内で得られた遺伝子の発現データだけでなく、他の実験室で得られたデータを利用することができる。そこで本研究では、複数の比較実験データを統合したメタデータセットを作成し、そこから得られる実験側で見た変動と遺伝子側で見た変動の両方向で遺伝子の発現変動を判定する、two-way AIC という手法を開発し、真に生物学的繋がりがある遺伝子群をより選択的に検出することを目的とした。

2. 準備

2.1 DNA マイクロアレイ

DNAマイクロアレイは生物の全遺伝子の発現を同時に俯瞰することができる技術であり、現在様々な研究で利用されている[1]。実験では、生物が転写したmRNAから逆転写され蛍光標識が施されたcDNAと、各cDNAと相補的になるようにデザインされた、基盤上に幾つも並べられたプローブ間でハイブリダイゼーション反応が起こるという原理により、蛍光強度の値からその生物の遺伝子発現の状態を間接的に観察することができる。

実際の実験では1チップでの蛍光強度の値そのものを利用することはなく、試薬の投与等なんらかの刺激を与えた生物のデータ（処置群）と、何も処置をしなかった生物のデータ（対照群）とで比較して、刺激によりどの程度値に変化がみられたのかを調べる。全遺伝子のうち値が急激に変化したものは、発現変動が起きたといい、発現変動が起きた遺伝子のことを発現変動遺伝子という。

† 東京理科大学大学院 薬学研究科 薬科学専攻
Graduate School of Pharmaceutical Sciences, Tokyo University of Science
†† 産業技術総合研究所 生命情報科学研究センター
Computational Biology Research Center, National Institute of Advanced Industrial Science and Technology
††† 東京理科大学 薬学部 生命創薬科学科
Faculty of Pharmaceutical Sciences, Tokyo University of Science

Vol.2011-MPS-86 No.4
Vol.2011-BIO-27 No.4
2011/12/1

2.2 オペロン遺伝子群

試薬の投与、培養条件の変化等、各実験条件で与えられた刺激によって遺伝子は発現変動を起こす。例えば、ある 2 つの遺伝子に注目した時に、生物学的に何らかの繋がりがその遺伝子同士にある場合、それらは同じ刺激に反応して、同じタイミングで発現変動を起こすはずである。本研究では生物学的繋がりとして、原核生物に特有の機構であるオペロンに着目した[2]。オペロンはゲノム上に存在する同時に転写される遺伝子セットであり、複数の遺伝子から構成される（図 1）。同じオペロンに属する遺伝子同士は代謝経路上で反応が隣同士である酵素[3]や、同じトランスポーターを形成するサブユニット[4]を転写する等、機能的に同じである場合が多い。そのため適切な統計手法を用いることができれば、マイクロアレイの遺伝子発現データからも共に検出されるべきであるといえる。

実際に本研究で使用した緑膿菌のデータにおいて、同一のオペロンに属する遺伝子を 2 つ選び出し、相関係数を計算するという操作を全ての組み合わせ 857 ペアで行ったところ、0.734 とかなり強い正の相関が確認された。乱数をもとに同じペア数だけゲノム上からランダムに 2 つの遺伝子を選び出して相関係数を計算しても 0.182 と、それほど相関はみられないため、オペロンの遺伝子ペアはどれかが発現変動遺伝子であった場合、オペロン全体としても発現変動遺伝子である可能性が高い。そこで本研究では、この同一のオペロンに属する遺伝子をどれだけ選択的に検出できるかをみることで、各統計手法の性能を比較した。

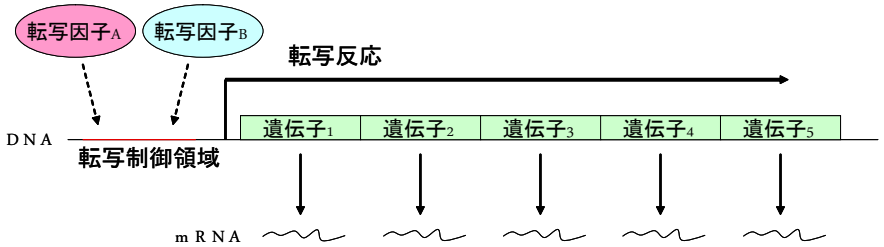

図 1　オペロンの転写反応

3. 方法

3.1 メタデータセット作成

① マイクロアレイデータ取得

全てのマイクロアレイデータは Affymetrix 社の GeneChip P. aeruginosa Genome Array から計測されたもののみを用いた。また 1 つでも欠損値を含む実験データは利用しなかった。データはマイクロアレイのデータベースである、NCBI の GEO[5]から 282 実験、EBI の ArrayExpress[6]から 7 実験、合計 289 実験分のデータを RAW 画像データ形式として取得した。(GEO での Accession: GPL84、ArrayExpress での Accession: A-AFFY-30)

②データの正規化

マイクロアレイの生の蛍光強度の数値には、全体的なシグナルの傾向の差やプローブ本来の蛍光強度の差などに起因するあらゆるバイアスが混在している。そのため、このバイアスの影響を減らすために全てのマイクロアレイデータに対して、RMA 法[7]を用いて正規化を行った。

上記の手順から、289 の比較実験データをまとめたメタデータセットを作成した。以後説明のために、作成したメタデータセットに対し、各実験データにおいて全遺伝子を解析する方向を“実験側”、各遺伝子における全実験での発現を解析する方向を“遺伝子側”と記す（図 2）。

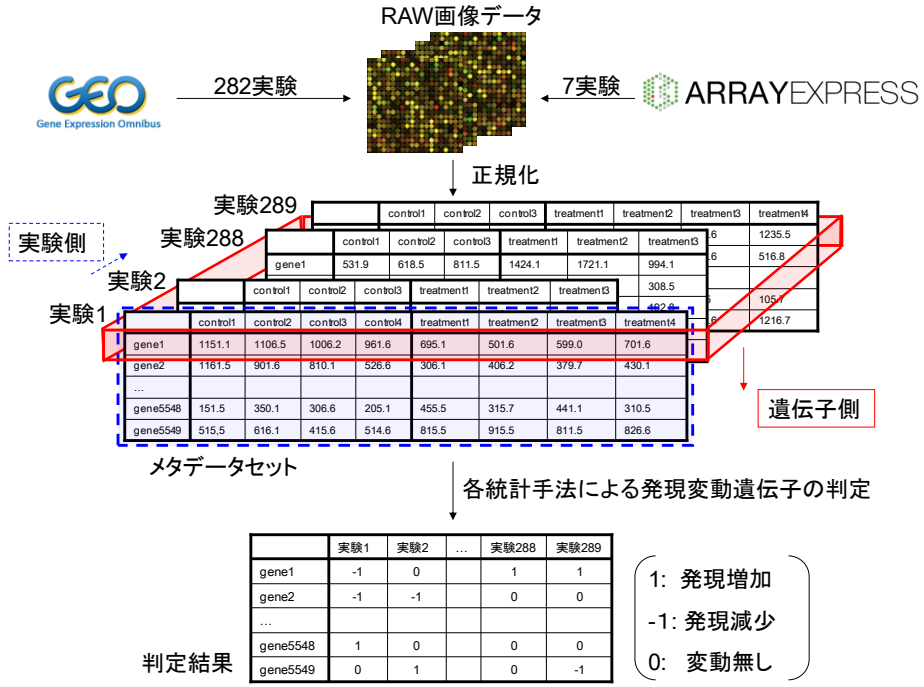

図 2　メタデータセットの作成

Vol.2011-MPS-86 No.4
Vol.2011-BIO-27 No.4
2011/12/1

3.2 two-way AIC

① AIC による分布の裾の切り出し

マイクロアレイのデータは正規分布を仮定したパラメトリックな統計手法を利用して、発現変動遺伝子を割り出すことが多い。例えば以下に示す、ある遺伝子において、対照群と比較して処置群が何倍発現量が変化したかを示す Fold Change（FC）も、正規分布に従うとして利用されている統計量の 1 つである。ただし、log は底 2 の対数、$\hat{t}$ は処置群平均値、$\hat{c}$は対照群平均値とする。

$$FC = \log \frac{\hat{t}}{\hat{c}}$$

しかし、各実験で全遺伝子のFCを計算した結果を、5549 行（遺伝子）×289（実験）のFC行列とし、縦横で各々閾値 p=0.05 の正規性検定（Kolmogorov-Smirnov 検定）を行ったところ、縦（685/5549）、横（1/289）共にほとんどの場合において分布の正規性が確認できなかった。このようなアレイデータの正規分布に従わない事例は他の研究でも数多く報告されている[8][9]。本研究においては、これは実際のデータが正規分布と比較して裾が広いことが原因であり（図 3）、裾にあるデータを最適な数だけ棄却すると、正規分布にあてはまることが確認された。

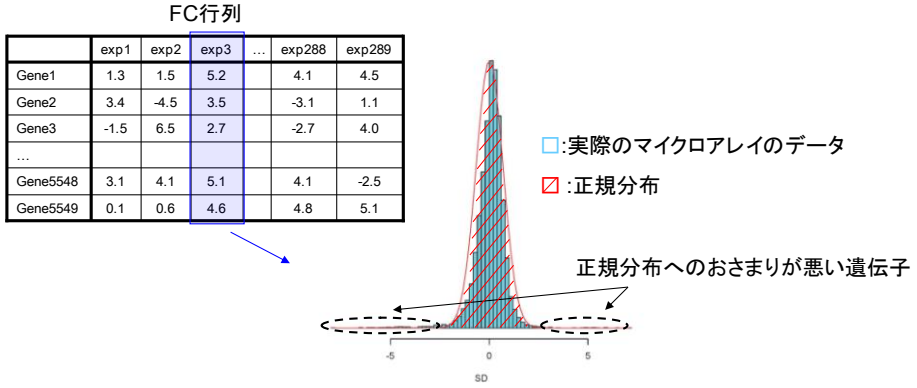

図 3　実際のデータの正規分布からのずれ

解析者によってはこの裾の部分を数%程度外れ値として棄却してから、各種の解析を行う場合もある[10]。しかし、本研究で検証用データとして用いた、オペロンの遺伝子群データを見てみると、変動を起こした時にこの広い裾の部分に集団で移動している事例が幾つも確認された（図 4）。

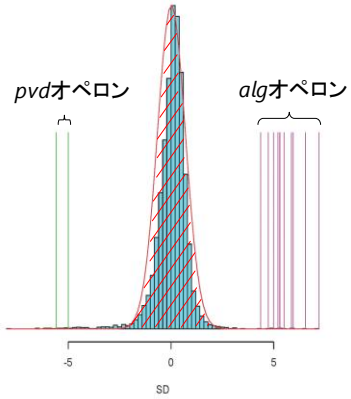

図 4　*pvd* オペロン、*alg* オペロン遺伝子の正規分布からのずれ

そのためこの広い裾の部分のみを切り出すことができれば、検出した遺伝子のうち、オペロンをより多く含むことになるのではないかと考えた。正規分布へのおさまりが悪い分布の広い裾の部分だけを選択的に切り出す手法は、先行研究で用いられている上田の統計量Uという、AIC を利用した統計量を用いた[11]。ただしここで、nは非外れ値数、sは外れ値数、σは外れ値を含んだ全てのデータでの標準偏差を意味する。

$$U = \frac{1}{2}\mathrm{AIC} = n\log\sigma + \sqrt{2} \times s \times \frac{\log n!}{n}$$

分布の平均値から離れている順に、そのデータを外れ値とした時のUの値を計算し、最小のUをとるデータまでを外れ値として判定した。

② two-way ANOVA について

従来手法では遺伝子側、実験側のどちらかからデータを扱うことがほとんどである。遺伝子側、実験側の変動の両方向の情報を利用するものとしては、two-way ANOVA が挙げられる[12] [13] [14]。two-way ANOVA では、全てのデータは以下の線形式に従うとする。

$$x_{i,j,k} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{i,j} + \epsilon_{i,j,k}$$

ただしμは全データの平均値、α_iは遺伝子側の主効果、β_jは実験側の主効果、$(\alpha\beta)_{i,j}$は遺伝子と実験の交互作用、$\epsilon_{i,j,k}$は誤差である。ここで遺伝子側の主効果とは、遺伝子側のデータによる全データの平均値からのずれ、実験側の主効果とは、実験側のデータによる全データの平均値からのずれ、交互作用とは、主効果によらず、ある遺伝子

Vol.2011-MPS-86 No.4
Vol.2011-BIO-27 No.4
2011/12/1

がある実験条件におかれることでみられる、全データの平均値からのずれを意味する。そのため、交互作用に寄与する遺伝子は発現変動遺伝子である可能性が高いといえる[14]。しかし、two-way ANOVA では、全データのうち交互作用がみられたかどうかを判定することはできるが、具体的にどの実験のどの遺伝子が交互作用に大きく寄与したのかを特定することは難しく、さらなる解析が必要となる。

③　two-way AIC のアルゴリズム

two-way ANOVA の交互作用に相当する、各遺伝子、各実験のもともとの変動の大きさを除いて、ある遺伝子がある実験条件におかれることでみられる大きな変動を検出するために、上田の統計量 U による裾の切り出しを遺伝子側、実験側の両方向で適用した（図 5）。以下に計算手順を示す。

ⅰ. メタデータセットから FC 行列を作成する。
ⅱ. FC 行列のうち遺伝子側と実験側とで各々平均値が 0、標準偏差が 1 となるように、標準化（Z-スケーリング）を行う。
ⅲ. 各々の方向で標準化されたデータに対し、AIC を利用して正規分布へのおさまりが悪い分布の裾にある遺伝子を検出する。
ⅳ. 各々の方向で判定された結果、両方向で共に発現変動が起こった場合を、最終的な発現変動遺伝子と判定する。

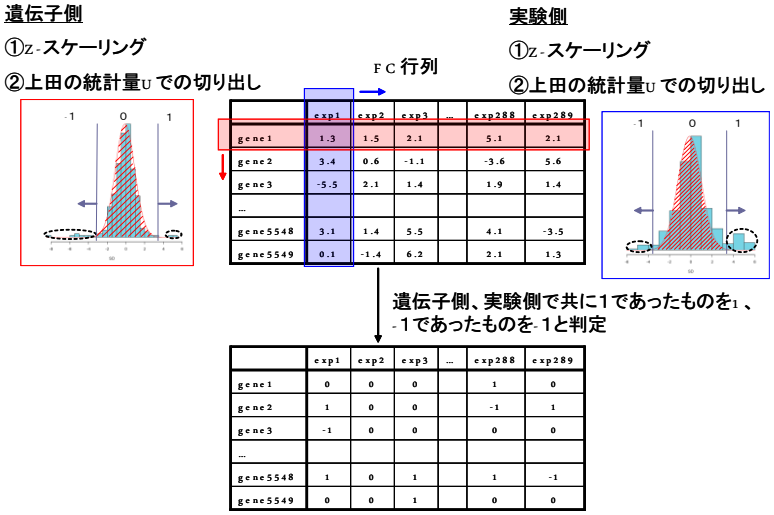

図 5　two-way AIC のアルゴリズム

3.3 比較する他の統計手法

本研究での手法 two-way AIC と、以下の 6 手法との比較を行った。

①　F 検定/t 検定<実験側>

まず F 検定で対象群と処置群の分散に差があるかどうか検定し、次に F 検定の結果から等分散であるとした場合、student の t 検定を、不等分散であるとした場合、Welch の t 検定を行う。

②　RankProducts 法<実験側> [15]

対照群内標本と処置群内標本のあらゆる組み合わせでのFCをもとに計算する RankProducts という統計量を用いて遺伝子の発現変動を順位付けする手法である。

③　AIC<遺伝子側> [11]

ある遺伝子がどの実験系で発現変動したのかを AIC を利用して検出する手法である。

④　AIC<実験側>

ある実験系でどの遺伝子が発現変動したのかを AIC を利用して検出する手法である。

⑤　標準偏差2σ <実験側＋遺伝子側>

遺伝子側での発現変動、実験側での発現変動がともに正規分布に従うと仮定し、両方向において、平均値から2σ（σ: 標準偏差）以上離れた遺伝子を発現変動遺伝子として判定する。

⑥　標準偏差3σ <実験側＋遺伝子側>

遺伝子側での発現変動、実験側での発現変動がともに正規分布に従うと仮定し、両方向において、平均値から3σ以上離れた遺伝子を発現変動遺伝子として判定する。

①、②はマイクロアレイで広く利用されている従来手法である。③、④は AIC の片方向だけでの適用と比較するために行った。③は先行研究で既に利用されている。⑤、⑥は two-way AIC と同様に遺伝子側と、実験側に対して検定を行うが、AIC ではなく標準偏差を用いることで、AIC と標準偏差とでどちらが適した統計量であるかを検証するために行った。

Vol.2011-MPS-86 No.4
Vol.2011-BIO-27 No.4
2011/12/1

3.4 手法の性能の比較

①オペロンの遺伝子座取得

緑膿菌のオペロンの遺伝子座のデータは京都大学の Operon Database[16]、及び University of British Columbia の Pseudomonas Genome Database[17]から取得した。遺伝子座とは遺伝子がゲノム上のどの場所に位置しているかを意味する通し番号である。

② 感度、特異度、超幾何分布のp値の計算

各手法により発現変動遺伝子と判定された遺伝子のうち、オペロン遺伝子群をどれだけ選択的に検出できたかを検証するために、各オペロンにおいて感度、特異度、超幾何分布のp値を計算した。ここでの感度とは、オペロンに属する遺伝子のうち何割を検出できたかを意味する。特異度とは、オペロンに属する遺伝子以外の遺伝子をどれだけ検出しないで済んだかを意味する。超幾何分布のp値とは、判定結果が偶然に起こる場合どの程度まれなことであるかを示す確率を意味する。これら指標の計算は各オペロンで行い、最後に全オペロンでの平均をとった。ただし、ある実験で与えられた刺激に対して、オペロンが常に発現変動するわけではないため、そのオペロンに属する遺伝子のうちの 1 つでも検出された実験系のみを対象とし、289 実験で一度も検出されなかったオペロンに対しては考慮しないこととした。以上から、全オペロンにおける平均感度（$\overline{sens}$）、平均特異度（$\overline{spec}$）、平均超幾何分布p値（$\overline{hgp}$）をそれぞれ以下のように設定した。

$$\overline{sens} = \sum_{k=1}^{N}\sum_{s=0}^{M}\left(\frac{TP_{k,s}}{T_k NM}\right)$$

$$\overline{spec} = \sum_{k=1}^{N}\sum_{s=0}^{M}\left(\frac{TN_{k,s}}{FNM}\right)$$

$$\overline{hgp} = \sum_{k=1}^{N}\sum_{s=0}^{M}\left(\frac{HP_{k,s}}{NM}\right)$$

ここで、Nは一度でも遺伝子が検出されたオペロンの数を意味する($0 \leqq N \leqq 93$)。Mは一度でもオペロンが検出された実験数を意味する($0 \leqq M \leqq 289$)。T_kはk番目のオペロン、s番目の実験での最大検出されうる遺伝子ペア数を意味する($0 \leqq T_k \leqq {}_{\text{オペロン遺伝子数}}C_2$)。$F$はオペロンに属さない遺伝子ペア数を意味する($F_k$ = 15392069)。$TP_{k,s}$はk番目のオペロン、s番目の実験での検出されたオペロン遺伝子ペア数を意味する。$TN_{k,s}$はk番目のオペロン、s番目の実験での検出されなかった非オペロン遺伝子ペア数を意味する。$FP_{k,s}$はk番目のオペロン、s番目の実験での検出された非オペロン遺伝子ペア数意味する。$HP_{k,s}$はk番目のオペロン、s番目の実験での検出されたオペロン遺伝子ペア数、検出された全遺伝子ペア数、そのオペロンで最大とりうるペア数、全遺伝子ペア数をもとに計算した超幾何分布のp値の値を意味する。

4. 実験

4.1 各統計手法との比較

本研究で開発した two-way AIC、従来の手法である F 検定/t 検定、RankProducts 法、AIC の遺伝子側での適用、AIC の実験側での適用、実験側と遺伝子側とで AIC ではなく2σ、3σだけで検出する方法の全 7 手法について性能を比較した。ただし、F 検定/t 検定、RankProducts 法は閾値によって検出される遺伝子の数が変わるため、比較に際しては two-way AIC と同等の感度になるような閾値に設定した。比較した結果を表 1 に示す。

表 1　各手法の性能の比較

手法	感度	特異度	超幾何分布のp値
two-way AIC(両側)	0.58578	0.99998	2.721×10^{-5}
F検定/t検定(実験側)	0.58477	0.99821	7.901×10^{-3}
RankProducts法(実験側)	0.58597	0.99717	1.123×10^{-2}
AIC(遺伝子側)	0.65665	0.68416	2.085×10^{-1}
AIC(実験側)	0.75034	0.99967	5.325×10^{-4}
2σ(両側)	0.65270	0.99871	5.202×10^{-3}
3σ(両側)	0.65488	0.99990	4.030×10^{-4}

この結果から、two-way AIC は同程度遺伝子を検出するように感度を設定した、従来の手法 F 検定/t 検定及び RankProducts 法と比較して、特異度が非常に高いことがわかった。これは誤って検出する数がどの手法よりも少ないことを意味する。

またデータのうち正規分布への収まりが悪い裾のデータを発現変動遺伝子とみなす AIC を遺伝子側、実験側のどちらか片方向で行った手法と比較しても、two-way AIC の方が最も超幾何分布のp値が低く、よりオペロンの遺伝子を選択的に検出できていることが確認された。

遺伝子側、実験側の両方向で発現変動遺伝子を検出するのに際しても、単に標準偏差（2σ、3σ）を利用したものよりも、AIC を利用した two-way AIC のほうがより選択性は優れていた。

これらの結果から、マイクロアレイのデータに関しては、遺伝子側、実験側の両方向で共に変動が大きい遺伝子に絞り込むことが有効であり、その中でも two-way AIC

Vol.2011-MPS-86 No.4
Vol.2011-BIO-27 No.4
2011/12/1

は、どの手法よりも、生物学的繋がりがある遺伝子群を選択している。また AIC は閾値を設定する必要が無いことからも、実用上においても優れているといえる。以上より共発現している遺伝子群をまとめて検出するという点において two-way AIC は強力な手法であることが示された。

5. まとめと今後の課題

本研究で開発した手法は、比較したどの統計手法よりも、特異度が高かった。また超幾何分布のp値も 2.721×10^{-5} と非常に低いことから、偶然では考えられないほどオペロンの遺伝子群を検出することができた。このことにより、two-way AIC は、他の手法と比べて相対的にも優れており、絶対的にもオペロンを選択的に検出する能力に長けているといえる。

two-way AIC のオペロン検出の選択性が優れている理由としては、以下のことが考えられる。まず、理論的な正規分布は、変数間が独立であることが前提にあるが、実際の遺伝子発現のデータにおいては、独立のものと従属のものが混在している。独立なものは、互いにまったく影響を及ぼさない遺伝子同士である。例として、代謝経路上で遠い遺伝子同士や、他の生物には存在するがその生物内では機能しているか不明なホモログ等が挙げられる。一方従属のものは、本研究のオペロンのように生物学的に繋がりがある遺伝子同士である。オペロンのデータでみられたように、従属な遺伝子は集団での変動が顕著であるため、実験で与えられた刺激に対して変動を起こした場合、正規分布の裾の方までまとまって移動し、正規分布から大きく逸脱することが多い。よって正規分布から逸脱した広い裾のデータのみを発現変動遺伝子と判定する two-way AIC の性能が優れていると考えられる。そのため引き続き変数間の独立性という観点からも、更なる解析を行っていく予定である。

また今後は本研究で用いたオペロンに加え、転写因子遺伝子とその転写因子に転写制御を受ける遺伝子同士、パスウェイ上で隣り合う反応を担当する遺伝子同士、転写するタンパク質同士で相互作用が確認されている遺伝子同士など、生物学的に繋がりがある遺伝子群のデータを増やし、加えて他のモデル生物でも同様の手法の性能比較を行う予定である。

参考文献

[1] D. Gershon, Microarray technology: An array of opportunities, *Nature*, **416**, 885-891, 2002.
[2] F. Jacob, and J. Monod, Genetic Regulatory Mechanisms in the Synthesis of Proteins, *J. Mol. Biol*, **3**, 318-356, 1961.
[3] K. Kin, Y. U. Kin, B. H. Koh, S. S. Hwang, S. H. Kim, F. Lepine, Y. H. Cho, and G. R. Lee, HHQ and PQS, two Pseudomonas aeruginosa quorum-sensing molecules, down-regulate the innate immune responses through the nuclear factor- κ B pathway, *Immunology*, **129**, 578-588, 2010.
[4] S. Jain, and D. E. Ohman, Role of an Alginate Lyase for Alginate Transporter in Mucoid ,Pseudomonas aeruginosa, *Infection and Immunity*, **73**, 6429-6436, 2005.
[5] R. Edgar, M. Domrachev, and A. E. Lash, Gene Expression Omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Research*, **30(1)**, 207-210, 2002.
[6] A. Brazma, H. Parkinson, U. Sarkans, M. Shojatalab, J. Vilo, N. Abeygunawardena, E. Holloway, M. Kapushesky, P. Kemmeren, G. G. Lara, A. Oezimen, P. R. Serra, and S. A. Sansone, ArrayExpress – a public repository for microarray gene expression data at the EBI, *Nucleic Acids Research*, **31(1)**, 68-71, 2003.
[7] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. B. Barclay, K. J. Antonellis, U. Scherf, and T. P. Speed, Exploration, normalization, and summaries of high density oligonucleotide array probe level data, *Biostatistics*, **4(2)**, 249-264, 2003.
[8] O. G. Troyanskaya, M. E. Garber, P. O. Brown, D. Botstein, and R. B. Altman, Nonparametric method for identifying differentially expressed genes in microarray data, *BIOINFORMATICS*, **18(11)**, 1454-1461, 2002.
[9] L. Hunter, R. C. Taylor, and S. M. Learch, and R. Simon, GEST: a gene expression search tool based on a novel Bayesian similarity metic, *BIOINFORMATICS*, **17(1)**, S115-S122, 2001.
[10] 藤渕航,堀本勝久, マイクロアレイデータ統計解析プロトコール, 洋土社, 2008.
[11] K. Kadota, S. Nishimura, H. Bono, S. Nakamura, Y. Hayashizaki, Y. Okazaki, and K. Takahashi, Detection of genes with tissue-specific expression patterns using Akaike's information criterion procedure, *Physiological Genomics*, **12**, 251-259, 2003.
[12] G. S. Churchill, Using ANOVA to Analyze Microarray Data, *Biotechniques*, **37(2)**, 173-177, 2004.
[13] L. Barrera, C. Benner, Y. Tao, E. Einzeler, and Y. Zhou, Leveraging two-way probe-level block design for identifying differential gene expression with high-density oligonucleotide arrays, *BMC Bioinformatics*, **5(42)**, 2004.
[14] S. Dudoit, Y. H. Yang, M. J. Callow, and T. P. Speed, Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments, Technical report of Stanford University of Medicine, 2000.
[15] R. Breitling, P. Armengaud, A. Amtmann, and P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, *Federation of European Biochemical Societiles*, **573**, 83-92, 2004.
[16] S. Okuda, T. Katayama, S. Kawashima, S. Goto, and M. Kanehisa, ODB: a database of operons accumulating known operons across multiple genomes, *Nucleic Acids Research*, **34**, D358-D362, 2006.
[17] G. L. Winsor, T. V. Rossum, R. Lo, B. Khaira, M. D. Whiteside, R. E. W. Hancock, F. S. L. Brinkman, Pseudomonas Genome Database: facilitating user-friendly, comprehensive comparisons of microbial genomes, *Nucleic Acids Research*, **37**, D483-D488, 2009.