

発現量分布の類似性に着目した マイクロアレイ測定データの統合手法の提案

福士 直翼^{1,a)} 斎藤 秀² 内田 真人^{1,b)}

概要：遺伝子の発現量を測定する方法の1つとして、DNA マイクロアレイ法が挙げられる。マイクロアレイを使用して行われた過去の実験により得られた大量の実験データは NCBI GEO (National Center for Biotechnology, Information Gene Expression Omnibus) のような公的なリポジトリにて公開されている。しかし、これらのデータで使用されているマイクロアレイには多くの種類が存在し、種類の異なるマイクロアレイ間には検出遺伝子数の差異、および検出遺伝子の種類の差異が生じる。そのため、異なるマイクロアレイを使用したデータ同士を単純に比較することはできない。一方で、複数の異なるマイクロアレイのデータを同時に解析することができれば、創薬や医学研究において標本数の増加による統計解析の信頼性向上が期待できる。そこで本論文では、発現量分布の類似性に着目したマイクロアレイデータの統合手法を提案し、その有効性を検証する。

NAOKI FUKUSHI^{1,a)} SHIGERU SAITO² MASATO UCHIDA^{1,b)}

1. はじめに

遺伝子発現の解析は、創薬や医学研究の分野で大きな役割を担っている。例えば、健康な人と病気の人で血液中の遺伝子発現を比較すると、病気の人では特徴的な遺伝子発現がある。これより、遺伝子発現が異なることを利用した病気の早期発見が可能となる [1]。また、同一の病気でも薬の効果がある人とならない人とは遺伝子発現が異なる。そのため、新薬開発において薬剤作用を遺伝子発現量の変化として捉えることは極めて重要である [2]。

遺伝子発現量を測定する方法の1つとして、DNA マイクロアレイ法がある。マイクロアレイを使用することで、細胞内に含まれる多数の遺伝子の発現量を同時に測定することができる。マイクロアレイを使用した実験により得られた大量の遺伝子発現データは NCBI GEO のような公的なリポジトリで公開されている。しかし、これらのデータを測定するために使用されたマイクロアレイには多くの種類が存在し、マイクロアレイの種類が異なると検出遺伝子の数や種類も異なる。そのため、異なるマイクロアレイで

測定されたデータ同士を単純に比較することはできない。一方で、もしも複数の異なるマイクロアレイで測定されたデータを統合して同時に解析することができれば、創薬や医学研究において標本数の増加による統計解析の信頼性向上が期待できる。

本論文では、異なるマイクロアレイで測定された発現量を関連付けて統合するために、種類の異なるマイクロアレイで測定された遺伝子発現データであっても、遺伝子の発現量と発現量の順位との間の関係が類似していることに着目した。この類似性を利用することにより、少なくとも1つのマイクロアレイで検出することのできる遺伝子の集合に対して、発現量の順位を一意に決定することが可能となる。本論文では、発現量の順位に対する発現量の分布傾向を利用することで種類の異なるマイクロアレイで測定された遺伝子発現データを統合する手法を提案する。

本論文の構成は以下のとおりである。第2節では、マイクロアレイの原理や種類、および品質管理について概観する。第3節では、世界最大の遺伝子発現データリポジトリである NCBI GEO における登録データを本論文における提案手法を適用できるように事前処理する方法について説明する。第4節では、本論文における提案手法の詳細な手順について説明する。第5節では、提案手法の評価実験と考察を行う。第6節は本論文のまとめである。

¹ 早稲田大学大学院 基幹理工学研究科 情報理工・情報通信専攻
Waseda University, Shinjuku, Tokyo 169-8555, Japan

² 株式会社 SIGNATE

a) n.wing.fks.0319@akane.waseda.jp

b) m.uchida@waseda.jp

2. DNA マイクロアレイ

2.1 マイクロアレイの原理

マイクロアレイとは、遺伝子発現量を測定するために、数万から数十万に区切られた樹脂やガラスなどの基板上に多数の DNA の部分配列を高密度に配置した分析ツールのことである。一般に、基板上に配置された DNA の部分配列のことをプローブと呼ぶ。マイクロアレイを使用すると、数百から数十万に上る網羅的な遺伝子発現解析を、少量の検体を用いて比較的短時間で行うことができる。このような利点から、マイクロアレイは遺伝子発現解析の分野で広く用いられている。

マイクロアレイによる遺伝子発現解析では、サンプル中の RNA を蛍光ラベルし、マイクロアレイ上のプローブと結合させる。この結合のことをハイブリダイゼーションと呼ぶ。ハイブリダイゼーションの結果、サンプル中に多く含まれる RNA はプローブとの結合数が多くなり、その部分の蛍光強度は強くなる。一方で、サンプル中に少量しか含まれない RNA はプローブとの結合数が少なくなり、蛍光強度は弱くなる。このような蛍光強度の強弱を測定することで、サンプル中での各遺伝子の発現量を解析することができる [3]。

2.2 マイクロアレイの種類

マイクロアレイには一色法アレイと二色法アレイの2種類がある。一色法アレイは、1枚のアレイに対して1種類のサンプルを用いて遺伝子の発現量を測定する。一色法アレイでは、研究対象であるサンプル中の遺伝子の発現量に関する情報を蛍光強度から得ることが可能である。一色法アレイを使用する利点は、コストが低く実験デザインも柔軟なことである。一方で、一色法アレイ同士のデータを比較する際には、基板の違いによりばらつきが生じる可能性を考慮し、アレイ間での補正が必要となる。

二色法アレイは、それぞれ異なる色の蛍光色素で標識した2つのサンプルを、等量かつ同時に1枚のアレイ上のプローブとハイブリダイゼーションさせ、色の比較で発現量を分析する。一色法アレイと比較して、1つのスポットで2つのサンプル間での遺伝子の発現比を直接比較することができるため、精確でわかりやすいという利点がある。一方で、二色法アレイでは蛍光色素の蛍光強度や取り込み効率の違いなどを考慮した標準化を行う必要があり、実験デザインが複雑になりがちになるという問題がある。

2.3 マイクロアレイの品質管理プロジェクト

マイクロアレイ技術の導入は、遺伝子発現研究に大きな影響を与えてきた。また、マイクロアレイを使用した実験により得られた大量の遺伝子発現データは NCBI GEO のような公的なリポジトリで公開されている。その一方で、

異なるマイクロアレイを使用すると結果に差異や矛盾が生じる可能性があるといったマイクロアレイ技術の信頼性への懸念も増加した [4]。例えば、同じ遺伝子名の発現データの再現性は、プローブ、プラットフォーム、実験ラボ、実験実施者、実験時間などの様々な要因によって人工的な影響を受けることが知られている [5]。

このような問題点を検証するために、マイクロアレイの品質管理を目的とした MAQC (MicroArray Quality Control) プロジェクトが米国食品医薬品局 FDA (Food and Drug Administration) 主導で発足した。その第1フェーズ (MAQC-I) では、同一のプラットフォーム内および異なるプラットフォーム間での遺伝子発現の測定の再現性や互換性について報告した [4]。MAQC-I では、同一の実験ラボで繰り返し実験を行ったとき、および実験ラボを変更して実験を行ったときに実験結果に再現性があることが示された。さらに、異なるプラットフォームを使って同一のサンプルを測定した際に、実験結果に矛盾が少ないことも明らかにされた。なお、MAQC-I における遺伝子発現データも NCBI GEO において公開されている。

3. 分析データ

本論文では、同一のサンプルを測定した複数の異なるマイクロアレイの遺伝子発現データを統合する手法を提案する。提案手法に適用する遺伝子発現データは、公的なリポジトリである NCBI GEO から入手する。しかしながら、提案手法を実行する上で必要となる各マイクロアレイにおける各検出遺伝子の発現量 (遺伝子発現の強度) は、NCBI GEO から入手できる遺伝子発現データからは直接わからない。そのため、NCBI GEO のデータはそのままの状態で使用することはできない。また、各マイクロアレイ間で検出遺伝子の識別子が異なる場合もある。次章で説明する提案手法においては、すべてのマイクロアレイのデータにおいて各検出遺伝子の発現量がわかっていること、および遺伝子の識別子が統一されていることが条件である。そこで本章では、この条件を満足するようにマイクロアレイの遺伝子発現データを前処理する。その一連の手順を以下に示し、その概要を図1に示す。

3.1 NCBI GEO における遺伝子発現データの前処理

NCBI GEO から入手できるデータの中で、本論文において必要なものが2つある。一つ目は、各マイクロアレイにおいて、どの遺伝子発現がどのプローブによって検出されるように設計されているかを表すデータである。二つ目は、各マイクロアレイにおいてサンプルを測定した際に各プローブで検出される遺伝子発現量を表すデータである。本論文では、以上の2つのデータを関連付けることにより、各マイクロアレイにおける各検出遺伝子の発現量を求める。ただし、1つの遺伝子が複数の異なるプローブによって検

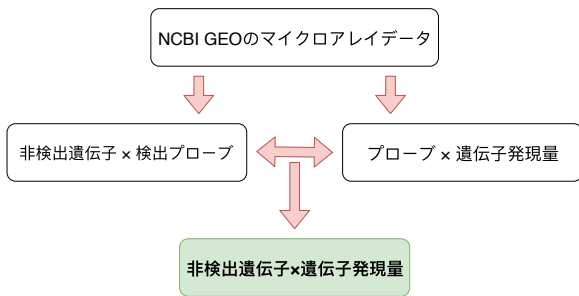


図 1 遺伝子発現データの前処理

出される可能性があるため、被検出遺伝子と検出プローブの関係は必ずしも一対一にはならない。これは、異なるプローブがそれぞれ同一遺伝子の別の部分配列を検出することがあるためである。本論文では、1つのマイクロアレイ内におけるすべての遺伝子に対して検出プローブを1つとした。この検出プローブは、同一遺伝子を検出するプローブ群の中で、発現量の計測値が最も大きい値であったプローブとした。なお、今回使用したデータでは遺伝子の識別子として、「Entrez Gene Database UID」,「GenBank Accession Number」の2種類が使用されていた。すべてのマイクロアレイ間で遺伝子の識別子を統一するために、識別子として「GenBank Accession Number」が使用されているデータについては、識別子を「Entrez Gene Database UID」に変換した。

3.2 異なる実験条件における測定データの一元化

MAQC-Iでは各マイクロアレイで共通の4種類のサンプルを測定し、複数の実験ラボにおいてそれぞれ複数回の繰り返し実験を行っている。そのため、ある1つのサンプルを測定したデータは、マイクロアレイごとに「実験ラボ数」×「繰り返し実験数」個存在する。本論文では、このような複数の実験データから得られる同一遺伝子の発現量の中から代表値を1つ決めることとした。具体的には、代表の発現量として、同一遺伝子の発現量群の最大値、最小値、中央値、平均値の4種類を試すこととした。

3.3 異なるマイクロアレイ間における測定データの正規化

一般に、異なるマイクロアレイで測定できる発現量のダイナミックレンジは異なる。したがって、同一遺伝子の発現量であっても、異なるマイクロアレイから得られた測定値を直接比較することはできない。そのため、異なるマイクロアレイにおけるダイナミックレンジが同等となるように、発現量の測定値を正規化する必要がある。正規化の方法にはさまざまなものがある。本論文では、total intensity normalizationを参考とし[6]、すべてのマイクロアレイにおいて共通に検出することのできる遺伝子の発現量の合計値が1になるように発現量の測定値を正規化した。その詳細は以下の通りである。

いま、 M 種類のマイクロアレイについて考える。また、マイクロアレイ j ($\in \{1, \dots, M\}$) で検出することのできる遺伝子の集合を S_j 、すべてのマイクロアレイにおいて共通に検出することのできる遺伝子の集合を $S_{\cap} = \bigcap_{j \in \{1, \dots, M\}} S_j$ と定義する。さらに、マイクロアレイ j における遺伝子 g ($\in S_j$) の発現量を $P_j(g)$ と定義する。このとき、マイクロアレイ j の遺伝子 g の発現量 $\hat{P}_j(g)$ を

$$\hat{P}_j(g) = \frac{P_j(g)}{\sum_{g' \in S_{\cap}} P_j(g')}, \forall g \in S_j \quad (1)$$

のように変換することで、 $S_{\cap}$ に含まれる遺伝子の発現量の合計値が1になるように正規化することができる。

4. 提案手法

4.1 提案手法 1

4.1.1 発現量分布と順位の関係

各マイクロアレイについて、計測された遺伝子の発現量の特徴を調べるために、横軸に発現量の順位、縦軸に発現量を取ったグラフを作成した。グラフを図2に示す。本図より、種類の異なるマイクロアレイで測定された遺伝子発現データであっても、遺伝子の発現量と発現量の順位との間の関係が類似していることがわかる。本論文では、このような発現量分布の類似性を利用することで、種類の異なるマイクロアレイで測定された遺伝子発現データを統合する手法を提案する。さらに、この提案手法が、発現量分布の類似性を利用せずに直接的に遺伝子発現データを統合する手法よりも妥当性の高い統合を可能とすることを示す。このことは、発現量分布の類似性には、遺伝子発現データを統合する上で有用な情報が含まれていることを意味する。

本論文では、MAQC-Iで使用された5種類のマイクロアレイの遺伝子発現データを用いる。これらの5つのマイクロアレイの検出遺伝子数は表1に示すように異なる。したがって、発現量の順位に関する統計量を求める場合には、マイクロアレイの検出遺伝子数が異なることを考慮しなければならない。なぜならば、マイクロアレイの検出遺伝子数が異なる場合には、発現量の絶対的な順位ではなく、検出遺伝子数の全体の中での相対的な順位で比較するのが自然であるためである。そこで本論文では、マイクロアレイの検出遺伝子数を相対的に揃えることを考える。検出遺伝子数は、統合に使用するすべてのマイクロアレイの中で、最も多いものを基準とする。本論文では、マイクロアレイが M 種類あった際に、マイクロアレイ j ($\in \{1, \dots, M\}$) における発現量の本来の順位が r である遺伝子を $G_j(r)$ と定義する。この場合、その遺伝子の正規化後の発現量は $\hat{P}_j(G_j(r))$ となる。以下、簡単のため $Q_j(r) = \hat{P}_j(G_j(r))$ と定義する。

4.1.2 同一順位の発現量の統計量の計算

例として、本来の検出遺伝子数が N_j 個のマイクロアレイ

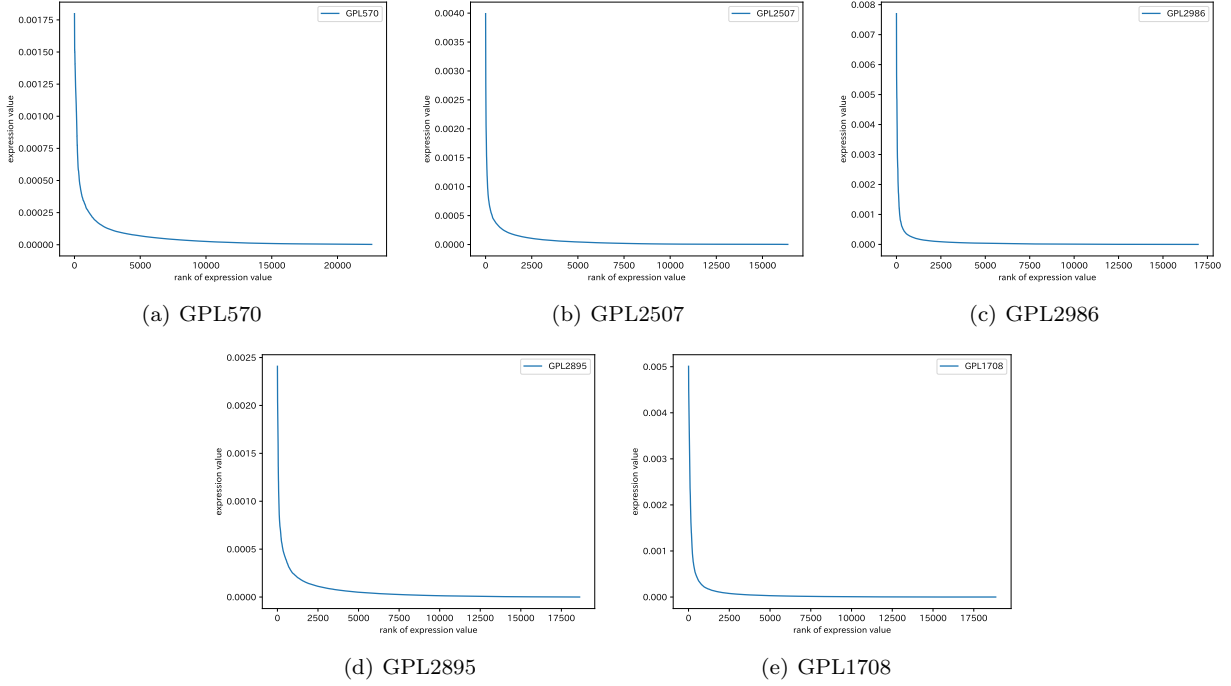


図 2 各マイクロアレイのデータにおける順位と発現量の関係

表 1 各マイクロアレイの検出遺伝子数

マイクロアレイ	検出遺伝子数
GPL570	22600
GPL2507	16373
GPL2986	16976
GPL2895	18625
GPL1708	18848

イ j の検出遺伝子数を、相対的に N ($\geq N_j$) 個に揃える場合を考える。ただし、 $N = \max_{j \in \{1, \dots, M\}} N_j$ である。本論文では、検出遺伝子数を相対的に揃えた後のマイクロアレイ j の順位 r の発現量 $Q'_j(r)$ を以下のように計算する。概念図を図 3 に示す。ただし、 n は 0 以上 N_j 未満の整数であり、 $Q_j(0) = Q_j(1)$ とする。

$$Q'_j(r) = \left\lfloor \frac{N_j}{N} r - (n+1) \right\rfloor Q_j(n) + \left\lceil \frac{N_j}{N} r - n \right\rceil Q_j(n+1),$$

$$\forall r \in \{1, \dots, N\} \text{ s.t. } n < \frac{N_j}{N} r \leq n+1 \quad (2)$$

次に、すべての j ($\in \{1, \dots, M\}$) について計算された $Q'_j(r)$ を統合した発現量 $\tilde{Q}(r)$ を定義する。本論文では、 $Q'_1(r), \dots, Q'_M(r)$ の最大値、最小値、中央値、ないしは平均値として $\tilde{Q}(r)$ を定義する。

4.1.3 各遺伝子の発現量の計算と順位の付与

統合された発現量 $\tilde{Q}(r)$, ($r \in \{1, \dots, N\}$) から、それに対応するマイクロアレイ j の発現量 $\tilde{Q}_j(r)$, ($r \in \{1, \dots, N_j\}$) を以下のように計算する。これは、式 (2) の逆の計算を行っていることに相当する。ただし、 n は 1 以上 N 未満の整数である。

$$\tilde{Q}_j(r) = \left\lfloor \frac{N}{N_j} r - (n+1) \right\rfloor \tilde{Q}(n) + \left\lceil \frac{N}{N_j} r - n \right\rceil \tilde{Q}(n+1),$$

$$\forall r \in \{1, \dots, N_j\} \text{ s.t. } n < \frac{N}{N_j} r \leq n+1 \quad (3)$$

本論文では、 $\tilde{Q}_j(r)$, ($r \in \{1, \dots, N_j\}$) を用いて、統合後の遺伝子 g ($\in S_U$) の発現量 $\tilde{P}(g)$ を計算する。ここで、 $S_U = \bigcup_{j \in \{1, \dots, M\}} S_j$ は少なくとも 1 つのマイクロアレイで検出することのできる遺伝子の集合である。以下では、マイクロアレイ j ($\in \{1, \dots, M\}$) における遺伝子 g の発現量の順位を $R_j(g)$ と定義する。マイクロアレイが異なると、検出することのできる遺伝子の数や種類も異なるため、 g を検出しないマイクロアレイが存在する場合がある。そのため、たとえばマイクロアレイ k では g を検出しない場合には、 $R_k(g)$ は定義しないこととする。本論文では、 $\{\tilde{Q}_j(R_j(g)) \mid R_j(g) \text{ が定義可能}, j = 1, \dots, M\}$ の最大値、最小値、中央値、ないし平均値として $\tilde{P}(g)$ を定義する。最後に、すべての g ($\in S_U$) に対して $\tilde{P}(g)$ を計算し、この値の降順に g を並べることで順位を付与する。

4.2 提案手法 2

前節で提案した手法よりも簡便な手法として、各マイクロアレイにおける同一遺伝子の発現量を直接統合する手法が考えられる。すなわち、マイクロアレイ j ($\in \{1, \dots, M\}$) の遺伝子 g ($\in S_U$) の正規化後の発現量を $\hat{P}_j(g)$ の最大値、最小値、中央値、ないし平均値により $\tilde{P}(g)$ を定義することも考えられる。この手法は、遺伝子の発現量と発現量の順位との間の関係が種類の異なるマイクロアレイ間で類似しているという特徴を使わない方法であり、前節で提案し

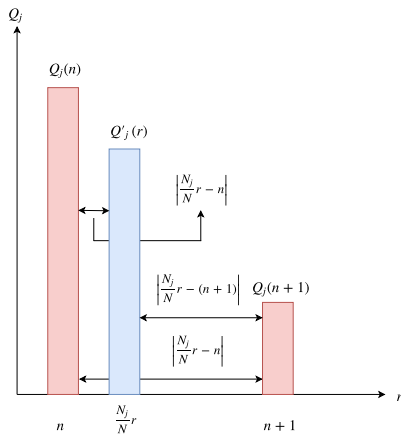


図 3 式 (2) の概念図

た手法と比較して手順が簡単であり，統合にかかる計算時間も短い．

5. 評価実験と考察

5.1 実験データ

本論文では，MAQC-I で使用された 5 種類のマイクロアレイの遺伝子発現データを用いて第 4 章で説明した提案手法を評価する．その際，第 3 章で示した手順により予めデータを加工する．なお，各マイクロアレイで測定されたサンプルは 4 種類存在する．本論文ではそのうち，Sample A (Stratagene Universal Human Reference RNA) を測定したデータを使用した．

図 4 はこれらのマイクロアレイで測定できる遺伝子の種類の違いを表したものである．図の縦軸にある A, B, C, D, E は，それぞれ GPL570, GPL2507, GPL2986, GPL2895, GPL1708 で測定できる遺伝子の集合を表す．例えば， $S_U \setminus (B \cup C \cup E)$ は， S_U のうちマイクロアレイ A および D のみで測定できる遺伝子の集合を意味する．また，横軸は対応する集合の要素数，すなわち遺伝子数を意味する．なお，これらの 5 つのマイクロアレイの少なくとも 1 つで測定できる遺伝子の集合である S_U の要素数は 25026 個，これらの 5 つのマイクロアレイのすべてにおいて測定できる遺伝子の集合である $S_{\cap}$ の要素数は 14246 個であった．このように，測定できる遺伝子の種類はマイクロアレイ間で大きく異なることがわかる．

5.2 統合前のマイクロアレイのデータ同士の比較

はじめに，MAQC-I で使用された 5 種類のマイクロアレイデータの整合性を確認する．これは，提案手法の評価におけるベースラインとなるものである．

表 2 は，任意の 2 つのマイクロアレイで測定された発現量の順位に関するスピアマンの順位相関係数を示したものである．スピアマンの順位相関係数が高いほど，マイクロアレイ i と j において，遺伝子の発現量の順位に線形の関

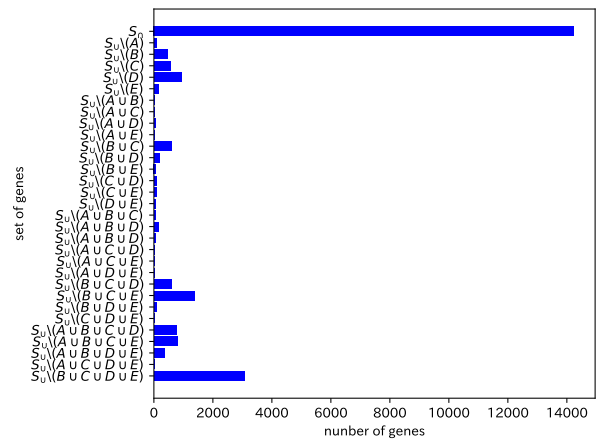


図 4 マイクロアレイごとの遺伝子検出状況

係があるということを示す．なお，マイクロアレイ i とマイクロアレイ j に関するスピアマンの順位相関係数 ρ は以下のように計算される．

$$\rho = 1 - \frac{6 \sum_{g \in S_{i \cap j}} \{R_i(g) - R_j(g)\}^2}{|S_{i \cap j}|^3 - |S_{i \cap j}|} \quad (4)$$

ここで， $S_{i \cap j}$ はマイクロアレイ i とマイクロアレイ j で共通に検出できる遺伝子の集合を表す．また， $R_i(g)$ と $R_j(g)$ はそれぞれ，マイクロアレイ i とマイクロアレイ j における遺伝子 g の発現量の順位を表す．ただし，スピアマンの順位相関係数を計算する際には，各マイクロアレイにおいて，それぞれ $S_{i \cap j}$ に含まれる遺伝子のみについて発現量の順位を振り直した．

また，視覚的に整合性を確認するために， $S_{i \cap j}$ に含まれる全ての遺伝子に対して，マイクロアレイ i での発現量の順位を横軸，マイクロアレイ j での発現量の順位を縦軸とした散布図を利用する．図 5 はその一例であり，GPL570 と GPL2507 を比較した際の散布図をヒートマップで表示したものである．ヒートマップ右横のカラーバーの数字は，その区画に含まれる遺伝子のプロットの個数を表している．また，ヒートマップ上には $y = x$ を理想直線として青色の実線で表している．理想直線 $y = x$ に近い部分の色が濃いほど，比較している 2 つのマイクロアレイの両方において発現量の順位が近い遺伝子が多いということを示す．これを定量的に評価するために，散布図における各プロットと理想直線 $y = x$ との距離の中央値を表 3 に示す．

発現量の順位に関するスピアマンの順位相関係数（表 2），散布図（図 5），および散布図における各プロットと理想直線 $y = x$ との距離の中央値（表 3）より，MAQC-I で使用された 5 種類のマイクロアレイデータ間には，発現量の順位に関するゆるやかな相関があることを確認できる．以下では，第 4 章で説明した提案手法により 5 つのマイクロアレイデータを統合することで，これらのマイクロアレイデータと強い相関を持つ発現量の順位を構成できることを示す．これは，統合前のデータ間における発現量の順位

表 2 元データ同士のスピアマンの順位相関係数の計算結果

組み合わせ	相関係数
GPL570, GPL2507	0.731
GPL570, GPL2986	0.782
GPL570, GPL2895	0.647
GPL570, GPL1708	0.766
GPL2507, GPL2986	0.715
GPL2507, GPL2895	0.599
GPL2507, GPL1708	0.709
GPL2986, GPL2895	0.636
GPL2986, GPL1708	0.761
GPL2895, GPL1708	0.622

表 3 元データ同士の理想直線との距離の中央値の計算結果

組み合わせ	距離中央値
GPL570, GPL2507	1312.39
GPL570, GPL2986	1145.51
GPL570, GPL2895	1700.59
GPL570, GPL1708	1324.41
GPL2507, GPL2986	1328.50
GPL2507, GPL2895	1519.22
GPL2507, GPL1708	1334.31
GPL2986, GPL2895	1446.03
GPL2986, GPL1708	1205.62
GPL2895, GPL1708	1557.76

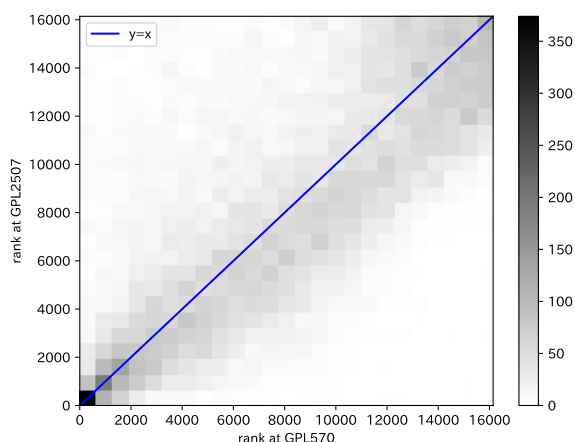


図 5 統合に使用する元データ同士の比較

には互いにゆるやかな相関のみがあるのに対し、それらのデータを統合した発現量の順位については統合前のすべてのデータと強い相関を持つ、という一見両立しない結果であり自明ではない。

5.3 提案手法の評価

次に、提案手法により 5 つのマイクロアレイデータを統合した発現量の順位と、統合前の各マイクロアレイデータにおける発現量の順位を比較する。なお、第 3 章および第 4 章で説明した通り、本論文ではマイクロアレイごとに 1 つのサンプルを測定したデータを 1 つに統合する際 (3.2

表 4 提案手法 1 を用いた時の比較結果 (相関係数上位 3 位以内)

パラメータの組み合わせ	相関係数 (平均)	距離中央値 (平均)
median,median,median	0.865	672.46
median,max,median	0.865	671.19
median,mean,median	0.865	670.48

表 5 提案手法 1 を用いた時の比較結果 (距離中央値上位 3 位以内)

パラメータの組み合わせ	相関係数 (平均)	距離中央値 (平均)
median,mean,median	0.865	670.48
median,max,median	0.865	671.19
median,median,median	0.865	672.46

表 6 提案手法 2 を用いた時の比較結果 (相関係数上位 3 位以内)

パラメータの組み合わせ	相関係数 (平均)	距離中央値 (平均)
mean,median	0.859	935.22
median,median	0.858	956.15
max,median	0.856	912.59

表 7 提案手法 2 を用いた時の比較結果 (距離中央値上位 3 位以内)

パラメータの組み合わせ	相関係数 (平均)	距離中央値 (平均)
max,median	0.856	912.59
mean,median	0.859	935.22
median,median	0.858	956.15

節), 同一順位の発現量の統計量を計算する際 (4.1.2 節), および各遺伝子の発現量を計算する際 (4.1.3 節, 4.2 節) にそれぞれ最大値, 最小値, 中央値, ないし平均値の 4 種類の統計量を使用することとしている。そこで, 提案手法 1 ではこれらを調整パラメータとして計 64 パターンの実験を行い, 順位相関係数もしくは理想直線からの距離の中央値が最適となる組み合わせを調べる。また, 提案手法 2 については同一順位の発現量の統計量を計算する過程がないため, 計 16 パターンの実験を行う。

まず, 提案手法 1 により 5 つのマイクロアレイデータを統合した発現量の順位と統合前の各マイクロアレイデータにおける発現量の順位を比較した結果を示す。表 4 は順位相関係数の平均値が上位 3 位以内であった結果, 表 5 は理想直線との距離の中央値の平均値が上位 3 位以内であった結果である。次に, 提案手法 2 を使用して同様の比較を行った結果を表 6 および表 7 に示す。さらに, 提案手法 1 および提案手法 2 から得られた発現量の順位について, 理想直線との距離の中央値の平均値が最適となったときの場合の散布図を図 6 および図 7 にそれぞれ示す。

5.4 考察

表 2~表 7 および図 5~図 7 に示した発現量の順位に関する相関係数と散布図より, 提案手法 2 よりも提案手法 1 の方が明らかに良好な結果を示していることがわかる。提案手法 2 を用いた場合, 順位相関係数は提案手法 1 を用いた場合よりもやや低い程度であるのに対して, 理想直線との距離の中央値の平均値は提案手法 1 を用いた場合の約 1.34

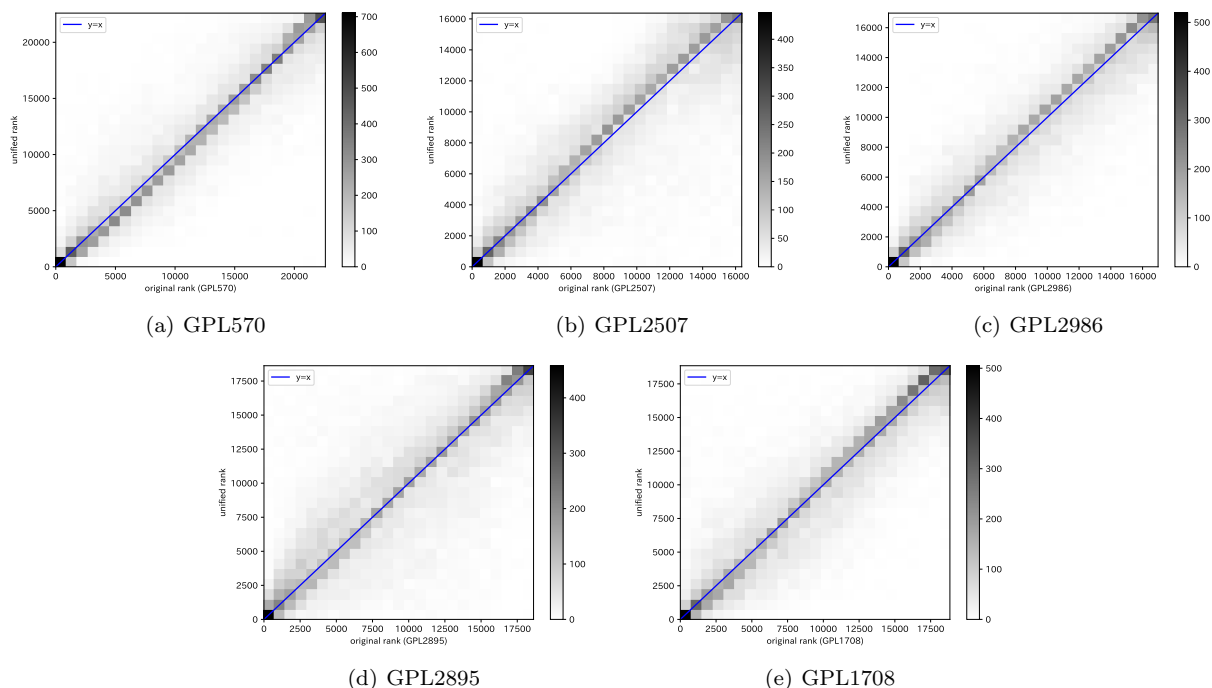


図 6 提案手法 1 を用いた場合の散布図

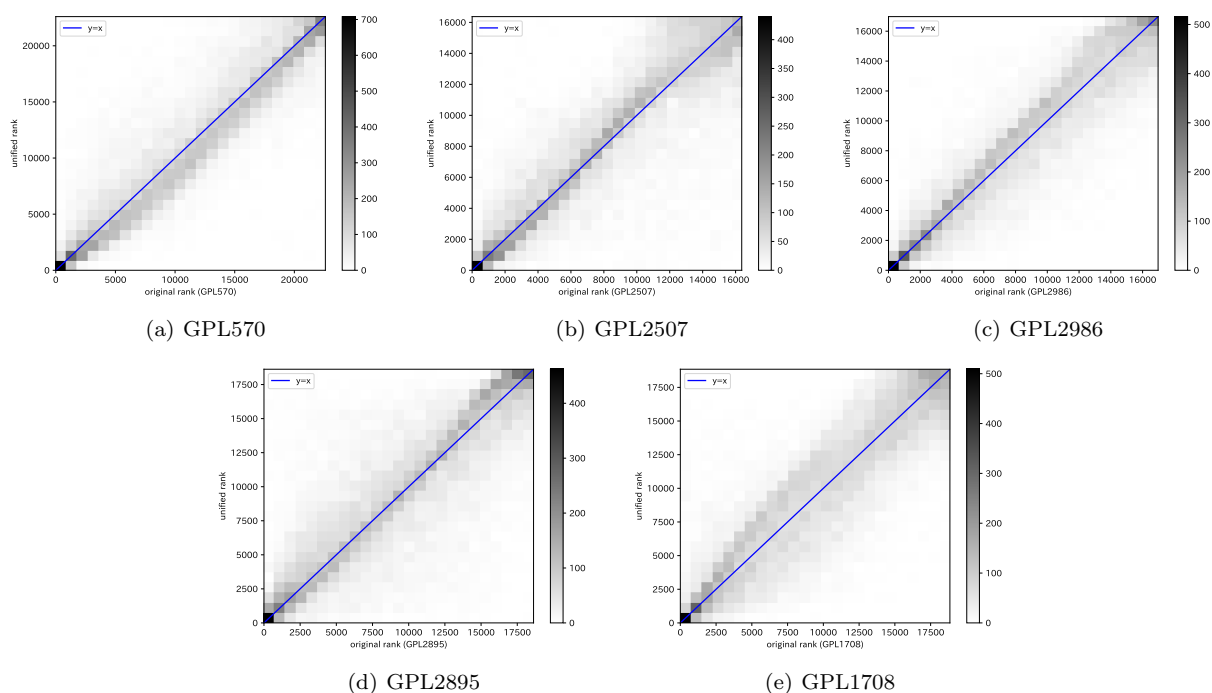


図 7 提案手法 2 を用いた場合の散布図

倍になっていることがわかる。実際に図 6 および図 7 を比較すると、提案手法 1 を用いた場合では理想直線 $y = x$ の付近に多くのプロットが集中しているのに対して、提案手法 2 を用いた場合では提案手法 1 と比較して $y = x$ の付近にプロットが集中していないことが確認できる。これより、順位相関係数が近い値を示す場合でも、統合後の順位が統合元のデータにおける順位と十分に整合しない場合があることがわかる。提案手法 1 は提案手法 2 と比較すると

統合にかかる手順も多く、提案手法 2 のように直感的な方法とはなっていない。しかしながら、以上の結果より、提案手法 1 のように各マイクロアレイのデータにおける発現量分布の類似性に着目した統合をおこなうことが非常に重要であるということがわかった。

なお、表 4～表 7 をみると、順位相関係数、および散布図における理想直線との距離の中央値が上位になっているパラメータの特徴として、すべて各遺伝子の発現量を決定

する際(4.1.3節, 4.2節)に中央値を用いているということがわかる。これより, 統合モデルを最適化するためには, 各遺伝子の発現量を決定する際に中央値を使用することが適切であると考えられる。しかしながら, この結果がデータに依存しない傾向であるかについては, さらなる検討が必要である。

6. おわりに

本論文では, 同一のサンプルを測定した複数の異なるマイクロアレイの遺伝子発現データを統合する手法を提案した。はじめに, NCBI GEO から入手することのできるマイクロアレイのデータの加工方法を説明した。次に, 各マイクロアレイの遺伝子の発現量分布の類似性に着目した提案手法1と, 同一遺伝子の発現量を直接用いる提案手法2の2つの手法を提案した。それぞれの手法で統合時のパラメータを変更しながらマイクロアレイデータの統合を行い発現量の順位に関する評価を行った結果, 発現量分布の類似性に着目した提案手法1の有効性が確認された。

一方で, 発現量の類似性に着目してマイクロアレイデータを統合するというアプローチが有効となる原因を完全に理解できているとはいえない。統合前のデータ間における発現量の順位には互いにゆるやかな相関のみがあるのに対し, それらのデータを統合した発現量の順位については統合前のすべてのデータと強い相関を持つ, という結果は自明ではなく, その原因を正確に理解する必要がある。提案手法によるマイクロアレイデータの統合を行うことで, 各マイクロアレイデータに含有するノイズ成分が相殺されるということがその一因として考えられるが, 本論文での結果のみからはこのことを断言することはできない。したがって, これを検証するためには, 提案手法を多様なマイクロアレイデータの統合に使用し, その妥当性を検証することが必要である。また, 提案手法により統合したマイクロアレイデータを統合に使用していない別のマイクロアレイデータと比較し, 相関を調査することも重要である。なお, 本論文では一色法のマイクロアレイのデータを統合する手法を提案したが, 二色法のマイクロアレイのデータも統合することのできる方法についても検討する必要がある。これらの検討は今後の課題とする。

謝辞

本論文を行うにあたり, 早稲田大学内田研究室の堀江輝樹嘱託研究員に多くの助言を頂いた。また, 本論文の一部は, 日本学術振興会における科学研究費補助金基盤研究(C)(課題番号 17K00135)による支援を受けている。ここに記し謝意を表す。

参考文献

[1] 株式会社 DNA チップ研究所: 遺伝子の発現量を調べる意味, <https://www.dna-chip.co.jp/diagnostic/>

rheumacheck/patient_new05.html.

- [2] 理化学研究所: 薬の効果がもたらす遺伝子発現の変化を網羅的・定量的に捉える, http://www.riken.jp/pr/press/2013/20131003_3/.
- [3] ThermoFisher: マイクロアレイとは? | マイクロアレイを用いた発現解析の原理と解析方法, <http://www.learningatthebench.com/start-microarray.html>.
- [4] Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., Baker, S. C., Collins, P. J., De Longueville, F., Kawasaki, E. S., Lee, K. Y. et al.: The MicroArray Quality Control (MAQC) project shows inter-and intraplatform reproducibility of gene expression measurements, *Nature biotechnology*, Vol. 24, No. 9, p. 1151 (2006).
- [5] Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G. et al.: Multiple-laboratory comparison of microarray platforms, *Nature methods*, Vol. 2, No. 5, p. 345 (2005).
- [6] Quackenbush, J.: Microarray data normalization and transformation, *Nature genetics*, Vol. 32, p. 496 (2002).