# 疾患原因解明基盤
# —遺伝子発現データクオリティ解析ツール—

## 1. 背景

近年、遺伝子に基づいたオーダーメード医療の実現が期待されているが、その基盤となるデータの信頼性は根本的に重要な問題である。例えば癌患者と健康な人で遺伝子発現がどのように異なるかをマイクロアレー技術を用いて解析し、癌患者特異的に働く遺伝子群を突き止めることによって、体質に応じた安全な抗癌剤の発見や最適な投薬計画、早期癌リスク診断システムの開発等が期待されている。しかし、その基盤となる遺伝子発現データは測定精度等の問題があり、現状ではデータの信頼性を客観的に評価する基準は確立されていないのが実情である。またノイズデータを除去（補正）する方法も一般的な指針はなく経験則やアドホックなルールによってなされている。

## 2. 目的

本プロジェクトでは、蓄積された過去のデータの統計解析に基づき発現解析データの客観的評価指標を構築し、精度の低いデータを除去（補正）し、解析の量と質のトレードオフを定量的に考慮した最適なフィルタリングパラメータを推定する方法を提案し、実用的なクオリティ解析プログラムを開発する。このプログラムはインターネット上に公開し WEB データベースに統合可能なものとする予定である。本研究開発の成果が基盤となり世界中の遺伝子解析データが標準化され、結果として病気や健康の問題で苦しむ多くの方々に希望を与えるものになれば幸いだと考えている。

## 提案の詳細

### 従来手法とその問題点

マイクロアレー解析は近年、広く普及し有望な技術であるが、測定精度の向上が大きな課題である。目視によるイメージデータの確認が一般的であるが、これには大きな手間がかかり、人によって基準が異なるという問題点がある。信頼性の劣るデータを同定し取り除く（或いは修正する）ことは、間違った解析結果をもたらすリスクを軽減する上で非常に重要である。

これまでマイクロアレー技術の開拓は、データマイニングを用いた統計解析により大量のデータから有用な知識を抽出することに主眼がおかれていたが、測定結果から質の高いデータを獲得する手法については必ずしも多くの研究はなされていなかった。その結果、最初の段階でのデータの品質が、解析の最終段階の解釈に多大な影響を及ぼすという問題が生じている。

マイクロアレーの解析では元になる測定データはスポットとよばれるイメージ（画像データ）である。ある条件下で働いている遺伝子は強く発光し、条件の違いによって発光強度が著しく変化する(log Ratio 値が大きい)遺伝子を検出することによって、遺伝子の働きを探ろうというのが基本原理である。しかし測定イメージデータの解析は必ずしも容易ではなく、実験上のノイズも含まれるため、信頼できるスポットを検出し、真のシグナルとノイズを分離し、信頼性に劣るデータを除去あるいは補正する技術が重要になってくる。しかしどういうデ

ータが信頼性の高い高品質なデータであるかという基準がなければ、このフィルタリングは難しい。

第一のアプローチは、スポットのイメージ属性を解析し信頼性の高いデータの基準を定める方法である。スポットは、フォアグランドの平均発光強度、バックグラウンドの平均発光強度、スポットの大きさ、形状、など数十の画素の属性で表現される。これらの属性は測定結果の信頼性に大きな影響を与える。例えば、発光強度が低すぎるとシグナルに対するノイズの割合が大きくなり測定結果が安定しない傾向がある。またスポット内の画素の発光強度分布の一様性やスポットの形状（真円度）の異常は、RNA や試薬の不良、不均一なハイブリダイゼーションなど、実験上、何らかのバイアスが働いている可能性を示唆する。イメージ属性に何らかの問題が認められるスポットは、フィルタリングの際に信頼性の低いスポットとして解析から除外あるいは補正されるべきであるが、イメージデータの属性は非常に多く複雑であり、どの属性を用いて、どの点をしきい値としてフィルタリングをすればよいのかは、自明ではない。発光強度やシグナルノイズ比などのスポット属性情報を用いた幾つかの判定基準や評価実験も報告されてはいる(Ritchie, BMC Bioinformatics 2006; Novikov, BMC Bioinformatics 2005; Ghosh, Bioinformatcs 2003)が、まだ研究途上にあり決定的なものはない。

目視によってスポットの良し悪しを判定する分類規則を明示的に規定することは難しいが、例えばベイジアンネットワークや Partial least squares (PLS)などの機械学習手法を用いてエキスパートが行う信頼性の高いデータの分類基準が獲得できれば、フィルタリングの自動化に応用できる可能性がある (Hautaniemi, et. al. Bioinformatics , 2003)。近年、例えば Trygg らが開発している MASQOT(Trygg et. al., BMC Bioinformatics, 2005)では、3人のエキスパートが8万スポットを良いスポットを悪いスポットに独立して判定した結果の解析をもとに、機械学習が行われ、主にスポットのシグナル強度や形状などの情報から自動的にスポットの Quality を判定する手法が提案されている。MASQOT は、大量のスポットの QUALITY を機械的に判定するシステムとして画期的であるが、エキスパートによって目視確認されたトレーニングデータの存在を前提としており、ユーザーが機械の自動判定結果を信用するかどうかという課題も残る。MASQOT の評価基準がすべての状況で妥当なものかどうかの確証はなく、評価基準の妥当性をいかに説得力あるものにするかが課題である。

第二のアプローチは、複数の同一遺伝子の測定データの分散をみることで QUALTIY を評価する方法である。理論的には、同一条件で同一の遺伝子を測定した複数の測定スポット(Replicate Spots)の測定結果は一致するべきであるが、実際には、測定時の実験環境（オゾンや温度など）、測定機器のバイアス、生物学的バイアスなどの様々な要因から、測定結果は一致するとは限らない。信頼性の高いデータとは、複数の測定結果の分散が小さいということであり、これをデータのクオリティの指標としてフィルタリングに用いるのである。この評価方法は、実際の測定データに基づいているので、客観的かつ実用的であるが、なぜ測定データの品質がよくないのかといった理由は説明することは出来ない。またどの属性に着目してフィルタリングを行うか、しきい値を何に設定すればよいのかという問いに関しては合理的な説明は出来ない。また1回しか測定できなかった遺伝子についてはスポットの良し悪しを評価することは出来ないという問題点がある。

**解析のトレードオフを考慮した Q-Score の提案**

前述したように従来研究においては、測定されたスポットの良し悪しをどう評価するべきかという点に重点がおかれていた。これに対し、本研究では、フィルタリングの定量的側面に注目し、フィルタリングはデータの品質とデータ量のトレードオフの問題であると捉える。情報理論的な視点で合理的なトレードオフカーブを定量的に理解し、適切なフィルタを選択することが重要である。過去の研究を調べてみても、トレードオフを考慮した定量的なフィルタリングアプローチは、提案者の知る限り、これまでなされておらず、本プロジェクトがはじめての試みである。

本プロジェクトでは、SMD 上の実際のデータを解析した統計をもとに、データの Quality 評価指標を開発する。SMD グループでは、同一の遺伝子を測定した複数の測定スポット（Replicate Spots）を用いて Quality のスコアを計算する”**Q-Score**”を提案している。”Q-Score”の計算式を図 1 に示す。基本的なアイデアは、同じ遺伝子を測定した複数の測定値(log raito 指標を用いる)の分散を見ることでそのデータの QUALITY を評価しようという考え方に基づいている。 Replicate Spots の測定値(log ratio 指標)の分散を Quality 評価指標として用いること自体は、斬新ではないが、オリジナルな点は、データをフィルタリングしていく過程で Q-Score がどのように変化していくかを動的に追うところにある。

図 1,図 2 に Q-Score の定義とダイナミクスの例を示す。この例では、測定スポットの属性の一つ(例えば Regression Correlation)を用いてフィルタ値を変化させながらフィルタリングしていく過程で、Q-Score とフィルタリング後に残されたデータの割合(Fraction)の関係がどう変化していくかを表している。興味深いことに、非常に多くの場合、Q-Score のグラフは下に凸のカーブ状をしている。つまり、最初の段階では急激に Q-Score が下がっていくが、ある程度から Q-Score は下がらなくなっていく。この Q-Score の変曲点をフィルタのしきい値として用いれば、測定アレイごとに、測定精度に応じた適切なフィルタリングを行うことが可能となる。

$$\mathbf{Q-Score} = \sqrt{\frac{\sum_{k\ 1}^{m}\left(\sum_{i\ 1}^{n_k}\sum_{j\ i+1}^{n_k}\left(x_{ki}-x_{kj}\right)^2\right)}{\sum_{k=1}^{m}\frac{n_k\times\left(n_k-1\right)}{2}}}$$

図 2　Q-Score の計算式
$n_k$: ある遺伝子 k の重複測定回数 m: 重複測定した遺伝子の数 $x_{ki}$:(重複測定遺伝子 k の i 番目の測定データの log ratio)-（重複測定遺伝子 k の平均 log ratio)

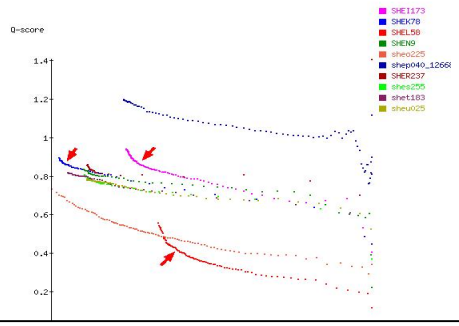



図 1　Q-Score の変化の例　(Q vs Fraction graph)
横軸は Fraction(あるフィルタ属性を用いてフィルタリングした後に残ったデータの割合)、縦軸は Q-Score、スライド毎に Q-Score のカーブをプロット。変曲点（赤矢印）が認められる。

## 3. 開発の内容

本プロジェクトでは様々な疾患の原因解明の基盤となる遺伝子発現データのクオリティ解析、及びフィルタリングシステムを開発した。遺伝子発現データの品質を評価する指標と

して QScore を提案し、様々なフィルタに対して QScore に基づき解析の量と質のトレードオフを定量的に考慮した最適なフィルタリングポイントを自動抽出する手法を提案し、クオリティ評価結果の可視化機能、自動フィルタリング機能、フィルタリング境界領域データの表示機能、及び GUI を実装した。本プロジェクトで開発した“遺伝子発現データクオリティ解析システム”は下記の機能を実装している。

**マイクロアレーデータの入力インターフェース機能**

測定したマイクロアレーの画像データ及び測定属性データをファイル、又はデータベースから取得するインターフェース。

**マイクロアレーデータのクオリティ評価機能**

マイクロアレーデータのクオリティ評価指標の開発。スポット毎のクオリティスコア及びアレイ毎のクオリティスコアの提示機能の開発。また複数のクオリティ評価指標の整合性を比較検証しクオリティ解析結果のサマリーを提示するインターフェース。

**マイクロアレーデータのフィルタリング機能**

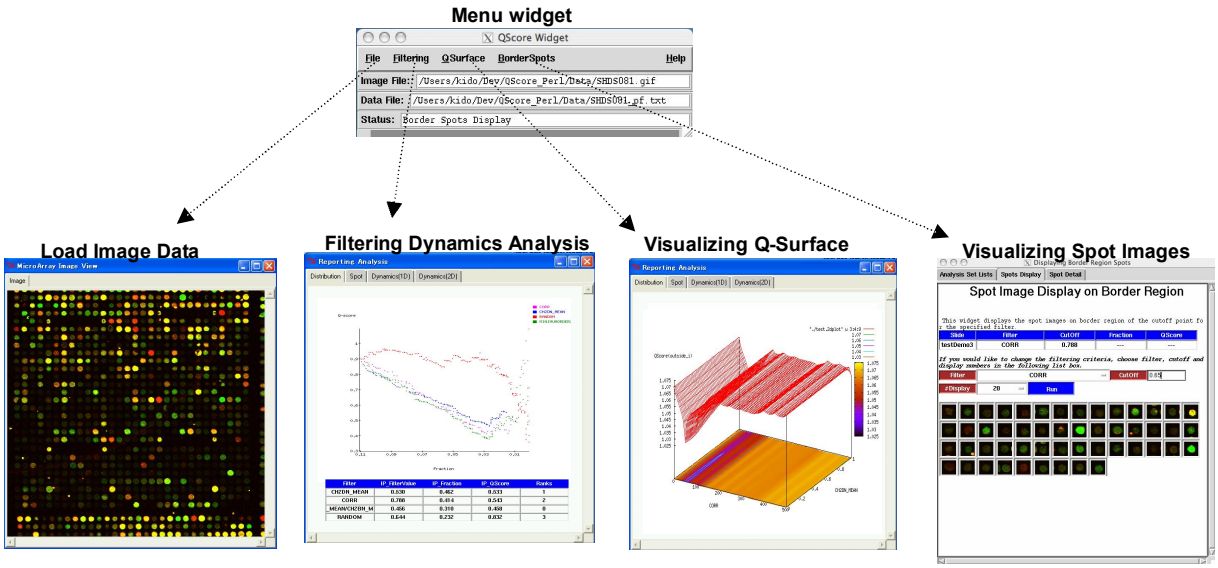
フィルタリングパラメータを動的に変化させ、除外するデータ量とデータのクオリティのトレードオフカーブの変曲点を解析することにより、フィルタリングカットオフの最適推奨値を自動抽出するアルゴリズム、多次元のフィルタパラメータの相互作用を解析し、複数フィルタの組み合わせによるフィルタリングを可能にするアルゴリズムの開発。解析から除外すべきスポットの提示機能、その後の解析に有用なクオリティ情報提示機能。

**クオリティ解析結果の可視化機能**

QScore グラフ(QScore-Filter, Fraction-Filter, QScore-Fraction)の表示機能、多次元の QSCORE Surface の可視化機能、マイクロアレー画像データとクオリティ解析結果をインタラクティブに対応づけて解析結果を提示する GUI。

**クオリティ指標の比較評価実験データの集積、及び、表示機能**

マイクロアレーデータの品質を評価するための有用な指標を開発するために、様々な指標属性の比較、属性間の相互作用や特性、フィルタリングへの効果などについて実データを用いた実験による統計的解析結果の蓄積。蓄積された解析データを表示するための GUI。




本システムの実装には主として perl 言語を用い、Stanford Microarray Database

Group で開発された Perl ライブラリモジュールとデータフォーマットを利用した。解析データの可視化には Perl GD モジュール及び gnuplot を用いる。GUI には、Perl TK を用いた。(詳細は添付のポスター論文資料を参考のこと。)

## 4. 従来の技術(または機能)との相違

### 1. 遺伝子発現データのクオリティを評価する QScore 指標

遺伝子発現データのクオリティ評価に関する研究はその重要性にもかかわらず、これまで客観的で体系的な研究が十分に報告されておらず研究者の主観的な評価に頼ることが多かった。本プロジェクトで実装された QScore 指標を用いることで、遺伝子発現データの品質を主観に依らず同一基準で横断的かつ客観的に比較することが可能になった。

### 2. 様々なフィルタのフィルタリング特性の比較とカットオフポイントの推定

一般的なマイクロアレーの解析ではスポットとよばれるイメージ(画像データ)が元になっているが、スポットには数十(30 程度)の属性があり、この属性値をフィルタとして数多くのフィルタリングがなされてきた。しかし、どのフィルタを用いるべきか、どこをフィルタリングのカットオフ値とすべきかについては客観的なコンセンサスがなく、歴史的慣習や個人の主観に頼っていた。本システムで実装されたフィルタリングダイナミックスの比較グラフの生成機能とカットオフポイント抽出機能を用いることで、多くのフィルタの中から最も効率的に QScore 値を改善するフィルタを客観的に見出し、マイクロアレーデータの特性に応じて適切なフィルタとカットオフ値を定量的に選択することが可能になった。

更に我々は複数のフィルタを組み合わせて、最も効果的かつデータの特性に応じて適応的に最も効果的なフィルタリングを行うハイブリッドフィルタを自動生成する手法を検証している。現時点ではハイブリッドフィルタの生成機能は提供していないが、2つのフィルタと QScore の関係を3次元で表示するインターフェースを提供している。またフィルタリングの境界領域に属するスポットイメージとその属性情報を表示するインターフェースも提供している。これらの機能はオリジナルであり、筆者の知る限り、同様な機能を有するシステムは現存しない。

### 3. PC 上で稼動するコンパクトな解析結果表示機能と GUI

Stanford Microarray Database (SMD) はアカデミックとしては世界最大であり、様々なマイクロアレーデータ解析ツールを Web 上で提供しているが、対象ユーザは基本的には SMD にアカウント登録をした個人または研究グループである。SMD の Web データベースシステムはオープンソースで公開されているので、SMD と同様な Web ベースのシステムを立ち上げることは可能ではあるが、これは個人にとっては敷居が高い。本プロジェクトで実装された解析結果表示機能と GUI を用いることで、SMD ユーザでなくても、本プログラムをインストールすれば、個人の PC 上でマイクロアレーデータのクオリティ

解析を行うことが可能になった。

## 5. 期待される効果

本プロジェクトでは様々な疾患の原因解明の基盤となる遺伝子発現データのクオリティ解析、及びフィルタリングシステムを開発した。遺伝子発現データの品質を評価する指標としてQScoreを提案し、様々なフィルタに対してQScoreに基づき解析の量と質のトレードオフを定量的に考慮した最適なフィルタリングポイントを自動抽出する手法を提案し、クオリティ評価結果の可視化機能、自動フィルタリング機能、フィルタリング境界領域データの表示機能、及びGUIを実装した。

本研究開発の成果が基盤となり世界中の遺伝子解析データが標準化され、結果として病気や健康の問題で苦しむ多くの方々に希望を与えるものになれば幸いだと考えている。

## 6. 普及(または活用)の見通し

1. Stanford Microarray Database
   新たに開発したモジュールは Stanford Microarray Database (http://genome-www5.stanford.edu/) に組み込み、Web上のサービスとして提供したいと考えている。
2. **論文**
   ICSB07 (International Conference on Systems Biology, 2007)においてポスター発表を行った。今後、クオリティ解析やフィルタリング手法に関する研究データの詳細をまとめてバイオインフォマティクス関連のジャーナルに投稿する予定である。
3. **標準化**
   MGED 会議などの場においてマイクロアレーデータのクオリティ指標やデータ標準化の提案に貢献していきたいと考えている。
4. **パートナーシップ**
   ベンチャー会社や研究機関とも連携し開発した技術やソフトウェアの普及に努めたい。

## 7. 開発者名(所属)

城戸　隆 (スタンフォード大学)

(参考文献,URL)

- Takashi Kido, et. al, “Quality Assessment of Microarray data and Optimal Filtering Criteria”,International Conference on Systems Biology, 2007, Poster Paper
- http://genome-www5.stanford.edu : Stanford Microarray Database