



2006 年度下期未踏ソフトウェア創造事業 採択案件評価書

1. 担当PM

北野 宏明PM(ソニーコンピュータサイエンス研究所)

2. 採択者氏名

開発代表者:城戸 隆 (スタンフォード大学 客員研究員)
共同開発者:なし

3. プロジェクト管理組織

有限会社アカデミア

4. 委託金支払額

5,492,877 円

5. テーマ名

疾患原因遺伝子解明基盤:遺伝子発現データクオリティ解析ツール

6. 関連Webサイト

なし

7. テーマ概要

近年、遺伝子に基づいたオーダーメイ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。

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な評価指標を構築し、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。

8. 採択理由

マイクロアレーの品質という重要な問題に着目し、突破口を開こうと考えている。学術的な内容であるが、この分野での標準化やツールの提供は重要な課題であり、何らかの進展が期待される。

9. 開発目標

本プロジェクトでは、遺伝子発現データのクオリティを評価する指標として QScore を提案し、QScore に基づく遺伝子発現データのクオリティ評価とフィルタリング機能を開発する。QScore を効率的に改善する様々なフィルタの特性やフィルタ間の相互関係を調べるためのツールを開発し、データの特性に応じて適応的にフィルタを組み合わせ最適なフィルタリングを行うための手法、解析の質と量のトレードオフを考慮して最適なカットオフ値を自動抽出する手法等を開拓することを研究目標とする。手法開拓の過程で実装したツール、プログラムモジュール、及び GUI を本プロジェクトの成果物として報告する。本プロジェクトで開発する“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することを開発目標とする。

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の開発。

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の開発。

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の開発。

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 の開発。

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 の開発。

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する。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 を用いる。

10. 進捗概要

本プロジェクトは、(1)探索的なアプローチによる手法開発やシミュレーション実験、(2) GUI やプログラムモジュール等を含むシステム開発 を平行して進めている。(1) に関しては新たなアイデアに基づいた手法提案や評価実験の結果が蓄積されて

きているが、探索的研究と検証を現在も継続して行っており詳細は別途、論文としてまとめる予定である。**本報告では、未踏ソフトウェアプロジェクトの趣旨に沿い(2)のシステム開発成果を中心に報告する。**

開発したシステムは、perl で実装されたプログラムモジュール群及び、perl/TK で実装された GUI からなる。GUI は、Windows 2000 (cygwin) 及び Machintosh (OS 10)で動作検証を行っており、perl, perl/TK, gnuplot, perl GD ライブラリがインストールされた個人の PC 上で単一のマイクロアレースライドの解析がコンパクトに実行できることを主眼においている。

プログラムモジュール群は、個人の PC だけでなく、例えば、Unix サーバ上で実装された Web サービスの一部として組み込まれることも想定している。イメージとしては、大規模なマイクロアレーデータセットの解析をユーザが WEB を介してバッチ処理で依頼し、サーバ上での解析終了後に解析結果がユーザーに e-mail で通知され、ユーザーが結果を Web で参照するというものである。このシステムはすでに Stanford Microarray Database (SMD)上のサービスとして稼動している。現在、オフィシャルな公開サイトで稼動しているシステムは旧バージョンのプログラムモジュールを用いているが、本プロジェクトで新たに開発された機能モジュールもいずれアップデートしていく予定である。

以下では、前段で述べた開発目標機能がどこまで実現出来たかについて簡潔に述べる。

マイクロアレーデータの入力インターフェース機能

マイクロアレーのイメージデータファイル及び測定属性データファイルをシステムにロードしオブジェクト化するプログラムモジュールを実装した。また PC 上で稼動する GUI を実装した。

マイクロアレーデータのクオリティ評価機能

前述した入力インターフェース機能により生成されたアレイデータオブジェクトに対して QScore の計算やフィルタ統計処理、新規フィルタの作成、フィルタ間及び QScore との整合性の評価などを行うプログラムモジュールを実装した。またフィルタリングを行うのに必要な PCL オブジェクトを生成する機能を実装した。

マイクロアレーデータのフィルタリング機能

前述したクオリティ評価機能により生成された PCL オブジェクトをもとにフィルタ値を

変化させながらそれぞれのフィルタに対しフィルタリングを実施しデータを収集するプログラムモジュール、このモジュールから得られたフィルタリングデータを解析しカットオフポイントの抽出とフィルタのランキング計算を行うプログラムモジュール、2次元の QScore Surface をプロットするために必要なデータを計算するプログラムモジュールを実装した。また PC 上で稼動する GUI を実装した。

クオリティ解析結果の可視化機能

前述したフィルタリング機能により生成されたプロットデータをグラフ化し Gif ファイルを生成するプログラムモジュールを実装した。散布図の作成には GD ライブラリと SMD で提供されているプロットモジュール、3次元図の作成には Gnuplot を用いた。

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

前述したフィルタリング機能や可視化機能によって生成されたデータやグラフ、イメージデータにアクセスし解析結果を表示する GUI を実装した。

11. 成果

(1) システム全体構成

以下では図表を用いて、本システムの目的、設計方針と基本的なアイデア、動作環境、処理フロー、入力データ、出力について説明する。

【目的】

このシステムはスポット型マイクロアレーを用いて遺伝子発現解析を行う研究者や研究グループ、また遺伝子発現解析の結果を診断や創薬に応用するために解析データの検証や解析クオリティの評価を行いたいと考えている方々に対して提供する遺伝子発現データのクオリティ解析ツールである。従来、主観に頼っていたクオリティの評価を出来る限り客観指標を定義して比較評価可能な形で提示し、データフィルタリングを通して、より解析精度を向上させるのに有益な情報を提供することを目的とする。

【設計方針と基本的なアイデア】

前述した QScore をマイクロアレーの各スライドデータに対して計算する。またどのフィルタがフィルタリング効率が高いか直感的に把握するために同一座標上に様々なフィルタリングカーブを表示するグラフを生成する。

適切なフィルタを選択すれば品質のよくないスポットが最初に除去され QScore 値が改善していくが、その改善効率はだんだんと小さくなっていくことが予想される。ど

の地点をフィルタリングのカットオフとすべきかは、解析の量と質のトレードオフの問題となるが、このシステムでは各フィルタに対して、フィルタリングカーブの変曲点を抽出しユーザに提示する。あらゆるケースで変曲点が明確に出現するとは限らないが、明確な変曲点が存在する場合には、効果的なカットオフ値の指標となる。また明確な変曲点が存在しない場合でも変曲点が明確に現れないという情報はクオリティを吟味する上で参考になる。従来、よく利用されてきたフィルタや慣習的なカットオフ値と比較することで、従来よりもより客観的でデータの特性に合ったフィルタリングを行うことが可能になる。

また既知のフィルタよりも改善効率が高く適切なクオリティ指標となるフィルタを見出すために複数の既知フィルタと QScore の関係を表示するグラフを生成する。この詳細については後述する。

更に実際のマイクロアレーのスライドイメージやスポットと本システムで表示するクオリティ指標が容易に対応づけられるように、カットオフ値周辺のフィルタリング境界に位置するスポットイメージを表示できるようにする。

【動作環境】

本システムは、MAC OS10, 及び Windows2000 (Cygwin)上で開発し動作確認を行っている。Windows2000 と MAC OS10 では GUI におけるフォントや色などに若干の違いがあり、設定を変える必要があるが提出したソースコードは MAC OS 10 上で動作検証をしたものである。

また本システムは Perl と Perl/Tk を用いて実装されており本システムを稼働させるには、Perl 5.0 以上、及び Perl/Tk をインストールする必要がある。またグラフ作成に Perl GD ライブラリを用いており、これもインストールが必要である。(Perl GD ライブラリに必要なライブラリをプラットフォームに応じてインストールする必要があるが、詳細は別紙の簡易マニュアルを参照のこと。)QScore Surface の3次元プロットの可視化には Gnuplot を用いており、これもインストールが必要である。

クオリティ解析の対象となるマイクロアレーのデータはスポットごとに様々なフィルタ指標の数値を表記したテキストデータでありサンプルデータも添付に含めている。データのフォーマットについては後述するが Genepix 仕様を基本としている。スポットイメージやアレイイメージを表示するための画像イメージも必要となるがこのサンプルも添付した。これらのサンプルデータやイメージは、Stanford Microarray Database のサイトで多くの解析事例が蓄積されており、すでに論文かされた多くの解析データにアクセスすることが可能である。

【処理フロー】

図 1 に本システムのデータ処理の流れと基本的な機能モジュールを示す。まず本

システムはマイクロアレー測定データファイルと測定イメージファイルを読み込み**入力インターフェース機能モジュール**がデータ処理を行い測定データをオブジェクト化する。Web サービスとして本システムを組み込む場合には、データベースへのアクセスモジュールを用いて測定データを取得するが本報告書では PC 上に測定ファイルが存在するという仮定のもとで解析を行う利用形態を想定する。通常、マイクロアレーの測定データファイルはスライド(あるいはアレイ)と呼ばれる単位でファイル化され一つのファイルには4万スポット(1つのスポットは1つの遺伝子発現の測定に対応する)程度の情報が格納されている。スライドの数は研究プロジェクトによって異なるが、実験条件の異なるスライドが数十程度ある場合が多い。

次に**クオリティ解析評価機能モジュール**が、QScore やその他のクオリティ指標や統計値を計算する。またクオリティ解析機能モジュールは、複数のフィルタ(あるいは複数のスライド)の情報をひとまとめにして PCL オブジェクトを生成し次の処理に手渡す。

本報告では PCL オブジェクトは複数のフィルタをまとめているものと仮定する。

フィルタリング機能モジュールは PCL オブジェクトを受け取り、各フィルタごとにフィルタパラメータを連続的に変化させフィルタリングのダイナミクスを調べる。複数のフィルタのフィルタリングダイナミクスを比較し、QScore の改善効率をもとにフィルタのランキングを計算する。また各フィルタリングカーブを解析し QScore の改善効率の変曲点を抽出する。また複数のフィルタを組み合わせた場合の QScore の改善効果を調べるために2つのフィルタと QScore 値を3次元空間上で可視化するためのデータを生成する。

解析結果可視化機能モジュールはクオリティ解析やフィルタリング解析で得られたデータを可視化して様々なグラフ(Gif ファイル)を生成する。またマイクロアレーのイメージデータから、フィルタリングの境界領域にあるスポットイメージを抽出する。

解析結果表示機能モジュール(GUI)は前述したモジュールによって生成されたグラフやデータを整理し、簡易な操作で表示するためのインターフェース(GUI)を提供する。

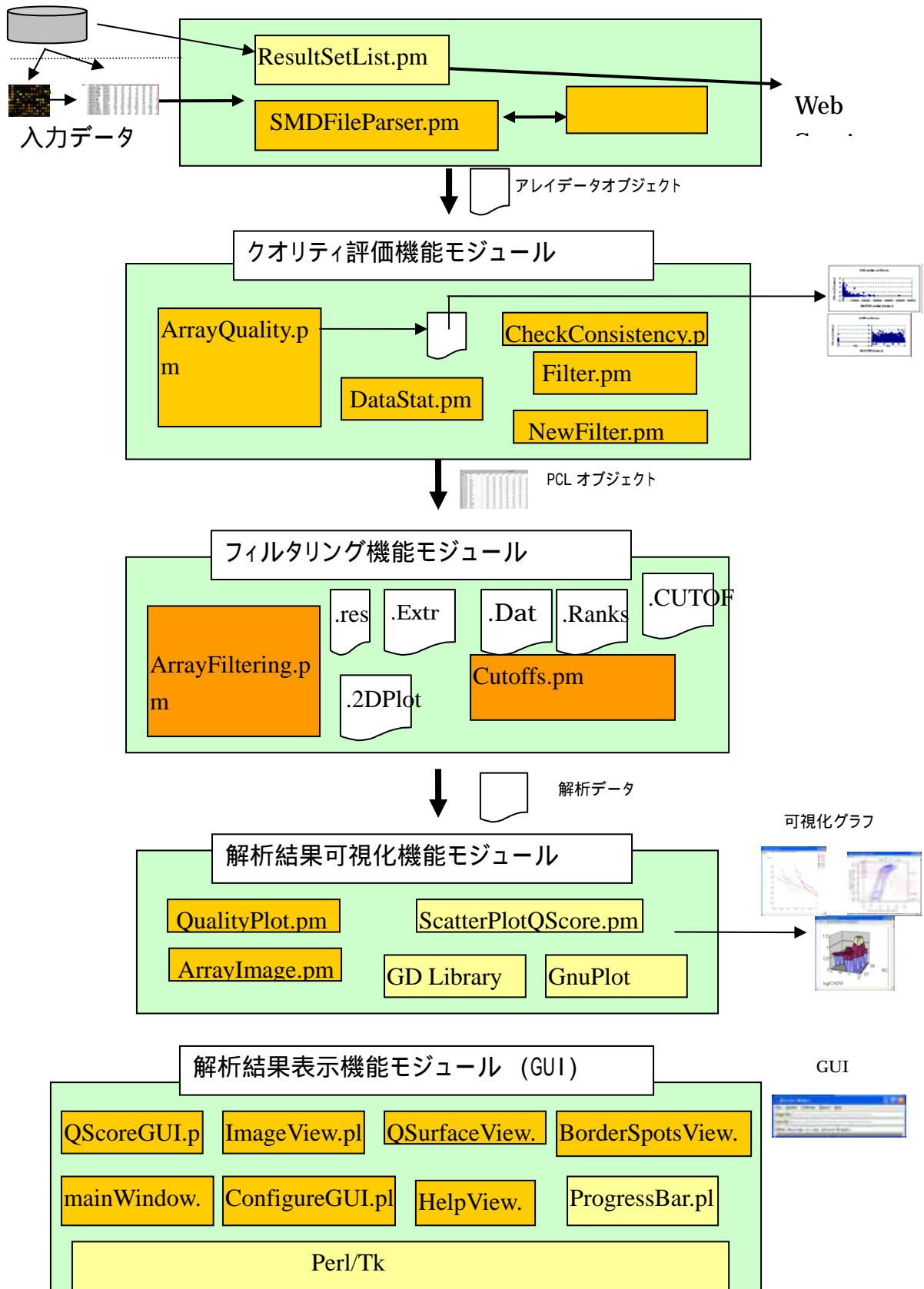


図 1 システム処理フロー

【入力と出力】

入力： 本システムの入力は、マイクロアレーのデータファイル及びイメージファイルである。マイクロアレーのデータはスライドとよばれる単位でファイル化されており、一つのスライドには約4万程度の遺伝子のスポット測定情報(遺伝子の重複を含む)が含まれているのが一般的である。通常、ある研究目的のために実験条件を変えて測定した数十のスライドファイルがあることが多い。イメージファイルは各スライドごとに、この約4万程度の遺伝子スポットの発現を画像情報として記録した Gif 形式の画像ファイルである。スタンフォード方式の2チャンネルのスポット型マイクロアレーでは、各スライドごとに Ch1(Red)と Ch2(Green)の発光強度を記録した白黒の Tiff ファイルがあり、これが合成されて1つのカラーの Gif ファイルが出来ている。

出力： 本システムの出力は、解析結果の各種グラフ及びデータファイルである。詳細については後述する。

(2) マイクロアレーデータの入力インターフェース機能

以下では図表を用いて入力インターフェース機能の目的、設計方針、実装、動作例について説明する。

【目的】

この機能モジュールは解析の対象となるマイクロアレーデータのフォーマットをチェックしシステムにロードする。また GUI でイメージやスポットを表示する際に必要となるイメージデータもロードする。これらのデータはオブジェクト化されクオリティ解析モジュールへ渡される。

【設計方針と基本的なアイデア】

マイクロアレーのデータはデータベースまたはファイルに記録されていると仮定する。Stanford Microarray Database ではデータベース中から特定のプロジェクトのデータを抽出するプログラムモジュールがあるが、本システムでは SMD ユーザ以外の方々もこのツールが利用出来るようにするためにファイルベースのモジュールを新たに実装する。一つの測定スライドごとにデータファイルとイメージファイルがあるものと仮定する。

本システムの利用形態としては、個人が PC 上で1アレイ事のクオリティ解析を容易

に出来るようなコンパクトなインターフェースと、サーバ上で大量の(例えば数百スライドの)アレイデータを対象にしてバッチ形式で統計処理するものの両方を想定している。複数(3~10程度)のフィルタを用いた場合、1つのアレイのクオリティ解析には数分から数十分かかると想定されるので、大量のアレイデータのバッチ処理をPCで行うのには向いていない。後者についてはStanford Microarray Database上のWebサービスとして実現することとし、本報告では前者の利用形態を前提に説明する。クオリティ解析の対象となる入力データフォーマットと例を図2に示す。このデータフォーマットはStanford Microarray Databaseで採用されているもので広く普及しているGenepixのファイル形式に準じている。

入力データの形式:

測定データ

マイクロアレーの測定データは2次元(N x M)のテキストデータを格納したファイルであり、N個のスポットに対してM個のスポットの属性値が計測されている。(図2)2次元テキストデータの一行目はヘッダ情報であり、属性名が記述される。2行目以降は測定データでそれぞれのスポットの属性値が記述される。

一枚のスライドで計測するスポット数Nは実験によっても異なるがヒトの遺伝子発現を対象とする場合4万程度になる場合が多い。スポット属性Mの数は30程度ある。その中には計測したスポットの遺伝子名に対応するclusterIDやスポットの発現活性を判断する指標となるlog Ratio値なども含まれる。またスポットのアレイイメージにおける座標もこのテキストファイルに記述されているものとする。

イメージデータ

これはマイクロアレーの測定結果の画像イメージである。広く普及しているスポット型のマイクロアレー(スタンフォード方式)の場合、Ch1(red)とCh2(green)の二つのチャンネルの発光強度が計測されそれらが合成されてカラーイメージ(gif file)が生成される。(図3)赤く光っているスポットは遺伝子の発現活性が通常より高く、緑のスポットは通常より発現活性が低いというように解釈されるが、この色情報に対応する2つのチャンネルの発光強度の比(log比)はratio値(log ratio値)として発現解析を行う際の中心的な指標となる。測定されたスポットの形状や色をイメージとして目視することは測定結果のクオリティを直感的に把握する上で有益であり、実際に多くの実験ではエキスパートによるスポットの目視確認によるフィルタリングが行われている。

フィルタ属性の例

マイクロアレーの測定データにはスポットの発光強度やバックグラウンドの発光強度、スポットの大きさなど30程度の測定指標があり、そのうちのいくつかはスポットのクオリティ属性値或いはフィルタ指標として用いられる。マイクロアレー スキャナーの

Genepix のファイルフォーマットは広く普及している。下記に Genepix のファイルフォーマットやフィルタ属性について解説してある URL を示す。

http://www.moleculardevices.com/pages/software/gn_genepix_file_formats.html

SMD のデータベースも Genepix のフォーマットに対応している。フィルタリングのフィルタとしてよく用いられるスポット属性には、たとえば、Ch1 Net, Ch2 Net, Regression Correlation, (Ch1 Foreground Intensity)/(Ch1 Background Intensity), (Ch2 Foreground Intensity)/(Ch2 Background Intensity) などの指標がある。これらの指標はそれぞれのチャンネルについて、フォアグラウンドとバックグラウンドの発光強度の差(Net)や、比をあるいは相関係数を評価指標とするものである。

これらの指標のどの値をカットオフ値としてフィルタリングするべきかは、一般的なコンセンサスはなくヒューリスティックや主観的な判断で行われている。歴史的な理由や慣習で、Regression correlation > 0.6, Net Intensity > -350, Foreground_Intensity / Background_Intensity > 2.5 などがよくカットオフ値として用いられるが、これらのカットオフ値を用いる科学的根拠があるわけではない。これらのカットオフ値に解析的意味づけを与えることは本研究の目的の一つである。

1	Spot	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
2	Gene ID	Gene Name	Gene Name	Gene Name	Gene Name	Gene Name	Accession	Preferred	Locustalk	Gene	Sequence	X Grid	Y Grid	Color	Fluor	Plate	Plate	Plate	Plate	Plate	Plate	Plate	Plate	Plate	Plate	Plate	Plate
1	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
2	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
3	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
4	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
5	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
6	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
7	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
8	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
9	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
10	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
11	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
12	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
13	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
14	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
15	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
16	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
17	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
18	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
19	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
20	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
21	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
22	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
23	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
24	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
25	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
26	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
27	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
28	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
29	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
30	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
31	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
32	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
33	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
34	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
35	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
36	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
37	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
38	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
39	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
40	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
41	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
42	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
43	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
44	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
45	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
46	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
47	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
48	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
49	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
50	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA	21	1	1	0	37225 D	1	1	1	1	1	1	1	1	1	1	1	1	
51	IMAGE017032	Integrin beta3	Ha025927	HS2291			2629	18254	CDNA																		

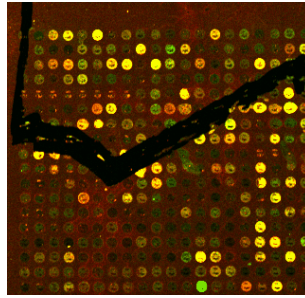
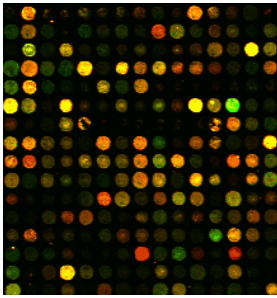


図 3 マイクロアレイイメージデータの例

【実装】

本システムでは下記のモジュールを実装(利用)した。実装言語は Perl である。プログラムの保守性と拡張性を考慮して各モジュールはオブジェクト指向で実装した。Perl を採用したのは、Stanford Microarray Database (SMD)グループで開発した Perl モジュールが再利用出来ることと、プログラムの開発効率を上げるためである。

SMDFileParser.pm

このモジュールはマイクロアレイの測定データ(一つのスライドデータに対応した Genepix 仕様の測定ファイル)を読み込み、アレイデータオブジェクトを生成する。主要メソッドである `_parseFile()` は入力ファイルの形式をチェックし全てのスポットのそれぞれのフィルタ属性対応するフィルタ値をオブジェクトに格納する。

NameMap.pm

このモジュールは測定ファイルのヘッダーに記述されたフィルタ名(スポット属性名)をチェックし、あらかじめ定義されたフィルタ名の変換テーブルを用いて、本システム上で利用するフィルタ名に変換する。

ResultSetList.pm

このモジュールは SMD グループで開発されたデータベースアクセス用のモジュールであり、Web サービスに本システムの機能を組み込む際にはこのモジュールを用いてデータベースからデータを取得する。

【実行結果例】

図 4 にイメージデータをシステムにロードし GUI ウィジェットへ表示した示す。テキストデータも同様にシステムにロードするが、ロード時間を短縮するためウィジェットへの

表示はデフォルトでは行っていない。GUI については、後述する解析結果の表示機能 (GUI) モジュールの項でまとめて説明する。

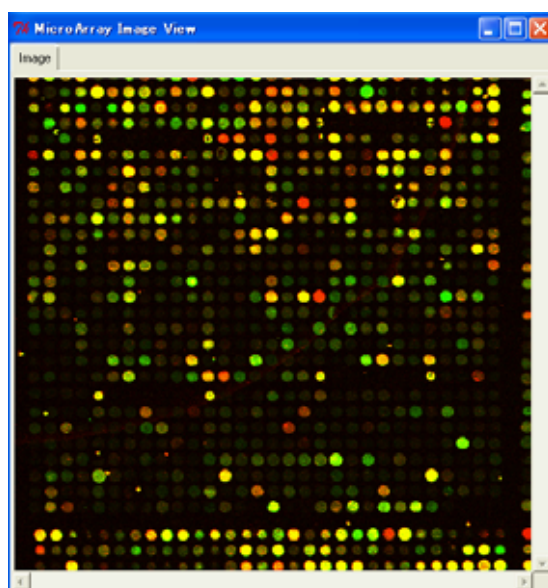


図 4 イメージデータのロード例

(3) マイクロアレーデータのクオリティ評価機能

以下では図表を用いてクオリティ評価機能の目的、設計方針、実装、動作例について説明する。

【目的】

この機能モジュールは前述したマイクロアレーのデータオブジェクトを受け取り、各フィルタに対して QScore(及び他のクオリティ指標値)を計算する。またその他のクオリティ解析に有益なデータやフィルタ間での相互関係を理解するのに有用な統計データを計算する。スポットのクオリティ属性を評価する様々な指標値を定義しそれらを新規フィルタとして登録する。またフィルタごと(あるいはスライドごと)のデータを一つのセットにしてまとめた PCL オブジェクトを生成し、後述するフィルタリングオブジェクトへ渡す。

【設計方針と基本的なアイデア】

このモジュールでは QScore を基本として、クオリティ解析に有用と考えられる様々な指標値や統計データを計算する。具体的には、マイクロアレーデータのクオリティ評価指標、スポット毎のクオリティスコア指標、また複数のクオリティ評価指標の整合性を数値化した指標などがある。下記に代表的なアイデアについて簡単に説明する。

クオリティ評価指標

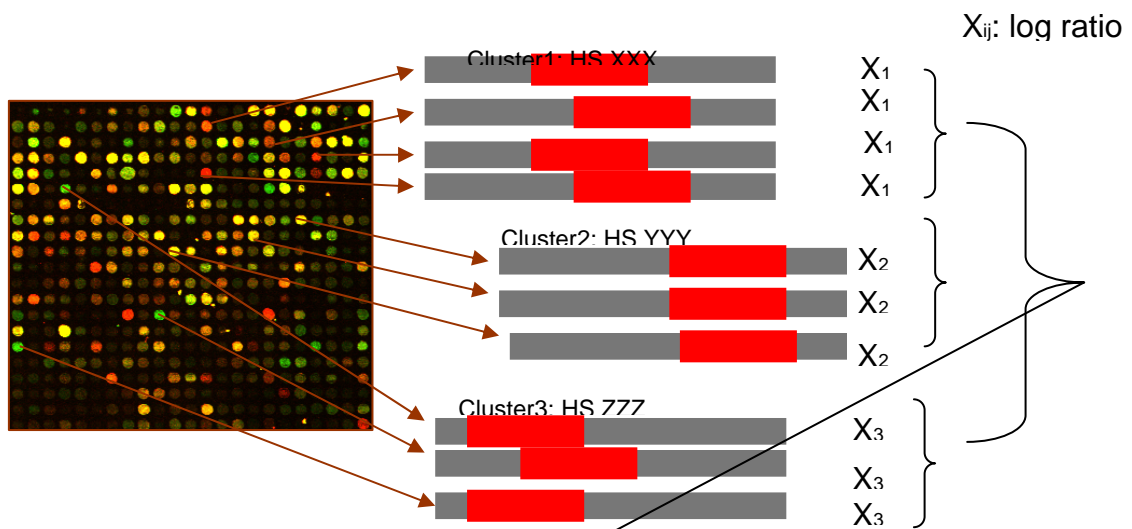
アレイデータのクオリティ評価には前述した QScore を用いる。図 5 に QScore の定義と説明図を示す。QScore の基本的なアイデアは、同じ遺伝子を複数回測定した重複測定スポット(Replicate Spots)の log Ration 測定値の分散を用いるというものである。QScore の定義式には幾つかのヴァリエーションが可能だが、基本的な発想は同じでありどのヴァリエーションを用いても、そのダイナミクスは極端には変わらない。

遺伝子の種類を示すアノテーションとして clusterID という属性値があり QScore を計算する際には同一の clusterID を持つスポット群(重複測定スポット: Replicate Spots)を最初にまとめる必要がある。次にそれぞれの clusterID について定義式に基づいて分散指標値を計算したものを平均して正規化したものを一つの実数値指標にする。図 6 に clusterID ごとに2つのフィルタ(Regression Correlation, Ch2N)と Log Ratio 値, QScore 値を並べて表示した例を示す。

QScore を計算するには、重複測定スポットが存在することが前提となる。SMD のデータベースには数多くの発表論文の測定データが蓄積されているが、一つのスライドには約4万スポット程度の発現情報があり、そのうち平均して約半数以上が重複測定スポットである場合が多い。(具体的な統計については本報告では詳述しない。)

重複測定数は2から5程度のものが多いがヴァリデーション用に数百の重複測定を行う場合も稀にある。SMD のデータベースには、人の遺伝子を測定したものも多いが、E Coli などのデータを扱う場合などには、隣接する3つの遺伝子が同じたんぱくをコードするケースがある。このようなケースでは、重複測定数を3倍程度に増やすことも出来るか可能性がある。

スポットの平均重複測定数kが大きければクオリティ判定の信頼性も向上すると考えられるが、どの程度のkがあれば十分な信頼性を確保できるかは重要な知見である。Ghoshらは彼らが提唱したqcomと呼ばれるスポットのクオリティ評価指標の信頼性を重複測定スポットを用いて定量的に議論している。(Wang X. and Ghosh S. et.al, *Bioinformatics* Vol. 19 no. 11 2003)このような理論的考察については、別途、論文として形にまとめていきたいと考えている。



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

m: number of cloneIDs with more than 1

図 5 QScore の定義

重複測定したスポットID (replicate)

遺伝子名に対応する **測定結果の Log** **clusterID ごと**

ClusterID	IDList (Key)	RC List	CH2N List	LogRatio List	STD(Log)	CV(LogF)	Qscore	log
Hs.445358	29661 26613	0.75 0.83	220 577	-5.21 0.896	3.053	-1.41 539	6.106	
Hs.537278	1699 41420 2612	0.63 0.94 0.13	165 3498 11	-0.18 4.366 -2.72	2.933422	6.02758	5.080836	
Hs.2420	25508 10046	0.64 0.56	687 130	3.886 -1.182	2.534	1.87426	5.068	
Hs.295459	11213 29670	0.18 0.64	65 171	0.145 -4.735	2.44	-1.06318	4.88	
Hs.584806	35142 18035	0.96 0.97	3701 62283	0.111 4.986	2.4375	0.956445	4.875	
Hs.561411	19116 42825	0.62 -1	128 10	-2.181 2.574	2.3775	12.09924	4.755	
Hs.25647	26032 3725 14356	0.37 0.85 0.6	70 426 1192		-7.433	2.712349	-1.09472	4.697926
Hs.494178	3085 34906	0.26 -1	57 3	-3.85 0.837	2.3435	-1.55559	4.687	
Hs.643432	27563 4084 40766	0.33 0.76 0.88 0.51	71 497 902 88	0.273 2.217 2.284	2.827318	104.7155	4.616991	
Hs.151219	7226 33798	0.77 0.91	319 2677	-0.292 4.323	2.3075	1.144877	4.615	
Hs.245044	22849 20943	0.93 0.69	4672 198	3.301 -1.261	2.281	2.236275	4.562	
Hs.221127	30367 15121 1227	0.04 0.72 0.26	1 225 67		-7.583	2.627524	-1.03951	4.551006
Hs.76224	31035 13496 3285	0.59 0.93 0.6	140 17843 195	0.225 6.456 1.969	2.624671	0.91029	4.546063	
Hs.213424	16185 17950 3631	0.79 0.76 0.38	727 207 117	0.218 -3.901 2.42	2.62052	-6.23438	4.538873	
Hs.512841	30732 28214	0.9 0.1	4689 393	3.088 -1.4	2.244	2.658768	4.488	
Hs.212051	3572 31392 43064	0.25 0.25 0.04	51 45 14	-1.675 -1.984 3.6	2.583785	-553.668	4.475248	
Hs.18048	41347 39584	0.75 0.97	178 2439	-4.249 0.17	2.2095	-1.08335	4.419	
Hs.519873	41857 24478	0.39 0.44	73 120		-5.94	2.209	-0.74377	4.418
Hs.79769	8147 22855	0.95 0.87	1472 1943	-0.845 3.562	2.2035	1.62201	4.407	
Hs.486798	27329 30025	0.84 0.07	388 11		-6.59	2.19	-0.66464	4.38
Hs.96459	3501 31339	0.14 0.17	14 61	3.645 -0.732	2.1885	1.502575	4.377	
Hs.23367	6765 29189	0.25 0.46	49 270	-1.103 3.242	2.1725	2.031323	4.345	
Hs.518089	38070 6811 35032	0.43 0.46 0.94 0.51	41 187 12718 165	-0.405 0.281 5.98	2.651778	1.878695	4.330336	
Hs.484551	16025 27119 3918	0.85 0.79 0.57 0.89	841 225 120 1959	2.069 -2.483 -2.3	2.823833	10.15767	4.313477	
Hs.453922	34874 42993	0.45 0.15	100 16	-0.456 3.837	2.1465	1.269743	4.293	
Hs.162807	27822 9099	0.89 0.61	7741 95		-5.352	2.136	-0.79821	4.272

図 6 clusterID ごとにフィルタ属性値(RC, CH2N)と QScore 値を集計した例

評価指標の整合性

このプロジェクトを開始する以前に筆者らは様々なフィルタの相関を計算し分類するという予備実験を行った。その際に用いた相関指標は Speaman correlation である。Speaman correlation のよる様々なフィルタのクラスタリングの解析結果やその考察については、別途、論文として報告予定である。

本システムでは新たに別の整合性の指標を提案した。基本的なアイデアは、重複測定スポットの中で、どのスポットを除去すればよいかという問いに対して QScore に基づく答えとフィルタに基づく答えが一致していれば両者に整合性があるとみなすものである。

スポットのクオリティ評価指標と新フィルタの定義

スポット属性指標を用いてスポットのクオリティ評価指標を構築しようという研究がある。スポットの形状やフォアグランド及びバックグランドの発光強度の平均値や分散値などの色々な指標からクオリティ指標の定義式を規定しようとするアプローチやエキスパートの目視評価の結果を教師信号にしてフィードバックして機械学習によりスポットクオリティの良し悪しを分類する規則を生成しようというアプローチもある。今のと

ころ、このようなスポットの評価指標の妥当性を客観的に評価する基準はなく、コンセンサスのとれた決定打はない。本システムでは、これまで論文などで報告された数々のスポットのクオリティ評価式を実装し QScore 値との関連を探る。また測定データの複数のスポット属性指標値からスポットのクオリティを定義する新たな指標を作成し、新指標をフィルタとして用いた場合に QScore の改善効果が従来、提案されていたスポット評価指標と比べてどの程度まで向上させることが可能なのかを検証する。また QScore を教師信号として効果的なクオリティ評価指標を機械学習により獲得することも試みる。

フィルタ統計

既知のスポット属性値(フィルタ)を用いて上述した新たなスポットのクオリティ評価指標を定義し、新規フィルタを構築するには既知フィルタの属性値や統計値を取得する必要がある。例えば、あるフィルタの属性値がそのフィルタの属性値の集合の中でどの程度の位置を占めるのかを平均値や分散値あるいは、ランキング情報をもとに算出した指標は有用である。

クオリティプロットデータの生成

フィルタ値(スポットのクオリティ指標値)と QScore 値にどのような関係があるのかを理解するために、clusterID ごとに両者の指標の組みをプロットした散布図を生成する。(図 8)一つの clusterID=i に対して k 個のフィルタの重複測定結果(f_1, f_2, \dots, f_k)があったとすると、この k 個の重複測定スポットに対して前述した QScore 値(QScore_i)が計算できる。また、 $\{f_1, f_2, \dots, f_k\}$ の代表値として最小値 $\text{MinFilterValue}_i = \min\{f_1, f_2, \dots, f_k\}$ をとり、QScore_i と MinFilterValue_i を全ての i についてプロットすることで散布図が書ける。この散布図はフィルタと QScore の関係を理解するのに有益であり、直感的には、適切なフィルタを用いれば MinFilterValue_i の増加に伴い(フィルタは増加するに従ってクオリティが高まるという仮定)、QScore_iは減少していくと予想される。このクオリティプロット散布図の例については、後述する。

PCL オブジェクト

クオリティ解析機能モジュールは後述するフィルタリング機能に PCL オブジェクトと命名したデータ構造を渡す。PCL オブジェクト(ファイル)の例を図 7 に示す。この例では、UID(スポットの ID)とアレイスライド(spo..)属性を用いて複数のスライドの実験結果を2次元マトリックス形式で記述しているが、アレイスライドのかわりにフィルタ属性を用いて一枚のスライドデータの内容を同様の形式で記述することも可能である。

sample.txt										
	A	B	C	D	E	F	G	H	I	J
1	UID	NAME	GWEIGHT	spo0	spo30	spo2	spo5	spo7	spo9	spo11
2	EWEIGHT			1	1	1	1	1	1	1
3	YAL003W	EFB1	1	0.23	-1.79	-1.29	-1.56		-0.27	
4	YAL004W		1	0.41	-0.38	-0.89	-1.06	-1.6	-1.84	-1.6
5	YAL005C	SSA1	1	0.61	-0.07	-1.29	-1.29	-2	-1.84	-2.25
6	YAL010C	MDM10	1	0.16	-0.15	-0.76	-1.25	-1.89	-1.74	-1.6
7	YAL012W	CYS3	1	0.03	1.39	-0.84	-1.64	-2.84	-2.47	-2.4
8	YAL015C	NTG1	1	-0.18	-0.18	-0.62	-1.32	-1.69	-1.43	-1.79
9	YAL018C	YAL018C	1	-0.51	-0.62	-0.76	3.74	4.54	3.22	4.33
10	YAL025C	MAK16	1	-0.14	-3.32	-1.84	-1.12	-2.4	-1.03	-0.6
11	YAL034C	FUN19	1	0.19	-0.03	-1.03	-1.29	-1.84	-1.94	-1.74
12	YAL035W	FUN12	1	0.01	-1.47	-1.15	-0.69	-1.36	-1.64	-1.29
13	YAL036C	FUN11	1	-0.15	-2.74	-1.79	-1.32	-2.12	0.3	-0.89
14	YAL038W	CDC19	1	-0.06	-1.89	-1.69	-2.32	-2.4	-0.81	-1.6
15	YAL040C	CLN3	1	-0.17	-2.25	-1.69	-2.25	-2.56	-0.3	-2.4
16	YAL054C	ACS1	1	0.51	2.6	1.9	1.7	1.35	-0.03	-0.23
17	YAL055W	YAL055W	1	-0.32	0.83	0.58	0.82	1.4	2.05	2.24
18	YAL062W	GDH3	1	0.3	2.59	3	1.44	0.31	0.34	1.36
19	YAL067C	SED1	1	-0.17	3.44	0.58	1.55	3.26	1.61	2.8
20	YAR003W	YAR003W	1	-0.29	0.54	0.6	1.08	1.42	1.86	1.42
21	YAR007C	RFA1	1	-0.14	1.74	2.41	2.1	2.04	0.57	0.84
22	YAR015W	ADE1	1	0.11	-1.51	-1.4	-1.36	-1.84	-1.89	-2
23	YAR027W	YAR027W	1	0.24	-1.06	-1.36	-1.56	-1.25	-0.94	-1.36
24	YBL009W	YBL009W	1	-0.01	0.62	1.04	1.3	2.52	2.15	2.24
25	YBL010C	YBL010C	1	0.01	0.21	0.7	1.45	2.25	1.77	1.24
26	YBL015W	ACH1	1	0.52	1.01	1.49	1.75	1.49	0.58	0.19
27	YBL027W	RPL19A	1	0.01	-1.84	-0.97	-1.47	-1.79	-1	-0.51

図 7 PCL ファイルの例

PCL ファイルは一般的には一つのフィルタ(通常、Log Ratio 値)の情報を複数のスライドに対してまとめたもので、よく一般的に普及しているフォーマットである。PCL とはクラスタリングを行う前の状態を意味しており、PreClustering の略である。本システムでは、PCL ファイルを一つのスライドに対し複数のフィルタ情報をまとめる形式として拡張して利用することにした。

【実装】

本システムでは下記のモジュールを実装した。

ArrayQuality.pm

クオリティ評価指標を計算するコアモジュールである。calculateQScore()メソッドはアレイデータに対し QScore 値を算出する。全てのスポットに対し clusterID(測定遺伝子名に対応するアノテーション ID)をチェックし重複した clusterID をもつスポット群を clusterID ごとにまとめ、測定値の分散を調べる。QScore 値は一つのアレイ(数万スポット)に対する一つの指標値であるが、clusterID ごとに QScore 値を計算し、フィルタの属性値と対応づけることはフィルタの特性を把握することに役立つ。これを実装したのが qualityPlot()メソッドで、clusterID ごとに qscore 値と最小フィルタ属性値の組を返す。clusterID ごとに測定されたスポット数が集計され、後述する可視化モジュールで散布図のプロットをこの測定数に基づいて色づけして表示される。(図 8)

CheckConsistencyWithFilter(), 及び CheckConsistencyWithQScore()メソッドは、前述した評価指標の整合性を実装したもので各フィルタ間、及びフィルタ値と QScore

値の整合性を数値化した指標を返す。この結果も後述する解析データ可視化機能モジュールでグラフとして可視化される。(図 9)

`makePCLObject()`メソッドはアレイデータオブジェクトを前述した PCL オブジェクトに変換しフィルタリング機能モジュールへ渡す。

DataStat.pm

このモジュールは各フィルタの統計指標を計算する。各フィルタに対し測定スポット数、平均値、分散などの基本統計量を算出する。また外れ値の検出アルゴリズムを実装するのに用いられる指標を計算する。例えば `calculateOutlierWeight()`メソッドは、与えられたスポットのフィルタ属性値の”外れ値度合い”を数値化して返すもので、フィルタ値分布におけるランキング計算情報をもとに0から1までの数値を返す。このフィルタ属性値の”外れ値度合い”の指標値は、MinimumOrder 法(仮名)により複数のフィルタを組み合わせたハイブリッドフィルタの一つを実装するときにも用いられる。

CheckConsistency.pm

このモジュールは各フィルタと QScore の整合性やフィルタ間の整合性を数値化した指標を計算する。前述した評価指標の整合性のアイデアを実装している。`_checkConsistencyWithFilter()`メソッドは、フィルタ間の整合性を0(整合性なし)から1(整合性あり)までの実数値に数値化する。また `_checkConsistencyWithQScore()`メソッドは、フィルタと QScore 値の整合性を0から1までの実数値に数値化する。`ConsistencyPlot1()`メソッドは clusterID ごとに全ての Replicate スポットに対し QScore 値との整合性を計算し、散布図データを生成する。このデータを可視化した例は後述する解析結果可視化モジュールの章で後述する。

Filter.pm

このモジュールはフィルタ処理に有益なメソッドを定義している。例えば `getBorderSpots()`メソッドは指定されたフィルタ属性値付近のスポット ID を検索して返す。これは後述する GUI モジュールにおいて、フィルタのカットオフ値付近のスポットイメージを提示する際に呼び出される。

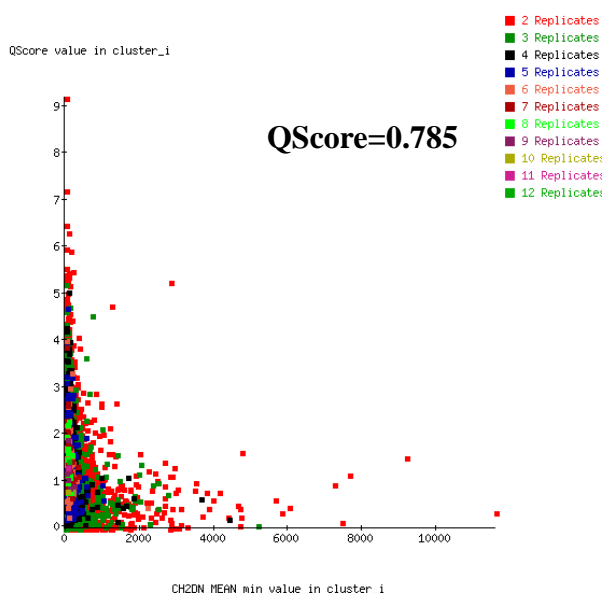
NewFilter.pm

入力データファイルにはない新たなフィルタを定義し登録するためのモジュールである。既知のマイクロアレイスポット評価に関する論文や開発者のアイデアに基づき様々な新規フィルタ生成アルゴリズムが実装されている。例えば、`cv()`メソッドは各フィルタについてフィルタ値の分散を平均値で除した Coefficient of Variation を計算して新規フィルタの属性値とする。`qcom01()`、`qcom02()`メソッドは、Ghosh らによって提案

されたスポットのクオリティ評価式を実装したものでスポットのサイズ、SN比、バックグラウンド強度に関する2つの指標、スポットのサチュレーションなどに関する5つのスポット評価指標から一つの指標値を算出し新規フィルタを作成する。(Bioinformatics, Vol. 21. 21 no. 8, 2005, pages 1573) mmd1l(), mmd2l() メソッドも論文から着想を得て実装したメソッドでスポット発光強度の平均値とメディアン値との差異をクオリティ評価指標として登録する。minimumOrder()メソッドは開発者が発案したスポットのクオリティ評価手法で、複数のフィルタに対し、それぞれのフィルタの外れ度具合を数値化し、それらの指標の最小値をクオリティ指標として登録する。スポットのクオリティを評価する方法は他にも色々考えられるが、現在、それらの特性について比較評価を行っているところである。

【実行結果例】

後述する解析結果可視化機能モジュールを用いてクオリティ解析結果解析モジュールが生成したグラフの例を示す。図 8 は前述したクオリティプロット図の例である。上の図はあるフィルタ(CH2N)のプロット図で下の図は別のフィルタのプロット図である。散布図の各点は、同じ ClusterID を持つ重複測定スポットに対応し(最小フィルタ値、QScore 値)の組がプロットされている。散布図の各点の色は重複測定数を表している。上の図ではフィルタ値が大きくなると QScore 値が明確に減少するという傾向がみられるが、フィルタによって、このプロット図の特性は異なってくる。



QScore=0.785

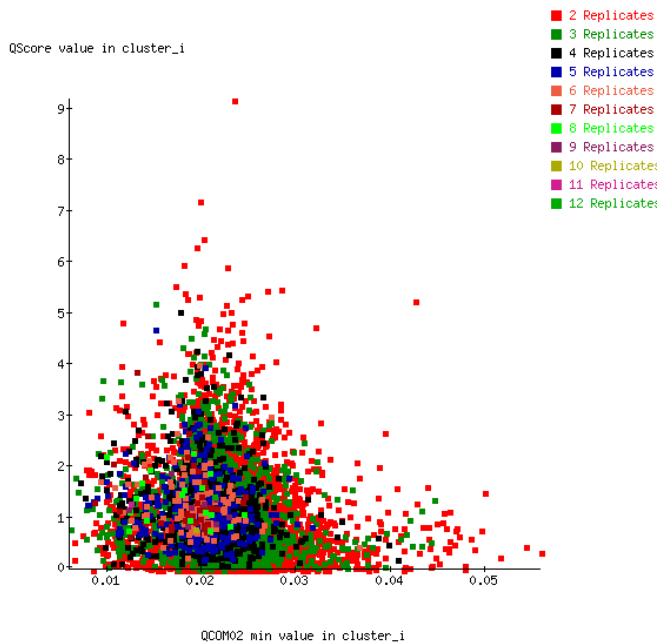
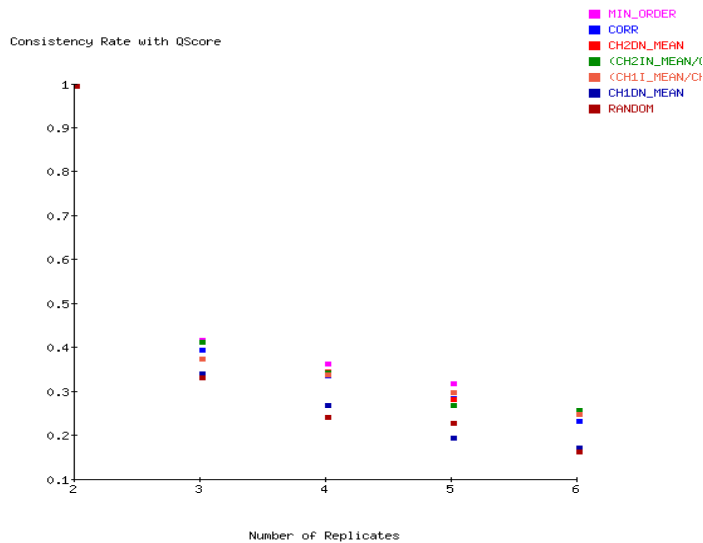


図 8 クラスタごとのフィルタと QScore クオリティプロット図の例

図 9 に評価指標の整合性の項で上述した実行例を示す。上の散布図は7つの異なるフィルタと QSCORE の整合性(正解率)の指標値をみたもので、横軸は重複スポット数 k を表す。 $k=2$ のときは正解率は1であるが、 k の増加とともに正解率は下がっていく。この正解率の減少率はフィルタによって差がある。ピンク色の点は、本研究で新たに発案した MinimumOrder 法によるフィルタ値でこのフィルタが他のフィルタに比べて正解率が高い傾向がみられる。下の表は正解率を各フィルタで比較したもので MIN_ORDER の値が最も高い。



filter	consistencyRatio	total	consistencyNum
MIN_ORDER	0.756211180124224	7728	5844
CORR	0.746376811594203	7728	5768
CH2DN_MEAN	0.75194099378882	7728	5811
(CH2IN_MEAN/CH2BN_MEDIAN)	0.751164596273292	7728	5805
(CH1I_MEAN/CH1B_MEDIAN)	0.742624223602484	7728	5739
CH1DN_MEAN	0.717520703933747	7728	5545
RANDOM	0.71195652173913	7728	5502

図 9 QScore とフィルタの整合性(正解率)のプロット例

(4) マイクロアレーデータのフィルタリング機能

以下では図表を用いてフィルタリング機能の目的、設計方針、実装、動作例について説明する。

【目的】

この機能モジュールは前述したPCLオブジェクトを受けとり、各フィルタごとにフィルタ一値を連続的に変化させながらフィルタリングを行い、フィルタリンググラフを生成する。このフィルタリンググラフから、それぞれのフィルタごとにQScoreの改善率を解析し、効率的にフィルタリングを行うことが出来るカットオフポイントを推定する。また複数のフィルタの関係やハイブリッドフィルタのフィルタリング効果を調べるために有用なデータを収集する。

【設計方針と基本的なアイデア】

この機能モジュールのアイデアは、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を調べ、そ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させようというものである。

また多次元のフィルタパラメータの相互作用を解析し、複数フィルタを組み合わせでより効率的なフィルタリングを実現する方法や、カットオフ値付近のボーダーラインにあるスポット群を提示する機能、その後の解析に有用なクオリティ指標値を算出する方法などを試みている。いくつかのアイデアについて下記に簡単に説明する。

フィルタリングアルゴリズム

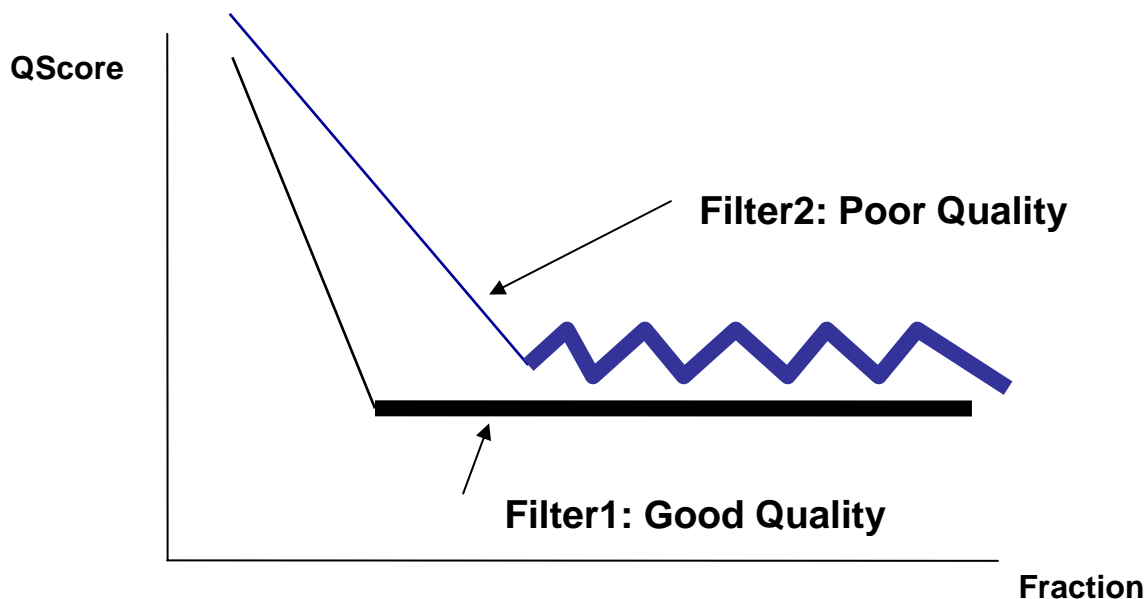
与えられたフィルタに対してフィルタ値を N ステップに分割し、それぞれのステップ i において、フィルタ値がステップ i における基準値より小さいものを順に除去していく。(説明のためにフィルタ値が小さいほどスポットのクオリティが低いと仮定する。)除去されてのこったスポット群に対して QScore 値を計算し、ステップ i におけるフィルタ値、Fraction ($=\{\text{ステップ } i \text{ における重複スポット数}\} / \{\text{ステップ } 0 \text{ における重複スポット数}\}$)、QScore 値の3組の値を記録する。複数のフィルタを用いたフィルタリングを実行する際には、そのアルゴリズムと等価な一つのフィルタを生成し、前述したフィルタリングを実施する。 K 個の複数フィルタを用いる場合でも、全てのスポットに対してフィルタリングで除去していく順序を決めることが出来れば、その順序番号を新たなフィルタ指標として用いることで一つのフィルタに対応させることが可能である。

カットオフ値抽出アルゴリズム

前述したフィルタリングのデータから、Fraction 値と QScore 値をプロットした Fraction vs. QScore グラフを作成する。このグラフの変曲点(Inflection Point)は理論的には各点における値の2階微分値(接線の傾きの変化率)をもとに定義できるが、プログラムとしての実装は、Janos Demeter 博士 (SMD グループの筆者の共同研究者)が実装したメソッドを用いて計算する。このプログラムは最小二乗法を用いて各点の前後の近傍の回帰直線を求め、2つの回帰直線の傾きの差を計算する。この回帰直線の傾きの差が最も大きくなるような点を計算対象の近傍をスライドさせながら探していく。ただし、近傍内に計算対象の点が少なすぎる場合や、回帰直線を求める際の最小二乗誤差が基準よりも大きい場合には直線のあてはめがよくないので除外する。従ってグラフの曲線が極端に振動していたり、はっきりとした変曲点が認められないような場合には、変曲点が抽出不能になる場合もある。変曲点が明確に抽出できるケースがどの程度あり、どのようなケースに明確な変曲点が現れるかなどについての実験データも蓄積されている。

フィルターランキングアルゴリズム

複数のフィルターのフィルタリングカーブを比較し、ランキングをつける。ランキングは0からはじまる整数値を割り振りこの値が小さいほどよいフィルターであるとみなす。理想的なフィルタリングカーブは最初に急激に QScore 値を下げ、それからフラットになるというものである。このような理想的なフィルタリングカーブを特徴づける指標として3つの基準を定めた。図 10 に概念図を示す。基準1はフィルタリングの最初の段階での QSCORE の減少率の傾きで、これが急であるほど良いフィルターであるとみなす。第2の基準はフィルタリングカーブの下部にある領域の面積で、この面積が小さいほど良いフィルターであるとみなす。第3の基準は、フィルタリングカーブの振動数でフィルタリングの過程で QSCORE が減少せずに増加してしまった回数をフィルタリングカーブの振動数としてカウントする。この振動数が小さいほど良いフィルターであるとみなす。これらの3つの基準それぞれにランキングスコアをつけ、3つのランキングスコアの合計をフィルターの良さを表すランキング指標として定義する。



Criteria for Ranking Filters

1. Steepness of initial drop
2. Area under the QScore curve
3. Monotonicity of QScore curve

図 10 フィルターのランキング指標値算出のための3つの基準

2次元の QScore Surface プロット

前述したフィルタリングは1次元のフィルタに対して QScore のダイナミクスを調べるものであり、QScore が多次元のフィルタ空間でどのような形状をしているのかは定かではない。多次元空間での QScore の形状が把握できれば、QScore を目的関数とした多次元のパラメータの最適化問題としてフィルタリングのカットオフ推定問題を捉えることが出来る。

このような着眼点から QScore を多次元的に可視化しようと考えた。図 11 にこのアイデアを示す。2次元のフィルタに対して一つの QScore 値を対応づければ3次元空間の局面として QScore Surface を立体的に可視化することが出来る。アイデアとしては2次元のフィルタ空間を格子状に分割して、それぞれの格子内で QScore を計算すればよい。しかし問題は、2次元のフィルタ空間を細かく分割するとそれぞれの格子内での重複スポットの数が少なくなり QScore が計算不能になるということである。このため発想をかえて格子内の QScore を計算するのではなく格子外の QScore を計算することにした。このようにすれば、QScore の計算対象となる重複スポットがなくなってしまうという問題を避けることが出来る。この指標を仮に各格子点 (i,j) について $QScore_outSide(i,j)$ と表記することにすると、 $QScore_outSide(i,j)$ の値が大きい程、アレ領域 (i,j) のクオリティは高いという解釈が出来るので、3次元で可視化された QScore Surface の山の部分を残すようなフィルタリングを行えばよいことになる。この山のふもとにあたる境界線を求めることが、多次元でのカットオフ境界を見出すアルゴリズム構築の目標となる。

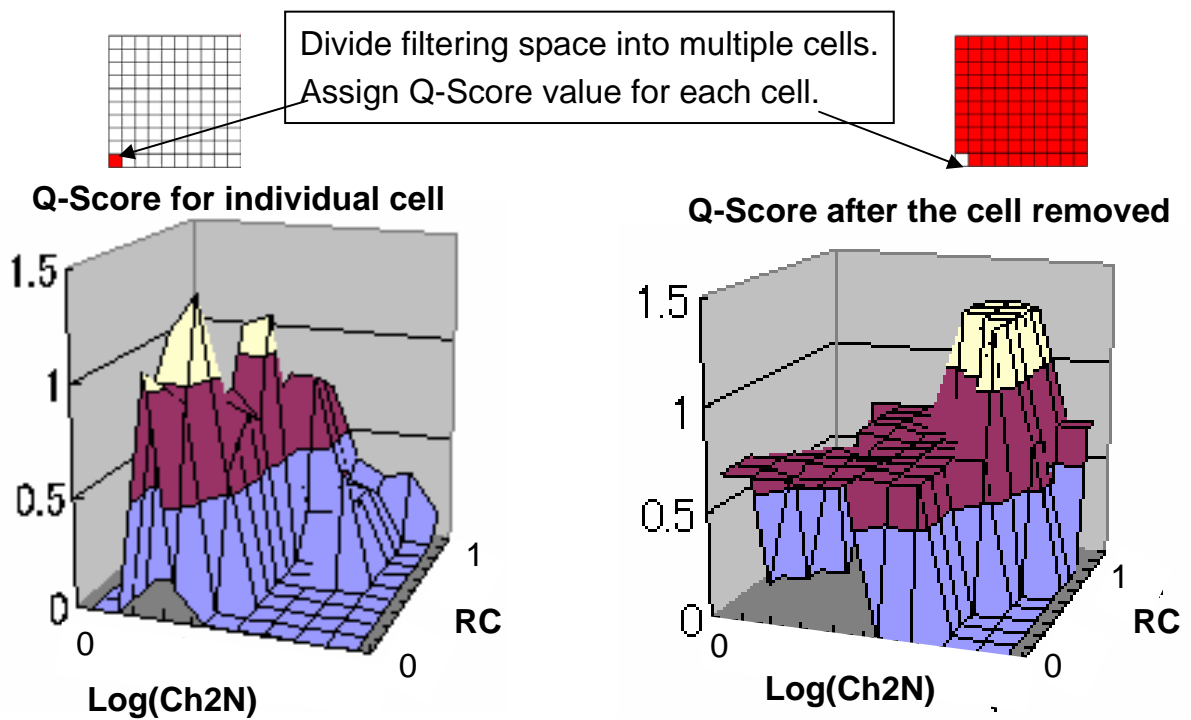


図 11 QScore Surface を可視化するアイデア

中間ファイル

この機能モジュールは実行過程で様々な中間ファイルを生成する。これらの中間ファイルのリストを表 1 に示す。これらの中間ファイルは後述する解析結果の可視化モジュールの入力となる。

拡張子	名称	説明
.pcl	PCL ファイル	PCL フォーマットの測定入力データ
.extr	Extreme ファイル	フィルタ値の範囲(最大値、最小値)
.res	Result ファイル	フィルタリングの解析結果
.data	Data ファイル	ランキング計算の3つの基準値の計算結果
.cutoff	Cutoff ファイル	カットオフ値の抽出結果
.ranks	Ranks ファイル	フィルタのランキング指標値

表 1 フィルタリング機能モジュールで生成される中間ファイル

【実装】

本システムでは下記のモジュールを実装した。

ArrayFiltering.pm

実際にフィルタリングを行うコアモジュールである。このモジュールは開発者の共同研究者である Janos. Demeter 博士が実装したモジュールを拡張している。オリジナルのモジュールはスライドごとにまとめられた PCL ファイルから指定されたフィルタに対し複数のスライドごとの QScore のダイナミクスを比較するグラフを生成するものであったが、このモジュールではフィルタごとにまとめられた PCL オブジェクトを読み込み一つのスライドに対し複数のフィルタの QScore ダイナミクスを比較できるように拡張した。また新たに、前述した QScore Surface を描画するのに必要なデータを生成するメソッドを新たに実装した。

このモジュールは前述したクオリティ評価機能モジュールで生成された PCL オブジェクトをロードする。コアメソッドは calculateScores() で PCL オブジェクトを受け取りフィルタの値を連続的に変化させながらフィルタリングの結果を記録していく。PCL オブジェクトには同一スライドの複数のフィルタ情報が格納されており、それぞれのフィルタについてフィルタリングダイナミクスを調べる。(PCL オブジェクトに同一のフィルタに対して複数のスライドが指定されていた場合には、スライドごとに特定フィルタのフィルタリングを実行する。) QScores 指標には、2.1.1 プロジェクト概要で定義した定義式の他に類似した他の4つの定義式もオプションとして実行出来るように拡張している。

QScore2DSurface()メソッドは前述した QScore Surface の構築アルゴリズムを実装したものである。指定された2つのフィルタ F_i, F_j に対し、2次元のフィルタ指標空間を格子状に分割し、それぞれの格子領域 Area(i,j) ごとにその格子外部領域の QScore 値、QScore_outSide(i,j) を計算する。gnuplotFor2DSurfaceGraph()メソッドは、QScore2DSurface()メソッドによって得られたデータを gnuplot ツールを用いて可視化する。

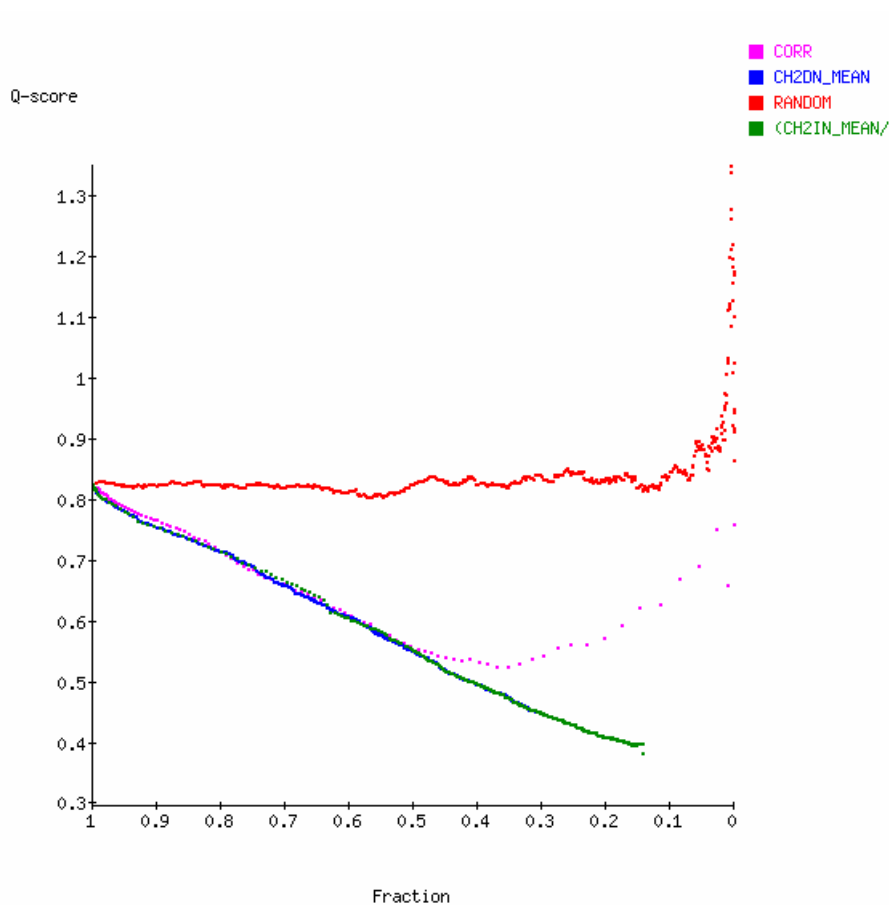
Cutoffs.pm

このモジュールは、QScore vs. Fraction グラフにおける変曲点を抽出し、フィルタの良し悪しを評価する3つの基準に基づいたランキングアルゴリズムを実装したものである。これも共同研究者の Janos Demeter 博士が実装したモジュールを開発者が本プロジェクトのために修正及び拡張したものである。オリジナルのモジュールでは、第2基準のみが動作していたが第1基準と第3基準も有効に働くように改良した。また PCL オブジェクトにスライドごとの複数のフィルタ情報が格納されている場合にも実行できるようにプログラムを拡張した。グラフの変曲点の抽出には、前述した最小二乗

法を用いたアルゴリズムを実装している。

【実行結果例】

上述したフィルタリングカーブとカットオフ抽出、及びフィルタランキングの計算結果の例を図 12 に示す。この例ではあるスライドデータに対して4つのフィルタが比較され、それぞれの変曲点に関する情報を抽出している。(このグラフの例では変曲点は明確ではない。)



Filter	IP_FilterValue	IP_Fraction	IP_QScore	Ranks
CH2DN_MEAN	265.000	0.462	0.533	1
CORR	0.788	0.414	0.543	2
_MEAN/CH2BN_M	4.560	0.310	0.458	0
RANDOM	0.388	0.508	0.824	3

図 12 データフィルタリングとカットオフ抽出及びランキング指標計算の実行例

(5)クオリティ解析結果の可視化機能

以下では図表を用いて解析結果可視化機能の目的、設計方針、実装、動作例について説明する。

【目的】

この機能モジュールは前述したクオリティ解析機能モジュールやフィルタリング機能モジュールで生成された解析結果オブジェクト(或いはファイル)を可視化しグラフ、プロット図、スポットイメージなどの Gif ファイルを生成する。これらの Gif ファイルのいくつかは後述する GUI モジュールから呼び出されユーザに提示される。

【設計方針と基本的なアイデア】

この機能モジュールのアイデアは、前述したクオリティやフィルタリングの様々な解析データをビジュアル化し、Gif ファイルを生成するというものである。

具体的には、フィルタリング機能で生成されたフィルタリング結果をグラフとして可視化し、QScore グラフ(QScore-Filter, Fraction-Filter, QScore-Fraction)を Gif ファイルとして生成することが出来る。また多次元の QSCORE Surface を可視化するために、2次元のフィルタと QScore 値からなる3次元プロット図も生成することも出来る。このように QScore の形状を一次的ではなく立体的に表現することは有益である。(図 11, 図 13)

また、マイクロアレー画像データとクオリティ解析結果を対応づけるために、それぞれのフィルタごとに、フィルタリング機能で抽出されたカットオフ値に近いボーダーライン上にあるスポットイメージ群をアレイイメージから切り出し Gif ファイルとして生成する。(図 14)

また GUI からは反映させていないが、前述したクオリティ解析モジュールやフィルタリングモジュールの解析結果を可視化する機能を実現している。例えば、クオリティ解析モジュールの章で述べたクラスごとに QScore 値とフィルタ値の代表値をプロットしたビジュアルな散布図や、QScore とフィルタの整合性(正解率)に関するプロット図も生成可能である。(前述したクオリティ解析機能モジュールの項を参照のこと。)

【実装】

本システムでは下記のモジュールを実装(利用)した。

ScatterPlotQScore.pm

QScore のプロットデータから GD ライブラリの画像処理機能を用いて散布図を作成し gif ファイルとして保存するモジュールである。このモジュールは SMD で開発されたものを再利用した。GD ライブラリを用いて散布図を描くための色々なメソッドを定義している。makeScatterPlot()メソッドは各種の描画メソッドを用いて散布図を作成し gif ファイルにセーブする。

QualityPlot.pm

クオリティ解析機能モジュールの項で説明したクオリティプロットを実装し可視化グラフ(散布図)を作成するモジュールである。前述した ScatterPlotQScore.pm モジュールを利用してプロット図を作成する。QualityPlot1()メソッドは散布図のデータを作成し、createGraphsForPlot1()メソッドは散布図のグラフ(gif ファイル)を生成する。

ArrayImage.pm

イメージデータを読み込み GD ライブラリの画像処理機能を用いて注目するスポットの部分だけを抽出し gif ファイルとして保存するモジュールである。フィルタリングのカットオフ値の境界領域にあるスポットイメージを抽出するためにこのモジュールを実装した。getSpotImage()メソッドは、スポット ID から対応するスポットイメージのアレイイメージ上の座標を識別し、対応するスポット領域を切り出した image オブジェクトを返す。getBorderSpotKeys()メソッドはフィルタ値と表示数を指定するとそのフィルタ値付近のスポット ID を表示数の数だけ抽出する。これは後述する GUI でフィルタリングのカットオフ領域付近のスポット群を表示する際に用いられる。

Gnuplot

3D 立体データの表示には、広く普及しているフリーのグラフ描画ツールである gnuplot を利用した。Perl script からパラメータを与えて gnuplot を呼び出しデータを可視化する。Gnuplot の公式サイトは下記にある。

<http://www.gnuplot.info/>

【実行結果例】

下記に上述した解析結果可視機能化モジュールで作成されたグラフやスポットイメージの例を示す。図 13 はあるスライドデータに対して Regression Correlation (CORR)と CH2DN_Mean の2つのフィルタと QScore の関係を可視化したものである。

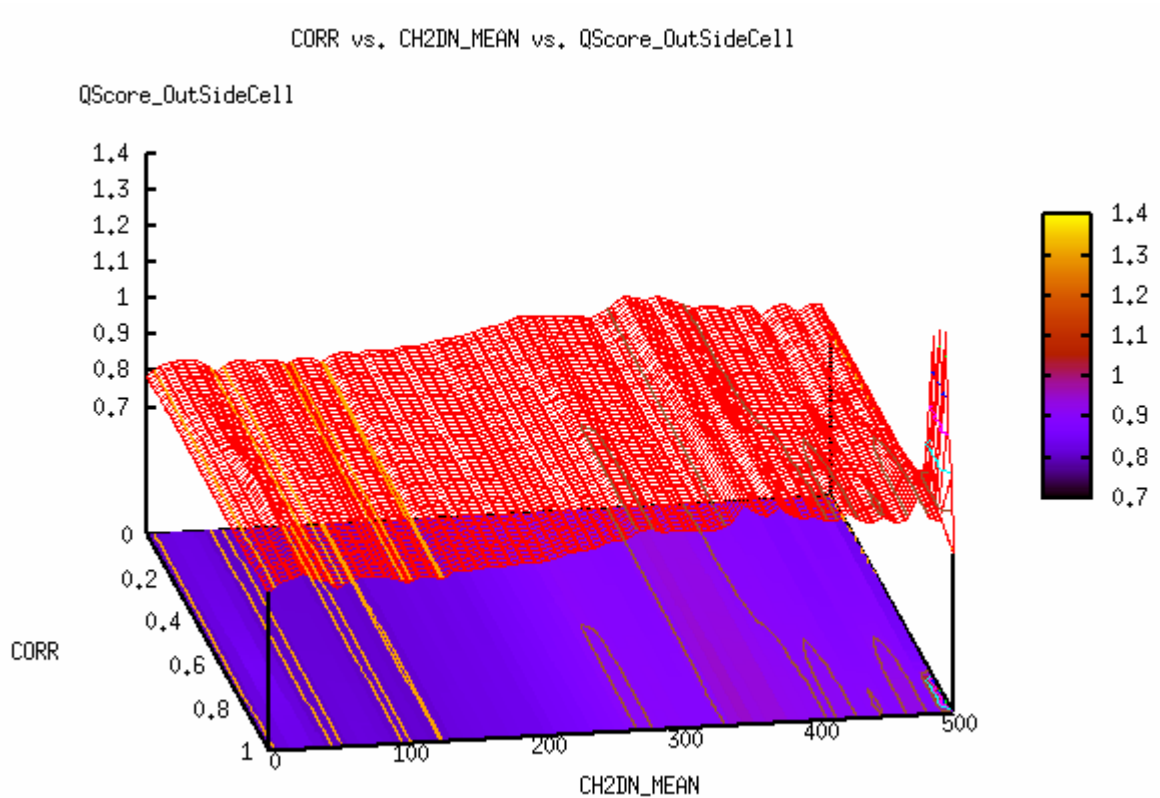


図 13 QScore Surface の3次元プロット図の例

図 14 にあるスライドデータに対して Regression Correlation (CORR) フィルタを用いた場合のカットオフ値(0.788)周辺のスポットイメージを抽出した例を示す。

Slide	Filter	CutOff
testDemo3	CORR	0.788

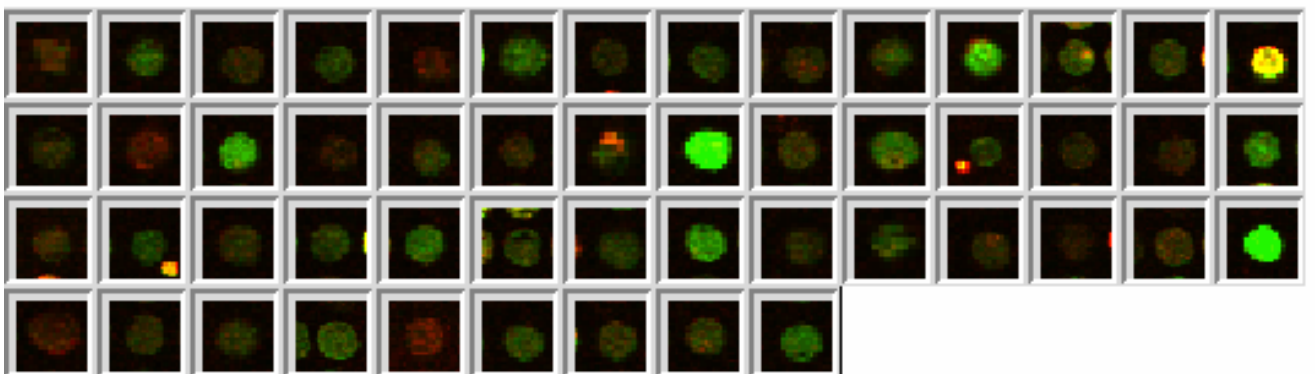


図 14 フィルタリング境界領域にあるスポットイメージ抽出例

(6)クオリティ指標の比較評価実験結果の集積、表示機能(GUI)

以下では図表を用いて解析結果表示機能の目的、設計方針、実装、動作例について説明する。

【目的】

この機能モジュールは、前述した各機能モジュールから生成されたデータやグラフを容易な操作でユーザに分かりやすく整理し表示する GUI を提供する。

【設計方針と基本的なアイデア】

上述した各モジュールはコマンドラインから Perl スクリプトとして呼び出され実行される。必要に応じて上述したモジュールを用いて Perl スクリプトを書き実行することが出来るが、これは初心者には敷居が高い。Stanford Microarray Database(SMD)では、Web サービスにより様々な解析サービスをオンラインで提供しているが、対象は SMD ユーザに限られており、SMD のロードされたデータが解析対象となるので、SMD ユーザ以外のユーザに対してコンパクトな GUI を提供することは有意義であると考えられる。これらのことを考慮して、実装したモジュールや解析結果を PC 上でコンパクトに呼び出すことが出来るような GUI を Web サービスとは別に開発する。

【実装】

本システムでは下記のモジュールを実装した。実装には Perl/Tk を用いた。Perl/Tk を GUI に用いた理由は、上述した他のモジュールが perl を用いて実装されていること、及び Tk のインターフェースを用いることで GUI の開発効率が高まると期待されたからである。

QScoreGUI.pl

GUI のコアとなるプログラムである。下記の他の GUI プログラムをロードしメインウィジェットを起動する。Perl/Tk モジュールを呼び出し MainLoop()メソッドにより GUI ベースのイベント駆動型の処理が開始される。

mainWindow.pl

GUI のメインウィジェットである。このウィジェットからデータファイルやイメージファイルの指定、フィルタリングプログラムの実行及び解析結果の参照を行う。データ入力機能とフィルタリング機能と QScore Surface 表示機能と境界スポットイメージ表示機能をメニュー画面から選択し実行出来るようになっている。処理対象となる測定データのファイル名とイメージデータのファイル名がウィジェット中のテキスト

画面に表示され、処理のステータスや進行状況がプログレスバーなどを用いてメインウィジェットに表示されるようになっている。

configureGUI.pl

PC 環境に応じて GUI の環境設定を行うプログラムである。ウィジェット画面時に利用する文字サイズやフォントなどをここで定義している。

ImageView.pl

イメージファイルやデータファイルを読み込むためのウィジェット。メインウィジェットの File メニューから呼び出される。Load Image 或いは Load Data が選択されるとユーザにファイルの指定を促し、ファイル選択後にデータがシステムにロードされる。イメージデータは新たなウィジェットを生成しロード時にユーザに表示する。データのロードには時間がかかる場合があるので、データロード中は、プログレスバーを表示する。

FilterView.pl

様々なフィルタでフィルタリングして生成された QScore のダイナミクスグラフを参照するためのウィジェット。メインウィジェットの Filtering メニューから呼び出される。Show Results メニューを選択すると新たなウィジェット画面が提示されこれまでに解析された QScore のフィルタリング結果のテーブルが表示される。このテーブルの各行の末尾には Graph ボタンと CutOffs ボタンが表示されていてこれをクリックするとフィルタリングダイナミクスのグラフや本システムで抽出されたグラフの変曲点に対応するカットオフポイントに関する情報が表示される。

Set parameters and Run メニューを選択するとフィルタリングのパラメータを設定し QScore のフィルタリング解析が実行される。

QSurfaceView.pl

二つのフィルタと QScore の関係を可視化して 3次元で表示するウィジェット。メインウィジェットの Qsurface メニューから呼び出される。Show Results メニューを選択するとこれまでに解析した QScore Surface の解析結果テーブルが表示される。このテーブルの各行の末尾には Graphs ボタンが表示されていて、これをクリックすると QScore 2D Surface の 3D イメージが表示される。また、この 3D イメージの視点を変えたり、他の指標を用いて 3D イメージを可視化するための設定ボックスも表示されており、ここでパラメータを選択することにより gnuplot を呼び出して新たな 3D イメージを表示させることが出来る。Set parameters and Run メニューを選択すると QScore 2D Surface のフィルタ設定画面が表示されパラメータを設定して QScore

Surface 計算プログラムを実行させることが出来る。

BorderSpotsView.pl

本システムによって抽出されたフィルタリングのカットオフポイント周辺のスポットイメージを表示するウィジェット。メインウィジェットの BorderSpots メニューから呼び出される。Show Results メニューを選択するとこれまでに解析したセットの一覧がテーブルになって表示される。テーブルの各行の末尾には Spots ボタンが表示されており、このボタンをクリックするとフィルタリングのカットオフポイント周辺のスポットイメージ群が表示される。またスポットイメージ表示画面にはパラメータ選択が出来るボックスも表示されており、フィルタやフィルタ値、表示数などを指定して対応するスポットイメージを表示させることが出来る。表示されたスポットイメージはボタンになっており、それぞれのスポットイメージをクリックすると、そのスポットの様々な属性値がテーブルとして表示される。

HelpView.pl

ヘルプ画面を表示するウィジェット。メインウィジェットの Help メニューから呼び出される。

progressBar.pl

進行に応じてプログレスバーを表示するウィジェット。ファイルのロード時やスポットイメージの表示時など処理に時間がかかる場合に進行過程をメインウィジェット中のプログレスバーとして表示する。

【実行結果例】

下記に上述した GUI の実行例を示す。図 15 にメインウィジェットとイメージデータ入力画面のイメージを示す。

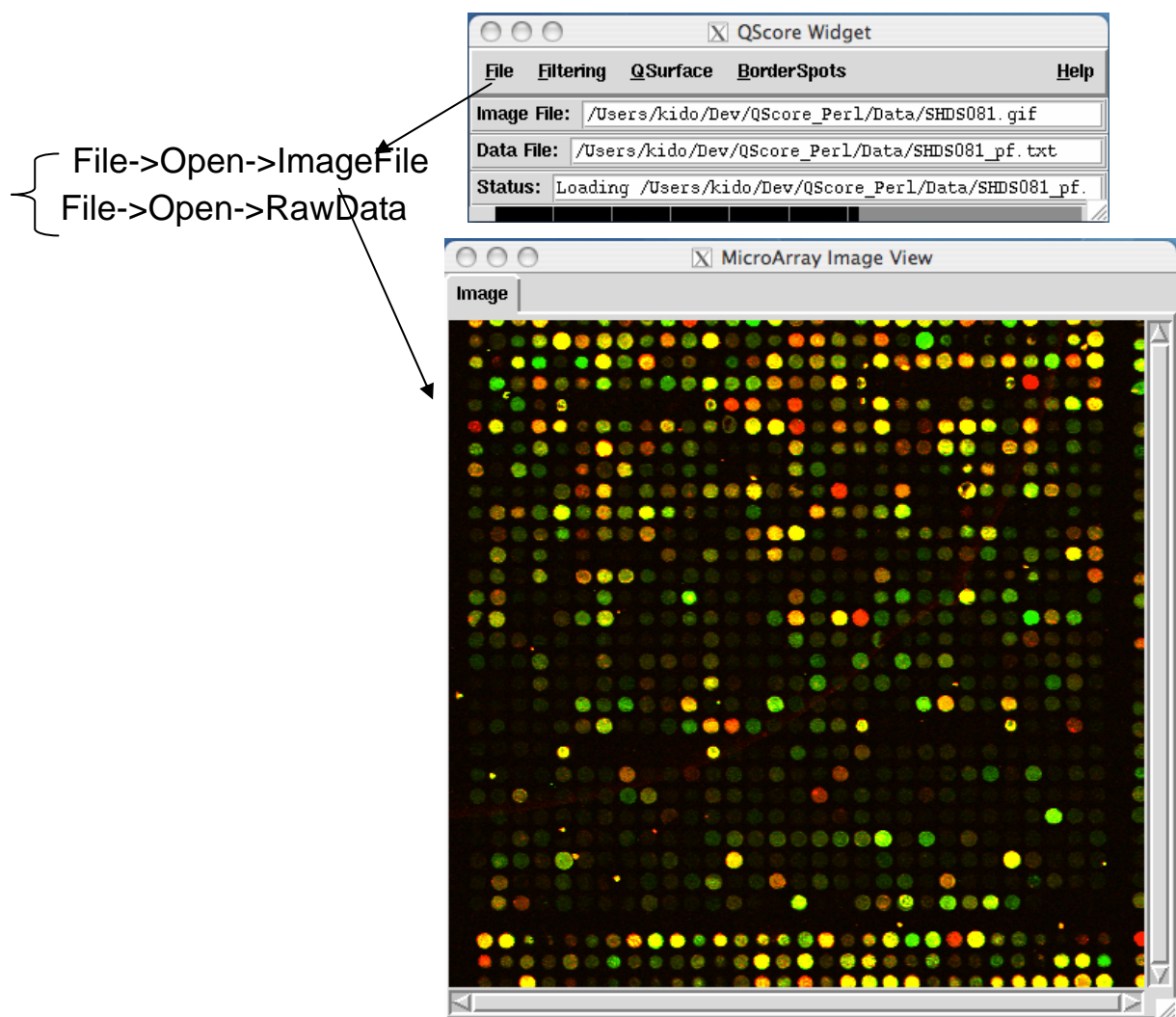


図 15 GUI の実行例 (メインウィジェットと入力インターフェース画面)

図 16 にメインウィジェットのメニューからフィルタリング解析結果を呼び出す画面イメージを示す。解析結果の参照テーブルからフィルタリングカーブの比較グラフと抽出された変曲点の情報がウィジェットに示される。

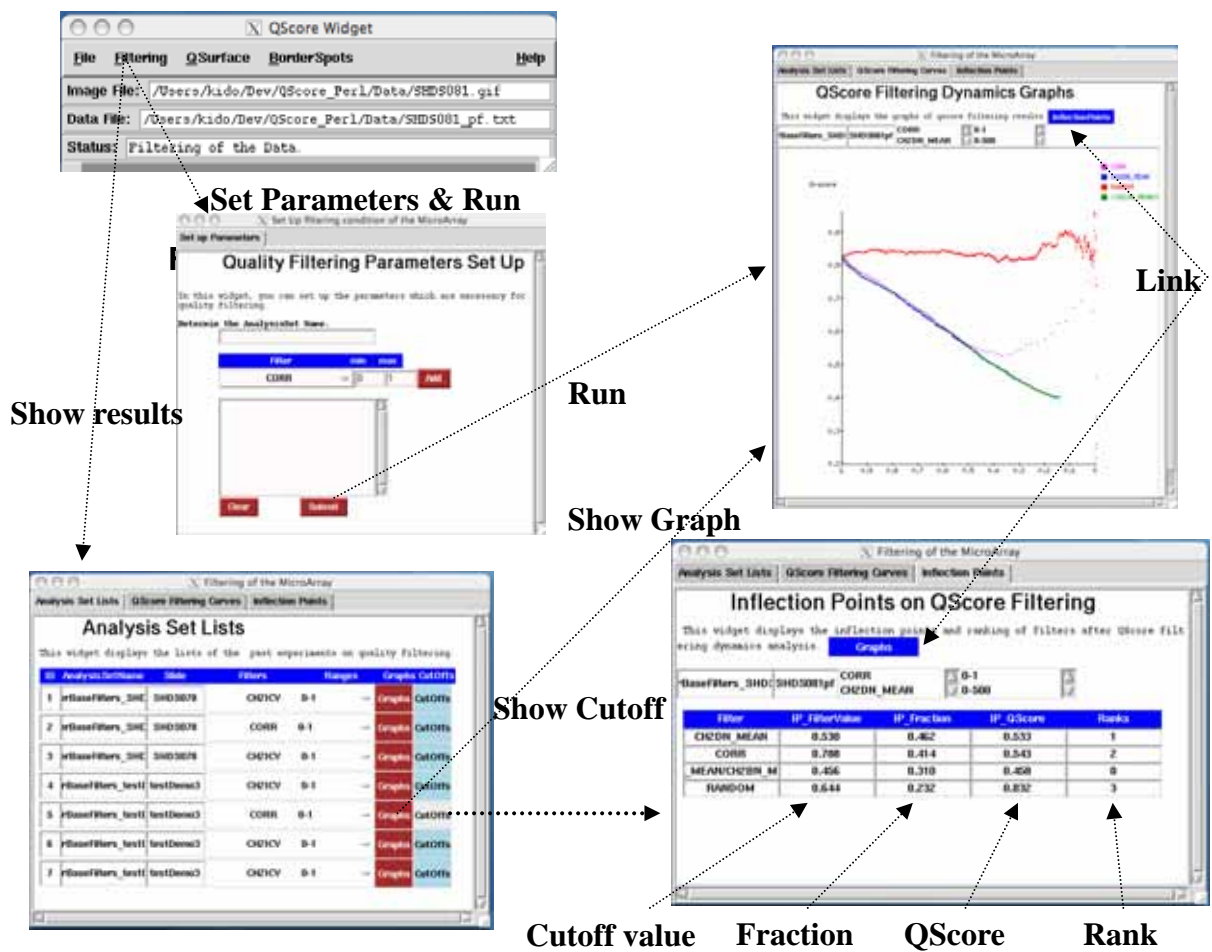


図 16 GUI の実行例 (フィルタリングの実行と結果参照画面)

図 17 に GUI のメニューから QSurface の実行及び結果参照を選択した時の画面イメージを示す。3D グラフは二つのフィルタに対して、QScore と Frequency(頻度)情報を可視化できるようになっている。3D 画像の視点(View Point)を変えて見ることも出来る。

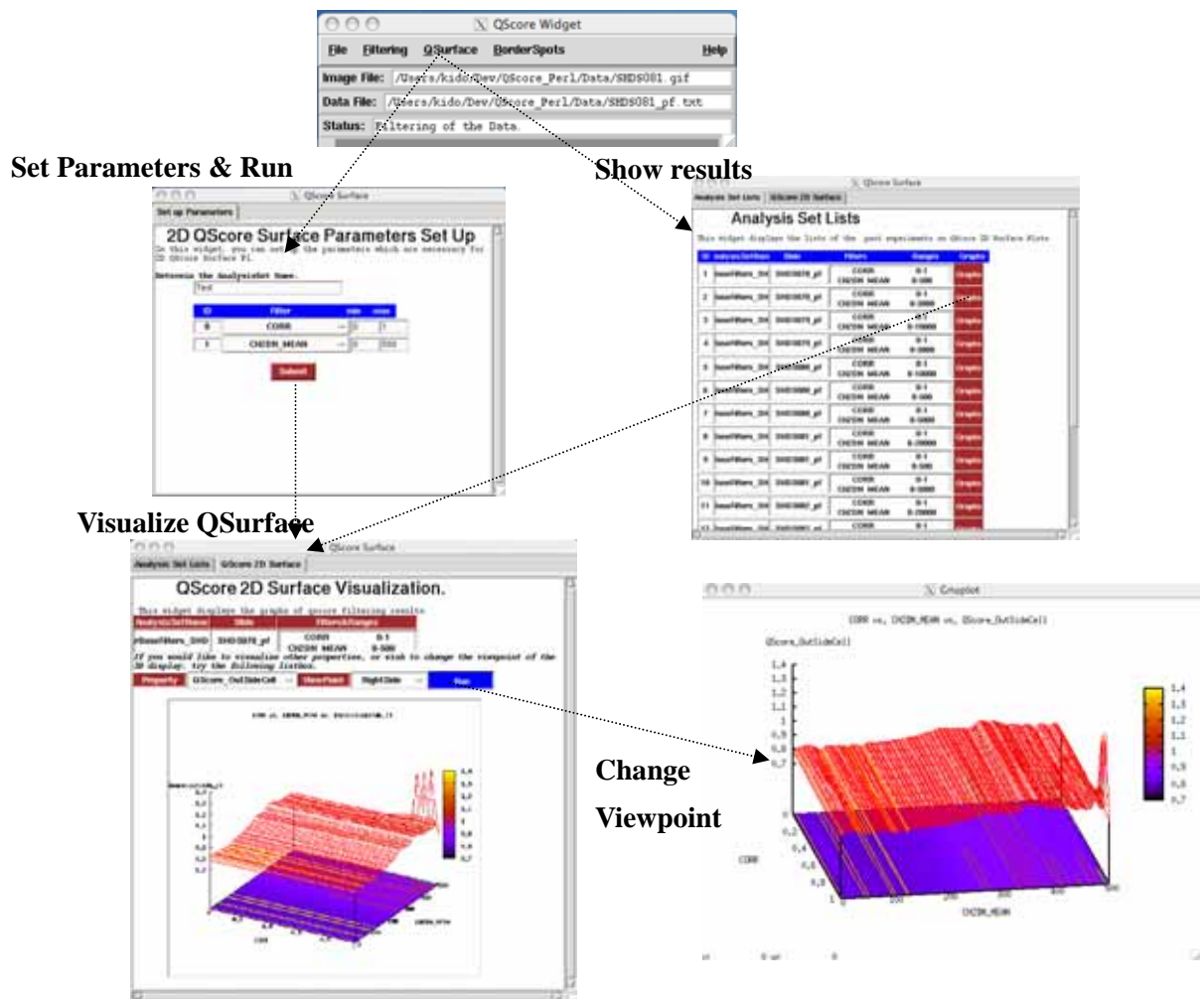


図 17 GUI の実行例 (QSurface 可視化の実行及び結果参照画面)

図 18 に GUI のメニューから解析結果を参照しフィルタリングの境界領域に属するスポットイメージを表示する画面イメージを示す。フィルタごとにカットオフ値周辺のスポットイメージが提示される。フィルタの値と表示数を指定して、フィルタ値周辺のスポットを表示することも出来る。提示されたスポットイメージをクリックすると、そのスポットの詳細な属性情報が表示される。

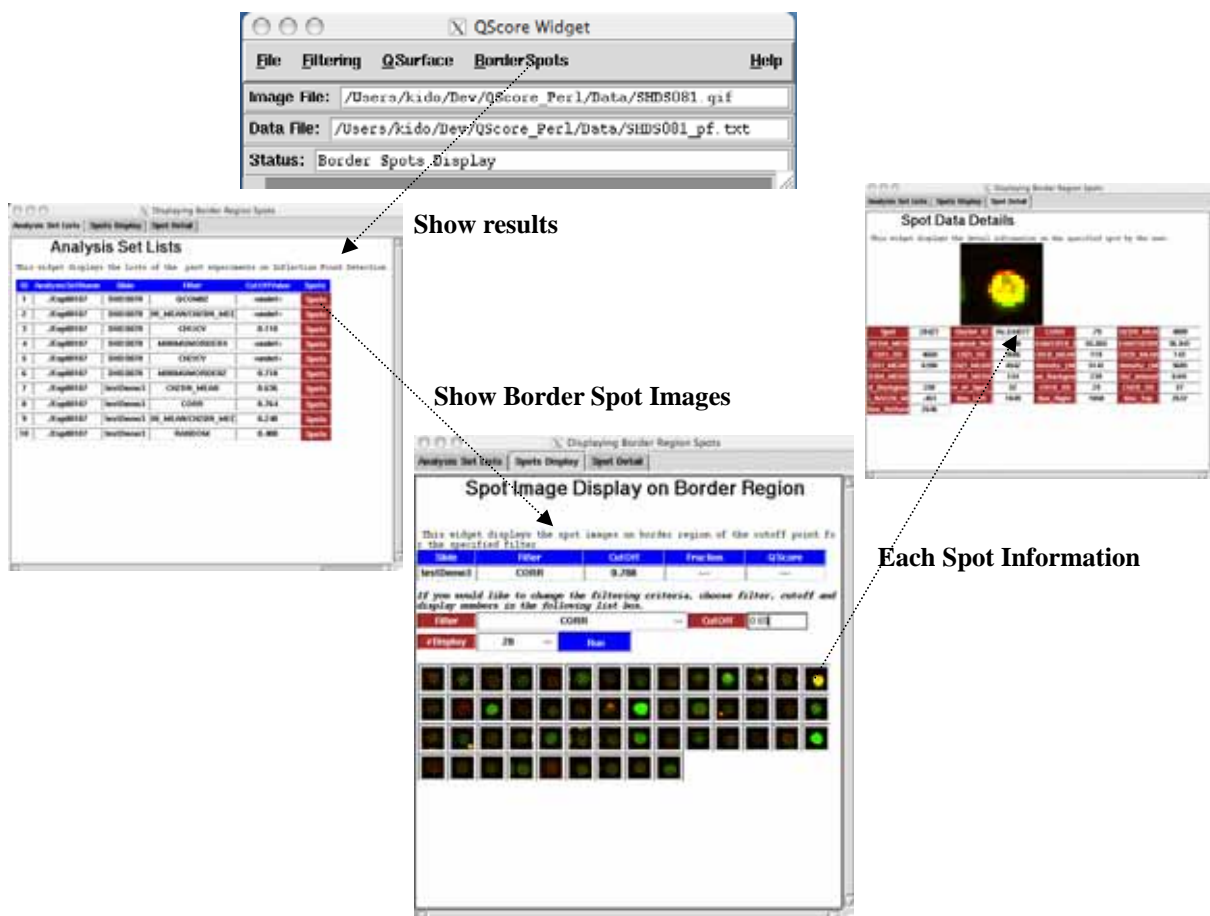


図 18 GUI の動作例 (境界領域のスポットの提示画面)

12. プロジェクト評価

ゲノム発現解析の品質評価に関して、着実に研究・開発を進めた。この分野での自分の立脚点をしっかり見定めて、今後の展開を構想することが重要である。

13. 今後の課題

本報告では本プロジェクトで開発したツールの機能モジュールやGUIを中心に内容をまとめ、フィルタリング指標の比較評価の実験結果やフィルタリング手法の比較検証結果、検証結果の解釈などについては詳述しなかった。これまでにマイクロアレーデータのクオリティ評価について様々なシミュレーション実験を通して多くのデータが蓄積されているが、これらについての考察は本報告とは別に論文としてまとめていきたいと考えている。今後の研究課題としては下記があげられる。

1. クラスタリング結果との整合性と QScore に基づくフィルタリングの評価

現在、提案している QScore は同一の遺伝子を測定すれば同一の測定結果が得られると仮定して複数の測定結果のばらつきをもとにクオリティを評価している。この方法は客観的で妥当ではあるが、同一の遺伝子を複数測定している事が前提であり、このような Replicates が十分に存在しない場合には信頼性が低くなる。期待する精度のクオリティを得るにはどの程度の Replicates が必要かを把握し解析精度の信頼性と Replicates の数との関係を定量的に把握することは重要である。

現在、行っているフィルタリングの評価はあくまで一つのアレイを対象に QScore をどれだけ改善したかという基準で比較をしているが、実際に本手法でフィルタリングした結果が最終的な解析結果のクオリティにどう影響を及ぼしているのかを評価することも重要である。

このため、現在、開発者はすでに正解がよく知られているデータセットを用いて、フィルタリングの結果がクラスタリングの正解率にどのように影響を与えるのかを調べるための研究計画を立てており、今後も研究を継続していく予定である。

2. 機械学習を用いた適応的ハイブリッドフィルタリング

現在、開発者は様々なマイクロアレーデータに対してどのようなフィルタが有効かについてのデータを集積している。また複数のフィルタを組み合わせることでフィルタリング効率と精度を高める手法についての研究と比較検証データを蓄積しているところである。前述したようにマイクロアレーのスポットには様々な属性があり、どのような属性に注目するとスポットのクオリティをうまく表現できるのか、QSCORE とフィルタやフィルタ同士の相関関係にはどのような特性があるのか、どのような方法でハイブリッドフィルタを構築すれば、フィルタリング効率を最も高められるのか等は興味深いテーマであり、今後、研究成果をまとめていく予定である。現在、QScore を教師信号として与えて Replicates セットをトレーニングデータとしてフィルタリング効率を高めるようにハイブリッドフィルタを機械学習により獲得させるというアイデアを検証する研究を進めている。

3. クオリティ評価結果の分類とノイズ原因の推定

2 で述べたフィルタ特性の分析や適応的ハイブリッドフィルタリングの研究は様々な解析結果を比較分類することでノイズの原因を推定したり、解析データを補正し標準化していくための有効な手がかりを与えられる可能性がある。今後の重要な研究課題である。