平成 25 年度新潟大学脳研究所
「脳神経病理標本資源活用の先端的共同研究拠点」
共同利用・共同研究報告書

# 大規模ゲノムデータを基盤とした情報解析手法の開発

研究代表者　近山　英輔 1)
研究分担者　宮下　哲典 2)
研究分担者　原　範和 2)
研究分担者　桑野　良三 2)

1) 新潟国際情報大学・情報システム学科
2) 新潟大学脳研究所・生命科学リソース研究センター・遺伝子機能解析学分野

**研究要旨**

多くの病気の発症と経過は、パーソナルゲノムの多様性が大きく関与していると考えられている。近年、次世代シークエンサーに代表される高精度・高速分析機器が開発され、出力されるゲノム情報は極めて大容量になってきているが、これを効率的に活用するアルゴリズム、又は遺伝子の時系列変動に対応した解析ソフトウエアの開発は充分ではなく、このような情報解析手法は、パーソナルゲノムに関わる研究に今後重要になってくると考えられる。そこで、大規模ゲノムデータとして、ヒト神経培養細胞の刺激応答に対する時系列エクソンアレイ解析結果を用いて、遺伝子発現量の時系列データの観点から、遺伝子ネットワークを解析するシステムの構築を試みた。その結果、ENCODE プロジェクト 41 種細胞の転写制御ネットワークと比較するシステムを構築することができた。また、次世代シークエンサーの大容量 DNA 配列情報を整理するプログラムも開発した。

## A. 研究目的

細胞内の全 RNA 発現状態は、ヒト細胞の多様な状態を反映する主要な指標の 1 つである。ゲノムからの遺伝子発現量をエクソン単位で包括的にカバーできるエクソンアレイ解析を用いることで、ヒト神経細胞の全遺伝子発現状態を定量し、細胞の複数の内部状態の区別を行うことが可能になる。ヒト神経細胞は様々な外部刺激に応答し、その内部状態を変えることから、外部刺激に対する応答を、細胞内部状態の時間変化として、つまりその代表指標としての遺伝子発現量の時間変化として、捉えることが可能になると考えられる。

複数の遺伝子全発現量を統計的に解析し、遺伝子ネットワークを構築するアルゴリズムが様々に開発されてきており、現在では高速計算機を用いて、大量データ解析を行い、多量の遺伝子間ネットワークの推定が行われている。しかしながら、そのような手法は相関値しか算出できず、実際の物理化学的な相互作用や動力学を推定できない欠点もあり、遺伝子ネットワークの解析手法はまだ発展途上にある。

本研究では、ヒトニューロブラストーマ SH-SY5Y、およびヒトグリオーマ U-251 MG を用い、それらに化学的な刺激を与えた初期状態から、時間を追って細胞内全 RNA 量が変化する様子をエクソンアレイ解析で定量し、その定量結果をゲノムワイドに解析するシステム構築を試みた。

## B. 研究方法（倫理面への配慮を含む）

大容量 DNA 配列情報を整理するプログラムは perl で実装した。次世代シークエンサーGenome AnalyzerIIx(GA IIx)から得られる複数のシート

に跨がる個別データを SNV または indel 情報ごとに整理するプログラムである。GA IIx から生成される FASTAQ ファイルを BWA、SAMtools、Picard でマッピングし、補正、duplicate marking、アラインメントを Picard、GATK で行い、SNV、INDEL を GATK で検出したファイルを入力した。アノテーションは GATK、snpEff を用いた。

ヒト培養細胞のエクソンアレイ解析では、GeneChipR Human Exon 1.0 ST Array を用いた。時間軸を初期値以外で 3 点設定した。転写産物データポイントは 1 サンプル（1 ジーンチップ）当たり約 100 万であった。培養細胞は SH-SY5Y（ニューロブラストーマ）、U-251 MG（グリオーマ）を用いた。SH-SY5Y に対し、KCl（濃度：50 mM）で処理し、脱分極を行った。SH-SY5Y、U-251 MG に対し、Dexamethasone による処理を行った。SH-SY5Y の KCl 処理では、処理時間 0（hr）0（3 点 [control]）、0.5（3 点）、2.0（3 点）、6.0（3 点）を行った（ジーンチップ計 12 枚）。Dexamethasone 処理では、0（3 点 [control]）、0.5（3 点）、1.0（3 点）、6.0（3 点）を行った（ジーンチップ計 12 枚）。

Java を用いて、エクソンアレイ解析結果ファイル、ヒト転写制御因子データ、ENCODE プロジェクト 41 種細胞のヒト転写制御ネットワークデータを読み込むシステムを開発した。遺伝子ネットワーク推定プログラム SiGN によるネットワーク推定を行った。

## C. 研究結果

以下に Java で構築した解析システム（ソフトウエア）の処理フローを示す。

1. エクソンアレイ解析結果読み込み
2. HumanTF 読み込み（Vaquerizas et al., *Nat. Rev. Genetics* (2009)）
3. ENCODE TF 読み込み（Neph et al., *Cell* (2012)）
4. ENCODE TF 細胞固有 network 読み込み（Neph et al., *Cell* (2012)）
5. エクソンアレイ FDR 計算
6. 41 種細胞固有制御関係と時系列変動の前後関係のマッチスコア自動計算
7. 細胞固有を 1 つ指定して、詳細を可視化、又は、ネットワークを時系列を考慮して可視化

図 1 に解析システムの実行結果の例を示す。この例は、SH-SY5Y/KCL の時系列エクソンアレイ結果を入力とし、ENCODE の SKNSH 培養細胞の転写制御ネットワークを指定した解析システムの出力である。この結果には、SRF 遺伝子発現量の時系列データ（代表値）、SRF を制御する転写制御因子遺伝子の発現量の時系列データ、SRF が制御する転写制御因子遺伝子の時系列データが図示されている。また、発現が見られなかった転写制御因子遺伝子については遺伝子シンボルがリストされている。遺伝子発現量時系列データの時間軸の重心が縦線で示されており、これは時間的順序を参照するためのものである。SRF 以外の他の転写制御因子遺伝子についてはタブをクリックすることで参照できるシステムになっている。

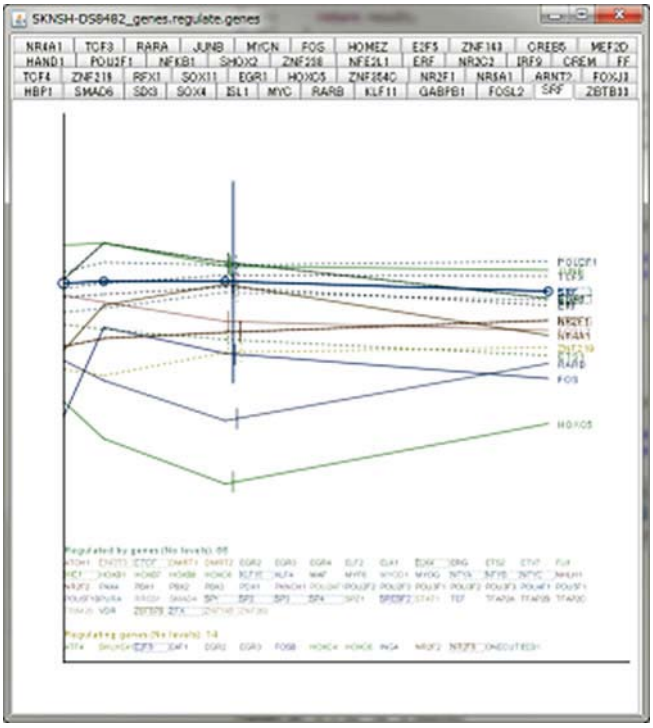


図 1　解析システムの実行結果

## D. 考察

ENCODE プロジェクトで明らかになった転写制御ネットワークとエクソンアレイ解析結果を比較すると、時系列で順序だった転写制御の ON/OFF 関係を見出すことは一見困難であった。これは、転写制御ネットワークのメカニズムに潜在するものか、ネットワークのノードとリンク数が大きいことによる複雑性に由来するものか、あるいは協同的よりも確率的な事象であることによるものであろう。

時系列解析を行うために、時間軸の重心計算を試みたが、時間的順序を定量する指標としては不十分であった。これを解決するには、具体的な微分方程式の数理モデルに時系列データをフィッティングさせ、転写制御の物理化学的な相互作用関係を推定する必要がある。そのためには、研究代表者の開発した数理的手法（発表 1）を今後用いたいと考える。

今回、ENCODE プロジェクト 41 種細胞のヒト転写制御ネットワークをエクソンアレイ解析結果に統合するシステムを実装でき、柔軟に遺伝子ネットワーク記述が可能な Java の生化学ネットワーク解析コアクラスライブラリを実装できた。しかしながら、遺伝子ネットワーク推定ソフト SiGN を用いた遺伝子相関ネットワークの解析結果を統合することができなかった。今後の展望としては、転写制御される非転写制御因子遺伝子、コアクチベーター、コリプレッサー、タンパク-タンパク相互作用、シグナル伝達、代謝経路、GO 等のデータベースからのデータを統合解析するような改良が期待される。

## E. 結論

SH-SY5Y（ニューロブラストーマ）、U-251 MG（グリオーマ）の化学刺激応答の時系列エクソンアレイ結果を、ENCODE プロジェクト 41 種細胞のヒト転写制御ネットワークと統合解析するシステムを開発した。そこでは、柔軟に遺伝子ネットワーク記述が可能な Java 言語の生化学ネットワーク解析コアクラスライブラリを実装できた。また、次世代シークエンサーの大容量 DNA 配列情報を整理するプログラムを perl 言語で開発した。

## F. 研究発表

**1. 論文発表**　なし

**2. 学会発表**

1. 近山英輔, ″非線形関数のステップ関数表示の公式″, 日本生物物理学会第 51 回年会, 京都, 2013 年 10 月 28～30 日

## G. 知的財産権の出願・登録状況（予定を含む）

1. 特許取得　なし
2. 実用新案登録　なし
3. その他　なし