平成 28 年度　利用報告書

課題名（タイトル）：

Development of machine learning techniques for DNA sequencing data

利用者氏名：○二階堂愛, 尾崎遼, 露崎弘毅, 石井学, 圍野宏樹, 芳村美佳
所属：情報基盤センター バイオインフォマティクス研究開発ユニット

１．　本課題の研究の背景、目的、関係するプロジェクトとの関係

次世代 DNA シーケンサー(NGS)は大量のデータを出力するが、そのデータから知識を取り出すには大規模な計算が必要となる。また NGS は生命現象の様々な階層(RNA, DNA, クロマチン状態)の情報を出力する。これらの情報をいかに統合し新規知見に結びつけるかが課題となる。そこで我々は深層学習を始めとする機械学習アルゴリズムを用いて、エピゲノムデータの統合に挑む。また大量の 1 細胞 RNA-Seq のデータから細胞タイプを予測するアルゴリズムの開発を行う。アルゴリズムの高速な実行ために GPU を利用した開発を行う。

２．　具体的な利用内容、計算方法

近年重要視されるようになった新しいタイプのエピゲノムデータの性質を詳細に理解するため、データ解析パイプラインの整備と基礎的な性状の解析に取り組んだ。

３．　結果

(1) 新規なクロマチンアクセシビリティ測定手法 ATAC-seq のデータを解析するパイプラインを作成し、ATAC-seq の実データに適用した。その結果、既存の測定手法 DNase I-seq に比べて、より少ない細胞数でオープンクロマチン領域および転写因子ネットワークを推定できることが判明した。(2) total RNA-seq および polyA RNA-seq データから新規転写単位を発見するパイプラインを開発し、さらにエンハンサー領域のアノテーションと統合することで、エンハンサーRNA の転写単位を発見することができた。

４．　まとめ

新しいタイプのエピゲノムデータに機械学習手法を適用する上で理解すべき基本的な性質を明らかにすることができた。

５．　今後の計画・展望

今後、ATAC-seq データおよびエンハンサーRNA のデータを含むエピゲノムデータの統合を目指す。また、隠れマルコフモデルを用いた 1 細胞 RNA-Seq のデータから細胞タイプを予測するアルゴリズムの開発に関しては、公共データベースにある RNA-Seq データの収集とデータ整形を進めている。