平成 27 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

深層学習の一種である stacked autoencoder を実装し、エピゲノムデータを入力とすることで、中間層の特徴表現としてクロマチン状態（エピジェネティック修飾の組み合わせ）を学習できるかを検討した。 米国 ENCODE project より公開されている種々のヒストン修飾を測定したデータを用いた。

３．　結果

stacked autoencoder の速度を CPU と GPU で比較した結果、GPU を用いた方が高速であること分かった。実データに適用した結果、エポック数（パラメータ更新回数）を当初の見積もりよりも増やす必要があることが示唆された。また、欠損値がある領域については、入力データから除いた方がよいことが示唆された。

４．　まとめ

stacked autoencoder をエピゲノムデータへの適用することで、深層学習を用いてクロマチン状態を学習する際に解決すべき課題を明らかにできた。

５．　今後の計画・展望

今後、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。