平成 30 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(1)、尾崎遼(1)，露崎弘毅(1)，石井学(1)，芳村美佳(1)

理研における所属研究室名:

(1)生命機能科学研究センター　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大規模な 1 細胞 RNA-seq データの類似性検索アルゴリズムの開発に取り組んだ。

３．　結果

(1) 大規模な 1 細胞発現データを高速に検索するためのデータ構造を開発した。またそれを Julia 言語を用いて実装した。 (2) Locality sensitive hashing を用いた遺伝子発現類似性検索ソフトウェア実装した。以上の 2 つのソフトウェアにより、検索精度を下げることなく、従来法の 100 倍程度の検索速度を実現した。

４．　まとめ

巨大な 1 細胞発現データを高速に検索するアルゴリズムとそのソフトウェアの実装に成功した。

５．　今後の計画・展望

現在は、数百万エントリのデータベースへの検索であるが、今後、数億エントリーのデータに対する検索が実施できるアルゴリズムを開発する。RNA-Seq データの収集とデータ整形を進めている。

６．　利用がなかった場合の理由

今年度はアルゴリズムの開発に注力したため、HOKUSAI での大規模な計算には至らなかった。

平成 30 年度　利用研究成果リスト

【雑誌に受理された論文】

なし

【会議の予稿集】

なし

【口頭発表】

二階堂愛．ライフサイエンス研究の生産性を向上させるオンデマンドクラウド．国立情報学研究所学術情報基盤オープンフォーラム 2018．2018 年 6 月 20-21 日．東京．

【ポスター発表】

なし

【その他（著書、プレスリリースなど）】

なし