2019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: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チーム

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データの次元圧縮アルゴリズムの開発に取り組んだ。

3.　結果

(1) 大規模な1細胞発現データを次元圧縮するための最適化アルゴリズムやデータ構造を検討・実装した。これにより、従来、一般的なメモリ規模で次元圧縮が困難であった100M個の1細胞トランスクリプトームデータな高速・省メモリで次元圧縮を実現した。

4.　まとめ

100M細胞規模の巨大な1細胞発現データを高速・省メモリで次元圧縮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、テンソル分解への応用を進めている。これを用いて1細胞トランスクリプトームに異質データが統合されたデータから知識抽出するアルゴリズムを開発する。

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。