2020 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)，市川巧(1)，芳村美佳(1)，亀田健(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。このような解析を実現するには多様なソフトウェアやデータベースを複雑に組み合わせて、解析ワークフローを実装・運用する必要がある。またそれぞれのデータベースやソフトウェアはバージョンアップがあり、ワークフローに改良や結果の評価が頻繁に必要となる。ゲノムデータは時間に対して定常的にデータが得られるわけではなく、生物や実験の都合によって突発的に得られるため、計算環境も突発的に用意する必要がある。このようなゲノムデータの解析環境を取り巻く問題を解決するために、データ解析ワークフローの標準化、計算環境のソフトウェアによる自動構築などに取り組む。

2.　具体的な利用内容、計算方法

ワークフローを標準ワークフロー言語 NextFlow によって実装する。ソフトウェアについてはコンテナ仮想技術を利用して実行再現性やポータビリティを担保する。これらの技術を用いて 1 細胞 RNA-seq のデータ解析ワークフローを開発する。HOKUSAI SellingShip 上に DevOps 技術を用いて自動的に計算環境を整えるソフトウェアを開発する。分子動力学計算については GPU ノードを利用するワークフローを実装する。

3.　結果

1 細胞 RNA-seq のワークフローとして RamDAQ を開発した。標準ワークフロー言語 NextFlow によって実装した。SellingShip の管理に使われている OpenStack をソフトウェアから操作し、必要な計算資源を動的にオンデマンドに得るプログラムを開発した。その結果、数分程度で PC クラスタを構築できることを確認した。さらにその計算環境で、RamDAQ を実行できることを確認した。これらの計算速度等を比較するため、パブリッククラウドやベアメタルの PC クラスタで同様の計算を行い、結果を比較したところ、クラウドや PC クラスタを同等の結果を実務上遜色ない計算時間で得られた。分子動力学については GPU ノードを利用するワークフローを開発し実行できた。

4.　まとめ

ゲノムデータ解析環境を HOKUSAI 上に自動構築できるソフトウェアの開発に成功した。

5.　今後の計画・展望

ワークフローの機能拡張を行う。BDR を始め全理研でゲノムデータ解析を実施している研究センターや研究室に技術提供を行う。またほかのシーケンス技術のワークフローに実装をすすめる。

6.　利用がなかった場合の理由