平成 29 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

理研での所属研究室名：本所 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

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収集と整形する。このデータセットを用いて 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するアルゴリズムの開発を進める。

これらの計算パイプラインを docker とジョブスケジューラーを利用して構成する。これらのシステムは、DevOps 技術を利用して自動的に構成できるようにする。

３．　結果

ローカルのコンピュータを利用し、データの収集と整形を進めた。エンハンサーRNA の定量法、可視化法を開発し、プログラミング言語 Julia で実装した。

計算パイプラインで利用される計算機やソフトウェア設定を Chef を用いてコード化した。これらのソフトウェアが正しく構成されることをローカルの PC クラスタとパブリッククラウドを用いて確かめた。

４．　まとめ

・データの収集、整形、アルゴリズム開発を進めた。

・計算環境の自動構成については完成し、ローカル環境やクラウド環境でのテストが終了した。

５．　今後の計画・展望

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アルゴリズムの開発を進める。同様に、深層学習を用いた 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。さらに、HOKUSAI のシステムで自動構成が可能をテストする。

６．　利用がなかった場合の理由

研究室内のローカルのクラスタやクラウドでのテストを先行させたため。

平成 29 年度　利用研究成果リスト

【論文、学会報告・雑誌などの論文発表】

1. Sasagawa Y, Danno H, Takada H. Ebisawa M. Hayashi T. Kurisaki A and Nikaido I. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol. (in press)
2. MuhChyi Chai, Tsukasa Sanosaka, Hironobu Okuno, Zhi Zhou, Ikuko Koya, Satoe Banno, Tomoko Andoh-Noda, Yoshikuni Tabata, Rieko Shimamura, Tetsutaro Hayashi, Masashi Ebisawa, Yohei Sasagawa, Itoshi Nikaido, Hideyuki Okano, and Jun Kohyama. The chromatin remodeler CHD7 regulates stem cell identity of human neural progenitors. Genes and Development. 2018.
3. Koki Tsuyuzaki and Itoshi Nikaido. Biological Systems as Heterogeneous Information Networks: A Mini-Review and Perspective. HeteroNAM18.
4. Hayashi T, Ozaki H, Sasagawa Y, Umeda M, Danno H, and **Nikaido I**. Single-cell full-length total RNA sequencing uncovers dynamics of non-polyadenylated RNAs, recursive splicing and enhancer RNAs. Nature Communications. 2013.
5. Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, **Nikaido I.** SCODE: An efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. **Bioinformatics**. 2017.

【国際会議などの予稿集、proceeding】

特になし

【国際会議、学会などでの口頭発表】

1. Itoshi NIKAIDO. Disassembling regionalization of neuroectoderm by high throughput single-cell RNA-sequencing. The 12th International Workshop on Advanced Genomics. 2017.

【その他】

- 研究開発クラウドの衝撃 - 1 クリックでビッグデータ解析環境を展開できるクラウドでオープンイノベーションを加速 - 日経ビジネスオンライン. 2017/08/22.
- 国際競争の激しいゲノム研究に新たな活力を。オープンソース × Microsoft Azure 活用で、先進のデータ解析環境をより多くの大学・研究機関へ提供. Microsoft for Business. 2017/05/12.