平成 23 年度 RICC 利用報告書

課題名（タイトル）：

次世代シーケンサーのバイオインフォマティクス解析

利用者氏名：二階堂 愛
理研での所属研究室名：神戸研究所 発生・再生科学総合研究センター 先端技術支援・開発プログラム ゲノミクス解析室 機能ゲノミクスユニット

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

本課題では、RICC 上に次世代シーケンサーのデータ解析パイプラインを構築することを目的とした。次世代シーケンサーは一度の実験で、数百 Gb のデータを出力する。このデータを高速にデータ解析し、生物学的な知識に結び付けるには、大量の計算リソースが必要とされる。次世代シーケンサーのデータは生体サンプルや遺伝子単位など細かいユニットに分割することができるため、計算も RICC のように大規模な CPU 数を誇るコンピュータで分散計算しやすいデータである。これらのことから、私は RICC を利用し、次世代シーケンサーデータに対して一連の解析が行えるようデータ解析パイプラインを構築することを目指した。特に、タンパク質がゲノム上のどの位置に結合しているかを調べることができる ChIP-seq のデータ解析パイプラインを RICC に構築する。

２． 具体的な利用内容、計算方法

解析パイプラインを以下の 4 つのステップに分け、パイプラインを構築した。(1)前処理（データクオリティ評価、シーケンスデータのゲノムへのマッピング、データフィルタリングなど)、(2)ピーク発見（タンパク質が結合した場所を特定する計算)、(3)タンパク質結合サイトの DNA モチーフ検索、(4)異なるタンパク質の結合地図の比較。このステップを約 20 個以上の Java や Ruby, C/C++, Python, R で書かれたプログラムを Rake で繋ぎ合せ、Grid Engine によるジョブスケジューラを利用し、分散計算ができるように工夫した。

３． 結果

パイプライン構築の結果、3 つのタンパク質(転写因子)のそれぞれ 14 データポイント、合計 42 の ChIP-seq データ（約 400 GB 相当)のデータ解析を同時に行うことができた。ただし、RICC がとても混雑しているため、ジョブ投入待ちのコストがボトルネックになった。またデータが大量なので RICC に送る時間がかかることも問題となった。

４． まとめ

RICC 上に次世代シーケンサーデータ解析パイプラインを構築し運用した。

５． 今後の計画・展望

今年度は、ChIP-seq のデータ解析パイプラインを中心に実装・運用してきたが、今後は、遺伝子の発現量を包括的に調べる実験技術である RNA-seq 解析のパイプラインについても準備を進める。

６． RICC の継続利用を希望の場合は、これまで利用した状況（どの程度研究が進んだか、研究においてどこまで計算出来て、何が出来ていないか)や、継続して利用する際に行う具体的な内容

RICC の利用により ChIP-seq データ解析パイプラインを完成させることができた。ただし、この分野のプログラムや解析アルゴリズムは日々改善されており、パイプラインの更新やチューニングを継続する。また RIKEN CDB にシーケンサーが導入されたため、データ解析の需要はますます増えるため、今後も実データでのパイプラインの運用を続ける。前項目の繰り返しになるが、RNA-seq のデータ解析パイプラインの需要も増えているため、この実装を進める。

７． 利用研究成果が無かった場合の理由

RICC で構築したデータ解析パイプラインを利用して、データ解析を行なった結果に基づいた実験が終了しており、論文を投稿済みである。またデータ解析パイプラインを構成するプログラムの一部、特にタンパク質結合地図を比較するプログラムについてもプログラムの公開と論文投稿の

準備を行っている。