平成 28 年度 先端的計算科学研究プロジェクト 成果報告書
大規模情報処理パイプラインによる次世代シークエンサーによるエピゲノム解析

大川恭行（九州大学生体防御医学研究所トランスクリプトミクス分野）

1. 背景

生命の設計図であるゲノムの遺伝情報は、クロマチン構造上の DNA やヒストンの化学修飾により選択的な発現制御が行われる。クロマチン構造上のゲノムの化学修飾はエピゲノムと呼ばれ、個々の細胞の状態、発生分化における運命決定を担う。

エピゲノムの本質を理解するためには、動的な構造変換機構担うクロマチン制御因子、クロマチン構成分子等の構造解析から、4 オングストロームの大きさでおよそ隙間なく全長２ｍにもおよぶゲノムに密に配置された全ヒストンの分布および特性を調べる分子スケールのエピゲノム解析と、特定の組織及び細胞レベルでの全遺伝子の測定による個体スケールでの表現型解析までを横断的に行う必要がある。

そこで、本研究プロジェクトでは、構造生物学、エピゲノム、情報生物学、そして、個体レベルの解析を専門とする４つの研究グループにより、九州大学情報基盤センターを拠点としたマルチオミクス解析を行うことを目指した。

2. 本年度の経過と成果：計算パイプラインの構築とその実行環境の整備を踏まえて

分子レベルでの解析では、神田グループが担当し電子顕微鏡を用いた単粒子解析を行った。本プロジェクトでは CX400 を用いた並列計算による電顕単粒子解析の高速化を行った。電顕単粒子解析では数万～数百万の蛋白質等の電顕像から立体構造を算出する手法である。近年、検出器及び解析法の発展により結晶解析レベルの分解能に到達しており今後の解析数が劇的に増大することが見込まれている。構造計算ソフト relion が進展に大きく寄与、ベイズ統計を利用するため計算量が増大しており、一般的に数百～千コアの計算機リソースが必要である。本年度は試行として小規模データを用いた予備検討を進めた。(72x72 画素の粒子像 27,624 枚)からクラス平均像（クラス数１００）を算出した。バッチジョブによる Xeon E5-2680 の 512 core の占有使用によるクラス平均像(Class2D)解析を行った結果、計算リソースにほぼ比例した速度向上が認められローカルの計算資源と比して 5-8 倍程度の高速化が達成された。実際のデータは今回のテストケースの 100-1000 倍となることからデータ量の増大 (数 TB～数十 TB)が見込まれ、センター利用のためのデータ送受信がボトルネックとなることが予想され、高速回線の利用等が課題となりそうである。

エピゲノム解析、個体レベルの転写量測定は、大川・林グループ担当し、エピゲノム情報解析を高速シークエンサーにより行った。多能性幹細胞である ES/iPS 細胞から、個体に発生しうる卵子を試験管内で分化誘導することを目的とし、卵子誘導培養法を確立した。

ES 細胞由来の卵子と生体由来の卵子との遺伝子発現比較を行ったところ、両者は非常に近い発現パターンを示した。このことから ES 細胞由来卵子は機能的であることが推測された。実際に ES 細胞由来卵子を生体に移植すると、健常な産児が得られた(Hikabe et al *Nature* 2016)。本解析において、従来 350 時間かかっていた、総計 1,805,494,592 塩基のゲノムへのアラインメント、種々の統計処理が、CX400 の利用により 25 時間に短縮された。得られたシーケンスデータのプロファイルのために、国際エピゲノムプロジェクト(iHEC)等で解析された 5000 件のシーケンスデータとの網羅的比較解析を行っている（取得データ数×５０００件の解析)。現在、1 件の比較解析あたり PRIMERGY CX400 環境下で 20 分の演算時間を必要としているが、並列計算の導入により現実的な計算時間となることが見込まれた。一方で、海外サーバ（遠隔地）で公開されている解析前の RAW データの総量が約 1PB、解析終了後データが約 10TB 程度以上でデータ転送の問題が起こり現在高速なデータ転送技術の導入検討を情報通信機構の村田グループとともに検討している。
本研究プロジェクトでは、学内外の参加研究室の解析環境から解析結果までを互いに共有し、分子構造レベルから個体表現型をつなぐための多階層解析拠点の形成を目指した。生物資源の有効利用、癌などの病気の解明と治療法開発、再生医療の実現などの諸問題に対して成果が短期的に上がったものや今後大きな成果を上げるインフラ整備に大きく貢献したと考えている。

今後の展望

本プロジェクトのエピゲノムパイプラインの解析は、今年度も 200 解析を超え、1 件当たりの解析量が増大しつつある(Kato et al *Science*, 2017)。本プロジェクトでは今後連携が見込まれる単粒子解析のチームも参画し、エピゲノム分野における多角的かつ横断的な計算環境の樹立を図った。課題として、外部あるいは内部の巨大データの効率のよい転送が挙げられる。今後の大規模データ解析を勘案して本問題にも取り組んでいきたい。

成果論文

Crystal structure of the overlapping dinucleosome composed of hexasome and octasome.
Kato D, Osakabe A, Arimura Y, Mizukami Y, Horikoshi N, Saikusa K, Akashi S, Nishimura Y, Park SY, Nogami J, Maehara K, Ohkawa Y, Matsumoto A, Kono H, Inoue R, Sugiyama M, Kurumizaka H.
Science. 2017 Apr 14;356(6334):205-208.