［報　告］計算科学・計算機科学人材育成のためのスーパーコンピュータ無償提供利用報告

農学部植物生命科学コースでの大量 DNA 塩基配列解析演習

— 学部 3 年生「学生実験 II」における次世代シーケンスデータ解析演習 —

宮下脩平
東北大学大学院農学研究科応用生命科学専攻植物病理学分野

東北大学農学部植物科学コースでは、学部 3 年生の演習「学生実験 II」(必修)の一部として、サイバーサイエンスセンターの「計算科学・計算機科学人材育成のためのスーパーコンピュータ無償提供」制度を利用した大量 DNA 塩基配列解析演習を行った。この解析演習は、近年農学研究において一般的な手法となってきた「次世代シーケンス」で得られる大量データを解析するための初歩を学生が体験することを目的としたもので、具体的な解析そのものよりは、CUI でのスパコンとのやり取り、データ形式やそれぞれの解析手法の軸となる考え方、実験手順とデータの実際的な関係性を、演習を通して理解することに主眼を置いた。今年度初めて行う演習であったが、サイバーサイエンスセンター関係各位のご協力により演習の目的を達成することができた。まず御礼申し上げたい。

生物は DNA を遺伝情報の担体とする。DNA は 4 種類のヌクレオチドが連なったものであり、それぞれがもつ塩基(アデニン、シトシン、グアニン、チミン)の並びである「塩基配列」を調べることは、遺伝情報を知り、ひいてはその生物を知ることにつながる。例えばヒトは約 30 億塩基対の DNA をゲノム(遺伝情報の総体)とする。2003 年に完了したヒトゲノムプロジェクトでは、この 30 億塩基対の配列を従来法で数百塩基ずつ決定したため、膨大な時間と労力、コストを要した。しかし 2005 年頃に「次世代シーケンサ」が実用化して以来、状況は一変した。次世代シーケンスでは大量の DNA 断片の塩基配列を並列的に決定する工夫がなされており、例えば illumina 社で最も大量の塩基配列を決定できる NovaSeq システムでは、44 時間で 3 兆塩基もの配列を決定できる(2019 年時点)。また従来法に比べて塩基当たりの配列決定コストが非常に低いことから、ヒトやシロイヌナズナといったモデル生物以外の生物のゲノム塩基配列も続々と決定されているほか、mRNA(DNAの情報が一時的・部分的に写し取られたもの)を逆転写して DNA に変換後に大量解析することで、mRNA の定量解析も可能となっている。これにより例えば、異なる生育条件での植物の遺伝子発現量の網羅的な比較を行うことなどが可能であり、そのような研究例は実際に増えてきている。次世代シーケンサを用いることは、生命科学研究・農学研究において近年では日常的になりつつあるといえる。

大量の塩基配列データが得られるようになったことで、それに対応する解析が必要となった。次世代シーケンサが実用化した当初、そのような解析は主にデータ解析の専門家が行うものであったが、近年では一般的な解析については GUI で簡単に行うシステムも存在し(ただし高額な料金をとられる)、多くの研究者がそのようなシステムを利用している。しかし解析の仕組みを全く理解せずにシステムに生データを投入して「結果」を得ることは、間違った解釈をもたらす可能性があって非常に危険である。同様に、解析ができる共同研究者にデータを「丸投げ」して結果を得る場合もミスコミュニケーションを生じることが多く、問題が多い。一方で解析の仕組みを知っていれば、適切な実験計画を立てて過不足ない量のデータで効率的に研究を進められるほか、お仕着せではないオリジナルの実験・解析系の開発が可能となりうる。共同研究者とのコミュニケーションも良好になり、新しい発見につながるかもしれない。しかし多くの生物学系・農学系研究者にとって CUI での解析は心理的ハードルが高い。そこで本演習では、研究の実戦経験のない 3 年生のうちからそのような解析を体験し、雰囲気に慣れる機会を提供することで、将来先入観なく実験と解析の両方に取り組める人材の育成につなげることを目指した。

演習は 2019 年 11 月 13 日、14 日の 2 日間、計 4 講時にわたって以下のスケジュールで行われた。

- 1 講時目： 公開鍵暗号方式によるスパコンとの通信および Linux の基本的な操作の習得
- 2 講時目： 次世代シーケンスデータ形式の理解と簡単な分析
- 3 講時目： DNA 断片配列の Trinity による de novo アセンブル
- 4 講時目： blastn による類似配列検索

なお演習にあたっては各学生・TA および教職員用に教育用アカウントを発行していただいたほか、解析に用いるソフトウェアのインストール場所、解析に用いるデータの置き場所としてのアカウントも発行いただいた。

演習に用いたデータは、東北大学植物園や農学研究科実験圃場の植物から抽出した二本鎖 RNA を逆転写し、増幅して得られた DNA の塩基配列を次世代シーケンサ（illumina 社 MiSeq）で網羅的に決定したものである。二本鎖 RNA は RNA ウイルスに特徴的であることから、得られた DNA 配列をつなぎ合わせ（de novo assemble: 参考になる配列なしにつなぎ合わせを行うこと）、それと似た配列を既知のウイルス配列データベースから探索することで、上述の植物に感染している RNA ウイルスを検出し、同時に配列決定できることが期待される。実際に演習参加学生は、上述の解析を通して植物園のシロツメクサおよび実験圃場のアカツメクサから複数のウイルスが検出できることを体験した。慣れない解析にもかかわらず全員が意欲的に取り組み、想定時間内に演習を終えられたほか、特に意欲の高い学生が追加の演習課題に取り組む様子も見られた。また演習の合間には、私共の研究室から TA として参加した佐々木稜太氏（修士 1 年）と武田萌氏（学部 4 年）が次世代シーケンスを利用して得た成果を含む自らの研究の紹介をそれぞれ行った。実際に身近で行われている研究を知ることで、演習参加学生の理解が深まったものと期待している。

謝辞

上述のように、本演習を行うにあたってサイバーサイエンスセンターの皆様には多くのご協力をいただきました。御礼申し上げます。また、東北大学植物園でのサンプリングでは技術職員の津久井孝博氏をはじめとする皆様にお世話になっております。この場を借りて御礼申し上げます。

左　　演習の様子
下左　　佐々木稜太氏による研究紹介
下右　　武田萌氏による研究紹介