解説

医療，農学，環境分野におけるビッグデータ解析

石井　一夫

はじめに

次世代シーケンサーの出現と普及により，医学，薬学，農学，生物学などゲノム科学を基盤とするライフサイエンス，グリーンサイエンスでは，ゲノム塩基配列などデータ産生量が爆発的に増えている．そのデータサイズは，数ギガバイトから数テラバイトにおよび，しばしばペタバイトレベルに達することもある．我々は，これらの大規模データ解析の高速化，最適化のため，（1）乱数発生によるデータの標本抽出法であるモンテカルロ法による標本抽出法を用いたり，（2）クラスター分散処理システムによる並列処理を行なうためHadoopなどのビックデータ処理のための解析システムを用いたりしている．これにより従来では考えられなかった高速大量の生物学データの計算が可能になり，新たな知識発見が可能となった．本稿では，これら分散処理システムなどを用いたゲノム科学におけるビッグデータ処理について概説する．

大規模データを処理するために

ビッグデータという言葉は，FaceBookなどのソーシャルメディアにおけるウェブログの解析やAmazonなどのレコメンデーション，ウェブ広告などにおいて増え続ける大量のデジタルデータを意味することが多い．これらの多くはテキストデータである．次世代シーケンサーから産生されるゲノム塩基配列データやそれらを扱う各フォーマット形式のデータの多くもテキストデータであり，ビッグデータ処理のために開発されたシステムを適用できる[1]．我々は，以下の4つの方法を用いて，これらのデータを処理している．

（1）モンテカルロ法により大量データから無作為にサンプルを抽出し，元のデータをシミュレーション[2]．
（2）マルチスコアマルチスレッドのCPUを搭載した大容量メモリサーバによる大量並列処理（ハイパフォーマンスコンピューティング）．
（3）Hadoopと呼ばれるビッグデータ処理のためのフレームワークを用いてコンピュータクラスタにファイルを分散し，計算プロセスを分散して処理．
（4）上記のHadoopを用いずに，シェルなどの簡単なスクリプトを組み合わせてコンピュータクラスタにファイルを分散し，計算プロセスを分散して処理．

ビッグデータ処理のためのフレームワーク

ビッグデータ処理専用のフレームワークとしてApache Hadoop（以下，Hadoop）というオープンソースシステムがよく用いられる．これは，検索システムGoogleのコンポーネントである分散ファイルシステムの「GFS（Google File System）」，分散ロックシステムの「Chubby」，並列プログラミングモデルの「MapReduce」，キー・バリュー型データストアの「BigTable」，プログラミング言語の「Sawzall」などがオープンソース化され誰でも利用できるようになったものである．

Hadoopは，これらのうち分散ファイルシステムHDFS（Hadoop Distributed File System）[3]および，ビッグデータ処理のための分散処理モデルであるMapReduce[4]を基盤とするビッグデータ処理フレームワークである．数百台，数千台のコンピュータからなるクラスターを用いて計算プロセスを分散して別々に処理し，それを集計して計算結果を得る（図1）．

通常の大学などの研究室ではこのような大規模システムを構築し維持することは現実的でないので，AWS（アマゾンウェブサービスAmazon Web Service）のようなクラウドシステムを利用してデータ解析を実施する．AWSでは，Hadoopフレームワークを利用して莫大な量

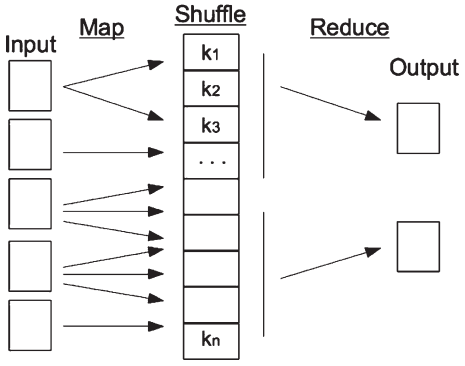

図1．Hadoop MapReduceの分散データ処理行程

著者紹介　東京農工大学農学府農学部農学系ゲノム科学人材育成プログラム（特任教授）　E-mail: kishii@cc.tuat.ac.jp

表1．アマゾンウェブサービスで利用できる生物データセットの例

データセット	内容説明
1000 Genomes Project	約200TB，1700人分超のゲノム
Ensembl	ヒトおよび約50種の生物のゲノム
Genbank	NIHのゲノムデータバンク
YRI Trio Dataset	ナイジェリアのヨルバ人のゲノムシーケンスデータ
The Cannabis Sativa Genome	「麻」のゲノム情報
UniGene	NCBI提供の遺伝子情報
Influenza Virus	NCBI提供のインフルエンザウィルス情報

のデータ処理ができるサービスであるAmazon Elastic MapReduce（Amazon EMR）が利用できる．今後，類似のサービスが増えて行くと思われ，クラウド環境を利用したビッグデータ分析の実施例が増えて行くと思われる．

クラウドシステムにおけるゲノム科学データ解析

クラウドシステムのゲノム科学への応用例としては，200TBにも及ぶ1000人分のゲノム配列情報データベースを作成するために開始された1000人ゲノムプロジェクトの成果が，AWSから利用できるようになったことが上げられる[5]．AWSでは1000人ゲノムプロジェクト以外にも多くのゲノム解析ツールや公共データベースが利用可能になっており，今後クラウド環境でのゲノム科学に関するデータ解析が普及してくるものと考えられる（表1）．

次世代シーケンサーのデータ解析は，大きく三つのステップに分けて考えられている（図2）．（1）一次解析：画像データから配列データを抽出するまでの行程である．この行程は次世代シーケンサーが正常に動けば，ほぼ自動で生成してくれる場合も少なくない．普段の解析ではあまり意識することがない．（2）二次解析：新規解析のDNA配列データのアセンブリや，既知の配列に対して整列を行うマッピングを行なう行程である．（3）三次解析：マッピングやアセンブリで得た結果を計数したり，有意差検定を行ったりしてデータ解析を行なう行程である．

クラウド環境においてHadoop上で動作する次世代シーケンサー解析ソフトウェアは，論文で確認できるだけでも数十から数百種類存在し，数百以上のノードをもつコンピュータクラスタを利用することで，従来型のサーバーで用いられるソフトウェアよりもかなりパワフルな解析が可能となっている．おもなものを以下に紹介する．

（1）Crossbow: Hadoopを使用して次世代シーケンサーデータをマッピングし，多型を検出するソフトウェア．
（2）Contrail: Hadoopを使用してde Brujinグラフ理論によりゲノムアセンブリを行なうソフトウェア．
（3）Myrna: Hadoopを使用してBowtieを用いてマッピングを行ないR/Bioconductor を用いてRNAの発現定量解析を行なうソフトウェア．
（4）GATK（Genome Analysis Toolkit）：Javaベースで動作するMapReduceのフレームワークを取り入れた遺伝子多型解析用ソフトウェア．

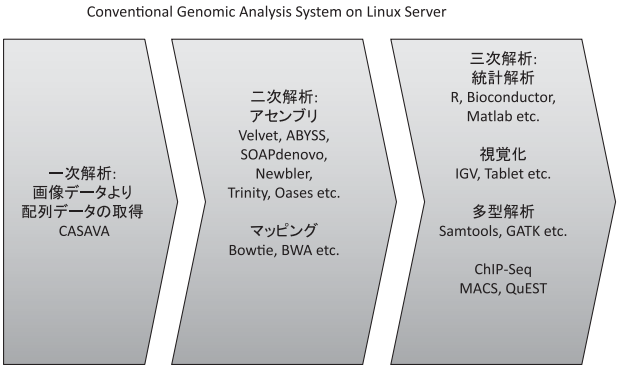

図2．次世代シーケンサーのデータ解析のフローチャート

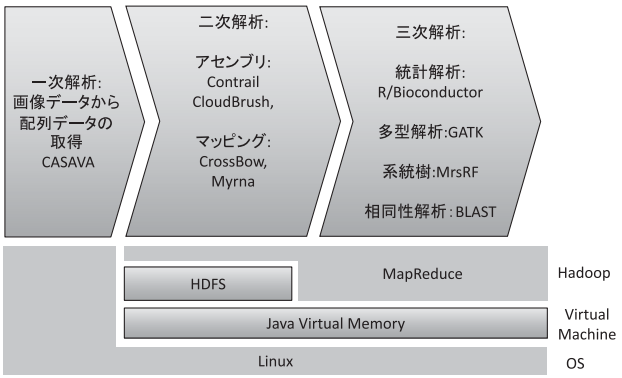

図3．クラウド上で動作する次世代シーケンサーのデータ解析ソフト

筆者らは，これらの解析環境を用いてモンテカルロ法や並列分散処理を利用し，従来では達成できなかった大規模計算を実施し，新規の生物知識の発見を可能にしている．たとえば，次世代シーケンサー配列データのより詳細なクオリティチェック，ゲノムスケールの進化系統樹の最適化，超多次元データを用いた臨床診断薬の開発などである．

文　献

1) 石井一夫ら：日本統計学会誌，**43**, 90 (2013).
2) Rizzo, M.著，石井一夫，村田真樹共訳：Rよる計算機統計学，オーム社 (2011).
3) Ghemawat, S., *et al.*: *19th ACM Symposium on Operating Systems Principles, Lake George, NY, October* (2003).
4) Dean, J. and Ghemawat, S.: *Simplified Data Processing on Large Clusters, Symposium on Operating Systems Design and Implementation* (2004).
5) 1000 Genomes Project: http://aws.amazon.com/jp/1000genomes/