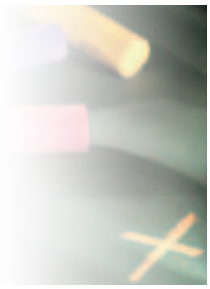
小特集

生命情報学が直面する大規模ゲノムデータ時代の課題

編集にあたって



片山 俊明
東京大学医科学研究所
ヒトゲノム解析センター

情報爆発は生命科学の分野においても生じている．2003 年のヒトゲノム計画の完了宣言を受けて，生命科学以外の分野からはゲノム科学は終わったと捉えられている向きもあるようだが，実際はむしろ真逆である．これから最も大量のデータを生み出す分野として成長を続けており，今まさに情報爆発が始まるスタート地点に立っていると考えてよい．特に，物理学や天文学などの大規模データと比較したとき，生命科学における情報爆発は単に技術革新による高速化と量的な大規模化だけではなく，縮約の難しい質的に多様なデータを扱うという点に特長がある．多様かつ大規模なデータをいかに目的に応じて統合し，生命の本質に迫るか，という科学的にも技術的にもチャレンジングな課題が待ち受けているのである．

ゲノム配列と関連情報の大規模化

1990 年代に開始された国際ヒトゲノム解析プロジェクトにおいて，ヒトゲノム DNA の塩基配列に関する概略情報を取得するまでに 10 年（高精度の配列情報取得までは 13 年）と数千億円規模のコストがかかった．その結果として解読されたヒトのゲノム DNA の長さは約 30 億塩基である．DNA に含まれる塩基（base）はアデニン（A）・チミン（T）・グアニン（G）・シトシン（C）の 4 種類であり，通常は各 1 文字で表記されるため，情報量としては 3Gb（Gigabases < Giga bytes）にすぎない（ただし，ゲノム解読の過程では実験手法の制限により，30 億文字の塩基配列を 1,000 文字程度の断片に分断し，それぞれを DNA シーケンサ☆1 で読み取り，断片の重複を元にアセンブルを行う，というプロセスを経ているためこの何倍〜何十倍もの DNA を解読している）．

近年，DNA の解読技術は格段に高速化し，ヒトゲノム程度の情報であれば 6 週間のうちに 600 万円程度のコストで取得できるようになった．次世代 DNA シーケンサでは数十〜数百塩基の短い配列を 1 日に数百 Mb 〜数 Gb 解読することが可能で，この短い大量の配列を解読済みのヒトゲノム配列と対応させることによって個人のゲノムの差異を明らかにする「1,000 人ゲノムプロジェクト[1]」も 2008 年にスタートした．DNA シーケンサの技術開発と情報処理の手法は進化を続けており，近い将来では個人のヒトゲノムを数日中に千ドル（約 10 万円）あるいは 1 日以内に数万円で解読できるようになりそうである．ゲノム解読が進められているのはヒトに限った話ではない．病原菌から動物や植物まで実に多様な生物種のゲノム解読も継続して進展しており，現在約 1,000 生物種のゲノム解読が完了☆2，計画が進行している生物種の数は約 4,000 を超えている[2]．実に 3 年間で倍増する進展ぶりである．

しかし，問題は個々人の差異に応じたテーラーメイド医療を目指す何万人もの個人ゲノムや，地球上の生物多様性☆3 から生命の本質や変動とその起源を究明するために生物種単位で増えていくゲノム配列の数だけではない．高速な次世代シーケンサの登場により DNA シーケンサの運用コストが下がったため，ヒトやマウスの臓器ごと・発生段階ごと・病状ごと等における遺伝子発現データを経時的に解読するなど，ゲノム以外の用途にも幅広く使われるようになり，ゲノムの何十倍にもなる大量の実験データが日々生み出されるようになってきているのである．一説によると，2012 年までには CERN の生み出す年間 15 ペタバイトのデータ量を超える DNA 配列データが生成されるとの予測もされている[3]．さらに，ゲノム配列を 1 次データとし，それを元にした生命の機能を理解するための実験手法・解析方法，さらにその結果と解釈が次々と発表されており，このような状況において生み出される実験データと文献データは質・量ともに膨大なものとなっている．

得られた大量の DNA 配列を解析するソフトウェアについても，DNA 配列のアセンブルやマッピングから遺伝子発見・機能予測・リピート探索・多型解析など多岐に及び，関連する遺伝子発現データや分子間相互作用の

☆1 DNA 断片から塩基配列を読み取るための実験機器．

☆2 これとは別に，昨今話題になったインフルエンザを含む，2,200 種を超えるウィルスのゲノム配列も次々と解読され公共データベースに登録されている．

☆3 SOS レポートによると，現在までに 180 万生物種が記載されているが，これはまだ地球上の生物の一部分でしかなく，2007 年度に発見された新種だけで 18,516 生物種にのぼる．http://www.species.asu.edu/SOS/

実験データ，知識ベースとしての文献データパスウェイのデータなど，実に多様なデータを統合的に利用する必要がある．ここでも，実験機器の発達により，または，研究の進展に伴う領域細分化と研究者の増加により，日々生み出される情報は急激な勢いで増え続けている．このため，生命現象の理解という大きな目的のためには，一研究者による手作業ではとても把握しきれないほどのさまざまな関連情報を分野横断的に，さまざまな視点から整理して効率的に取得できることが望まれている．これらの関連情報を効率良く利用するためには，それぞれの研究室や研究機関で得られた結果を，標準化された手法と形式で流通させる必要があるが，現状ではそこに大きな課題が残されている．

国際開発会議 BioHackathon について

これらの課題を効果的に解決していくため，ライフサイエンス統合データベースセンターでは，BioHackathon と呼ばれる国際開発会議を主催し，国内外における生命科学系の情報処理技術の開発者およびその利用者が一堂に会して議論や求められるツールの開発を行った．BioHackathon は歴史をたどると，オープンバイオファウンデーション[4]とスポンサーの主催により，バイオインフォマティクスのオープンソースライブラリである BioPerl, BioJava, BioPython, BioRuby などの開発者を参集して 2002 年に開催されたのが始まりで，その後システムバイオロジーやフィロインフォマティクス（系統情報学）などの分野でも取り入れられてきた，ソフトウェア開発を主体とした国際会議である．通常は 1 週間程度の期間，目的に応じて開発者を招待し，合宿形式で共通の課題について議論し同時に集中的なソフトウェア開発を行う機会を提供することで，対象や時差によるギャップを超えた効率的な共同作業が可能となり，プロジェクトの垣根を取り払ってさまざまな新しい成果を生み出してきた．これまで国際会議といえば，学術分野では最新の研究成果をお互いに発表し情報交換する場であったが，インターネットの恩恵により情報の流通自体にはそれほど困らなくなっており，大規模な会議で顔を合わせても個別のテーマについて議論を深める時間は十分には取れないのが現状である．一方で BioHackathon のように，共通の課題を抱えている研究者や実務的な開発者を一堂に会し，その場で顔を突き合わせて集中的に解決を図る国際開発会議は非常に生産的な会議のスタイルであり，今後新しい国際会議のパラダイムとしてもっと普及してほしいと期待している．

さて，ライフサイエンス統合データベースセンターが主催した BioHackathon 2008（第 1 回目）は平成 20 年 2 月 11 ～ 15 日，BioHackathon 2009（第 2 回目）は平成 21 年 3 月 16 ～ 20 日の日程で開催され，それぞれ約 60 名（うち海外から 30 名程度）の参加者を集めた．BioHackathon 2008 は産業技術総合研究所 CBRC との共催で東京お台場にて開催され，配列データからタンパク質間相互作用や糖鎖情報学まで多岐に及ぶデータの標準交換方法と解析手順の Web サービスによるワークフロー化を中心にさまざまな課題について議論が行われた．これを受けて，BioHackathon 2009 は沖縄科学技術研究基盤整備機構との共催で沖縄にて開催され，多様なデータの統合的な利用環境の整備を目指し，マリンゲノミクスなど実験研究者の求めるデータベース統合とさまざまなクライアントソフトウェア間の連携について幅広い議論が行われた．その中で，大規模データの取り扱いについての技術開発だけでなく，データの統合的な利用に必要となるセマンティック Web の技術，さらにセマンティックな解釈の基盤となる文献データからのテキストマイニング，大量の情報を統合した結果を理解するためのビジュアライゼーションなどの技術を整備していくことが不可欠であることが明らかになってきた．

本特集のねらい

本特集では，2 回の BioHackathon で行われた標準化と統合化の議論を念頭に，「生命情報学が直面する大規模ゲノムデータの課題」というテーマで下記 4 つのトピックを取り上げ，バイオインフォマティクス分野における情報学的な問題点について俯瞰したい．第 1 稿「分散データの統合とセマンティック Web」では，世界中のサーバに分散したさまざまなフォーマットの生物学的データの統合的利用についての技術的な課題を，第 2 稿「巨大データの扱いと解析」では，その際に問題となる巨大データの取り扱いとグリッドコンピューティングについて，第 3 稿「生命科学分野におけるテキストマイニング」では，大量に得られたデータの生物学的な意味付けに必要となる科学論文を中心としたテキストマイニングの課題について，第 4 稿「生命科学分野・ゲノムデータの可視化技術」では，得られたさまざまな情報を研究者が把握し理解するためのビジュアライゼーションについて現在の取り組みを，それぞれ解説している．本特集が，すでに大規模データの情報処理技術を持つ異分野の研究者の参入につながれば望外の喜びである．

参考文献
1）http://1000genomes.org/
2）http://genomesonline.org/
3）The need for speed. Paul Flicek, Genome Biology 2009, 10:212
4）http://open-bio.org/

（平成 21 年 6 月 30 日受付）