3.2.3 ビッグデータの有効利用例②：

ゲノム解析・バイオインフォマティクス

(1) 課題概要

分子生物学では、細胞内の大量の遺伝子発現を同時計測可能な DNA マイクロアレイなどの登場により、実験で得られるデータの量が飛躍的に増加し、それら大量のデータ、いわゆるハイスループットデータから計算機を用いて生物学的な発見を行おうとする研究手法（＝バイオインフォマティクス）が行われるようになった。ヒトの染色体にあるすべての DNA 配列を解き明かしたヒトゲノム計画においても大量の計算機が投入され、「計算機を用いた分子生物学」が成果を挙げてきた。大規模なデータに基づく研究・解析はこの分野ではハイスループットデータが出始めた当初から重要な課題として考えられてきたことであるが、近年このような大規模データはビッグデータと呼ばれるようになり I/O 性能への要求が高いなどこれまでになかった特徴を備えていることから、HPC 分野においても計算機科学と連携が必要な重要な一分野として認識されるようになってきている。

ヒトゲノム計画ではたった一人の DNA 解読に長い年月を要したが、近年になって次世代 DNA シークエンサーと呼ばれる超高速かつ低コストで DNA を解読する装置が開発され、個人個人のゲノム解析あるいは組織毎など多種多様な細胞のゲノム解析を行うことが現実的になった。国民すべてが自分自身のゲノム配列を知り、ゲノムに基づく最適な医療を受けるいわゆる個人ゲノム医療の時代が目前に迫っている。次世代シークエンサーは、DNA を非常に短い断片にし、それを並列に読み込むことにより DNA を高速に読み取ることができる装置である。ゲノム解読は読み取った DNA 断片をつなぎ合わせる処理が必要なため、それに大量のストレージと I/O 性能の高い計算機を必要とする。具体的には現在の DNA シークエンサーで 1 サンプル（細胞種）当たりおよそ 1TB のストレージが必要である。病気とゲノムの関係性を明らかにするためにはある程度まとまった量の病態サンプルにおけるゲノムと正常組織でのゲノムとの比較が必要である。そのためには最低 1000 サンプル程度のゲノム解析が必要であり、それだけで 1PB のストレージが必要となる。2020 年には 200,000 人規模のサンプルの解析を行うと予想されており（後述）、また 1 サンプル当たりのデータ量も増えることが予想されているため、必要なストレージ量は膨大なものとなることが容易に予想される。また、病気の原因となる変異の同定などのためには、これらのストレージにあるデータに大量にアクセスすることが必要である。ゲノム解析に求められている計算機性能は、データ大量アクセスの必要性の点で、他の HPC 分野で必要とされているものと大きく異なる。個人や病態サンプルの DNA 配列データが蓄積されるとそれらのデータを用いた新しい解析が行われると考えられる。ゲノムと病気の関連を解析する、いわゆる全ゲノム関連解析（GWAS）も、より大規模化・複雑化し、それに必要な計算リソースもそれに応じて膨大になると予想される。

生命科学におけるデータ解析ではゲノム配列データだけではなく冒頭で触れた遺伝子発現データや DNA 修飾のデータ（エピゲノム）、タンパク結合など多種多様で膨大なデータを組み合わせて解析が行われる。今後開発される観測技術によってより多くのデータが蓄積される

ことが予想され、観測技術の進展にしたがって多様な解析ソフトウェアが組み合わされて利用される。このようにこの分野では常に新しい計算・解析手法が提案されているため一つのソフトウェアの寿命（利用期間）が短いのも特徴である。

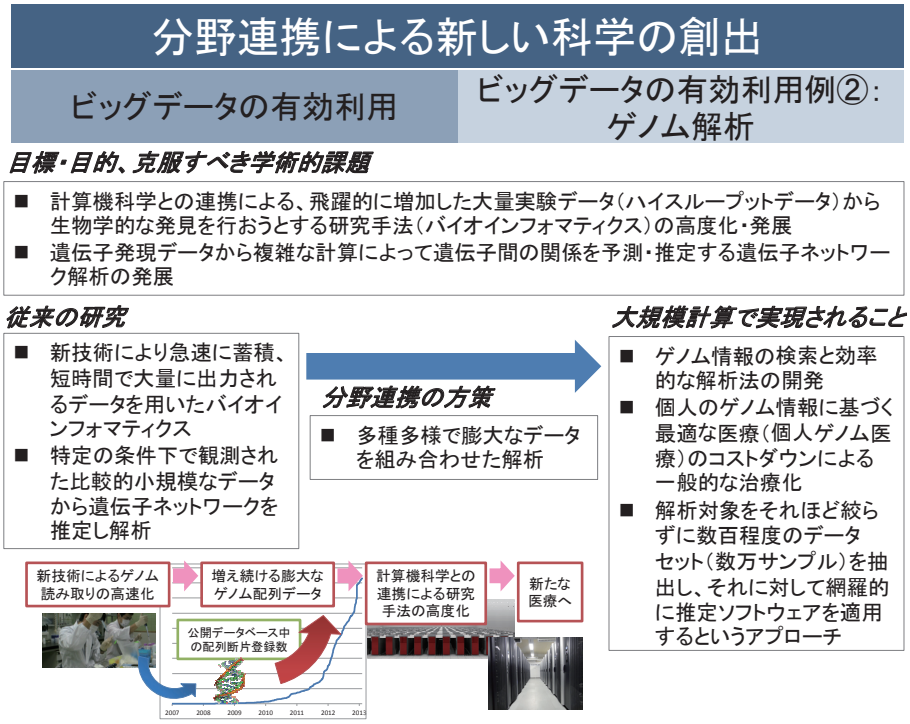

図 3.2.3-1　ビッグデータの有効利用例②：ゲノム解析

(2)　サイエンスの質的変化と長期的目標

生命科学において長期的にどのような観測技術がどの程度のコストで可能になるかは非常に流動的であり不明確な部分が多い。しかしながら次世代 DNA シークエンサーについては技術的な発展のロードマップが示されていることもあり比較的予測しやすい。また次世代 DNA シークエンサーが扱うデータ量もその他の従来のハイスループットデータと比較して膨大である。実際、NCBI（National Center for Biotechnology Information）のデータリポジトリの 98%以上がすでに次世代シークエンサーデータとなっている。したがって、ここでは次世代 DNA シークエンサーデータ解析をバイオインフォマティクス（生命科学データ解析分野）での代表的な課題として扱うこととする。次世代 DNA シークエンサーの普及により、それらのデータを生かした新たな解析需要も当然出てくると思われるが、ここではそれらを含めた値として概算値として議論をする。

現在、大規模なヒトゲノム解析プロジェクトとしてさまざまながんの細胞のゲノム配列を国際的に手分けして解読するプロジェクト、International Cancer Genome Consortium（ICGC）による国際がんゲノム計画が進んでおり、日本でも肝がん 500 人分の DNA 配列解読をこのプロジェクトの担当として行っている。このプロジェクトでは東京大学医科学研究所ヒトゲノム解析センターの持つ大型計算機を中心に利用し、2〜3 年間でヒトがん細胞 500 症例 1000 サンプル

分のゲノム配列を解読する計画が進行中である。その他にも、日本は参加していないがさまざまな人種の 1000 人分のゲノムを国際協力の下、解き明かそうとする 1000 人ゲノム計画というものもある。更に、これらの計画の終了などを受けて、現在は 100 万人ゲノム計画が米国・中国を中心に呼びかけられている状況である。したがって、日本においても今後 2015 年頃には 10,000 サンプル（人）規模の、2020 年頃までに 200,000 サンプル（人）規模のゲノム解析が行われると予測される。また 2015 年頃には更に世代の進んだ第 3 世代シークエンサー、2020 年頃には第 4 世代シークエンサーが登場すると言われており、1 個人について得られる情報量も飛躍的に増加していくと予想されている。そのため、ゲノム解読だけでも非常に多くの計算機リソースが必要になるのは自明であり、早急に研究体制を整備する必要があると言える。

以降では、次世代 DNA シークエンサーデータ解析と疾患関連遺伝子発見のための統計解析を具体例として述べる。

(i) 次世代 DNA シークエンサーデータ解析

次世代 DNA シークエンサーは、DNA 配列を 50 から数百塩基程度の長さに断片化し、それを同時に読み込むことによって超高速に DNA 配列を読み取る装置である。したがって、得られるデータは大量の断片化された DNA 配列であり、ヒトの場合およそ 1 億個程度、ファイルサイズにしてー100TB（2020 年頃に開発が期待されている第 4 世代シークエンサーを仮定）になる。連続した DNA 配列を得るためにはこれらの断片を、次世代 DNA シークエンサー解析ソフトウェアを用いて染色体ごとに 1 本につなげる必要がある。現在、主に行われている方法は、配列が既知である参照元となる DNA 配列（いわゆるリファレンスゲノム）のどの部分に最も一致するか各断片 1 個 1 個について探索し一致させる計算である。その際、個人差や変異などにより完全に一致することはないため、ミスマッチを許す文字列探索法を用いている。読み込んだ断片の各塩基には信頼度の数値情報がついているが、DNA シークエンサーの世代が進むにつれ付加情報が豊富になり計算が複雑化すると予想されている。計算処理に膨大な計算リソースが必要となるため、現在は読み取る DNA 配列をタンパクをコードしているエクソン領域に限定するなど、読み込む対象を絞ることで対処している。次世代には個人個人が複数の組織のすべてのゲノムをシークエンスすることが当たり前になると予想され、それに必要な計算リソースもまた膨大となる。ゲノム解析では個人の DNA 配列を得るだけではその個人が既知の変異を持っているかなどの診断は可能であるが未知の変異を見つけることはできない。したがってある程度まとまった量のデータが必要になり、そこから統計的解析を行う必要がある。特にがんゲノム解析では、がんはゲノムの異常であることから、がんの特異的な変異の同定だけでなく DNA の特定部位のコピー数解析なども非常に重要である。

(ii) 疾患関連遺伝子発見のための統計解析

疾患関連遺伝子とは、病気のなりやすさに影響を与えている遺伝子である。疾患関連遺伝子がわかれば、その遺伝子をターゲットにした医薬品や治療法の開発へとつなげることができる。疾患関連遺伝子発見のための統計的解析として最近広く使われている手法が、ゲノムワイド関連解析である。関連解析とは、患者群と、非患者群の間で、対立遺伝子の頻度の差を統計検定することにより、疾患関連遺伝子を探る手法である。関連解析の対象となる人数は、2013 年現

在では数千人～十万人ほどである。ゲノムワイド関連解析とは、関連解析を全ゲノムを対象に行う手法である。遺伝情報はすべての人が同じではなく、個人ごとに違っている部分がある。個人ごとの塩基配列の違いを遺伝子多型と呼び、1 塩基の違いを SNP という。ゲノム配列上、近隣にある SNP 同士は連鎖し、親から子へと一続きのまま遺伝する多いため、すべての SNP を見なくても、代表的な SNP を選べば、ヒト集団の多様性がある程度は明らかになる。このために選ばれる代表的な SNP を tagSNP という。2013 年時点では、ヒトの全ゲノムの中から 100 万個ほどの tagSNP を選び、それを対象に関連解析が行われていることが多い。今後は、次世代シークエンサーなどの技術革新により、対象人数が増えて行くと予想される。また、tagSNP ではなく、ヒトの全ゲノム配列を用いた解析が行われていくようになると考えられる。現時点でも千人規模の全ゲノム配列を決定するプロジェクトが多国間共同研究として進行中である。今後はこのようなプロジェクトが日本国内、そして世界中で行われることが予想される。そのため、今後のゲノムワイド関連解析は、数万人以上・全ゲノム規模の解析へと進んでいくと予想される。そのための高速計算が必須である。

(3) コミュニティからの意見

2013 年 3 月 10 日に大阪にて文部科学省科学研究費補助金新学術領域研究「システム的統合に基づくがんの先端的診断、治療、予防法の開発」プロジェクトの公開講演があり、そこで実際に現場でがん研究を行っている研究者と、将来がんゲノム解析やがん遺伝子ネットワーク研究で必要となるスーパーコンピュータなどについてパネルディスカッションを行った。

そこで出た現場からの主な意見は以下のとおりである。

- 現在のゲノム解析を支えているヒトゲノム解析センターの現在および将来の計算能力が将来予想されるデータ規模に対して危機的なこと
- 計算機の能力だけでなく、それを使いこなす人材が不足すること、が挙げられる。またスーパーコンピュータを用いた研究が確実にがん研究を変えており今後の発展に大いに期待していること

(4) 必要な計算機資源

(i) 次世代 DNA シークエンサーデータ解析

次世代シークエンサー解析に必要な計算スペックを以下に示す。DNA シークエンサー技術の今後についてはかなり流動的であるため一部あいまいな表現を残している。例えば想定しているのは解析対象となる人数であり、一人当たり何種の細胞のサンプルが取得できるかはコストや技術的な制約により可変なため、ある程度幅を持たせる意味で明示していない。

2015 年頃に、第 3 世代シークエンサーにより 10,000 人規模の、2020 年頃に第 4 世代シークエンサーにより 200,000 人分規模の解析が必要であると想定しているため、その 2 つの時期における必要スペックを記す。なお、ここに挙げた計算スペックの見積は計算資源の総和として必要な物であり、10 拠点ほどに分割されていても問題ない。また、複数台のシークエンサーからのデータを独立に処理可能なことから、1 システムで実現が必要なスペックは導入されるシークエンサーの台数分で分割可能である。効率的に解析を行うためには、ゲノム解析に用いら

れる計算機はDNAシークエンサーに併設されている必要がある。また、必要演算性能については、ゲノム解析では整数演算が中心であるため、FLOPSではなく、Operation数で記述している。

【ネットワークバンド幅、レイテンシ】

2015年：3〜6GB/s（シークエンサー＝ストレージ間の総和）

2020年：600 GB/s〜1.2 TB/s （シークエンサー＝ストレージ間の総和）

（100台のシークエンサーで分散させることを仮定すると〜12 GB/s/台）

データ並列性があるため、計算機の計算ノード間の通信バンド幅およびレイテンシはあまり重要ではない。その代わりに、シークエンサー＝ストレージ間のネットワーク帯域がボトルネックの一つとなり得る。第3世代および第4世代シークエンサー1人分のデータはそれぞれ1〜10 TBおよび10〜100 TBであり、想定している10,000人および200,000人分のデータを1年間（31,536,000秒）でシークエンサーから転送することを考えると最大で1〜10TB×10,000=10〜100PBまたは10〜100TB×200,000＝2〜20 EBの転送が必要とされ、必要な帯域は実効でそれぞれ3GB/sおよび600GB/s必要である。実際に倍程度必要と考えると第3世代機時代には3〜6GB/s、第4世代機時代には600 GB/s〜1.2 TB/sの帯域が必要であると言える。これは総計であるので、10拠点に分散されれば必要な帯域はこの1/10となる。また100台のシークエンサーに分散されれば、1台当たりは1/100である12GB/sほどになる。実際にはもう一桁ー二桁ほど多い台数のシークエンサーが導入される可能性もある。

【1ケースあたりの総メモリ容量】2015年：0.016PB〜0.16PB、2020年：0.16〜1.6PB

現行既存アプリケーションのプロファイル結果より入力ゲノムサイズをx [MB]、分割数をnとしたとき，消費メモリ量y [GB] は $y = (0.015 x / n) + (1.4 * n)$と表すことができる。想定している2010年頃は1ケース（サンプル）あたり1ー10TB，2020年頃は10ー100TBであることからそれぞれ0.016PBー0.16PB、および0.16PBー1.6PBという見積もりになる。実際には入力ファイルを分割可能であるため分割実行が可能である。その場合、同時に必要なメモリ量は分割数で単純に割った値となる。

【ストレージ容量】2015年：10〜100PB、2020年：2〜20EB

単純に1人当たりの必要データ量に想定される人数を掛け合わした値である。例えば、第4世代機による1サンプル100TBの場合、200,000人では20 EB必要となる。

【ストレージ速度】2015年：〜0.56 TB/s、2020年：ー5.56TB/s（分割処理可能）

現行システムにおける1ファイル80MBの入力データに対するプロファイル結果はIO量が11.204GBである。これを1サンプル100TBに外挿するとIO量は14,005,000GBとなる。想定している2020年での200,000人（サンプル）規模の解析を1年間で行うと仮定すると1サンプルの実行時間は2520秒となる。以上より，必要なIO性能は5.56TB/sとなる。2015年頃のデータ量はこの1/10であるので単純にIO性能も1/10である。またこれは入力データが分割可能

なことから同一システムで実現されなくても総計として必要な IO 性能であって，システム（解析拠点）の数で分割可能である。

【CPU 速度】2015 年：～0.54 TOPS、2020 年：～5.4TOPS

現行システムにおける 1 ファイル 80MB の入力データに対するプロファイル結果では総（実行命令数（整数演算量）は 10.890GOP (GI)であった。ここから外挿すると 1 ケース（サンプル）あたり 13,612,500GOP の演算量が必要であり，必要な整数演算性能は 5.4TOPS である。

(ii)　疾患関連遺伝子発見のための統計解析

疾患関連遺伝子発見のための統計的解析ソフトウェアに必要な計算スペックを以下に示す。なお、シークエンサー技術の今後についてはかなり流動的であるため、それに応じ、疾患関連遺伝子発見のための統計的解析に使えるデータ量は、今後予想外に増大することもあり得る。そのため、一部あいまいな表現を残している。

2015 年頃に第 3 世代シークエンサーにより 10,000 人規模の、2020 年頃に第 4 世代シークエンサーにより 200,000 人分規模の解析が必要であると想定しているため、その 2 つの時期における必要スペックを記す。

【ノード当たりのメモリ容量】2015 年：40GB、2020 年：800GB

関連解析は、連鎖している領域をまとめて 1 ノードで解析したほうがよい。連鎖している領域の数は、ゲノム全体でおよそ 10 万である。しかしながら、連鎖している領域の大きさはさまざまである。そのため最大 1 千万塩基程度の長さのゲノム領域をセットで解析することが想定される。この領域に対して、部位一つ当たり 1 人 2 塩基を持つとすると、1 万人を対象に関連解析する際も、年齢・性別・サンプルされた地域ごとに、患者群千人、非患者群千人ほどに分けて並列的に解析することが考えられる。すると、1 千万×2 塩基×2 千人で、40GB ほどあるのが望ましい。人数が 20 倍になると単純に 20 倍となる。

【ストレージ容量】2015 年：100TB、2020 年：2PB

ゲノム上の部位一つ当たり 1 人 2 塩基を持ち、一人一人のデータの区切りのために 1 バイトを使うとすると、一人当たりのデータ量が 9GB となり、これを人数で掛けると、2015 年には 10,000 人でおよそ 100TB、2020 年には 200,000 人で 2PB 必要となる。

【ストレージ速度】2015 年：25GB/s～250GB/s、2020 年：500GB/s～5TB/s

ヒト一人のゲノムサイズは、父親由来と母親由来がそれぞれ 3GB、計 6GB であり、データにする際には一人一人の間に区切り文字が入るので一人当たりは 9GB となる。想定している 10,000 人および 200,000 人分のデータをストレージからメモリにロードする時間が、実際に計算機を占有できる時間を圧迫してはならないと考えると、許されるのは最大でも 1 時間ほどであろう。とすると、2015 年には 25GB/s、2020 年には 500GB/s が最低でも必要となる。仮に、ロード時間を 6 分で計算すると、2015 年には 250GB/s、2020 年には 5TB/s が必要となる。ただし、これには分散 I/O での対応も可能である。

【CPU 速度】2015 年：～50 PFLOPS、2020 年：～1EFLOPS

現在、10PFLOPS の能力を持つ京速コンピュータ「京」の上で、tagSNP 数十万に対して、患者群 1500 人、非患者群 2000 人規模の解析を行っている。全計算能力の 10 分の 1 を利用しても、解析に数時間かかる。今後、tagSNP から全ゲノムに解析対象を広げることで一人当たりのデータ量が 100 倍になり、さらに 2015 年には人数も 3 倍ほどになることから、計算時間は 300 倍になる。現在の「京」の 5 倍のコンピュータで、使用割合を 10 倍に増やし、占有時間を 6 倍に増やすことで何とか対応できる。2020 年にはさらにこの 20 倍必要となる。

課題	要求性能 (PFLOPS)	要求メモリバンド幅 (PB/s)	メモリ量/ケース (PB)	ストレージ量/ケース (PB)	計算時間/ケース (hour)	ケース数	総演算量 (EFLOP)	概要と計算手法	問題規模	備考
個人ゲノム解析	0.0054	0.0016	1.6	0.1	0.7	200000	2700	シーケンスマッチング	がんゲノム解析200,000人分のマッピングおよび変異同定	1人分の解析を1ケースとした。入力データを分割することで、細かい単位での実行、拠点をまたいだ実行も可能。整数演算中心のため「総演算量」はInstruction数とした。総浮動小数点演算量は45.864EFLOPとなる。
疾患遺伝子発見のための統計的解析	10	0	200	2	140	5	25000	ゲノムワイド連鎖解析(GWAS)	ヒトゲノム3Gbp x 200,000人分. 1ケース4万人	メモリ量は800GB/node. ノード数25万を仮定