東京大学 医科学研究所

the Institute of Medical Science
the University of Tokyo
Human Genome Center

お客様情報

東京大学 医科学研究所

●所在地
〒108-8639
東京都港区白金台4-6-1
http://www.ims.u-tokyo.ac.jp/

1892年に設立された伝染病研究所を前身とし、附属の研究病院を持つ国内随一の医学・生命科学のための研究所。感染症、がんなどの疾患を対象に、基礎研究の成果を医療に直結させることを使命としています。

ヒトゲノム解析用スーパーコンピューターShirokane3に100ペタバイトまで拡張可能な省電力テープ・アーカイブを実装

東京大学 医科学研究所(以下、東大医科研)は、附属病院を有する国内最大規模の生命科学研究所であり、1892年に北里柴三郎氏が創立した私立衛生会附属伝染病研究所を前身とします。現在は、分子生物学や細胞生物学、発生工学の著しい進展とゲノム科学の急速な展開などを背景に、がんや感染症、その他の難治疾患を対象とした、最先端の研究と医療に取り組んでおり、基礎研究の成果を医療に直結させることを目指しています。特に生命・医科学の基礎研究、ゲノム医療、細胞・遺伝子治療などの先端医療開発を推進しています。
東大医科研は、先進的ながんゲノムの研究に取り組んでいる国際がんゲノムコンソーシアム(ICGC)の日本参加メンバーとして国際的な貢献を積極的に行っています。また、ICGCに加え、アメリカ国立衛生研究所(NIH)の一機関であるアメリカ国立がん研究所(NCI)とアメリカ国立ヒトゲノム研究所(NHGRI)が資金提供するがんゲノムプロジェクトThe Cancer Genome Atlas (TCGA)のデータベースも活用しながら、ゲノム研究を推進しています。
ヒトゲノム解析センターは、DNA塩基配列を読み取ることができるシークエンサーから得られる膨大な情報を、スーパーコンピューター「Shirokane」で解析し、がんをはじめとする難治性疾患の先端研究を実施しています。全ゲノムに対するシークエンス処理(ゲノムを構成するDNAの塩基配列決定)がより安価で高速化されつつある中で、急激に増え続けるデータへの高速アクセスを実現しつつも、消費電力を抑えコスト効率良く保管するため、最新の「Shirokane3」では新たに100ペタバイト(PB)まで対応可能な大規模テープ・アーカイブ・システムを実装しました。

100万人分のゲノム解析データを保存できる電力消費量の少ないストレージシステムが必要

ヒトゲノム解析センターで最も注力しているのは、がん細胞のゲノム解析、および臨床シークエンス(クリニカル・シークエンス)です。がん細胞のゲノム解析は、がん細胞において遺伝子の機能異常を引き起こしているDNA変異を調べる研究です。がん細胞のゲノムを1回シークエンスすると、標準にしているリファレンスゲノムとの相違点が数百万カ所も出てきます。そこから、がんに関連した変異であろうと思われるものを絞り込んでいくと、がん種にもよりますが、1万以上の変異を見つけ出すことができます。この変異が、なぜ生命システムに異常を引き起こすのかを、研究しています。一方、シークエンスを解析した結果を患者にフィードバックするのが臨床シークエンスです。例えば、IBM Watson Genomic Analytics などで、がんゲノムの解析結果を解釈し、有効な治療方法や薬の候補を提供することで、最終的に医師が判断して患者にフィードバックするためにゲノムシークエンスの解析を

【事例概要】

・大規模テープ・アーカイブ・システム

【ハードウェア】

・IBM TS1150 テープ・ドライブ
・IBM TS3500 テープ・ライブラリー

【ソフトウェア】

・IBM Spectrum Protect
・IBM Spectrum Scale (GPFSテクノロジー)

【ビジネスメリット】

・100PBまで拡張できる高いスケーラビリティー
・多量データの保管と省電力性の両立
・100万人のゲノム解析データの高速な読み書き
・テープを意識せずに利用できる高い利便性

研究しています。

医学研究のために全ゲノムシークエンスを実施すると、膨大なデータが発生します。ヒトゲノム解析センターでは、非常に高いスループットの次世代シークエンサーを利用していますが、これを1回走らせると10人分の全ゲノムを読み取ることができ、全部で1テラベース(Tb)のデータが出力されます。このデータがもとになり、さまざまな解析結果が付加されることで、データ容量は更に膨らんでいきます。これまでヒトゲノム解析センターでは、2009年に導入された「Shirokane1」、2012年に導入された「Shirokane2」で、ヒトゲノムの解析を行ってきました。Shirokane1はホームディスクが1PB、ニアラインディスクが600TBのストレージ構成で、Shirokane2はホームディスクが2PB、ニアラインディスクが500TBの階層化ストレージ構成でした。

ヒトゲノム解析センター長で理学博士でもある宮野悟氏は、「Shirokane1、Shirokane2のストレージは、その使用量が常に100%近くで推移しており、ゲノムデータの解析に支障が出始めていました。Shirokane1、Shirokane2を計画した時点では、運用期間の途中でいっぱいになってしまうほど解析データの容量が増大するとは予想できませんでした」と話します。

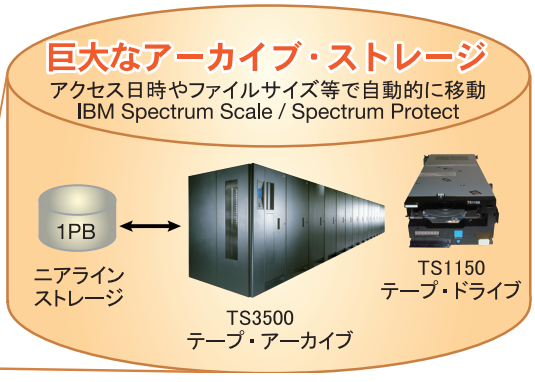
ストレージ使用量増大の背景として、2011年ごろからシークエンサーの精度が著しく向上しコストが減少したことからデータ量が急速に増えたことが挙げられます。宮野氏は、「患者さん1人のゲノムの解析に必要なストレージ容量は、最低でも200ギガバイト(GB)程度で、高精度になると500GB以上が必要になります。このゲノム解析データを100万人分保存して、高速に読み書きをすることができ、さらに消費電力が少ないストレージシステムを構築することが必要でした」と話しています。Shirokane2 の次のスーパーコンピューターでは、ストレージシステムの大容量化、高速化に伴い、消費電力量の増大などが解決すべき課題の1つであり、膨大なデータ量に対応しつつ消費電力量が、全体で1メガワットを超えないことが必要でした。こうした大規模データを効率的に保管するソリューションを検討するにあたり、ゲノムを研究している国内外の研究所における大規模ストレージの利用状況も確認し、最終的にShirokane3の入札仕様を策定しています。

Shirokane3にIBMテープ・アーカイブを採用 100万人分のゲノム解析データの保存を実現

ヒトゲノム解析センターでは2015年4月1日より、Shirokane1、Shirokane2の後継となるスーパーコンピューターShirokane3の運用を開始しています。バイオメ

スーパーコンピューターシステムの比較と Shirokane3 の巨大アーカイブ・ストレージの概要

システム名称	Shirokane1	Shirokane2	Shirokane3
運用開始	2009 年 1 月	2012 年 1 月	2015 年 4 月
処理能力	75 TFLOPS	150 TFLOPS	422 TFLOPS
ホームディスク	1 PB	2 PB	12 PB
ニアラインディスク	600 TB	500 TB	1 PB
テープ	n/a	n/a	21 PB (最大 100PB)



ディカル・ビッグデータに対する多様な要求に対応したShirokane3は、ヒトゲノム解析を高速に分散処理するため、サーバーは計1万1,160個のCPUコアで構成される分散メモリ型サーバーと、ヒトゲノムアセンブリのような大容量メモリを必要とする解析用に1ノードあたり2TBのメモリを搭載した大規模メモリサーバーで構成。従来比で約10倍となる422テラフロップス(TFLOPS)の総合論理演算性能を実現しています。

ストレージは、ゲノム研究における膨大なデータアクセス要求に対応可能な12PBの高速ディスクアレイと22PBの階層型アーカイブで構成しました。階層型アーカイブストレージは、1PBの大容量分散ストレージと21PBのテープ・ライブラリーを単一のファイルシステムとして透過的なデータ移動ができるよう、IBMのSoftware Defined Storage製品であるSpectrum Storage(Spectrum ScaleおよびSpectrum Protect)を採用。今後のデータ増加に対応できるよう、テープ・アーカイブの容量はテープ・カートリッジを追加するだけで最大100PBまで柔軟に拡張でき、従来比で約33倍にあたる100万人分のゲノム解析データの保存が可能になりました。

宮野氏は、「ゲノム解析や臨床シークエンスでは、膨大なデータを保管できるストレージを確保しておく必要があります。すべてをハードディスクで構成することも技術的に可能ですが、問題は消費電力でした。白金台キャンパスおよびShirokane3の契約電力量は決められています。そのためハードディスクだけで100PBのストレージを構成するのは現実的ではありません」と話しています。こうして、Spectrum Storageと容量・スピードともに優れたIBMエンタープライズ・テープ(TS1150ドライブ搭載のTS3500テープ・ライブラリー)による大規模テープ・アーカイブ・ストレージの採用を決定しました。さらに限られた電力を最大限に活用することを目的に、国内の大学や研究機関のスーパーコンピューターでは初の採用となる間接蒸発式冷却装置を採用。東大サステイナブルキャンパスプロジェクト(TSCP)の最優先課題である、温室効果ガス排出削減による低炭素キャンパスづくりへの貢献を目指しています。

省電力、高性能、直感的な操作性で研究を加速
世界各国の研究所で採用された実績も高く評価

大規模なテープ・アーカイブ・システムを導入した効果について、「大量データの高速な読み書きと省電力化に高く貢献しています。一般的にテープ・ドライブは、大量データの読み書きの速度が遅いというイメージがありますが、今回の大規模テープ・アーカイブは、特に書き込みが高速化されています」(宮野氏)。

高速分散ファイルシステムSpectrum Scaleは膨大な解析データの高速な読み書きを実現し、Spectrum Protectは通常のストレージと同様にファイル単位でのテープ・アクセスを可能にします。その結果、1PBの大容量分散ストレージ領域と、21PBの大容量テープ領域の間を、アクセス頻度やファイルサイズなどの条件で、データが自動的に移動する仕組みになっています。この階層化ストレージ構成では、利用者は大容量テープ領域をまったく意識することなく、1つの巨大なストレージシステムを利用しているかのように解析データを読み書きすることができます。

また、高速ディスクアレイでゲノムデータを高速に解析し、解析した結果を大規模アーカイブストレージに保存しておき、必要なときに大規模アーカイブストレージから解析結果を高速ディスクアレイに移動させて解析処理を継続することもできます。

宮野氏は、「パブリックに公開されている研究データで、ダウンロードに半年程度かかるものもあります。こうした大容量のデータは、高速ディスクアレイに保存するのではなく、大規模アーカイブストレージに保存することで効率的な管理が可能になります」と話します。

東京大学 医科学研究所
ヒトゲノム解析センター
教授 理学博士
宮野 悟 氏

“大量データの高速な読み書きと省電力性に高く貢献しています。一般的にテープ・ドライブは、大量データの読み書きの速度が遅いというイメージがありますが、今回の大規模テープ・アーカイブは、特に書き込みが高速化されています”

アーカイブ・ストレージ

IBM Watsonでがんの原因を分析
手回し計算機の世界から脱却

今後のIBMへの期待については、「IBM Watsonのような次世代のテクノロジーを気軽に使える世界が来ることです」と宮野氏は話します。2015年4月から運用を開始した「Shirokane3」は、100万人分のゲノム解析データ保管にも対応する大規模テープ・アーカイブを備えており、今後、クラウド基盤で稼働する「IBM Watson Genomic Analytics」と連携し、研究を進めていくためのビッグデータ解析基盤となります。がん関連だけでも年間20万本近い論文が発表されるなか、ゲノム情報を解析し、それに関する文献を調べ、数百の論文を読んでも、なぜその患者ががんになったかは多くの場合簡単には分かりません。研究者や医師1人ひとりの知識や経験で判断するのは困難な時代になっており、ゲノム情報に基づくがん研究は人知を超えた世界に突入しています。「将来的に期待されるのは、IBM Watsonに変異情報を入力すると、変異データベースや文献データベースなど、あらゆる情報から今後どのような病気になるか確率を表示してくれることです。また、蓄積された日々のヘルスデータから、今後病気になる可能性があるのかを確率とともに予測してくれることです。IBM Watsonのようなテクノロジーによって、まるで何十年も昔の手回し計算機の世界からようやく脱却できるのです」(宮野氏)。

ハーバード大学とマサチューセッツ工科大学が共同で運営するブロードインスティテュートが1年間でシークエンスするデータ容量は、2014年の時点で300ペタベース(Pb)ほどにもなるといわれていましたが、血液検査のような感覚で、一人のゲノムデータを時系列で収集し、分析をしていく世の中が来るとさらに増えることが予想されます。「近い将来、エクサバイト、ゼッタバイトの世界に間違いなく突入し、容易にデータシェアリングができる環境も重要になります。引き続きバイオメディカル・ビッグデータ時代に対応できるITシステムを研究し、開発していかなければならないと考えています」と宮野氏は語り、人々の健康医療の向上を目的とした更なるゲノム研究の発展を見据えています。

日本アイ・ビー・エム株式会社
〒103-8510　東京都中央区日本橋箱崎町19番21号

10-15 Printed in Japan