ポスト「京」重点課題 2

個別化・予防医療を支援する統合計算生命科学

NEWS LETTER

Vol. 6

Contents

INTEGRATED

COMPUTATIONAL

LIFE

SCIENCE

TO SUPPORT

PERSONALIZED AND

PREVENTIVE MEDICINE

ポスト「京」
重点課題2
個別化・予防医療を支援
する統合計算生命科学

■ Research Report

Subtheme A

がんの遺伝子解析とスーパーコンピューティング

サブ課題A
東京大学医科学研究所　伊東 聰
東京大学医科学研究所　矢留 雅亮
（左から）

DNAシーケンスの高速化と低価格化は、医学・生物学にとどまらず、医療の在り方を変えようとしています。サブ課題Aでは、次世代シーケンサーが生み出す膨大なDNA情報を高速に解析するため、ポスト「京」の性能をフル活用できるソフトウェアや解析環境を開発しています。これにより、がんを始めとした遺伝子疾患に対する的確な医療の実現が強く期待されています。

遺伝子とDNAと次世代シーケンサー

メンデルが遺伝の法則を発表したのは1865年。その約90年後の1953年、ワトソンとクリックがDNAの2重らせん構造を発表しました。それから約65年、DNAと遺伝子の研究は急速に発展してきました。2017年現在では、個人の全ゲノム（DNAに含まれる全遺伝子情報）の読み取りが数日、10万円程度で可能となっており、これらの技術をベースに遺伝子診断サービスを行う会社も出てきています。米国が1990年に開始したヒトゲノム計画が、たった一人の全ゲノム解読に13年と30億ドルかかったことと比較すれば、どれほど急速な発展をしてきたかが容易に想像できるでしょう。

この発展に最も大きな貢献をしているのが次世代シーケンサーです。DNAは数種類の塩基が連結した鎖のような構造（塩基配列という）をしています。この塩基配列を調べる機械をシーケンサーと呼びます。次世代シーケンサーの登場は2000年代中頃であり、従来型のシーケンサーが使用していた電気泳動の代わりにガラス基板とレーザー蛍光を用いて塩基読み取りを行います。この方法は一度に大量の塩基配列読み取りが可能であり、シーケンスの時間短縮とコスト低下を急速に推し進めることになりました（図1）。

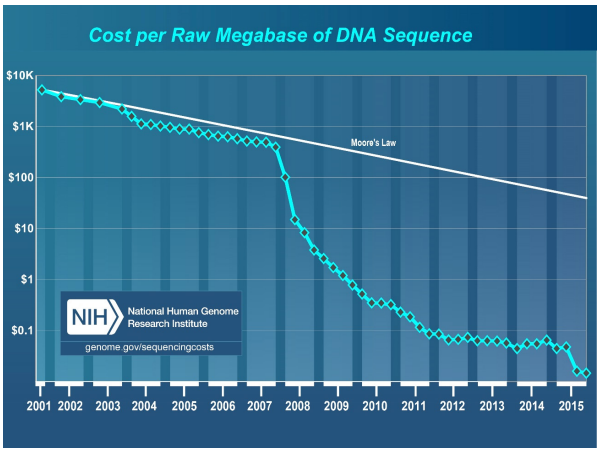

図1： DNAシーケンスコストの推移[1]
１００万塩基対（Megabase）当たりのシーケンサーのDNA読み取りコストの推移。2007年から急激にコストが下がっていることが分かる。

シーケンス解析とスーパーコンピュータ

次世代シーケンサーによって全ゲノムを低価格で高速に読み取れるようになりましたが、実はそのままでは研究に使えるデータにはなりません。現在最も利用されている次世代シーケンサーは、DNAの長い塩基配列を百数十文字程度の断片に分解し、それぞれの塩基の並びを読み取っています。ですから、読み取った断片（リード）が全ゲノムのどの場所の物であるかを探し、元の配列に再構成する必要があるのです。

ここで、全ゲノムシーケンスした際のデータサイズがどのくらいあるのかを考えてみましょう。人のDNAは約30億塩基対あると言われています。全ゲノムを対象にしたがんの研究では一般的に、1サンプル当たりDNA30～40コピー分（900億～1200億文字のデータ）の塩基配列情報を読み取ります（次節参照）。次世代シーケンサーで読み取るリード当たりの塩基配列長を100塩基と仮定すると、リードの本数は10億本前後になります。この膨大なリードそれぞれの位置

■ Research Report

Subtheme A

を探すことは、例えるならば 10 億ピースのジグソーパズルを解くようなものです。人力はもちろんのこと、パソコンでもとても計算しきれません。そのため、スーパーコンピュータを使って高速に処理することが必要なのです（図 2）。

がん研究と遺伝子変異解析

厚生労働省によると、2015 年の日本人の死因第 1 位は悪性新生物（がん）となっています。国民病ともいえるがんの克服は、健康長寿社会を目指すうえで避けては通れない重要な課題なのです。

がんは DNA の変異が引き起こす病気です。変異というと、とても珍しく深刻な現象に聞こえるかもしれません。しかし、人体の中ではそれほど珍しい現象ではありません。DNA にはたんぱく質をコードしている部分やその産生の制御などに関連する部分が数万あり、「生命の設計図」ともよばれています。細胞中の DNA は飲酒や喫煙などの生活習慣や加齢、紫外線やアスベストなどの外的要因、ウィルス感染など様々な要因によって傷つきます。小さな傷であれば自己修復機能によって回復します。修復できない大きな傷や修復に失敗した場合、その細胞は死んでしまうようにプログラムされており、新しく作られた細胞に置き換えられます。このようにして、通常は DNA に異常が起きても正常な状態に戻ります。ところが、ごく稀に修復されなかった細胞が生き残ることがあります。修復も細胞死も免れて生き残った細胞（変異細胞）の DNA に変異が蓄積していくのです。

最も単純な変異である点変異（DNA 中のある 1 塩基が別の塩基に置き換わる変異）を例に考えてみましょう。がんの腫瘍から採取したサンプルにはたくさんの細胞があります。そのすべての細胞に同じ変異が入っているわけではなく、また、腫瘍の中にも正常細胞が少量残っています。シーケンスの際にはそれらが入り混じった DNA が読み取られます。このような条件のなかで、変異の入っているリードの割合を腫瘍と正常組織の間で比較し、統計的に有意な差があるかどうかで判別したり、複数のサンプルを同じように統計的に比較することで変異を見つけ出しています。

我々のグループが開発している Genomon[2] は、このような統計解析などを駆使してがん化に強い関連性のある遺伝子変異を高精度・高感度に検出するソフトウェアです。これまでにヒトゲノム解析センターの医学・生物学向けスーパーコンピュータ「Shirokane」[3] を用いて多くの成果 [4,5,6] を挙げています。

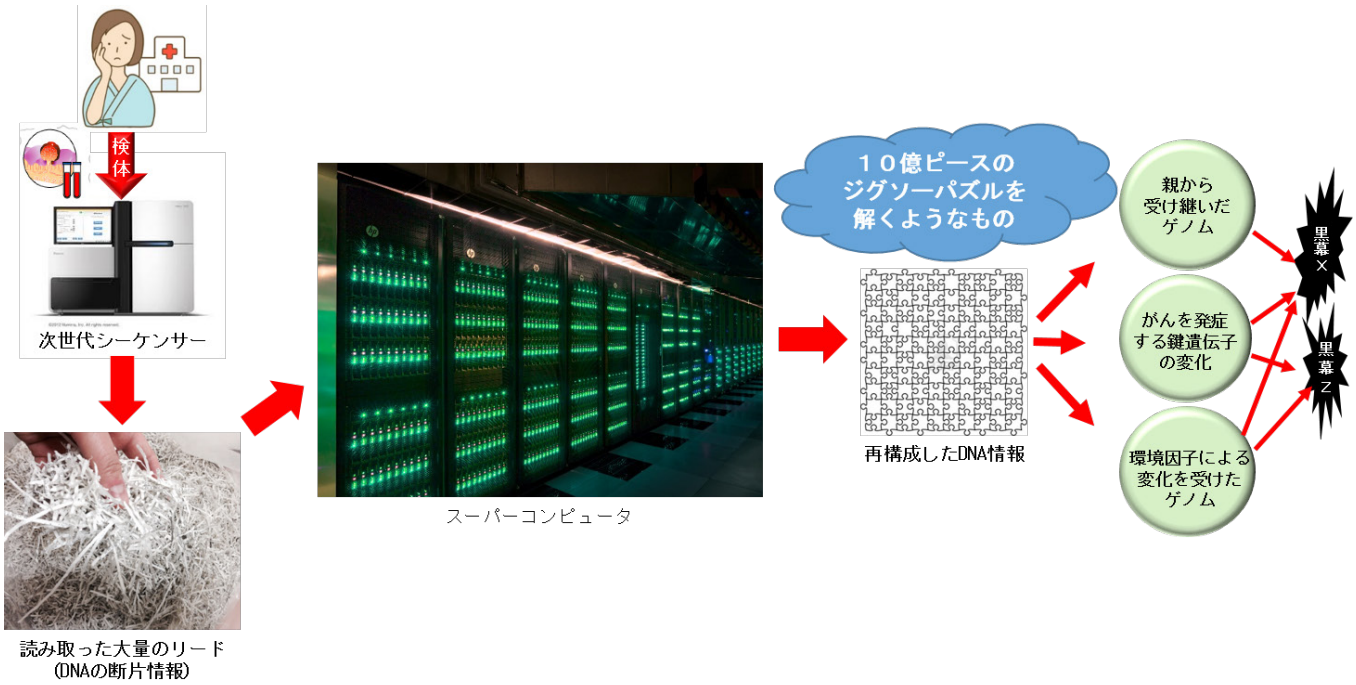

図2： シーケンス解析の流れ
採取したサンプルを次世代シーケンサーにかけると、大量のDNA断片情報（リード）が得られる。シュレッダーにかけた紙片のようなイメージだ。紙片の量は、リード長が100で30本分のDNA情報と仮定すると、30億/100*30=9億リードのデータになる。

■ Research Report

Subtheme A

個別化医療実現に向けた高速解析環境の開発

2015 年1月、オバマ米国大統領（当時）がプレシジョン・メディシン・イニシアチブを議会において発表しました。この流れを受け、個別化医療の取り組みが全世界で急速に進展しています。個別化医療とはその名の通り、各個人に応じた医療を意味します。これまでのがん医療は、患者多数に対して、多くの点で有効性が認められた治療法（抗がん剤など）を施すものでした。しかし、このような治療法は患者個人の視点では有効でなかったり、副作用が強すぎたりすることがあります。

がんの個別化医療では、患者個人のがん細胞の DNA のシーケンスを調べ、有効な抗がん剤や分子標的薬をより的確に選択できるようになるのです。ポスト「京」コンピュータが完成する2020年頃には、シーケンサーの性能が大きく向上し、シーケンス解析による的確な医療を求めるたくさんの患者さんの需要を満たせるビッグデータ解析能力が必要になると予想しています。

ポスト「京」はこのビッグデータ解析に十分な性能を持ちますが、解析ソフトウェアはそのままでは動かせません。スーパーコンピュータで運用するソフトウェアの開発は高度な専門性を要求されるため、医学 / 生物学の研究者が開発・運用することはとても難しいのが現実です。彼らが利用しやすいスーパーコンピュータ環境には専用の支援ソフトウェア（グリッドエンジン：GE）が搭載されています。この GE はシーケンスデータ解析における世界標準環境ではありますが、物理や工学といったこれまでのスーパーコンピュータユーザのソフトウェア利用効率を下げてしまうため、ポスト「京」には搭載しないことになっています。

そこで我々は、ポスト「京」上で GE の機能を提供する、医学 / 生物学者用支援ソフトウェア「Virtual Grid Engine」(VGE)[7] を開発しています。

VGE は、スーパーコンピュータのシステムからは物理や工学などのシミュレーションソフトウェアと同様の並列ソフトウェアに見えますが、ユーザからは GE の機能を提供する外部プログラムに見えます。このため、他ユーザのソフトウェア利用効率を妨げるようなことは無く、Genomon を始めとする医学・生物学用解析ソフトウェアの多くを GE の搭載されたスパコンと同じように利用することが出来るようになります（図 4）。VGE のようなソフトウェアで重要な点は、VGE の動作自体に掛かる処理時間（オーバーヘッド）をどれだけ小さく出来るのかという点です。VGE で採用しているアルゴリ

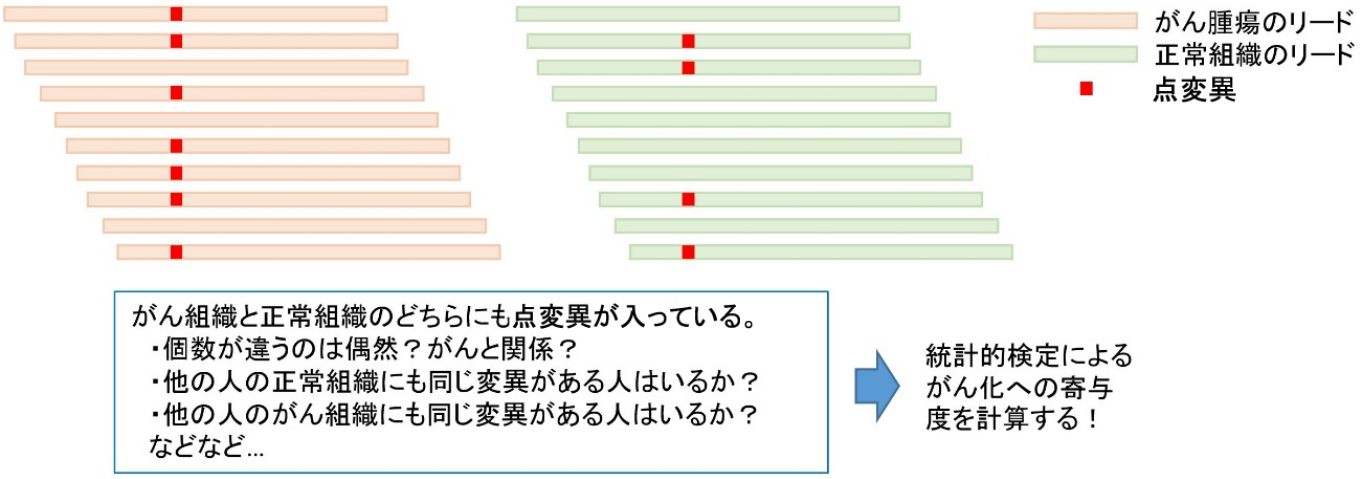

図3： シーケンス結果における点変異の見え方の例

■ Research Report

Subtheme A

ズムは一般にマスターワーカーと呼ばれています。このアルゴリズムでは基本的に計算資源（CPU 数や計算ノード数）とジョブ数（ソフトウェア実行完了までの計算単位総数。図 4 参照）を管理します。この管理コストはオーバーヘッドの一部ですが、計算資源量やジョブ数が大きくなるとコストが増大することが知られています。VGE ではポスト「京」の大規模資源とそれを用いた解析を念頭に、オーバーヘッドを出来るだけ小さくするよう設計に工夫をしています。「京」を用いた予備性能評価では、数千ノード・数十万ジョブのテスト計算においても、十分に小さなオーバーヘッドになることを確認しています。

ポスト「京」のもたらす大規模計算能力とそれを生かすことのできる解析ソフトウェアの開発により、大量のシーケンスデータを一度に解析できる環境が実現しつつあります。将来的には、病院でのがん遺伝子検査が血液検査のようにすべての患者さんに安価かつ高速に実施できるようになると期待しています。

＜参考文献＞
[1] https://www.genome.gov/sequencingcostsdata/
[2] http://genomon.readthedocs.io/ja/latest/
[3] https://supcom.hgc.jp/japanese/
[4] *Nature* 478, 64–69 (06 October 2011), doi:10.1038/nature10496
[5] *Nature* 534, 402–406 (16 June 2016), doi:10.1038/nature18294
[6] https://github.com/SatoshiITO/VGE

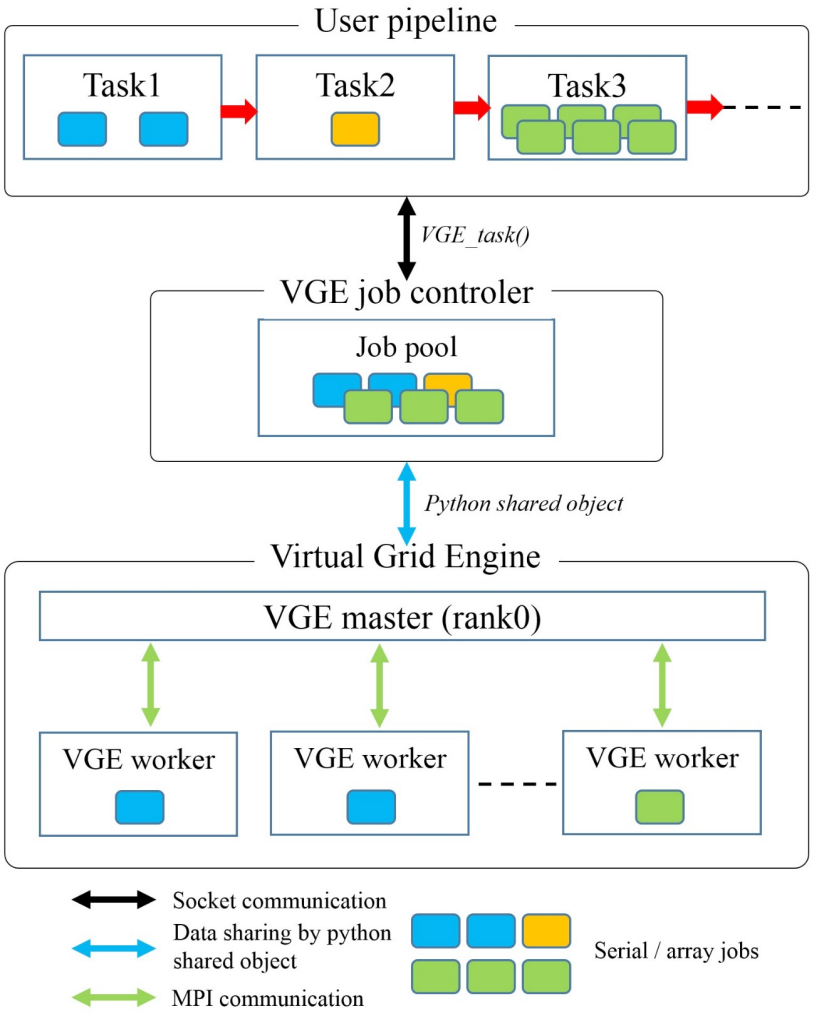

図4： Virtual Grid Engineシステム
ユーザの利用したいソフトウェア（user pipeline）は複数のタスクで構成されている。各タスクは逐次ジョブ（黄色）の物もあれば並列ジョブ（青や緑）まで、規模も計算内容もバラバラである。VGEはスーパーコンピュータ上にMPIプログラムとして計算機リソースを確保し、ソフトウェアからのタスクを順番に受け付け、確保した計算機リソースに各タスク内のジョブを自動的に割当、実行させる。

■ Information

News & Events

出展報告：2017 年度理化学研究所計算科学研究機構 一般公開

〇ブース出展

日程：10 月 14 日 (土) 10:00 ～ 16:00
場所：理化学研究所計算科学研究機構 6F 講堂 (神戸)

理研神戸キャンパス一般公開の第 2 地区となる理研計算科学研究機構 (AICS) の 6F 講堂に設けられたポスト「京」重点課題の研究紹介や工作体験などを行う「神戸スパコンシミュレーション」にて、ブースを出展しました。今回で 2 回目の出展となります。

今年は、「「健康×医療×スーパーコンピュータ」の衝撃　ゲノモンと一緒にのぞいてみよう！」をポスト「京」重点課題 2 のテーマとして、研究概要を紹介するポスターを展示しました。また研究対象である「生命・ヒトのからだ」に興味を持ってもらおうと参加型ゲーム「人体の臓器パズル」を実施しました。このゲームは、身体の主な器官 16 パーツを「ヒトのからだ台紙」に合わせて、すべて正しい位置に置くことができればゲノモンシールが 1 枚もらえる構成です。3 才～小学生の子どもたちに家族が寄り添って試行錯誤しながらも一緒になって楽しむ姿が印象的でした。老若男女を問わず多くの方がパズルゲームに参加され盛況のうちに終了しました。

「人体の臓器パズル」を楽しむ来場者

サブ課題 A 小川誠司教授が、2017 年度武田医学賞を受賞

小川誠司教授が、武田科学振興財団が実施する 2017 年度武田医学賞を受賞しました。この受賞は、がん免疫研究に大きく貢献したとして、「成人 T 細胞白血病の分子基盤とがんの免疫回避に関わる新たなメカニズムに関する研究」の研究業績が高く評価されたものです。

「武田医学賞」は、医学界で顕著な業績を挙げ、医学ならびに医療に優れた貢献を果たした学者・研究者に贈呈されます。

贈呈式は、11 月 13 日 (月) にホテルオークラ東京にて執り行われました。

・武田科学振興財団の武田医学賞に関するホームページ
　http://www.takeda-sci.or.jp/business/doc/2017_prize.pdf
・小川誠司分担責任者の研究内容は、NEWSLETTER Vol.2 でご覧いただけます。
　http://postk.hgc.jp/_media/library/newsletter2.pdf

平成 29 年度 ポスト「京」重点課題 2 シンポジウム

その予防・医療、時代遅れです
ビッグデータ×シミュレーション×ポスト「京」＝ ＊ † ∞

2017 年 11 月 13 日に第 2 回目となるポスト「京」重点課題 2 のシンポジウムを開催しました。参加者はおよそ 100 名にのぼりました。アンケートにご記入いただいた貴重なご意見やご感想等は講演者にフィードバックし、今後の研究に役立てていきます。ご来場いただきました皆さま誠にありがとうございました。

シンポジウムの様子

文部科学省　ポスト「京」開発事業
重点的に取り組むべき社会的・科学的課題に関するアプリケーション開発・研究開発
重点課題 2　個別化・予防医療を支援する統合計算生命科学
Integrated Computational Life Science to Support Personalized and Preventive Medicine

■ 問い合わせ先
国立大学法人東京大学医科学研究所　ヒトゲノム解析センター　DNA 情報解析分野
ポスト「京」重点課題 2 個別化・予防医療を支援する統合計算生命科学 事務局

〒 108-8639 東京都港区白金台 4-6-1　TEL：03-5449-5615　FAX：03-5449-5442
E-mail：icls-office＠hgc.jp　　URL：http://postk.hgc.jp/

ポスト「京」重点課題は、国家基盤技術としてスーパーコンピュータ「京」の後継機となるポスト「京」を活用し、国家的に解決を目指す社会的・科学的課題に戦略的に取り組み、世界を先導する成果の創出を目指す文部科学省の事業です。重点課題 2「個別化・予防医療を支援する統合計算生命科学」は、東京大学医科学研究所を代表機関として、ポスト「京」によって初めて実現できる「情報の技術」、「物理の原理の応用」、そして「ビッグデータの活用」により、病態の理解と効果的な治療の探索法の研究を行い、その成果を個別化・予防医療へ返す支援基盤となる統合計算生命科学を確立することを目的としています。

COVER PHOTOGRAPH : Autumn leaves of Karasawa Cirque
NEWS LETTER VOL.6 2017.11