＜ テクノロジー×サステナブルな未来 ＞

東大医科研 ゲノムデータ解析で挑む、未知の病気解明と健康な日常

登壇者
東京大学医科学研究所
ヒトゲノム解析センター・センター長　井元 清哉 氏

聞き手
日本アイ・ビー・エム株式会社
常務執行役員 最高技術責任者　森本 繁典

**東京大学医科学研究所(東大医科研)ヒトゲノム解析センターでは、
がんや腸内細菌、新型コロナウイルスなど、未だ解明されていないことが多い医療分野において、
膨大なゲノムデータの解析により、真実を追求し、新しい発見に取りくまれています。
本記事では、10月5日～8日に開催したThink Summit Japanでの対談をノーカットでお伝えします。**

データがあればわかる。問題はどうデータを集めるか？

森本　2020年の1月から始まった新型コロナウイルスの感染症の世界的な大流行に、井元先生は最前線で戦ってきました。医療関係の領域に、データサイエンティストのプロとして対応されてきたことについてお聞かせください。

井元　私はもともと統計学を学んでいました。データサイエンスが専門ですから、データがあれば解析する方法を考えます。ただし、データがなければ話になりません。新型コロナウイルスで最初に困ったのは、データを集める方法でした。新型コロナウイルスに感染した患者さんの情報と、感染したウイルスの情報を同時に集める必要があります。それを集めるためのネットワークをつくったり、検体から採集したウイルスゲノムの情報を集めたりしたあと、今度はそれを「価値化」するという研究を進めることができるようになります。まずはデータを集めることが大変な課題でした。

森本　2020年の1月頃は症例も少なかったし、データ自体も少なかったわけですよね。そのころと今と比べると、解析をする上でどのような違いがありますか。

井元　2020年の4月ぐらいからだったでしょうか、日本ではまだ数十人くらいの規模だったんですけども、GISAIDという国際的なデータ・ベースがあり、それが

コロナのウイルスゲノムを集め始めたのです。

あれよ、あれよという間に数千から万を超えてウイルスゲノム配列が集まりだしました。われわれはすぐにアカウントを申請してデータを入手しました。そうすると早くもウイルスゲノムが少しずつ変化していることが分かってきたのです。これは大変なことになったと思い、解析するプラットフォームが必要だと考えました。

森本　ウイルスゲノムと言った場合、それはどの程度のデータ量なのでしょうか。

井元　ヒトのDNAというのは30億塩基対あります。一方、今回の新型コロナウイルスはRNAウイルスで約3万塩基しかありません。ごくごく小さいゲノムサイズなのですけれども、それがあっという間に変化するのがヒトゲノムと違うところなんです。

森本　なるほど、その変化を捉えたり、先回りして予測したりすることで、ウイルスがどういう性質を持つかが分かるということですね。

井元　その通りです。ウイルスゲノムの解析は世界中で行われていて、ウイルスゲノムのどういう箇所がヒト細胞に感染する際に重要な役割を担っているのか、ある程度分かっているのです。重要なゲノム領域に変異が生じて感染力が強まったりするのですから、できるだけ多くの感染者からウイルスゲノムを調べていくことが必要だと考えています。

人の記憶に頼らない確度の高い分析へ

森本　3万の塩基配列と患者の数やデータが増えることによって、マトリックス状のすごく大きなデータ・サイズになるわけですね。そういう中で、やっぱりコンピューターというのは非常に解析にとって重要だと思うのですが、いかがでしょうか。

井元　そうなのです。最初にウイルスゲノムのデータを免疫学者の先生と一緒に見始めました。われわれは何もツールを持っていませんでしたから、もうATGCの3万配列をとにかく並べて、そして多くの配列で少し違った塩基を持っている場所を見つけ出して、それをリスト化してディスカッションしました。しかし、あっという間に限界がきてしまいました。「これは何かしらの可視化ツールが必要だ」とすぐに気づきました。

森本　なるほど。井元先生にはIBMの「SARS-CoV-2 Variant Browser」というツールを一緒に開発していただきましたが、そのあたりのいきさつについて少しお話してください。

井元　ツールを最初に見せてもらったのが4月ぐらいでした。当時、可視化ツールを持っていなかったわれわれは、Variant Browserを見た時に「これは使える」と思いました。ただ、いくつか追加で欲しい機能がありました。特に新たな感染者からウイルスゲノム配列を得たときに、その配列の祖先に当たる配列が、データ・ベースに登録されていないかを瞬時に調べたかったのです。

日本では濃厚接触者の調査をしていましたが、人の記憶に頼ったレポートであるため限界がありました。「3日前にだれとどのくらいの時間会っていたか、その時にマスクを付けていたか、2人の距離がどれくらい離れていたか」といったことの記憶は曖昧です。でも、ウイルスゲノムを調べれば、より大きな意味を持つ情報が手に入ります。それを基に感染場所を調べることができれば、感染確率の高い場所や行動が分かると考えたのです。そのためには、ウイルスゲノムをトラッキングできる機能が必要でした。そういういくつかのアイデアがあって、ブラウザー開発に一緒に携わるようになりました。

森本　今ではウイルスの遺伝子を解析することによって、どの世代かということまで分かるのですね。そうすると、その同じクラスターでかかった複数人の人も、具体的に誰からどのように感染したか、その経路まで分かるようになったのですね。

コンピューティング・パワーは不可欠

井元　その通りです。1つのクラスターと判断されても全員が同時に感染したわけではない場合があります。例

えば、Aさんに感染したウイルスとBさんに感染したウイルスのゲノムを調べた結果、共に20箇所の変異があることが分かったとしましょう。そのうち19個はAさんのウイルスにもBさんのウイルスにも共通の変異です。Aさんのウイルスの残り1個の変異はBさんのウイルスには見られません。また、Bさんの残りの1個の変異はAさんのウイルスには見られません。この場合、AさんからBさん、もしくはBさんからAさんに感染したわけではなく、ほかの人から感染して、新しく変異が獲得されたと考えることができます。逆に、全く同じウイルスゲノムだとすると、感染関係にあるかもしれませんよね。つまり、ウイルスゲノムの情報とAさんとBさんの行動情報があれば、ハイリスクな場所や行動が見えてくるのです。そうした知見がたまっていくので、感染対策に役立つと考えました。

森本　なるほど、まさに今データサイエンスの力でそういった新しい事実や現状をあぶり出すわけですね。最近のコンピューターの進化はめざましく、ムーアの法則に従えば1年半に2倍ずつ価格性能比が上がるので、15年間で1000倍近くに達します。逆に考えれば15年前はコンピューティング・パワーが今より1000分の一も貧弱だったのですが、このような仕事はできたでしょうか。

井元　15年前の技術ではまったく不可能です。ウイルス以外の例もお示しします。私はゲノム領域で研究をしていますが、がん細胞からゲノムを調べると、がんの原因となる変異が分かってきます。そのゲノム解析技術がこの10年でとてつもなく進歩してきていて、その速度はムーアの法則を超えているのです。そのため、データはどんどんアウトプットされてきて、われわれはそれを解析し、患者さんにその結果を還元しようとします。そうすると、計算に必要となるリソースもムーアの法則を超える速度で拡張しないと追いつかないことになってしまいます。

森本　コンピューティング・パワーが上がっていくと同時に、処理すべきデータ量も増えていくわけですね。

井元　はい、さらに言うと、解析したい対象の種類も増えていくのです。

腸脳相関という不思議

井元　コロナウイルスのRNAが3万に対して、人間は30億塩基対の中に2万個の遺伝子があります。この遺伝子からタンパク質が作られて、それが私たちの身体をつくっています。しかし、もっとたくさんの遺伝子を持つ生物たちが私たちの身体の中というか表面にいることをご存じですか。それが腸内細菌叢です。腸内フローラとも言われていますね。なぜ腸内フローラと呼ばれているかと言うと、腸の中にはさまざまな細菌が私たちと共生しており、同じような種類の細菌は集まり存在し、その様子がさもお花畑（フローラ）みたいに見えることから、そう呼ばれているのです。その腸の中にいる細菌の数は100兆を超えると言われています。われわれの体は40兆くらいの細胞によって作られていますが、私たちの体の中には100兆個、1000種類を超える細菌がいて、食べたものの分解を助けて、私たちがつくることができない必須アミノ酸や短鎖脂肪酸をつくって、身体の恒常性を保っているのです。このバランスを崩壊させること、例えば

抗生物質を飲みすぎたり、油ものをたくさん取りすぎたりといった生活習慣によって、菌の種類のバランスが壊れ、腸内細菌が大切な物質をつくれなくなってわれわれの体に悪影響が出てきます。

腸内細菌はじかに腸に接しているわけなので、腸内細菌が悪くなれば腸も悪くなります。腸炎などに直結するわけです。しかしそれだけではありません。気分が滅入るとか、メンタルをやられるといった心の調子にも腸内細菌が関わっていることが分かっています。それを脳の状態と腸の状態が相関する「腸脳相関」といいます。

森本　1000種類で100兆個以上あるということは、その組み合わせを考えると天文学的な数になるわけですね。

井元　はい。僕はその解析が21世紀の大きな科学的な挑戦だと思っています。腸内細菌叢のバランスの崩壊で病気になってしまう、逆に自分が最高のパフォーマンスを発揮できたときに、腸内細菌がどんな状態だったのかを知ることができたらすごいことですね。ヒトゲノムと腸内細菌のバランスによって、調子の良さがつくりだされるわけです。

森本　そうなると、現在のAI（人工知能）はパターン認識を得意としていますが、ゲノムと腸内細菌バランスを考える場合は同じ「調子がいい」という時にも、対応パターンが無限にあり、組み合わせ問題のさらに複雑なものになってくるのですが、今のコンピューターの性能が100万倍になっても、完全な解析は難しい気はします。

井元　やはりコンピューティングのパワーが最初に必要です。それから、腸内細菌は便から調べるのですが、年一度の検便っていやですよね。それが腸内細菌叢を調べるときの一番のハードルだと思うのです。それをITで解決できるといいですね。スマート・ウォッチで心拍や睡眠時間などさまざまなバイタルが簡単にとれるようになりました。腸内細菌もそれくらい簡単に調べられるようになると思います。調子の良い時はどんな腸内細菌叢になっていて、どんな生活習慣から組み立てられているとか、個人個人でデータを分析することでわかってくれば、個人に対応した生活習慣を提案する将来が見えてきます。

森本　大量データとその因果関係の分析や、より細かなデータを個人単位で取ることも重要ですね。

井元　はい。日本では数十年前からコホート研究が行われてきました。コホート研究とは、特定の集団（コホート）を追跡することによって生活習慣と病気などの様々な傾向を知る研究です。どういう生活習慣がどういう病気につながるのか、という傾向を知る意味ですごく大切です。そして今、その集団の情報から個人にどのようにフィードバックするかが問題になっています。コホート研究を成立させるためには研究参加者の協力が必要不可欠です。研究参加者を募りデータを集めるために国から研究費を受けて多くのコホート事業は継続されていますが、参加者に何らかの利益がないと持続性があるとは言えないと思います。今は個人情報の保護やセキュリティーに注目が集まりがちですが、データを提供した人に何らかの利益を提供する仕組みを作る必要があります。その1つが解析結果を個人へ健康維持・向上の具体的なアクション・プランとしてフィードバックすることだと思います。

森本　今は、個人情報の保護とかプライバシー、セキュリティーっていうところに注目がいきがちですが、そのデータのオーナーシップと実際にデータを提供するメリッ

ト、動機づけをするということが非常に重要だということですよね。

井元　はい、そうなのです。セキュリティーやデータのオーナーシップを考え、今後個人にデータを返すという観点で真剣に話し合わなければなりませんね。

森本　それと、いかに簡単にデータが取れるかが重要ですね。1つの鍵は、AIなどを持ったセンサーの小型化だと思います。半導体の世界もどんどん微細化が進んでおり、最新のテクノロジーだと50ナノメーターのトランジスターが作れるようになっています。昔は真空管1個がそれくらいのサイズでしたけど、今はコロナウイルスの直径より小さく、その中にトランジスターを埋め込めるような時代になっています。そのまま進んで、例えば生体の中の情報を取っていくと、もっと進化するのでしょうか。

井元　体の中って結構難しいんですけど、ナノ・マシンみたいなものも今開発されていますし、今までは侵襲性が高い、例えば採血しなくては分からなかった血糖値といったデータも非侵襲で取れるようになってきました。そういう観測デバイスに加えて、その場で解析できてしまうようなエンジンが付いていると、またちょっと別次元の話になってきますね。

森本　さらに、各センサーの性能が上がって、感度が高まっていくということですね。

井元　例えば、その場で演算ができれば、血糖値を測って必要なインシュリンをリアルタイムに注入するといったことも可能になるかもしれません。

量子コンピューターと新しいテクノロジーにかかる期待

森本　先ほど井元先生が言っていたように、日々取ったデータと自分のゲノム情報、それから100兆の単位で存在する腸内細菌の組み合わせによって、コンピューティング・パワーがさらに必要になってきます。最近IBMでは量子コンピューターの発展に取り組んでいます。2021年は、日本に世界で2台しかない量子コンピューターの商用機を導入しました。これによって、先ほど話題に上がった組み合わせ爆発を起こす可能性の高い複雑性の高い計算を、将来的には量子コンピューターで今のスパコンよりも速く計算できるようになると期待されています。先生はそのあたりの可能性についても考えていますか？

井元　私の関連する分野で言うと、例えば人には2万個の遺伝子があります。しかし、この2万個の遺伝子に生じたゲノムの異常だけががんの原因とは限りません。ほかの領域に起きている変異がその病気に関連している可能性があるわけです。しかも、複数の変異が組み合わさってがんは生じていることが多いです。さらに、対象はヒトゲノムだけではありません。腸内細菌叢には100兆個の細菌がいると言いましたが、その細菌たちが持っている遺伝子数は、人よりも100倍～1000倍くらいあるんです。

1000～2000万種類の遺伝子を細菌は持っていて、その細菌遺伝子から作られるさまざまな酵素が私たちのヒトゲノムの遺伝子が作る酵素と協働してアミノ酸の合成や代謝反応が成立しています。細菌叢のバランスが崩壊し、私たちの体の恒常性に必要な遺伝子を持つ細菌が消失することが病気につながることがあるわけです。だから、その組み合わせ問題を解いていく必要があるんです。そのために量子コンピューティングのような次の時代の計算機リソースは必須になっていくと思います。

森本　計算機の分野では、まだまだ技術革新が重要ということですね。小さいところでは、体内に入れるセンサーやマイクロ・マシン、さらに少しデータを集めてきたら、今度はそれを集合的に観察する、さらには腸内細菌やそれ以外の要因、組み合わせた時の体調や因子などを考えると、非常に膨大な情報になっています。さらに、統計値というよりは、個人の生活や体調管理をするということですから、コンピューター・リソースはいくらあっても足りないですね。

井元　個人個人もそうですが、個人を評価するために集団の中での個人を考えなくてはいけません。そうすると集団の解析をしなくてはいけないわけですね。そうする

とさらにコンピューターの能力が必要になるわけです。

森本　先生もご自身でハイパフォーマンス・コンピューティング（HPC）の「SHIROKANE」を運営し、実際に研究されていますが、これからの社会でのHPCの重要性をどのように考えていますか。

井元　今データを観測するセンサーの発展はめざましいですが、それを生活に生かせているかというと、持っているデータの価値の100分の1も使えていないと思います。1つの理由は解析できていないからです。センサーの能力が日々向上していて、新しい要因を観測できるようになっています。データを解析して、われわれの生活の質が高くなるようにしたいですね。コンピューターがよりパワフルになり、巨大なデータを瞬時に転送し、解析し、個人にフィードバックできるネットワークの強化が必須と考えます。

テクノロジーと医療。求められる「バイリンガル」

森本　テクノロジーの開発も重要ですし、道具や技術を作って実際の医療や健康の研究に供するというコラボレーションが重要です。しかし、テクノロジーと医療の両方を理解する井元先生みたいな「バイリンガル」が増えて欲しいと思います。そのあたりの技術人材の状況はいかがでしょう。

井元　今は随分と分野の融合が進んでいます。実際に医師免許を持って診療に当たっている医師が、私の研究室に来てディープ・ラーニング（深層学習）を勉強しているのです。そういう時代になっていますから、今後バイリンガルのような人が増えていくと思います。しかし、やはり人材育成コースとか学部などは必要になるでしょう。

森本　最後に、そういう分野を目指す若い技術者や医師にメッセージをお願いします。

井元　ビッグデータの時代といわれて久しいです。ビッグデータは構築されつつありますが、それを活用するデータサイエンティストにとっては、やることが山積みになっています。データが眠っているという状況が、特に医療・健康分野では見られます。非常に重要なデータを、例えば医師の方だと、自身の臨床的な観点から解析することができるわけですから、医療関係の人にはどんどんこの分野に入ってきてほしいです。

　また、もっとコンピューター・サイエンスから医療分野に入る人も増えていくべきだと考えます。さまざまなデータ形式、構造化および非構造化データが山ほどあります。それをいかに保存し、どんなインフラで解析すればより効率的か、より高速なアルゴリズムの構築や新たなデータ解析技術を作り新しい知識の発見といったコンピューター・サイエンスの問題もあります。そうした複数の方面から医療のビッグデータを解析するようなコミュニティができないと、活用は難しいと思います。いろいろな研究分野から、健康医療のビッグデータ解析に入ってきてもらいたいです。

森本　われわれの健康に関する問題は新型コロナウイルスだけでなく、これからも続きますし、多くの技術的な課題もあることが分かりました。また、計算機や情報産業の新しい技術を生み出すのはもちろん、道具として使いこなして価値に変えることの重要性がよく分かりました。井元先生、本日は誠にありがとうございました。