バイオサイエンスデータベースセンター・ワークショップ 報告書

「データ駆動型研究の推進と課題」

開催日：2020 年 12 月 1 日（火）

2021 年 5 月

国立研究開発法人科学技術振興機構(JST)
バイオサイエンスデータベースセンター(NBDC)企画運営室

目次

I. エグゼクティブサマリー

国立研究開発法人科学技術振興機構（JST） バイオサイエンスデータベースセンター（NBDC）は、2020 年 12 月 1 日、ワークショップ「データ駆動型研究の推進と課題」を開催した。本報告書は、本ワークショップでの発表、議論をまとめたものである。

近年、生命科学研究では、計測技術の驚異的な進歩により、研究データが爆発的に増加し、情報のデジタル化、コンピューテーショナル化が加速度的に進んでいる。こうした研究データを集積、整理・統合し、生命現象を包括的に理解する研究開発は既にライフサイエンスの潮流を形成しているが、一方で、データの共有・公開基盤の構築や公共データの高度利用では課題も多い。

こういった認識は 2000 年のヒトゲノム解読に端を発したオミクス研究の勃興から現在に至るまで当該分野の基本思想として定着しているが、とりわけわが国ではいまだにデータを活用したライフサイエンスの取り組みは限定的で、論文レベルでの国際競争力の低下も著しい。

ワークショップの開催に先立って有識者へのインタビューを実施したところ、上記課題を解決するために次のような研究スタイルの推進が有効であるとの仮説が浮かび上がった。その研究スタイルとは、これまでに公共データベースに膨大に蓄積され、今後もさらに増していく多種多様な生命科学データを複合的に解析して仮説を立て、実験的に検証して新たな発見を導くというものである。公共データを活用することで、仮説立案から検証までを、効率的にかつハイスループットに実施できるようになる。また、公共データを統合利用することで、探索範囲を、従来の方法論では到底なしえないほどに広げることができるため、これまで考えも拠らなかったような想定外の発見に至ることが期待される。ここでは、こうした研究スタイルを「DX（デジタル・トランスフォーメーション）型データ駆動研究」と呼ぶ。DX 型データ駆動研究を推進するには、データを収集・整理し、公開する基盤を開発して構築・運用する「データベース構築者」、データを統合・解析する技術を開発し、またその技術を応用して新たな仮説を生成する「データ科学者」、仮説を検証する「実験研究者」の協働が不可欠であると考えた。

本ワークショップでは、DX 型データ駆動研究の実行可能性や課題点を実例に基づいて議論することを目的とし、有識者から話題提供をいただいたあと、総合討論を行った。

石濱 泰氏（京都大学）は、特にプロテオームデータの再利用を中心としたデータ駆動型研究の実例について、国際的なデータ集積、公共データベースのリン酸化プロテオームデータを用いた大規模再解析、公共データベースのデータを用いたキナーゼ予測ツールの構築といった実例を紹介した。また、データベースの種類として、生データの集積基盤のほかに知識体系の集積基盤があり、世界全体の学問の下支えとして重要であること、DX 型データ駆動研究の推進には、データベース構築者とデータ利用者が協働できる仕組み作りが必要であると指摘した。

竹本 龍也氏（徳島大学）と沖 真弥氏（京都大学）は、ドライ研究者とウェット研究者の協業の実践例を紹介した。ただ、2 氏の事例は、おのおのが有する技術・特徴がうまく噛み合った希有なものであって、広く推進するためにはデータベース構築者と実験科学者とが協働するファンディングと、前提を相互に理解しあう機会が重要であるとの認識を示した。

坊農 秀雅氏（広島大学）は、公共データベースを活用し、低酸素変動遺伝子候補のメタ解析研究の実例につ

いて紹介した。隘路として、公共データベースのデータを利用するためには十分な人手を介する必要があると指摘した。

鎌田　真由美氏（京都大学）は、データベースの構築と収録データの解析についての実例を紹介した。DX 型データ駆動研究を成功させるためには、「データベース構築者」とデータ提供者との相互理解の場を充分設ける必要があったこと、また一般に、オントロジーが充分整備されてないためにデータを活用するための前処理に時間がかかると指摘した。さらに、データを収集、変換し、データベースとして公開する「データエンジニアリング」の重要性が、データサイエンスが注目を集めるのと対照的に、十分認知されていないと指摘した。

山西　芳裕氏（九州工業大学）は、さまざまな公共データベースを組み合わせた解析から知識を抽出し、化合物の新たな効能を予測した例を複数紹介し、公共データに基づいたデータ駆動型研究の可能性を示した。共同研究に際して「データ科学者」と「実験研究者」双方の意思疎通が重要であること、質、量ともデータが増え続ける状況下にあって公共データの利用には ID、フォーマットの不統一がボトルネックであり、継続的な対応が望まれることなどを指摘した。

6 名のコメンテーター、伊藤　隆司氏（九州大学）、小安　重夫氏（理化学研究所）、菅野　純夫氏（東京医科歯科大学）、瀬々　潤氏（ヒューマノーム研究所）、永井　良三氏（自治医科大学）、平井　優美氏（理化学研究所）からは、ご自身の経験に基づき、幅広い視点でのコメントがあった。

全体議論では、三者が連携して推進する DX 型データ駆動研究が重要かつ効果的な施策であると示された。DX 型データ駆動研究において、データベース構築者には、データの質、量が年々大きく変わっていくなかで、複数プロジェクトの研究データを、利用イメージを充分理解したうえで統合していくことが期待されている。一方、データ科学者には、アルゴリズムそのものを新たに開発するというよりも、既存アルゴリズムを如何に生命科学分野へ応用し、データから仮説を如何に絞り込むかについての道筋を付けることが期待されている。また、国内の取り組みは現状ゆっくりであるが、海外の事例を踏まえるとスピード感をもった対応が不可欠だとの指摘があった。DX 型データ駆動研究の推進方法としては、次の 2 つが示された。1 つめは、実験研究者とデータベース構築者、データ科学者の協業を促進する方向性である。2 つめは生命科学とデータ解析の双方について極めて高い見識を有する研究者に投資し、有効性を提示する方向性である。また、推進の過程で、データ解析の概念・方法論をツールとして構築し、提供することでより多くの研究者の参入が期待されるとの意見もあった。

本ワークショップを踏まえて JST では、今後具体的な研究開発課題や研究開発の推進方法の議論をさらに深めることとした。

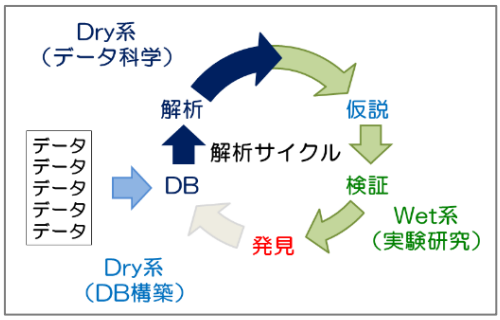


図 1-1 データ駆動型科学の解析サイクル

瀬々　潤氏 トーゴーの日シンポジウム 2020 発表資料を改変

II. 趣旨説明

1　データ駆動型研究の現状と課題

太田 紀夫（JST-NBDC）

(1)　発表内容

本日のワークショップ開催の背景には、近年、多種多様でかつ膨大な量の生命科学情報が蓄積してきているという点と、生命科学分野でのデータ駆動型研究が世界的に注目され、重要性が認識されてきているという点がある（図 1-2）。

図 1-3 には「データ駆動型研究」の解析サイクルを図式化したものを示した。本ワークショップでは、データ駆動型研究を「データに基づいて仮説を立案し、それを検証することで、新たな発見に結びつける研究手法」と定義する。

データ駆動型研究の最近の例として、猛威を振るう新型コロナウイルスの進化系統解析が挙げられる（図 1-4）。世界各地で採取された新型コロナウイルスの配列データを使って、いま世界のどこでどのように変化したウイルスの感染が広まっているかといった状況が、ほぼリアルタイムに明らかにされた。

我々は、ライフサイエンス分野において、

- データのオープン化が進み、それらの統合が試みられてはいるものの、十分には活用されていない、
- COVID-19 関係でも、単一の情報を使った、比較的単純な解析されているケースが多かった、
- 今後、どの分野でも多様なデータベースを活用して新たな価値を提供する時代になっていくと考えられているが、
- ライフサイエンス分野ではデータを活用した新たな価値創出の事例が少なく、問題である、

と考えている（図 1-5）。

そこで、

- これからの生命科学研究の発展には、多種多様かつ膨大な生命科学情報を活用したデータ駆動型研究の促進が必要である。
- そのためには、「データベース基盤」、「データ科学」、「実験研究」の 3 者の融合が、今まで以上に求められる。
- そして、これらの真の融合には、データ駆動型研究の解析サイクルを効率良く回すための研究開発投資と、それを動かす人材の掘り起こしが必要である。

という 3 つの仮説を設定し、どうしてこれまでこうした「データ駆動型研究」が進んでこなかったのかについて、30 名近くの有識者の先生方にご意見を伺った（図 1-6）。その結果、技術的な問題、研究支援の仕組みの問題、人材の問題、など、データ駆動型研究を進めていく上で障害となっているさまざまな課題が挙がった（図 1-7）。

図 1-8 に、本ワークショップの趣旨を示した。本ワークショップでは、生命科学分野において、研究データを高度に利活用するための研究開発上の課題と、その解決に資する研究開発投資を議論し、今後の研究支援のあり方を検討したいと考えている。

本日ご出席の皆さんの意識合わせのために、データ駆動型研究について、実例を当てはめながら類型化を試みたい。

データ駆動型研究の典型的な例としては、山中４因子の発見が挙げられる（図 1-9）。この因子の発見は、理化学研究所の FANTOM データベースを使って、ES 細胞で特徴的に発現している 24 種類の転写因子を絞り込み、そのいずれか、あるいはその複数の組合せにより幹細胞性が既定されているという仮説を立て、それを検証することによって実現した。

近年は、より多様で複雑なデータを扱ったデータ駆動型研究もなされている（図 1-10）。ゲノムやトランスクリプトームなど、さまざまなオミックスデータをシステマティックに取得し、それらを複合的に解析して仮説を立て、実験的に検証して新たな発見を導いているケースだ。本ワークショップでは、こういったタイプを、仮に「マルチ型のデータ駆動型研究」と呼ぶ。このケースは最初に実験をデザインする段階で予め仮説が置かれていることから、「データ駆動型研究」であると同時に、仮説駆動型研究の側面も持つ。

図 1-11 に示したのは、今後、推進していくべきと考えているデータ駆動型研究だ。先ほどの「マルチ型のデータ駆動型研究」との違いは、あらかじめデザインした実験で取得したデータセットを解析するのではなく、公共データベースに膨大に蓄積している多種多様な生命科学情報を解析対象としている点だ。ご存知の通り、公共 DB（データベース）に登録されているデータは、さまざまな研究者が登録したもので、実験条件が違っていたり、場合によっては測定系が異なっていたりもする、とても不揃いなデータの集団だ。従来の科学の考え方では、こうした異なる実験条件で取られたデータの比較は、原則厳禁とされている。しかし、こうしたデータでも、膨大に集積してくることで、適切な解析により、十分意味のある知識が抽出できるようになってきている。AI を使った SNS ビッグデータ解析で、人々の動きや社会のトレンドを解析した事例などは、最近、皆様もよく目にすると思う。生命科学分野でも、公共 DB にこのように膨大なデータが蓄積してきていることから、これらの情報をもっと活用出来るようになってきているのではないかと考えている。そこで、こうした公共 DB のデータを使ったデータ駆動型研究を、本ワークショップでは仮に「DX（デジタル・トランスフォーメーション）型のデータ駆動型研究」と呼ぶ。本ワークショップでは、この DX 型データ駆動研究の実現可能性や課題などについて議論したい。

「DX 型のデータ駆動型研究」の利点の一つは、膨大かつ多種多様なデータを探索範囲にすることで、これまで考えも拠らなかったような想定外の発見や、思いがけない成果が生まれる可能性が期待される点にある（図 1-12）。また、時間と共に新しいデータが継続的に追加されていくことから、探索範囲は将来もずっと拡大していく。

一方で、大きな課題もある。公共データベースに登録されているデータは、実験条件や測定系の違い等から、とても不揃いなデータ集団になる。したがって、P 値を指標とした従来の統計学的手法による解析が馴染まない。このため、データの正規化や QC、AI 関連技術の活用なども含めた新しい解析手法の開発が必要になる。

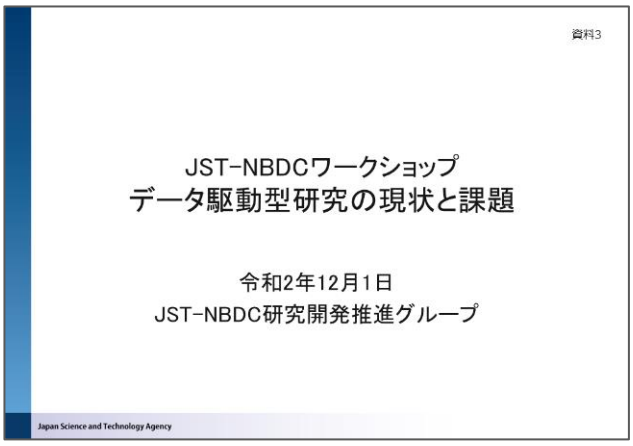
「DX 型のデータ駆動型研究」のもう一つの利点は、解析サイクルのスピード感だ。新たに実験をして取ったデータを解析するのではなく、既に DB にある情報を解析して仮説を立案していくことから、スピード感を持って研究の解析サイクルを回していくことができる（図 1-13）。

本ワークショップのゴールは、以下の 3 つと考えている（図 1-14）。

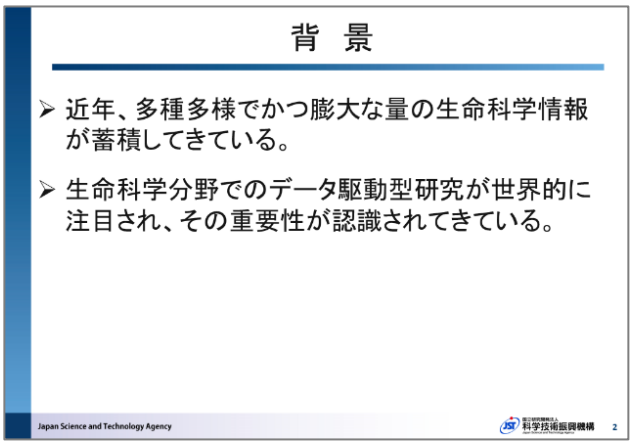
1. 公共 DB の膨大な情報を活用したデータ駆動型研究の意義
2. こうしたデータ駆動型研究の解析サイクルの隘路と、その対応策
3. 上記議論を踏まえた上での、研究支援のあり方

本日は、こうしたデータ駆動型研究に実際に取り組んでいらっしゃる５名の先生方にご登壇いただき、DX 型データ駆動研究の解析サイクルの隘路とその対応策を議論したい（図 1-15）。

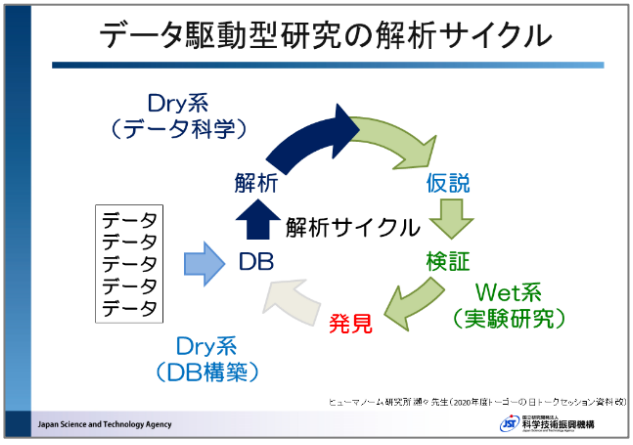
(2)　資料



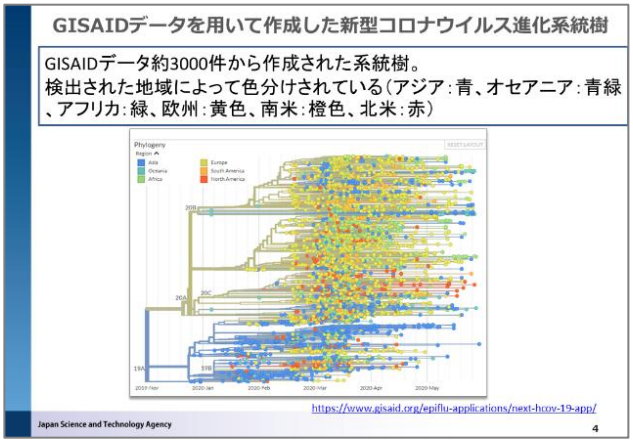
1-1 表紙



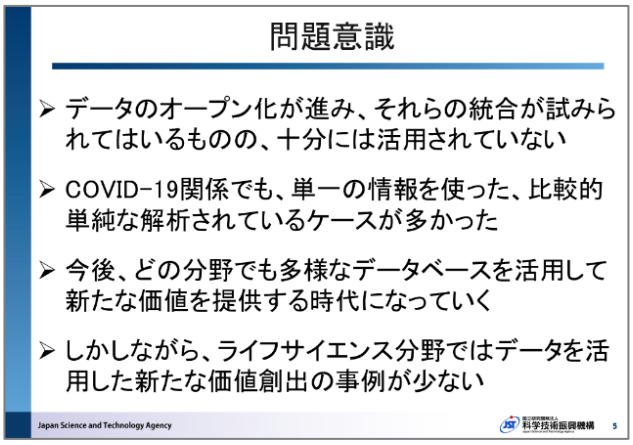
1-2 背景



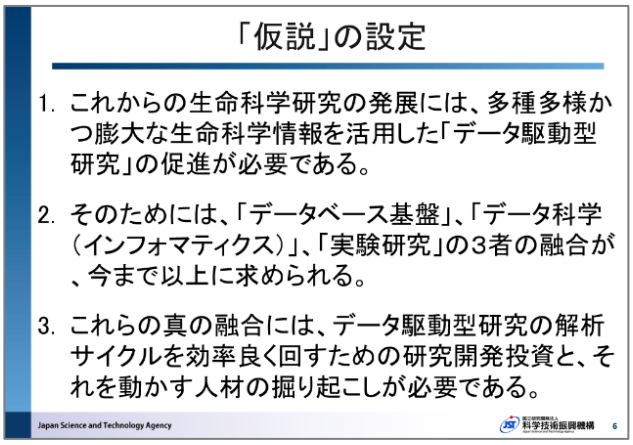
1-3 データ駆動型研究の解析サイクル



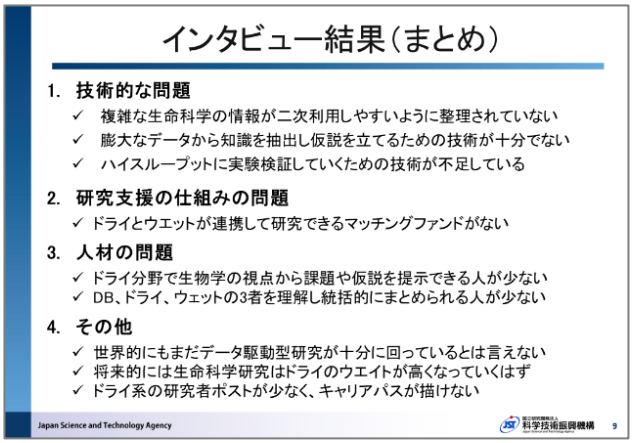
1-4 GISAID データを用いて作成した新型コロナウイルス進化系統樹



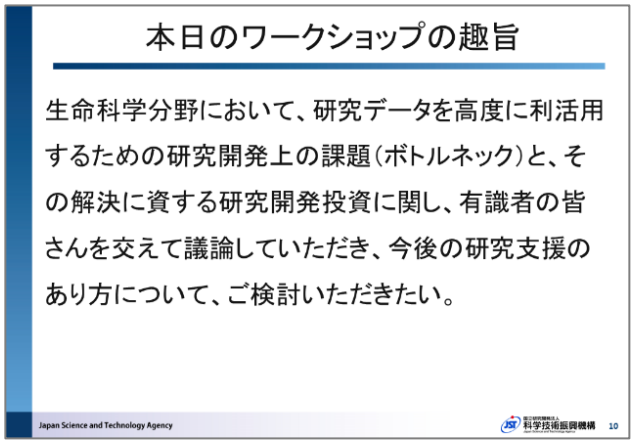
1-5 問題意識



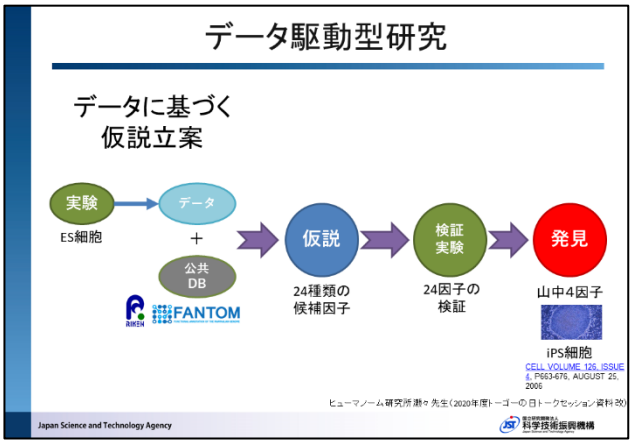
1-6「仮説」の設定



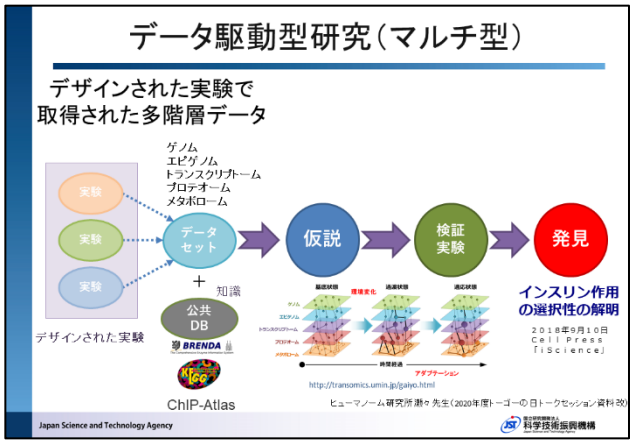
1-7 インタビュー結果（まとめ）



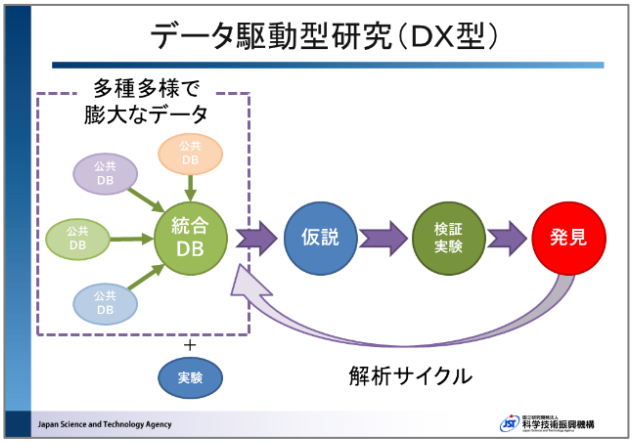
1-8 本日のワークショップの趣旨



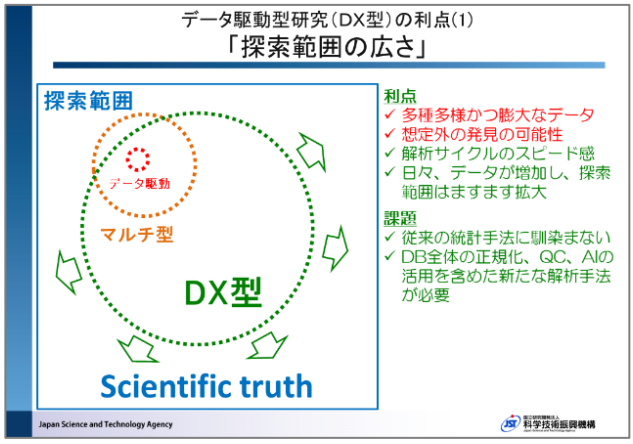
1-9 データ駆動型研究



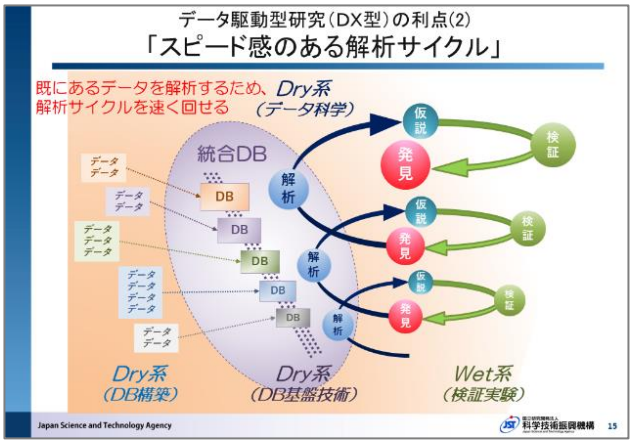
1-10 データ駆動型研究（マルチ型）



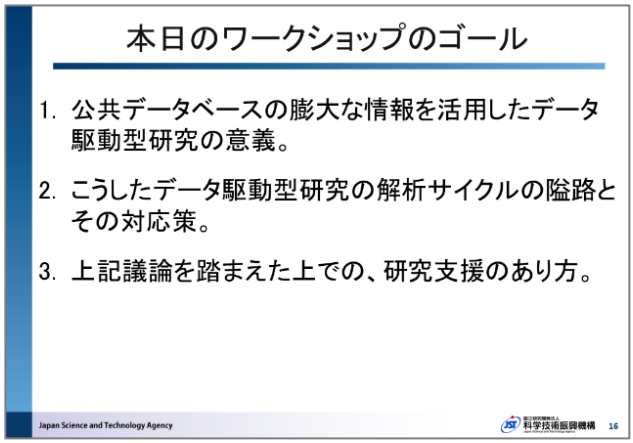
1-11 データ駆動型研究（DX 型）



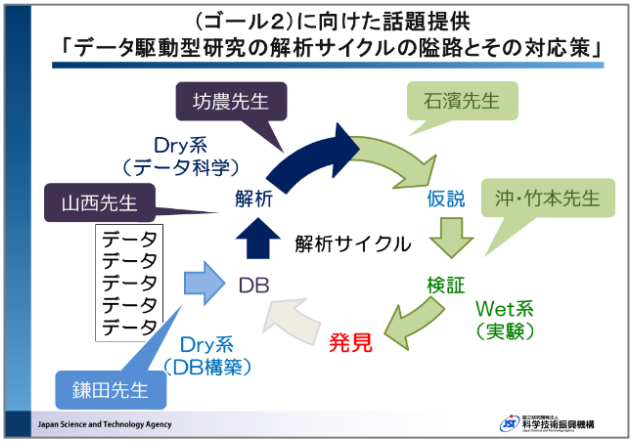
1-12 データ駆動型研究（DX 型）の利点(1)



1-13 データ駆動型研究（DX 型）の利点(2)

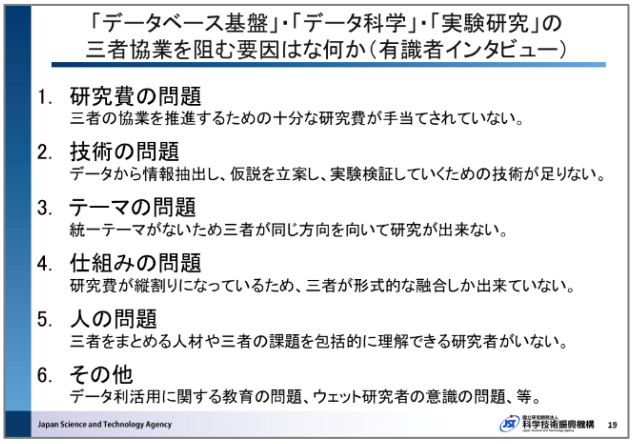


1-14 本日のワークショップのゴール



1-15 ゴール 2 に向けた話題提供

1-16 補遺



1-17 DB 基盤・ﾃﾞｰﾀ科学・実験研究の協業を阻む要因

III. 話題提供

1　jPOST/ProteomeXchange を用いたデータ駆動型科学

石濱　泰（京都大学）

(1)　発表内容

jPOST と言うプロテオームデータベース――国際コンソーシアムの中では ProteomeXchange というもののコアメンバーとなっている――のデータを用いたデータ駆動型科学を紹介する。

図 1-2、1-3 に示した、Cancer MoonShot 2020 を御存じだろうか。次の大統領になるはずのバイデン氏は、2006 年、前政権時代の副大統領時にがんのムーンショットプロジェクトを立ち上げ、この 2020 年までにがん撲滅を目指すという非常に大きなプロジェクトを立ち上げた。息子を脳腫瘍で亡くし、残りの政治生命を全てかけてがんを撲滅すると宣言した。当時、ダブリンで世界のプロテオーム学会があったが、それにも参加し、非常に大きなインパクトを与えた。

最終的に、世界 16 カ国のメンバーが集まって MOU を取り交わした。日本も署名した。この枠組みで、プロテオゲノミクスに取り組むこととなった。これはゲノムの変異情報とプロテオーム情報を組み合わせることによって、がんを撲滅するための何か新しい方向性を見つけようというものである。

図 1-3 は日米韓で記者発表したときのプレスリリース文だ。がんプロテオゲノミクス研究のデータを大量に取得し、深層学習アルゴリズムを利用して世界中のデータサイエンティストを呼び込んで、がん撲滅の道筋をつけることが非常に大きな目標であった。

残念ながら日本では、その後全く動きがなかったが、世界では一部動いていて、データ・シェアリングが重要な項目になっている。NIH が資金を拠出し、Proteomic Data Portal、もしくは Proteomic Data Commons を構築している。図 1-4 に示した Proteomic Data Portal を見ていただきたい。このように、データが既に蓄積されていて、どんどん論文がいいところに出ている。プロテオームデータだけではなくて、リン酸化プロテオームや、ユビキチン、アセチル化、グライコプロテオーム等のデータも集積しているとともに、米国内のデータだけではなく、国際コンソーシアムのデータもどんどん入れていくことになっている。

問題は資金提供元、もしくはデータ提供元である NIH、もしくは NCI そのものがこの管理をしていることだ。独立性に大きな疑義がある。例えば、この多くの論文では非常にいろいろなデータ処理をしながら発表されているが、生データを第三者が解析すると実は違う結果が出たりする。都合の悪いデータがいつでも隠せるような仕組みはよくない、というのが我々のコンセンサスだ。

今、世界のプロテオームデータの標準は ProteomeXchange だ(図 1-5)。主としてはイギリス EBI の PRIDE、米国の MassIVE、PASSEL。jPOST は第 3 番目に ProteomeXchange コンソーシアムへ加盟し、アジア・オセアニアの中心リポジトリとすべく運営している。その後、幾つか新しいものが加わり、また先ほどの Proteomic Data Commons も加わりたいとのことだが、独立性に大きな疑問があり、現在のところ、オブザーバーということになっている。

jPOST は、我々は世界中の生データをリポジトリに取り込み、それを再解析して世の中に公開することで、データの統合と共有のための仕組みをつくっている（図 1-6）。

この一例だが、2014 年の Nature にヒトプロテオームのファースト・ドラフトが発表された。このファースト・ドラフトは、この ProteomeXchange からデータを取り込んで、1 万個以上のファイルを再解析して発表したものだ。しかし、いろいろな問

題があった。この研究論文をひとつのきっかけにして jPOST でも再解析をおこなった（図 1-7）。これはもともと日本のデータだが、オリジナルデータに比べてはるかに質も量もいいものができていると考えている。

もう１つの成功例としては、ProteomeXchange の約 6,800 件のリン酸化プロテオームのデータを全部集め、それを再解析した研究がある（図 1-8）。オレンジの線が引いてあるのが EBI のデータベースの研究者、青い線で引いてあるのはデータサイエンティストの名前だ。ペドロ氏が中心になって解析し、2020 年 3 月にヒトのリン酸化プロテオームのランドスケープとして発表された。今まで見つからなかった、五十数種の新しい特徴を見つけ出したと報告している。

また、日本の状況を見るために、NBDC が所掌するライフサイエンスデータベース統合推進事業において、毎年開催されているトーゴーの日シンポジウムのキャッチフレーズを並べた。この事業は、データベースを使う人が部外者で、先述の研究にあるような体制にはなっていない。データベースをつくる人と使う人がごっちゃになっていないと、こういうプロジェクトは出てこないのではないか。

先ほどのペドロ氏だが、ヒト・リン酸化プロテオームのランドスケープを既に手に持っていたので、COVID-19 の問題が出てきたときに、いち早くデータを取り込み、ウイルス感染時に何が起きるかをヒト・リン酸化プロテオームの観点から再解析をした。これが 2020 年８月に発表されている。また、彼は、タンパク質・タンパク質相互作用の、COVID-19、SARS-CoV2 の影響を見た、という論文の共著者にもなっている（図 1-9）。当該論文は４月に発表され、既に引用数は 800 回を超えている。

話が戻るが、本論文のメインのアウトプットは、どのキナーゼが COVID-19 に対して何か影響を及ぼしているか、だ。キーはキナーゼの予測だ。我々はペドロ氏と長年コラボレーションし、2016 年には論文を発表している（図 1-10）。タンパク質-タンパク質間相互作用のデータと公共データベースにあるリン酸化プロテオームのデータを使って、どのキナーゼがどういう基質を持つかを予測している。

これを発展させ、我々が取得したイン・ビトロのキナーゼと基質の関係をベースにしたキナーゼ予測、もしくは基質予測ができるようになっている。図 1-11 に示したのは jPOST のツールのひとつだ。こういうものがあると、先ほどの COVID-19 の話が出てきたときにも素早く対応できる。

役に立つデータベースというのは一体何だろうか。図 1-12 は大規模計測データ、オミックスのデータを集積したようなデータベース、DepMap だ。がんにフォーカスして、その大規模データを集めて整理して、研究者に提供している。米国の Broad Institute が DepMap プロジェクトとしてこれを運営している。新しいデータセットが継続的に入り、またオープンサイエンスへのコミットメントをベースにしているため、データが全て CC-BY 4.0 になっている。データサイエンティストにとっても、一つ一つのタンパク質ベースで仮説を持って研究する研究者にとっても、こうしたデータベースは非常に有用になると考えられる。

DepMap は生データを集積してつくられたデータベースだが、そうではなく知識を集積したデータベースというものもある。タンパク質関係では UniProt が一番知識を集約しているものだと思う。UniProt の中にいろいろなデータベース・文献からのデータが集約されている。例えば Syndecan-1 のページには、例えば我々の jPOST も掲載されており、どういう実験条件で、どういうペプチドが何回測定されているかというデータも入っている（図 1-13）。もう一例で、先ほど少し触れた KEGG がある。KEGG は色々な知識を集約しているが、一番よく使われているのはパスウェイの解析で、この Syndecan-1 が、どういうパスウェイのどこにいるかといったことが分かる。

こういうものは世界全体の底上げにつながる知識体系だ。公共性を持って独立で運営されていて、やはりフリーであることが重要だ。あるデータベースは、利用条件の問題から人気を失い、残念ながら別のプロジェクトで構築されたデータベースに置き換わりつつある。

生データとメタデータが組み合わさって初めてデータサイエンスに供するためのデータとなるということが重要で、データ共

有とデータ統合を進めていくことがデータサイエンスを進めることになると思う（図 1-14）。常にアップデートされている、どんどん新しいデータが入ってこなければいけない、ここも非常に重要だ。先ほどの NCI のサイトなどでは独立性が担保されておらず、よくないと思う。独立した研究機関がデータを公開すべきだ。公共性と独立性がなくてはいけない。百科事典のようなものも必要だ。

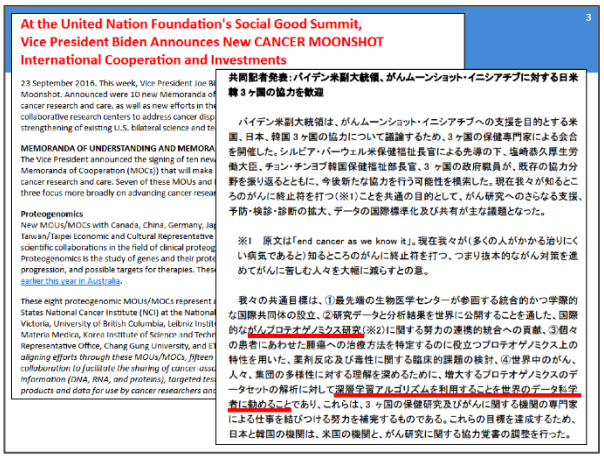
利益を考えるのであれば、セレラ社のヒトゲノム配列データを武田薬品が購入し、配列データ全体がオープンになる前にオーファン GPCR を同定してゲノム創薬につなげた例がある。こうしたデータは売れるであろうが、こういうことをやるために我々はデータベースをつくっているのではないということも言えるかと思う。とにかく何に使えるのか分からないけど集積することも重要で、学問全体の下支えになる。国で囲い込むメリットはなく、国際展開すべきであると思う。

最後に、データを出す人、データベースをつくる人、使う人は三位一体になっている必要がある。特に、計測のところで、ゲノムのように 4 文字で表わすことできないデータは、データ産生者がつくる人、もしくは使う人と三位一体になっていることが重要だと考える。

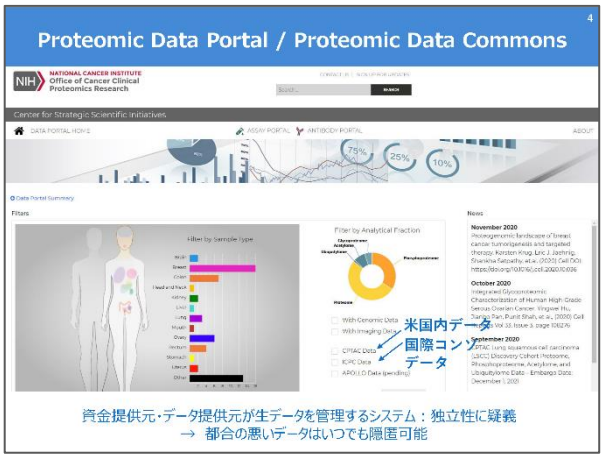
(2)　資料

1-1 表紙

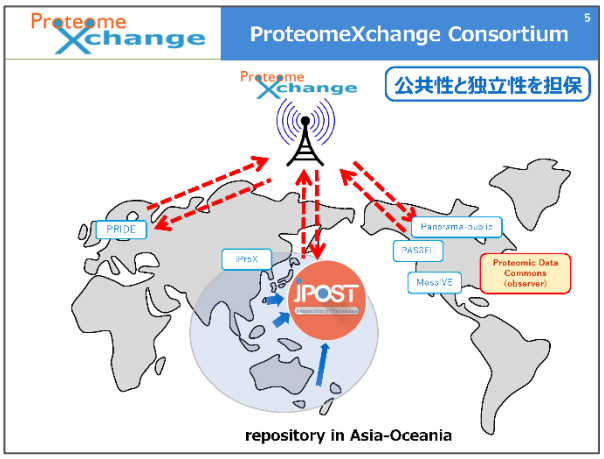
1-2 Cancer MoonShot 2020(1)



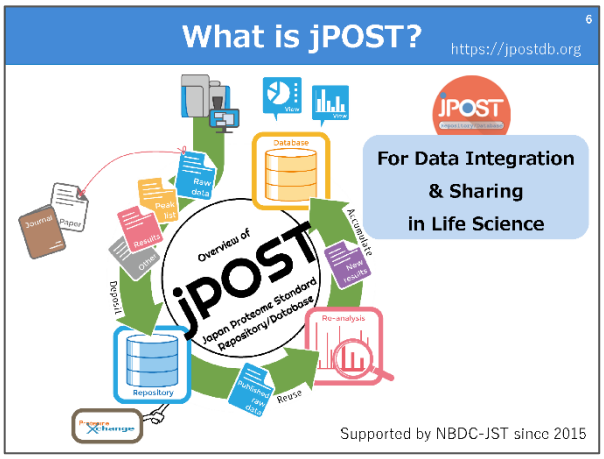
1-3 Cancer MoonShot 2020(2)



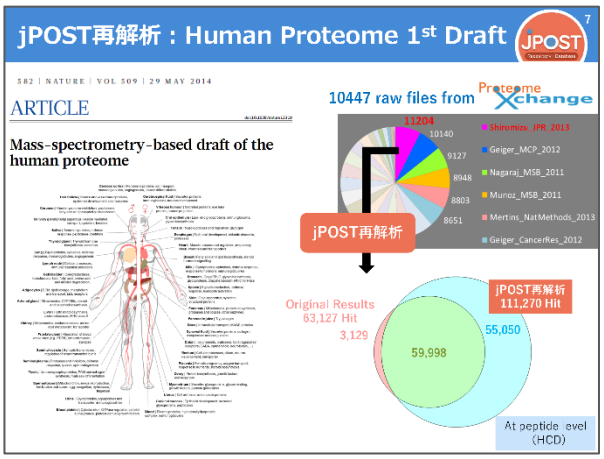
1-4 Proteomic data Portal



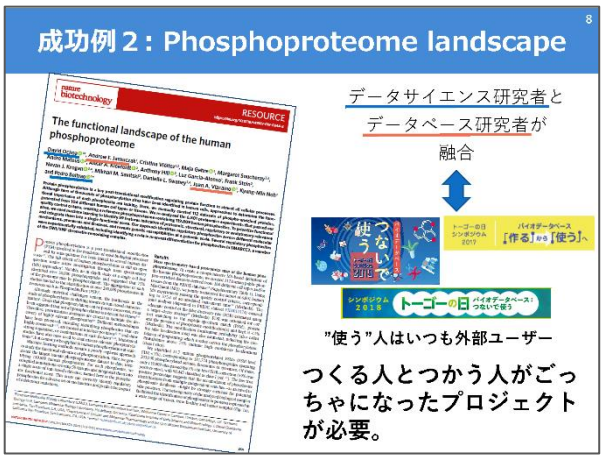
1-5 ProteomeXchange Consortium



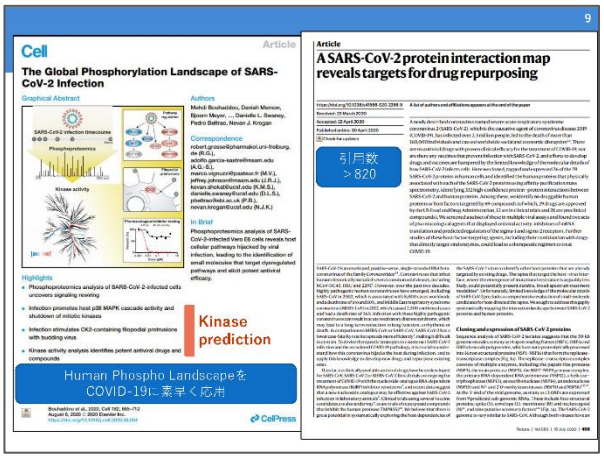
1-6 What is jPOST?



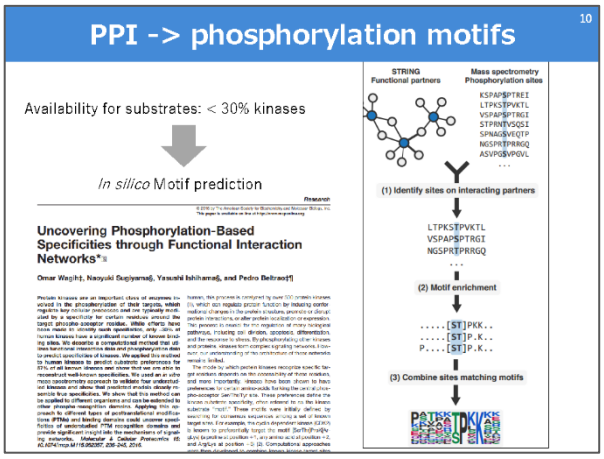
1-7 jPOST 再解析



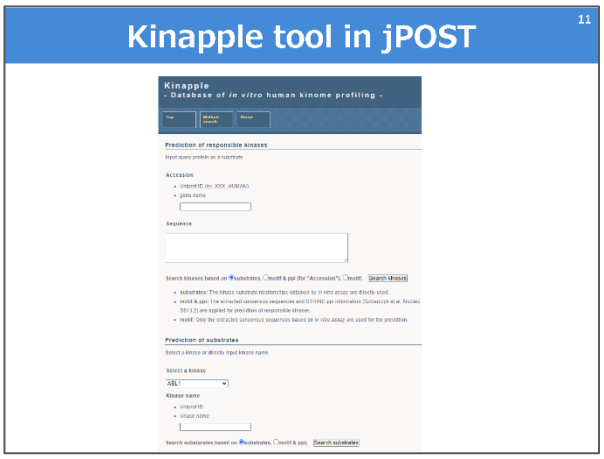
1-8 成功例 2



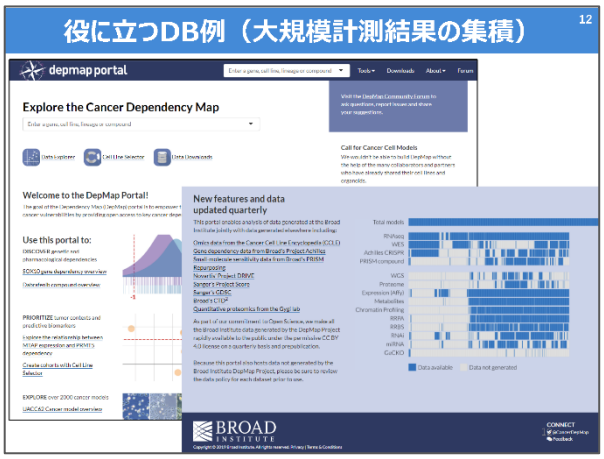
1-9 SARS-CoV-2



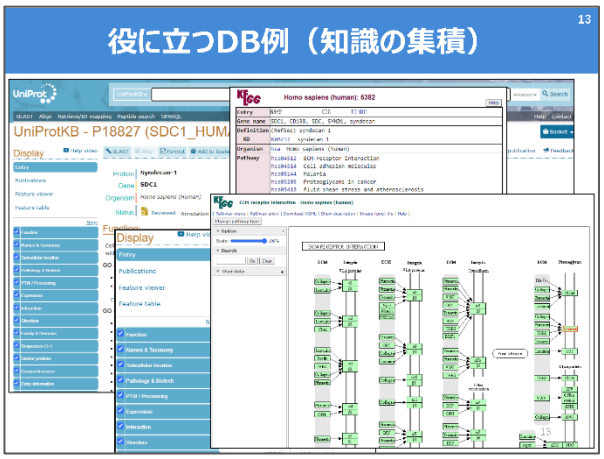
1-10 PPI→Phosphorylation motifs



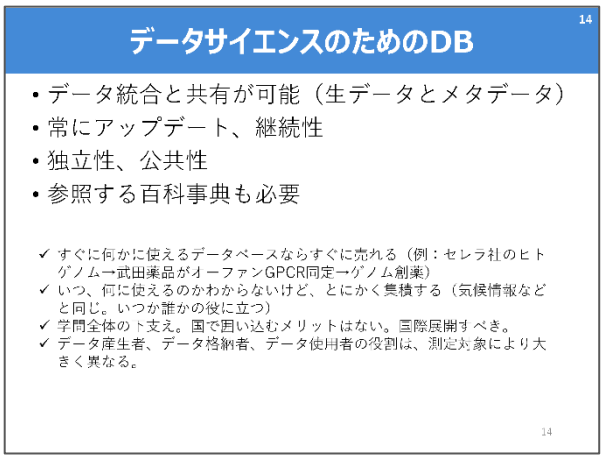
1-11 Kinapple tool in jPOST



1-12 役に立つ DB 例（大規模解析結果の集積）



1-13 役に立つ DB 例（知識の集約）



1-14 データサイエンスのための DB

(3)　質疑応答

（質問）データベースの独立性ということを非常に強調されていた。データベースの構築・運営の費用は誰が持つべきかに関し、コンセンサスが世界的に得られているのか。

（回答）得られていない。プロテオーム分野で最初に公共データベースとしてできたのは「トランジェ」だが、資金がなくなり、途絶えた途端にデータが散逸し、５年分ぐらいのデータが全部なくなってしまった。プロテオーム分野では、先例を踏まえ、いまは欧州だと EU が、米国だと NIH が大きなサポートをしている。

（質問）資金をサポートしていても、そこが所有しているのでは独立性がないのでは、なかなか運営が難しい気がする。

（回答）NIH が研究資金を提供している研究者が、NIH がサポートするデータベースにデータ登録してはいけないということではない。要するに、データ産生者が圧力をかけて「いったん公開したデータを引き戻させてくれ」ということが言えないように独立していなければいけないということだ。

（質問）先ほどの成功例、データサイエンティストやデータベースの研究者が「ごっちゃ」になる必要があるとのことであったが、「ごっちゃ」になるためにどんなことが必要か。

（回答）データベースを作った人が、その作ったということだけが成果になるような仕組みになっていてはいけないと思う。逆もそうで、使う方はそこにデータがあれば幾らでも勝手に使うと思うが、使いやすいデータフォーマットにして提供しておけば、後は使う方が勝手に使ってくれるだろうというのが今までの、トーゴーの日で発表されているような形だと思う。それだとなかなか利用につながらない。

（質問）つくる人はつくるのが目的化しているところが結構強い。バイオロジーの観点での議論ができないという意見もある。

（回答）使う側がつくる側とタッグを組めるような受け皿をつくっておく必要があると思う。つくった側も「外部から一切委託は受けません」となっていれば、いつまでたってもインタラクションは進まない。一緒にやりましょうといったときに「はい、やります」という受容体を持っていなければいけないと思う。

（質問）融合の実例で示された研究では、データサイエンス研究者とデータベース研究者の融合になっていて、ウェットな研究者がいない。データベースがあっても、ウェットな研究者には敷居が高くて、どう使っていいのか分からないというのがあると思う。UniProt は、本当に誰でも知識をまとめて得られる。一方、例えばオミックスデータだと、データをどう使っていいか分からない。「nature biotechnology」の論文について、マイニングされてきたデータをどうやって検証したのかなど、どうやって本当にウェットな人が使えるようになるかのお考えをお聞きしたい。

（回答）確かに、「nature biotechnology」論文は、著者のなかにウェット研究者がいない。ただ、あれだけ大規模な解析をされたら、その一部に自分のデータが含まれていても諦めるなと感じた。自分たちではこんなものできないというぐらいの大規模性を持って新しい情報を出してきている。その情報をもとに、例えば UniProt などで、そのリン酸化はこういう機能があって、こういうことをやっているかもしれないと考えていくことができる。そのように個々の研究者がいろいろな研究で役に立てたりすることもある。あるいは、その大規模データをまとめて検証するようなウェットの研究者と組むことも可能だと思う。「Nature Biotechnology」論文は、検証が要らないぐらい多くのアウトプットが出ていることも一つ大きな特徴だ。私は、次の可能性をすごく感じた。ウェットの研究者なしでもここまでデータサイエンティストでアウトカムが出るんだと思い、紹介させていただいた。

（質問）ウェット研究者は、生物学分野で認められている実験手段で検証された情報でないと信用しないのではないか。

（回答）例えば、いろいろなデータが出たときに「これはイン・ビボで証明されているデータです」、「イン・ビトロで証明されているデータです」という。それと同じように「イン・シリコで予想されているデータです」という形で出てくるのであれば、どこまで信用するかは論文を見る人次第だ。イン・ビトロでも、どこまで信用するかはその人次第だ。アンキャラクタライズのものに対して、少なくともイン・シリコの情報を提供するということに大きな意義があると考える。

（質問）エビデンスのレベルを、ラベルとしてデータにつけるということと理解した。

2　ChIP-Atlas によるデータ駆動型研究

竹本　龍也（徳島大学）
沖　真弥（京都大学）

(1)　発表内容

ChIP-Atlas を使って私と沖特定准教授がどのような研究をしているかという具体的な内容をお話しする。

図 2-2 は、JST 事務局からの話の中で我々がどこにいるかを示したものだ。沖氏はドライ系の研究者として ChIP-Atlas を作成している。今回、実験仮説の立案を私とともに行った。私はウェット系の研究者として検証を行っている。

我々は、炎症性腸疾患を 1 つのターゲットとしている。炎症性腸疾患をはじめとした遺伝子系の疾患は、ゲノム上に蓄積した変異によって引き起こされる。これまで、疾患原因となる変異、いわゆる SNP（一塩基多型）を同定するために、ゲノムワイドな解析、GWAS が大規模に行われている。しかしながら、SNP の多くは遺伝子以外のノン・コーディング・リージョンに存在することが多いため、発症までのプロセスがほとんど解明されていない。

また、疾患の候補となる SNP は数多く同定されているが、疾患の直接的な原因となる SNP はほとんど分かっていない。したがって、疾患と関連する SNP のうち、遺伝子本体に対して優先的に研究が行われていて、このノン・コーディング・リージョンに対する SNP は、ほとんど手つかずのままとなっている。

私が沖氏と実施している予備的な研究において、疾患原因となる SNP にはエンハンサーと呼ばれる遺伝子発現調節領域が多く存在していて、なおかつ、そこに様々な転写因子が結合するのではないかということを見いだした。ChIP-seq 解析、いわゆる Chromatin Immunoprecipitation Sequencing を行ったいくつかの論文から導いた仮説だ。ただ、ChIP-seq 解析を扱う論文は数万もあって、それら一つ一つを精査することは不可能だ。これは我々も試みたが、大変な作業だった。以前から交流のあった京都大学の沖氏に相談したところ、沖氏が開発している ChIP-Atlas を使えば、数万件のデータからとても確からしいデータを、情報を得ることができると分かってきた。

図 2-7 に沖氏が開発した ChIP-Atlas の画面を示す。ChIP-Atlas は、転写因子が、どの遺伝子の周りに、どれぐらい結合するのかという疑問を全て解消してくれる。ただ単に 1 つの ChIP-seq データだけではなくて、世界中の研究者が報告した約 14 万件の、1,200 テラバイトにも及ぶ ChIP-seq の実験データを統合したデータベースである。全てのデータを同じ手法で開発しているので、多彩なデータを同一の尺度で俯瞰することができる。さらに、統合したデータから生物学的に重要な仮説を提案してくれるのが ChIP-Atlas の大きな特徴である（図 2-8、2-9）。

先ほどの話に戻るが、我々は疾患候補となっているノンコーディングリージョンに存在する多くの SNP のうち、特に転写因子が結合するものを ChIP-Atlas を用いて抽出した。その結果、炎症性腸疾患と関連する 100 個以上の SNP の中から、このように転写因子が結合するような SNP、特に「転写因子結合ホットスポット」と我々は呼んでいるが、それを 20 個程度まで絞り込むことができた（図 2-10）。

我々が次に行ったのは、この ChIP-Atlas で見いだされた転写因子結合のホットスポットが真に疾患原因の SNP かの検証だ。ここから私のウェット系の研究の話になる。我々は受精卵にゲノム編集因子を導入してマウスをつくっている。従来は、このようにマウスの受精卵一つ一つにガラスのキャピラリーを差し込んで、昔だとトランスジーン、最近だとゲノム編集因子をマウスの受精卵の核に入れることでゲノム編集マウスや遺伝子改変マウス、今回だと疾患の変異を導入するということを行ってきた。

ところが、この方法は非常に時間と技術が必要で、マウス系統をつくるのに大変な時間と費用を要する。私は、図 2-11 の左に示すように、受精卵を電極の間に並べて、ゲノム編集因子で満たした液の中に並べておいて電気をかけるこ

とで簡便かつハイスループットにゲノム編集因子を受精卵に取り込む方法、受精卵エレクトロポレーション法を開発した。この方法を用いると、いろいろな系統、先ほどお話しした 20 系統でも、ある程度の時間をもって作成することができる。この方法を用いて、私は自分がもともとやっていた発生の研究だけではなく、免疫の研究やクロマチンの研究、ほかにも幾つかの系統をこれまでつくってきた（図 2-12）。

疾患候補の SNP が 100 以上となると、もう止めてしまおうと思うものの、この方法を使えば、20 個程度、もしくは 30 個程度であれば、全てゲノム編集、先ほどの方法で変異マウスをつくって実際に大腸炎になるのかを精査できる。

既に多くの系統についてはゲノム編集が作成完了していて、今後はホモマウスの作成を行って解析をする予定である。まだゴールはしていないが、この方法を用いることで仮説に基づいた検証が完了すると考えている（図 2-13）。

以上が研究の内容だが、今日お話ししたデータ駆動型の研究を進めるにあたり、私たちが感じていることを述べる。

今回はドライ系の研究者として沖氏、ウェット系の研究者として竹本が炎症性腸疾患の疾患 SNP を同定すべく研究を行っている（図 2-14）。1 つ目の課題は、データベースに基づいて出された仮説を証明していくにも、残念ながらかなりの費用が必要だということである（図 2-15）。沖氏を含めたドライの研究者の方々の御尽力で公共データベースの統合や、それに基づいた仮説がようやくできるようになってきて、非常にウェットの研究者からするとありがたい。しかし、それでもなお仮説を検証するにはお金がかかる。

我々の場合、数年前に科研費・基盤 B で数理・複雑系や情報・ゲノムとウェットを融合する領域、時限付の特設領域で採択され、研究を進めてきた。それに加えて、徳島大学の共・共拠点の事業、それから私が運営しているベンチャー企業の研究費を投入することで何とか今年までやってきた（図 2-16）。

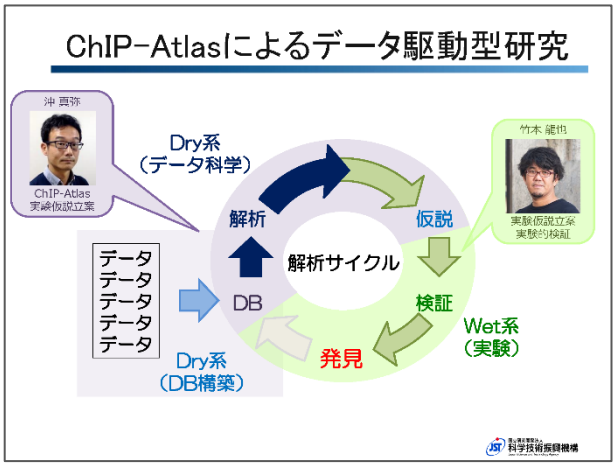
今後、データ型駆動研究をサポートするための政策やグラントが必要なのではないか。そうした、データから仮説、検証までを俯瞰するサポートがあってもよいのではないかと考える。

もう 1 つ、なぜ我々が共同研究を行っているのかというと、実は私と沖氏は大学院生のころから同じ大学の同じ建物で発生生物学を研究し、ともにノンコーディングリージョン――エンハンサーによる遺伝子発現制御の重要性について、昔からその重要性について分かち合ってきたからだ。我々は発生生物学の研究をそれぞれ行うために、私自身はノンコーディングリージョンの編集ということでゲノム編集を、沖氏は ChIP-Atlas の開発を行ってきた。その古くからの信頼があって我々が共同研究を行うことができている。ただ、そうじゃない人たちにとってはこのようなマッチングが難しいのではないかと考える。私も正直、沖氏を知らなければ、ドライの人たちと組む機会があまりないという実感がある（図 2-17、2-18）。

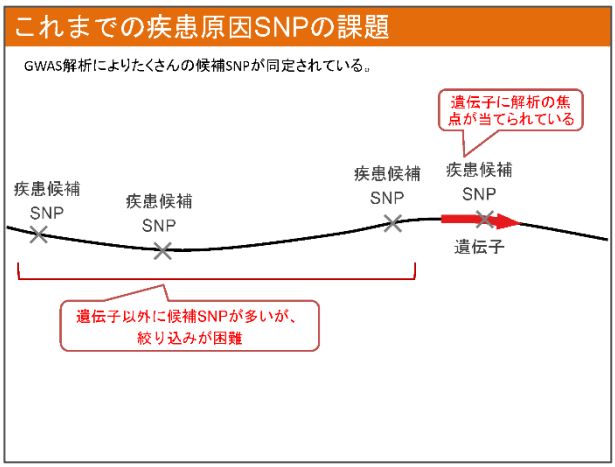
そこで 1 つの提案としては、データベースとドライ、ウェットが双方向的にマッチングする場や交流の機会が必要なのではないかと考える（図 2-19）。データベースから出された仮説を実験的に検証することで新たな発見ができて、それをまたドライに還元できるサイクルが重要だと思う。こうした体制を支援する政策やグラントへの期待を込め、この 2 つについて今日はお話しした。

(2)　資料

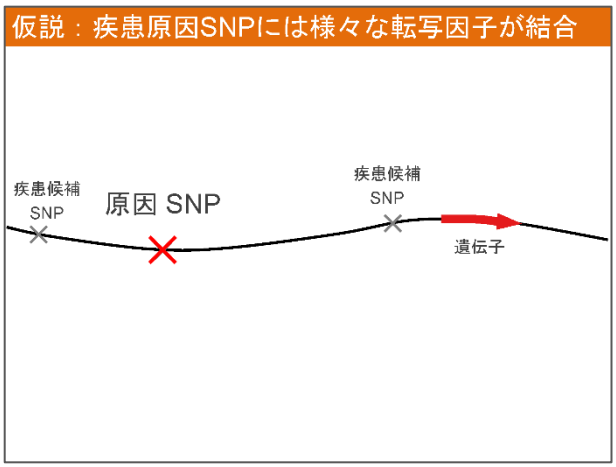
2-1 表紙



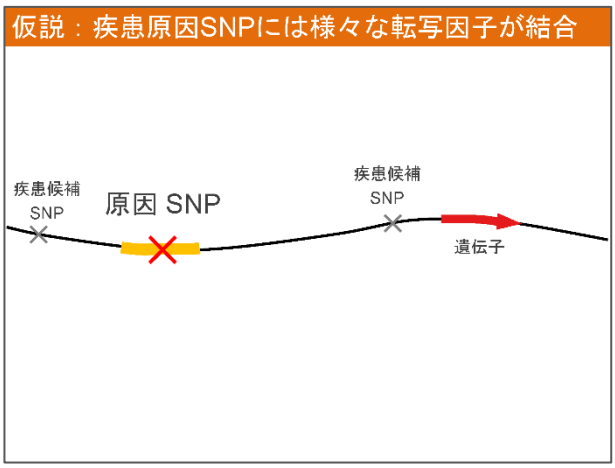
2-2 ChIP-Atlas によるデータ駆動型研究



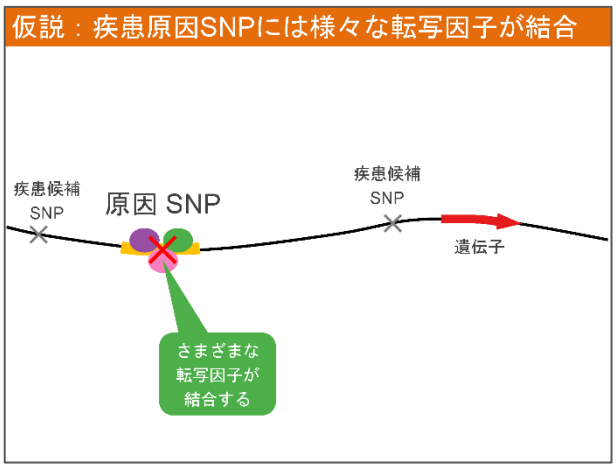
2-3 これまでの疾患原因 SNP の課題



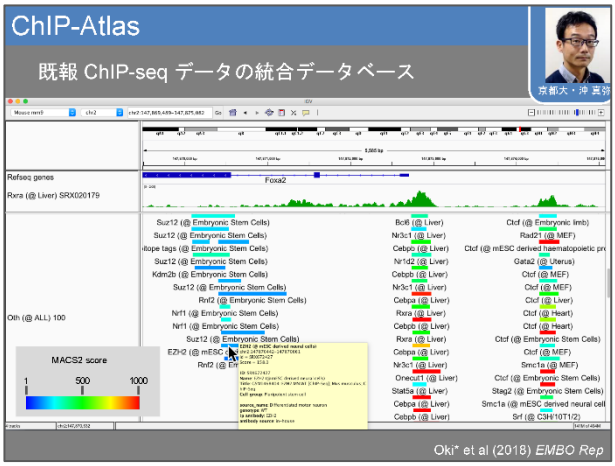
2-4 仮説（１）



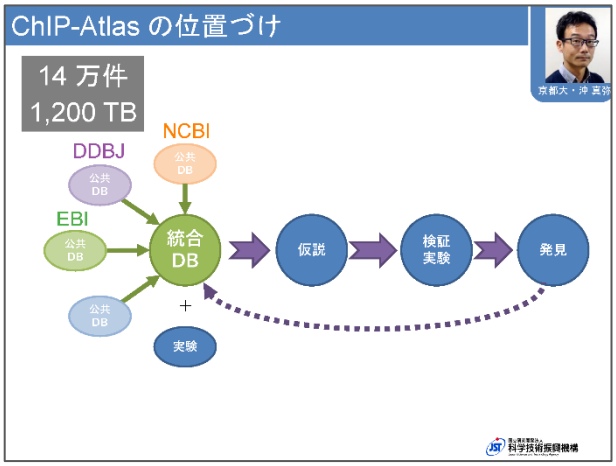
2-5 仮説（２）



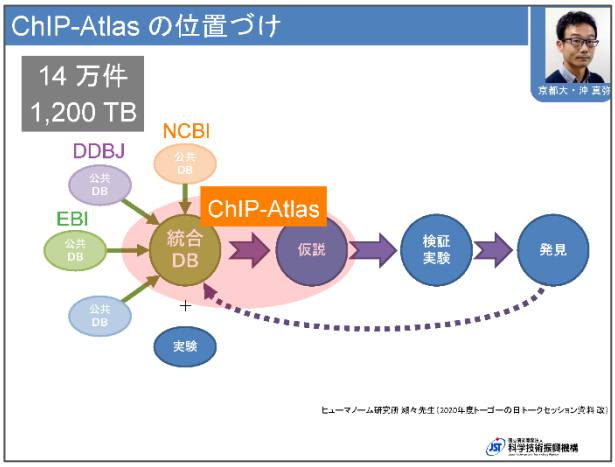
2-6 仮説（３）



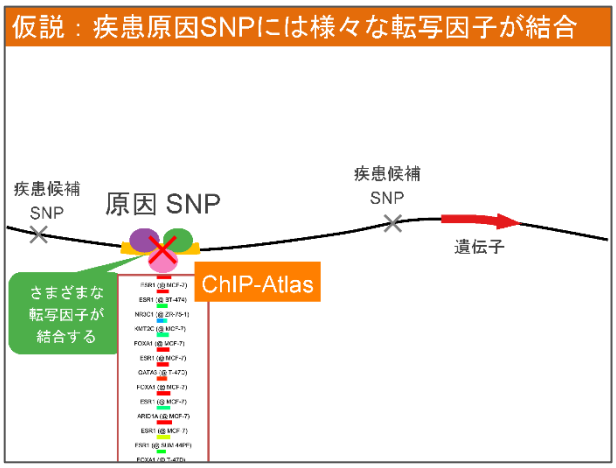
2-7 ChIP-Atlas



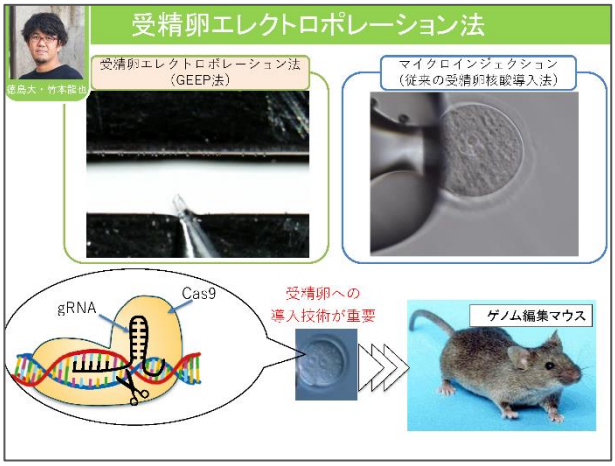
2-8 ChIP-Atlas の位置づけ（1）



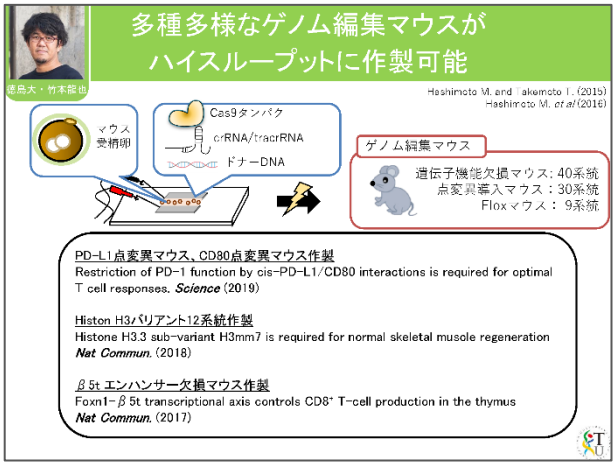
2-9 ChIP-Atlas の位置づけ（2）



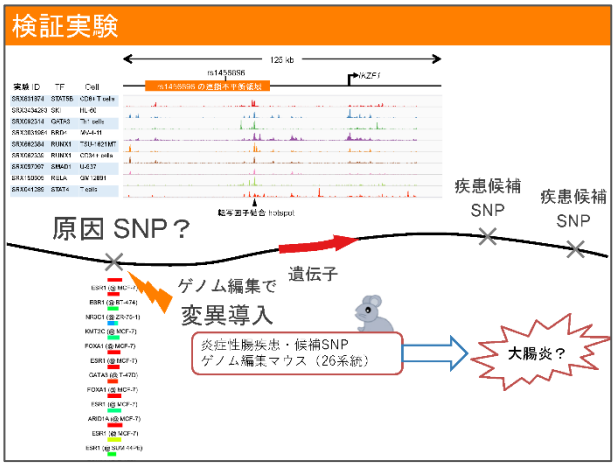
2-10 仮説（4）



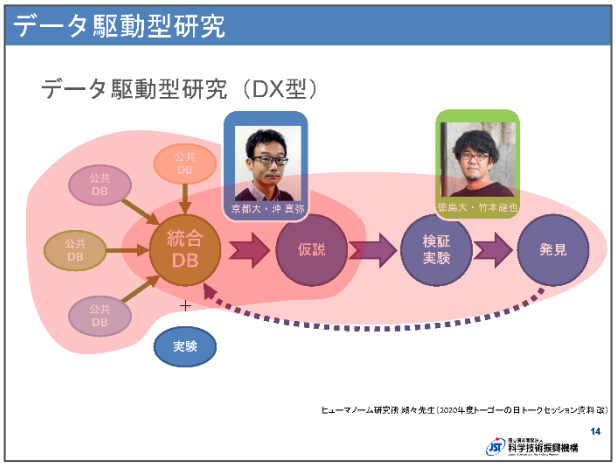
2-11 受精卵エレクトロポレーション法



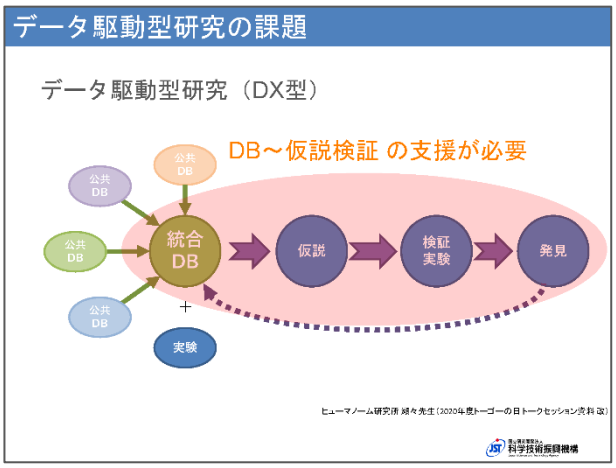
2-12 多種多様なｹﾞﾉﾑ編集ﾏｳｽがﾊｲｽﾙｰﾌﾟｯﾄに作成可能



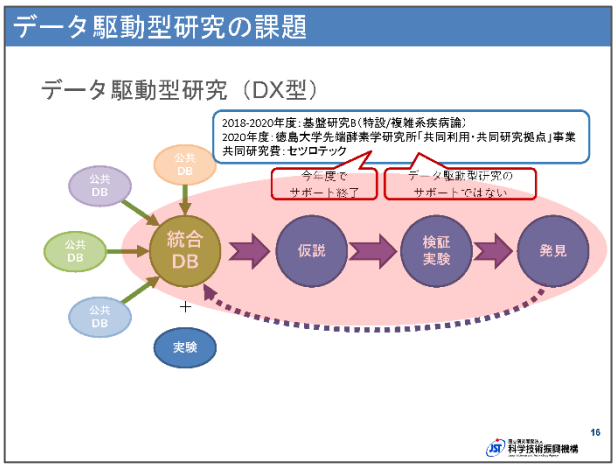
2-13 検証実験



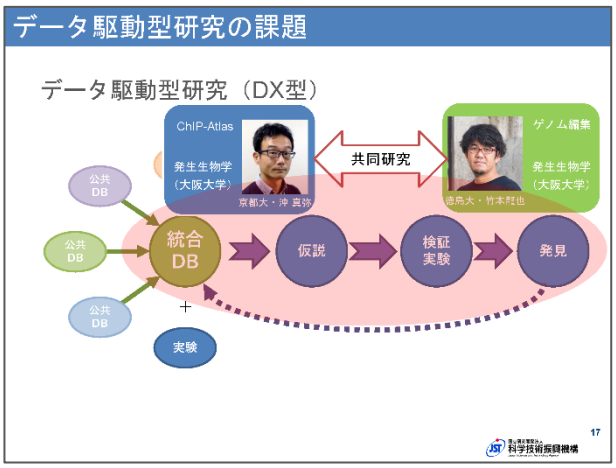
2-14 データ駆動型研究



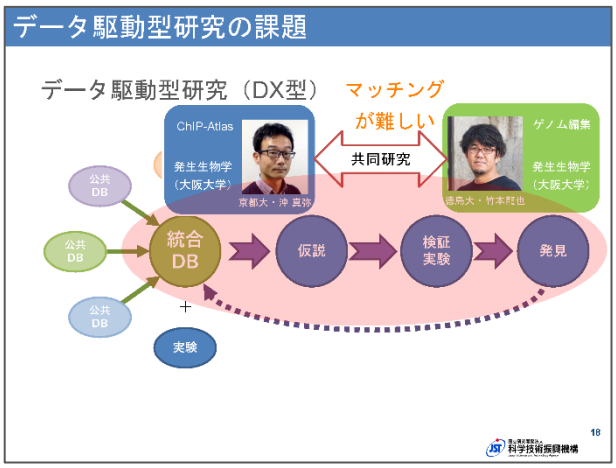
2-15 データ駆動型研究の課題（1）



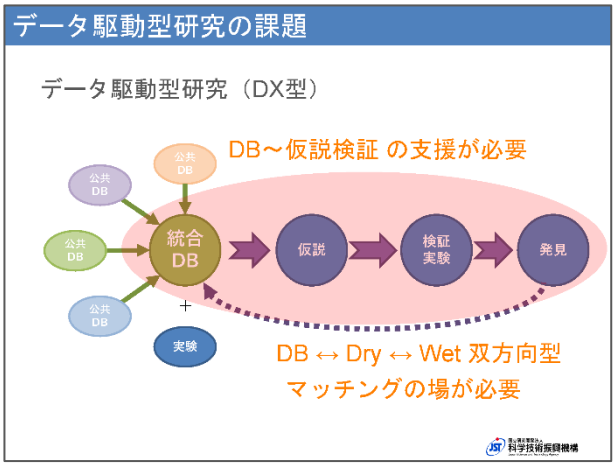
2-16 データ駆動型研究の課題（2）



2-17 データ駆動型研究の課題（3）



2-18 データ駆動型研究の課題（4）



2-19 データ駆動型研究の課題（5）

(3)　質疑応答

（質問）いわゆるヒトの GWAS から絞り込んできた SNP で、特に転写因子結合部位みたいなものに相当するマウスのゲノム領域を編集する場合、単純には「ここかここ」と言いづらいと思うが、目をつけた領域の近傍で、マウスで同様な転写因子が集まって結合しているところを狙うということか。

（回答）はい。ヒトとマウスで保存されていて、なおかつヒトの細胞でも、マウスの細胞でもいろいろな転写因子を結合している場所に絞り込んだ。

（質問）今回、竹本先生が研究を進めるに当たり、沖先生は必要だったか。データベースは公開されているわけで、それでも沖先生と研究者と組まないとうまくいかなかったと思うか。

（回答）私自身、ドライに対する壁を感じており、沖先生がいたからできたというのは正直なところある。ChIP-Atlas は周知活動を積極的にやっているが、どうしてもウェットの研究者からすると、操作するなかで、そもそも使い方は合っているのかなどと不安を感じる。大丈夫だよと背中を押してもらえることがあれば、安心できる。

（質問）第２、第３の竹本先生を生むために、周知活動がまだまだこれから必要ということか。

（回答）その通り。新型コロナウイルス感染症の影響で最近、止まっているが、ぜひ継続していきたい。

（質問）竹本先生がこういった実験をやる上で、ChIP-Atlas によって単純に数が絞られただけではなくて、その裏に、炎症に関係する転写因子が結合するといったことを、実際に in vivo で検証してみたいと思うモチベーションになったものがあるのではないかと思うがどうか。

（回答）ウェット研究者としては、単純に配列情報だけだとか、変異が多いとか、P Value でどうだとかというよりも、転写因子の結合がたくさん、しかもその転写因子が何か疾患に絡んでいそうだなという、においというとあれなのだが、そういうエビデンスが蓄積している様子が見せられると優先的にやりたいと思う。

3　公共データベースからの低酸素発現変動遺伝子のメタ解析

坊農　秀雅（広島大学）

(1)　発表内容

「公共データベースからの低酸素発現変動遺伝子のメタ解析」として、遺伝子発現のデータベースを再利用した研究についてお話しする。

図 3-2 は私自身を低酸素にさらした実例だ。八ヶ岳の頂上、2,899 メートルで、自分を低酸素状態にした。地上の85%ぐらいに酸素が薄くなるだけで大分変になって、この後、高山病のような症状になってしまった。こうした低酸素に我々が置かれたとき、どういうシグナル伝達が動いているだろうか。実際に、がん細胞でも低酸素状態になっている。そういう状況下でのシグナル伝達の経路、仕組みを解明したいということで、ずっと研究されてきた。2019 年のノーベル生理学・医学賞は、低酸素の仕組みに関する研究が受賞理由になったので、記憶にある方も多いのではないかと思う。

2006 年頃から「がんとハイポキシア研究会」に参加し、研究をずっと続けていた（図 3-3）。2007 年には DBCLS に移籍し、データベース統合化事業に関わるようになった。研究会でいろいろな人に話を聞いて、いろいろ学んできたが、低酸素で発現が変異する遺伝子とは何かと聞くと、研究者が言うことが違う。ターゲットにしている遺伝子が遺伝研究者ごとに異なるが、仕方の無いことだ、という状況であった。しかし、それでは困る。データ駆動型研究を通じ、低酸素変動遺伝子を見つけたいというモチベーションで研究を進めてきた。

当時、遺伝子発現解析をするために、マイクロアレイという実験手法が広まっていた。そこで、マイクロアレイのデータを公共データベースから取得し、解析するために、低酸素状態にした時とそうでない時とでペアをつくり、発現が上がった遺伝子、下がった遺伝子を数え上げる、ヒトとマウスの全部の遺伝子に対してやる、ということを行った（図 3-4）。最終的には図 3-4 右に示したように、可視化した。やり方としては、公共データベースから探してきて、数としては少ないが、AOE（遺伝子発現の目次のデータベース）を自分でつくってデータを探した。

実際に共同研究者に使っていただいたところ、いろいろな実験でよく発現が下がっている遺伝子には DNA ダメージ、DNA が損傷を受けたときに応答する遺伝子が多いと分かったという論文に貢献した（図 3-5）。論文は「PLOS ONE」に掲載された（図 3-6）。アブストラクトには、「Comprehensive gene expression and database analyses」などと書かれている。DBCLS からもプレスリリースされ、そのなかで全体のデータ駆動型の研究から発見したと主張した（図 3-7）。

時代が下ると、次世代シークエンサー由来のデータが数多く蓄積されるようになってきた（図 3-8）。遺伝子発現もこの次世代シークエンサーで取得された RNA 配列から RNA-Seq で測れるようになってきた。そこで、この次世代シークエンサーの測定データで同じことをやってみた（図 3-9）。

図 3-10 に示したのは 2019 年に実施した研究だ。AOE を用いてデータを探索し、45PB もの次世代シークエンサーのデータ、そこからデータをかき集めた。「DBCLS SRA」が常に最新の Sequence Read Archive のデータを保持してくれているからできたことだ。次世代シークエンサーのデータから遺伝子発現のデータだけを取得し、定量した。低酸素の状態のものとそうでない普通の酸素の状態のもののペアをつくり、遺伝子発現の変化をいろいろな人の実験からカウントした。

図 3-11 は横軸がヒトで、縦軸がマウスのもの。右のほうに行けばヒトで発現が上がっているし、上のほうに行けばマウスで上がっている。ハイライトした部分の遺伝子は低酸素の専門家も遺伝子発現があると言うものになっている。ということをプロットして、データ駆動型に低酸素で発現が上がる遺伝子を見つけることができた。

次のステップは検証だが、検証するにしても、ヒトとマウスで保存されて発現量が上がる遺伝子以外に、ほかの要素も入れようと検討している。図 3-11 はさっきのマイクロアレイのときと同じような可視化だが、マイクロアレイのときのデータ（図

3-12）よりも NGS のほうがよりクリアに遺伝子が特定できている。

そうして見つかってきた遺伝子発現のリストは、もちろん論文にサプリメントデータとして掲載している。しかし、そのままでは再利用しにくい。また、その論文誌がなくなってしまってはまずい。そこで、公共レポジトリの figshare に掲載している（図 3-13）。塩基配列やタンパク質の配列にはデータベースがあるが、そうではない「軟らかい」データをとにかく入れておく図書館みたいな役割をするサイトがあって、そこにエクセル型のデータ、スプレッドシート型のデータとしてデータをアップロードしている。figshare の利点は、誰でもアクセスできることと、DOI がつくので引用ができることだ。

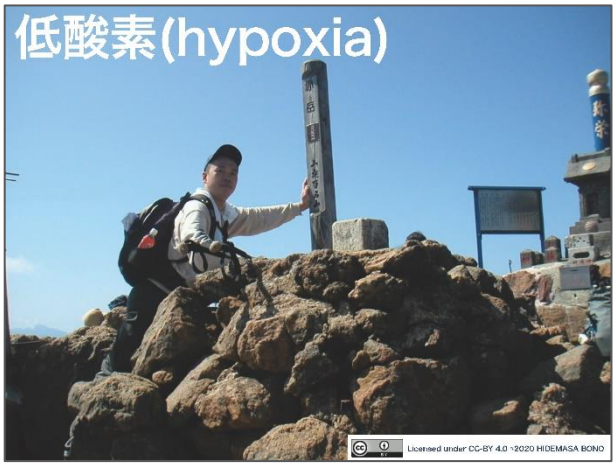
図 3-14 は沖氏の ChIP-Atlas を使った例である。低酸素を支配していると言われているものに非常に関わっていると言われている HIF1A――HIF-1αの転写因子のデータを合わせた。横はヒトで、右に行けば行くほど低酸素刺激をしたときに発現が上がる遺伝子であることを示す。縦軸が ChIP-Atlas から得た ChIP-seq HIF1A の遺伝子のスコアで、縦軸が上に行くほど ChIP-Atlas から得たスコアが高い。この図から、HIF-1α依存的な発現変動をしている遺伝子であることが分かるのに対して、逆にほとんど上がっていないものに関しては HIF-1αに依存しない形の発現変動する遺伝子だと分かる。さらなる絞り込みに使えるのではないかと思い、今、これを見て、ここにあまり知られていない遺伝子を見て精査している。

そうして見つけた候補遺伝子を、ゲノム編集で 1 個ずつノックアウトしたり、逆にノックインしたりして検証しようとしている。ゲノム編集イノベーションセンターの研究者にゲノム編集の実験を手伝ってもらってやろうと、今、新しい研究室を立ち上げているところである（図 3-15）。

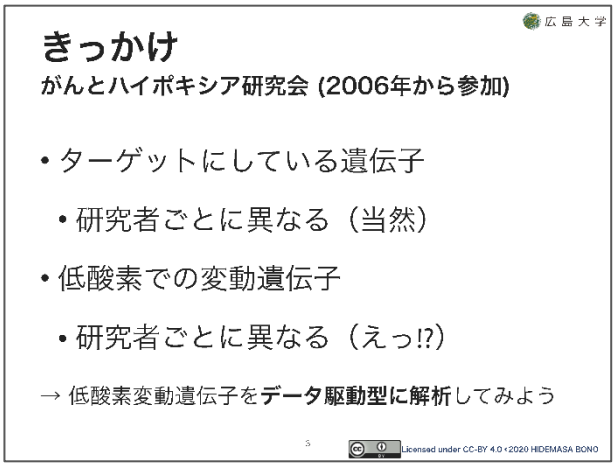
まとめとしては、公共データベースからデータを集めてきて解析するのは、現状、機械的には無理だ。人、特に研究者が見ることが大事になっていると思う。データセットさえ集められれば、あとはコンピュータのプログラムで自動的にできるはずだが、実際にはどういうふうに処理するべきかというパラメータの検討、データの解釈について専門家がやらないといけない（図 3-16）。

(2)　資料

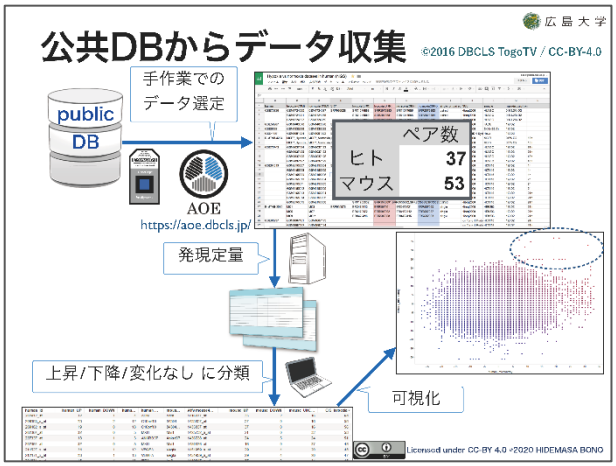
3-1 表紙



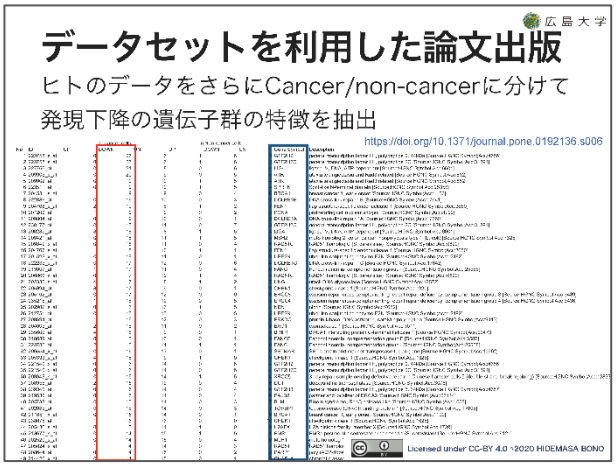
3-2 低酸素(hypoxia)



3-3 きっかけ

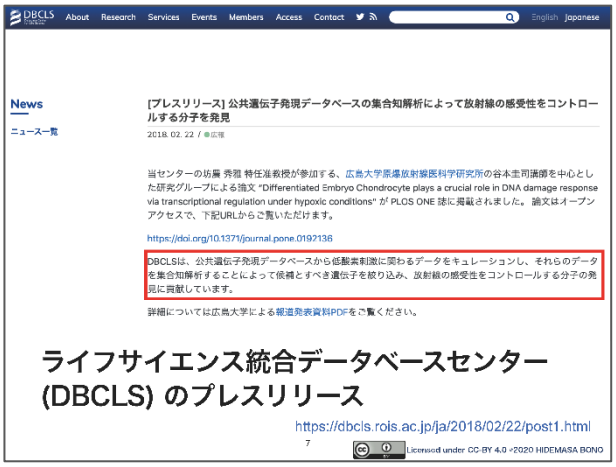


3-4 公共 DB からデータ収集

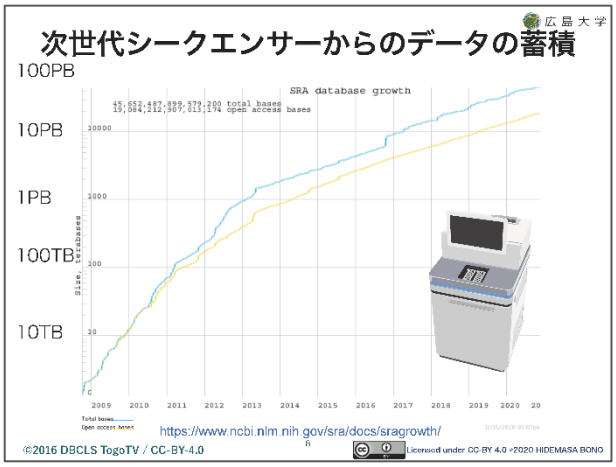


3-5 データセットを利用した論文出版

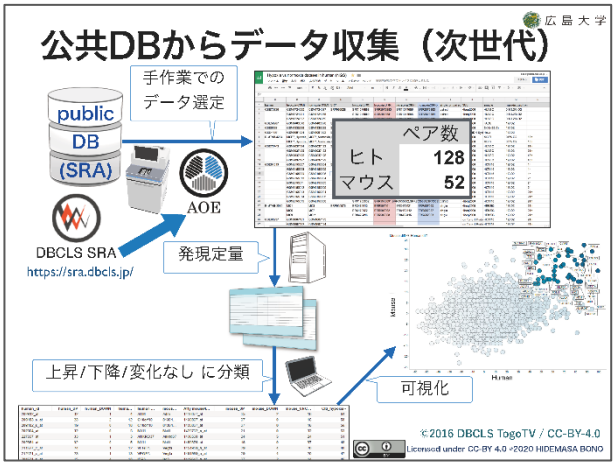
3-6 PLOS (2017)



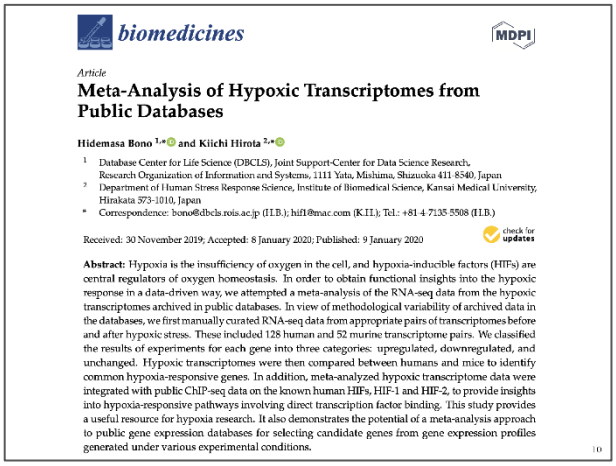
3-7 プレスリリース



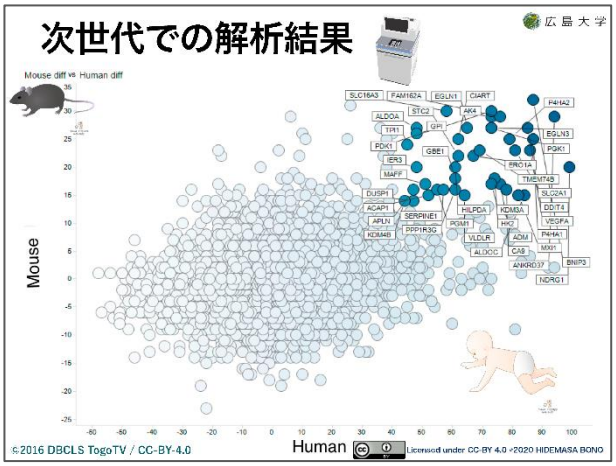
3-8 次世代シーケンサーからのデータの蓄積



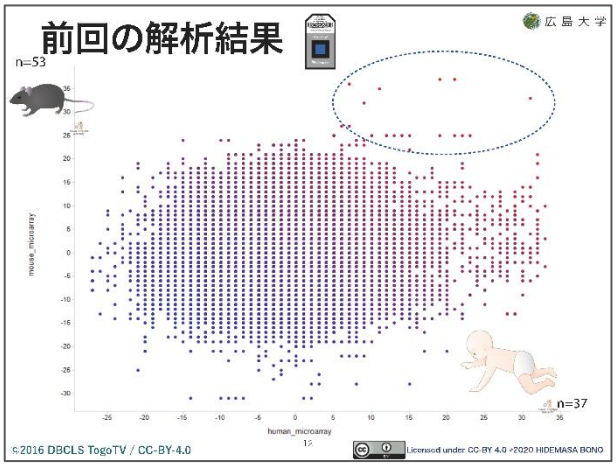
3-9 公共 DB からデータ収集（次世代）



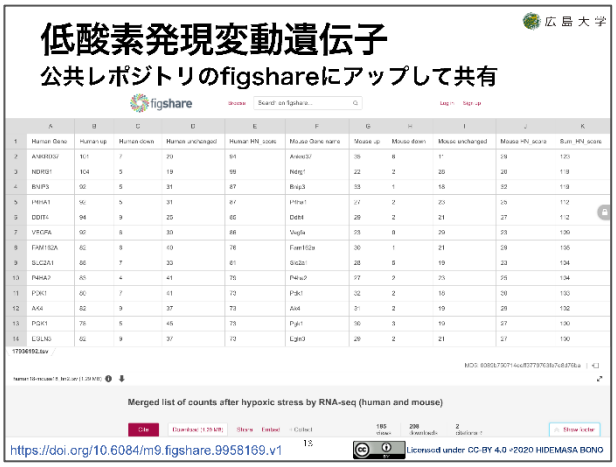
3-10 biomedicines (2020)



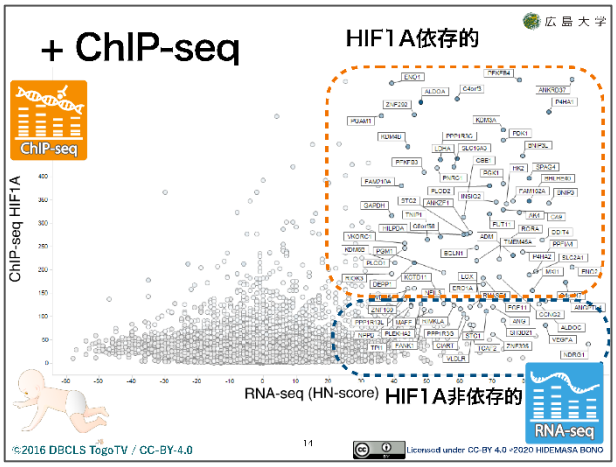
3-11 次世代での解析結果



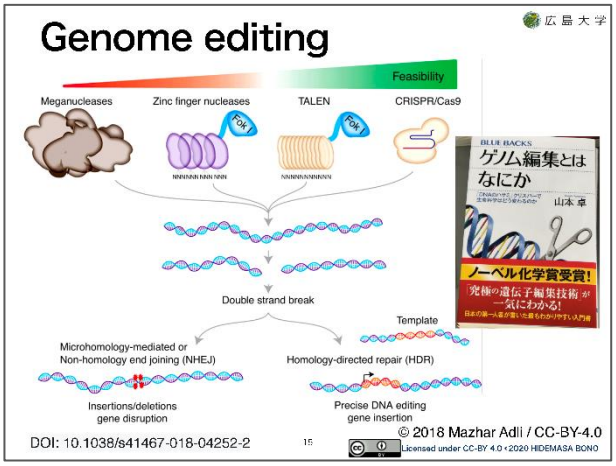
3-12 前回の解析結果



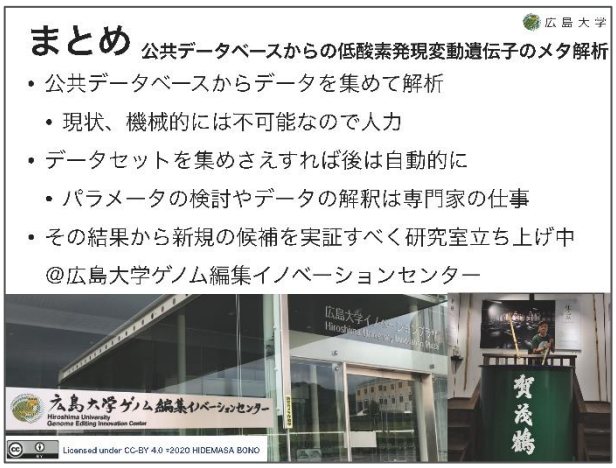
3-13 低酸素発現変動遺伝子



3-14 +ChIP-seq



3-15 Genome editing



3-16 まとめ

(3)　質疑応答

（質問）AOE や DBCLS SRA の立ち上げ・運営上でどのような点に苦労があるか。
（回答）公共データベースを集めてくるのだが、公共データベースは結構変わるということがある。昔からのデータは今のデータと整合性がとれないということがよくある。特に実験データの記述に関しては非常に変わってきている。
（質問）それを一つ一つ御覧になりながら、これは使える、使えないと判断しているということか。
（回答）基本的には、ある程度絞り込みはコンピュータを利用する。しかし、最終的には人間が判断しないと使えるデータにならない。先ほどお見せしたようなペアのデータ、低酸素とそうじゃないもののペアのデータは一つずつ自分の手でつくらざるを得ないというのが現状だ。

（質問）そのように苦労され、公共データを再利用しやすくしている AOE について、現在の利用状況をどのように捉えておられるか。
（回答）まだ、あまり知られていない。サービスそのものの周知活動もすべきかも知れない。ただ、自分自身で活用事例をつくることで広めることがより大事だと思っており、いままで活動してきた。

（質問）こういった解析をしたときに期待していることといえば、今まで言われている HIF1 パスウェイ以外の新しいパスウェイが出てくるかどうかだと思うが、そうした観点でなにか現時点でご発言いただけることはあるか。
（回答）出てくると思う。それは今、遺伝子としてセットとしてやっているのはタンパク質コード遺伝子だけだが、ノンコーディング遺伝子に関しても同じ計算が RNA-Seq に関してはできる。実際、非常に同じような挙動をしているノンコーディング RNA が数多く取れてきている。関与しているノンコーディング RNA がほぼ間違いなくあるだろうと考えている。
（質問）先ほどの竹本先生の解析とも重なるが、特にノンコーディングリージョンのところでは、どういった転写因子が重なってくるかといったことも非常に有用な情報になってくると感じた。
（回答）そのとおりだ。ChIP-Atlas は非常にありがたく使わせていただいている。

4　DB 基盤整備の重要性

鎌田　真由美（京都大学）

(1)　発表内容

私からは、このサイクル（図 4-2）の中の赤破線で囲んだ部分に関してお話したい。主にはデータベースの構築についてだが、そこからの解析や、今は医学研究科に所属し、こういうウェット研究者と一緒に研究しているので、この部分もお話しする。データベースをつくっていく、データを整備していく中で感じることもあったので、その辺りも述べたい。

まず、データ基盤構築の重要性に関し、データベースをつくってきた体験を通してお話しする。これまで、我々、京都大学では MGeND というデータベースをつくってきた。日本人の疾患ゲノムの情報を集積して、臨床現場においてゲノムの臨床的な解釈、キュレーションを行うために役立てていただく、そういったデータベースとしてつくってきたものだ（図 4-3）。

MGeND は AMED の臨床ゲノム情報統合データベース整備事業の一環で開発してきた（図 4-4）。本事業は、図 4-4 に示す疾患領域において、各疾患領域のエキスパートが採択されている。その先生方が患者のリクルートやゲノムの解析、そして各バリアントに対する疾患関連性の判断などを行った上で、公開可能なものを未制限公開の統合データベースに提供いただき、公開する。我々は、この統合データベースのシステム開発の担当として参画した。

本事業は 2016 年 12 月に始まった。まず、このデータベースをつくるに当たり、各先生方のところに行き、様々な疾患領域における研究の目的や、どういう研究を進められているのか、また、具体的にどういうデータを取るのか、どういうデータだと提供いただけるのか、またそれはどういう形なのかを細かく聞いた。また、疾患領域においてやりたいことや見たいものは様々なので、データベースのユースケースがあるのかを知る意味で、どういうふうに活用したいのかを聞いた。こういうインタビューを介してデータのモデル及びデータフォーマット、登録の形式などを策定し、2017 年からデータを集約し、2018 年にデータベースを公開した（図 4-5）。

データ登録の作業は、実験をされている先生方も特に感じられるところかと思うが、かなり手間がかかる大変な作業だが、これまでに各拠点の先生方に御協力をいただいて多くのバリアントを提供いただいている。現在、粛々と公開作業を進めているところだ。

疾患関連バリアントのデータベースとしては ClinVar が世界的に有名だと思う。この ClinVar に収載されているバリアントのデータと今回構築したデータベースに収載されているバリアントのデータを比較した結果が図 4-6 だ。MGeND のデータが３月で少し古いが、比べてみると、ClinVar には収載されていないバリアント情報が MGeND には半数以上あることが見ていただけると思う。

図 4-6 左側のプロットは、横軸が ToMMo から取ってきた日本人におけるアレル頻度、縦軸が gnomAD から取得したアジア人以外のアレル頻度だ。グレーでプロットしているドットは ClinVar に収載されているバリアントで、色がついているものが MGeND に収載されているバリアントだ。御覧いただいて分かるように、両集団において頻度が高いものも多いが、アジア人以外の集団では頻度が低いようなバリアントも多くある。

また、図 4-6 右下側は、それぞれのバリアントに付与される臨床的意義を示している。ClinVar でその臨床的意義が明らかになっていないものであったとしても、MGeND には病原性なし、もしくはありとして登録されているものがある。また、病原性が疑われるというところで止まっているものに関しても Pathogenic というラベルがついているようなバリアントが多く見受けられる。こういった意味でも、日本人でのデータを集めるデータベースをつくってきた意義があると感じている。

MGeND は臨床で使っていただけるものをということでつくってきた。活用事例としてはゲノムの解析の後のキュレーションがメインになる。私自身は、「富岳」を使った大規模シミュレーションへの活用を進めている（図 4-7）。

このプログラムは、ゲノムで臨床的意義が分かっていないもの、また、有効な治療法が分かっていないものに対し、分子動力学計算を用いて分子への影響及び薬剤とのインタラクションへの影響をシミュレーションして予測するものだ。「富岳」を使うことで大規模に（ここには 1,000 種と書いたが、現在より多くのバリアントをシミュレーションできないかと検討している）機序の分からないバリアントに対する推定をすることで、メカニズムの解明につながるカタログをつくれないかと思っている。

何でも計算対象にできるわけではない。計算に回すバリアントを選別する工程で MGeND に登録されているアノテーション情報などを基に、併せて既存のデータベースの情報を統合し、優先度の基準をデータから策定することを進めている。また併せて、先ほども見ていただいたように、MGeND のみで病原性が明らかになっているものもあるので、それらを学習データとして使った疾患管理予測の機械学習モデルの構築も、このプロジェクトの中で進めている。まだ具体的な成果はお見せできないが、このようにして活用していきたい。

これらの経験を踏まえた上で、それぞれデータベースの構築の部分と解析の部分での課題について感じたところを述べたい（図 4-8）。データベースの構築に関し、「うまくいった」と書いた。MGeND を、臨床的意義の疾患管理バリアントのデータベースとして構築することはできたと思っており、実際、評価もしていただいている。

なぜうまくいったのかについて要因を考えてみたときに、今回、AMED 事業でデータ提供が義務づけられたことがある。皆さんにしっかりと提供いただいている。図 4-2 のサイクルの中の発見する方々が、提供者に含まれているため、データベースの活用のイメージが共有できているところが、非常に協力的にデータを出していただけた要因だと思う。

インタビューしながら関係構築をしたところも大きかったなと思う。実務者の方と会話をすることで、具体的にどういうデータベースにすると活用していただけるのかというユースケースの蓄積にもなったので、具体的な開発項目も非常にクリアであった。

また、それぞれの拠点ではゲノム解析を実施されていることもあり、データエンジニアリングできる人材がいらっしゃったことも大きかった。そういった人材がいない拠点に関しても、できる限りサポートして進めたことでデータの集積ができた。

これらを踏まえた上で感じたこととしては、データベースに対する理解不足がある。「とりあえず集めれば何とかなる」というような、ふわっとした活用イメージでは、なかなかこういう形でデータベースを構築する基盤をつくるところは難しいと感じた。

次に活用の部分における課題について、自身もこのデータを集めてきて解析して感じたことも踏まえてお話したい（図 4-9）。データ活用の一歩であるデータベースから解析の部分、この矢印の部分を加速するには何が必要なのかと考えたとき、これまでにデータを統合する技術やオントロジーといったものは整備されているが、実際に自分が解析をしようと思ったときに欲しいレベルでのオントロジーが整備されていないという点がある。そのため、活用がしにくい。また、データ解析の目的に応じたデータが統合されていたとしても、そこから目的に応じてデータを抜き出してくる、そして成形する ETL 処理（Extract/Transform/Load 処理）に手間がかかると常々感じる。これらの処理は、目的・ストーリーが違えば、その研究・タスクごとにつくる必要がある。

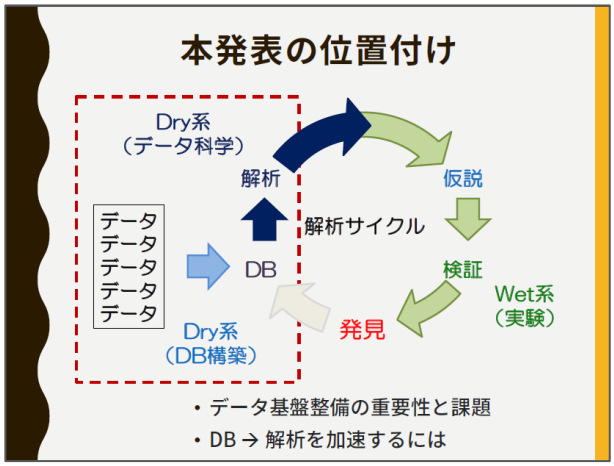
サイクルを推進するために何が必要だろうか。図 4-10 左上に示したのは、一般社団法人データサイエンティスト協会による、データ分析人材に必要なスキルだ。データを活用してデータサイエンスとして生かしていくとき、Data Science とドメイン知識、データサイエンスに使える形にする Data engineering が必要とされている。今回のサイクル（図 4-9 右上）で考えると、Data engineering は「基盤」をつくる、そして解析に持っていくまでのところを整備するスキルだ。Data Science は「活用」するところ、そして、Business Problem Solving はウェットのドメイン知識だ。サイクルを回すためには、石濱先生もおっしゃっていたように、データサイエンスをする人と基盤をつくる人との協働が必要だと感じる。また、各領域における課題、すなわち、どう使いたいのか、どうデータを活用したいのかを共有することで、どんなデータが必要で、どういう抽出をすべきなのかということが具体化されてくる。その事例を積み重ねていくことで汎用的なデータ抽出、解析までのフローもでき

る。それを活用インフラとして提供していくことで、サイクルを加速できるのではないか。

最後、これも常々感じているところで、データサイエンティストの育成ということはよく謳われる。私も実際、本学においてそういうプログラムをやっている。しかし、データを使える形にするまでのデータエンジニアリングの重要性も周知していく必要があると感じている。重要性を周知するにあたって、エンジニアリング部分だけでの成果は発信しづらい。協働する中で事例を成果として発信していくことが重要だ。

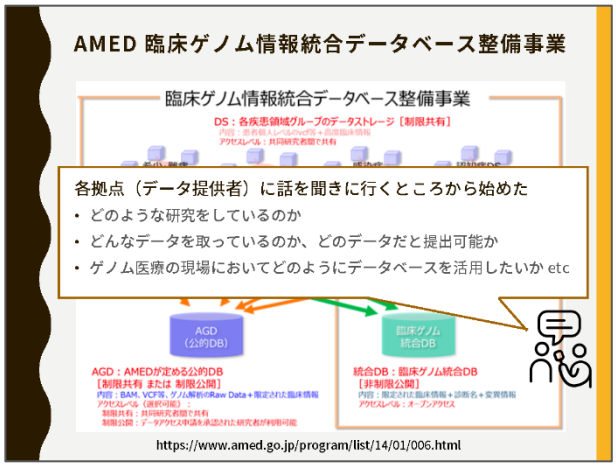
(2)　資料

4-1 表紙

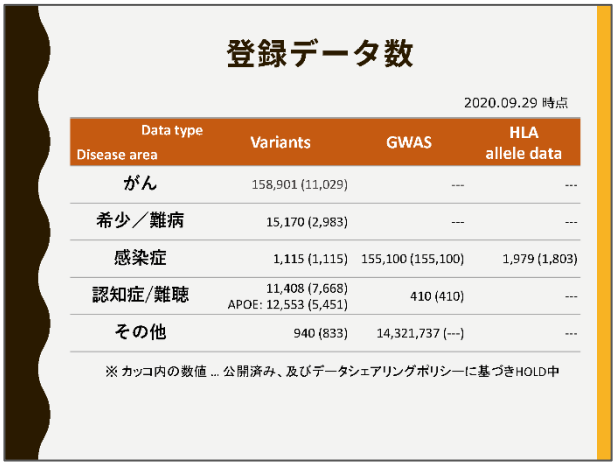


4-2 本発表の位置づけ

4-3 データの基盤整備の重要性と課題

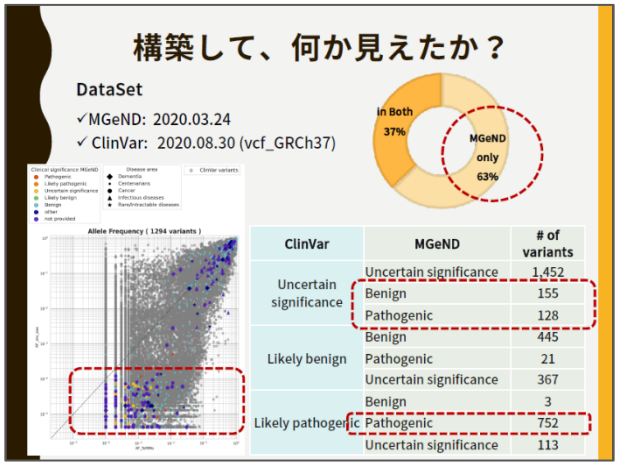


4-4 AMED 臨床ゲノム情報統合データベース整備事業



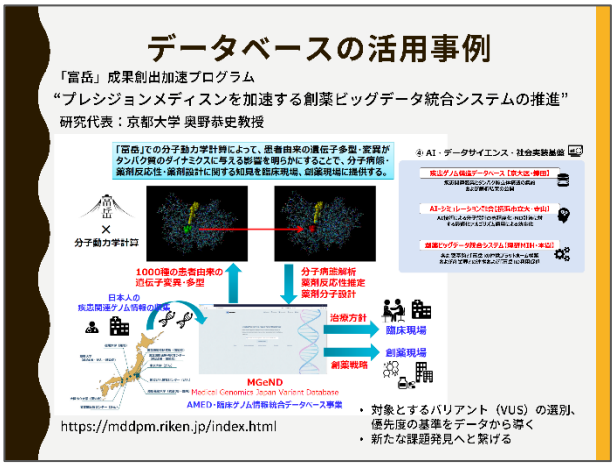
Data type / Disease area	Variants	GWAS	HLA allele data
がん	158,901 (11,029)	---	---
希少／難病	15,170 (2,983)	---	---
感染症	1,115 (1,115)	155,100 (155,100)	1,979 (1,803)
認知症/難聴	11,408 (7,668) APOE: 12,553 (5,451)	410 (410)	---
その他	940 (833)	14,321,737 (---)	---

4-5 登録データ数

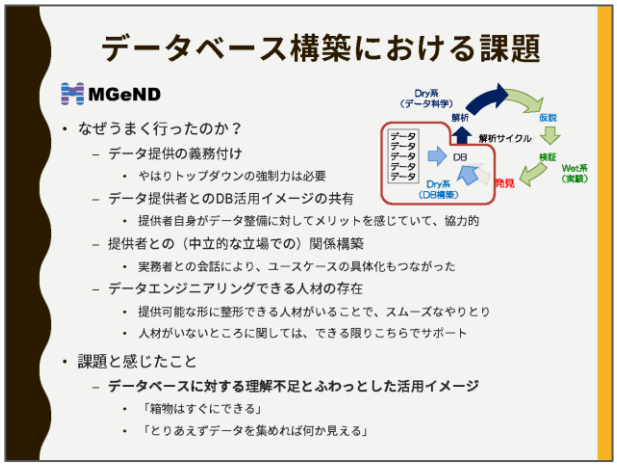


ClinVar	MGeND	# of variants
Uncertain significance	Uncertain significance	1,452
	Benign	155
	Pathogenic	128
Likely benign	Benign	445
	Pathogenic	21
	Uncertain significance	367
Likely pathogenic	Benign	3
	Pathogenic	752
	Uncertain significance	113

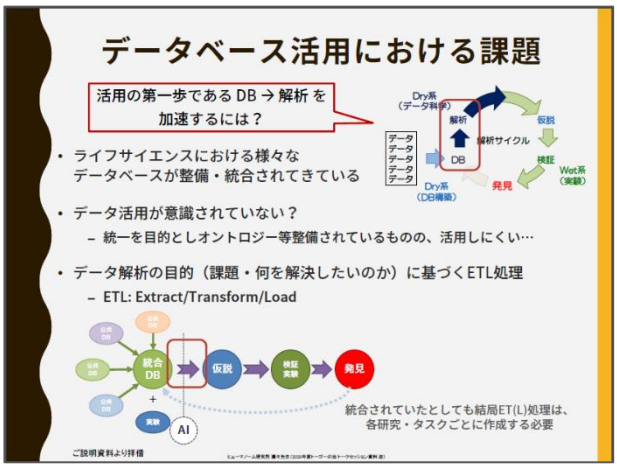
4-6 構築して、何か見えたか？



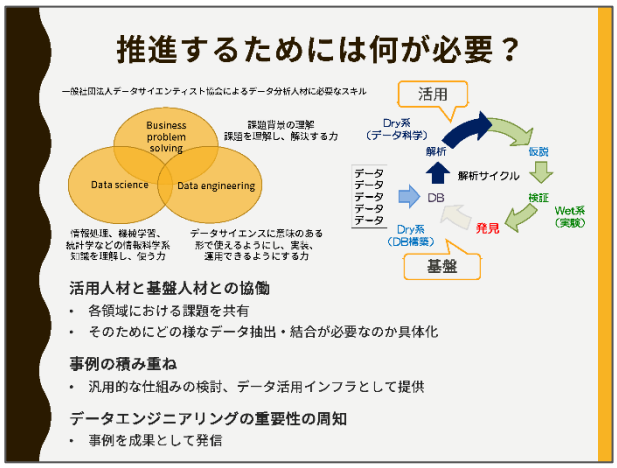
4-7 データベースの活用事例



4-8 データベース構築における課題



4-9 データベース活用における課題



4-10 推進するためには何が必要？

(3) 質疑応答

（質問）キュレーションに関しては、意見が分かれているときは基本的に全部載せる方針ということであった。時がたつといろいろと解釈が統一されていく場合があるが、どのぐらいの頻度で見直しをするのか。

（回答）提供いただいたデータは、そのまま載せている。提供者から更新の依頼があれば、データの更新ということで対応としている。

（質問）提供者の側から更新依頼がなければ、そのままずっと残るということか。

（回答）その通りだ。データを見るなかで、そこは課題だと感じる。

（質問）MGeND を利用し、「富岳」でバリアントと化合物との関係を研究しているとのことであった。実際に MGeND の中のバリアントでコーディング・リージョンに落ちているものはどのぐらいの数で、どのぐらいの頻度か。

（回答）コーディング・リージョンのほうが、現状は多い。アミノ酸が変わってしまうようなミスセンス変異であっても、今は収録対象の疾患が絞られているが、現状収載されているなかで数千以上のバリアントの候補は出ている。

（質問）データ提供者から「このデータをウィズドローしたいのですが」との要望を受けた場合、ウィズドローするか。既に公

開されているものでも、後からウィズドローは可能か。

（回答）先ほどのデータ更新と同様に、変更がある場合には全てのデータをまとめて提供いただき、置き換えている。ウィズドローしたい場合、抜いて提供いただければ、抜いた状態で公開される。

（質問）論文投稿の際、生データの公開要求を受ける。そこに公共データベースとしての役割があるとすると、論文公開後に、根拠となる情報を差し替えてしまうとかウィズドローしてしまうのは、公共性という意味では非常によくないように思う。何か整備されていたり、もしくはその予定があったりするか。

（回答）ご主張は、もっともと思う。現状、MGeND でも、論文に記載できるような ID の発行について、議論をずっとしている。現在はサブミッションいただいた段階で ID を発行し、更新があればそれにサブバージョンをつける形にしている。ウィズドローされた場合のデータのサブバージョンで区別が付くには付く。しかし、それをどのようにして公開するかというところは、現状、議論中だ。

（コメント）この論点は重要だと思う。ウィズドローを含めて version1.1 などにする必要がある。

（コメント）規模の大きな研究会開発プロジェクトで、その成果としてデータベースを公開されることがあるが、「公共性」を担保できない。同じ資金提供元であっても独立したプロジェクトでデータベースが作製されるべきだと思う。

（コメント）大変重要なご指摘を誠にありがとうございます。きちんと担保される形で運用、情報提供できるよう、議論を進めていく。

（質問）データエンジニアリングをするエンジニアの養成も大切だというお話だった。どういうスキルの人がどういうキャリアパスの可能性があるかといったことを、どのように考えていくのがよいと思うか。

（回答）難しい問題だと思う。要件の定義は難しくて、今、データベースの専門家がいないのではないかと感じる。企業には、データウェアハウスの整備などにおいて、そうしたエンジニアリングをやっている方々がいると思う。ライフサイエンスデータベースにおいて、研究者として従事されている方がなかなかいないのが現状だろうが、人材を配置する必要があるのではないか。ただ、解析に持っていくために、分野によってはドメイン知識が必要で、そう簡単にはいかないはずだ。答えが見つからない。

（質問）データサイエンティストとデータエンジニアの違いが分かっていないのだが、データエンジニアは 1 次データを加工して使いやすくする人たちのことか。

（回答）今回の話において、データサイエンティストはデータを情報処理したり、統計学を情報学的な手法を使ったりすることで利用する人で、エンジニアは、データを利用できる状態にする人だ。

（質問）ウェット研究者の方とコミュニケーションをとっていく上で苦労されていること、こういうことをしたらいいのではないかと感じることはあるか。

（回答）解析の背景や手法の重要性をいかに理解するかが重要だ。必ず、ウェット研究者の研究を理解しなくてはいけない。解析結果を検討する際に、そうした視点をいかに議論に含めていくかをいつも気にかけている。そのためには、会話の量が重要だと感じる。

（質問）私の経験では、インフォマティクスを、打ち出の小槌か魔法の玉手箱みたいな感じに思っているウェット研究者がいた。

（回答）正直に申し上げれば、そうした研究者にお会いすることもある。そうした意味でも、先ほど述べた、背景の共有の重要性を主張するようにしている。

5　データ駆動型研究が拓く創薬と医療

山西　芳裕（九州工業大学）

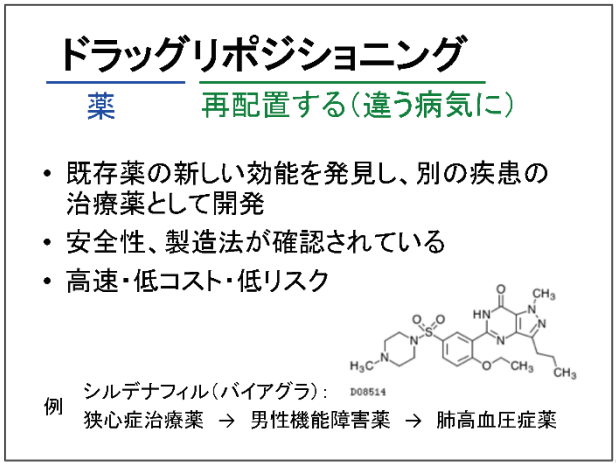
(1)　発表内容

「データ駆動型研究が拓く創薬と医療」というタイトルで発表する。

図 5-2 は、私のデータ駆動型研究における位置づけだ。既存のデータベースに入っている多種多様なデータを解析して、予測して、仮説を立てる。この辺りが私の範疇になる。検証は実験系の研究者と共同研究して、このサイクルを回している。

創薬研究は、年々、医薬品開発が難しくなっている。お金も時間もかかるし、ほとんどが失敗に終わってしまう（図 5-3）。

そこで最近、ドラッグリポジショニングという考え方が注目されている（図



5-4）。既存薬の新しい効能を発見して、本来とは別の疾患の薬として開発するというアプローチだ。既存薬なので安全性が担保されているので、高速で低コストで低リスクの創薬が可能になる。有名な例だと、商品名のバイアグラで知られるシルデナフィルという薬は、当初は狭心症の薬として開発されていたが、現在では男性機能障害や肺高血圧症の薬として使われている。

私は、このようなドラッグリポジショニングの問題、薬と疾患の関係性、これを自動的に予測するような機械学習の手法を開発している（図 5-5）。数多くの薬があり、疾患もまた数多くある。狙いは、この薬はこの疾患に効くことは分かっているが、こちらの疾患にも効くかもしれないといったことをコンピュータ上で一網打尽に予測することだ。これによって偶然の発見から脱却したい。

解析するデータだが、私は公共データベースをフル活用している（図 5-6）。薬に関するデータは DrugBank や、KEGG DRUG、SIDER（図 5-7）、低分子化合物に関するデータは PubChem や ChEMBL、BindingDB（図 5-8）、遺伝子・タンパク質に関するデータは UniProt や、KEGG PATHWAY、Reactome（図 5-9）、疾患に関するデータは OMIM や GWAS catalog、GEO（図 5-10）など、様々なデータベースを駆使している。このような多種多様なデータを統合解析して、薬とタンパク質と疾患の 3 種間のネットワークを予測するような情報技術を開発している（図 5-11）。つまり、どの薬がどのタンパク質を標的として、どの疾患に効くのか、これを予測することが目的になる。薬の空間があって、タンパク質の空間があって、疾患の空間があるわけだが、異なる空間の間にどういうつながりがあるのかという、この未知の関係性を情報的に予測している（図 5-12）。今日は薬とタンパク質の相互作用の解析について紹介したい。

図 5-13 には、ゲノムワイドに薬とタンパク質の間の相互作用を予測するような機械学習アルゴリズムの開発を示した。

この薬とタンパク質の相互作用の予測問題を情報学的に捉えると、薬とタンパク質のペア、化合物とタンパク質のペアがあるとしたら、それが相互作用するペアなのか、そうではないのかに分類する問題と考えることができ、それを実現するような機械学習の手法、モデルの開発などをしている。入力情報は化合物の情報、タンパク質の情報になる。どういうデータを入力とするのか次第で予測結果は結構変わる。例えば、化合物の化学構造もしくは薬が人体に与えるフェノタイプを入力とする解析や、薬や化合物を細胞に限ったときの遺伝子発現パターンの解析など、いろいろな解析をしている。

図 5-14 には 8,000 個の日本やアメリカの薬、これを全て解析した事例を示した。これらが約 1,300 種類の疾患のどれに効きそうかという大規模な予測を行ってきたが、この図はその一部だ。例えば、左上に Pioglitazone という薬がある。これは、本来、糖尿病の薬だが、MAOB というタンパク質に相互作用することが予測されたので、パーキンソン病に効くと考えられるケースになる。MAOB はドーパミンを分解する酵素なので、これを抑えるとドーパミンが増えて治療につながるだろうと考えられる。実際、最近のコホート研究でも、Pioglitazone を飲んでいる人はパーキンソン病の発症率が有意に低いことが報告されていて、これは一つのエビデンスになるかと思う。

今、グラフ上で説明したのは化学構造に基づく予測であった。機械学習モデルの学習データに入っている化合物で構造が似ている化合物がこれだが、結構似ているので、こういった構造の類似性を学習していると考えることができる。これは人間でも予想はできそうだ。

次に薬物応答の遺伝子発現データに基づく予測の例を示す。精神病のお薬でフェノチアジンという薬がある（図 5-16）。これはアンドロゲンの受容体である AR に対する相互作用が予測されたので、前立腺がんにも効くのではないかと考えられる。学習データに入っている化合物の中で一番遺伝子発現パターンが似ていたのはこの化合物だったが、結構構造が違うことが分かる。つまり、化学構造に依存しないような予測ができるのが特徴だ。本当にそうなのかを確認するために実験検証をしてもらったところ、予測した AR に対する阻害効果を確認することができた（図 5-17）。IC50 が 3.6μM ぐらいのレベルで阻害作用が確認できている。

また、もともとは抗がん剤ではないけれども、隠れた抗がん作用を持つような安全な薬の発見ができないかということにも取り組んでいる。図 5-18 に示したものは東大の谷先生との共同研究だが、がん関連パスウェイを制御するような薬の探索などを行っている。ここでは LINCS や、CMap、TG-GATEs など薬物応答の遺伝子発現データを収集し、統合解析をして、その薬に応答してどういうパスウェイに作用していそうかを明らかにして、がんにつなげる解析を行っている。

実際に抗がん剤以外の薬から抗がん作用が予測されたものトップ 10 個を実験検証してもらったところ、半分ぐらいで抗がん作用を確認することができた（図 5-19）。例えば、精神病の薬で Penfluridol、神経遮断薬で Promazine などを同定できている。図右上の一番下の細胞が正常細胞だが、がん細胞では減少したり死んだりする傾向があって、正常細胞は毒性がないという特性を確認できている。

このような考え方は漢方薬にも適用することができる。例えば、腹痛の漢方の大建中湯の大腸がんの一種への効能、またその作用機序を予測して、マウス実験で検証した（図 5-20）。富山大学の門脇先生との共同研究だ。この写真で、このピンク色がかっているのが腫瘍だが、縮小している様子が分かるかと思う。

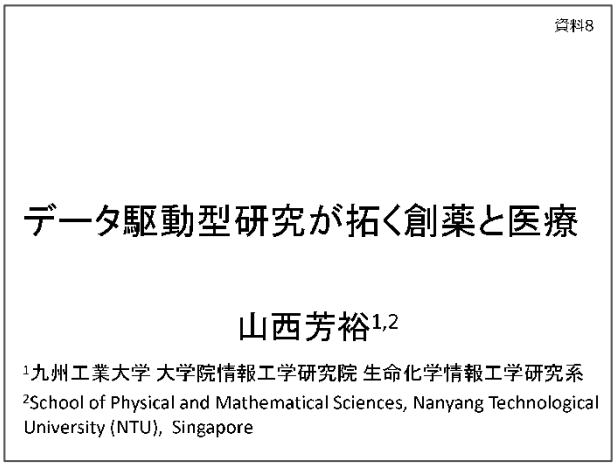
また、データ駆動で予測した薬を臨床試験で検証しようということも共同研究している（図 5-21）。

沖先生が開発されている ChIP-Atlas（図 5-22）をフル活用し、大規模な ChIP-seq データを用いた創薬研究、再生医療の研究も行っている。ChIP-seq のデータだけではなくて、多階層のオミックスデータと統合解析することによって、細胞を直接変換するダイレクトリプログラミングに必要な転写因子のセット、低分子化合物のセットを予測するような手法の開発などを行っている（図 5-23）。

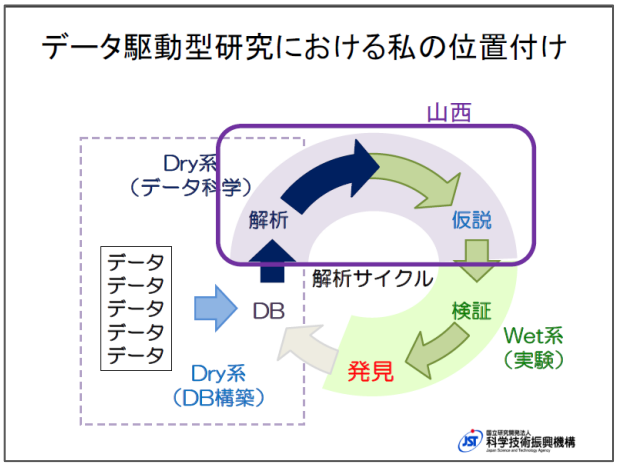
皮膚の細胞から神経細胞への直接変換、ダイレクトリプログラミングを誘導するような低分子化合物の予測にも取り組んでいる（図 5-24）。

様々な公共データに入っている多種多様なデータを機械学習で統合解析することで、化合物の新しい効能をデータ駆動で予測できることを示した。それは治療効果であったり、健康効果であったり、分化誘導能などに相当する。こういった考え方は漢方や食品などにも展開可能だ（図 5-25）。どちらにしても、データサイエンスの研究とウェット研究のうまい連携が重要だと、日々、実感している。

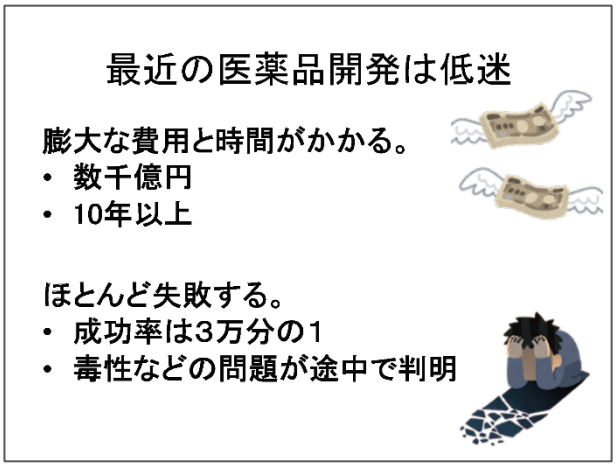
(2)　資料



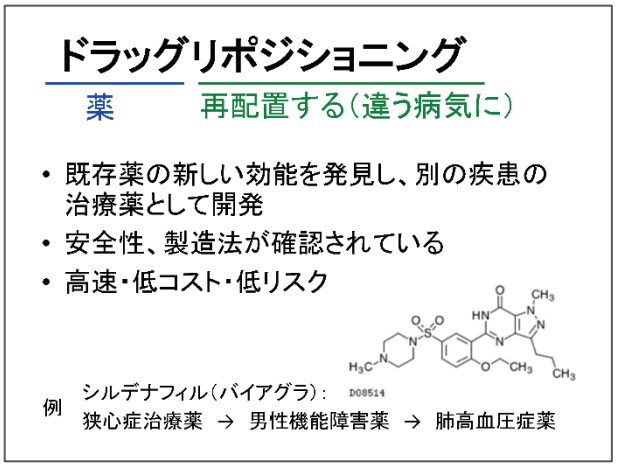
5-1 表紙



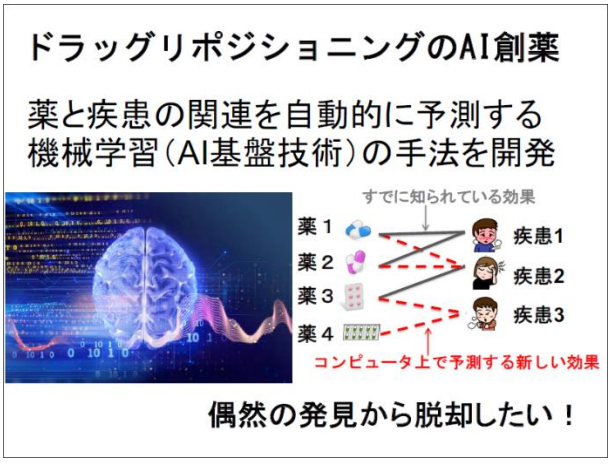
5-2 データ駆動型研究における私の位置付け



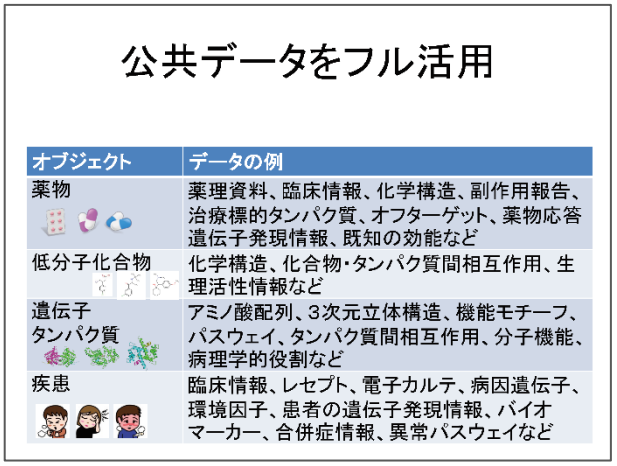
5-3 最近の医薬品開発は低迷



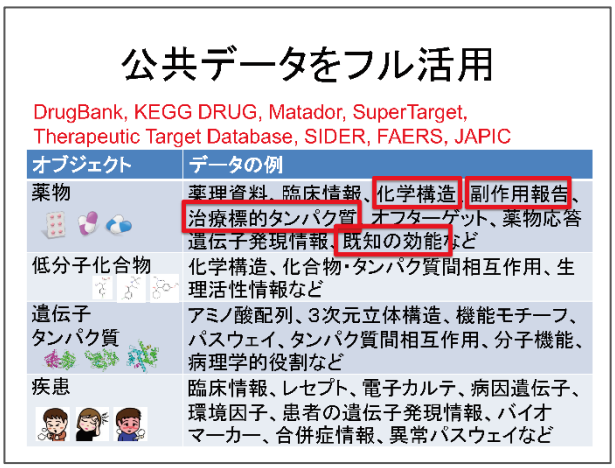
5-4 ドラッグリポジショニング



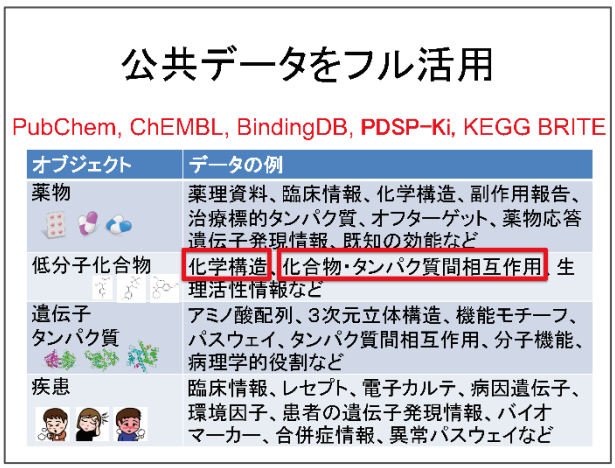
5-5 ドラッグリポジショニングの AI 創薬



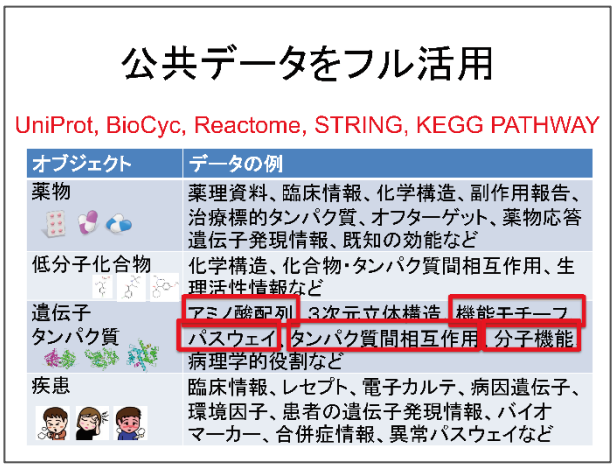
5-6 公共データをフル活用



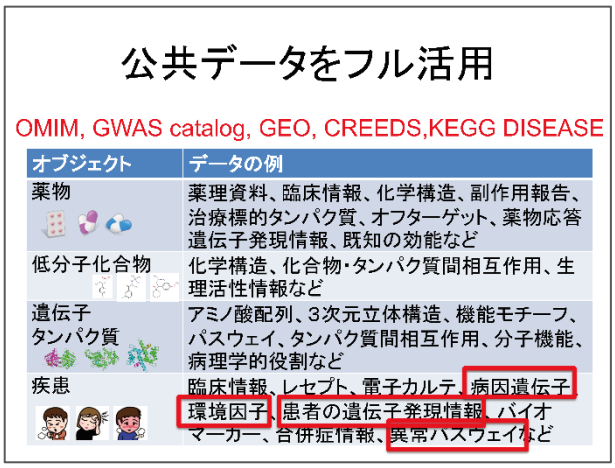
5-7 公共データをフル活用（薬物）



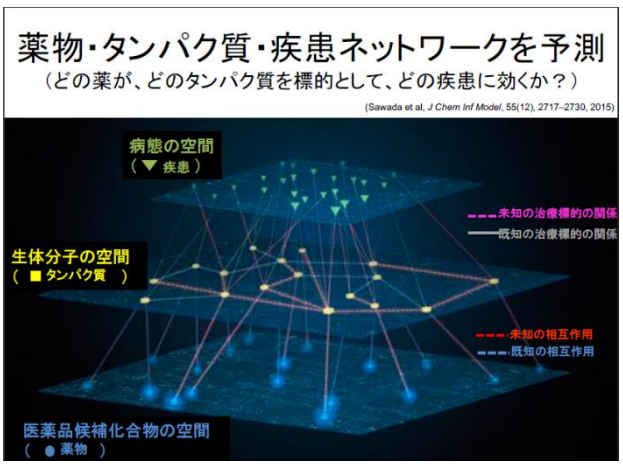
5-8 公共データをフル活用（低分子化合物）



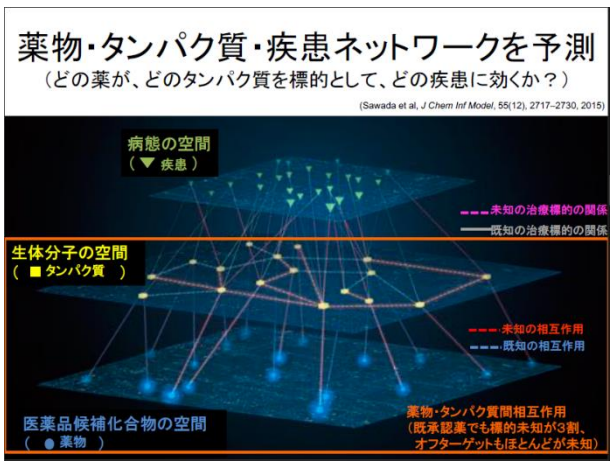
5-9 公共データをフル活用（遺伝子・タンパク質）



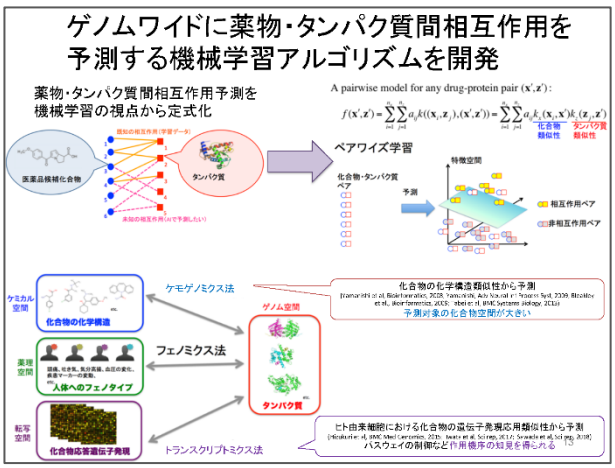
5-10 公共データをフル活用（疾患）



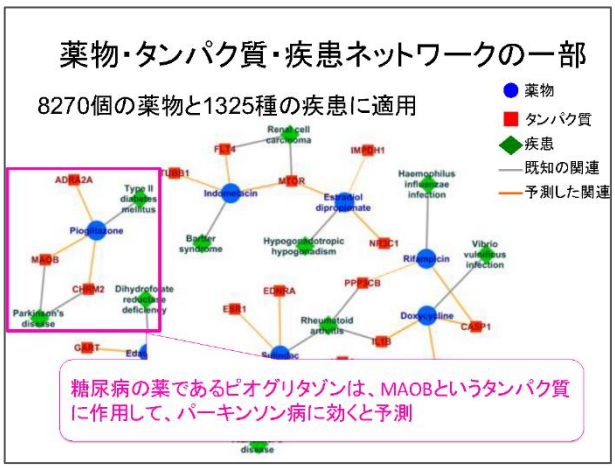
5-11 薬物・タンパク質・疾患ネットワークを予測（1）



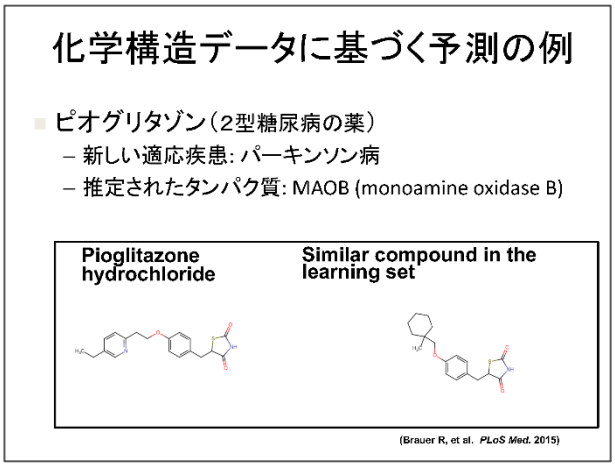
5-12 薬物・タンパク質・疾患ネットワークを予測（2）



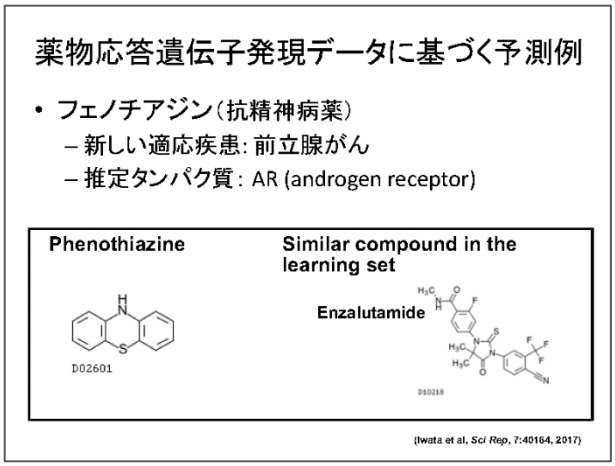
5-13 ゲノムワイドに薬物・タンパク質間相互作用を予測する機械学習アルゴリズムを開発



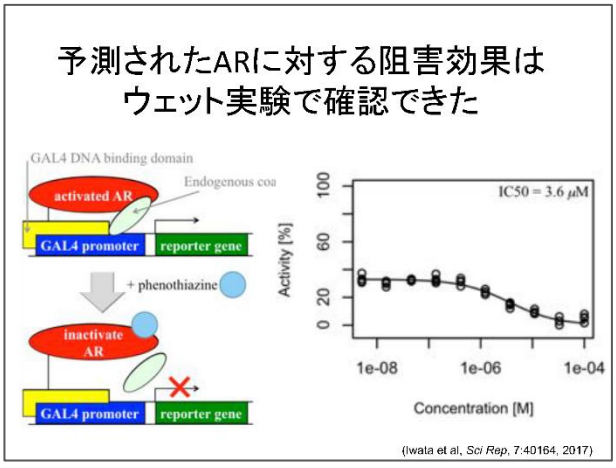
5-14 薬物・タンパク質・疾患ネットワークの一部



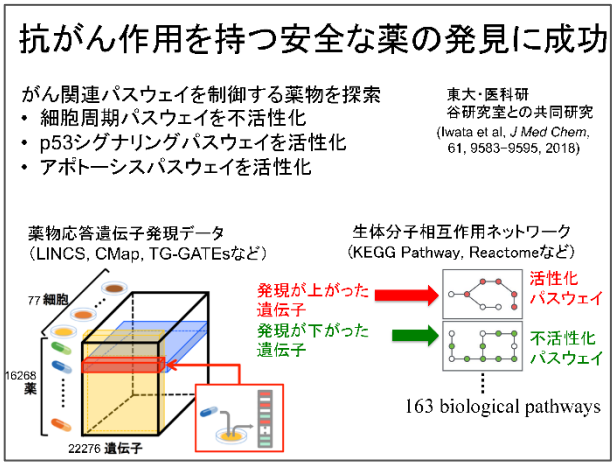
5-15 化学構造データに基づく予測の例



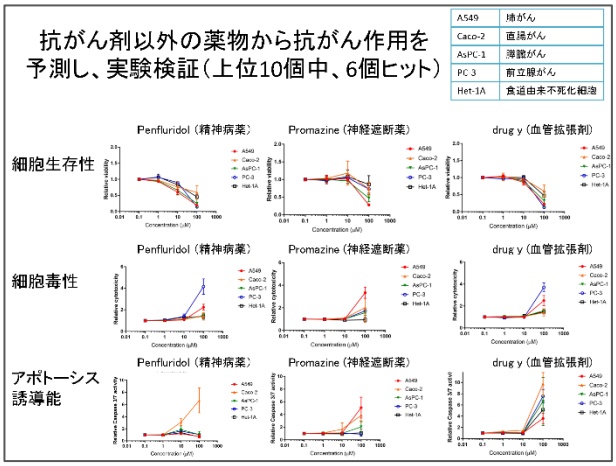
5-16 薬物応答遺伝子発現データに基づく予測例



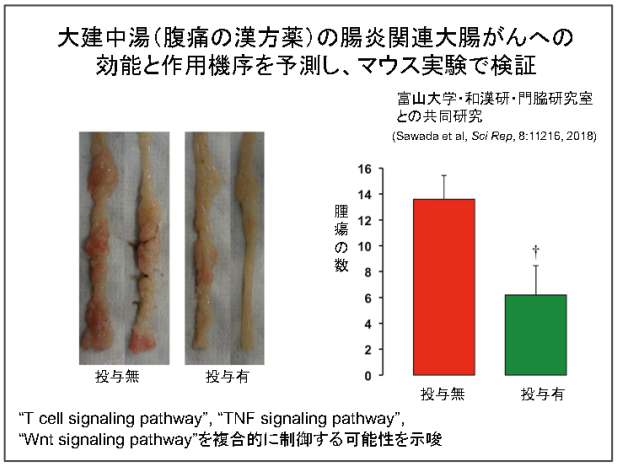
5-17 予測された AR に対する阻害効果はウェット実験で確認できた



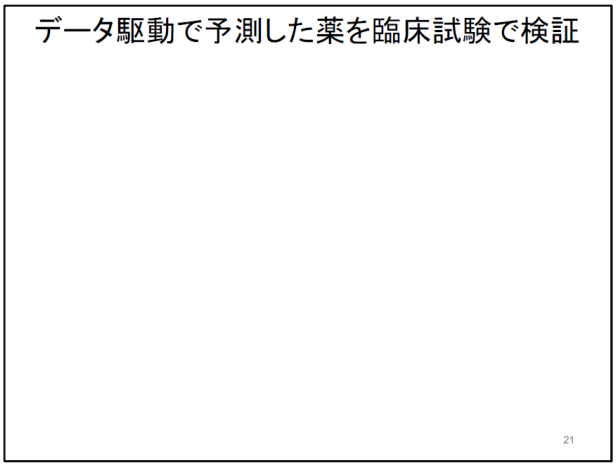
5-18 抗がん作用を持つ安全な薬の発見に成功



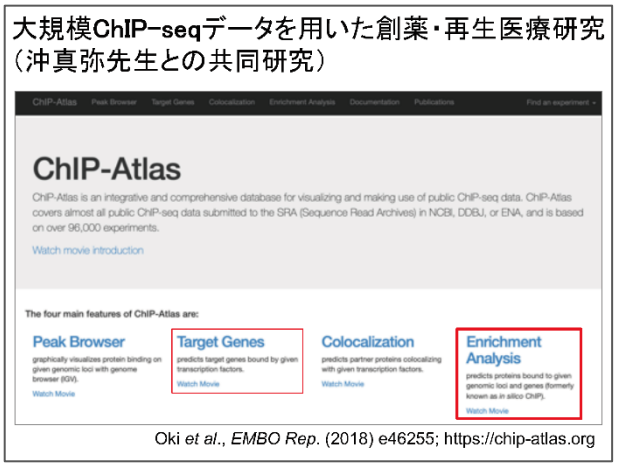
5-19 抗がん剤以外の薬物から抗がん作用を予測し、実験検証



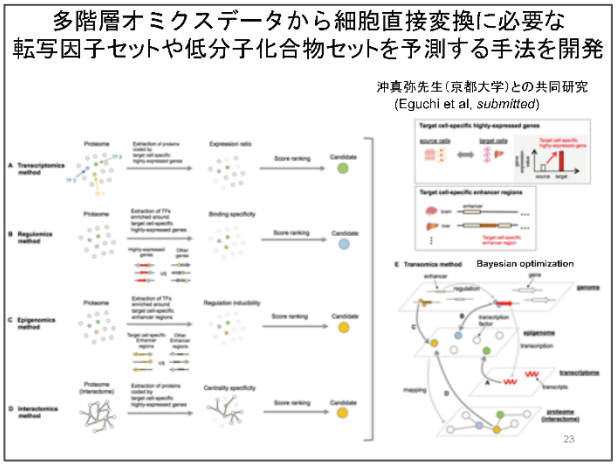
5-20 大建中湯（腹痛の漢方薬）の腸炎関連大腸がんへの効用と作用機序を予測し、マウス実験で検証



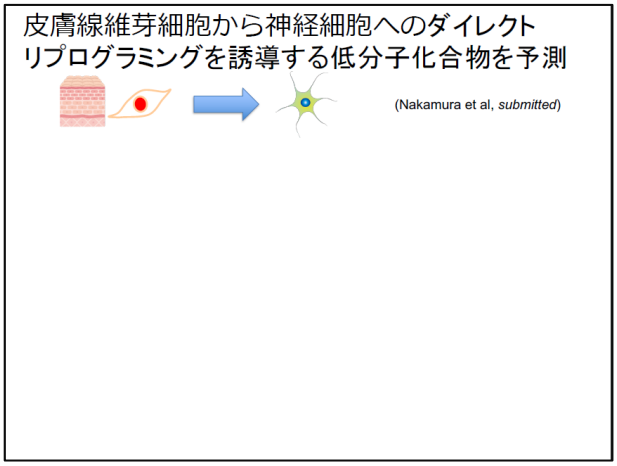
5-21 データ駆動で予測した薬を臨床試験で検証



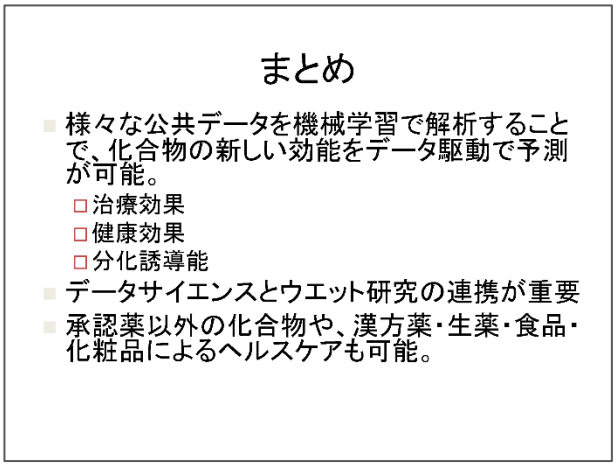
5-22 沖真弥先生との共同研究



5-23 多階層ｵﾐｸｽﾃﾞｰﾀから細胞直接変換に必要な転写因子ｾｯﾄや低分子化合物ｾｯﾄを予測する手法を開発



5-24 皮膚繊維芽細胞から神経細胞へのダイレクトリプログラミングを誘導する低分子化合物を予測



5-25 まとめ

(3) 質疑応答

（質問）今回の研究は、例えばタンパク質と化合物や、タンパク質とフェノタイプなど、いろいろなペアで機械学習モデルをつくって、それで全体のネットワーク構築をされるということかと思う。しかし、膨大な手作業が発生すると思う。機械学習の関係を見いだしてくるところは計算だが、一個一個モデルをつくってということは膨大な手作業となり、ウェットな実験研究者には、自力ではできないところになってしまうのではないか。ツール化といったことは可能だろうか。それとも、ツール化は難しく、専門家と組まないといけないだろうか。

（回答）ネットワーク予測をするための機械学習モデル、これを学習するためのデータ整備について回答する。今回の場合、化合物とタンパク質の相互作用のデータ自体は、既存の公共データベースに入っているデータを収集したものだ。ただ、御指摘のとおり、マニュアル作業は度々発生する。データベースごとに化合物の名前や ID、遺伝子の名前、タンパク質の名前、ID がばらばらで、フォーマットもばらばらなので、まず統一して、機械学習で解析できるフォーマットにしてやるところに結構時間がかかる。そこは御指摘のとおり。ただ、それが終わったら、情報解析するところはスムーズにできると思う。

今回は公共データを使った例だけを紹介したが、もちろん実験系、医学系の研究者が持つ独自の実験データを生かして学習データに入れて予測することも可能だ。その辺りは実験系の研究者がどう考えるかに依存していて、研究者が持つデータをより重要視して予測してほしいとか、これは補助的なもので公共データベースの網羅的なデータを重要視して予測してほしいとか、その辺りをリクエストに応じて解析している。

（質問）モデルもいろいろ選択肢があると思う。ウェット研究者の要望で決めるのか、それとも様々試すのか。

（回答）ウェット研究者がリクエストする場合もある。例えば、最近は AI ブームで、深層学習が流行っていることもあって、ここにあるネットワークを用いて予測してほしいとか、そういう依頼も多くはなっている。しかし、現場の人間としては、データが同じで、学習データが同じで、データの表現方法も同じだったら、どんな方法を使っても予測精度自体は差がないと感じてはいる。ただ、手法によって解釈性や、なぜその予測結果が出たのかを提供できる方法とできない方法がある。そういったところで、この方法だったらこうした解釈ができますよといった助言をしながら、どういうモデルを使うのかを決めていくような共同研究を行っている。

（質問）先ほどの回答のなかで、山西先生ご自身でいろいろな公共のデータベースからデータを集めてきて、相当の手作業によって整えられたということであった。ということは、有用な公共データベースが世の中にいろいろあるが、それらが必ずしも統合して使いやすい形にはまだまだなっていないという理解でよいか。

（回答）そうだと思う。本当に個々のデータベースはすばらしいと思うが、独自の ID が使われていたり、独自のフォーマットになっていたりするので、まずはそこから必要な情報を取ってくるところから始めないといけない。そこが大変で、また、フォーマットが異なる複数のデータベースがあって、そのフォーマットも年によって変更があったりする。これまで使えていた、パーシングできるプログラムが急に使えなくなるといったことも起こるので、苦労している。データベースの数は収束せずにどんどん増えていく傾向があるので、そういう作業は終わることがないと個人的には考えている。

本当に世界中のいろいろなデータが収集されて、統一された ID や名前、フォーマットになっているものがあるならば、私としては非常にありがたい。

（質問）そうして収集した公共データに加え、例えば共同研究者の方のプライベートなデータも交ぜながら研究をされていることと理解した。

（回答）その通り。

IV. 総合討論

司会：川口 哲（JST-NBDC）

1　発表内容

(1)　趣旨説明

○司会：最初に、データ駆動研究による新たな価値について議論したい。

今回の議論の役者は３者、すなわち、ドライ系のデータベース構築（鎌田先生が言及されたデータエンジニアリング部分）、ドライ系のデータ科学（これは一般的にインフォマティクスと言われていると思うが、ここではこういう言葉を使う）、そしてウェットの実験科学者だ（図 IV-3）。この３者が今後どのように協業していくのかが論点になる。

図 IV-3 に示すサイクルに対し、「この議論は 20 年前からやっているでしょう、今さらではないか」と思われる方もおられるかもしれない。実際に現代のウェット研究者は、データの多寡は別としてデータベースを使っている。

それでも、あえてこの課題を設定した理由は２つある。１つはワークショップ冒頭の趣旨説明で示したように、実験環境が大きく変化している。20 年前に比べてデータの量も質も大きく変わっているし、技術も大きく進歩している。２つ目は、データ駆動型科学の価値を確認したい。ライフサイエンスの分野において本質的にデータ駆動科学が研究開発に変化をもたらすのかを確認したい。

前述の３人の役者について、日本のライフサイエンスの研究者の割合を、ざっくりではあるが整理した（図 IV-4）。分母は分子生物学会員の 8,744 名。分子は 2 つで、1 つめはいわゆるドライ系のデータサイエンティストの方が所属しているバイオインフォマティクス学会の約 500 人、2 つめはデータ構築に関わる方が参加する、トーゴーの日シンポジウムの参加者の約 400 名。全体の約１割がいわゆるドライに位置づけられる。分母については、日本生化学会やその他の農学系、微生物学系の植物を含めればもっと大きくなるので、いわゆるドライの実際の割合はさらに少ないはずだ。

この構成比でどうサイクルを回すか、ということになる。これは、人材育成だけが解決策ではない。今いる人材が活躍しやすい環境を整えたり、データベース開発とデータ解析をする拠点を形成したりすることも考えられる。実際、ウェットのライフサイエンスからはそうした要望が多い。

まず、データ駆動型研究の価値を、具体例で議論したい（図 IV-5）。具体例で、といったのは、例を出さないとウェットの方に賛同してもらえないと思うからだ。これまでの発表を聞いていても、ウェットの方とドライの方がかみ合っているのかどうか、懸念を感じる面がある。

図 IV-5 に名前を挙げた方々に、こういうデータサイエンスに今後どういう期待をしているのか、どういう方向性が考えられるのか、御意見をいただきたい。

本日は、日本全体の課題として捉えて議論したいと思っている。

(2)　具体事例に基づく課題の洗い出し①

○司会：山西氏の事例をベースに DX の課題を整理した（図 IV-6）。山西先生の研究は、我々が考えているデ

ータ駆動型に非常に近い。その理由は、公共データベースを活用し、そこから新たな統合データを自分で作成されている点、また、それを踏まえてリポジショニングに向けた新たな因子を発見しているという点の 2 つだ。

山西氏は遺伝子発現や化合物や GWAS といったデータを御自身でダウンロードし、御自身でデータをつなぎ、そこから実際に疾患の制御の可能性のある因子を持ち込んでいる。普通のウェットの方ではできないのではないか。データベースから統合データをつくるところに一つの隘路があると思うが、ここが自動化されていたり、公共データベースとして提供されていたりといったことはないか。

○山西教授：様々なデータベース中の情報から統合データをつくるのは難しい。データベースのフォーマットや、ID、名前が本当にばらばらだ。データベースによっては、「ほかのデータベースの ID にはこの ID が対応します」といった ID と ID の対応表を提供しているが、その対応表自体が間違っているケースもある。結局、自分で化合物の構造を見たり、遺伝子の配列を見たりして確認しないといけない。ウェットな人、プログラミングの知識がない人がやろうとするのは、かなり大変な作業なのではないか。

○司会：汎用的なデータを整備するにしても、それぞれによってニーズが異なるだろうか。国として汎用的な統合データを用意していくべきか、それともニーズに応じた形で使いたいデータを整備していくべきか。

○山西教授：個人的には、医薬品関係の統合データベースがあったらありがたい。ただ、例えば、創薬の現場で重要なデータというのは、新しいものがどんどん出てくる。例えば 2010 年頃に薬物応答の遺伝子発現プロファイルのデータが Broad Institute から CMAP という形で出たが、2017 年頃には、網羅性を大幅にカバーする形で LINCS ができた。新しいデータベースができると、測定プラットフォームや濃度、時間などの条件が異なり、またその表記方法も異なったりする。ウェット主体の人がそれを見て、いろいろなデータと統合して独自の、自分の目的を達成するための価値ある統合データをつくるのはハードルが高い。そうしたデータを統合してくれるとよい。

○司会：図 IV-6 の②の部分の、研究成果のバイオロジカルな価値をどうやって判断するのかが気になっている。先生はどちらかというとデータサイエンティストの部類に入ると思うが、生物学的なクエスチョンやニーズはどうやって取り込んでいるのか。もしくは、どなたかと連携する形か。

○山西教授：両方やっている。私はデータサイエンスの人間で、私ができることは何かを予測するところまでだが、それだけだとサイエンスの価値は低い。そこで、できるだけそれを実証するために実験系の人と共同研究している。とはいえ、「これを検証してください」と言ってもやってくれないケースがほとんどだ。実験系の人が本当に興味を持っている疾患、生命現象、臓器といったニーズを把握した上で議論を進めると、うまくいくケースが多いように感じる。インフォ側もウェット側も、どちらも自分が下請的に使われるのは嫌だと思う。ウィン・ウィンの関係になるような共同研究を心がけている。

○司会：ウィン・ウィンの関係を構築することがとても難しいと思う。後ほどウェットの方からも御意見をいただきたい。

(3) 具体事例に基づく課題の洗い出し②

○司会：図 IV-7 に竹本氏の研究を整理した。主に ChIP-Atlas と GWAS のデータをつなぎ、そこから転写因子のスポットを絞り込み、さらにそれを CRISPR-Cas でノックアウトマウスを作成しているという理解で正しいか。

○竹本教授：その通りだ。

IV 総合討論

○司会：2 つのデータは最初からつながっているわけではなく、それぞれ独立に絞り込むプロセスが必要だと思う。これは手作業か。情報量が少ないなら手作業で済むだろうが、今後、データの種類を広げて解析する場合、課題となるのではないか。

○沖特定准教授： ChIP-Atlas と GWAS データをつなぐのは、ほとんどコンピュータで処理している。GWAS のデータは GWAS カタログを用いている。どのゲノム座標に SNP があるか、何の疾患と関連しているかというデータベースだ。ChIP-Atlas は、これもゲノム座標が主体となったデータベースだが、染色体何番のどこにどんな転写因子がひっつくかというデータなので、座標と座標を突き合わせることで、この疾患の SNP にはこんな転写因子がひっつきますよという統合データをつくることができる。なので、ここは非常に簡単だ。

統合したデータから、竹本氏との共同研究で優先順位をつけた。ヒトとマウスの間で保存されているかどうか、保存されていた場合に、マウスのゲノムにおいてもいろいろな転写因子を結合しているか、という点を判断材料にした。あまり余計な知識を入れないようにした。単純に、例えば、自己免疫疾患だったら血球系の細胞における血球分化とかに重要な転写因子が何個ついているかを順位付けし、順位の高いものを竹本氏が変異導入するという流れだ。

○司会：ドメイン知識に基づく遺伝子の絞り込みは、誰でもできるものではないのではないか。

○沖特定准教授：おっしゃるとおり、どのデータを組み合わせてどうやるかというプランを立てる部分は、ウェットの知識がかなり必要になる。プランを立てさえすれば、その後の統合データの作成や予測の作成はほぼ自動的にできるということだ。

○司会：お二人は近い研究室にいらして、お互いをよく理解しているからこういう取り組みにつながった。しかし、一般化するには、出会いの場をつくるのがよいのではないかという御提案との理解でよいか。

○沖特定准教授：その通り。竹本氏と私が採択された研究費は、基盤 B の特設分野で時限付だった。数理や、データサイエンス、ウェットを組み合わせた提案を募集していた。現在は、科研費のなかでドライとウェットの融合を求める審査区分がないと思う。政策なり、グラントの審査の仕方が重要になるのではないか。

○司会：今回、候補数が 20 程度にまで絞り込めれば、竹本氏の新技術によって表現型解析ができるとのことであった。しかし、先の手法で絞り込めない場合、データ統合や解析のステップでさらに絞り込むための技術なりが要ると思う。そうした課題に直面した例あるか。先ほどの手法で候補数を絞り込めない場合のプランについて、お考えがあれば。

○竹本教授：私は、候補数を絞り込むのに苦労した記憶はない。他の研究者との共同研究では、コーディング領域に特化し、絞り込むことができた。

○沖特定准教授：私は、候補がたくさん出てきて困ったことはある。今回の竹本氏との共同研究でも、最初に出てきた候補数が多過ぎたので、いろいろなフィルターを試した。ただ、私自身、データサイエンス自体があまり得意ではない。データベースをつくる人がいて、ウェットの人もいて、山西氏のようなデータサイエンスができる人が 3 者で組まないと、限界がある。

(4) 具体事例に基づく課題の洗い出し③

IV 総合討論

○司会：平井先生が 10 年前に御自身でドライに取り組まれたご経験について伺いたい。

○平井チームリーダー：植物の代謝研究において、数理モデルをつくり、代謝ネットワークの推定をしていた。具体的には、バイオケミカルシステム理論という理論によってモデル式を構築し、メタボロームデータを使ってそのパラメータを推定するということをした。私は自前のデータだけを用いたが、公共データを使うともっと普遍的なモデルがつくれるのではないか。しかし、メタボロームデータの公共レポジトリが不十分であるのが隘路のひとつであり、現在は世界的にその構築の取組みがなされている。

また、私は理論の専門家と共同研究させて頂いたが、ウェットの人間だけでいかにデータを使い倒せるかというようなことも考えていた。そのためには、データからどうやって知識を抽出してくるかという方法論、コンセプトを確立する必要がある。そのためのツール開発をして、それを一般化するということが必要だと思っていた。しかし、そこも隘路だと思う。理論の構築とか数式を扱うのは実験研究者には難しい。そこがツールになっている、あるいは「こういうことをしたらこういうことが分かる」という概念・方法論が構築されて、広くウェットな人に普及されると、ウェットな人だけでも自立してできるのではないか。現状、インフォマティクスは谷間というか、実験研究者から見ると、うまく使えていない。

(5) データ駆動型研究の将来展望

○司会：データ駆動型研究は、本当に想定外の発見に繋がりうるか。今後の可能性をお聞きしたい。
山西氏の事例は、ドラッグリポジショニングの観点から今後も新しい展開が描けると思う。竹本氏、沖氏も、GWAS と ChIP-seq の組み合わせで新しいターゲットを広げるような、いわゆる産業的な方向性及びサイエンティフィックな意味でも非常に典型事例になりえる。

また、jPOST やほかの多様なデータベースがつながっていくと、さらにそういった想定外の発見の可能性も広がってくると思うが、どう思われるか。

さらには、今、医療分野では、それこそゲノムを中心に既にがんの医薬品が迅速に同定される時代になり、オミックスレベルに上げた形で様々な疾患に対応していくという流れがある。その中で、データ駆動型研究の可能性について、科学技術への貢献及び医療の観点で改めてお聞きしたい。

○永井学長：ある現象に意味があるかというのは、「たまたま」かどうかということとの戦いだ。「たまたま」を脱却するためにいろいろなデータベースを見ながら方向を決めていく必要がある。

臨床医の立場からお話しする。最終的にいろいろなバイオメディカルサイエンスが医療で評価される。問題は、ヒトを対象とした計測データは、実験動物と比べて非常にばらつくということ。そのばらつきの中で評価することは容易ではない。大規模臨床試験、3,000 例や 5,000 例の比較研究が行われるが、あれもばらつきが大きいことが理由だ。非常に大きなばらつきのある集団を対象に統計的な有意差を見い出すためには、N を大きくするのが定石である。しかし、いつでもできるものではないし、長い時間と多くのコストが非常にかかる。

そういう中で出てきたのが、個別化医療である。特にリアルワールドデータも使って事前確率を高めれば、より少ない数で個別化医療がある程度決められるのではないかという考えが指摘されてきた。情報は、ゲノムだけでなく、多ければ多いほどよい。人間のばらつき・多様性をどう制御していくかというストラテジーが必要だ。

○菅野非常勤講師：データサイエンス分野の成果を医療分野に応用していく研究開発を、地道にではなく、飛躍的に、規模を拡大して推進する必要がある。

ビッグデータが、医療とは関係ない分野で非常に大きく話題になり、情報解析の手法も様々なものが生み出されてきている。そうした手法群がデータ駆動型研究に直結してくるのではないか。医療分野ではデータサイエンスの進歩の取り込みがゆっくりだ。日本は特にだ。みんなピンと来ていないが、そのうち「えっ」となるような結果が数多く出てくる。今や、医療分野では電子カルテや、ゲノムデータなど、かなりの量のデータがあちこちで利用可能になりつつある。画像処理に使われているような技術が応用され、思いもかけないような発見がされるなど、ウェットの研究をやっている人が広範な情報技術を日常的に使って研究する時代がすぐそこまで来ているのではないか。

(6) アルゴリズム開発について

○司会：CRDS がまとめたデータ科学の課題を図 IV-8。に示す。アルゴリズム部分に谷があり、戦略的な手当が必要なのではないか。国の全体を見ると、AI や IT 関係で、多くの施策が打たれている。ライフサイエンス分野における課題と日本の対応策についてどのように考えるか。

○瀬々社長：個人としては、アルゴリズムは大分出来上がっていると感じる。それよりは、どうやって適応していくかだ。機械学習、データサイエンスに取り組む人が、そのデータの利用の先を理解することが必要だ。また、データベースの人たちもアルゴリズムをどんどん使っていく時代になっている。したがって、谷になっているイメージはあまりなく、ちゃんとつないでいくことが必要だ。両者がちゃんと歩み寄らなければいけない。

Nature のトップページに、昨日（2020/11/30）、CASP の結果が掲載された。Deep Mind が AlphaFold2 をつくって、驚愕するような結果を出した。もしかしたら、タンパク質の立体構造解析はもう要らないかもしれないというレベルだ。そうした時代が、すぐそばに来ている。もう、データを使わない研究者全員を圧倒する勢いでライフサイエンスを革新しないと、世界で立ち遅れてしまうのではないか。

(7) 人材育成

○小安理事：ウェットか、ドライかではなく、両方できる人材が育ってきている。いかに我々が後押しできるかが大切だ。また、解きたい問題があるということが大事だ。

理化学研究所でデータサイエンスに取り組む若者を紹介したい。川上氏は東京大学 医学部を卒業した。しかし、数理に非常に長けていて、高校時代に数学オリンピックに出たほどだ。人工知能のことも分かる。清田氏は、筑波大の医学部を卒業して外科医をしていたが、基礎研究に転向し、数学や人工知能に取り組んでいる。こうした人材の下には、例えば数学科を出てから医学部に入ったような学生が集まってくる。彼ら・彼女らは生物学的な知識も持っているし、数理もできるし、AI のテクノロジーも使えるし、自分たちでアルゴリズムを書ける。こういう人材をどうやって集め、育てていくかが大事だ。

川上氏らは、血液検査のデータだけで卵巣がんの患者さんの予後を予知するため、機械学習を用いてアルゴリズム開発を行った。また、糖尿病疑いの患者を層別化した。また、清田氏らは、深層学習を使って 6 時間後に敗血症を発症するか否かを血液データから予測するコンペで、世界で約 10 位の成績をおさめた。こうした技能をもった人材が集まっている。

こうした場をつくっていくことが大事だ。私自身は古典的な仮説ベースの免疫学者だが、彼らと一緒に議論していると色々新しいものが見えてくる。アトピー性皮膚炎の患者について、血液データだけを用いて層別化し、ある薬剤に対して効く患者、効かない患者、必ず悪化する患者などで分ける。すると、アレルギー分野でよく知られたマーカー以外に、全く知らないようなマーカーが出てくる。実際、そこに新しいメカニズムが隠れていそうだということ

が見えてきた。実例に鑑みると、こういうアプローチがとても大事だと感じさせられる。

ドライのツールが使えても解きたい問題がないのではだめで、解きたい問題があってドライのツールを使うということ。ウェットにもドライにも両方に興味があって、自分の解きたい問題をどうやって解くかという、総合的にアプローチすることがこれからやられていくのではないか。

○伊藤教授：私は、約 30 年前に、データ駆動型研究を始めた。本質は当時と変わっておらず、周りの環境が著しく変わったという印象がある。

データ駆動型研究をはじめた当時に私が期待したことは２つあった。１つは、量的な違いだ。発見のスピードが上がる。もう１つは、質的に異なる発見。全体が見えてくることで、これまで知られていなかった特徴が新たに見つかってくるならば、科学の質を変えることにつながる。

今、こうした変化が訪れてくるようになってきた。環境が変わってきた。本ワークショップの議論を聞き、いよいよ 2 点目が実現しうる時代だと思った。そのためにはいろいろな人の協働が必要だ。一方で、若い人の中では１人でどんどんと越境している人はもう確かに出てきている。そういった人材をさらに後押しするようなことができるといい。

2001 年開始の、JST が実施した BIRD 事業では、情報科学と生物科学との融合型アプローチによる研究開発が公募された。いま、らせんをぐるりと１つ上に上がった段階で、この時代に即した取り組みを講じることによって、越境しようとしている人たちを後押しすることになるだろう。また、後続者をも生んでいくことになるのではないか。

○坊農特任教授：個別のデータ統合の手法については、NBDC 事業において、チュートリアル動画を発信する等が行われている。この活動は大変評価されていると理解している。

(8)　まとめ

○司会：今日は全体を通して、それぞれの隘路も分かった。また、その隘路を打破するための人材育成や拠点の中での組み合わせ、さらにはファンディングの仕組みについて示唆いただいた。新たな問題を掲げた上で、関係者がデータ主導で解決していく。もちろん、最終的には検証のための実験だ。そうした取り組みを支援する仕組みが国としても必要だと感じた。

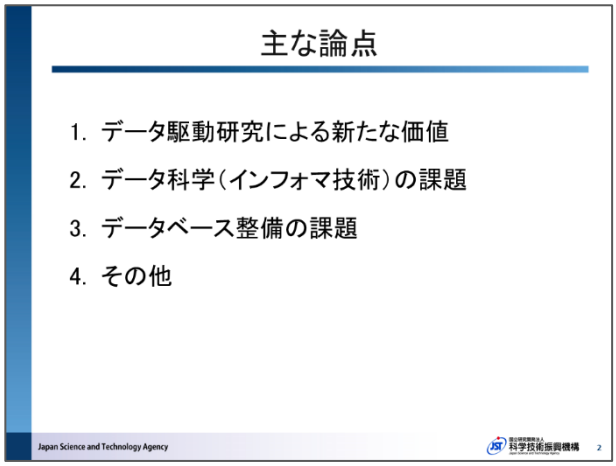
ただ、アルゴリズム開発は隘路なのではないか、いわゆるインフォマティクスの技術がもっと必要なのではないかという点については、充分に議論できなかった。

また、国としてデータベースをどう整備していくのか、１次データ、２次データ、そして NBDC がやっているような統合データについて、基盤を今後どう整備していくのかについても議論する時間がなかった。

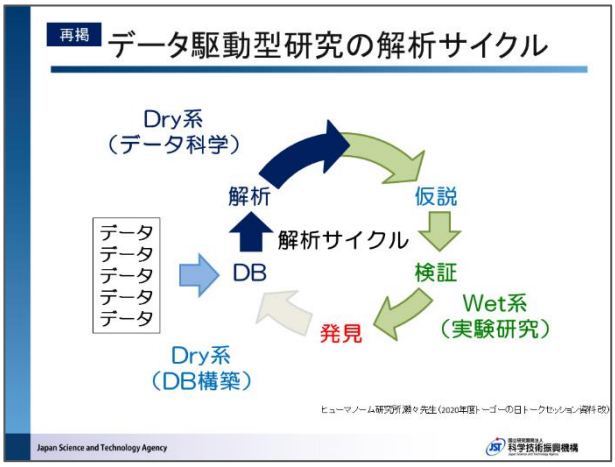
整理し、今後の対応策を考えたい。

2　資料

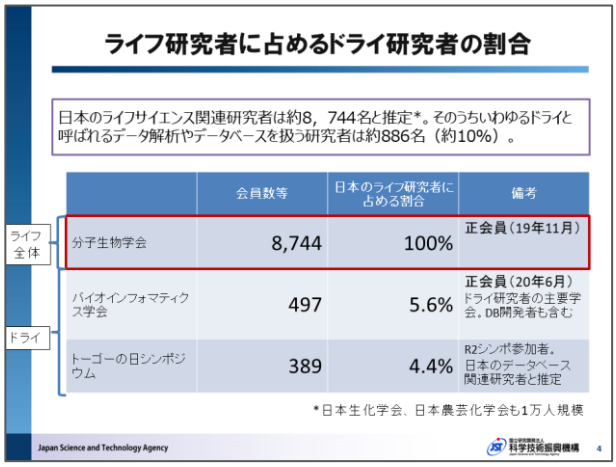
IV-1 表紙



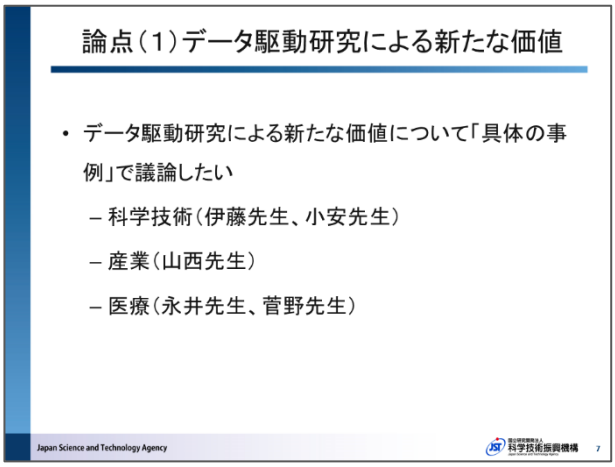
IV-2 主な論点



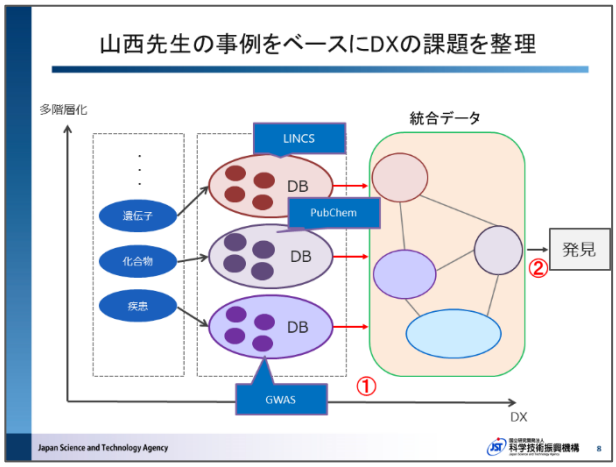
IV-3 データ駆動型研究の解析サイクル



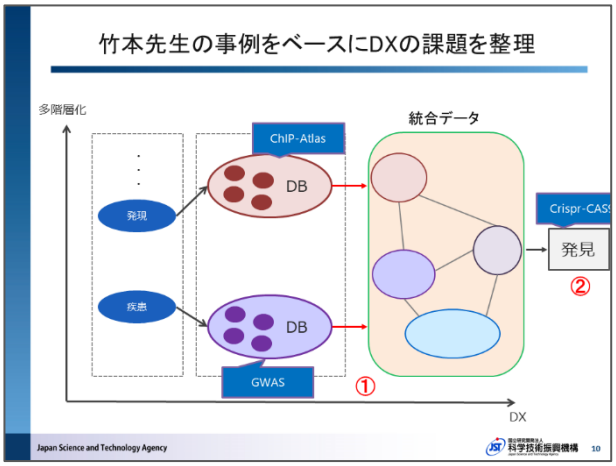
IV-4 ライフ研究者に占めるドライ研究者の割合



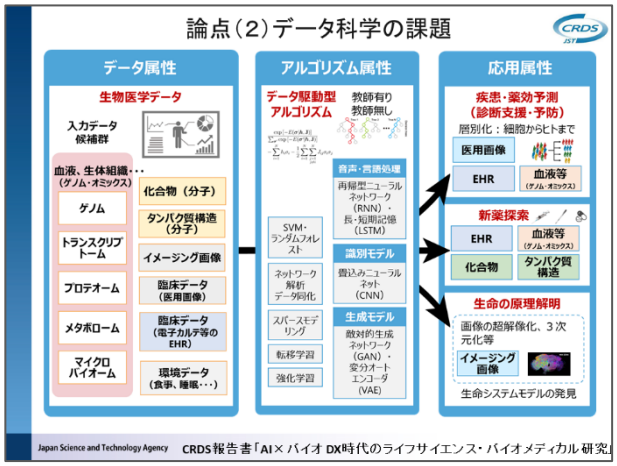
IV-5 論点（１）



IV-6 山西先生の事例



IV-7 竹本先生の事例



IV-8 論点（2）

V. 付録

1　ワークショップ概要

(1)　開催概要

件名：データ駆動型研究の推進と課題

日時：2020 年 12 月 1 日（火）15:00～18:00

形態：オンライン

(2)　目的

近年、生命科学研究では、計測技術の驚異的な進歩により、研究データが爆発的に増加し、情報のデジタル化、コンピューテーショナル化が加速度的に進んでいる。こうした研究データを集積、整理・統合し、生命現象を包括的に理解する研究開発は既にライフサイエンスの潮流を形成しているが、一方で、こういった研究開発の推進では、データの共有・公開基盤の構築のみならず基盤を高度に活用する取り組みが重要である。

こういった認識は 2000 年のヒトゲノム解読に端を発したオミクス研究の勃興から現在に至るまで当該分野の基本思想として定着しているが、わが国ではいまだにデータを活用したライフサイエンスの取り組みは限定的で、論文レベルでの国際競争力の低下も著しい。

以上を踏まえ、本ワークショップでは、研究データを高度に利用する研究開発の課題（ボトルネック）と、その解決に資する研究開発投資について有識者を交えた議論を行い、関連動向を踏まえた支援のあり方についても検討する。

(3)　プログラム

議題	時間帯	発表者	備考
開会あいさつ	15:00　～　15:03		
注意事項説明	15:03　～　15:13		
ワークショップ趣旨説明	15:13　～　15:23		
自己紹介	15:23　～　15:47		
話題提供 ①	15:47　～　16:00	石濱氏	発表 10 分、質疑 3 分
話題提供 ②	16:00　～　16:13	竹本氏・沖氏	
話題提供 ③	16:13　～　16:26	坊農氏	
話題提供 ④	16:26　～　16:39	鎌田氏	
話題提供 ⑤	16:39　～　16:52	山西氏	

議題	時間帯	発表者	備考
休憩	16:52 ～ 17:02		
総合討論	17:02 ～ 17:59		
閉会あいさつ	17:59 ～ 18:00		

(4) 出席者（敬称略）

外部有識者

石濱 泰	京都大学 大学院薬学研究科 教授
伊藤 隆司	九州大学 大学院医学研究院 教授
沖 真弥	京都大学 大学院医学研究科 特定准教授
鎌田 真由美	京都大学 大学院医学研究科 准教授
小安 重夫	理化学研究所 理事
菅野 純夫	東京医科歯科大学 難治疾患研究所 非常勤講師
瀬々 潤	ヒューマノーム研究所 代表取締役社長
竹本 龍也	徳島大学 先端酵素学研究所 教授
永井 良三	自治医科大学 学長
平井 優美	理化学研究所 環境資源科学研究センター チームリーダー
坊農 秀雅	広島大学 大学院統合生命科学研究科 特任教授
山西 芳裕	九州工業大学 大学院情報工学研究院 教授

文部科学省

辻山 隆	ライフサイエンス課	生命科学専門官
山田 和輝	同	技術参与
寺本 敏紀	同	生命科学研究係長
本間 椋	同	生命科学研究係員

JST

高木 利久	NBDC	センター長
星 潤一	NBDC 企画運営室 室長	
川口 哲	NBDC 企画運営室 研究開発推進グループ	調査役
眞後 俊幸	同	主査
太田 紀夫	同	主任調査員
伊藤 桂子	同	主任調査員
金山 晋司	NBDC 企画運営室 企画・外部連携グループ	調査役
佐藤 早苗	同	副調査役
安達 澄子	同	主査

島津　博基	研究開発戦略ｾﾝﾀｰ(CRDS) ﾗｲﾌｻｲｴﾝｽ・臨床医学ﾕﾆｯﾄ ﾕﾆｯﾄﾘｰﾀﾞｰ	
中村　輝郎	同	フェロー
宮薗　侑也	同	フェロー
蔡　慧玲	戦略研究推進部　ライフイノベーショングループ	副調査役
江島　亜樹	同	調査員

以上