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

「NBDC で今後取り組むべき データベース整備の検討」

開催日時：2017 年 11 月 5 日（日）

2017 年 12 月

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

目次

I. エグゼクティブサマリー

1　開催経緯

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）が推進するライフサイエンス統合推進事業では、広く研究者コミュニティに共有かつ活用されることにより基礎研究や産業応用研究が活性化することをめざし、日本の生命科学データベースの統合を進めている。

統合化推進プログラムは、ライフサイエンスデータベース統合推進事業におけるファンディングスキームであり、平成 23 年度の NBDC 発足以来、ライフサイエンス分野のデータベース（DB）を幅広く整備し、多様な科学的知見を繋ぐことによって日本の生命科学研究の推進に貢献するために実施されてきた。本プログラムにより、現在、多様な分野の DB において、一定程度整備が進んだ状態にある（実施状況、主な成果等については、平成 28 年度に実施した事業評価に係る報告書「ライフサイエンスデータベース統合推進事業　事業報告書」（https://biosciencedbc.jp/gadget/unei/jigyou_houkoku.pdf）参照）。今後は、これまでの基盤整備の継続に加えて、日本の生命科学研究の競合優位性を高めるため、さらに産業応用に資するために戦略的な DB 構築を進めていくべきと考えている。

そこで、ライフサイエンス分野の有識者にご参加いただき、各分野の最新動向と今後整備が必要な研究データについて意見交換を行うことで、NBDC、特に統合化推進プログラムにおいて、今後どのような分野・領域に重点をおいて推進すべきかについての戦略の検討を目的として、「NBDC で今後取り組むべきデータベース整備について」と題したワークショップを開催した。

ワークショップの開催に先だって、NBDC は、DB 整備とその統合的な利用によって資するべき応用分野として、疾患の発生機序解明と創薬研究、農作物等の育種研究、有用物質生産法の開発等を掲げ、主に、統合化推進プログラムにおいてこうした応用分野の研究の加速に貢献するようなデータの整備を進めることを目標とした。また、上述に示したような、応用分野においては、ゲノム情報から表現型までの深い理解が必要であり、DB として整備の進むゲノム情報から表現型までの間に位置する階層の多様な情報の統合が必要である、との問題意識を提示した。

その上で、各分野の有識者には、応用を見据えた上で、ゲノム情報から表現型までを繋ぐために必要となる研究データ、DB を挙げていただき、議論の端緒とした。

2　ワークショップで出された主な意見

2.1　応用を見据えた際に整備が必要となる DB

（疾患解明・創薬）

※この分野は、AMED との有機的な連携・取り組みの位置づけ整理を行う必要がある。

・コホート研究データが統合化され、ゲノム情報等の一元的な検索、情報取得が可能な DB。

・医療関連の画像データおよびゲノム情報等とのリンクがされた DB。

・疾患関連タンパク質としての発現量情報や翻訳後修飾情報等のプロテオーム情報が整備された DB。

（育種）

・ゲノム情報や発現情報と形質情報がリンクされた DB。

・農林水産省の各プロジェクトで個々の研究目的に沿って作成されている DB、およびフィールドでの表現型データ、気象データ等を統合した DB。

（有用物質生産）

・バイオエコノミーという観点から、産業競争力にどう寄与していくかを強く意識した DB。

(例えば合成生物学的アプローチなどを考えた場合には、現在の DB では不足している情報が多く、これからのニーズを踏まえた DB の構築が必要。)

（食品）

※本ワークショップにて、上記項目に加えて着目すべき応用分野として挙げられた。

・複雑な物質循環を理解するための、メタボローム DB。

（メタボロームは、研究の特殊性等からデータ産出が難しく、データ量が少ないのが現状であるが、目的を明確にして整備を進めていくことが必要。）

2.2　新たなデータ整備、解析手法の観点から整備が必要となる DB

・画像データは多くの情報を含み、また生命現象の様々なスケールでの情報を取得できる。特に、医療、育種の分野での有用性が高く、整備が必要。

・データセット間における高い接続性の確保が必要。それにより全体で大きなデータとして一体的に扱うことができるならば、AI 研究、システムバイオロジー研究としても使い易いデータとなる。

・AI 研究では教師データ、システムバイオロジー研究では高精度の予測を可能にするためのデータが求められており、キュレーションされた質の高いデータ整備が重要。

2.3　DB 整備を進めていくにあたっての留意点

・目的ごとに整備すべきデータ内容、精度が異なるため、具体的な目的設定が必要。

・利用者からのヒアリングを継続的に実施し、求められている DB を常に意識しながら整備を進めていくことが重要。特に、応用研究に資することを目指すなら、産業界との密なコミュニケーションをとり、産業界、応用分野との共同研究を進めながら実例を示していくことが求められる。

・整備されたデータを解析するためのツールの整備や、使い易い DB 整備の工夫(インターフェース等)も重要。

・生物系の研究者と解析の研究者、両者が一緒に研究をすすめるような仕掛けが望ましい。

・データ統合・活用の具体的なフラッグシッププロジェクトを企画して推進するのもよいのではないか。

2.4　その他

・わが国としてバイオ戦略を検討している今まさにこの時期に、NBDC が実施している基盤整備の重要性をアピールすることが必要。

・公的資金研究データを十分集めるには、データ共有の義務化も重要。

・DB を利用する側としての多くのバイオ研究者が、DB を十分使い切れていない状況も踏まえ、利用方法のさらなる周知が重要。

・データ規模の増大に対応するような技術開発と環境整備（クラウド利用等）が重要。

・日本人固有のデータ等、NBDC にしかないデータ、コンテンツ、ツールの整備ができるとよい。

・DB 整備の人材への配慮（DB 整備は中長期的に継続的なエフォートを要する一方で、継続的な論文発表には直結しない等）が重要。

NBDC では本ワークショップで挙げられた意見を踏まえ、統合化推進プログラムでの公募領域の設定、さらに、NBDC 全体として今後の DB 整備の方策等についての検討を進める。

II. ワークショップの背景および概要

1　NBDC の活動

バイオサイエンスデータベースセンター（NBDC）は、ライフサイエンス分野における我が国のデータベース（DB）の一元的な統合を目指し、2011（平成 23）年 4 月に発足した。以来、日本のライフサイエンス研究から産出されるデータや DB が最大限に活用されるように、また、研究終了後に死蔵されることの無いように、これらのデータや DB を統合・整理するための研究開発とサービス提供を行ってきている。

具体的には、NBDC が実施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、以下の 4 つの柱に沿って事業を推進している（図 1）。

（1）戦略立案

ライフサイエンス分野の研究データに関する課題対応と我が国の事情を踏まえた事業運営戦略の立案・実施

（2）ポータルサイトの構築・運用

DB が散在する現状でも、必要な DB の探索や一括検索利用できる環境の整備、及び継続的に DB 公開を維持し、再利用を促進する機能

（3）DB 統合化基盤技術の開発

分野を超えた DB 統合化を推進するための基盤技術の開発

（4）統合化推進プログラム

ファンディングによる DB 統合化推進：分野毎に核となる統合 DB の整備の推進

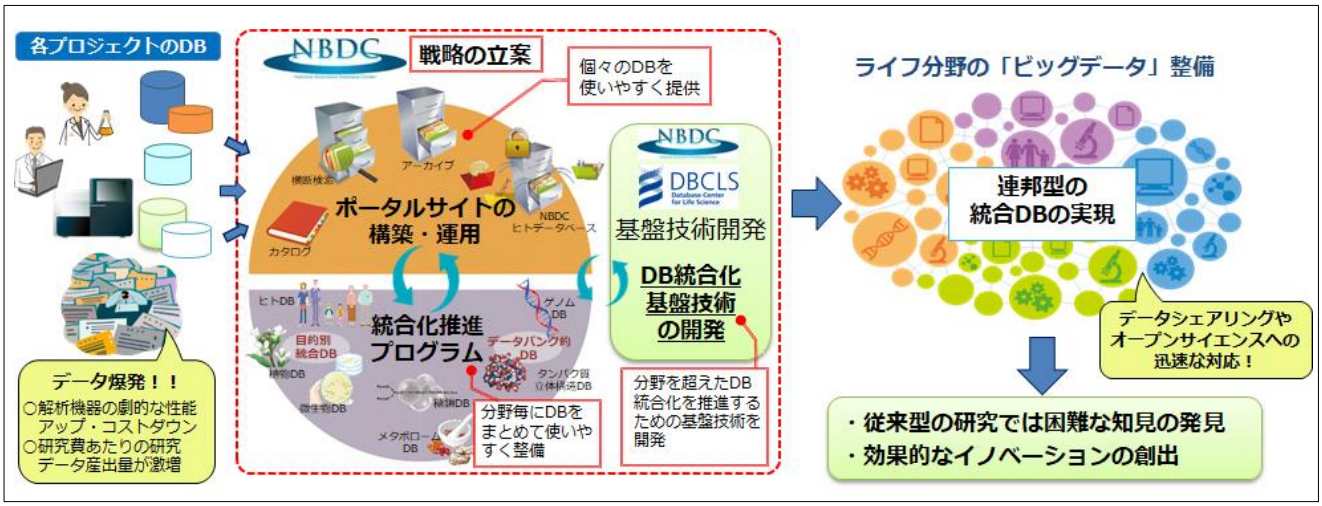


図 1. NBDC が実施するライフサイエンス統合推進事業の全体像

このうち、（1）～（3）については、NBDC 内、および共同研究、機関間連携等により実施をしてきているが、（4）の統合化推進プログラムは、ファンディングという形で進めてきており、公募により支援先を決定している。特に、平成 23 年度の事業開始以降、継続的に DB 整備を支援している分野も多く、各分野の DB 整備が整い、より一層ユーザと協働して DB の統合的な利用を進めていくべき段階にきている。そのため、今後、DB の統合的な利用を一層進めるためには、今まで以上に戦略的に DB の整備を進めていく必要があると認識している。

今回のワークショップでは、「NBDC 全体としてのデータベース整備」と題して議論をするものの、それを具体的に実行していく施策としては、統合化推進プログラム（図 2）（https://biosciencedbc.jp/tec-dev-prog/funding-pro

gram） を念頭において、議論を進めることとした。

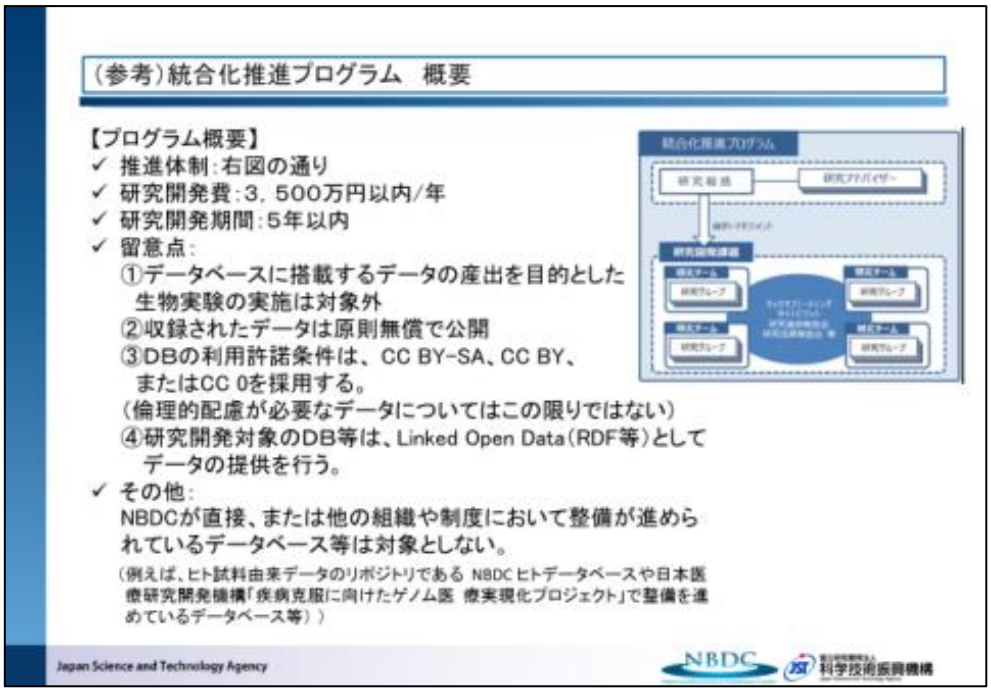


図 2. 統合化推進プログラム概要

2　NBDC のこれからの活動方針（事業評価および提言をうけて）

2.1　事業評価

NBDC では昨年度センター運営 6 年目を迎え、今後の NBDC の事業推進のために、これまで進めてきた事業推進状況の検証を行うことを目的として、外部評価委員による事業の評価を行った。その結果、「全体として優れた成果を出していると結論づけられ、近年のオープンサイエンスや情報技術等に関わる動向にも配慮しつつ、今後も継続して事業を推進することが望まれる。」との評価を得た。なお、今後の事業推進にあたっては、特に、ファンディング等に関連するものとして以下の意見をいただいている（https://biosciencedbc.jp/gadget/unei/jigyou_hyouka_23_28.pdf, https://biosciencedbc.jp/gadget/unei/jigyou_houkoku.pdf, https://biosciencedbc.jp/gadget/unei/jigyou_houkoku_gaiyou.pdf）。

- 医科学、医療、創薬の分野に加え、食、環境、バイオマスエネルギー（グリーンバイオ）などの分野におけるデータベース整備も重要である。
- ライフサイエンス分野の研究開発には、構造化データだけではなく、画像、テキストなどの非構造化データ、さらに、RDF 化が困難な膨大なデータ 等も重要であり、引き続きそれらデータの取扱いについての検討を進めてほしい。

2.2　提言

上記に記載の事業評価を受け、さらに、我が国のライフサイエンス研究におけるデータ基盤整備とその中で NBDC が果たすべき役割に関して NBDC 運営委員会が議論を行った。その結果が、「バイオサイエンスデータベースセンターの今後のあり方について（提言）」（https://biosciencedbc.jp/gadget/unei/unei_teigen.pdf）としてとりまとめられた。研究の進展等による変化には柔軟に対応しつつ、我が国全体としての課題には関係する省庁や機関と連携し

て積極的に解決にあたるよう提言がされているが、特にファンディング等に関連するものとしては以下のとおり。

・データベース構築者の視点から、利用者の視点に転換する－【新たな利用】
　・ヒト由来データの共有と利活用は喫緊の課題であり、ヒトとモデル動物の研究データをゲノムから表現型まで統合し、利用者の要請に応えられるデータ基盤を確立する
　・これまで主としてオミクス別に統合されてきたデータベース同士を統合し、ヒトとモデル動物の様々なオミクスデータや研究用試料情報を統合的に利活用できるデータベース等を開発する
・新たな知識やイノベーションを生み出すデータベースを構築する－【新たな価値】
　・大規模データ解析や人工知能等も含めた多様な分野の研究者と研究企画段階から密接に連携して潜在的ニーズを発掘し、利用者とともに統合データの利活用に取り組む
　・データベース統合に向けた研究開発を継続し、我が国として国際的に主導的地位を得ることが可能な統合化前の新興領域についても時期を逃さず統合化を支援する
　・データベースやサービスの利用状況をよりよく把握するとともに、今後重要性が増すと考えられる画像やテキストなどの非構造化データへの対応についても引き続き検討する

2.3　事業評価、提言を踏まえた NBDC の活動方針

事業評価および提言を踏まえ、NBDC では、事業方針および具体的な進め方を以下のとおりとした。

＜事業方針＞

1．大型プロジェクト研究と連携し、未公開データのプロジェクト内での共有を支援する

公開前からの連携により、公開しやすく利活用性の高いデータの整備を支援するとともに、大規模かつ統一した様式で利活用しやすいデータの集積を目指す。

2．応用につながる領域に焦点をあて、基礎研究データの統合を行う

AMED や CREST との連携や、ファンディングで開発するデータベース、NBDC ヒトデータベースなどを活かして、利用者からのフィードバックを得つつ、国内外の既存データや知見の統合を目指す。

3．利用者との協業により、統合データの利活用に取り組む

大規模データ解析や人工知能の分野の研究者・機関との共同研究に取り組むほか、ファンディングによる研究開発においても利用者との協業を加速する。これらにより、データ解析・インフォマティクス分野と、生物科学分野の双方の利用者の観点をデータベース整備に反映し、データ利活用による仮説立案とデータ解析を支援する。

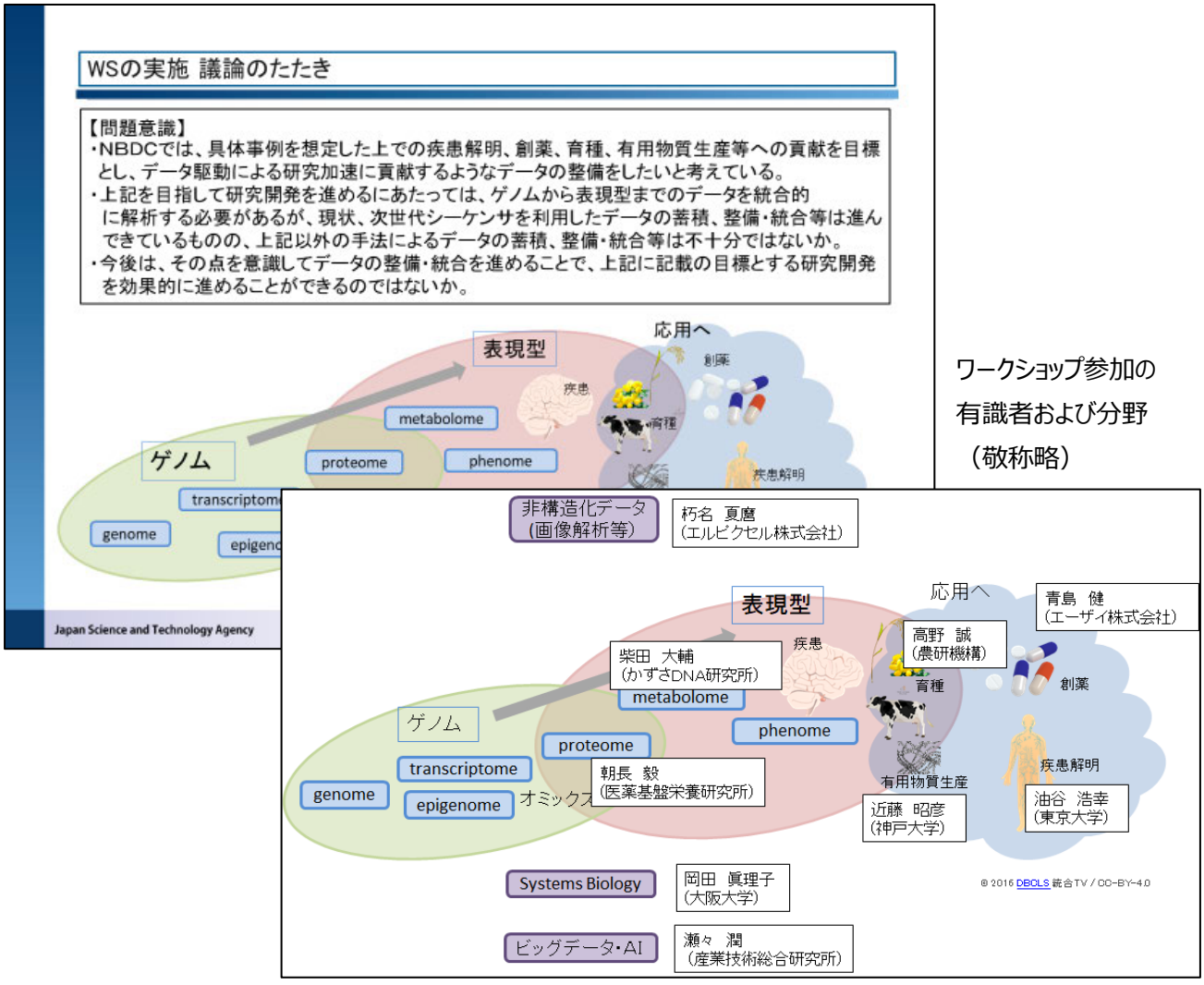
（参考）ライフサイエンス委員会（第 86 回）資料 3
（http://www.lifescience.mext.go.jp/files/pdf/n1934_04-2.pdf）

＜具体的な進め方＞

【目標】	疾患解明、創薬、育種、有用物質生産等への貢献を目標として DB を整備する。
【整備の進め方】	**出口に向けた、目的志向でのデータ統合を目指す。** ・今まで困難であった表現型（機能、病態、形質）の統合を行う。 ・ゲノムから表現型の統合を図る。 ・疾患解明、創薬に向けて、ヒトとモデル生物等のオミクスデータを統合する。 ・育種、有用物質生産に向けて、モデル生物/非モデル生物等のオミクスデータを統合する。 **長期的展望に基づいた DB 構築を企図する。** ・先端的研究の中からも DB 化のニーズを先取りする。 ・画像、動画等の非構造化データの DB 整備を進める。 ・注力すべき新興領域への支援を行う。

3　ワークショップ開催に当たっての議論のたたき

上記のことを踏まえ、NBDC の問題意識を以下のように提示し、ワークショップを実施した。



ワークショップ参加の有識者および分野（敬称略）

ワークショップでは、国内外の DB 整備状況を把握し、NBDC で取り組むべき課題、特に統合化推進プログラムで今後重点的に推進すべき分野・領域を、短期・中期的な時間軸も踏まえて抽出することを目的とした。なお、ワークショップにご参加の有識者の方には、「応用につながる具体事例を想定した際に、どんなデータ（ベース）が不足しているか。また、データ整備を進めるにあたって、どのような点に留意すべきか」。また、「上記のデータ整備を日本で実施する意義、重要性」について、発表をお願いした。

III. 有識者発表

1 医薬品開発におけるビッグデータの活用

エーザイ株式会社 データクリエーションセンター 青島 健

＜発表内容＞

私は製薬会社から来た人間ということで、きょうは医薬品開発におけるビッグデータ、データベースの活用について、創薬現場のニーズについてのお話ができたらいいなと思います。

まずその前に、私自身とそれから私が今所属している組織のご紹介を簡単にさせていただきたいと思います(図 1－2)。皆さんもご存じのとおり、最近薬の開発は非常に成功度が低くて、しかもお金がかかるということで、どうにかしてこの成功度を上げていく必要があります。製薬会社の中ではさまざまなデータ、ゲノミクスのデータ、薬理のデータ、化合物のデータ、リアルワールドデータがあり、これらのデータをこれまでそれぞれの部署で解析とか活用をしてきました。けれども、やはり限界はあるということで、1 年半ぐらい前に、要は探索からマーケティングまでのデータを横串でビッグデータとして全部活用し、さらに人工知能も活用して価値を生み出していくという趣旨で hhc データクリエーションセンターが作られました。私自身も創薬研究、探索研究でつくばに 10 年ぐらい、また、3 年くらいの臨床開発の経験からもやはり上流から下流までのデータ、情報を見渡せないと、成功確度の高い創薬はできないと思っています。 このように前臨床、臨床研究、生産、販売、市販後調査など創薬のプロセスの中で、さまざまなデータベースを我々は既に活用しています(図 1－3)。これは非常に多岐にわたりますので、きょうは主にコホート研究とリアルワールドデータについて、私たちはどういうふうに活用しているかについて説明したいと思います。

これはいわゆる次世代医療基盤法といい、今年の 5 月に公布されていて 1 年以内で施行される予定です(図 1－4)。基本的にはレセプトとか電子カルテのデータをきちんと活用できるようにしていきましょうという趣旨です。実はこの法律の前に、個人情報保護法の改正がありまして、ゲノムも個人識別符号に当たるなど、なかなか簡単に電子カルテのデータが使えないと聞いております。これに対して、国が指定した機関で匿名化されれば、我々製薬会社とか大学の研究機関も使えるようにしていきましょうという法律になっています。 直近の動きとしてはナショナルデータベース（NDB）が公開されました。特定健診データも含めたレセプトデータです。またこれは、PMDA 主導の MID－NET で、製薬会社からの薬の副作用の報告と電子カルテをつないで活用していきましょうというもので、我々もこのデータベースに注目しています。

NDB を使って何をするかですが、イメージできるように、これは糖尿病予備群の例です（図 1－5）。皆さんご自分の HbA1c の値はご存じですか。実は 6.5 以上は糖尿病になりますが、5.6～6.5 というのは糖尿病予備群と言われています。このスライドは特診に入っている方で、全国都道府県の HbA1c の値分布になっています。東京都、神奈川県は人口が多いですけれども予備群も多い。この統計だと 825 万人ですけれども、実際は全国で 1000 万人ぐらいの糖尿病予備群はいると言われています。

そこで経産省の研究班では、予備群に対して介入あり、介入なしで有意に違う結果が得られている。介入あり群では 0.56、HbA1c の値が改善されている。こういったデータを使って予防につながるという活用例です。

我々はこういったデータを使って、薬づくりにも活用しています（図 1－6）。創薬、開発、申請、上市、市販後まで、特に臨床開発、プロトコルの最適化とかまた申請のところに関しても、医療経済、差別化、薬価の交渉、さまざまなところでこういった医療ビッグデータを活用しています。

医療ビックデータは結構いろんな種類がありまして、レセプトデータ、調剤薬局のデータベース、DPC 病院つまり急性期の病院の電子カルテ、コホート研究などあり、特徴、利用可能性もそれぞれです。例えばコホート研究のところに”○”が多

い、このようなデータ（コホート研究）は私たちは非常に使いたいと思っています（図 1－7）。こういったバイオバンク、コホート研究のデータを NBDC が統合し、製薬会社の研究者が利用できるようにしていくことは重要で、我々は関心を持っています（図 1－8）。

アメリカのベンチャーはさまざまなバイオバンクを連携するインターフェースをつくっています。たとえば、このバイオバンクのこの検体、このデータが欲しいと選択していくと契約できるような、スピーディにできる仕組みになっています。この仕組みを NBDC によって実現できるか考えていただきたいです。

もう一例、これはリアルな悩みですけれども今の薬には、いろんな名前、様々なコードがあって、お互いに全然互換性もなく、非常に使いづらい状態です（図 1－9）。現在各社でそれぞれ対応表を作っていますが、それこそ NBDC でこういったデータベースを用意していただけると非常に助かります。

またこれから、たとえばアップルウォッチのようなウエアブルデバイスからライフログが大量に生成されます。こういったデータもこれからどういうふうに整理かつデータベース化していくか重要で、予防とか薬による治療経過を観測するなどに役に立てる重要な分野の 1 つだと思います（図 1－10）。

たとえば、先日のネットニュースで、アメリカの男性がアップルウォッチをつけていて異常な心拍数が何回も出たところで、病院に行ったら、肺血栓と診断され、命を救われたというような実例があります。こういったウエアラブルデバイスから出てきたライフログをこれから NBDC 主導でどういうふうに統合していくか、個人的に興味を持っているところです。

あまり時間がないので最後にこれまでの話をまとめさせていただきます（図 1－11）。ジェノタイプ、フェノタイプ型のデータベースは創薬研究に不可欠で、いかに各種バイオバンク、コホート研究のデータを統合していくかというのは、私は非常に重要だと思っています。

リアルワールドデータに関しては探索から市販後までの非常に幅広い分野で利用できますので、関連分野と連携して高品質なデータベースをぜひ構築していただきたい。我々もリアルワールドデータを活用し始めていますが、医薬品コードとか医薬品の名前の正規化は至急整備していただけたらと考えています。

最後に先ほどお話ししたような、ウエアラブルデバイス、IoT などリアルタイムなライフログは治療効果のモニタリング、病気の予防に非常に有用です。今後、こういったデータの収集とかデータベース構築を今から準備しておく必要があるのではないかと思います。

図 1－1

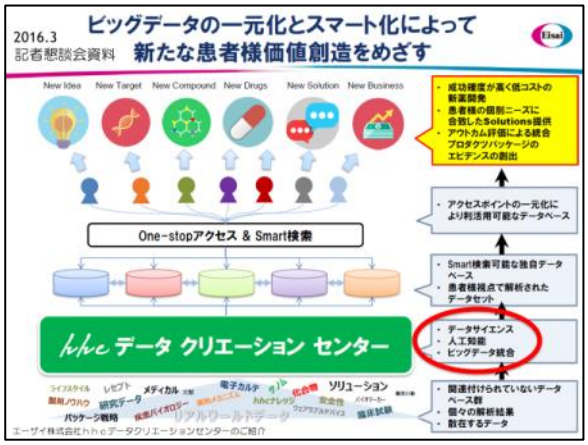


図 1－2

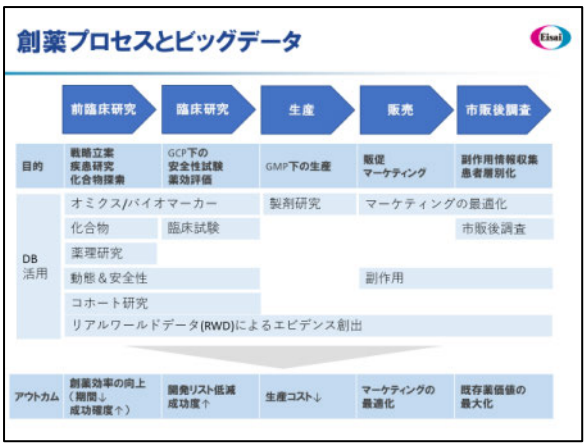


図１－３

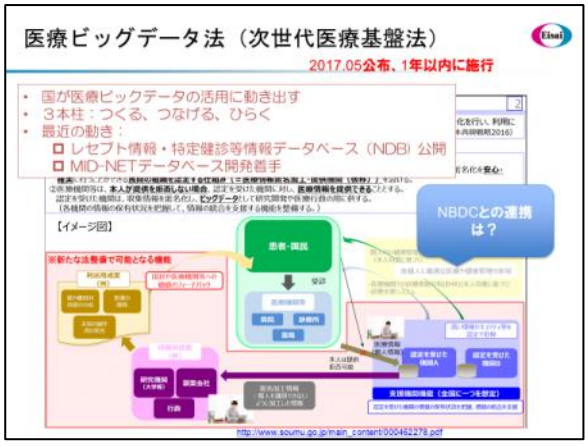


図１－４

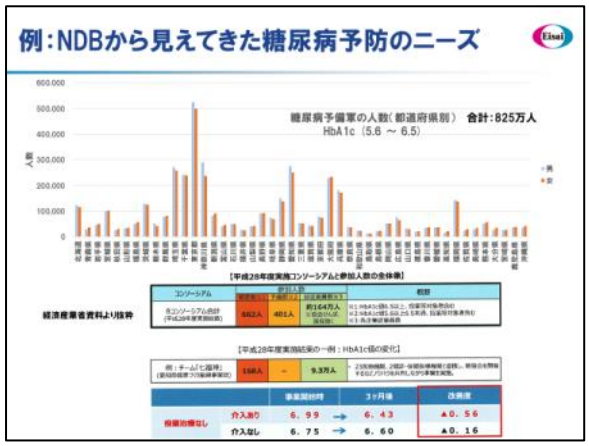


図１－５

図１－６

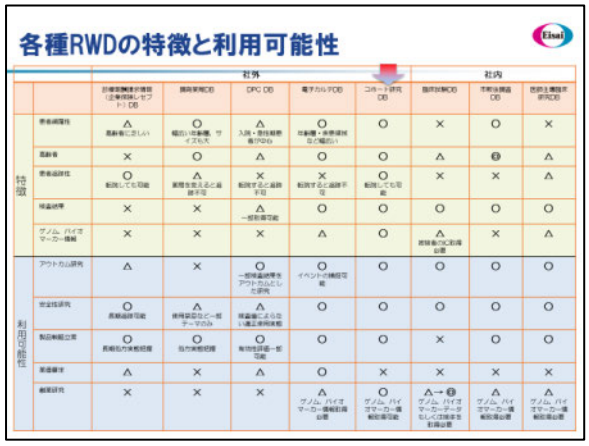


図１－７

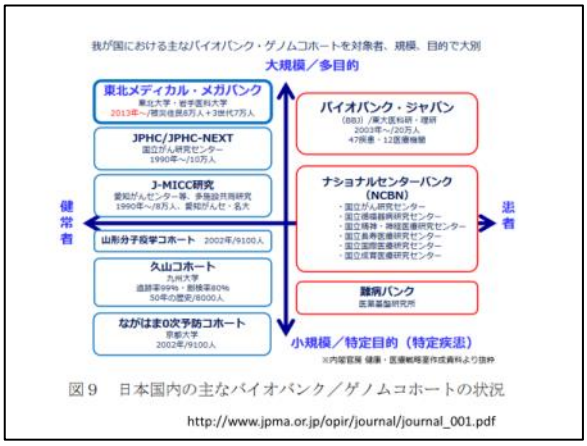


図１－８

図１－９

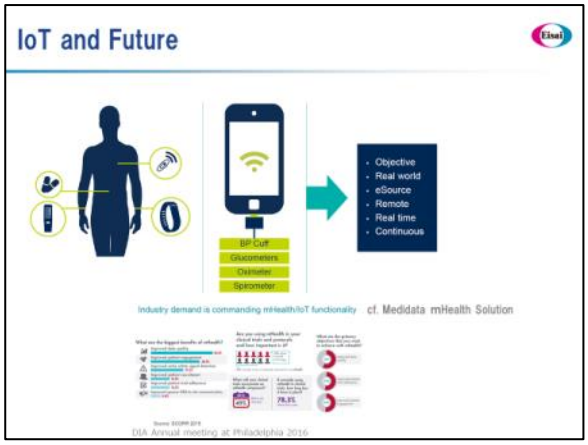


図１－１０

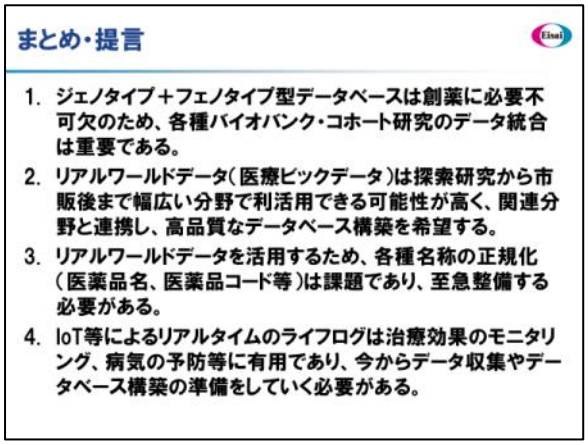


図１－１１

＜質疑応答＞

（質問）ウエアラブルデバイスの話と、ゲノムの話をつなぐという点に関して、公共データに期待をするのか、それとも自分のところでとっていくのか、その辺はどういうバランスになっているか。

（回答）公共のデータはボリュームのところ、またバイアスをなくすというところに期待している。独自で収集するデータに関しては、非常によくコントロールされているが、集団は限られている。両方をうまく活用していきたいと考えている。

（質問）コホート研究をつないでいくベンチャーの件、どういうやり方か興味がある。例えばベンチャーはユーザからつなぐことに関してお金を得ているのか等、全体の構造はどのようになっているのか。

（回答）ベンチャーとバイオバンクの契約形態は不明だが、アメリカのさまざまなバイオバンクについて公共機関も含めて契約をしているようだ。そこにある情報全体からユーザが見たい情報を得られる仕組みになっている（もちろんインフォームドコンセントの関係等で出せないデータもある）。その後の個別機関等との情報のやりとりもベンチャーが仲介している（ベンチャーはこの部分をビジネスとしている）。個別のバイオバンクとの契約等には非常に時間もかかるし労力もかかるので、このような仕組みはユーザとして非常に便利。

（質問）ゲノム情報と結びつかないリアルワールドのデータは、どのように利用しているのか。

（回答）例えば薬剤の使われ方とか、アンメットメディカルニーズとか、治療の実態等に関しては非常に参考になる。その他マーケティング戦略等にも結構活用している。

2　ライフサイエンス基盤としてのゲノムクラウド

東京大学　油谷　浩幸

＜発表内容＞

病気の疾病の解析ということでお題をいただきました。実際自分たちが収集・利用しているデータはゲノムとトランスクリプトーム、あるいはエピゲノムデータでありまして、データの肥大化という点に関連づけてお話します。本日の資料作成にあったては、彼らにも情報提供してもらっています（図２－２）。

世の中にデータベースはいっぱいあって、日本のゲノムコホートのデータからがんのデータや GTEx、最近論文も出ましたけれども、データベースごとに立派なサイズがあって、それをまたつないでいこうというようなことが今求められていると思います（図 ２－３）。 解析コストが下がってデータ量が上がったことによりリソースを提供する研究プロジェクトは増えています。新たなシーケンサーである NovaSeq が日本でも今稼働し始めていますけれども、1 人の全ゲノム解析が 10 万円、RNA も解析できるプラットフォームということで HiSeqX10、X5 から見ると非常に使い勝手がよく、ユーザから見るとハイスループットのデータをさらに出せる一方で、それらをフルに活用していく基盤が日本で非常に脆弱なことは、これまで何度も指摘されているところでありますが、もう本当にパッチワークでやっていてはどうしようもない状況です（図２－４）。

何年か前にアメリカ、海外でもこういうビッグサイエンスを進める際に、拠点に集約した場合、若い研究者たちをどうやってモチベートしようかということが結構議論されていました（図 ２－５）。ただし、生物学は物理学とは異なって、マシンラーニングなどの解析をしても、それぞれのサブジェクトごとにデータの解析方法が違う（図２－６）。

ちょっと前の Nature からとったレビューです（図 ２－７）。それぞれのプロジェクトごとにアナリシスが違うというようなことがユニークである、Routinely Unique という点において、データが大きいといっても先ほどの物理学のようなものとは違うというところが生物学の特徴です。

先ほどのキーワードに出てきましたが、ビックデータとか AI とかを回そうとしても、個別の情報の解析の手法が違うというところが 1 つの問題。昨今話題になるデータの解析の再現性については、実験系であればデータの写真をちゃんと撮っておくなり、証拠を残せるんですけれども、情報解析というのは、それぞれの人間が行った解析ログは保存されておらず、同じシーケンスデータを複数の人間が扱って違う結果が出てくるということは、決して珍しいことではありません。

今自分が感じていることとしては、情報解析の再現性と、データが巨大になっているということでデータが動かせないこと、この 2 つが本日申し上げたいことになります。データサイエンスからビッグデータ、これも 2014 年のマップですけれども、クラウドコンピューティングなどのテクノロジーはアベイラブルに普通のビジネスになっていますので、そういうものをどんどんアカデミアで動かしていくことを加速していただければなと思います（図２－８～11）。

つまりデータ解析の再現性に関しては、Docker とか Notebook とかビジネスのほうで出てきたクラウドのテクノロジーを使って、データ解析の再現性を高める必要があります。1000 人のゲノムをデータベースからダウンロードしようということになると、それだけで普通の研究室ではできなくなります。使用したいデータベースに自分の解析パッケージを持っていけば解析ができて、他人とも解析データを比べられるということが自由にできるようにしないと、小さい研究室や学生の場合にはデータそのものになかなかアクセスができないことになります。もちろん、データ管理を行う環境整備ヒトのデータの臨床倫理の問題とかあります。

山中さん（現オラクル）は、Pitagora-Galaxy プロジェクトを日本の中で DBCLS の方々と一緒に進めてきています。Galaxy 自体はずっと Penn 大が中核になって進めてきていてうちの研究所でも 2 年ほど前にワークショップを開催しています（図２－12～16）。

いろいろな解析プラットフォームを Galaxy の上で動かすようにしようという試みはよいのですが、ツールができてもそれぞれ

のユーザーの計算機資源には限界があるというところでなかなか Galaxy のプラットフォームを広めることが難しいようです。1 つのソリューションはクラウド上で開発したパッケージが動くようにするということかなと思います。

繰り返しになりますけれども、データのダウンロードはほとんど不可能ということです。ICGC では 2700 例くらいのがんゲノムの Whole Genome データを解析するために、Broad Institute や Sanger Centre やドイツが中心になって全世界で解析をしてきたわけですけれども、非常に苦労も多かったです。

一方で、クラウド自体が今までコストが高いというイメージがありましたけれども、相当に安くなってきた。海外のいろいろな疾患、自閉症だけで Google Genomics に 7000 人のゲノムデータベースがあったりとか、そういう個別のデータベースが多数できており、それらをつなげる必要があります。クラウドビジネスはほかの分野でも SE の需要が非常に高いので、彼らをバイオに呼び込むのはなかなか大変かもしれません。バイオ分野に安定的な職がないと人材が入ってこないというのも、本日のテーマ選定とは違うかもしれませんが、優秀な人材を呼び込むためには安定的な組織構築が必要かと思います。

クラウドの構築については、西村さん（テンクー）がまとめてくれています。コンセプトとして様々なデータの種類があって、生データから 2 次的なデータ、もちろん医療データなどをつなぐ場合に、ヒトの疾患データは共有の手法が課題です。公共ゲノムデータと独自データ、あるいは製薬企業さんが治験の際に収集したデータもぜひここに入れていただき、データへアクセスできる研究者を増やすことが非常に大事だと思います（図 2－17～22）。

最初はコンピュータをとにかくクラウド上に置きますよと。今、日本ではゲノム情報をクラウドにおくのは危険だとかいわれますが、アメリカの NIH のデータベースアカウントをつくる際に Google のアカウントを使うことでセキュリティを担保しており、クラウドはむしろ安全ということで、まずこの辺の今までの認識を全体的に改めていかなくてはいけない。

まずデータをクラウドに置く。そして最終的には、先ほど言ったようなツールをのせて解析を共有化するというようなステップで進んでいく必要があります。現在のように個別のデータバンクに VPN の回線をつないで共同研究のデータを見せてもらうのでは研究は進みようがないかなと思います。

米国では既に、Findable, Accessible, Interoperable and Re-usable、FAIR というコンセプトのもとにさまざまなデータセンターがあります。先ほど言った Google Genomics などの大手が乗り出していて、いずれデータを共有化したときに、メリットを受けるのはこの辺の会社になるかもしれませんし、日本として一番弱いところかなと思います。

アカデミックのクラウドである SiNET5 については、クラウド利用促進機構が全国の研究機関をつなぐ仕組みを構築していて、その中にクラウド接続サービスが昨年ぐらいから動き始めていると聞いています(図 2－23)。けれどもどういうデータをそこに載せていくかというところで NBDC とうまく連携をして、ヒトのゲノムや公共オミックスデータに加えて独自データをそこに載せて使っていく仕組みをつくることが重要です。

EU の European Open Science Cloud というプロジェクトでは、ゲノミクスだけでなくて全てのゲノムデータ、バイオロジーのデータも載せていこうという、非常にチャレンジングなものです。推進する人たちも本当に実現できるのだろうかと思いながらも進めているアクティビティと聞いています。本年 6 月に開催した国際ゲノム会議で来日された Ewan Birney (EBI) らが中核となって進めているものだと思います（図 2－24）。

ゲノムデータは国境を越えないというのは、最近の風潮になってきている一方、各国のデータをつなぐ Global Alliance for Genomic Health には NBDC も参加されていると聞いていますし、解析パッケージをつくるなどのアクティビティをどんどん強めていっていただきたい。がんゲノムデータの場合は、統合データベースあるいは TCGA/ICGC にデータを納めていますが、殆どの研究者はそれらを個別のサーバーにダウンロードして解析するわけです。今までの 50 とか 100 症例程度のサイズの情報であれば研究室のサーバーや各大学にあるスパコンで解析することが可能だったわけですが、それ以上になるとパンクしてしまうし、コストも時間もかかる。例えば、EGA から最大 10 ギガぐらいの回線でダウンロードしても、100 症例のがんのゲノムデータをダウンロードしようとすると数週間かかるということで、研究のスピードの上でもボトルネックになっています。

さらに新たにヒトゲノムや疾患を研究したいという研究者がいたとしても、データへのアクセスや計算環境が整わなければそれだけでも参入はやめておこうということになってしまうのではないかと思います（図２－25～30）。

辻さん（東大先端研）が推奨する Notebook 形式のプログラムは元来 Python 言語で開発されていますが、ブラウザ上でメモの作成ができるツールとして利用者が増えています。実行結果を記録しながら、データの分析作業を進められるので実験系研究者にとっても比較的アクセシブルなテクノロジーかと思います（図２－31～34）。

いまだに、見えないものにお金を払いたくないというのが日本人の何となく性のような気もいたしますけれども、データの保管、計算機資源から人材、どれもないないづくしかもしれませんけれども、ライフサイエンスのインフラとしてのゲノムクラウドの整備をぜひ、先ほどのような全国のネットワークと NBDC との間でお進めいただけないかなと思います（図２－35）。

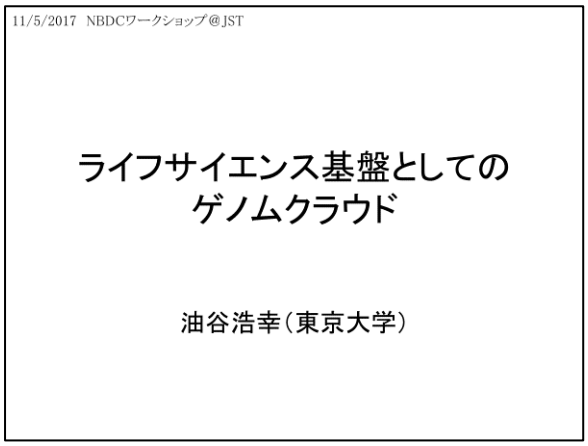


図２－１

図２－２

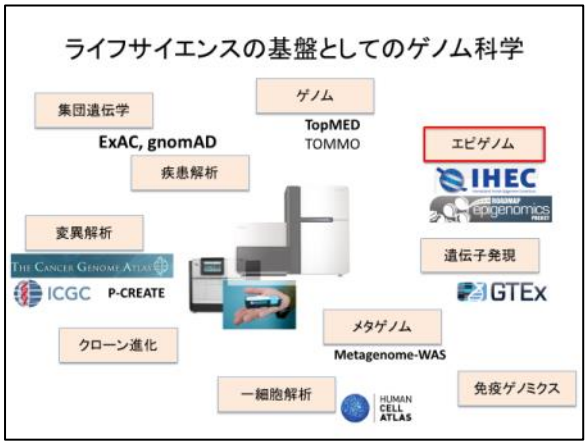


図２－３

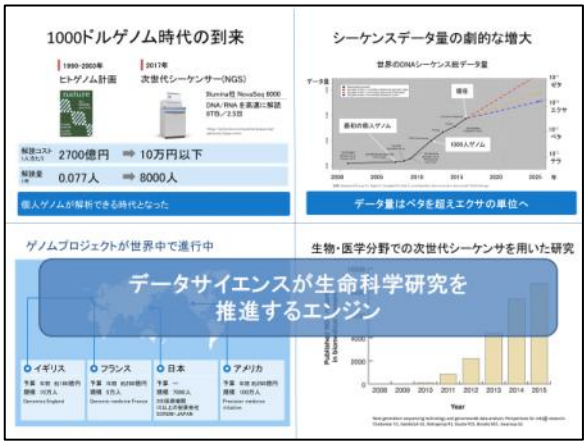


図２－４

ビッグサイエンス?

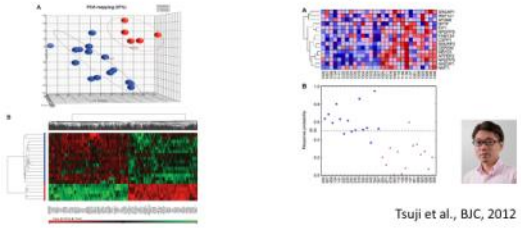
- 物理学
  - Fermi National Accelerator Laboratory
  - Spring-8
  - カミオカンデ
- 生物学
  - 大規模ゲノムセンター(ヒトゲノム計画)

?

図 2－5

What is machine learning?

- Unsupervised learning
  - 教師無し学習
  - Principal component analysis
  - Cluster analysis
- Supervised learning
  - 教師あり学習
  - Prediction model for FOLFOX (anti-cancer drug) therapy efficacy



Tsuji et al., BJC, 2012

図 2－6

データを基盤とするサイエンス

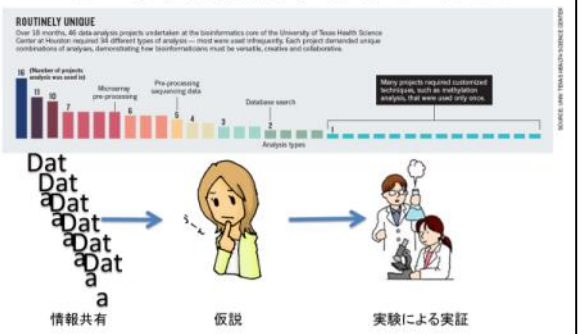
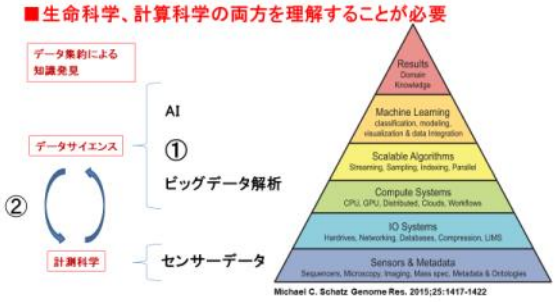


図 2－7

■クラウド技術を活用した大規模解析が必要
■生命科学、計算科学の両方を理解することが必要



Michael C. Schatz Genome Res. 2015;25:1417-1422

① データサイエンスの停滞　　(専用アルゴリズムの不足)
② フィードバック による品質管理

図 2－8

Gartner's 2014 Hype Cycle for Emerging Technologies Maps

http://www.gartner.com/newsroom/id/2819918

図 2－9

データ共有・解析環境

1. データ解析
   - 解析ツールを共有できる→データ分析の再現性
     - Docker
     - Jupyter Notebook
   - リンクされた付随データから必要な情報だけを利用
2. データを動かすことの困難
   - 時間と手間がかかる(1000ゲノムを超えると非現実的?)
   - データ管理を行う環境整備(倫理承認を含めて)
   - 臨床情報など付随データを一緒に動かすことは不可能

→ゲノムクラウドの整備が急務

図 2－10

データ解析再現性

- データ解析（結果）の再現性
  - ツールのコンテナ化
- データ解析環境（構築）の再現性
  - 解析環境の構築コストが大きい
  - 大規模化

図 2－11

Galaxy

Penn Stateのメンバーを中心としたGalaxyチームにより開発されている

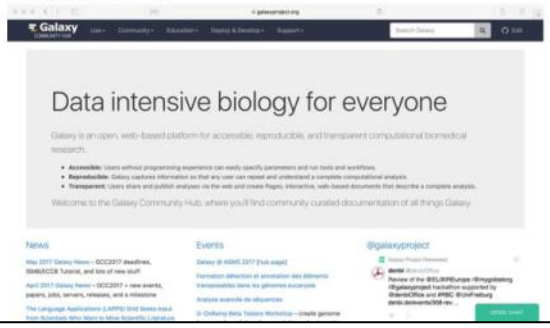


図 2－12

Galaxy Community Conferences

In 2010 National Science Foundation has initiated yearly gathering of Galaxy users and developers by proving the project with a grant supplement. Since then this gathering grew into an annual event known as the Galaxy Community Conference (GCC). GCC alternates between US and Europe and provides a forum for sharing knowledge and building collaborations.

Event/Year	Location	# Registered
2010	Cold Spring Harbor, New York	69
2011	Lunteren, The Netherlands	148
2012	Chicago, Illinois	203
2013	Oslo, Norway	219
2014	Baltimore, Maryland, United States	214
2015	Norwich, United Kingdom	230
2016	Bloomington, Indiana	206
2017	Montpellier, France	
2018	Portland, Oregon (joint conference with BOSC)	

図 2－13

図 2－14

Time	Title	Speaker	Workflow	Slide	Video
09:30-10:00		(Registration)	–	–	–
10:00-12:00	**Preparation Hands-on:** Installing Pitagora-Galaxy	Tazro Ohta (DBCLS, ROIS)	–	pdf	video
12:00-13:00		(Lunch)	–	–	–
Time	**Title**	**Speaker**	**Workflow**	**Slides**	**Videos**
12:30-13:00		(Registration)	–	–	–
13:00-13:05		Opening	–	–	–
13:05-13:25	**Opening Remarks:**	Hiroyuki Aburatani (RCAST, Univ of Tokyo)	–	–	video
13:25-13:30	**Message from Galaxy Team**	Dave Clements (Johns Hopkins University)	–	pdf	video
13:30-13:45	Introduction of Community Galaxy	Ryota Yamanaka (RCAST, Univ of Tokyo)	–	pdf	video
13:45-14:00	Workflows: RNA-seq (1)	Mika Yoshimura (ACCC, RIKEN)	workflow	pdf	video
14:00-14:15	Workflows: RNA-seq (2)	Shinji Nakaoka (IMS, RIKEN)	workflow	pdf	video
14:15-14:30	Workflows: CAGE	Makoto Nasuno (ASCADE)	workflow	pdf	video
14:30-14:45		(Coffee 1)	–	–	–
14:45-15:00	Workflows: ChIP-seq	Ryo Nakaki (RCAST, Univ of Tokyo)	workflow	pdf	video
15:00-15:15	Workflows: 8S-seq	Yutaka Saito (CBRC, AIST)	workflow	pdf	video
15:15-15:30	Workflows: Variant Calling (1)	Norio Shinkai (CBRC, AIST)	workflow	pdf	video
15:30-15:45	Workflows: Variant Calling (2)	Makoto Ikeda (Percipere)	workflow	pdf	video
15:45-16:00	Workflows: Variant Calling (3)	Hideki Nagasaki (NIG, ROIS)	workflow	pdf	video
16:00-16:15		(Coffee 2)	–	–	–
16:15-16:55	**Keynote 2 (Overseas Galaxy Project):** Genomics Virtual Laboratory	Andrew Lonie (Life Sciences Computation Centre, VLSCI)	–	pdf	video
16:55-17:10	Infrastructure: Galaxy on AWS Cloud	Yuichi Yoshiara (Amazon Data Service)	–	pdf	video
17:10-17:25	Infrastructure: Galaxy on PC Clusters	Tomohiro Uchida (NABE International)	–	pdf	video
17:25-17:40	Platforms: Illumina BaseSpace	Eri Kibukawa (Illumina)	–	pdf	video
17:40-17:55	Platforms: Garuda Alliance	Nikos Tsorman (The Systems Biology Institute)	–	pdf	video
17:55-18:00		Closing	–	–	–

https://galaxyproject.org/events/tokyo2015/

図 2－15

クラウド利用の背景

- Big Data 化によりデータのダウンロードは困難に
  -例：TCGAデータ 2.3PB, 1000 genome 230TB
  シーケンスデータは指数関数的に増加している。
- クラウド利用によるより安価な解析。
  -SNSなど他産業でのBIGDATA利用が進みクラウド利用は安価に
- コンテナ/仮想化技術の標準化が進みどこでも同じ解析が可能に
  - Docker, Hadoop など

従来：データをダウンロードして解析

今後：データのあるクラウド上で解析

図 2－16

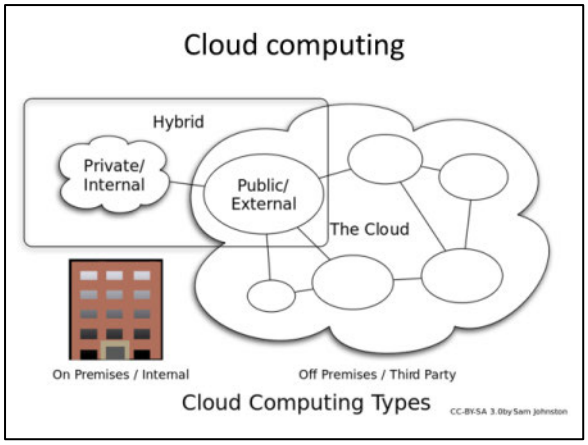


図 2－17

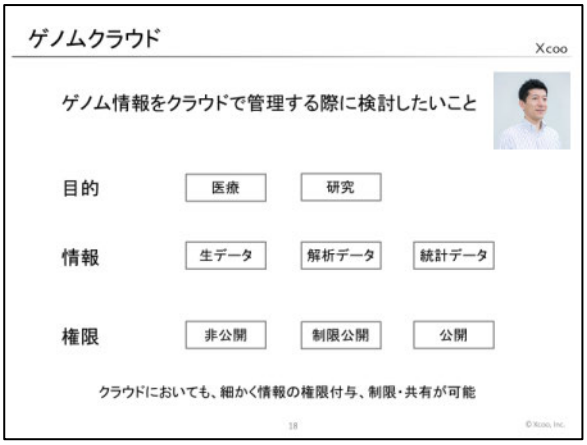


図 2－18

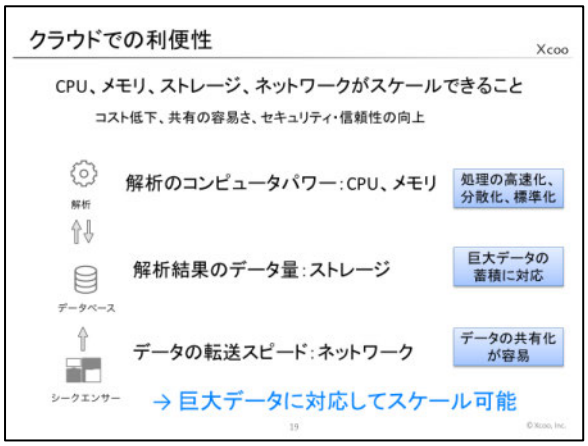


図 2－19

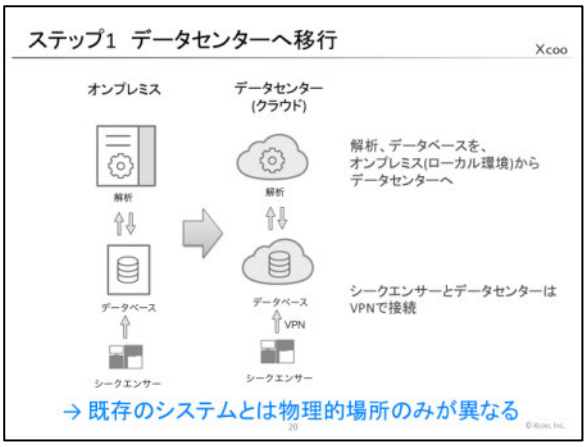


図 2－20

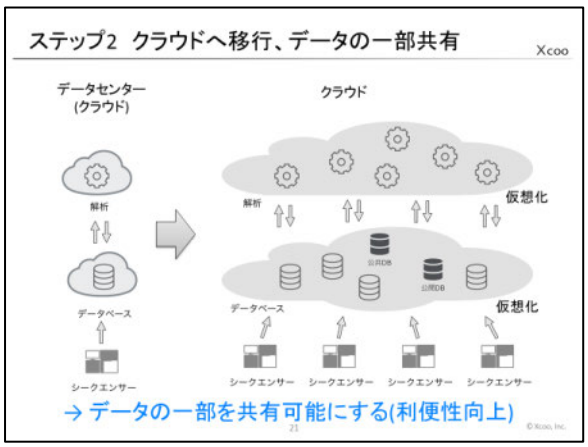


図 2－21

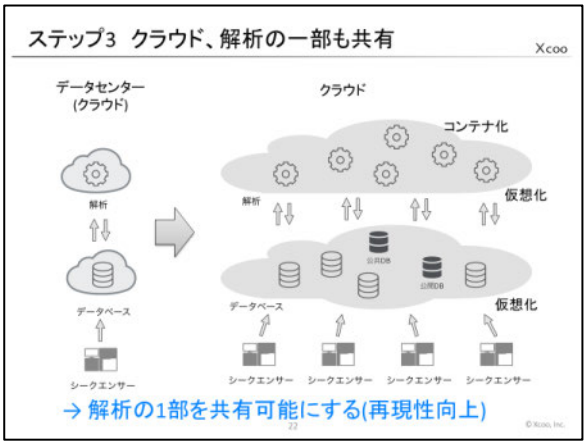


図 2－22

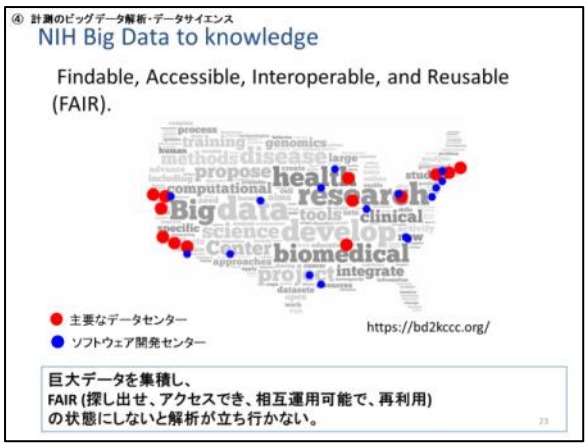


図 2－23

図 2－24

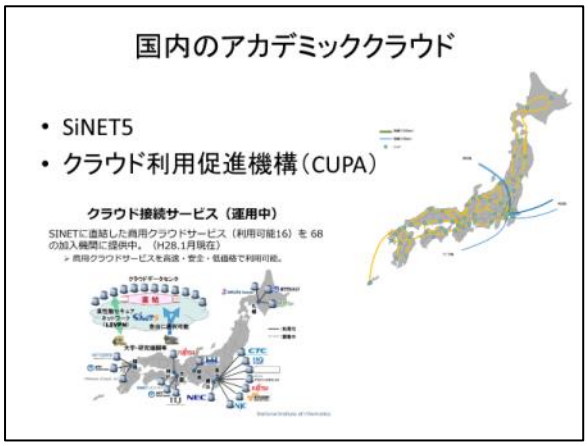


図 2－25

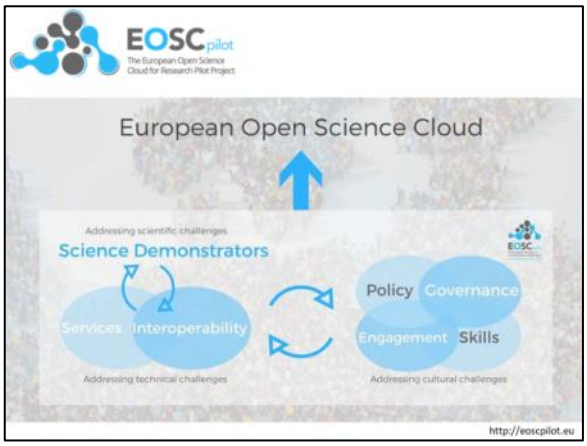


図 2－26

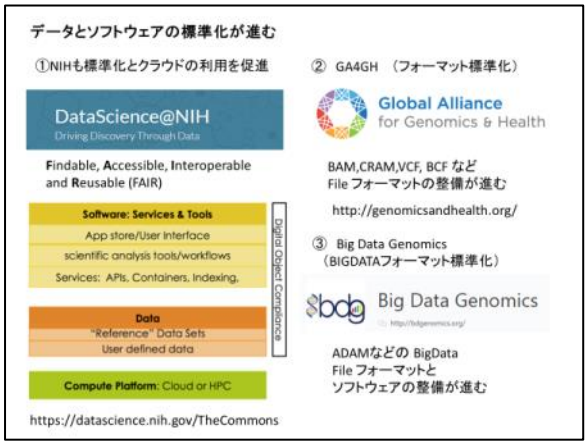


図 2－27

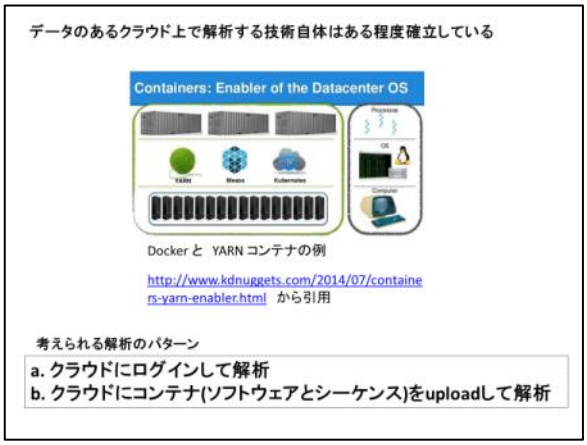


図 2－28

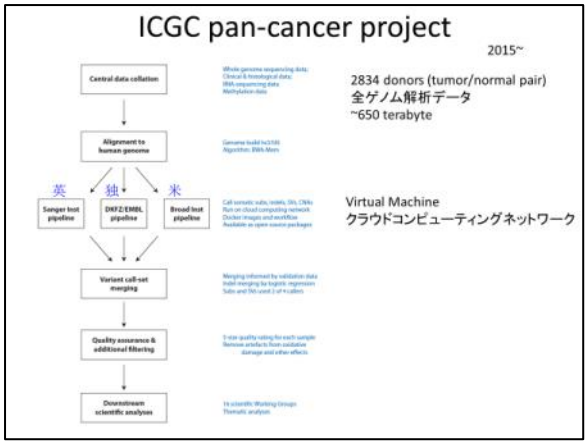


図 2－29

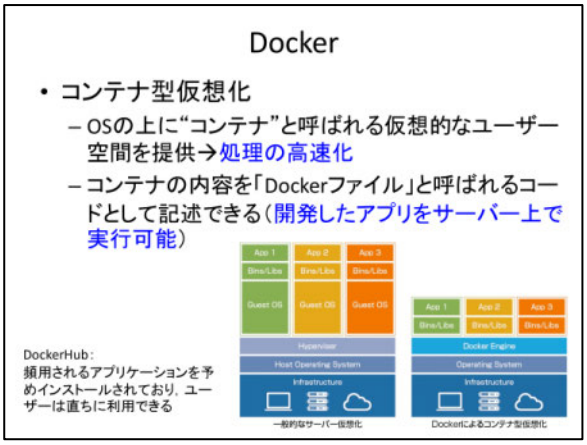


図 2－30

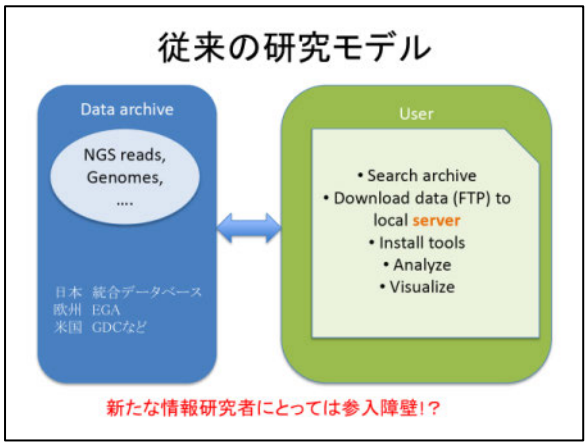


図 2－31

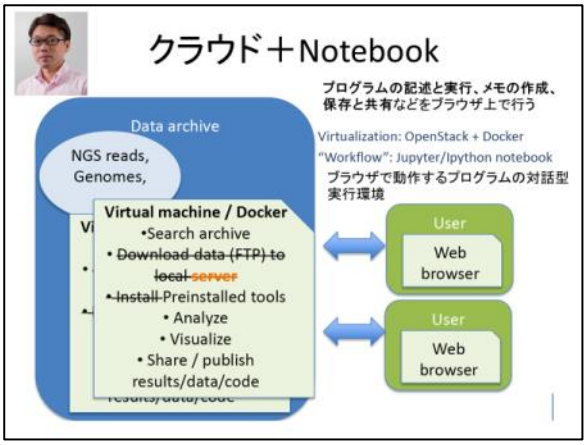


図 2－32

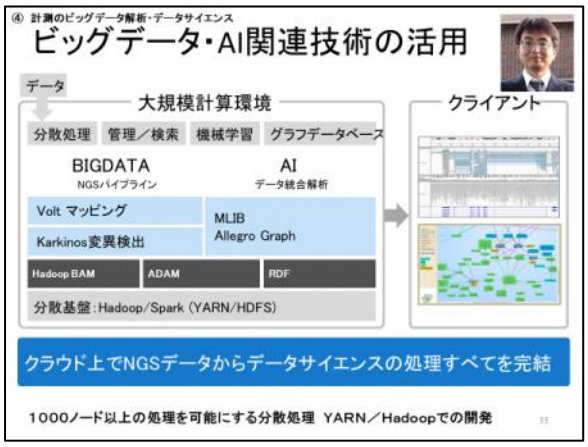


図 2－33

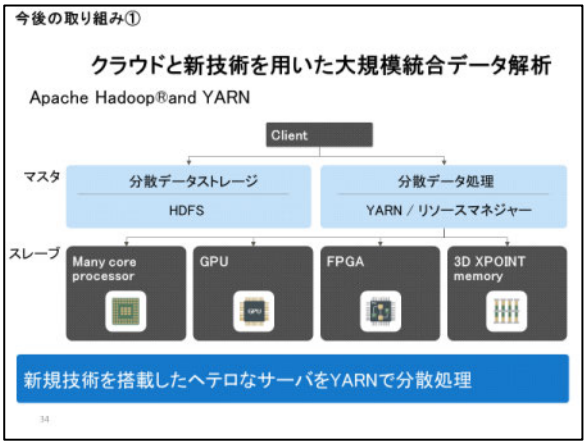


図 2－34

図 2－35

＜質疑応答＞

（質問）実際バイオの人を教育してクラウドを使ってもらうというのは、結構ハードルが高いのではないか。

（回答）ツールもある程度パッケージ化していかないと利用は難しいと思う。そろそろ情報解析のところも標準化というか、あるいは再現性という視点が、実験と同じように必要になってくるのではないか。

（質問）EOSC のとりくみについて、また、個人ゲノムデータのシェアリングについてもう少しお話伺いたい。

（回答）EOSC はパイロットプロジェクトとして動いている。クラウドでデータをシェアしていく際の、ポリシー作り等にあたって、例えば倫理の問題等を議論するためのコミュニティを作っている。ヨーロッパの国はそれぞれ小さいから、当然シェアしないといけないし、EU のファンディングもいくつかの国でという仕組みになっているということも背景にあるかと。個人ゲノムデータのシェアリングについては、診療あるいは医療情報の部分は、アクセス制限がついていたとしても、誰でも申請すれば見られる形にするのには抵抗があるだろうと思う。ICGC-MED 等では、その次のフェーズをどうするかという議論は続いているけれど、臨床情報が国境を超えるというのは、難しいと思う。

（質問）データのシェアにあたってだが、自然にデータが集まってくるかというと、日本の場合、アメリカに比べるとデータシェアリングポリシーが若干弱いように思うがどのようにお考えか。

（回答）次世代がん（文部科学省　次世代がん研究シーズ戦略的育成プログラム）では、NBDC とも協力して、プログラム終了後 2 年でデータを全部出すという方向で進めてきたので、データは全部出るはず。ただ、TCGA などのデータは、病理所見などの臨床情報がそこそこついているが、日本だとそこまではシェアリングされていないのが現状。例えばTCGA では、データセット毎に論文を出しているが、あくまでもそれはリソースとしてのデータ紹介に近いと思う。純粋なリサーチではなくて、リソース提供に対して研究費をファンドしているというところを明確にして、リソースを出した研究者には、アクティビティを続けられるようにしていくしくみが必要。それで出さなかったら次のファンディングはないと、ファンディング機関がきちんと言っていくのがいいのではないか。なお、データの登録、ID を合わせる等にも、結構手間がかかるので、そのようなところも、プロジェクトのほうで事業開始時点からきちんと対応してもらうように指導する一方、それらのコスト負担も、ファンディングエージェンシーが予め配慮するということが大事になるのではないか。

3　有用物質生産に有効なデータベース整備

神戸大学　近藤　昭彦

＜発表内容＞

私からは今までの医療とは違いまして、工業的ないろんなものづくり、そちらのほうのお題をいただいていますので、そちらで発表させていただきたいと思います。

ここにありますように、今バイオインダストリーは非常に革命の時期に来ておりまして、バイオエコノミーというのが全体のそれをくくる言葉ですけれども、これはバイオテクノロジーを利用した産業、これは近々に 200 兆円を超えていく規模になっていましてその中身は農業、医療、工業生産が 1 対 1 対 1 ぐらいです。3 分の 1 等分ぐらいが各領域のエコノミーのサイズと言われていまして、そういう意味で非常に伸びてきている（図 3－2）。

なぜそれが非常に急速に伸びていくかというと、先ほどありましたように、DNA のシーケンス技術が非常に急速に発展して、AI や IT 技術が発達しまして、出てきたデータをより解析できるようになりました。そして、ゲノムを合成したり編集したりすることが自在にできるようになりましたので、こういうことを使って新しい医療技術、新しい工業、ものづくりとか新しい農業というのが革新的に進む可能性があるということで、こういったことが今行われようとしています。

スマートセル・バイオインダストリーということで、経済産業省もそういった産業化を進めていますけれども、従来の遺伝子組み換えでは自然界の探索から試行錯誤によっていろんなものづくりを行ってきたわけですけれども、時間がかかる、いつできるかわからない、コストもかかるということで、ビジネス的には要請が強くてもできないことが多かったわけです（図 3－3）。

それを合成バイオにしようということです。探索からスタートしてものづくりをするのではなくて計算機上にありますデータを使って、そこからいろんなものを設計できるようになれば、自然界の探索は時間がかかることでありますので、このデータを使っていろんなものをスタートしていこうということです。計算科学的に自然界にない人工的な代謝経路や遺伝子回路とかいろいろなものが設計できますと、今までつくれなかったものとか、今までできなかった医療というのができるようになるだろうということです。

さらにロボットとか高度解析技術によりまして、そのスピードが飛躍的に上がってきているということであります。そういうことによって人類に必要なもの、健康も含めて全てつくるような形ができるのではないかと期待されているわけです。

しかしながらこういったデジタルからスタートしようとしても、完全なデジタル化は当然できていないわけです（図 3－4）。したがいまして、不完全な情報から、DBTL というのは、Design Build Test Learn、普通の工学的なアプローチですけれども、デザインをして、とにかく不完全であるけれどもデザインしてそれをつくってみて、それをテストしてそこから学ぶという、普通の工学のアプローチです（図 3－5）。これをやってその中で知識を集積することによって、生命の理解とかデジタル化を両方推進していくということが、ポイントになってまいります。

ここにございますように、世界的にもビジネス的にもかなり発展してきている。バイオのイメージが、今まで試験管を人がピペットでやっていたというイメージから、ロボットがいっぱいいてコンピュータでという感じにバイオテクノロジーの研究開発が変化していくということです。いろんな情報を使いましてデザインをして、ものづくりの生命とかをデザインして、それを高速に遺伝子合成したりゲノム編集をしたりするということで、高速につくり込んでそれを 100、1000、10000 というオーダーでつくったものを高速に高精度に評価して、計算どおりにならないわけですけれども、そこの結果から学んでさらによくしていこうということが、DBTL と先ほど言われたもので、これはゲノムデザインサイクルプラットフォームと呼んでおりますけれども、こういう形のものが非常に世界で進んできております。

つまり AI や IoT 技術を使った統合システム、ネットワーク化によりまして、人が手作業でやっていた試行錯誤をコンピュ

ータやロボットが一瞬でやることによりまして、開発スピードや高精度なデータの集積スピードを飛躍的に上げることができているということです。

そうしますとシステム、プラットフォーム内に知識や高精度データが高速に集積していきます。それを使いましてデジタル化を推進すれば、より高い設計精度でいろんなものづくりの細胞とかをデザインできるようになりますので、少ないサイクル数で目的を達成できるということになりますから、どんどんやればやるほど好循環が回っていくことになります。このような形のシステム化というのが、世界で大競争になっているというのが現状であります。こういうところをいかに制していくかということになります。

NEDO のほうでもスマートセルのプロジェクトが行われておりまして、統合システムプラットフォーム化、競争力のあるような先ほどのプラットフォームをつくることによりまして、一例として細胞をつくる時間、工業的に使えるようになるまでの時間を、従来 10 年ぐらいかかっていたやつを 1 年以内でできるようにするとか、そういった具体的な目標を持ってこういったものが進められているわけです（図 3－6）。

そういう意味では今のような全体のシステムプラットフォーム化というのが非常に重要になるということです。そしてこれが、さらに加速していくというようなサイクルになっていくということです。

実際に細胞をつくるとはどういうことかといいますと、ここにありますように微生物のゲノムがわかりますと、どんな代謝経路があるか、全部コンピュータ上で描くことができます。これは生物はつくっていないんですけれども、これをつくろうと思うと、新しい代謝経路をコンピュータ上のデータを使ってシミュレーションをしてつくり出します（図 3－7）。だけどここへ行く炭素の収率が悪いものですから、この辺は要らないというところを見つけ出してそれを切ってここへの収率を上げて、でもどこかにボトルネックがあると全体のスピードが遅いものですから、ボトルネックを見つけてスピードを速くすることによって、高収量でスピードも速く目的のもの、つくれなかったものをつくることが可能になるということを具体的にやっていこうということです。

これは先ほどありました、ゲノムデザインサイクルのプラットフォームをつくろうということで、さっきのようなデザインツールなどを全部コンピュータ上にこういうアイコンのような形で使いやすく整理いたしまして、新しい代謝経路をつくる。こういうことをいろいろと今整理しておりまして、ユーザとのワークフローを自在につくれるような、こういうものをつくったりしてきております（図 3－8）。

こういった開発を進めてきているわけですけれども、非天然の人工代謝経路をつくるってどんなプログラムかというのを、この中で 1 つだけ簡単にご紹介したいと思います（図 3－9）。例えばこの物質からスタートして、あるものをつくりたいというときに、実際にはどこを見ても天然の生物はつくっていないというときに、KEGG のデータベースとか PubChem データベースからデータを持ってまいりまして、いろんな酵素反応を組み合わせることによって、ここからここに行けるようなものを見つけてまいります。ただし、非天然の酵素を入れておかないと、実際生物がつくっているものはつくれないわけです。

そういう意味では非天然の酵素、化学反応は新しくつくれませんけれども、酵素の特異性は変えられるというような制約条件を置きまして、そういうもの入れ込んでやりますと、実際に生物がつくっていないようなものでも、限られた反応ステップでつくることができるようになる。そういうパスを見つけましょうということ。そして見つけたものが、人工酵素をつくるのは非常に今、ボトルネックになっているんですけれども、つくりやすさとか、その代謝経路の収率から、例えばこの 2 番目が一番いいということを選んでやる。

次に実際にデータベースに行きます。そしてどの生物が持っているどのゲノムのどの酵素を使ってそれをつくっていけばいいのかとか、全体をつなげるようなシステムをつくっていくのは非常に重要になるということです。

こういったことをやるにあたって、私たちなりにどういったことをやるのが必要になってくるんだろうかというのを以降、何枚かのスライドで考えてみたのがこちらであります（図 3－10）。

育種、有用物質生産に向けて、モデル生物／非モデル生物のオミックスデータを統合する必要があるということです。一

応データベースはあるんですけれども、ばらつきが多かったり、精度が難しかったり、あるいは学習データにも使いたいんですけれども、そういうことを使おうと思ってもなかなか使えないということがあります。

それから既存のデータベースに情報の付加の必要性があるということです。例えば遺伝子があってもそれはどういうものかとか、酵素遺伝子、宿主情報、ツール情報など、そういったさまざまなデータを付加的に使ってつないでいかないと、なかなか使いにくいということがあります。

より具体的にということで、例えば KEGG のことを考えてみますと、ここにありますように導入する遺伝子の情報と各酵素反応と保有生物の関係（図 3－11）。例えば検索でここの EC 番号からそれを持つ生物を検索できないとか。いろいろ欠けている検索がありますのでこういったことをしていくと、より我々のような分野でも非常に使えるようになるのではないかと考えています。

次に、いろいろな生物をいろいろな形でプロモーターとかをつくっていく分子パーツが必要になります。分子パーツとかのデータベースはないですし、そういった構築の文献とか特許をまとめたデータベースができると、我々として非常に使えるかなと。あるいは、どんなものがつくられているかとか培養条件、合成して生物をつくっていくという観点に基づいた、産業利用に基づいたものはまだ不十分かなという感じで考えております（図 3－12）。

例えば既存の中にも非常にいいところまで、例えばNITEさんがつくっているようなデータベースの例です（図3－13）。こういったものはゲノム情報配列から微生物の機能を推定するデータベースで、目的の機能を持つと推察される生物を検索できたりするんですけれども、まだまだ不十分なところがあります。こういったものをさらに拡充すればいいのではないかというものもございます。

それから全体的なところとして、これはあくまでも私見ですけれども、より明確に利用を意識した統合を進めるべきではないかということで統合範囲です（図 3－14）。国境を越えてどうするかとか、あるいは統合によってどのような検索ができるとよいかということが、本当に考えられてつくられているかということで、特に産業界のユーザを含めた共同開発がきちんとできているかとか。

それから継続的な集積の取組ということで、何とかこういうことができるようにならないかということ、国内のデータベースを公共的に維持する仕組みが必要でしょうと。先ほどもありましたけれども、データはただではないので、信用できる国内のデータベースというのは非常に重要になります。ただ、そこで全部をオープンにするというのは難しいと思いますので、オープン・クローズ戦略で有料のところは有料にしたらいいのではないかということ、どこまでをどうするんだというのを本格的に議論してやっていく必要があるのではないかというふうに考えております。

さらにつけ足しですけれども、Genome Project Write というのが動き出しまして、読む時代から書く時代（図 3－15）。ゲノムをつくるのに 1000 分の 1 のコストでできるようになったら、いろいろどんどんつくって、いろんなことが研究できるようになるといったときに、こういうのをつくるためにもそれに備えたデータベースの構築も必要になるのではないかというふうに考えています。

図 3－1

図 3－2

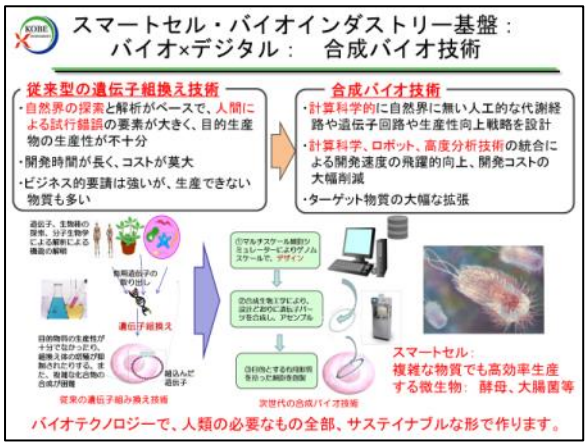


図 3－3

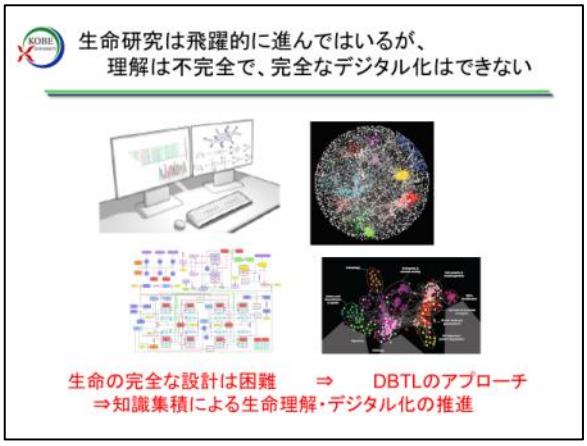


図 3－4

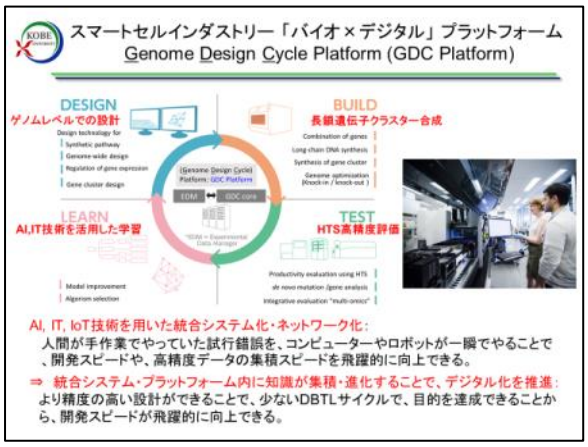


図 3－5

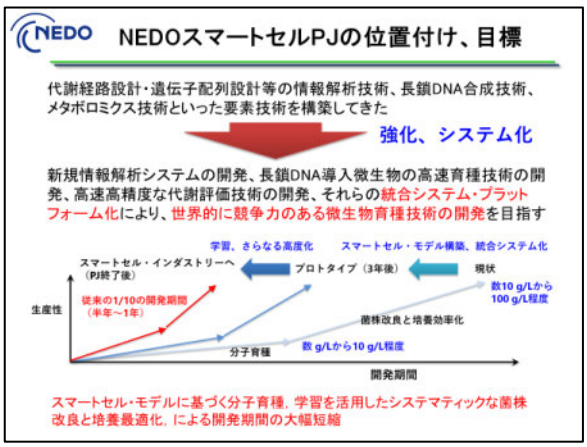


図 3－6

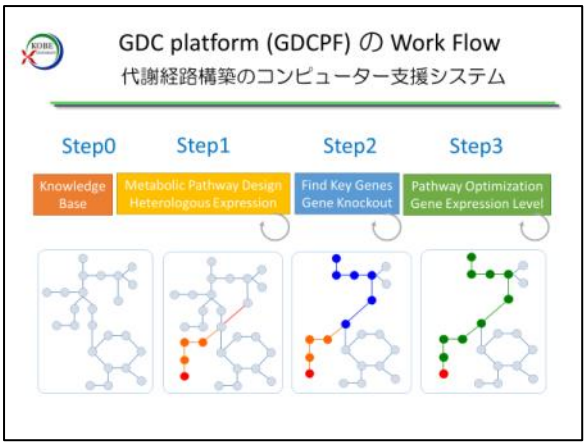


図 3－7

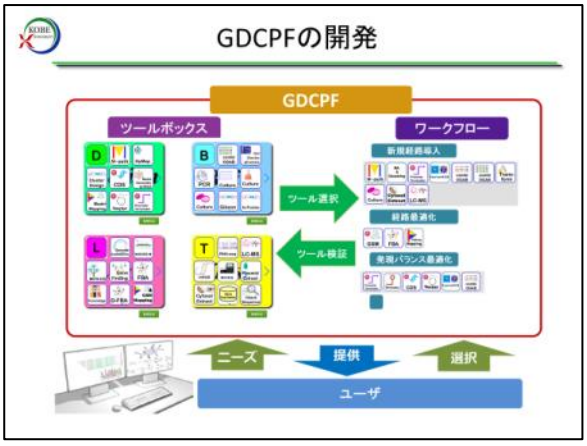


図 3－8

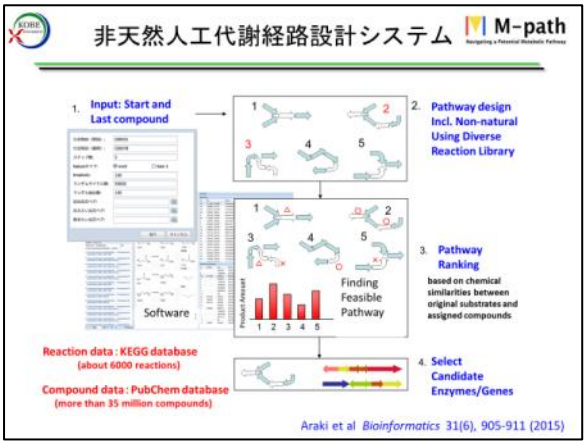


図 3－9

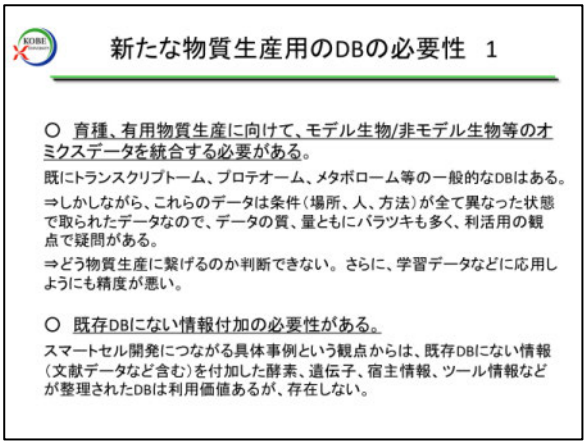


図 3－10

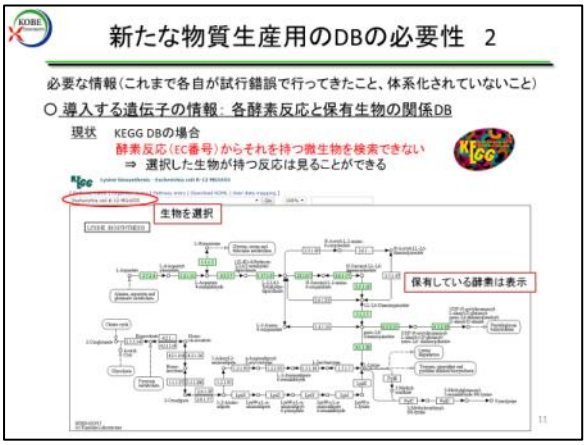


図 3－11

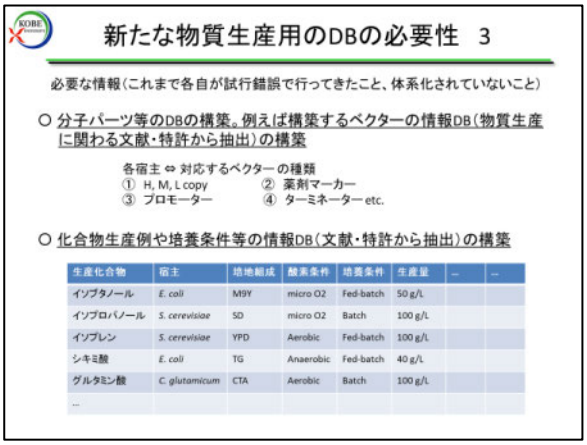


図 3－12

図 3－13

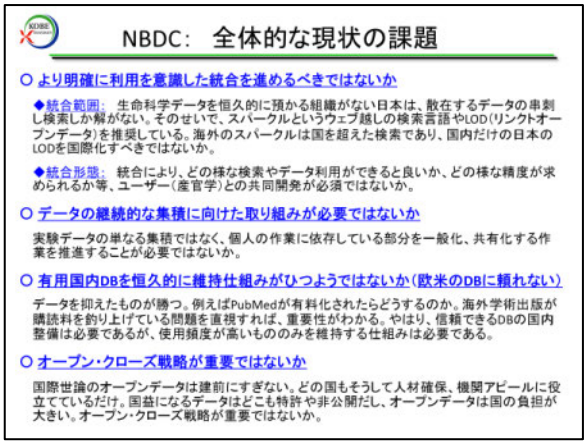


図 3－14

図 3－15

＜質疑応答＞

（質問）「個人の作業に依存している部分を一般化、共有化する」ということについての具体例と、共通化の範囲（目的が異なっても共通化できるか）についてはいかがか。

（回答）例えばどんな微生物のどういう機能を発現させるかを考えた場合、そういったデータが付随したものがないのが現状。そのため、キュレーションするとか、情報を付加するとか、独自に作業をしている。そういうのを全部についてやることはできないと思うが、産業的に有用なものについて、選択的に整備してもらえるとよい。共通化については、医療分野と工業ものづくりの分野では異なると思うが、例えば大きく農業とか工業ものづくり、医療とか分けると、そこではかなり共通した情報があると思う。

（質問）DBTLサイクルを回すというのはとても重要なこと。一方で、結構失敗するので、失敗した例がデータベースの中に残っていると他の研究者も研究を進めることができるかと思うがいかがか。

（回答）実は、そこが DBTL サイクルの一番重要なところ。いろんな機械学習をさせようと思うと、正しいデータだけだと学習にならないので、莫大な量の、いろんな程度のレベルの失敗データを含めて、保証された精度でロボットが実験をしたデータが蓄積され、それが全体としての精度を上げていくことにつながる。それらデータを解析することで、知識ベースが蓄積された予測器としての精度が上がることにつながる。

（質問）GDC プラットフォームは、微生物だとすごくサイクルが速くて、変異を微生物にかませて、コロニーをピックアップし

てそれでアウトプットを調べるということを全部やっていくというのは、結構大変ではないか。

（回答）そのためにロボティクス化が非常に重要になる。ゆくゆくはこういったことをもっと高等生物のほうに広げていく方向。例えば、いろんな酵素をどうやったらより汎用的なアッセイ系で全部分析できるようになるのか、あるいはもうちょっと特異的なバイオセンサーみたいなものがいいのかとか、そういったところもあわせて研究されている。全体としてこのスピードが速くなると、この中に情報が莫大に高速にためられるので、どれだけのものになり得るのか、よりデジタルで記述できるようになるかということが、今まさに挑戦されている。

（質問）製薬会社の中でも、化合物とタンパクとドッキングシミュレーションをやりながらこういうふうに回していくのだが、例えば創薬に関してどこまで汎用性が高いものか。

（回答）基本的な全体構成は変わらないということで、実は日本だけではなくてヨーロッパを初めとしてアメリカももちろん、全てのところに有効。デザインしてロボットさえ増やせば高速にどんなスピードでもできる。それでデータをどんどん蓄積していけばいいわけで、各国が莫大な投資をしてきている。情報系の会社にとっても重要な開発になっており、世界中で、非常にスピードアップして、ベンチャー会社も出てきているという状況。

4　育種に役立つデータベース整備

農業・食品産業技術総合研究機構　高野　誠

＜発表内容＞

私に与えられた課題は、育種に役立つデータベースということで、これからは世界の人口が 2050 年には 100 億人に迫る、あるいは日本で見ると逆に今後どんどん少子化、高年齢化が進んで生産人口が減っていくという中で、食料生産というのは非常にこれから重要になってくるわけです。そのときに品種改良、育種をいかに効率よく進めるかというのがとても大事になりますので、そこにデータベースを整備することによって効率的な育種に役立てたい。

あるコントロールされた中でいろんな代謝経路がわかったものについては、きれいにいろんなことができていますけれども、食料生産、作物等の場合はやはりある意味ブラックボックスということと、周りの環境によって遺伝子発現が変化しますのでその辺も含めたものをつくっていかないと、なかなか役に立つものはできないと考えています。

現在、農研機構のほうでいろんなデータベースをつくってきています。主に作物中心に遺伝子、ゲノムのデータベースそれからイネについては発現データベース、全遺伝子の突然変異を網羅するミュータントパネル、あるいはすべてのイネの品種の特性データベース、こういうもののデータベースをつくってきております（図 4－2）。そのほかにもありますけれども、農水省のデータベースの１つの大きな特徴として、情報があったとしても、それから実際、育種をする場合、材料がないといけないわけですから、そういう“モノ”、遺伝資源についてはジーンバンクがありますし、遺伝解析研究の過程でつくられた研究材料についても整備をしてデータベース化して、“モノ”もきちんとそこに蓄えています（図 4－3）。

そういうふうにいろんなデータベースの作成が、これまでの農水省のプロジェクトの中で進めてきてはいるんですけれども、これが育種に今役立っているかというとなかなかそこまでいっていない。育種の研究者というと、どちらかというと現場に近いところで、従来の交配育種、自分たちの目で見ていいものを選ぶということをやっている中で、情報をうまくそこに入れ込むということは、まだなかなかうまくやり切れていない（図 4－4）。

先端的な育種の方々も、研究者と連携して研究し、そこから得られたゲノム情報、あるいは発現情報をもとに、それを組み込んでいくという形で利用していますが、これを直接育種の方々が使えるような形になっていない。

今はイネのいろんな品種や野生種のゲノム解読が進められ、大量シーケンスのデータが出ているものがなかなか利用され切れていないということもあります（図 4－5）。

最後に一番大事なのは、遺伝子情報がいくらわかっても、それが実際にどういう形質とつながっているかということがわからないと、ターゲットにしにくいということで、この形質のデータベースをどうするかということと、その対応づけがいろいろ求められていると思っています。実際、こういうふうにイネを中心としてコムギも含めて、たくさんの情報が出ているわけです。こういうものをうまく使えば比較ゲノムによって、ゲノムマイニングで有用な自然変異のもとになるような遺伝子が結構すぐにわかったりするんですけれども、それをいかに育種につなげていくかというのはこれからの問題になってくると思います。

ですから育種の分野で今求められていると思うのは、これまでにあるようなゲノム関係のデータベースと関連したようなもの（図 4－6）。それから実際に育種の人たちにとっては、バイオインフォマティクスについて詳しくわかっていなくても、ちゃんと使えるようなデータベースにしなくてはいけないですし、それから最初に言いましたけれども、今いろんな形質のデータベースもできています。それはある意味、研究室の中で温度の条件を一定にしていろんなストレスをかけるとか。非常にデータはとりやすいんですけれども、データをとるために必要なデータベースになってしまっていて、実際に圃場でいろんな植物の発現を見るときに使えるかというと、なかなかそうじゃないということもあります。フィールドでのオミックスデータが必要になってくるというふうに考えています。

ですから全体としては、既存のデータベースがあり、そこにフィールドフェノミクス、フィールドオミックスのデータを加える（図

4－7）。あるいはここに、農水省はどちらかというと作物が中心ですけれども、かずささんが野菜のデータベース等をおつくりですのでそれも含めた上で、作物横断的なオミックス育種データベースをつくると、非常に育種にとって有用ではないか。

農研機構ではこういうのをつくっています。特に形質データにつきましては、実際に広いフィールドでいろんなデータをとっている。そういうのは例えば大学とか民間ではなかなかとりづらいということがありますので、そういうものを含めたデータベースとして、こういう方々に提供することによって、あるいは育種関係者も含めて利用できる研究基盤の整備を進めることによって、こういうデータベースに含まれているいろんなデータをうまく利用して研究が進み、それが実際の品種改良等に結びついていくというふうに考えています。

実際今、農水省のほうではこれまでもそうなんですけれども、プロジェクトとして、これは農研機構の強みですけれども、日本全国にいろんな拠点があります（図 4－8）。作物の場合は、環境との応答がとても大事になりますので、つくばでとったデータだけではなかなか品種改良に使いづらいということで、いろんな地域によってそれぞれの品種についてとった形質データがありますので、そういうものにオミックス情報、あるいは環境データ、そういうものを入れ込むことによって、いろんな地域できちんと使えるようにデータベース化する。

実際のデータを取るところは、農水省のプロジェクトの中で行いますので、このあとを統合データベースの中でうまくデータベース化することによって、ゲノム情報を利用した育種の効率化、あるいは汎用化、どんな人も使えるというものができてくると、育種にとっても非常に有効になるのではないかと考えています。

1 つの例としてこれは、フィールドオミックスです（図 4－9）。実際の田んぼで育てた稲の葉っぱから 1 日に 8 ポイント、それを生育期間ずっと通しながら発現データをとっています。それに対してその時期のいろんな気象のデータ、温度、風速、いろんなものがありますけれども、それらとの相関を調べることによって、それぞれの遺伝子の生育時期、あるいはそのときの 1 日の変動をモデリングしようということが実際やられています。このモデルは割と単純な線形でモデルができるんですけれども、こういうモデルをつくることによって、これはフィールドオミックスデータベースにある程度入っているんですけれども、ある時期にこういう気温になったときに、それぞれの遺伝子がどういう発現をするかということがある程度の正確さでシミュレーションできるようなことができています。

そういうことをすることによって、環境あるいは developmental ないろんなファクターを入れるだけで、そこにおける遺伝子の発現パターンが予測できる。これは今は個別の遺伝子ですけれども、それをあるシグナリングのパスウェイ、あるいはネットワーク、制御遺伝子のこういう prediction に結びつけば、実際にどういうところが育種のターゲットになるかということが明らかになってきますので、そういうところをねらった育種をすることによって、目的の品種を非常に効率よくつくり出すことができるのではないかと考えています。

そのときに、1 ついま我々の中で弱いところは、形質の表現型のデータベースがなかなかないということです。世界的にはフェノタイピングが注目されており、国際的な組織として、International Plant Phenotyping Network ができています（図 4－10）。拠点としてはここにあるようなところです。ヨーロッパが中心ですけれども、北米、中国、オーストラリアにもあるんですが、残念ながら日本にはその拠点はありません。フェノタイプというのは GxE、遺伝子と環境との関係によって、フェノタイプが決まるので、この辺の環境応答をうまく解明しないと、いくらゲノムを解析しても、それは使えないということになります。

もともとはドイツが始めたんですけれども、それから EU に広がっていって、アメリカでもできて、中国も今こういうふうな形で、フェノミクスに対する取り組みを強化しています（図 4－11）。残念ながら日本では、ここは遅れています。こういうところをうまくキャッチアップして、それを育種につなげていく必要があると思っています。

例えばオーストラリアでは、温室の中の非常に精密なフェノタイプ、もっと大きな中で作物、これはモデル生物ですけれども、作物を経時的にずっと連続的にいろんな形質について、単なる表現型だけでなく、光合成能力に関係するような蛍

光とかそういうものを自動的にモニターするようなものができています（図 4－12）。ヨーロッパでも、本部はイギリスにあるんですけれども、こういうところがシステムをつくっています。これはイギリスの Rothamsted Research Institute です（図 4－13）。我々が一番興味があるのはフィールドでの形質で、ドローンとかあるいはこういう Field scanalyzer というものを使うことによって、非常に精密に屋外における形質のデータを取ることができます（図 4－14）。こういうものがないと、なかなか遺伝子との関係というのは精密には結びつかないので、我々としてはこのようなハード面を整備するのは難しいんですけれども、こういうことも含めてデータをとることによって、将来精密で効率的な育種に使えるようなデータベースにしていくことができないかと考えています。

目的としては、今の既存のデータベースに新たにフィールドオミックスデータを加えることによって、日本の中でのこういう基盤を整備することによって、将来的に大学も含めて、育種につながるようなうまいデータベースをつくることが、今後必要になってくるのではないかと考えています。

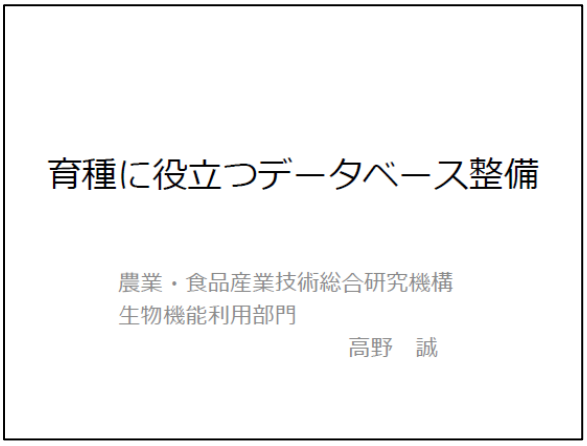


図 4－1

図 4－2

図 4－3

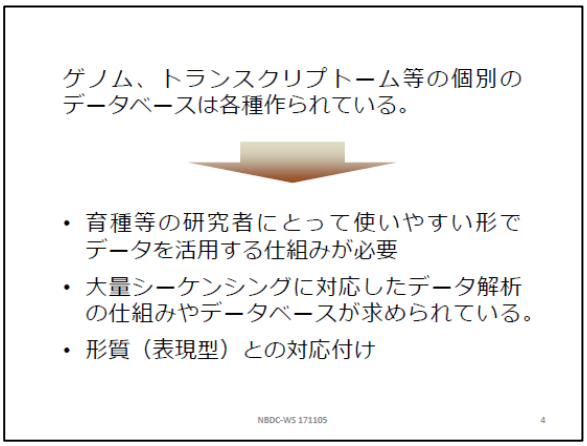


図 4－4

図４－５

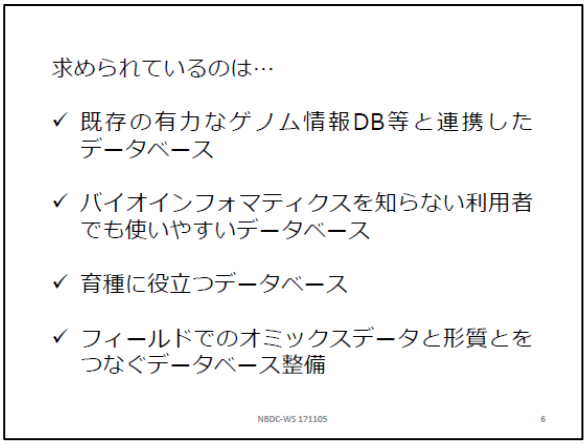


図４－６

図４－７

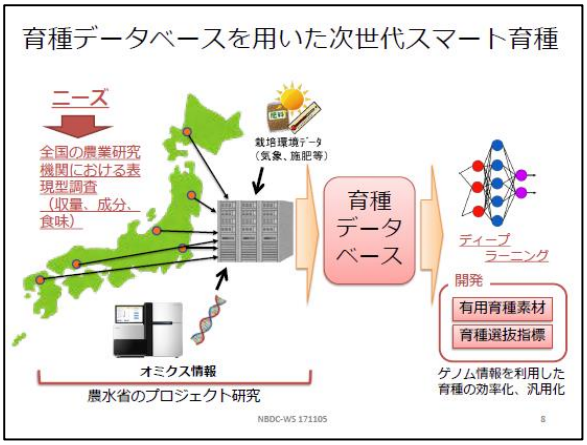


図４－８

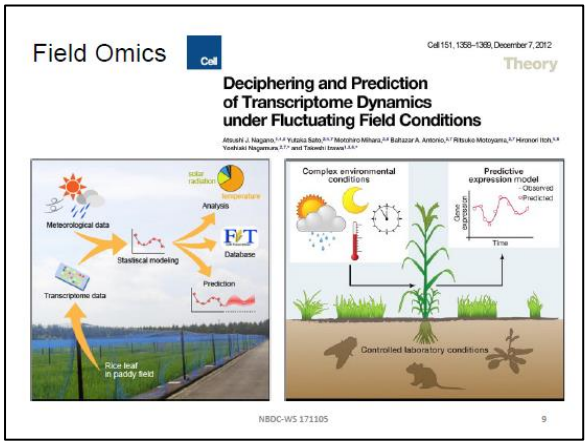


図４－９

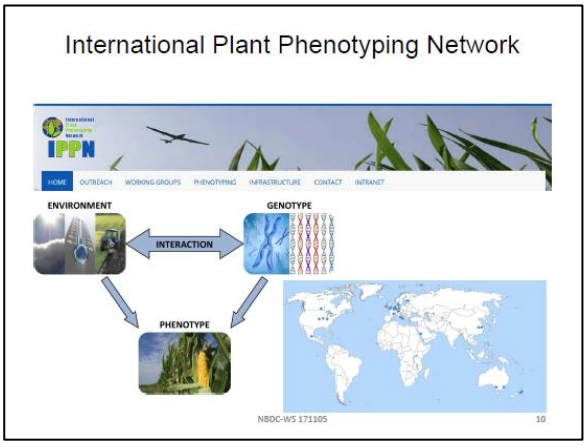


図４－１０

図４－11

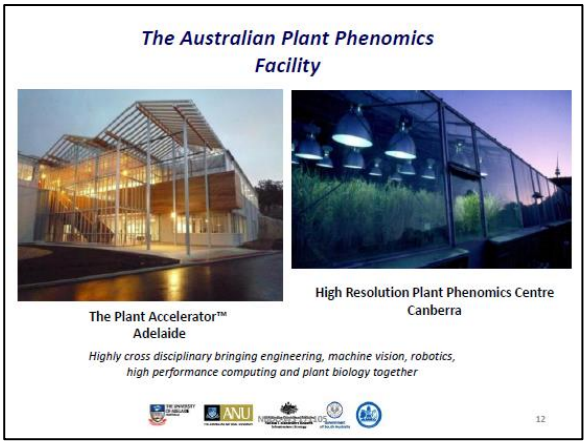


図４－12

図４－13

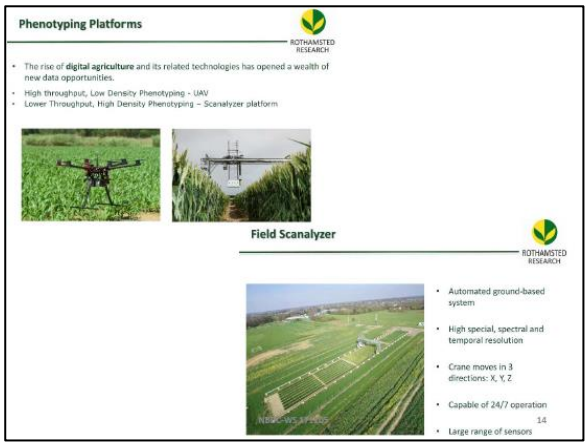


図４－14

＜質疑応答＞

（質問）屋外でのフェノタイプのコレクションに関しては、既存の品種とかそういったものを、例えば気候の違い等によるデータをとっていくのか、それとも遺伝子をノックアウトしてデータを集めているのか。

（回答）両方ある。１つはここのモデルに使ったのは、つくばの中の実験圃場で、ある品種のイネを使って３万ぐらいの遺伝子について１日８回、それから成育期間養苗から登熟までずっとマイクロアレイでのデータの積み重ね。そこで、次は応用に向けて、日本中で、そこまで精密なデータではないが、少なくとも目に見えるデータについて、いろんな育種についてとる。それをうまく組み合わせることによって、環境と生育ステージをある程度インプットすれば、そこでの遺伝子発現等も見えると考えている。そのようなモデルができることで、遺伝子発現データをもとに、ある程度早い段階から、今後イネはどういうふうになっていくか、どういうふうにしないといけないのかという予測が可能になりつつある。

（質問）野外といった場合に、年の変動もあれば１日の変動もあったりいろんな変動があると思うが、こういった環境による表現型のバリエーションというのは、どのようなデータをとっていくのか。

（回答）今回のモデリングに関しては、特にコントロールと比べているわけではなくて、まさにマイクロアレイの莫大なデータと様々な気象データを照らし合わせてみたところ、遺伝子の発現の変化と相関があるのは、日長と気温で、それ以外はほとんど関係ないことが分かった。そういう意味で結構単純なモデルでできている。そうするといわゆる

線形モデルを使って、実際のデータを入れ込んでいって、どういうふうなモデルだと一番それに合うかというところをつくっていくということになる。ただ、それだけだといろんなずれが出てくるので、もうちょっと精密な形質データの利用もいれていくことになるだろうが、大枠としては非常にシンプルなものでいけると考えている。

（質問）発現の場合は遺伝子が決まっているのでデータベースの整備がしやすいと思うが、フェノタイプになったときにフェノタイプの定義はどのようにすればいいか。

（回答）難しいが、例えば生産性を見た場合、光合成能力は例えば、蒸散の能力である程度見えてくるとすると、葉面の温度とかを見る。あるいは光合成で要らなくなったものが、蛍光で出るわけだが、蛍光の状況を見るとか、いくつかの要素に分けて、それぞれをモニターすることによって、ある程度そこで総合的に見ていくようにすると、もうちょっと正確にできるかなと。

（質問）フィールドオミックスなど農水省のほうでデータベースをつくるのではないかと思うが、NBDC にどのようなことを期待するか。

（回答）研究者レベルでは、オミックスを使っていろんな育種に使えるようなことを導き出していくことはできるが、実際品種改良に携わる人たちにとっては、そこはある程度ハードルが高い。そこをデータベース化することによって、ここをこう変えたいときにはどうするかみたいなことがわかるようなインターフェースというか、そういうところがうまく使えるようなものになるといい。それは必ずしも農業関係だけでなくて、例えば大学等の作物関係の研究者にとっても、自分たちが得られないようなデータがそこにあるわけなので、そこをうまく使えるような共通の基盤をつくるということを期待したい。

（質問）米で商品になっているものについては、各品種のゲノムのデータ、フェノタイプのデータはでているか。

（回答）実際に品種ごとのデータももちろんあるが、ただ育種にとっていえば、日本の中の品種はある程度でき上がっているものなので、それをどうこうというよりは、近縁野生種に近い品種を利用することになると思う。今まで日本の中で品種ができる過程で、おいしさを追求するために、例えば、病気に強い遺伝子など、落とされてきたものがいっぱいあるので、そういうものを昔の品種や近縁野生種からとってきて中に入れる、そういう方向にいくということになると思う。

5　メタボロームの視点から

かずさ DNA 研究所　柴田　大輔

＜発表内容＞

私自身は産業界に割と近く、バイオインダストリー協会の理事をやっています。その関係で、産業界の方などから、ご意見を伺うことが多いのですが、NBDC に対しての期待が高く、産業界に配慮した DB 整備に対する期待が高まっています。

メタボロームの話をする前に、我が国を取り巻く経済的な環境とデータ整備の必要性について、少し触れておきます。産業界が直面している課題として、国連が提唱している SDGs の問題があります（図 5－3）。また、気候変動に関するパリ協定、つまり、化石資源からの脱却という観点から、バイオエコノミーという議論が出てきている。世界の機関投資家の間では、環境とか社会とかガバナンスがちゃんとしていないと、もう投資しませんという動きがあります（ESG 投資）。このような世界的な動きから見れば、日本は、今は苦しくなっています。

2009 年に OECD がバイオエコノミー構想を出し、ほぼ全ての先進国は、バイオエコノミー政策を一丸としてやってきています。一方、日本でのバイオ政策は、文部科学省、経産省、厚労省、農林水産省などがばらばらに進めているわけですが、その背景には、2011 年に東北大震災があり、バイオエコノミーを進めるだけの余裕がなかったのでしょう。昨年度、経産省が NEDO に委託事業を出して、バイオエコノミーの調査を行いました。最近、その報告書が NEDO から公表されています。実はこの委員会の委員長を私がやっていました。その中で感じたことは、日本の産業界は、このままで大丈夫かという危機感がものすごく強いことです。とりわけ、欧州委員会 EC が規制と認証制度を始めようとしていることが問題です。この議論をすると時間がかかるので詳しくは話しませんが、SDGs のイメージを経済に取り込み、ヨーロッパ経済をうまく活性化しようという思惑が透けて見えます。正直言って、化石燃料の依存度が高い日本は、極めて不利であるとの状況があります。

このような経済的な状況の中で日本が勝つために、人工知能、IoT の話が出ていると理解すべきでしょう。人工知能は、一般の方は難しいことができるからすごいなという話ですが、そんなことではなく、誰でも簡単にできるからこそ、産業的に大きなインパクトがあるということです。10 月 12 日の総合科学技術・イノベーション会議では、各省庁が集まって、日本が勝てるための政府戦略に関しての議論がされています。その中で、データが肝だ、ということになっているので、NBDC に対して、大きな期待が寄せられています。

話がかなり横道に逸れましたが、メタボロームに話を移します。メタボロームという考え方は、セントラルドグマの議論にとどまるだけでなく、複雑な物質の循環系の議論に移っています（図 5－4、5）。微生物、植物が生産した化合物が動物に入っていって、それがまた循環していくという複雑な循環系が成立しているということです。そこが研究の主戦場になっているとの印象があります（図 5－6）。

生物が生産する化合物がいくらあるかという議論は昔からあって、植物研究者は昔 20 万ぐらいと極めてのんびりしたことを言っていましたが、奈良先端大の研究では、100 万種類は地球上にはあると結論しています（図 5－7）。しかし、それもかなりのアンダーエスティメイトとなっていると思います。植物をつくるだけでもそれの 10 倍以上、1000 万種類ぐらいとかあって、それが化学的な変化を受けたり、食品とか加工とか物理的、化学的な変化が起きていますから、それの結果としてできた化合物が体内循環しています。つまり、数億種類の化合物の世界をどう見ていくかということの認識がまず必要であると思います。

産業競争力懇談会では現在、バイオに関して様々な議論がされています。マテリアルインフォマティクスを使って、バイオでのものつくりをどのように考えるかという議論が 1 つあります。もう 1 つは、食品そのものが膨大な化合物の変換プロセスと

いうことになりますので、例えば、腸内メタボロームの問題、あるいは体内循環という議論が起こります。未病マーカーみたいなもの、例えば、アルツハイマーとか生活習慣病になる前にいかにしてその兆候を検出するかという議論がかなり大きい。機能性食品に関しても、日本の産業ということでいいますと、大体 12 兆円ぐらいです。医療が 7 兆円ぐらいでしょう。農産物は 6 兆円以下です。それに比べると食品はかなり大きいわけです。日本としての産業的な力として十分にある。土壌メタボロームの場合だと、農産物の増産にもつながっていくと思います。

実際、ヒトはどうなっているかというと、ヒトはそもそも遺伝子の数から考えてもそんなにたくさんのものはつくれないわけですが、カナダのデータベースでは、生体で検出される化合物の種類は、11 万になっています。食品として摂取したもの、あるいは作物由来、微生物由来、微生物が変換したもの、あるいは化学的、物理学的に変化したものが、こうやって循環しているということです。これをどのように検出するかが、1 つのポイントになっていると思います（技術的な部分については、図 5－8～12）。

この研究分野では、化合物にアノテーションをどうつけるかが大きな課題です。メタボロームあるいは成分というのは大変難しくて、なかなか簡単に 1 つの方法で解析できないということで、いろんな人たちでやっているんですけれども、日本でやっている人は少ないということもあってなかなか進んでいない。Lipid に関しては、東大の田口先生たちが作ったリピドームというものがあります（図 5－13）。かずさ DNA 研究所では、最近、フラボノイド 7000 種類を全て同定するためのデータベースを作っています（図 5－14、15）。このようなデータベースの作成は、かなり大変でして、これをやるだけでも 7 年ぐらいかかっており、その間に、担当者がお子さんを 2 人生んだというぐらい、長い時間がかかって大変な作業でございました。なかなかやる人がいないので、人材育成という観点でも課題がある。

食品関係の DB の例として、食品メタボロームのレポジトリを紹介します（図 5－16）。いろんな食品を質量分析装置で分析すると、図に示した 2 次元イメージとなります（図 5－17）。液体クロマトグラフィで分けて、こちらがマスの 2 次元イメージです。強度も入れると 3 次元イメージ、実際上は、マスマスパターンとかありますから、n 次元イメージになっています。こういうものをレポジトリとして現在つくっていっているわけです。

結構おもしろいことがわかってきました。例えば分子量が 611.16110 という化合物の配糖体はどこにあるかというのが、網羅的わかります（図 5－18）。最近、医学部附属病院と一緒にやっている例では、尿のメタボロームを解析しています（図 5－19）。入院したとき食べる流動食に関して、少人数で解析を進めたのですが、体内循環している成分は、個人によっても違うし、男女の性別によっても随分違います。共通したものもあるけれども、それ以外のもの、いろいろ見えてくるわけです。こういう中から最終的には何かのマーカーを見つけるということで、先ほどの議論でいくと、未病マーカーみたいなものです。将来的には患者さんを入れてデータベース化するということになるかなと思います。

最後にまとめますと、産業的利用の場面が多いにもかかわらず、成分とかメタボロームデータに対するデータベースの整備がかなり不足しているのが現状でしょう（図 5－20）。その原因としては、分析方法がばらばらだということもありますし、もう 1 つは、質量分析装置は 1 億円とか 2 億円ぐらいするために、誰もが簡単に解析できないという事情もあります。ですから集中的にどこかでやらないとなかなか難しいだろうと思います。NBDC の支援に実験は含めないということになっているので、これをどうするんだという話にもなります。

食品メタボロームレポジトリというのは、公開されたものは世界的には見当たりません。「健康に生きる」ということを考えた場合、このようなレポジトリを産業的に利用することは、かなり重要かと思います。農業生産の問題、バイオマスの問題は、今回、話すだけでの時間はありませんでしたが、最初に紹介したような世界的な動向と絡めて、戦略的にメタボローム、成分データを産業で活用することが、重要になってくると思います。

図５－１

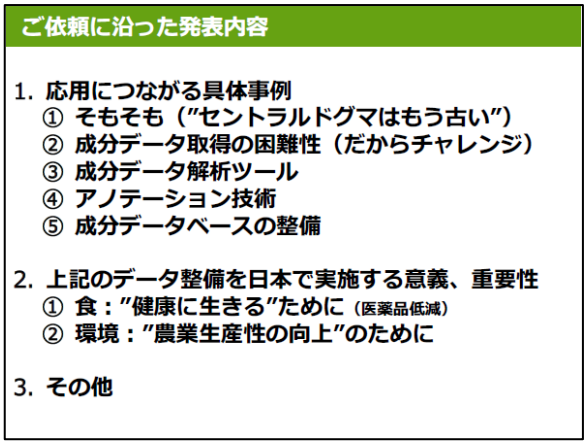


図５－２

図５－３

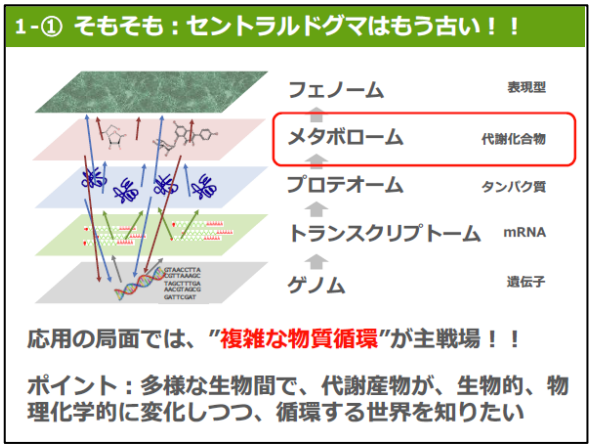


図５－４

図５－５

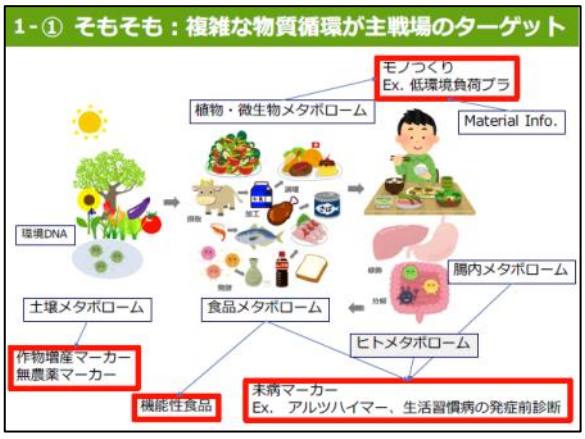


図５－６

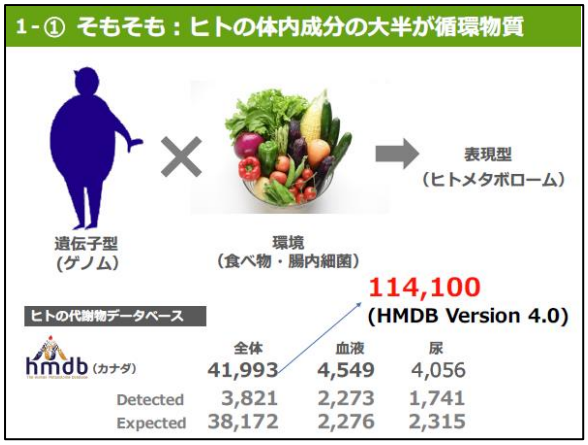


図 5－7

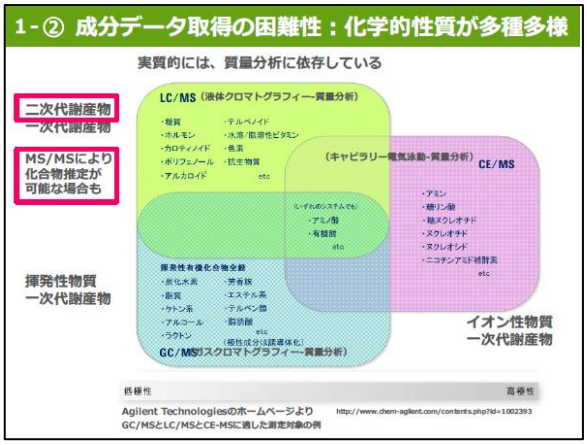


図 5－8

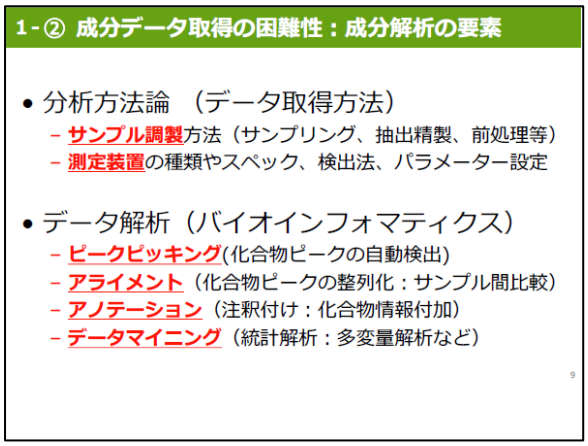


図 5－9

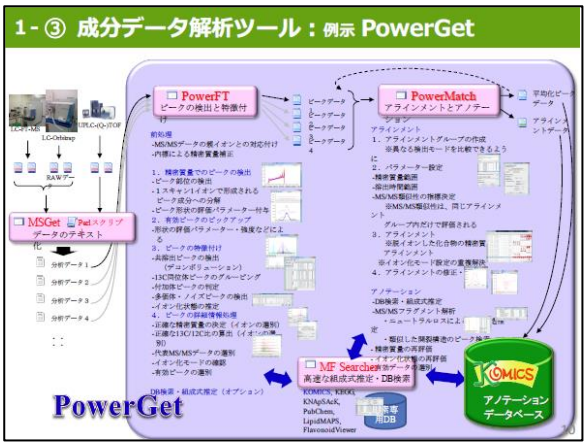


図 5－10

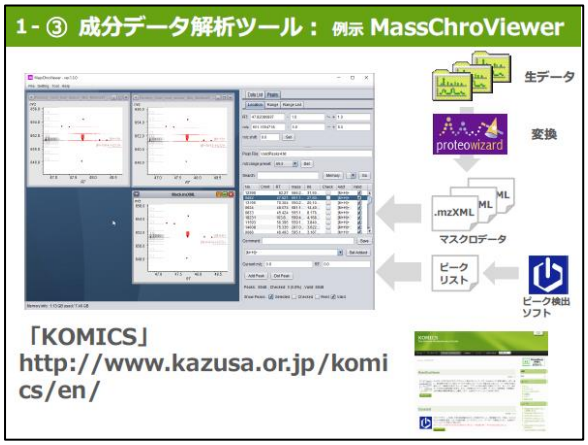


図 5－11

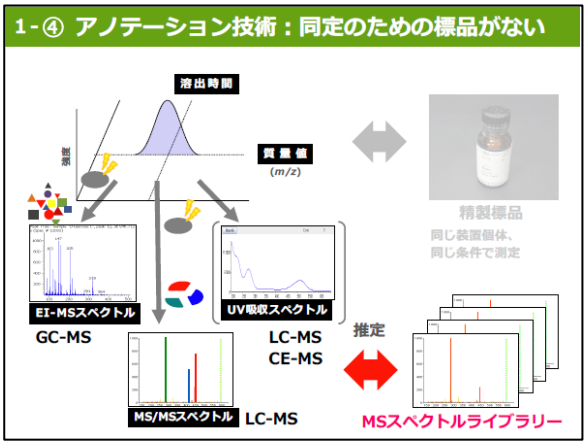


図 5－12

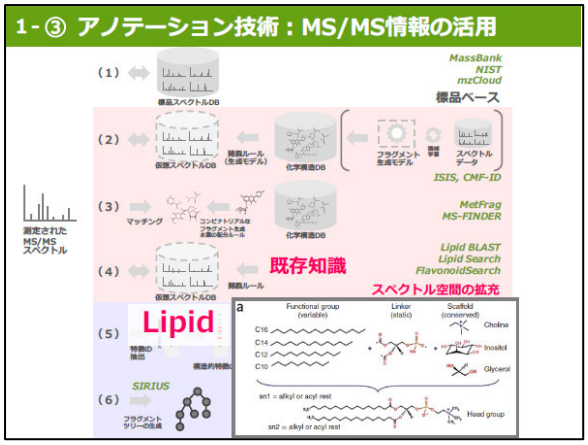


図 5－13

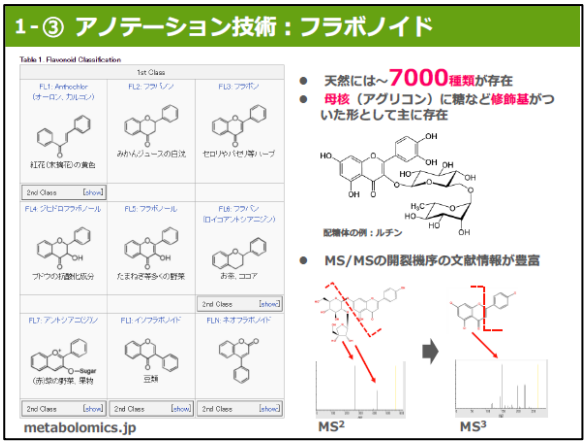


図 5－14

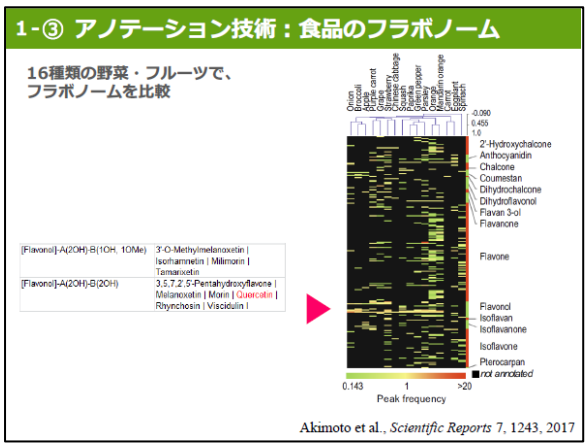


図 5－15

図 5－16

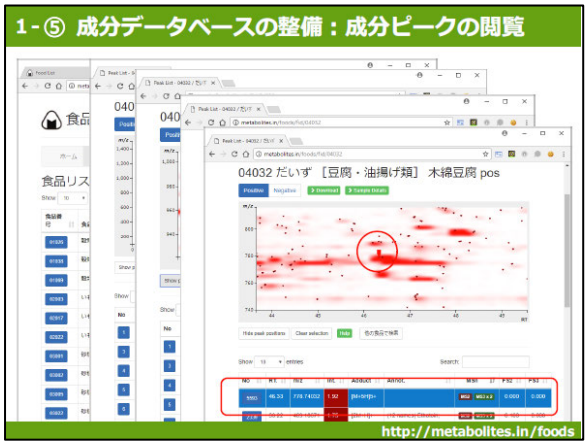


図 5－17

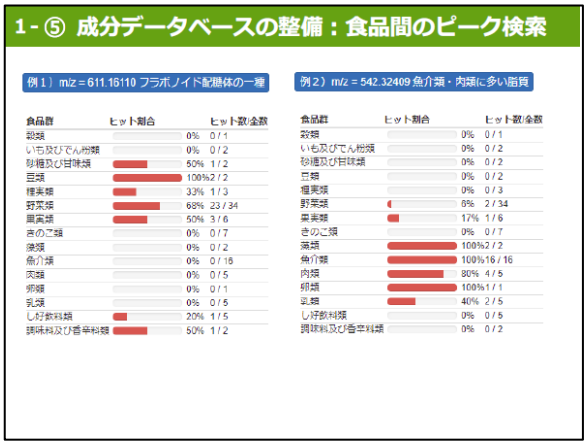


図 5－18

図 5－19

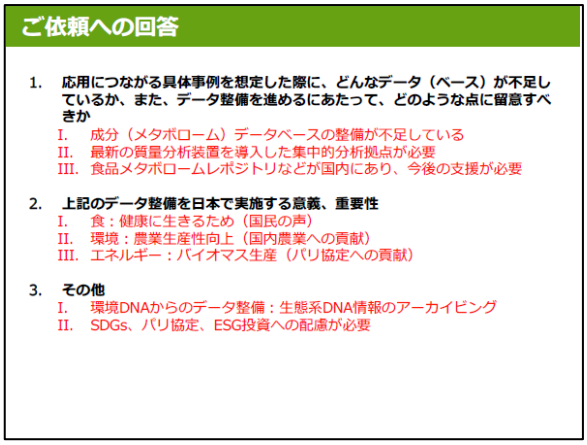


図 5－20

＜質疑応答＞

（質問）メタボロームのデータは不安定なものだと思うが、データベースに付加情報（測定機器や、食品だと作られてからの時間、その他）は入っているか。

（回答）研究室では基本的にはそれは全部入れている。メタボロノートという、メタデータのためのデータベースがあるので、そこに全て入れている。データは、RDF 形式になっている。

（質問）メタボロームは不安定ということについてだが、今の状況でいろんなところでいろんな装置を使っていろんなプロトコルでとられているデータを集めてきても、あまり意味がないということになるのか。

（回答）意味がないことはないと思う。最終的には分子量の正確な質量数が知りたいので、精密質量をはかろうとしている。例えば小数点 1 桁ぐらいまでいけば、組成式が大体推定できるし、小数点 4 桁くらいまでいくと、原子がいくつかという議論ができる。そのレベルまでないと実質的に化合物のキャラクターを特定できないが、大半のデータはそうなっていない。ただ逆にいえば、こういうところを整備すれば、それから逆算し候補が出てくる。トップダウン的にセンター化して、かなり質の高いデータを維持するということが多分一番大事なことになってくると思うが、できる研究機関は世界的にみてもものすごく少ない。逆に言えば日本が取り組めば、勝てることは十分あり得る。

（質問）食品のメタボロームをデータとしてとる場合に、定量値はとれるのか。

（回答）質量分析計の本質的な問題として、定量性が悪いということがある。また、メタボロームというのは、いわゆる標品が 100 万数のうち、数百しかないのが現状。

6　プロテオミクスデータベースの必要性と今後の方向性

医薬基盤・健康・栄養研究所　朝長　毅

＜発表内容＞

今日私に与えられたのは、プロテオームということなので、プロテオミクスの研究者の立場としてお話をするとともに、今回応用につなげるデータベースが欲しいということで、どうやったら応用につなげられるかお話しします。我々の研究所はアカデミア創薬を目指していますので、創薬につなげられるようなデータベースというような話をさせていただきたいと思います。

セントラルドグマの中で、タンパク質というのは生命機能に直接関与していまして、多くの疾患の原因につながっていますので、薬の直接の標的となっている。ですので、プロテオミクスというのは創薬に欠かせないということに関しては、皆さん異論はないと思います（図６－２）。

先ほどの 1 枚のスライドで創薬研究にプロテオミクスが欠かせないことがほとんど全て集約されているんですけれども、補足として創薬研究が抱える問題点についてお話しします。20 世紀の創薬というのは、表現型からスクリーニングしてきたような創薬（図６－３）、それから培養細胞やモデル動物を用いたスクリーニングをしたもの、このようなものは、まず表現型のスクリーニングとしては、標的や作用機序が不明でありますし、モデル動物などを用いたものはヒトに応用できるか不明ということで、21 世紀に入ってからゲノム創薬というものが使われるようになりました。

これは、ヒトの臨床検体のゲノム解析、もしくは遺伝子産物を解析して、どういうものが標的になるかというものです。ということでタンパク質のデータベースは非常に重要なんですが、現時点でその情報量は圧倒的に不足しています。その理由として、まずゲノムに対してかなり遅れているということです。

これはリン酸化タンパク質の同定部位を示したグラフですけれども、2000 年、ちょうどゲノムプロジェクトが終わってゲノムが全部解読されたころに、リン酸化の部位同定はほとんどできなかった。それが 10 年たってようやくかなりの、1 回の解析で数万のリン酸化の部位が同定できるようになりました（図６－４）。

ゲノムプロジェクトに遅れること 10 年、ようやく 2014 年にヒトのプロテオロームのデータベースが報告されて、その中では大体 9 割ぐらいの遺伝子に対応したタンパク質があるということがわかった。これはあるということがわかっただけで、例えばどういう臓器にどのくらいあるか、病気とどう関係しているかほとんどまだわかっていないという状態です。

創薬の話ですけれども、現在の創薬のターゲットとしてどういうものがあるかというと、大体キナーゼとか GPCR の膜タンパク質というのが主なものです。緑で示しているところが開発されていないタンパク、創薬ターゲットです（図 ６－５）。青が既に応用されて使われているもの、赤がクリニカルトライアル、臨床試験のほうですが、ほとんどのタンパク質は、半分以上のタンパク質がまだ創薬のターゲットとつながっていないということで、それは原因としてはタンパク質の情報が不足しているということです。

我々がプロテオミクスをやって創薬ターゲットを見つけるときに、どういう悩みがあるかという具体例を、2 つほどお示したいと思います（図 ６－６）。1 つはタンパク質の発現量情報を用いた創薬研究で、もう 1 つは、翻訳後修飾を用いた研究です。時間が限られていますので簡単にしか説明できませんが、我々の方法としては、ショットガンプロテオミクスを用いて、例えばがんであればがんの組織を使って、どういうタンパク質の発現が変動しているかというのをまず探索して、最終的にそれを検証していくわけです（図 ６－７、８）。例えばこれは大腸がんの組織の探索の例です（図 ６－９）。大体 5000 個のタンパク質が 1 回の実験で同定できて、これは膜タンパク質に焦点を絞ったんですが、6 割くらいが膜タンパクで、3000 個くらいのタンパク質が出てきて、そのうちの 400 個ぐらいが差があった。そのうち GO 解析で実際に膜に局在しているものが 100 ぐらい。その 100 の中で検証して最終的に 44 個まで絞り込みました（図６－10）。

この 44 個が一応大腸がんの膜タンパク質の創薬標的ということまで絞り込めました。その後我々のところでは、抗体を

つくって抗体医薬に応用しようとしているんですけれども、どのターゲットを選べばよいかというのがわからない。そこで悩むんですね。

このようにリストが出てきます（図 6－11）。PSM というのは、実際に同定されたペプチドの数ですが、タンパク質の発現量と考えてください。あとは局在だとか機能とか、こういうのを一個一個のタンパク質を文献情報で調べていかないと、どういうものかがわからない。そこがネックになっています。やはり発現量。抗体医薬をつくるときにどうしても発現量の多いものを、優先順位として挙げていくので、どのくらいの発現量があるのか。それから、局在はどうなのか。ちゃんと膜に局在しているのか、それともいろんな臓器に局在しているのか、それともある特定の臓器に局在しているのか。いろんな臓器に局在している場合は、薬をつくっても毒性が強くて使えないという可能性が高いです。

あとは機能はどうなのか、疾患との関連はどうなのかというような情報が必要です。こういう情報が、例えばワンクリックでタンパク質を見つけたときに得られるようなデータベースがあると、創薬に非常に役に立つと思います。

もう 1 つの例は翻訳後修飾の代表としてのリン酸化プロテオミクスの例です（図 6－12）。先ほども言いましたキナーゼというのは、創薬のターゲットの中で一番多いので、どのキナーゼが創薬標的になりうるか調べることが重要です（図 6－13）。我々としてはどういうキナーゼを創薬標的にすればいいかということを、このような実験を使ってやっています（図 6－14）。

具体的には、組織とか培養細胞、PDX モデルからリン酸化プロテオームを使って解析すると 1 回で数万のリン酸化サイトが同定できるんですが、そのリン酸化サイトがどのキナーゼによってリン酸化されているかという情報がなかなかない。今、Phospho Site Plus という民間の CST が出しているデータベースで、8000 個ぐらいのキナーゼと基質の情報が紐づけられていますけれども、これでは圧倒的に足りないということで、どのくらい足りないかといますと、我々の解析では、紐づけられているものが 5%しかない（図 6－15）。残りの 95%は、リン酸化サイトはわかってもどのキナーゼによってリン酸化されているかが全然わからない。なのでこれを埋めるようなデータベースが将来的に必ず必要になってくると思います。

ということで今までの話をまとめますと、タンパク質を創薬に応用するときに、発現量の情報とか修飾情報が必要です（図 6－16）。あとは局在。全身に発現しているのか局所で発現しているのか。最後に機能と表現型の情報が必須です。これらのデータの紐づけが絶対に必要です。そういうものを実際に自分で全てのタンパク質にやるのは無理なので、例えば文献情報でもいいから紐づけできればいいというふうに思います。

それとともにプロテオミクス以外のオミックスのデータが必要で、特にゲノミクスです。ある遺伝子変異があると、細胞内でどういうタンパク質が異常になるのか。あとタンパク質が異常になったときに、その遺伝子、ゲノムに変異があるのかどうかという情報があると、非常にありがたい（図 6－17）。今アメリカの、NCI ではゲノムとタンパク質を統合したデータベース、プロテオジェノミクスというプロジェクトが走っています。

現在プロテオーム解析については、jPOST が 2015 年から走り始めています（図 6－18）。これはアジア・オセアニアで唯一のデータデポジットリーサイトで、これまで、世界で 3 カ所ぐらいしかなかったですけれども、ようやくアジア・オセアニアに 1 つのサイトができました（図 6－19）。世界中の人がその恩恵を受けているんですけれども、その理由としてまず、統一したプラットフォームに全てまとめられているというのと、あと非常に速いので、世界各国からデータが集められています。プロテオームの論文を書くときに必ず生データを投稿しないといけないんですけれども、この jPOST は世界で一番速いので、世界中の人に使われています。

現在 1 次データベースを集めているんですが、それを最終的に 2 次データベースにするときに、例えば創薬にそのデータベースを使いたいというときに 2 次データベースの中で、疾患に関連するものとか発現量に関連するデータ、翻訳後修飾に関連するデータとかを集めて、それとゲノムのデータベースとも関連づけて、最終的に臨床応用に持っていきたいと考えているところです（図 6－20）。

図 6－1

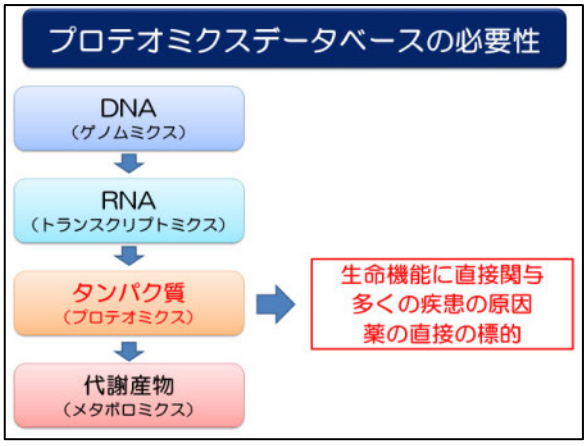


図 6－2

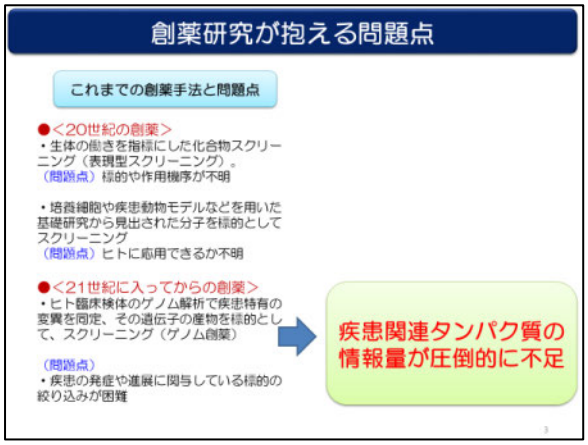


図 6－3

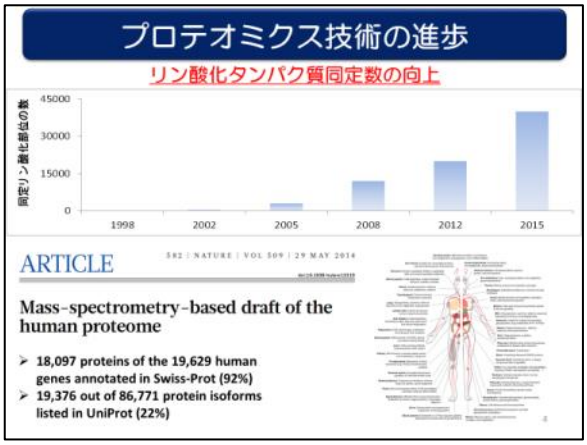


図 6－4

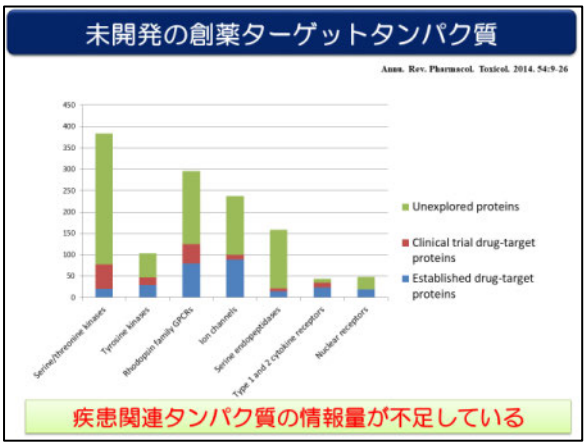


図 6－5

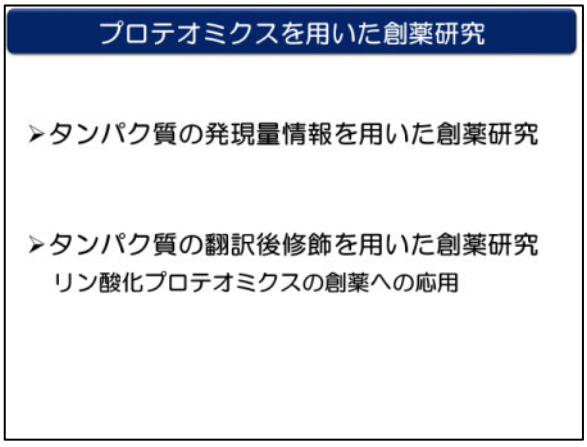


図 6－6

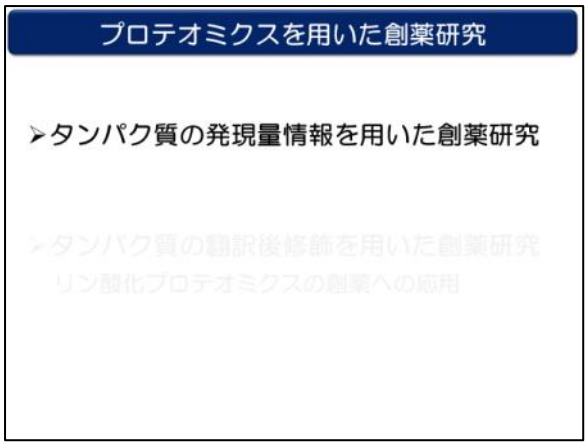


図 6－7

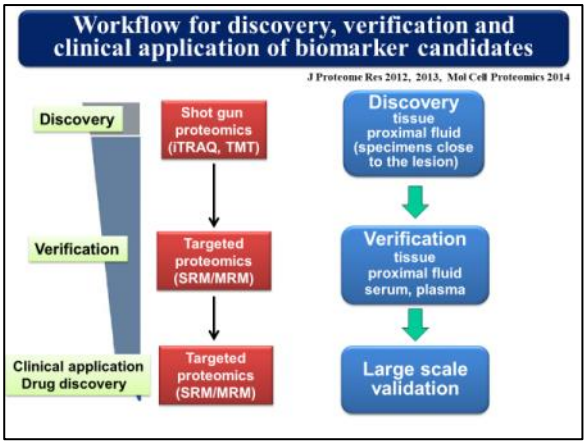


図 6－8

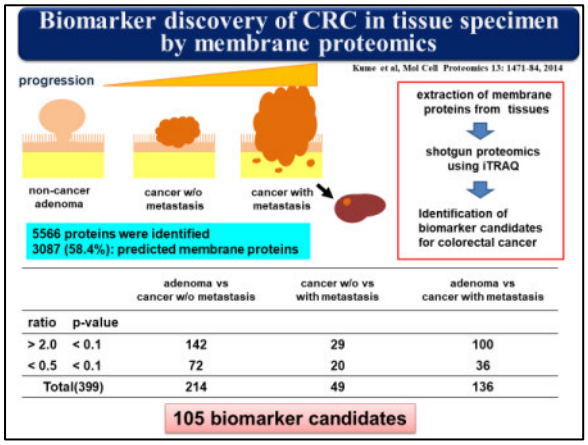


図 6－9

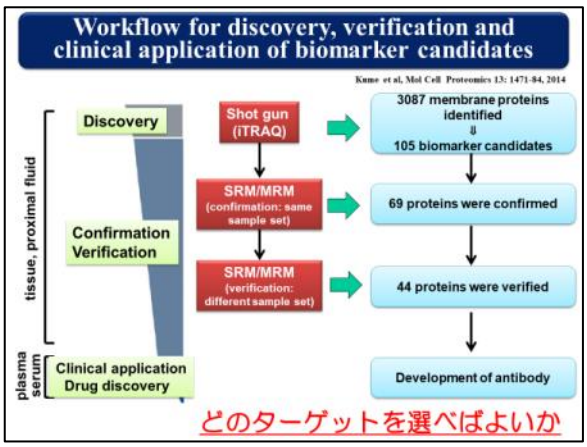


図 6－10

図 6－11

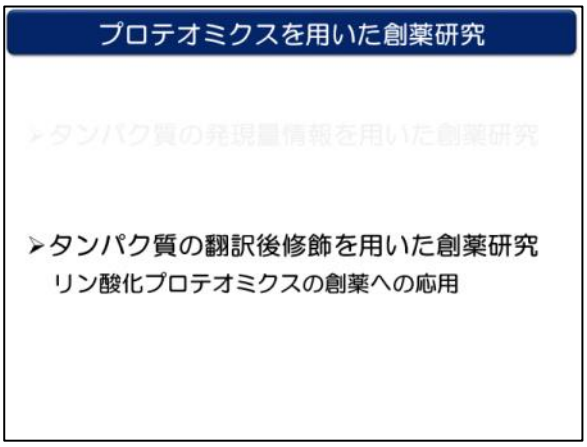


図 6－12

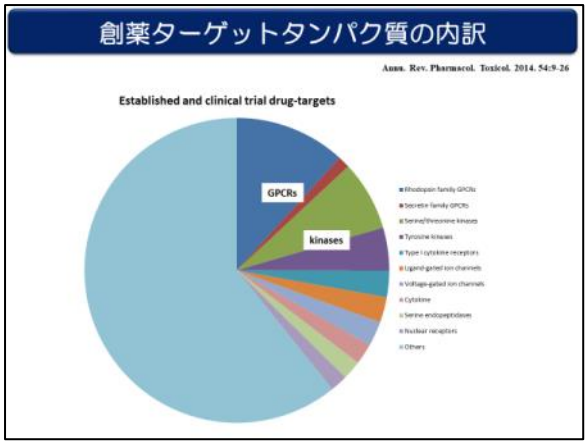


図 6－13

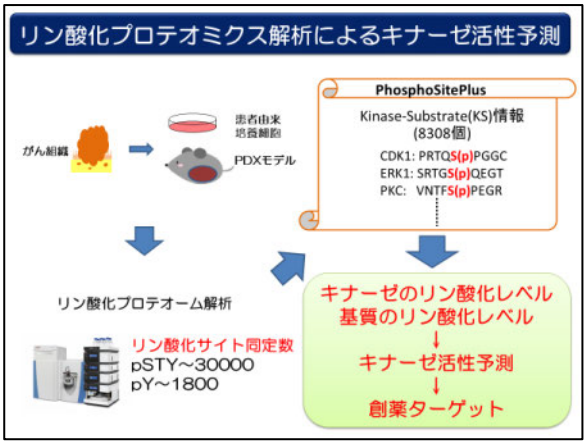


図 6－14

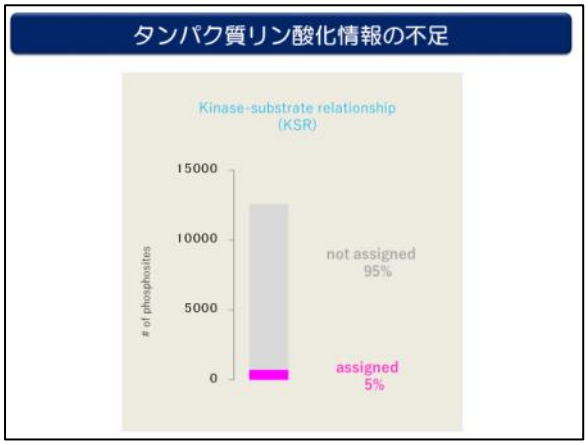


図 6－15

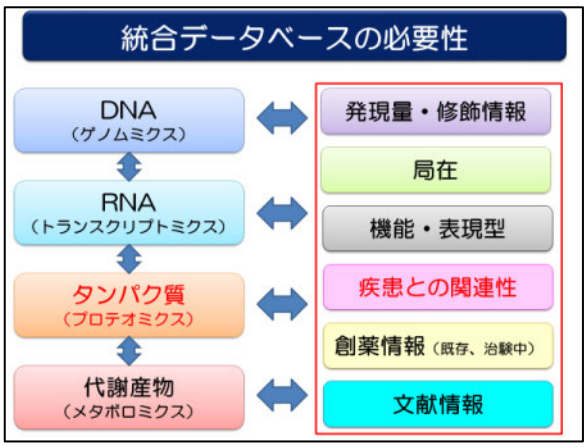


図 6－16

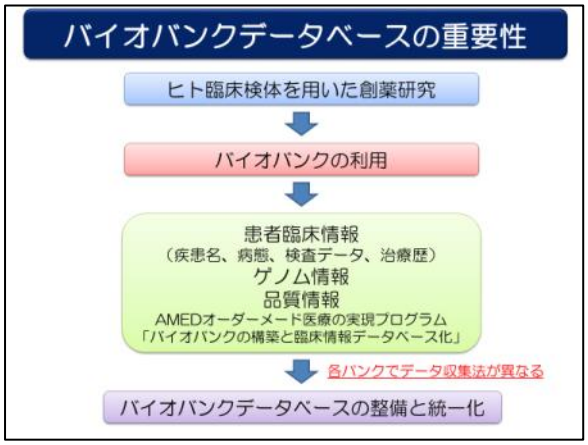


図 6－17

図 6－18

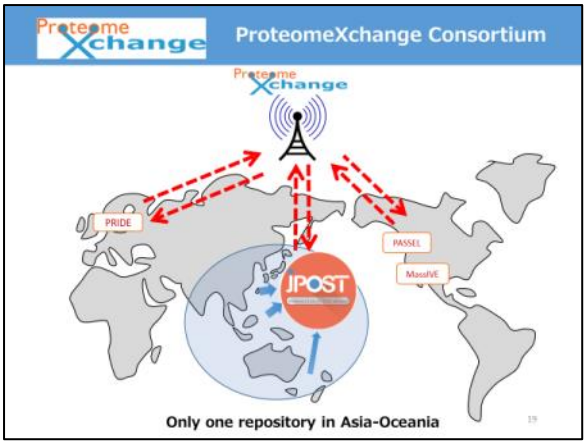


図 6－19

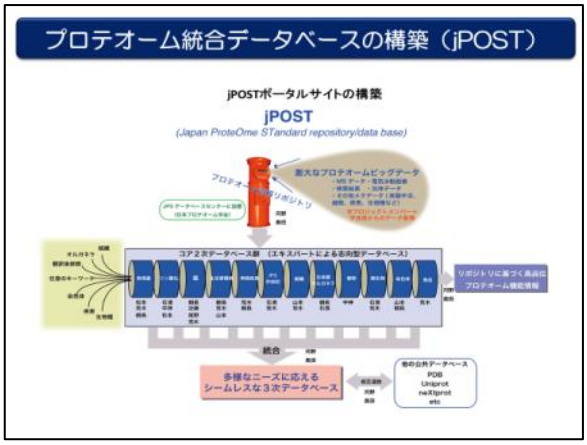


図 6－20

＜質疑応答＞

（質問）こういった統合データベースの構築というのは、スナップショットだけでは生命現象全体の説明はなかなか難しい。バイオマーカーは見つかっても、なかなか臨床まで応用がいかないというのはそういう原因があるのではと思う。いわゆるタイムコースのデータを計画的にしっかりととっていくというのが、また、パスウェイの解析と連携した形での整備が非常に重要になってくるのではないか。

（回答）それは各研究者がやっていくと思う。我々もそれは計画しているし、実際にタイムコースというと、例えば診断のマーカーとかそういうものは発症する前から、治療した後、それから再発した後とかそういうタイムコースを使って、どういうバイオマーカーが一番役に立つかというようなことはやる。パスウェイに関しては、先ほどリン酸化のデータを示したとおり、基本的なデータがそろってくると、パスウェイもおのずとわかってくると思うので、それを平行してやってきたいと考えている。

（質問）創薬というターゲットからはもちろん、キナーゼが一番、リン酸化プロテオームのテクノロジーから積み上げやすいと思う。アセチル化とかメチル化などの修飾も創薬の標的になりつつあるかと思うが、データのプロダクションはどうか。

（回答）リン酸化が先に走っているという状態で、それ以外の翻訳後修飾は追いかけているという状態。これからの課題。

（質問）リン酸化ということと膜タンパクをやるというのは、例えば抗体をつくるというところで出てくる話だと思うが、一般論として、本来ターゲットにすべきタンパク質は、細胞の中にあるのではないか、という考えについてはどうか。

（回答）それもあるが、まだ技術のほうが追いついていないところもある。丸ごと全部細胞をとってきて解析しようとすると、どうしても量の多いものからしか見えてこない。細胞内で創薬の対象になりそうなものは、タンパク質量の少ないことが多くて、まず局在で分画してそれぞれを解析するなどの工夫が重要である。それから、本日お話ししたリン酸化プロテオミクスによって見つかるキナーゼは細胞内のターゲットである。

7　システムバイオロジーとバイオデータベース

大阪大学蛋白質研究所　岡田　眞里子

＜発表内容＞

今日は「システムバイオロジーとバイオデータベース」ということでお話しさせていただきたいと思います。私は、システムバイオロジーという研究をしています。

システムバイオロジーというと、日本ではメカニカルモデリングのようなモデリング研究をイメージされる方が結構多いんですけれども、アメリカではシステムバイオロジー　イコール　オミックスと考えられています（図７－２）。システムバイオロジーといっても多様でして、研究内容で大まかに分けますと、データサイズに依存したデータ主導的なもの、モデリングのようにボトムアップでシミュレーションをやっていくものの 2 種類があります。どちらも利点がありますが、データ依存的なものですと、データドリブンに包括的な解析が可能で、ボトムアップのモデリングですと、分子メカニズムなどが理解できます。

データサイズとともに、研究アプローチも 2 種類ありまして、計算的にやっていくか、あるいは実験的にやっていくか、があります。また、システムバイオロジー向けのデータベースやシミュレータの構築も研究分野としてありますし、計算量をそれほど使わない理論重視のアプローチもあります。

システムバイオロジー研究では、データベースを日々使っています（図 ７－３）。例えばパラメータを決めるためにタンパク質の構造を PDB から取得して、利用することもあります。あるいはモデルが構築できますと、モデルをデータベースにデポジットしていくというようなこともあります。また、パスウェイのシミュレーションの際には、モデル DB からすでに構築されたモデルを取ってきて使ったり、パスウェイ DB、KEGG、代謝系ですと酵素 DB も使います。遺伝子発現データから何の転写因子が活性化されるのかを調べたりするときにも、データベースを使います。

これまでは、オミックスを中心としたシステムバイオロジー研究者とシミュレーションをやっているシステムバイオロジー研究者は、分野が別という感じでしたが、最近はこれらを融合したアプローチが多く用いられるようになってきました（図 ７－４）。

アメリカは静的なビッグデータ解析が多い一方で、日本あるいはヨーロッパは動的なデータを扱うことが多いです。ただ、同じアプローチだけ使っていると知見発見にどうしても行き詰まってしまう。そこで最近は、アメリカでも、メカニズムを知りたいときに、モデリングを使うということをよくやるようになりました。それだけいろいろな知識を持った人、いろいろなデータベースを使える人がシステムバイオロジー研究者の中にいるわけです。

まず自分の研究の立場からお話ししますが、私たちはシグナル伝達系の数理モデリングを行っています。シグナル伝達系は環境因子と細胞の中の遺伝子の情報がインテグレートされて細胞の制御を行うというようなシステムです（図 ７－５）。こういった情報処理の解析に、数理モデル、ここでは、微分方程式による時間変化のシミュレーションを行います。さらに、モデリングにはパラメータが必要になってきます。パラメータには、例えば分子間の結合、あるいは解離のスピード、酵素のミカエリス定数があります。パラメータを得るためには、実際にデータを実測するときもありますし、タンパク質の構造から予測するときもあります。あるいは時系列データからパラメータを予測するときもあります。

自分たちの研究ベースでもデータベースはたくさん使いまして、このパスウェイの構造、ネットワークの構造を抽出するのに、文献情報やパスウェイ DB を使います（図 ７－６）。分子イメージングデータ、タンパク質構造、プロテオームデータ、こういったものも使います。実際にデータを自分たちで収集するときと、データベースや文献情報からパラメータをとってくるときがあります。

こういったものがモデルになって、それが論文化されると、モデルというのは、BioModels というデータベースに登録されます。そうしますと自分たちが構築したモデルが、世の中で多くの人に使われるようになります。

またさらに遺伝子発現、シグナル伝達だけではなく、細胞運命まで説明しようとすると、トランスクリプトーム、エピゲノムあるいは疾患ゲノムというようなデータも使うようになります。

これは有名な Hanahan＆Weinberg の Cancer Hallmarks という図です（図 7－7）。この図の中で、赤く囲ったところを今まで自分たちで数理モデルとして構築しています。まだ作成していない部分も今後、数理モデル化し、これらを統合していこうということを今考えています。そうすると、かなり大掛かりなパラメータが必要になってきます。

よく計算系の同僚に、何でこんなに面倒くさいシミュレーションをやるのか（図 7－8）、もっと簡単にできるんじゃないか、と言われます。ただシグナル伝達系というのは、フィードバック制御とかそういうものが多く含まれていて、すごく非線形が高い。特に炎症とかアレルギーとかイメージされるとわかると思うんですけれども、閾値応答という、あるところまでは全然症状が出ないのに、あるところを超えると、急に症状が出るという閾値みたいなものがあります。そういったものは、AI とか統計モデルだけでは説明できない。微分方程式系のモデルだからこそ説明できる分子メカニズムがあります。そういった分子機構を知るためには、わざわざモデルを作る必要があります。

問題になってくるのは、こういうパラメータの大きさですけれども、こういうところは多分 AI でこれからどうにかできないかと思っています。ただ、いまのところは、地道にモデルをつくっていっているというのが現状です。

1 つのモデルをきちんとつくるのには 3－5 年かかりますが、そうやって一旦できたモデルが発表されると、いろんな人が使うようになります。例えばこれは私たちがモデルをつくった例ですけれども、Merrimack という会社では、薬剤スクリーニングに使ったり、エジンバラ大では、乳がんの耐性にかかわる遺伝子を予測したりというようなことに使っています（図 7－9）。

この Merrimack ですけれども、エンジニアリングの考え方で薬を開発していくという会社です（図 7－10）。ハーバードと MIT のスピンオフベンチャーで、シミュレーションによって、がん標的薬剤を開発しています。既に臨床段階に行っています。

ここまではシグナル伝達系を対象としたものですけれども、海外では、全細胞シミュレーションを行っているところがありまして、2012 年にはスタンフォードのコバートのラボが、マイコプラズマの全ての遺伝子について、論文、パラメータ、代謝、翻訳全てのプロセスをメタボローム、トランスクリプトーム、ゲノム、プロテオーム、これらのデータを統合してシミュレーションして予測するということをやっています（図 7－11）。これが、521 遺伝子から成る全細胞モデルですけれども、かなり注目を浴びた論文です。これからは多細胞生物のモデル化が進んでいくと思います。

アメリカの FDA ではインシリコの肝臓モデルをつくっています（図 7－12）。聞いた話によりますと、新しい薬を開発したときには、インシリコモデルを使って、あるいは知識データベースを使って、肝臓への毒性をきちんと調べなさいというようなルールが課せられつつあるように伺っています。肝臓モデルは、physiological なモデルですけれども、こういったモデルに対する研究者のハードルが低くなって、分子レベルでのモデリング研究が受け入れられやすい土壌になってきています。

モデルというのは、データベースの箱として使うこともできます（図 7－13）。これは簡単なモデルですけれども、S＋E が SE になって P になるという式ですけれども、ここに入るパラメータは例えば遺伝子発現量、タンパク質の量から得ることができます。こういうものは公共データベースからとってこられるものです。親和定数はちょっとハードルが高いんですけれども、タンパク質構造からある程度推定することは可能です。

こういったデータベース情報を統合するツールとしてモデルが役に立つ。私たちは今まで、例えば自分たちがつくったモデルがどの程度、予測精度があるだろうということを試していたりします。さっきから議論になっていますように、日本の臨床データはあまり公開されていないので、公開されているものをとってきます。例えば Oncomine というのはある程度はパブリックになったデータがダウンロードできるようになっていますので、そういったものをとってきます（図 7－14、15）。

例えば、疾患のサブタイプごとに、どういった遺伝子が発現しているかがわかり、さらにデータのキレーションというのがある程度きちんとされていますと、シミュレーション結果の精査もしやすいです。

こちらはインシリコモデルと臨床データを比較したものですが、どの遺伝子が予後に影響するかを調べると、ある程度一致する。ただ、一致しないところもあります。この結果からは、相互作用によって機能が発現する分子では、遺伝子発現量やタンパク量などが疾患の症状と割と合っている。ただ、キナーゼのような酵素で、変異が大きく影響するようなものは、濃度だけでは説明できないというような、そういう傾向がモデルリングからわかるようになります（図 7－16）。

どんなデータベースが必要かということですけれど、やはりつなぐためのデータベースが欲しいと思います（図 7－17）。そのときに先ほどのような疾患のサブタイプ、つまりフェノタイプなどの、キレーションがしっかりしているほうがいい。あとゲノム変異によるタンパク質構造の変化といったものもパラメータとして利用できるので、ゲノム配列とタンパク質構造をつなぐものがあるといいと思っています。

またネットワーク解析を行う上では、生物種間で保存されている遺伝子やネットワークといったものがある程度 DB でわかるとよいと思います。実験検証もやりやすくなります。また、さらにデータの質は高くあってほしいと思います。

データベースを日本で行う意義についてということを書かれていたので、思ったことを書いています（図 7－18）。まずデータベースと人材ということがありますけれども、維持すること、研究すること自体がデータ時代の人材育成につながると思います。今は、私たちは、日常的に次世代シーケンスなどのデータ解析をやっている。生物学とデータは切っても切り離せない時代になってきている。

例えば PD-1 などの開発例では、自然と基礎データというのは日本に蓄積されているというようなことがありますので、そういった日本発データを整えていく必要があると思います。

また、先ほどの育種の問題でも発表されていたと思いますけれども、環境とゲノムの相互作用によってフェノタイプが変わってきます。例えばアトピー性皮膚炎ですと、感受性遺伝子は欧米とアジアで全く異なるということがあります。欧米ですと皮膚の保湿性にかかわる遺伝子が感受性遺伝子だけれども、アジアに関しては炎症にかかわる遺伝子が感受性遺伝子、というように環境によって原因遺伝子がが違っています。そういったことから、日本ならではのデータを構築維持することによって、日本独自の新しい知見が生まれるのではないかと思います（補足資料　図 7－19～29）。

図 7－1

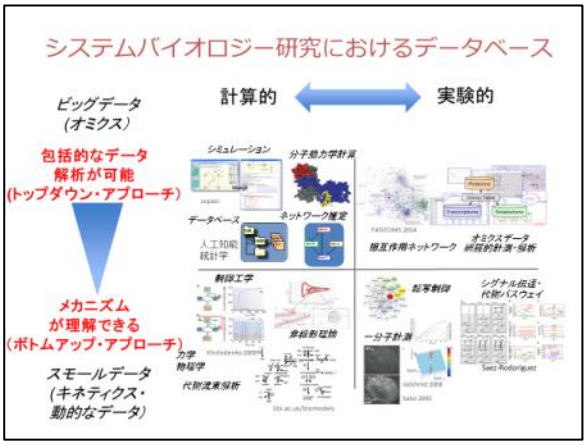


図 7－2

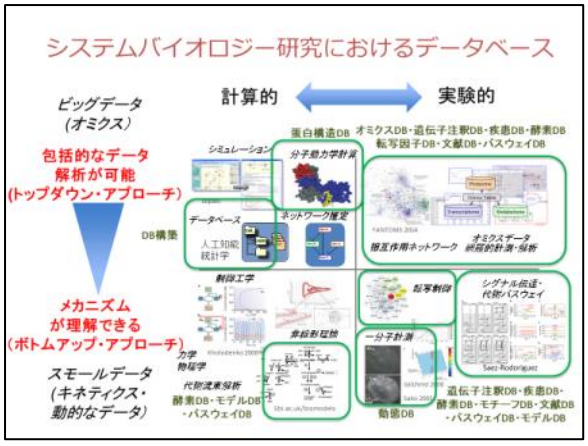


図 7－3

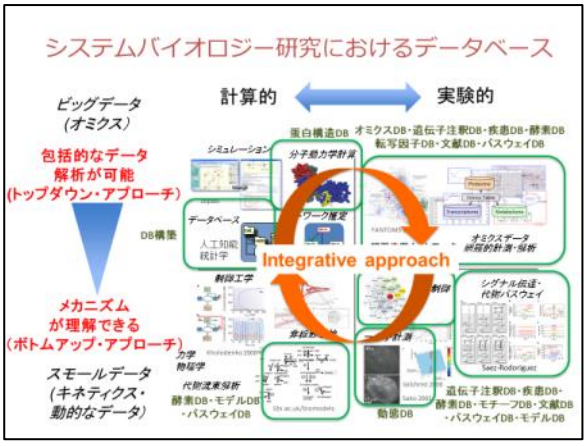


図 7－4

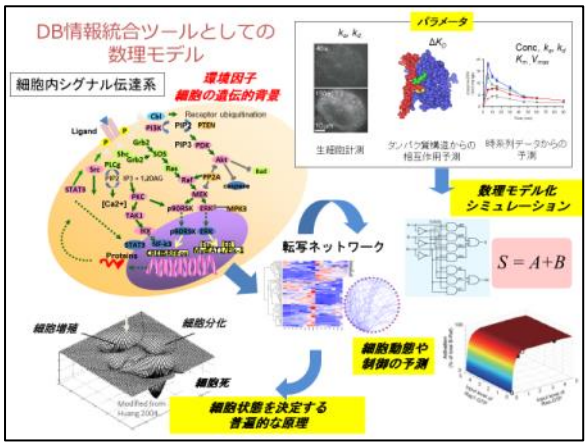


図 7－5

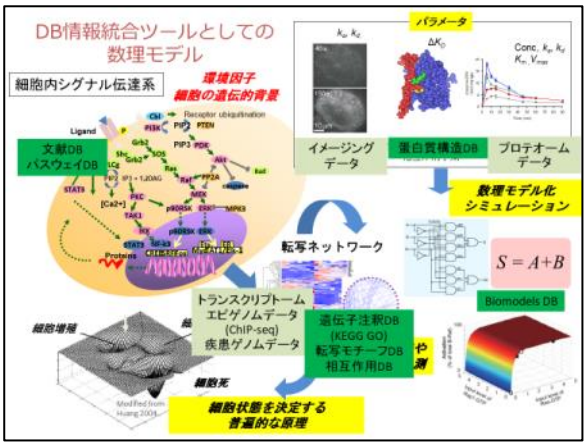


図 7－6

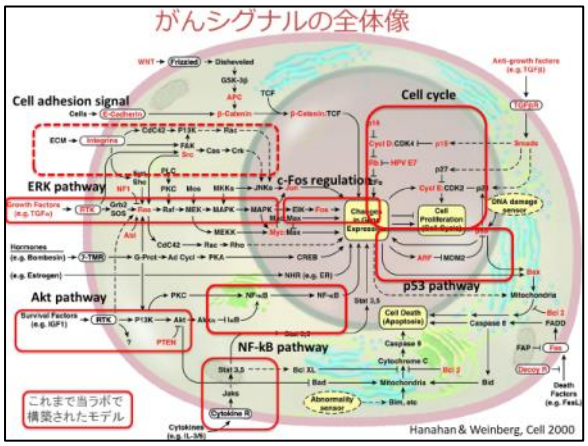


図 7－7

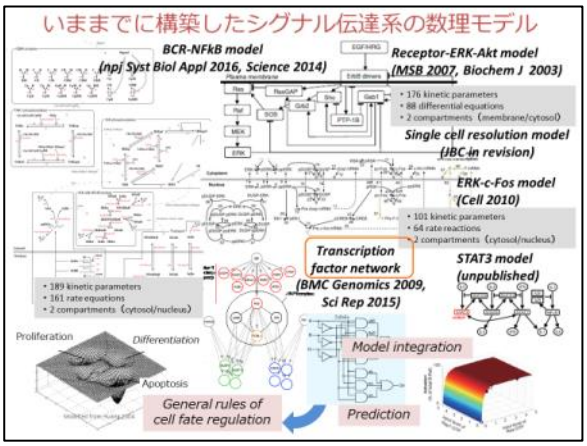


図 7－8

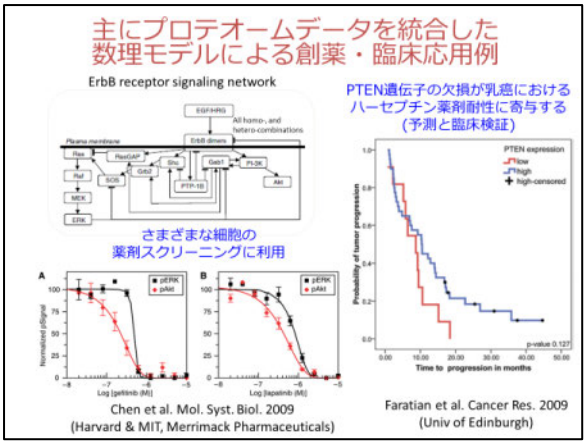


図 7－9

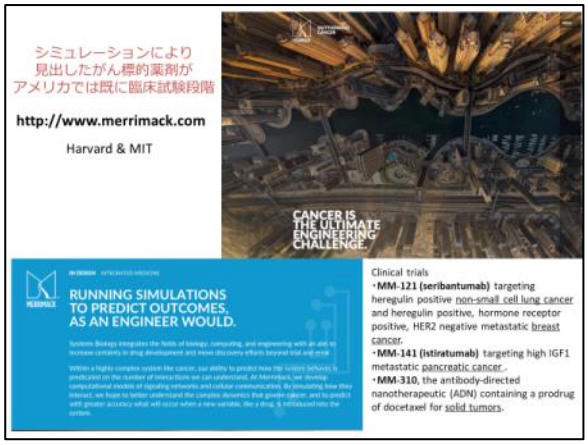


図 7－10

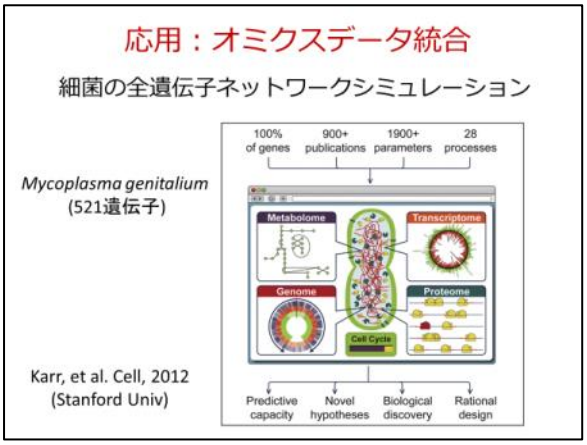


図 7－11

図 7－12

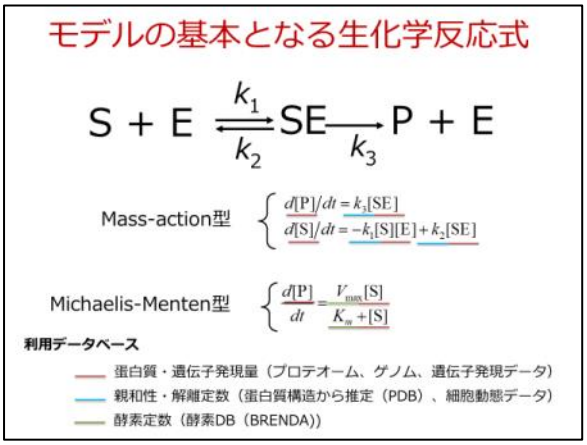


図 7－13

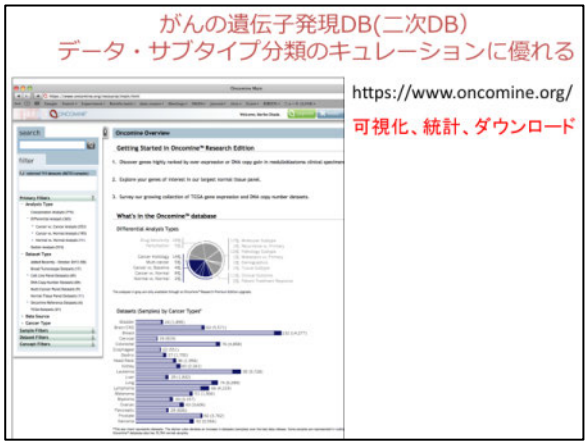


図 7－14

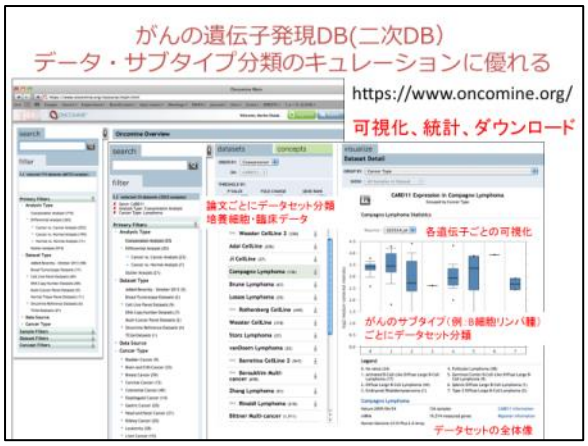


図 7－15

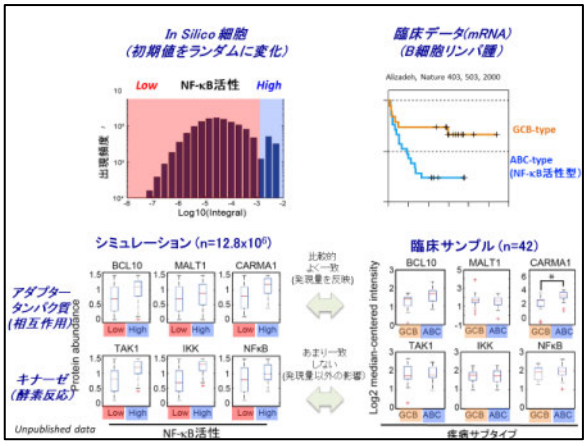


図 7－16

どんなデータ（ベース）が不足しているか
ゲノムー蛋白質ーフェノタイプをつなぎ、
高精度の予測を可能にするDB

- つなぐDB
- つなぐための生物学的なキュレーション（RDF化＋）
- ゲノム変異から蛋白質構造変異へ（創薬につなげやすく）
- 種間で保存されている遺伝子やネットワーク（線虫ーマウスーヒトのデータを行き来、実験検証をしやすく）
- 質の高いデータ、質の高いアノテーション

図 7－17

データ整備を日本で行う意義

- 維持すること・研究すること自体がデータ時代の人材育成に繋がる。
- 生物学とデータ（ベース）は切っても切り離せない時代となった（データ生産＋デポジット＋データ利用）。日本で生物研究が行われる以上、その近くでデータ（ベース）構造を理解し構築できる人は必須。
- 日本発の創薬では、初期データは自ずと日本に蓄積されていく(PD-1抗体の例）。
- 環境とゲノムとの相互作用によりフェノタイプが生まれる。
  疾病発症機構：例えば、アトピー性皮膚炎の感受性遺伝子は欧米とアジアでは全く異なるなど。
  育種：日本の風土や環境に適した動植物。
  物質生産：環境側としての微生物も風土により異なる(限りない亜種XX.sp)
- 日本ならではのデータ特性とデータベースの利用により、日本独自の新しい知見が生まれる。

図 7－18

追加資料

図 7－19

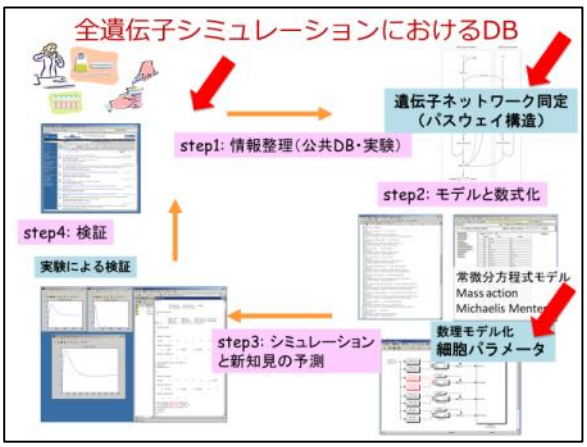


図 7－20

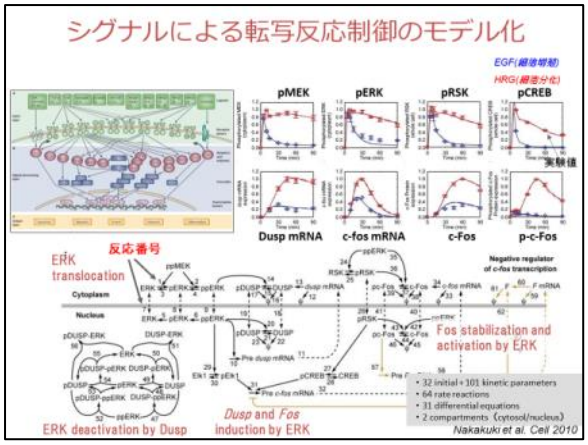


図 7－21

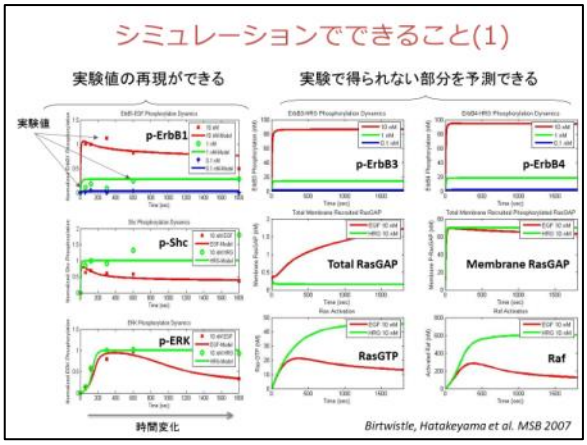


図 7－22

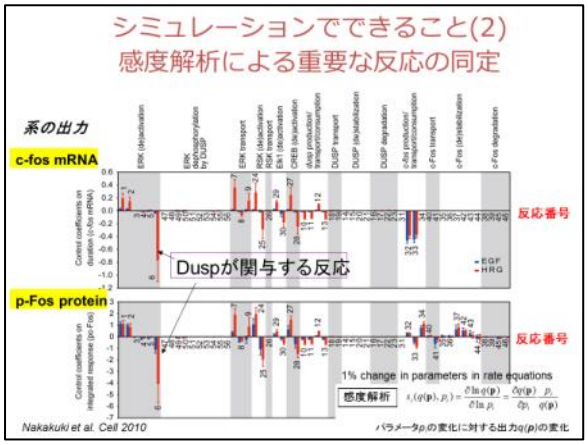


図 7－23

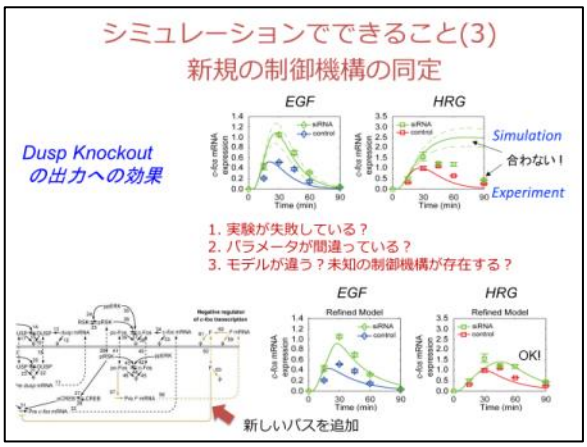


図 7－24

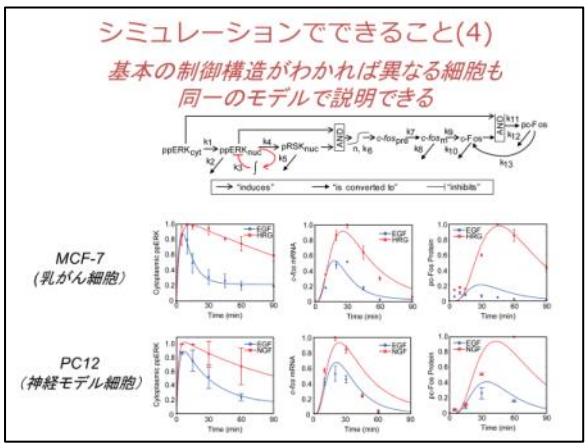


図 7－25

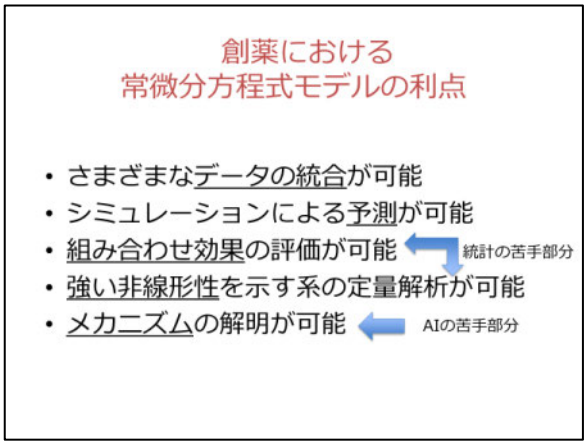


図 7－26

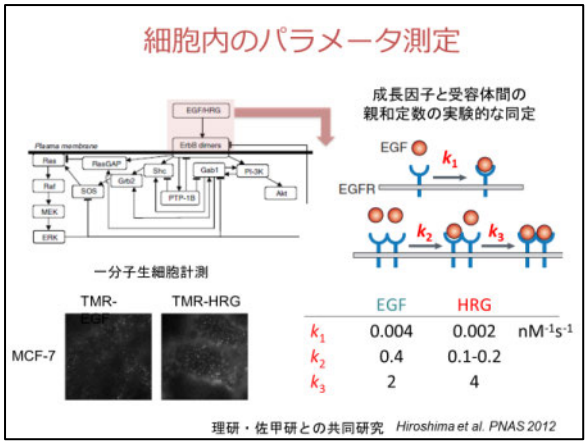


図 7－27

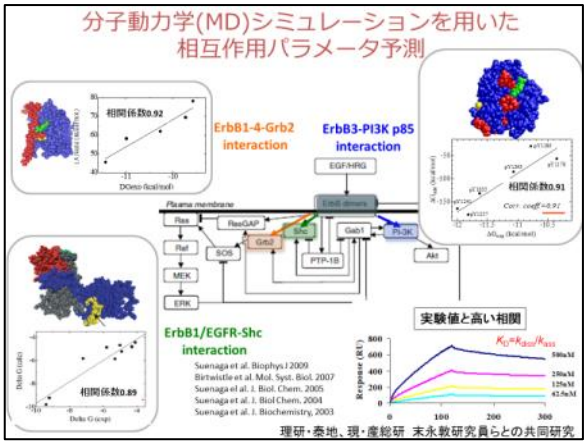


図 7－28

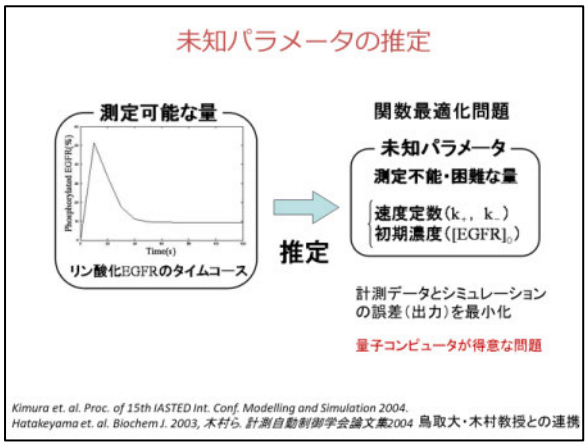


図 7－29

＜質疑応答＞

（質問）これから多分定量データとか動態データ、一分子計測とかそういうデータがこれからいろいろ出てくると思うんですが、その重要性や今後の展開はどう見ておけばいいか。

（回答）動態データ、例えばプロテオームみたいなタイムコースとか遺伝子発現のタイムコースというのは、割と扱いやすく、データベース化しやすいと思う。実際そういうものは私たちも使いやすい。ただ、イメージングデータとか、データそのものはとれているけれども、そこからパラメータを計算するときには、間にまたモデル式を挟んで数値化しているので、そういったものはデータベースには直接のりづらいのではないかと思う。方法論が一貫しておらず、シングルモレキュールトラッキングについてはスタンダードな方法がないので、結果として、目視の介在が大きい解析をしている。このような状態ではデータベース化は、しづらいのではないか。

（質問）そういった画像の動画像、そういうもの自体をデータベースとして上げることで普通の画像処理の研究者が、トラッキングアルゴリズムを開発してくれるというのが理想かなと思う。

（回答）その通りと思う。

（質問）ハーバードと MIT の会社で創薬で薬をつくったという話だが、彼らも独自なモデルをつくってそれでシミュレーションをしてということか。

（回答）彼らは、がんシグナルをメインにやっているが、実際には私たちがつくったモデルや、ほかの人がつくったモデルを Merrimack で統合し、それで自分たちの調べたい遺伝子と、自分たちの計測データをあわせ、その計測デー

タにパラメータフィッティングをしていって、実験と予測をやっていると想像している。

（質問）普通例えば日本の製薬会社だと、なかなか伝統的な創薬しかできていなくて、こういったシミュレーションから出てきた結果をいきなり臨床試験に入れるのは、非常に考えにくいと思うが、この会社のデシジョンメイキングのポイントはどんなところにあるのか。

（回答）ピンポイントの標的ターゲットを選び、データはかなりとっているはず。既存薬剤と自分たちの薬剤の比較をやって、それで自分たちの薬がほかの薬、抗体医薬がメインですけれども、ハセプチンなどと似た戦略が彼らは得意なので、本当に小さな定量的な違いによって薬の相互作用、特異性を予測しているのだと思う。抗体医薬とマウスの臨床、前段階の試験などは、また別個やっているんだと思うが、最初の候補選択の段階では、かなり定量的なことで決めているように想像している。

8　統合 DB のビッグデータ・AI 利用に向けた考察

産業技術総合研究所　瀬々　潤

＜発表内容＞

セントラルドグマと異なった軸としてシミュレーション、データベース、機械学習、AI を書かれたときに、これらが利用されている場面は多く、全部を網羅すると発表時間では終わらないボリュームがあるかと思います。では、だからといって AI 時代になったとき何か大きなブレークスルーが今のところあるかと考えると、大きなものはまだライフサイエンスに関しては、これからなのではないかと思います。

だからといって、囲碁の話をまじめにするのかと言ったら、ここで囲碁の話をしてもしようがないと思うので、僕の考える AI 利活用の問題点というようなものいくつかピックアップしてみました。

データ解析一般には、データベースがあり、そのデータを用いた機械学習、検証実験というふうに進みますが、バイオロジーのデータ解析も同様に進んでいくわけですけれども、そこに問題がいくつかあるかなと思って下に書いたのを順番に話をしていこうと思います。

1 番目が歯抜けのデータベース、2 番目が新 NP 問題、3 番目が骨董品データベース、4 番目がアノテーションの不在、5 番目が検証されない予測と 6 番目がフィードバックという話です（図 8－2）。

まず歯抜けのデータベースは何の話をしているかというと、データベースをつくりましたというとき、似たようなデータを集めてみても、実は全体が網羅されていないというケースが多々あります（図 8－3）。下にあらわしたのが、その雰囲気です。例えばイネのデータベースがあってコムギのデータベースがあって、合わせてみましたといっても合う遺伝子は少ないですねとか、あるいはヒトのデータベースでも、がんのデータベース、糖尿病データベースを集めても、遺伝子は合っているかもしれないけれども、人が違っていたりとか、とられているジェノタイプが違ったりするので、データを集めることはできるかもしれないけれども、意外と間に齟齬がいっぱいあって、とられている状況が違ったりしていて、統合することに意味があるかどうかもわからないし、そもそも統合できるかできないかもわからないことがあります。わからない値に関しては、補完をしていくしかなく、補完をするには目標をある程度明確にする必要があります。データをビッグデータにしたことによって、すぐに何か新しいことがわかるのかといわれると、そこにはギャップが存在していて一筋縄ではない場合がある。これが 1 つ目の歯抜けデータベースの問題です。

解決策はあるのかというふうに言うと、機械学習一般でこういう歯抜けの部分を埋めるような補完の技術そのものはあります。予測をする技術はあります。ですがそれが生かせるかどうかは、まだデータ量も含めて試していないのでわからないんですが、そういう補完技術がうまくできてくると、データが、歯抜けの部分が埋まってきて増えていくかな、きれいに使えるようになるかなと思っています。

2 番目は新 NP 問題というやつです。NP 問題というのは、計算機の人にとっては計算量の話で出てくる問題ですけれども、生物データの場合にはサンプル数より次元、例えば、遺伝子数とか SNP 数が大き過ぎる問題です。サンプル数が 1 万人に対して、SNP の数が 1000 万というふうになると、非常に横長のデータになります。そういうケースでは、なかなか予測をしても外れている、言い換えると、偽陽性のケースが非常に多くなります。これに関しては、生命科学のデータベースではよくある問題です。

ところが最近 GWAS などでは 100 万人規模の結果もあらわれていますし、シングルセルでは 1 万細胞×1 万とか 2 万遺伝子というふうになっていますので、この新 NP 問題は問題にならないデータも出てきていますが、とはいえ、一般のレガシーなデータベースに関しては問題になっていますので、こういう問題は今後も常にあるのかなと思います。

これに対しての解法は実際にはなくて、統計で言えば多重検定の問題を補正したりとか、あるいはシミュレーションを考

えてもう少し精度を上げていくとか、特徴量を選択していくとか、いくつか解法はあるにしても決定的なものはありません。

3 番目は、データベースの骨董品化の問題です。例えば、医学系の介入研究は介入終了まで第三者にデータを渡さないというのが医学系の不文律のように使われていて、例えば 1 年間の介入を行うとなったら、1 年間の観察期間を含めて 2 年間はデータは触れられません（図 8－4）。

例えばこういうアップルウォッチなりフィットビットのものをとって 1 年間の介入実験を行ったら、2 年後までデータに触れないということになるわけです。そういう状態では、出てきたデータはもう 2 年前のデータで、2 年後のアップルウォッチの性能は多分全然違うはずです。そういう状況でデータベースをつくってももうデータは腐っている、意味がないというふうになりますので、この辺は医学介入研究と時代の流れの速さの間のギャップを考える必要があります。そして、今の人工知能のリアルタイムに物事を解析していくという研究、特に強化学習のような、ロボットが学んでいくような研究が進んでいますので、データベースがどんどん学んでいくシステムが構築できるように、研究をするというシステムを変化させていく必要があります。

あと希少例を集めることはとても重要ですけれども、希少例だけに集中しすぎると、全体像が見えないデータベースになってしまう可能性があります。先ほど述べたデータの歯抜けの状態になるので、この辺の全体像を見るというのと希少例を集めるということのバランスについて考えながらデータベースを構築しないと、学習に耐えうるデータベースにはならないだろうと思っています。

4 番目は、アノテーションの不在です。AI や機械学習に関しては、アノテーション情報は必須です。深層学習が流行るきっかけとなったデータ解析の大会では、100 万枚のデータに対して、これは犬がいますとか、リンゴがありますというアノテーションがつけられています。そういうデータが存在することで、はじめて現代の機械学習は成果をあげることができます。

配列データベースなどデータベースへの登録では、レポジトリに登録する際に、アノテーション情報をつけるように求めていますが、人によって詳細まで書かれている場合もあれば、必要最低限のみが書かれている場合もあります。また発現量に関して言えば、ノーマライズされているのか、されていないのかに関しても全然書かずに登録されている例も見られますので、データが使われるように、どのような実験が行われ、どのように処理されたデータであるのかについて、生命科学コミュニティ全体として、整備される必要があると思っています。

また、そもそも、生命現象自身が曖昧なので、ちゃんとしたアノテーションをつけるということ自身、相当無理があるかなと思っています。できれば第三者が自分の思ったことというか、そのときの状況によってつけるシステムがあって、そこに逐次アノテーションを追加していくような、今のまさに GO なんかはそういうところがありますけれども、必要に応じて情報が追加できるシステムがあってもよいと感じています。

5 番目は、検証されない予測です。今の AI、機械学習の時代になって、実験して検証するスピードより、予測をするスピードの方がはるかに速く、検証されない予測というのが実は山ほどあるわけです（図 8－5）。また、予測をしても、実験ができるとも限らないという問題もあります。この遺伝子を止めれば疾患が治ると予測されても、そもそもその遺伝子をノックアウトするのが困難だったりするケースです。また、機械学習を 1 回やると精度を上げていって最後でき上がった予測に関して、例えば遺伝子でいえば、2 万遺伝子全部に関して一挙に予測ができるわけです。でも実際に検証されるのはその中の 2 個とか 3 個とかを検証して、合っていましたといって終わりにするということになるので、残りの予測に関しては、データベースをつくるのかというと、データベースをつくるほどのモチベーションもないですし、つくったところで多分論文にも通らないのでつくらないといって終わってしまうわけです。

こういう検証をされない予測は、実際にはデータがここにあって、こういう方法があって、こういうふうにやればこの結果が出ますという方法のデータベースのものがあれば、興味ある人が片っ端から検証してみるようなことができるかもしれないので、データのデータベースでなく、手法のデータベースのようなものがあってもよいのかなと思います。

機械学習全般に関して、多分今はいわゆる教師あり、教師なしのいずれかに分類される学習手法が多いですけれども、

今後可能性があるものとしては、1 つは強化学習、ロボットのように予測をしてそして賢くなるという手法が生命科学のデータにも適用されていきます（図 8－6）。先ほど少し申しましたが、さまざまな実験を例えば 100 人リクルートして実験をして、予測をつくって 100 人リクルートして予測して実験してというようなサイクルが走っていくような形で進むでしょう。

もう 1 つ時系列の解析に関してもどんどん増えてくると思います。生命科学のデータベースで時系列の実験結果もありますが、観測点の数が 20 点とか 30 点とか、非常に少ない。よく機械学習で扱われる時系列解析は、センサーデータのような頻繁にタイムコースをとられるものに対する機械学習が多いですけれども、数千点は取られるものなので、このギャップを意識してデータベースの作成や解析を行う必要があると考えています。

そしてちょっと僕の野望ですけれども、もし生体データベースが本当に生体情報を全部網羅できるとすれば、ある種シミュレーションに近いことができるのではないかと思っています（図 8－7）。例えば囲碁やチェスのような形では、計算機が勝手に囲碁を指してみて、勝ち負けを判断していました。それに対してデータベースを考えた場合、こんなデータがあったらいいなと思って突っ込んだときに、それに対するデータをほかの情報から、とられていないけれども補完をして、そしてつくるということができるのであれば、まるでシミュレーションのようなことが、データベースからできるのではないか。

AI の成功事例のポイントには、閉鎖系データとか成功失敗が明確であるとかあるのですけれども、結局は考えたことがすぐに判定して予測できるということが重要なので、そのサイクルをすぐに回せるように、データベースがまるでシミュレーションができるような、そういうシステムができ上がると結構おもしろいかなというふうに思っています。そういう意味でデータベースを集めてくることが重要かなというふうに考えています。

最後になりますが、今の時代のデータベース整備は何かというと、まず AI に関しては、今まるで AI が最後の地点のように考えられていますけれども、やはり AI は途中であって、それから出てきた技術、それから出た予測というものが最後重要だと思います（図 8－8）。AI もデータも土管の技術なので、それだけで何かできるわけではないので、それを地下に埋めてみんながどんどん利用できるシステムが必要になるだろうと思っています。

現状に関して言えば、きれいな実験を少数行って、結果を示したいということはとてもよくわかるんですけれども、多少汚くても、包み隠さず示した大量データのほうが、人工知能や機械学習にあっているケースが多いので、人手による取捨選択を行うのではなく、データをいっぱい集めることが重要だと思っています。

あとアノテーションはとても重要と考えています。完璧なアノテーションを一度で作成するのではなく、技術の進歩でデータの再取得が行われるように、データが新しくなったらアノテーションも更新し、AI、機械学習をもう 1 回回して予測をすればよい。このようなサイクルがどんどん回ることで、生命科学全体が進展していくのではないかと思います。

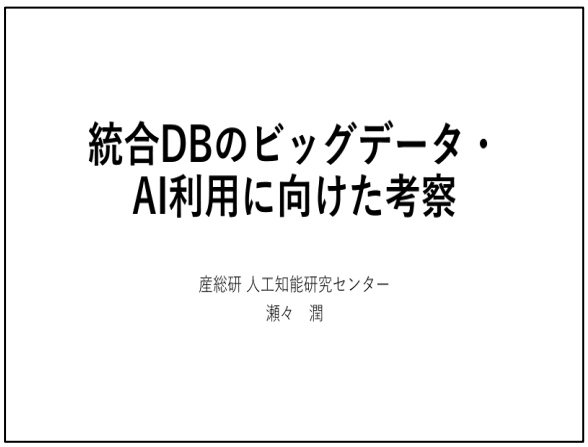


図 8－1

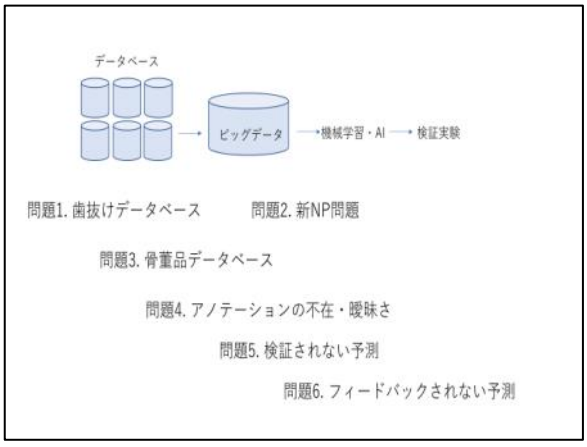


図 8－2

- 問題１．歯抜けデータベース
  - 似たようなデータを集めてみても、全体が網羅できない
  - そもそも結合できないか、分からない値は補間するしかない
- 問題２．新NP問題
  - サンプル数より次元（特徴量）が多すぎる問題。生物のビッグデータ（GWAS、発現量など）は大抵そう。
  - しかし、超大規模GWASやsingle cellの解析などでは、この問題はなくなってきている。（＝計算機屋にとって魅力的なデータ）

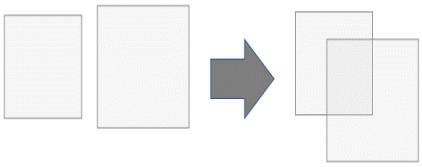


図８－３

- 問題３．骨董品データベース
  - 意味１：医学系の介入研究は、介入期間終了まで第三者データに触れないという内規。折角のデータが腐る。
  - 意味２：希少例を集めることに執着しすぎると、全体像が見えず、生命の全体像が見えないデータベースとなる可能性がある。
- 問題４．アノテーション不在・困難
  - AI・機械学習には、アノテーション情報が必須
    - 機械の予測が正しいか、誤っているかの判断材料が必要
  - 多くのケースで、そもそもレポジトリに登録する際のモチベーションが無いので、アノテーションが不在 or いい加減
  - そもそも、曖昧な生命現象
    - 例えば転写因子のBinding siteの様に、位置や状況が場合によって変わってしまう曖昧なものもあり、アノテーションとして、利用価値があるか曖昧なものも多い。
  - 第三者が追加できれば良いかもしれないが、データ生産者の意図を越えては難しいかもしれない。

図８－４

- 問題５．検証されない予測
  - 計算機を用いて複雑な生命現象など、予測することは可能かもしれないが、その結果を検証することが難しい
    - 教科書やデータベースにあれば、それを利用してチェックできる
    - データベースに無い場合は、新たな実験を実施する必要がある
      - ロボットサイエンティストの様なもので、解決できることを、個人的には望んでいる
      - そもそも実験自身が不可能な予測が出てしまうこともある。
  - 結果、やりっぱなしの予測が多い
    - データベースが生命現象を網羅すれば、全ては検証可能になるが・・・
- 問題６．フィードバックされない予測
  - 予測結果は、何らかの形でデータベースとして蓄積されても良いように思うが・・・
  - 現状は、データベースとしてまとめるのが、精一杯か。

図８－５

- 機械学習の一般的な問題分類
  - 教師あり（クラス分類）、教師なし（クラス分類、相関分析）
- 生命科学で需要が増してくる（いる）問題設定
  - 強化学習
    - ロボットが徐々に賢くなるようなアルゴリズム
    - 条件：頻繁に介入ができること。
      - 医学系でも、介入→経過観察→介入方法の変更、が高速にできるようになれば、利用価値が高まる
      - ただし、現状の介入実験は、介入期間中に第三者のデータ取得を許さない「お約束」があるので、難しい
    - 逐次的にデータを取るので、データを貯めるデータベースとは相性が悪い
  - 時系列解析
    - センサー等の解析技術。故障検知など。
    - 生命科学では、生存解析。長年のカルテ情報などやIoTの利用に従い増える。
    - 現状は（発現解析など）、時系列が取られていても、多くは高々数点。
    - 最近は数百点あるようなものも出てきていて、今後も増えるだろう。

図８－６

もし、生体データベースが、全生体情報を網羅できたら何が出来るか

- チェスや囲碁のような形での、生体シミュレーションが可能になる。
- 現状のAIの成功事例は、「閉鎖系」「大量データ」「成功失敗が明確」という特徴
  - 前者２つは、広い意味で迅速なシミュレーションが可能である、ということを指している
  - 生体シミュレーションの研究も進んでいるが、逆に、万有引力の法則や量子化学計算の結果を（数式はわからないけど）データから予測してしまう研究も進んでいる
- もし、データが全てを網羅できたら、どのような状態に対しても未来が予測できるようなAIが可能になる
  - 現実、「全て」は無理なので、近似的に全てに近いデータが集まったら、ということ。

図８－７

ビッグデータ・AI時代のデータベース整備とは何か

- まず、AIもデータも土管技術。それだけで、何かができるわけではない。
  - ガスが整備されることで、外食産業ができるようになるように、それを使った研究が展開できる。
  - 見た目は変わらなくても、ガスコンロが電気コンロになるようなことは起きる
- 間違いの無い完璧な少数を集めるより、多少粗くても大量データを集める
  - 狭い範囲の研究ならきれいな少数でよいかもしれないが、生命はそれほどまだ良くわからないことが多い。まだまだ大量のデータを集めて、そこから仮説を出していくフェーズ
- 加えたい人が加えたいデータを、付けたい人が着けたい注釈を着けられる整備
  - 注釈そのものは、データベース同様に、注釈を付ける人の主観が反映される。本来は網羅的に注釈が着けられるべきだが、現実的ではないので、主観が入っていて良いので、とにかく注釈をつけていく。
  - これを繰り返すことで、計算機が注釈を推薦できるようになる
- データの再取得を怖がらないデータベース整備
  - 一度取ったデータでも、目的や機材が変われば、利用法・利用価値が異なる。
  - 常にアップデートできる（ロールバックできる）データベース

図８－８

＜質疑応答＞

（質問）きれいな少数より汚い大量データ、についてもう少し説明を。

（回答）あえて言っているところもあるが、例えば 1 個チャンピオンデータを示して、これは合っていいますといわれるよりは、少しばらつきがあってもいいので 10 点とってこれですという分散がわかったほうが、後々きくということがあると思う。また、近年の AI の学習においても、大きな画像を与えても実は小さなデータに分割したり、小さく圧縮して扱っていたりする。その中で見える傾向としては、AI においては、一個一個のクオリティを上げるより、少し粗くても数多く集めた場合のほうが、できた AI の性能が高いということである。この法則が一般的かどうかは分からないが、AI の学習には成功例だけなく、失敗例も入力してあげないといけないので、きれいな成功例だけでなく、失敗っぽいものも含めた形でのデータベースが必要だと考えている。

（質問）汚いデータというのは言い方の問題かなと思う。例えばメタボロームの場合は、なるべくきれいなデータを出そうとしてはいるけれども、結果的に 1 対 1 の対応関係がつかない。曖昧性の議論がどうしても残る。特にメタボロームの分野というのは、分析化学の人たちが圧倒的に多く、その人たちにとっては、それは汚いとなるので、そんなことを議論していても仕方がないかと。

（回答）まさにそのとおりで、どうしても生物自身曖昧さの部分が残っていくので、無理やりきれにしていくというか、無理やりきっちりやっていくよりはある程度曖昧さを許した上で物事を進めていくことも重要。

（質問）ゲノム生物学に入ってから、いわゆる要素還元式的な考え方というのは大分減ってきている。それでも 95%ぐらいの人が要素還元論で研究しているわけだから、これは避けられない話。そこの認識はデータベースをつくるときの大変大きなポイントだと思う。

（回答）そのとおりと思う。

（質疑）汚いという言葉の定義にもよるが、一方で、人工知能にはアノテーションが必要だという話、その 2 つを考えると、データのクオリティ、シーケンスのクオリティは低くても、それにちゃんとアノテーションがついていればいいという意味なのか、それともアノテーションもなくていいという意味か。

（回答）配列に関して言えばはっきりと 0、1 が決まるものなので、これは明確にきれいな情報が集まっていたほうがよいだろうと思う。一方、トランスクリプトロームやメタボロームあとフェノームに関しては、1 点決めてもそれを定量するのが難しいというか、時間でも値が変わっているし、状況をあらわすのが難しいので少し大き目の情報をアノテーションをつけながら、少し雑でもよいというか、生物のゆらぎも含めて集めたほうがよいだろうというふうに思う。そういうときに、アノテーションを使うことが必要だろうと思っている。

（質問）失敗例は特にこういった学習にとっては、これから非常に重要になってくるのではと思うが、そういうアプローチはあるか。

（回答）今存在するかどうかと言われるとほとんどないように思えるが、実際には機械学習をするときにはどうしてもネガティブなデータも必要、あるいは成功例と成功以外という形での学習もある(半教師あり)。そういう意味では、データが失敗かもしれないし、失敗じゃないかもしれないけれども、いっぱいあるということが重要。

9　生命科学のための画像解析の現状と画像データベース整備を必要とする理由

エルピクセル株式会社　朽名　夏麿

＜発表内容＞

生物現象の画像からの解析というのを我々はやってきたんですけれども、例えばメタボロームとかゲノミクスとかを考えていると、突然ある分子なり塩基配列なり化合物のさまざまな生物種とか培養条件での解析なわけですけれども、画像というのはおもしろいことに、一分子イメージングから果ては人工衛星を用いた地球レベルでの生態系の解析といった、生物現象としては、実は我々が目にすることができるほとんどのスケールをカバーしているというのが特徴です（図 9－2）。もちろんこれは全部区別せず集めて、森羅万象をデータベースにするというのは、まだちょっと今は早い段階だと思います。

もう 1 つ特徴として、生物画像に挙げられるのは、先ほどもありましたが動きもある（図 9－3）。例えば構造解析にせよ分子の動きにせよ、動画像をとらえないといけない。動画です。立体的に CT とか共焦点顕微鏡をとることもありますし、最近ですとマスイメージングといって、質量を縦軸に撮った多次元画像というのが考えられます。また、枚数も非常に多くございまして撮影装置がよくなっていますので、数分で数ギガバイトのオーダーでとれてくる。こういった観点というのは人間にはなかなか 3 次元以上、4 次元以上になってくると見にくいのでコンピュータで何とかしようという話になりますし、枚数が多いということは自動化とかそういったものが重要になってくる。コンピュータで何かできないかという話になってくるというわけです。

一方、もう 1 つの特徴として余りにも大量であるということがあります。イメージデータベースというのがあります。もちろん生物データだけではないんですけれども、ものすごい数の生物データ、さまざまな種のものが含まれていますし、観察方法とか部位で分けてみると、さまざまになるのは当然です。

そもそも何で顕微鏡を見ているのかとか、何でイメージングをしているのかという目的も多々あります。シンプルなところでは、形とか位置を知りたいというところでしょうが、pH インジケーターとかカルシウムインジケーターを使って濃度をはかりたいということもありましょうし、動画像から動きをはかりたいということもあります。ですので、1 つのプロトコルであるいは 1 つの手法で全てのケースを網羅できないという話が先ほどありましたが、まさにこれは生物画像でもいえることです。

生命科学における画像解析のここ 10 年ぐらいの流れを、私なりにまとめてみたんですけれども、まず 1 つは一般的なイメージング装置がすごく普及してきている（図 9－4）。それから、イメージングデータ自体が、それに伴って増加していることです。研究者の数は変わらないけれどもイメージングデータは、種類も数も増えてきている。

これまで論文に出すときも、最後に奇跡となる画像を 1 枚載せておけばよかったものが、統計解析できちんと本当にそれが起きているんですかということを調べないとならなくなりましたし、逆に言えば、単に観察装置としてイメージングデバイスがあったのではなくて、ある種の測定装置としてデバイスが出てきたということになります。一生懸命きれいな画像を 1 個撮るのではなくて、SN 比が悪くてもたくさん撮ることに意義が出てきました。それはもちろん定性的なデータだったものが定量的なデータになるときに、必要なケースです。n を増やしていくということ。

画像解析をするのは誰かという主体の変化というのもあります。これまで誰かウェットな研究者がやってきたものが、共同研究でドライの研究と組むことになってきましたし、最近では電子顕微鏡の撮影は委託が進んでいます。また DNA の抽出とか統計解析は、キット化、パッケージ化が進んだことを考えれば、画像解析、画像データ処理についてもこういったものが進んでいくことは、十分考えられると思います。

画像解析の一例として、例えばこういったもともと撮った画像、原画像といいますが、それから複数のフィルターをかけていって、目的となる、この場合はアクチン繊維ですが、構造を撮って、そこから何か気孔という植物の構造ですけれども、気孔の孔辺細胞のアクチン繊維が気孔に対して何度で分布しているかを知ろうという研究ですが、やっています（図 9－5）。

こういったものは入力と出力があったときに、その間をどういうふうにつないでいくかというのは、これまでは経験と試行錯誤

でやってきたのですが、近年、こういったもの自体をデータベース化するという動きがヨーロッパのほうでは進みつつあります。画像解析手法のデータベース化です。

画像解析全般のお話をするには時間が十分ありませんので、画像解析の中でも最近トピックになっている自動分類とか自動評価といったところに焦点を当てていきます（図 9－6）。先ほどお話もありましたけれども、画像としてデータベースなり何なりが整備されている必要がまずあります。それは従来の昔ながらの方法ですと数枚でよかったんですが、ディープラーニング、深層学習を使うとなると、数千万枚はなかなか大変なアッパーリミットですけれども、多く必要になるということは明らかです。データが多ければ多いほど精度が上がってくるという関係性は、常にあります。

これまで我々は、そういった画像１枚とかあるいは数千枚から、面積とか輝度とかわかりやすい特徴のこともありましたし、もっと複雑な特徴、詳細は今回、説明する時間はありませんけれども、人間には何ともいいようがないけれども、コンピュータは画像から取得できると、数値化された特徴というものを持ってきました。それらが似ていれば、同一種別だろうという仮定に基づいて、分類器、クラシファイアをつくるという流れになっています。分類器のつくり方も従来は、古くは分類規範自体を検討して、こういうふうに枝分かれしていって、最終的にこういうふうになればいいじゃないかというようなデシジョンツリーといわれるものをつくっていましたが、自動的にこういった特徴と答えがわかっている画像があれば、徐々に分類マシンがつくれるようになってきました。

この場合はデータ整備、特徴として何を使うかも重要だし、分類器に何を使うかも重要だったんですけれども、ここ 5～6 年、ディープラーニング、深層によるネットワークというもののインパクトが非常に画像解析の分野では大きくなりました。ディープラーニングの一番インパクトされた分野が画像だったのではないかと私は思ってます（図 9－7）。

先ほどの 2 番と 3 番の仮定、画像からどういった見方をしたものを数値化すれば、それに意味があるかというものを、従来は試行錯誤なり何なりをやっていたんですが、それが面積なのか明るさなのか、それともよくわからない特徴なのかというものを、今のディープラーニングによるアプローチでは、コンピュータ側に任せることができるようになっています。

また分類器の作成についても、どういった決定の方法を使うかというのを考えずに、ディープラーニングにおいては、ある種のネットワークとして最終的には分類ソフトウェア、分類器を作出することもできます。ディープラーニングというからには神経細胞が複数、それも多段、4 段以上に複雑なネットワークなんですけれども、一般に上の段はディープラーニングによって、特徴は正しく考える過程に属していると考えられますし、ネットワークの後段、後の段というのは、分類の基準をつくっている過程だと考えられます。

そうなってきますと、一番のボトルネックというのは教師データの整備です。教師データさえあれば、ネットワーク用にフリーで存在しているディープラーニングのツールキットから自動的な分類ソフトウェア、あるいは解析ソフトウェアが誕生するという流れになってきています。いかにここを集めるかというところが、これからの我々画素を扱っている生物学者のテーマになってきます。

例えばこの論文は、Nature の今年の 2 月に出た論文ですけれども、皮膚がんの画像から自動的にそれがどういった種類の皮膚がんなのかを分類するという方法を、ディープラーニングでやりましたという研究ですが、書いてありますように皮膚科の専門医のレベルの分類ができましたといっています（図 9－8）。実際にこれは、ROC カーブという判断基準を決めるとき、分類基準を決めるときによく使うカーブですが、右上に行けば行くほど高い分類精度を示しているということになっていて、この点々が医者、専門医、青が今回彼らが作成した分類ソフトウェアですが、この青のカーブよりも左下に医師が来ているということは、つまり医師よりも高い精度のソフトウェアができてしまっているということになります。

ここで用いている画像の枚数は約 13 万枚です。13 万枚を集めるのは相当大変だと、我々の経験では思われます。さまざまな倫理規定を突破していかなければなりません。

もう 1 個最近出た論文ですけれども、脳腫瘍の画像を MRI の撮影装置から、どこが脳腫瘍の部分で、それに伴ってど

こが浮腫を起こしている部分かを予測するソフトウェアを作成したところ、やはり人のゴールデンスタンダードと同等、あるいはこれは横から見た図ですけれども、ここがゴールドスタンダードとして与えられた人間による分類結果、右がソフトウェアによる分類結果ですけれども、むしろ人間による分類のほうが一枚一枚やっていくうちに人間の分類規範が多分変わってしまったんですけれども、アーティファクトを含んでいるように見えます（図 9－9）。

こちらでは、先ほどと枚数が数桁小さい。こちらでは 13 万枚でしたけれども、こちらは 100 枚の画像からでも、ディープラーニングがうまくいったという例です。仕掛けとしては小さい、かなりカスタマイズしたネットワークを使ったことと、もう 1 個この場合は画像から 1 つのクラスを分類するのではなくて、1 ピクセル 1 ピクセルを分類するという問題なので、実際にインスタンスの数としては非常に多くなっているということがあると思います。

こちらについては実はコンピュータサイエンスの分野ではよくあると思うんですけれども、テストオブテスト、ベンチマークテストみたいなものが世の中に存在していて、それに対してコンテストを行っています（図 9－10）。そのコンテストのデータを使って、こういった研究が以降も行われているということです。世にある画像のデータベース的に集まっているものを見ますと、純粋に理学的な観点から集めたデータベースもありますし、こういった工学的な観点から集めた画像データベースもほどほど存在しているというふうに見られます。

まとめになりますが、生命科学における画像を初めとした非構造化データというのは、最初のほうに申し上げましたが、多様性というのが非常に大きな特色でして、これまでの画像処理技術に加えて、機械学習とりわけディープラーニングが日増しに存在感を増しています（図 9－11）。また、撮像系がこの研究室にしか世の中に存在しないといった特殊なものでなければ、その結果として開発された技術というのは、医療なり農林水産業への応用・展開は技術上は容易だと考えられます。

また機械学習による解析上、現在ボトルネックになっているのはデータ整備のステップです。特に、アノテーションのついたデータです。何でもいいから画像を集めればいいわけではなくて、その画像が何なのか。例えばある病気とタグづけされているのか、ある遺伝子発現状態とタグづけされているのか。それによって、データによって何がわかるのかは全く変わってきます。

我が国の場合、顕微鏡もそうですし、CT、MRI もそうなんですけれども、ハードウエアに関しては、実は世界に市場を展開して国際競争の中で高い存在感を保っているという状態です。アカデミアにおいても、ある研究室で顕微鏡があるのかどうか、どういう顕微鏡を持っているかというのを見聞きすると、いい顕微鏡を我が国の研究室では持っています。ただ、それを解析するソフトウェアなり、解析する人材というのは欧米と比べて後れをとっているというのが、私の把握しているところです。これらのことから、人材育成も重要ですし、画像についてはデータ整備を行うことによって、解析用ソフトウェアの開発を促進することがまさにできる状態にあると考えられます。

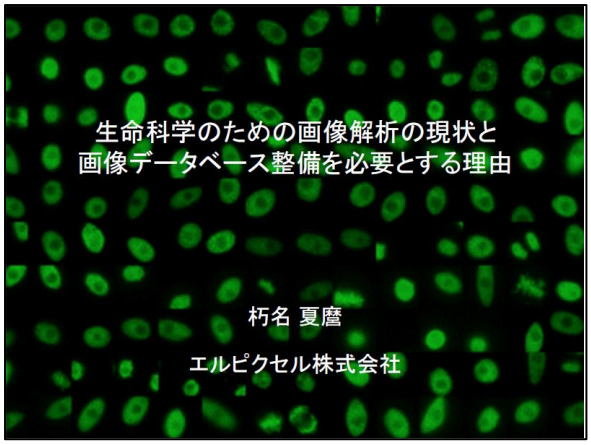


図 9－1

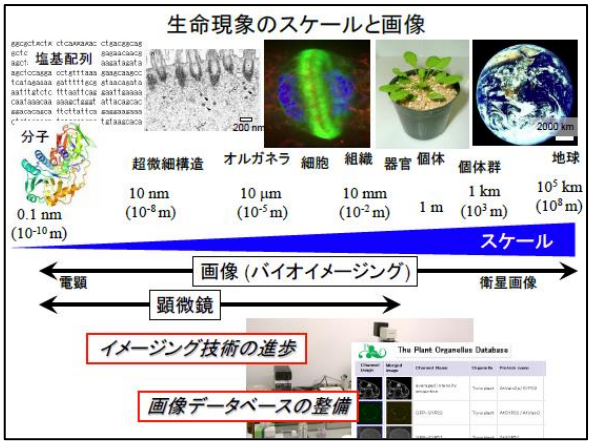


図 9－2

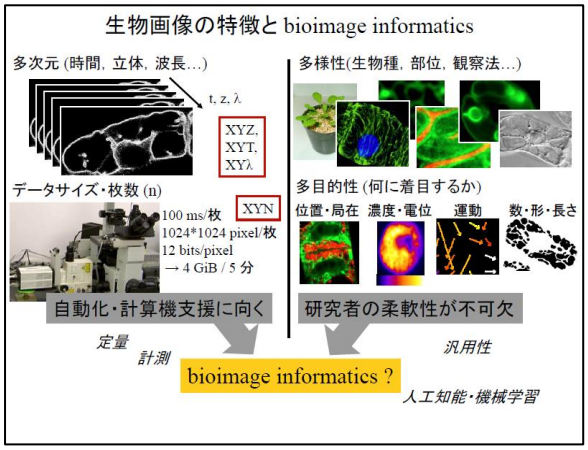


図 9－3

生命科学における画像解析の潮流

* 一般的なイメージング装置の普及
  デジカメ、スキャナ、USB カメラ、スマフォ、監視カメラ…
* 生命科学研究におけるイメージングデータの増加
  各種顕微鏡、電気泳動、質量イメージング、ドローン…
* Figure としての画像から、計測データとしての画像へ
  “奇跡の一枚” → 統計解析、観察 → 測定、定性 → 定量
* 画像解析技術の高度化・専門化
  - 共同研究
  - 委託化（例: シーケンス、プライマ合成、電顕撮影）
  - キット化・パッケージ化？（例: DNA抽出、統計ソフト）

図 9－4

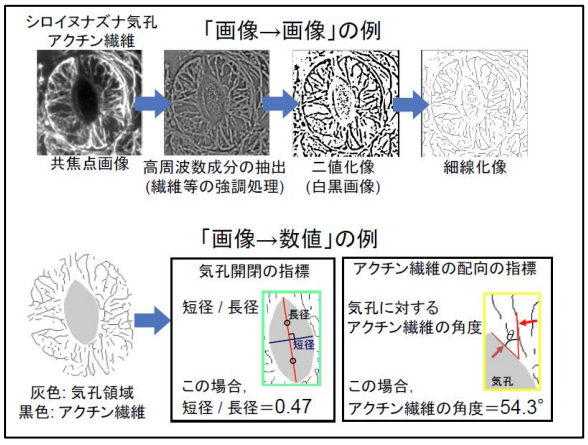


図 9－5

図 9－6

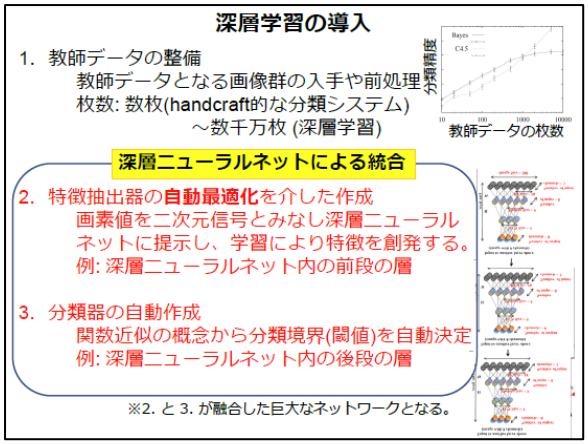


図 9－7

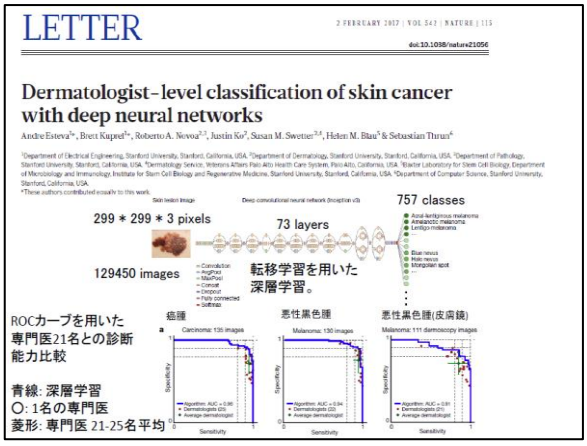


図 9－8

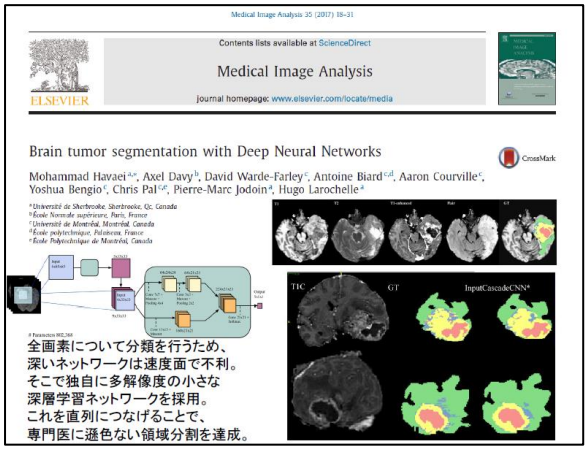


図 9－9

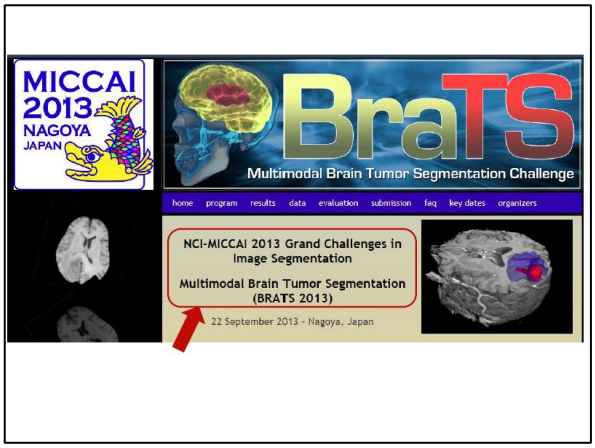


図 9－10

まとめ

* 生命科学における画像をはじめ非構造化データは、その多様性が大きな特色であり、解析手法として従来の画像処理技術に加えて、機械学習とりわけ深層学習が日増しに存在感を増している。

* 撮像系が特殊・高価なものでなければ、開発された解析技術の医療・農林水産業への応用・展開は技術上は容易である。

* 機械学習による解析上、ボトルネックになっているのは「データの整備」のステップである。使いやすいデータベースを一極化することは研究推進に大いに資する（ヒト由来の場合は個人情報保護法の遵守が求められる。画像全般については著作権法が適用されうる）。

* 我が国の多くの撮像関連メーカーは世界に市場を展開し、国際競争の中で高い存在感を保っている。アカデミアにおいても諸外国より高性能な撮像装置が我が国では普及している。一方、解析用ソフトウェアでは欧米に遅れている。これらのことから、我が国で画像についてデータ整備を行う事は極めて重要である。

図 9－11

＜質疑応答＞

（質問）どこまで精度を求めるかということと、データ量との関係についてどう思われるか。

（回答）実際、作成する過程においては、そもそもどのパーセンテージの分類精度があれば十分なのかというのが、先に与えられるべき情報であって、まず最初に 100 枚とかそのぐらいの枚数を集めて、だんだんこれを外挿して、それぐらいの枚数を集めれば所定の目的が達成されるのではないかというアプローチをとるのが理想的です。

（質問）中途半端な画像データベースをつくっても意味がないけれども、ある程度の数が集まれば意味があるということになる。問題にもよるでしょうけれども、100、200 の画像データベースをつくっても意味はないけれども、例えば万のオーダーの画像データベースがつくれるのであれば意義があるという理解でよいか。

（回答）おっしゃるとおりだと思う。動画に関して言えば、研究した方々が一生懸命手で打ったものというのは、まさにアノテーションなので、それと画像というのがペアになったものがあれば、コンピュテーショナルなアプローチによって、その学生の方々や院生の方々がやったことを再現するようなソフトウェアをつくるというアプローチがある。

（質問）例えばスキンキャンサーということで、大量のデータを使って予測した場合、さらに別のキャンサーにといろいろやっていこうとしたときには、また、ゼロからやっていく必要があるのか。

（回答）基本的には、この画像だったらこのレベルが必要みたいな経験則はある。もちろん、転移学習といって、似たような画像であれば実はちょっと枚数を減らすというのは、アプローチは既に確立している。例えば組織設定の画像に関していうと、ある程度胃がんだろうが食道がんだろうが、片方であればもう片方でも、少し転移をすれば似た画像であればそういうことは可能。この皮膚がんの例では、自然画像といって、あらゆる一般の物体の画像を使ったものを、これよりも多分 3 桁ぐらい多い画像であらかじめ学習した上でさらにこの枚数を学習している。

（質問）医療現場のリソースはバイオプシーをとるのもすごく大変で 1 回しかとれない。あれをいかに少ないサンプルで学習させるかというアプローチはあるか。

（回答）なかなか難しいが、案外少なくても 1000 枚ぐらいまで許されれば、それなりにワークするなと思う。ただ、やはり 10 枚では無理だなというそういう経験則はある。

IV. 総合討論

有識者からのご発表の全体まとめとともに、主に以下の 3 点について総合討論を行った（司会：高木センター長）。

1 有識者発表全体まとめ

【司会】いろいろお聞きして大変参考になりました。ありがとうございました。これをより具体的な活動、統合化推進のどういう分野をどういうふうに整備していくかとなると、なかなか難しいところがございまして、もう少しそのあたりを具体的に少しお聞きできればと思っております。

今日お話があったのは基本的には研究開発が世界的にどういう方向に進んでいて、どういう DB が必要かということだと思いますけれども、それを今日お話のあった分、全て対応することもできませんので、その中で少し優先順位をつける、あるいはどういう考え方で整備をすればいいかということで、お話をお聞かせいただければと思います。

例えば先ほどの画像データの皮膚がんのデータでも、10 万枚集まるということがなければ、DB をつくろうと思っても意味がないわけです。そういう意味では、今日お話になったところで、実際にデータ、きれいなデータか汚いデータかはともかくとしまして、出していくということがある程度見込めないと、DB として我が国として整備していくことはなかなか難しいという面もございます。そのあたりに関して、我が国としてこの分野は重点的にやったほうがいいとか、そういうことでもう少し絞って、もしご意見があればお聞きしたい。本来であれば私どもの考え方としては、DB でこういうものが大事だからこういうデータ生産プロジェクトをつくってほしいというのが、我々の DB の立場なんですけれども、なかなか国の施策としてそういうわけにもいきませんので、そのあたり独立しているところもございますので、ある程度データ生産が見込めて、我が国の強みなりどういうところを今後整備していけばいいか、もしお考えがあれば、お聞きしたいと思います。

2 どのような優先順位で DB 整備を進めていくとよいか

○さっきの画像のデータですけれども、これを一気に何千万枚集めるというのは非常に難しいと思います。たとえば、私たちも透明化技術を使って、がんの組織を透明化してそれを共焦点レーザでスキャンし、また人工知能を使って、漏れのないようながんの診断の技術の開発をやっています。その最終成果は、基本的には NBDC に登録しようと思っていますので、このようにコホート研究のデータをすべて NBDC に統合されるような仕組みはとても大事なのではないですか。一気に全部はできなくても、中長期的に企画していただきたい。多分製薬会社とか医療関係者は、みんなそれ（コホート研究データ）を欲しがっているのではないかと思います。ぜひそういった医療関連の画像や、ゲノムともリンクした形でできますと、さらに創薬につなげていけるというようなところがありますので、とても重要だと考えています。

○繰り返しになりますけれども、データの規模が個々の研究者ができるレベルではなくなってきているので、そこをつなぐ仕組みとして計算資源も各大学の情報基盤センターが持っているもの、なけなしのものを使っているみたいな状況ではなくて、使いたい人がお金を出せば使えるような形にしていかないと、そこはもうそれ以外で何か新しいものをつくるというよりは、本当にインフラとしてそこを強くしないと、日本の研究者が研究できなくなっているんじゃないかと思います。

○1 次 DB と 2 次 DB ということを考えると、1 次 DB は、先ほどの画像の話と高木先生の話を聞いて、確かに足りないのはそこだなというのは思います。理研 QBiC のほうでハイスループットの一分子計測の機械がつくられていますけれども、それを解析するツールが足りないというふうに思っています。イメージングデータは、日本のお家芸です

ね。アメリカに負けつつあるとは言いますけれども、やはり基礎の顕微鏡をつくるところは上手だと思いますので、解析ツールもあるとよい気がします。2 次 DB に関しては、日本で GWAS のデータなどをより高次に、実際の生物機能に関連したところで、タンパクのこの辺が変異になっているから危ないんだとか、そういうのが DB でわかるようになると、その都度、MD シミュレーションをやらなくても済むのでいいと思います。

○ゲノムから RNA タンパク質、その次にいろいろと上がっていって特にイメージングはサブセルラーのレベルから地球レベルまでとれると申し上げましたが、やはりだんだん上のほうにいくにつれて、種ごとの違いというのがどんどん明らかになっています。細胞のレベルぐらいまでは何とか違う種でも同じようなものが起きる、レベルががっと上がっていって、積み立て式じゃないですけれども、イメージングにおいて最も近寄れるのは、まずは細胞前後のレベルではないかと思っています。

○今後のいろんなことを考えますと、バイオエコノミーといいますか、どう産業界、あるいは基礎研究においても全てにおいてそういうところに貢献できるかというのを示していかないと、国からのサポートは得られ続けられないと思うんです。やはり産業競争力にどう寄与していくか。バイオエコノミーというのは、世界が推進しています。その上で何かということなんですけれども、1 つそれは何か新しいものをつくっていくための基盤にもなり得る基礎的なことということで、DB 的にいって先ほどありましたように、ユーザが検索しようと思っても検索できない情報というのがまだまだ多かったり、例えば先ほども言いましたけれども、酵素反応というのは、そこから微生物にたどり着けない。検索が十分でない。あるいは遺伝子についての付加的ないろんな情報が十分ついていなくて、使えないという DB を充実させていく、あるいは検索できるようにしていく。あるいは、だんだん合成というか、何かをつくっていこうという話になったときに、ゲノムをつくるとかいろんなことが出てくるんですけれども、それに使える新しい時代のデータベースがまだ欠けているところがあるんじゃないかという話です。もう 1 個は、AI という時代になってきて、学習のデータセットということ、いろんなデータを集めるときにどこまで何の目的で集めるのかという、学習なのかそれとも百科事典的にそこに存在していることに意味があるのかとか、そういうことを考えて集める。タイプが違うものが存在するのではないだろうかと思って、その目的と集め方とどれを優先するかというのは、かなりリンクする話だと思うので、そこをしっかり議論した上で戦略的にやらないといけないのでないかというふうに思いました。

○私は今回 3 つありまして、まず優先順位をつけるという話をしたときに、メタボロームなんてそもそもやっている人が少ないから、データ量が圧倒的に少ないんです。なぜかというと、特殊な装置が要るんです。大学なんかで買うという意味では、結構高いです。実は、食品会社とか医薬品会社は全部持っているわけです。その中に全部データがたまってはいるんです。それをどうやって集めてきてどうやるか。それをオープンにするかどうかというのは、国の戦略とかかわってしまってくるだろうと思います。ただ、国の戦略との関係という中で、バイオエコノミーの中でどう位置づけるかという議論が必要かなと。その中で、データ量は少なくても残すべきものはちゃんと残さないと、後々禍根を残すであろうということが 1 点です。2 点目は、1 番の優先順位の前の問題で 0 番というのがあって、皆さん、優先順位をつけるのはなぜかというと、政府から予算を削られていますと、あるいは今の産業界は役に立ちませんと言っていますと、それで皆さん、四苦八苦しているわけです。今回のワークショップもそうされたわけですけれども、そうではなくて実はデータベースはこれだけ重要だから、予算を拡大してくださいと。そのためにどうするんですかというような議論、その視点がどうしても必要かなと。バイオ予算を倍にしようとしているときに DB がなかったら、何もできませんと言えばいいわけです。だから予算を増やすための戦略を、皆さんが出せばいいのではないか。3 点目は先ほど汚いデータの議論があったんですけれども、邪魔くさいデータは皆さんとらないんです。私の研究室は変わった人しか採用しないということもあるので、そういうことを 3 年間、5 年間、6 年間と平気でやる人しか僕は採用しないんです。そういう人たちはなかなかペーパーを書けないという切実な問題もあります。そういうシステムの問題、つまり

日本の学問としてどう考えるかみたいな評価の仕方も必要で、少なくともそういう人たちも世の中にいるので、その人たちをうまく何とか採用できるような、生活できるような仕組みづくりというのも必要かな。

【司会】AI の方がいて、でもなかなかバイオの DB はとっつきにくい等、いろんな障壁があると思います。今日お話になったこともそのうちのいくつかだと思います。その意味で、今後、AI 研究者がとっついてくれて、何か成果を出すにはどういうふうにデータベースを整備していけばいいでしょうか。

○非常に難しい質問です。かつ自分も解析をしたいといいながら、いろいろハードルを越えないといけないことが多かったです。まず恐らく AI の人が植物の話とか微生物の話といってもなかなかとっつくのは難しいので、一番やりやすいのは多分ヒトの話なんですが、ヒトの話は一番プライバシーに関連するので、データはとっつきにくいと。少なくともGWAS なりのデータなり情報がある程度集まって、こういうふうに手続きを経れば扱えるんだという手順が明確になっているというのが、一つ重要かなと思います。そのときに東北はこれで、バイオバンクジャパンはこれでみたいな感じでやっていると大変なので、できれば一括の場所があるのが望ましいかなというふうに思っています。それがデータを扱うという上では重要かなと考えています。

【司会】フィールドオミクスというと、何か探索空間が広過ぎてどこを集めればいいのかというのは、なかなか DB にするにしても非常に歯抜けの DB になってしまうのではないかという懸念もあると思うんですが、そのあたりも含めてどういうふうに整備していけばいいでしょうか。

○例えば今形質データというのは例えばヒトの背丈をはかったりとか、あるポイントを見てそれを数値に直しているのが利用できるデータになっているんですけれども、先ほどの画像の話でいうと、1 回画像を撮っておけばそこからはいろんな計測値が抽出できるわけです。ドローンとか先ほどのフィールドオミクスみたいなのがあるんですけれども、今までの形質評価というのは、どちらかというと人がとりたいと思ってとっている数字データしかないものが、画像データとしてとることによって、後からいろんなところに注目して新しいパラメータがそこから生まれてくると思いますので、今まで人が手でやっていたものをまず画像で撮るということから始めると、形質の評価が、AI とかを使うことによってもっと今までわからなかったところが、実際にはそこに関係しているんだということが見えてくると、もっといろいろと評価軸が増えてくると思います。ですから今あるゲノムあるいは発現 DB と形質データ、今まではなかなか結びつかなかったものが画像をうまく利用することによって結びつくような形でいけば、いろんな品種改良のターゲットみたいなものとして、あるいはそれは、大学の先生にとっても、実際の圃場でのいろんなものと遺伝子の動きというものが、それによってうまくつながるという方向で発展していくのではないかというふうに思っています。

○そもそも論ですが、優先順位をつけるということを聞いたときに、何を目的として優先順位をつけるのか。そこで、応用につながるものを目指すと言われて、疾患とか創薬などの言葉が出てくると、AMED とかぶるんじゃないかというふうに思ったんです。AMED でも AI 創薬をやっていますので。AMED ではできないことをやらないといけないのではないかなと。あまり創薬とかそれだけに特化するようなことを考えるのではなくて、いろんな分野の人が、例えばプロテオミクスだったらプロテオミクスをやっているいろんな基礎の研究者も入れるような DB をつくれることを考えてほしいと思います。

3　データ整備を進めるに当たっての実現性、問題点について

【司会】データ整備を進めるに当たっての実現性とか問題点について、特に人材の問題とかは大きいと思います。人材というと先ほどちゃんと評価してポスト、給料をちゃんと用意すればいいんだろうと思うんですけれども、なかなかそうはならない仕組みも我が国にはあると思います。これだけは言っておきたいみたいなことがございましたら、ぜひご発言いただきたいと思います。製薬企業などは人集めは大変ですか。

○おっしゃるとおりで、例えば人工知能の利用とか、実は創薬でも細胞等の画像データはたくさんありますし、それをAI でトライしようと思ってもそういった人材がなかなかいない。グーグルとかヤフーなどの IT 企業に全部人材をとられてしまって、なかなか我々のところに回ってこないという、非常に困っている状況です。NBDC の本業ではないかもしれませんけれども、NBDC で、それなり分子生物ドメインナレッジを持っている方もいらっしゃいますし、さらにバイオインフォマティクスの人材もたくさんいらっしゃると思いますので、今後ともこういった人材育成も 1 つのポイントなのではないかと考えています。

○プロテオミクス研究者に関しては、日本プロテオーム学会に参加している研究者の中でまとまっているんです。まだ新しい学問というのもあるんですが、プロテオミクスでやっている人は結構学会に参加していて、必ずその中でデータベースとかインフォマティクスのセッションもつくっています。そういうところで話してくれる人を集めて、それに伴って聞く人も来るということで、学会主導で人材が育成できるのではないかなと、我々は思っています。

○うちのラボでは、自分たちでデータをとって自分たちで解析するというのが多く、高度な情報技術は自分たちでは持っていないんですが、そこは連携でやっていくというやりかたをしています。だから 2 段階ぐらい、うちのラボの生物がわかって情報もある程度わかる人、向こう側には、情報がわかって生物がちょっとわかる人という、間々に人がいるという感じで、動かしています。今まで理研にいて、阪大に行って思うことは、ドクターに行きなさいと先生方は思うんだけれども、学生さんは 6 年も長いと言ってあまり行かない。だけど見ていると情報解析、生データ解析ぐらいだと、修士の学生さんでも十分やっていける。入ってみないと学生さんもわからないけれども、一般的に安定だからといって企業に行くんだけれども、やはり心の中では基礎研究をやりたいと言ったりする。そういうところをうまく利用して、ぐるぐる回るようなコミュニティがあると、そんなに PhD、PhD といわなくても、人材はある程度回せるのかと思う。ただ、そのときにある程度お給料がよくないと、扶養家族ができたりすると、低い給料ではアカデミックでやれないというのもあるし、そこら辺が見ていて何か悩ましいなと思いますけれども、もうちょっと企業との間の風通しはよくしてもいいのかなという気がします。

○今のアカデミックな人材というのと、もうちょっとデータサイエンティストみたいな、SE というとちょっと言い過ぎかもしれませんけれども、データを動かすところの人材が、特に生命科学の分野で、今までは Computational Biology なりバイオインフォマティシャンと言われていた人たちがカバーしていたわけですが、それもカバーできなくなって専門職としてデータサイエンティストが全然いないというところが、今インフラを非常に弱くしているので、そこを DB をつくるのに DB を管理する人間だ何だというのが当然要るわけです。それをつなぐ役割とか、クラウドのメンテナンスとか、そうするとそこまではアカデミアのほうでは、今は日本の大学もそういうところに全く人材を割かないですから、そうするとそこが今ライフサイエンスをやる上でのすごく大きな、日本の研究者はみなそこで世界から後れをとってきつつある。そういうウェットの、あと機械のほうの、サーモのオービトラップとか自分たちのところで回していますけれども、もともとのあれをつくっているテクノロジーに対しては、情報解析のプログラムは非常に弱いですね。だから自分たちでつくらざるを得ない。そういうところ、でも日本の会社はもっと弱い。FEI の機械とかを見るとそこでの情報解析のところに割く力。日本はものづくりだ何だといって、機械を磨くところだけを考えているけれども、そこで現場でデータの欲しい、実験をやっている人たちは、ブラックボックスか何かできれいな処理されたデータが出てきてくれたほうが、画像処理も含めて、あるいはスペックのデータも含め、そこで日本の機械は、生の性能はやたら電圧は高くでいいんだけれども、何かデータをとるスピードは遅いよねというところでだんだん売れなくなるということで、そこの情報処理のところ。別にこれは高級な情報科学のテクノロジーではない。ただ、データサイエンスという新たな意味で、ちょっと日本の企業のほうにもない、何か育てていかないと、足腰が今は本当に弱くなっているかなという気がします。

○今ありましたように、データ整備をする方の人材も確かに足りないんですけれども、使う方が使おうということにいか

ない。教育がなされていないので、結局どんなにデータ整理をしても、ほとんどマジョリティのバイオの人は使わない、使えない。そうすると使われていないんじゃないかという話になって、そこはお金が出せないという話というのが循環するような気がします。周りを見ても、我々のところもかなり強制してやっとさっきのようなデザインツールとか、いろいろ使いやすくやって、やっと強制して使うということがあります。この前ケンブリッジに行って、合成バイオというのは新しい博士課程のプログラムを見たら、1 回情報の人はバイオの実験、バイオの人は情報のプログラムを必ず受けなければいけないみたいな、割とそういうことが世界でも結構行われています。ですから、ここで言ってどうかわからないことですけれども、使うことに抵抗がなくみんながどんどん使って恩恵が受けられるような入りやすいインターフェースを持った DB が必要であるし、それを使う側の教育、我々の責任になるんですけれども、もっと使ってできる。そうしないと世界からどんどん遅れているというのは、多分データを使いこなせない、DB を使いこなせない。インフォマティクスというか、こういった情報基盤を使いこなせないというところをもっと声を大きくして言っていかないと、結局、いくら整備しても、日本の中で使う人は少ないというみたいな状況にならないようにということは、声を大きくしていかなければならないのではないか。いろんな人たちがいろんなところから言っていかないといけないように思いました。

4　DB 整備の推進方策について

【司会】今私どもはいろんな分野ごとの DB をつくるのに、統合化推進というファンディングのスキームを使っております。それが年間 3,500 万円で 5 年間という今のプログラムですが、それで植物とか微生物とかプロテオームとかそういう整備をしているんですけれども、3,500 万円ではとても何もできないというお考えもあるかもしれませんし、かといってそれを大きくしますと分野が非常にまたそれこそ 2 つ、3 つしかできないとか、おっしゃるように予算を拡大できれば違うんですけれども、どうしても限られたパイの中で議論をせざるを得ない。

○そうだと思います。現実的な選択肢というのはあると思いますけれども、予算を拡大するという議論はどうしても必要で、私はちょっとこの NBDC のパンフレットを見ていて、これではだめだなと正直思いました。これは結局ポータルサイトをつくっているだけみたいに見えてしまうわけです。先ほどの産業界の方々の中には、NBDC の活動をあまり評価していない方々もおりますが、1 つ原因をいろいろ見ていると、誤解が結構あると思うんです。ちゃんと発信していないんじゃないかという感じがします。これは一体誰に対して発信しているんですかというのが、明確じゃないんです。つまり、基礎研究、基盤研究はやっている人に対して発信している議論なのか、一般の国民、税金を払っている人に対してやっているのかはっきりしていないんです。NBDC は実施していることをちゃんと発信していくと。そのことがデータ整備をしているということの意味合いが、明確になっていくと思うんです。そのことが、結局人材の確保とかにもつながってくるような気がします。先ほどの人材の話ですけれども、データ整備というのは、私の研究室もずっとやっているんですけれどもかなりベタな、本当に単純なことをずっとやっているわけです。これも 1 年、2 年とかではなくて 5 年以上とかかけてずっとやっているわけです。ものすごく大変なわけです。私の立場では、それを例えば共同研究みたいな形で、うまくその人たちを共同研究のテーブルに入れてもらうとか、そういうことをしながら何とかやりくりをして人材に業績が残るようにしているんです。ただ、そういうことも含めて全体の宣伝といいますか、とりあえず予算の拡大が必要だと、産業界に思わせるようなことはぜひとも考えていただきたいと思います。

5　その他

○農水省のほうではこれまで農水のプロジェクトでいろいろな DB をつくってきて、例えば RAP－DB などはアノテーションもすごくしっかりしていて、使いやすい DB なんですけれども、なかなか育種現場ではそこが今まで使われていないということもあるので、その間をつなぐようなもの。今せっかくあるいろんなもの、表現型の DB も含めて、現場で使

えるような形のインターフェース的なものといいますか、そういうものまではなかなか個別のプロジェクトでは出来ません。個別の DB はできるんですけれども、それをつなげる、統合して例えばワンストップというと言い過ぎかもしれないですけれども、現場の人にここに行けばこういうことが出来る、実際にバイオインフォマティクスがよくわからなくても、自分の欲しいものがここからとれるみたいな、そういうものの整備がとても必要で、それができれば随分 DB そのものの意義も上がってきますし評判も上がるんじゃないかなと今思っています。

○全体としてですけれども、1 つ言えなかったことは、先ほどおっしゃったように、データ間のリンクの話がやはり重要だと思っています。個々の DB が散逸している状況で、この遺伝子とこの遺伝子は一緒であるとか、この形質、この形質は一緒であるというような、そういう DB それぞれの間をつなぐ整備は一つ必要なのかなと思っています。そうすることで、見かけ以上は一個一個小さいんだけれども、全体としては大きく見えるという DB になるので、AI という意味でも使いやすくなっていくのかなという印象を得ています。そこら辺の整備が一つ必要かなと思っています。NBDC 全体も AI も、先ほど少し触れましたけれども、全体が基盤技術になっていくと思います。なくなってしまったときに初めて、みんななくて困ったというふうに思うと思うんです。情報発信を含めて、もう少しどうすればいいのかはなかなかアイデアはないんですけれども、基盤でみんななくなると困るという言い方をするのはいいかどうかわかりませんが、ちゃんとそれが使われて必要なんだということをちゃんと説明していくことが重要なんだろうと思っています。

○実はバイオ戦略は 2001 年、2002 年ぐらいに、最後に政府が出したんです。今回政府が出して、もう 15～16 年たっているんです。実は私はこれが最後のチャンスだと思います。逆に言えば、今頑張れば、今というのはどういうことかというと、来年夏前です。各省庁の予算で動くわけです。それまでにとりあえず予算の拡大に関して、かなり真剣になってやる必要があるし、こんなことはほとんどない時期なので、そのチャンスを逃さないようにされたらどうかなと思います。

○ここに行かないと検索できない。総合的な検索というか、いろんな方々、産業界の方々にきちんとインタビューされて、個別に本当に使っている方に聞かれて、こういうワークショップではなくて、もっと現場の人たちの声をよく聞いて、何ができればここが使われるんだろうという、先ほど言ったように、酵素から微生物に行けないとか、あるいは遺伝子の名前はあるけれども産業的に使いたい情報が何も、そこら辺が充実していないとか、ここでしかというようなことがいろいろあると思うんです。それを充実させるというのが 1 つと。先ほどの議論で AI という時代になってきたら、いろんなデータが必要になってくるわけですね。正しいデータだけではなくて、何回も議論がありましたみたいに、正しくないデータもないと使えないという、そういう DB というかデータセットといったほうがいいんでしょうか。DB ではないかもしれませんけれども、目的がかなり違うものが、バイオ×デジタルという時代の中では求められてきていることは間違いないです。ただ、その時代が求めているものに合わないものをつくったって、何だ、データをいっぱい集めているだけと言われたら、予算の拡大どころか、激減の可能性もあると思うんです。言われたようにチャンスだと思います。バイオ×デジタルにちゃんと符合した形のものが提案できれば、それが多分世界が求めているもので、それが大きく変わろうとしているので、そこら辺が非常に重要かなと。何をプロジェクトでやるんですかとか、どういう人材を育成するんですかに全部関係してくると思いますけれども、チャンスでもあると思います。

○画像に関して言うと、いい顕微鏡を持っていて、たくさんとれるラボというのと、インフォマティシャンがいる研究室というのは、かなり違うことが多いんです。なのでコラボレーションでやっていくのが一番自然な形かと思います。そういったことを助長するような予算枠になっていると、金額なのか、それは条件なのかわかりませんけれども、よいのではというふうに考えています。弊社で 50 近くの大学を回って学生とかスタッフ、院生とかに画像解析のワークショップというか講義みたいのをやったことがあるんですけれども、アンケートをとると、授業とかで画像処理について学んだことがある学生というのは、ほとんどの人がウェットなので、1%ぐらいしかいないです。それではなかなか正しい処理とか DB

の作成というのは、できないだろうと思います。

○ある程度予算が限られた中で言えば、つなぐ DB があまりない。とにかく、個々の DB に行って、その都度目的に応じて検索せざるをえないので、つなぐものがあるといいだろうというのはまず 1 つあります。またそのときに、フラグシッププロジェクトみたいなものをつくってつなぐ。ただ、漫然とつなぐといっても、新しい生物的知見を求める場合には RDF 化したらつなげるというものでもない、生物の機能とある程度アウトプットを求めてつないでいく、そのために具体的なプロジェクトがあるといいんじゃないかと思います。例えば NCI 全体で、Ras というがんに関係する分子のプロジェクトをやっているというのを聞いたことがあります。今さら Ras なのかとも思うんですけれども、やはり病気に重要、だけどわからないことが多いから、やっているらしい。またさらに、さっきお話があったように、なにが足りないのか、何が必要かに関しては、各研究者からヒアリングをする必要があると思います。さらに、生物学はデータをたくさんとるんだけれども、データをとってそこで終わっているというのがすごくある。でも、いまは、それをつないで人工知能でも何でも使って、新たな知見の予測をしたいわけです。データをとってそれで終わりではないので、予測を可能にするようなもの、つないで予測するというところも合わせて、次の DB のプロジェクトして予算化も考えていく必要があるのかなと思います。

○大体議論は尽くされてきているんですけれども、やはり使ってもらえる DB にする必要があって、やはりデータがあって、それが今までは使いたい人はまた登録してダウンロードしてくださいとかいうのが難しくなった。そうすると、個々の研究者がデータを見たいときに、全部のリレーションを見たい、もちろん人工知能を回すときには全部のデータを手元に持ってくる必要か、コンピューターの上にある必要があると思いますけれども、その一部だけでも、ヒトのデータになると個人情報云々の問題があって、ここの遺伝子だけ、あるいはこの領域だけをサーチしたいというものに対応するような、例えば遺伝子の発現であれば、お母さんの染色体とお父さんの染色体は、やはり違うでしょうというようなところのデータだけを見たいという人は、いっぱいいると思うんです。いわゆる GA4GH なんかがやっているようなモレキュラービーコンとか、そういうアクティビティとうまくタイアップしながら、特定の場所だけの、そうすると結構そういうところからこういうのがわかるのだったら、全体をまた見てみようとか、データを本当に全部は見られる必要はないと思うので、種 1 つでは少ないと思うので、ただユーザをふやしていくという。そこにいろんなアノテーションをつける。今だとみんな NCBI とかサンタクルーズのブラウザにどうしても行ってしまいますけれども、その中で日本人固有の、例えば SNP 情報に関するファンクショナルなデータは NBDC のここに来ればとか、そういうようなここにしかないコンテンツ、アノテーションを含めたものが検索できるとかいうのがあると、そこに全部資源を投入する。どうしてもゲノムやエピゲノムにちょっとバイアスしますけれども、といいかなと。もちろんあとは、全体のクラウド化するというようなところは別個に、それこそ予算を拡大してやっていただけるといいかなと思います。

○私からは 2 点ほど。1 点目は統合データベースというのは、データを同じところに集めるというふうに思いがちけれども、実はその必要がなく、要はインターフェースをつくればいいので、つまり物理的に同じところにデータを集める必要はなくて、先ほど US バイオバンクベンチャーのように、インタフェースによる統合で十分です。高木先生もそういう仕組みをつくられているとおっしゃっていますが。このような統合は、私は非常に重要ですし、そうするとユーザー数もふえてくるのではないかと思っているところが 1 つ。もう 1 つ、やはりこれが役に立っているという実例をたくさん出していくしかないと思っています。産業界、応用分野との共同研究例をふやしていくのとその実績を発信していくことで、たとえば、NBDC のこのデータを使って、こういう結果が得られましたとの実例をたくさんつくっていけば、予算などの話もしやすくなると思っておりますので、ぜひそういったところも考えていただければと思います。

6　おわりに

【司会】いろいろと貴重なご意見をありがとうございました。本日ご指摘いただいた点は、実は既に私どもで取り組んでいたり、あるいは計画中のものだったり、いろいろなものがございますけれども、多分そのあたりのプレゼンテーションの下手さも、言っても言い訳になりますので、今後実例を、パンフレットの作り方も含めて、ご指摘に従うように応えられるようにしていきたいと思います。予算の拡大に関しても、本日のご意見を建設的に捉えて進めていきたいと思います。本日はどうもありがとうございました。

V. 付録

1　ワークショップ概要

1.1　開催概要

件名：NBDC で今後取り組むべきデータベース整備の検討

日時：2017 年 11 月 5 日（日）13 時～17 時

場所：JST　東京本部　4 階会議室

1.2　目的

NBDC では、データベースの統合化を進めているが、今後、応用につながる具体事例を強く念頭に置いた際に重要となる領域に焦点をあてて基礎研究データの統合を行いたいと考えている。今後どのような領域、分野に焦点をあてて、どのようなデータベース（2 次データベースも含む）を整備していけばよいのかについて、有識者から意見を伺い、議論を行う。

それにより、利活用可能な国内外のデータベース整備状況の把握、および、NBDC で取り組むべき課題、特に統合化推進プログラムで今後重点的に推進すべき分野・領域を、短期・中期的な時間軸も踏まえ抽出する。

1.3　出席予定者（敬称略）

1）　外部有識者

青島　健	（エーザイ株式会社　hhc データクリエーションセンターデータサイエンスラボ　部長）
油谷　浩幸	（東京大学　先端科学技術研究センター　ゲノムサイエンス分野　教授）
岡田　眞里子	（大阪大学　蛋白質研究所　教授）
朽名　夏麿	（エルピクセル株式会社　研究開発本部　取締役）
近藤　昭彦	（神戸大学　大学院科学技術イノベーション研究科　教授）
柴田　大輔	（かずさ DNA 研究所　バイオ研究開発部　部長）
瀬々　潤	（産業技術総合研究所　人工知能研究センター　機械学習研究チーム　チーム長）
髙野　誠	（農業・食品産業技術総合研究機構　生物機能利用研究部門　主席研究員）
朝長　毅	（医薬基盤・健康・栄養研究所　プロテオームリサーチプロジェクト　プロジェクトリーダー）

2）　NBDC

高木　利久	（NBDC　センター長）
長洲　毅	（NBDC　統合化推進プログラム　研究総括）
星　潤一	（NBDC　企画運営室　室長）
舘澤　博子	（NBDC　企画運営室　研究開発推進グループ　調査役）

1.4　プログラム

13：00～13：05　開会挨拶

13：05～13：15　ワークショップ企画趣旨説明

13：15～14：45　個別発表（発表 10 分、質疑 5 分）

青島　健　（エーザイ株式会社）

「医薬品開発におけるビックデータの活用」
油谷　浩幸　（東京大学）
「ライフサイエンス基盤としてのゲノムクラウド」
近藤　昭彦　（神戸大学）
「有用物質生産に有効なデータベース整備」
髙野　誠　（農業・食品産業技術総合研究機構）
「育種に役立つデータベース整備」
柴田　大輔　（かずさ DNA 研究所）
「メタボロームの視点から」
朝長　毅　（医薬基盤・健康・栄養研究所）
「プロテオミクスデータベースの必要性と今後の方向性」
（14：45～15：00　休憩）
岡田　眞里子　（大阪大学）
「システムバイオロジーとバイオデータベース」
瀬々　潤　（産業技術総合研究所）
「統合 DB のビックデータ・AI 利用に向けた考察」
朽名　夏麿　（エルピクセル株式会社）
「生命科学のための画像解析の現状と画像データベース整備を必要とする理由」
15：45～17：00　総合討論

以上