# 理研ライフ系データベース基盤の長期的な方向性に関する提言（最終報告）

平成 27 年 10 月 29 日
データベース作業部会

目次

## 1. これまでのデータベース作業部会の経緯

データベース作業部会では、平成 26 年 8 月、理研内オンラインアンケート(回答数約 200)に基づく理研データベース基盤の短期的な方向性に関する提言(中間報告)を行った。その概要は下記の通りである。

i. 情報基盤センターは、オールジャパンのデータベースカタログの活動とも連携して、理研の研究者が作成したデータベースを見つけやすくする仕組み(電話帳機能等)を整備すべき
ii. 各々のデータが何かを記述する「メタデータ」を体系的に管理する技術開発を情報基盤センターが主体的に行うべき
iii. データ統合ではなく活用を第一義としたデータ共有・公開の枠組みとなる新たなデータベース基盤を、情報基盤センターが主体的に検討すべき

しかしながら、上記中間報告では、生命科学研究、データベースの世界的な動向に基づいた、長期的、具体的な提言に至っておらず、作業部会は、そのような提言をまとめるべく、さらなる検討が必要だと結論した。平成 27 年 2 月に開かれた平成 26 年度第一回バイオインフォマティクス委員会では、この方向が承認され、データベース作業部会から本提言を行うこととなった。

## 2. 生命科学および医科学研究、データベースの世界的な動向

### i. データ本位生命科学

生命科学、医科学では、従来より研究者の仮説を証明する「仮説本位」アプローチが主流であった、しかし、近年では、網羅的なデータを大規模に取得し、統計的事実に基づいて新たな生命現象、あるいは仮説を発見していく「データ本位」アプローチが様々な分野に広まっており、今後の生命科学を先導していくと考えられる。以下に、各分野における動向を述べる。

#### ① オミックス解析

##### A) ゲノム DNA 解析

次世代シーケンサー(NGS)の高速化により個人の全ゲノム配列を容易に決定できるようになった（whole genome re-sequencing)。またエクソン領域だけをシーケンスするエクソーム解析の登場により大量のヒトゲノム配列が生み出されるにようになった。2008 年から 2012 年に行われた国際研究協力プロジェクト 1000 Genoms Project (http://www.1000genomes.org/) ではヒト約 2,500 人のゲノムシーケンスを行い配列データベースが公開されている。英国では、Genomics England (http://www.genomicsengland.co.uk/) という 100,000 人ゲノムプロジェクトを進めている。米国の公的資金やベンチャー、サウジアラビアでも同様のプロジェクトが進行中である。国内では東北メディカルメガバンク (http://www.megabank.tohoku.ac.jp/)、ナショナルセンターバイオバンク事業（NCGM, http://www.ncbiobank.org/) などにより日本人のヒトゲノム配列が収集されている。NCGM には理研の貢献度も高い (理研 IMS) 。これらの一次データの一部は NCBI SRA (http://www.ncbi.nlm.nih.gov/sra)、DDBJ DRA（http://trace.ddbj.nig.ac.jp/dra)、EBI ENA (http://www.ebi.ac.uk/ena) にストレージされているが、表現型のデータとの関連は DDBJ Japanese Genotype-phenotype Archive (JGA, https://trace.ddbj.nig.ac.jp/jga/index.html) に保存されている。このデータを利用するには利用目的などが審査され、個人情報保護と研究の進展を両立させようとしている。

さらにがんのゲノム決定も盛んに行われており、ICGC (International Cancer Genome Consortium, https://icgc.org/)が中心となり、11 万症例以上のがんゲノムシーケンスを実施しデータベース化している。ICGC へは理研 (旧 CGM、現 IMS) からの貢献度も高い。

ヒト以外の非モデル生物のシーケンスも容易になったため、ゲノムプロジェクトが盛んである。5 千種の昆虫ゲノムを決定する 5,000 Insect Genome Project（i5k, https://genome10k.soe.ucsc.edu/) などが進行中である。

ゲノム解析のデータベースは一次データ (FASTQ ファイル) の保存が重要になる。近年、標準ゲノム配列に対して多型のある部分だけをグラフ構造で保存する基本ファイルフォーマットが提案されるなど、データ圧縮・簡潔データ構造技術とそのファイルからの検索が今後の重要な分野となるだろう。また大量のゲノムデータやその他のオミックス情報を同時に可視化する方法も求められている。さらにそれらの情報を統合し検索したり統計モデル化する方法も重要となるだろう。

今後は大量のヒトゲノムが決定されることにより、遺伝子間や疾患との関連、エピゲノムとの関連、カルテ情報との統合などのデータ本位な解析アプローチが重要な課題となるだろう。

B) メタゲノム解析

NGS を用いて環境中に生息する生物群衆由来のゲノム混合物を解析し、群集の構成を計測する手法である。この方法により、土壌、腸内環境など、主に微生物生態系とその動態を、難培養菌も含めて正確に計測することが可能となり、環境保全、土壌改善、健康増進等、様々な分野への貢献が期待されている。理研においても、バイオマス生産性向上 (理研 CSRS)、腸内免疫 (理研 IMS) 等、様々な観点でメタゲノム研究が行われている。

他の NGS 解析と同様、メタゲノム解析でもさらなる大規模データの解析が発展の鍵であり、データ保存量や解析スループットの拡大が求められている。また、生物種の特定や分類を行なうための相同検索の高速化が大きな課題の一つである。さらに、得られた群集組成データから、環境の機能を推定するデータ解析手法の確立が求められる。そのためには、遺伝子機能、パスウェイ、さらには各生物種の特徴等、入手可能なすべての関連情報を統合的かつ定量的に解析し、新たな知識発見につなげる技術が必要である。理研においても、多種のゲノム配列混合物から高効率に生物種を特定する解析ツールの開発 (理研 QBiC) が行われている。また、サンプリングした環境のメタ情報がデータベース化されていることが重要であり、多様な情報を統合可能なセマンティックウエブ技術への期待も高い。国内の生命科学データ統合を目指すバイオサイエンスデータベースセンター (NBDC) 統合化推進プログラムでは、プロジェクトの一つとして、同技術を用いた微生物メタゲノム情報の統合が行われている (東工大／NBDC 統合化推進)。このプロジェクトでは環境、宿主、生物種を始め様々なメタ情報の統合が行われている。

メタゲノムと同様な手法で環境中の生物の機能的動態を捉える「メタトランスクリプトーム」、土や海水などを直接シーケンスする環境 DNA (eDNA) 等の新たな手法も開発されており、環境と生物の相互作用をさらに詳細かつ動的に捉えることが可能になると期待される。そのため、さらなるデータ量への対応、解析効率、メタ情報(データベース)の整理が必要になると考えられる。

C) トランスクリプトーム解析

RNA の発現情報を網羅的に定量するトランスクリプトーム解析はマイクロアレイによって実現化されたが、NGS を用いた近年の技術開発によって現在では定量性・網羅性が飛躍的に向上し、さらに RNA の構造も同時に決定できるようになった。広く利用されている RNA-seq 法の場合にはスプライシング部位の決定を行える他、理研で開発された CAGE 法 (理研 CLST) では転写の開始位置を決定し定量することができる。さらに、解析対象をより微細なものや 1 細胞を対象とする技術開発も進んでおり、理研でも 1 細胞 RNA シーケンス法が開発されている (RIKEN CLST、ACCC など)。平成 27 年度ライフサイエンス研究概算要求の大きなテーマとしても採択され、今後はデータの増加が考えらえる。特に、細胞バーコード技術 (理研 IMS) と混合反応、マイクロ流路技術などにより現在では、数百から数万細胞からの 1 細胞 RNA-Seq の開発が

進められている。これにより組織に含まれる未知の細胞サブタイプが発見されてくるであろう。

RNA のシーケンス技術は多様化している。microRNA のシーケンス、核や細胞質局在している RNA をシーケンスする方法や転写中 (nascent-seq)、翻訳中の RNA (ribosomal profiling) だけを読む方法、RNA とタンパク質結合を観測する手法などが登場しており、RNA の代謝・動態を計測できる。これらの情報を統合して理解するために、データベースの整備が重要となるだろう。今後は現象に関わる RNA 量を時空間で網羅的に計測する動きもあり、データの増加が見込まれている。

新規に取得された RNA プロファイルを解釈するためには、既知の遺伝子発現量データベースとの比較が重要である。現在、公共の遺伝子発現レポジトリとしては、NCBI GEO (http://www.ncbi.nlm.nih.gov/geo/)、EMBL-EBI Array Express (https://www.ebi.ac.uk/arrayexpress/) があるが、その多くがマイクロアレイのデータであり、今後は RNA-Seq や CAGE のデータを取りこんだデータベースが必要となるだろう。また発現データベースのサンプル情報などが正しく十分に整理されていなければ、発現量の類似検索などはできない。今後はセマンティックウエブやテキストマイニング、エキスパートによるヒューマンキュレーションなどによる発現データベースのアノテーション整備が必須となるだろう。データジャーナル (Scientific Data など)が主導するデータ標準化の動きもみられる。

RNA のシーケンスにはリファレンスとなる RNA 配列、構造のデータベースが必要となる。現在、Ensembl (http://ensembl.org/)、RefSeq (http://www.ncbi.nlm.nih.gov/refseq/) などのリファレンストランスクリプトームが整備されてきたが標準化の動きがあり、GENCODE (http://gencodegenes.org/) が中心的な役割を果しつつある。

理研からは FANTOM プロジェクト (http://fantom.gsc.riken.jp/jp/) によって多数の細胞における プロモータレベルでの発現情報が公開されているが、日本 (DDBJ)には NCBI GEO や EMBL-EBI Array Expression に対応する公的な発現データベースが存在せず、日本やアジア周辺での発現データベースに関するイニシアティブは相当低い。この点も解決されなければならない点である。

D) エピゲノム解析

世代を越えて表現型や疾患が遺伝するがゲノムに変化がないエピジェネティクスのメカニズムの理解は、今後の生命科学における極めて重要な課題のひとつである。ヒトゲノム計画終了後、日本はポストゲノムに舵を切ったが、欧米ではゲノムの非遺伝子コード領域などに着目した ENCODE 計画 (https://www.encodeproject.org/)、Epigenome Roadmap (http://www.roadmapepigenomics.org/) などを進め、現在のエピゲノム研究の基盤的な役割を果たすデータベースを構築した。日本ではこのような計画が遅れており、海外のデータベースを利用するだけになっている。ただヒトのエピゲノム情報を集める国際ヒトエピゲノムコンソーシアム (http://ihec-epigenomes.net/) では、日本チーム (http://crest-ihec.jp/) が実験技術を提供するなど活躍している。今後、エピゲノムデータベースのさらなる充実が期待される。

課題は、エピゲノム (DNA やヒストンの化学修飾) と遺伝子発現量の関連を網羅的に結びつけ、疾患や環境応答を理解するためのデータの統合技術であり、時間軸に

沿ったエピゲノム変化と発現量の関連性解析も今後盛んになってくるだろう。さらにエピゲノム変化がどのように世代を越え、表現型に影響を与えるのかを調べることは生命科学の最重要課題のひとつであろう。理研ではこの点に着目し、眞貝エピジェネティクス（独創的研究提案制度）、疾患エピゲノムプロジェクト(平成 27 年度ライフサイエンス概算要求)などが進行中である。

データベースにあるエピゲノム情報、特に DNA とタンパク質結合、ヒストン修飾データ（ChIP-Seq）は抗体や試薬キットによる再現性に問題があり、データの再利用にはそれらのメタデータを整理しデータベース化することが重要である。

ヒストン修飾や DNA メチル化はクロマチンの高次構造に影響を与えて、その結果、遺伝子発現量を変化させる。クロマチン高次構造は HiC によりシーケンスでき理研 CDB が実施している。今後はヒストンやクロマチンのライブイメージング技術（RIKEN QBiC）など画像データとの統合も重要になってくるであろう。

E） メタボローム解析

メタボローム解析は、NMR や各種質量分析法を用いて、生命現象に関係する低分子代謝物を網羅的に解析する 研究手法である。生理現象の解明などの基礎研究に加え、バイオマーカーの発見など医療分野における応用例も多く、日本は高い測定技術を背景に世界のメタボローム情報の約 6 割を保有している。さらに、メタボローム解析では、ピークやスペクトルを標準品との比較によって化合物同定を行なうため、実験から得られた定量値を伴う網羅的な代謝物検出データ、および各化合物知識を、大規模に収集、連結、共有化することが極めて重要である。

理研では、脂質メタボロームの計測技術を開発しており(理研 IMS)、今後のデータ増加が見込まれる。また、メタボロームデータを集約した統合データベースを構築（理研 CSRS／NBDC 統合化推進）しており、化合物、スペクトル、活性の情報とゲノム情報の統合化を進めている。また京都大学では、代謝パスウェイのデータベース、KEGG（http://www.genome.jp/kegg/）を提供している。

F） フェノーム解析

「フェノーム」とは、オミックス階層の最上位として、表現型＝Phenotype の総体を示す言葉である。フェノームは、単なるデータの集合としてだけではなく、表現型同士の関連性、すなわち、疾患と個別病態とのつながりなどの個体レベルでの関連性、いわゆる「モデル生物」概念の根拠となる生物種を超えた進化的なメカニズムの保存性、また、生態系における個体間相互作用など、様々な関連性を検出、可視化することにその意義がある。生命メカニズム全貌の理解のためには必須とも言える解析であるが、障壁も多い。画像解析や、計測技術の標準化を通し、様々な試みが行われている。

F-1. 網羅的表現型解析

従来、特に高等生物では、表現型解析は仮説に従った個別研究として行われてきた。しかしながら、このアプローチでは個別論文になりにくくネガティブデータが未発表として埋もれてしまうこと、データの客観性や相互利用性に乏しいことが、生命メカニズム全体の理解のための障壁になりつつある。

網羅的表現型解析は標準化された解析手法を用いて、多項目の検査を、遺伝

子改変系統等に対して網羅的に行う手法であり、大規模解析を通じて客観的な生命機能データを得ることができる。近年では、微生物や極めて小さな動植物だけではなく、マウスやシロイヌナズナ、ヒトにおけるコホート解析をはじめとして高等生物でも行われている。

一般に、小型の生物では、後述の画像解析をはじめとする先端的計測技術を用いて、極めて客観性の高いデータ、あるいは従来計測できなかったデータの取得が行われ、未知の生命機能発見に貢献している。これに対して、いわゆる高等生物では、大規模化が難しく、相互利用性に課題を残しながらも、応用性の高いデータの取得が行われる。マウスでは、12 カ国が参加する国際連携プロジェクトとして、全遺伝子の機能解析が行われている (http://www.mousephenotype.org: 理研 BRC 参画)。ヒトコホート解析ではまさに医療データの取得が行われ、医科学への直接の貢献が期待されており、英国 Genomics England、米国 Precision Medicine Initiative 等もコホートデータが重要な位置を占めている。

これらのデータを生物、および分野横断的に連結させることも大きな潮流となっている。疾患と表現型との関連性を統合的に扱う国際ポータルサイト Monarch initiative (http://monarchinitiative.org/) では、医療に貢献するデータとして、マウス、ラットはもちろん、線虫やショウジョウバエのデータも、モデル生物、あるいは直接実験解析可能なデータとしてヒトと結びつけられている。国内でも、単なる疾患としてではなく、疾患に含まれるそれぞれの表現型の情報を医療研究に活用する必要性が高まっており、日本医療研究開発機構 (AMED) の掲げる医療の ICT 化における臨床情報の共有化においても、表現型をキーに様々な研究リソースを関連づけ、研究成果を最大化することが必要だとされている (http://events.biosciencedbc.jp/images/togo2015/p28-29.pdf )。理研では、モデル生物リソースの表現型統合データベース (理研 BRC／NBDC 統合化推進: http://jphenome.info)、植物の表現型データベース (理研 CSRS: http://metadb.riken.jp/metadb/db/SciNetS_ria61i) が作成されており、これらのデータのさらなる利活用の推進が求められている。

F-2. バイオイメージインフォマティクス

トランスクリプトーム、プロテオーム、メタボロームなどのいわゆるオーミックスの計測技術は、1990 年代以降のゲノム科学の急速な発展と並行して、自動化、高速化、定量化が急速に進んでいたが、表現型データの取得は実験者の目視による実測や認識に依存しており、客観性や定量性、スループットが低く、データ本位生命科学を実現する上での障害になっていた。しかし、今世紀初頭から、計算機制御を用いた顕微鏡画像の自動撮影と、画像認識を用いた顕微鏡画像中の生物試料の特徴の自動認識を組み合わせた計算表現型解析の技術が発達し、表現型データの量と質に変革をもたらしつつある。計算表現型解析は大きくハイコンテントスクリーニング (High contents screening) と生命動態の定量解析、アトラス構築の 3 つに分類できる。

ハイコンテントスクリーニングはマイクロプレート上で培養した細胞に自動撮影、自動画像認識を適用し、自動スクリーニングする手法で、薬剤スクリーニング等で広く活用されている。その発展系として、マイクロプレート上に組織や線虫などの

個体を培養する場合もある。ハイコンテントスクリーニングの装置開発は早い段階から企業が積極的に参入し、欧米諸国では多くの大学、研究所に普及しており、生命科学研究の共通基盤的な方法になっている。一方、我が国の大学、研究所での普及は遅く、欧米諸国に大きく遅れをとっている。

生命動態の定量解析では細胞や胚、組織の動態を多次元自動撮影と自動画像認識により定量計測し、これを定量的に解析する。細胞や胚、組織の動的な現象の解析に活用される。この分野では、標的とする生命現象に最適化された顕微鏡撮影装置と、それによって撮影された画像に最適化された画像認識手法の開発が必要となる。欧米諸国では 2000 年代の中盤頃から Janelia Farm Research Campus (米国)、Max Planck Institute of Molecular Cell Biology and Genetic s(独国)など、顕微鏡開発と画像処理技術開発を融合した研究組織を設立し、この分野をリードしている。これらの研究所では、「ショウジョウバエ胚の全細胞核の全胚発生過程における追跡」のようなフラッグシッププロジェクトを設け、先進的な技術開発を進めている。日本ではこのような研究機関は未だ設立されておらず、欧米に比べて大きく遅れを取っている。(理研 QBiC はこのようなアクティビティを持つ国内唯一の研究機関であるが、バイオイメージインフォマティクスによる生命動態の定量解析が QBiC の設立目的では無いため、そのアクティビティの規模は小さい)。生命動態の定量解析は、近年では、細胞生物学、発生生物学の個別研究においても必要不可欠な解析になっており、個別研究で使用可能な汎用的な画像解析のツール開発の要求が高まっている。現在、広く使われているものとしては、ImageJ(Fiji)を筆頭に、CellProfiler の他、Janelia Farm Research Campus や Max Planck Institute of Molecular Cell Biology and Genetics などが多くのソフトウェアツールを公開している。欧米に比べ、日本ではこのような汎用的なソフトウェア開発も大きく遅れをとっており、戦略性を持った早急な対策が必要である。理研 RAP では生物画像用の汎用的な画像解析ソフトウェアの開発が行われている。

アトラス構築では、蛍光顕微鏡や電子顕微鏡等で 3 次元の試料を断層撮影し、それらを用いて、試料の 3 次元構造や遺伝子の 3 次元的な発現状況を計算機で再構築したデータベースを構築する。マウスやヒトの脳の3次元アトラスの構築を目的に 2000 年代前半に設立された Allen Institute for Brain Science(米国)が有名である。特に、電子顕微鏡を用いた3次元断層画像から脳神経系のネットワークを再構築する「コネクトーム」の画像認識プログラムの開発は、当分野の最も重要な課題であるが、Allen Institute for Brain Science および前述の Janelia Farm Research Campus が技術開発をリードしている。当分野における日本のアクティビティおよび技術水準は低いが、2014 年から、理研が主導しマーモセット脳のコネクトームや周辺技術開発を開始した (理研 BSI、CSLT、RAP)。組織の透明化に関しては理研が世界を先導している (RIKEN BSI、CDB、QBiC)。

ここにあげた3つの分野に加えて、これらの研究の医療応用として、病理画像から病態を推定するデジタル病理学の分野が注目を集めてきている。デジタル病理学研究は、診断制度の向上および治療の病院間格差、地方格差の解消のために、今後ますます重要になっていく分野であると思われる。

現在、バイオイメージインフォマティクス技術を使って計測されたデータの多くは、プロジェクト毎に構築されたデータベースから公開されていることが一般的である。

また、公開の際のデータフォーマットもプロジェクト毎の独自フォーマットを使っていることが一般的である。理研では、生命動態の定量計測データの利活用を促進するために、生命動態の定量計測データ用の統一ファイルフォーマット BDML を開発し、世界各国の生命動態計測データを収集する SSBD データベース（理研 QBiC／JST-NBDC 統合化推進プログラム）を構築した。理研 QBiC では生命動態の計測データを全世界規模で統一的に管理する国際コンソーシアムの確立を目指して活動している。

共通資源である画像データを、研究コミュニティにおいてどのように共有していくかは、バイオイメージインフォマティクス分野の大きな課題である。特に、限られた研究機関にのみ整備されている最先端の顕微鏡装置を使用して取得した画像の共有化は、画像取得の能力が研究の優劣に直結する細胞生物学、発生生物学の研究者にとって重要な問題である。その意味で、画像データベースの充実化、標準化が求められる。画像データベースとしては、論文誌 Journal of Cell Biology による画像のデータベース JCB Data Viewer や、American Society of Cell Biology（ASCB)による The Cell: an image library が構築され、論文等で使用された高画質の画像の共有化が進められている。上記の米国の取り組みに比べ、我が国の画像データベースに関する取り組みは大きく遅れをとっている。本年度から、上記の SSBD データベースが画像データベースとしての機能を拡張し、細胞生物学および発生生物学分野の最先端画像の収集を開始している（理研 QBiC/JST-NBDC 統合化推進プログラム)。画像共有のプラットフォームとしては、国際連携プロジェクト Open Microscopy Environment（OME）により開発されたオープンソース画像データベースソフトウェア OMERO が（顕微鏡画像以外でも）広く用いられつつある。しかし、幅広いコミュニティからの画像の利用には、画像へのメタデータの付加と標準化が必要であるが、現状では不十分な状態である。理研 QBiC では、OME と連携して、画像メタデータ標準化に取り組んでいる。

また、画像データの高精細化、多次元化(空間軸、時間軸)により、データ量は増大の一途をたどっており、バイオイメージインフォマティクスを推進するには、画像フォーマットの大規模データへの対応(バイナリーフォーマット（HD5 等）への対応とデータ転送インフラの整備が大きな課題である。

② ネットワーク解析

生体内あるいは細胞内の分子の相互関係により構成されるネットワーク全構造の解明は生命の統合的理解に迫る核心である。一方、これまでに明らかにされたネットワークはそのごく一部ではあるが、これらの知見は新たなデータを読み解く上で強力な基盤になり得る。したがって、ネットワーク解析はそれ自体が生命科学における主要なテーマであると同時に、その結果を有効活用する枠組みは重要な研究基盤と位置付けることができる。

分子ネットワークデータは大きく、科学論文として発表された知見を専門家が吟味し相互関係を抽出したものと、上記オミックス解析に代表される網羅的手法により取得されたデータから相互関係を抽出したものに分類できる。前者としては日本が世界に先駆けて構築した KEGGpathway が代表的なものであったが、その利用ライセンスの変化などもあり現在では海外のデータベース（Reactome、Panther、WikiPathway、BioModels 等）

の利用が広がりつつある。また、民間会社により販売されているデータベースの利用も増加しており、内容だけでなくその利用形態が多様化していることも伺える。これまでの研究が及んでいない、あるいは専門家によるデータ収集が間に合わない領域に関しては後者のネットワーク、つまり網羅的データより推定された制御・類似関係によってアプローチすることが可能である。実際に近年はそのようなデータベース（String、CellNet、CMap 等）の構築と利用が進んでいる。

また、データベースの構築と共に、ネットワークモデルを記述する方式が、様々な階層を対象に提案されている。北野らが提案した分子の反応プロセスを記述する SBML（The Systems Biology Markup Language）はその先駆けであり、様々な方式同士の交流を促すために近年は COMBINE（http://co.mbine.org/）コミュニティが活発に活動している。

③ 計算構造生物学

タンパク質、RNA、糖鎖等からなる生体分子複合体の機能の理解、さらにはその理解を創薬等に応用する為に、分子構造・運動を、構造解析、分子シミュレーションにより解明していく研究分野である。大阪大学・日本蛋白質構造データバンク（PDBj: Protein Data Bank Japan）では、米国 RCSB、BMRB、および欧州 PDBe と協力して、生体高分子の立体構造データベースを国際的に統一化された PDB アーカイブとして公開している。理研では、NMR, Spring8 等の大型放射光施設等を用いたタンパク 3000 プロジェクトにて、タンパク質の基本構造の解明および機能の解析で大きく貢献してきた。最近では in-cell NMR や X 線自由電子レーザー（XFEL）を構造解析に使う研究も進んでいる。さらに、様々な計算機的手法と実験データを組み合わせ、京などの高性能計算機を利用して実験データから生体分子構造のモデル構築する手法の開発が行われている（理研 AICS）。

④ 人工知能的アプローチ

生命科学で扱うデータは大規模化、複雑化し、コンピュータによる知識抽出は必須ともいえる状況であり、人工知能による高度な解析の実現は極めて期待の高い課題である。2011 年にクイズ王に勝利した IBM の質問応答システム Watson（非構造データ分析、自然言語処理、検索の高速化等、既存技術の組み合わせで構成されている）は、医療診断支援、潜在的な薬物間相互作用の検査等における意思決定支援システムとしての応用が期待されており、実際に東大医科研との共同研究が開始されている（http://www-06.ibm.com/jp/press/2015/07/3001.html）。また最近では、ディープラーニング(深層機械学習)と呼ばれる多層化した無教師繰り返し学習により未知の特徴量を発見する手法への期待が高まっている。2012 年、Google は画像への多階層ニューラルネットワーク学習を用いて、人間が特徴点を指示することなしにコンピュータに「猫」を学習させる（猫に反応するニューロンを作成する）ことに成功した。このように、技術的背景は多様であるが、現代のコンピュータの卓越した計算力と記憶能力を用いて、膨大なデータの背後にある規則を見つけ出す点においては、実用に向けた基礎研究が確立しつつあるといえるだろう。特に生命科学において、これらの技術を駆使することで、いわゆるビッグデータからの知識発見が加速されると期待されており、今後に向けて極めて重要性の高い研究分野の一つである。

その意味で、近い未来の生命科学のデータベース基盤は、人工知能からの利用が比較的容易であることが重要である。ただし、人工知能技術は、データマイニング、テキストマイニング、ディープラーニング等の多様な技術の総称であり、処理可能なビッグデータについては明確な制限や標準は個々の技術に依存し、一概に述べることはできない。例えば、オントロジーでは人間の持つ知識を積極的に機械可読な形式でモデル化するが、ディープラーニングではモデル化以前に特徴点の指示などの「教師」は必要ないとされる。このようなアプローチは人間の積極的介入が少ないことで、人間では見つけにくい、「より新しい」知識が得られる可能性がある。

一方で、人間の知的活動に基づいたモデル化を行い、それに基づいてデータの均一化や標準化を行うことは、人工知能の機械処理に対して有利に働くことが多い。例えば、前述のディープラーニングの成功例でさえ、画像自体が高度に標準化された大規模データであり、これを膨大な計算リソースを用いて処理することが必要であった。また、知識のモデル化によって（バイアスは避けられないが）より人間のニーズに即した処理を行えるメリットがある。さらに、以下⑤で述べるメタデータとの対応で議論すると、メタデータ記述の世界標準であるセマンティックウエブの枠組みでは、より正確かつ体系的なメタデータ記述を行いながら、データ流通や統合を通して、人工知能からも比較的容易に利用可能な巨大な知識ベースが構築されると期待されている。より体系的で豊富な知識として記述されたメタデータは、人工知能の処理結果の妥当性の検討に使われ、高度な知識処理の実現に繋がると考えられる。また、標準化されたメタデータは、プログラムからの自動的なデータ利用を補助する役割もあり、docker 等のコンテナ技術と組み合わせることで、プログラムが自動的にデータを収集・解析するパイプラインが広く実現されることが期待される。

理研では、人工知能技術を用いた生命情報解析基盤が議論されはじめており、データ分類、関連解析、相関解析等の統計的マイニングのデータ基盤として、Garuda プラットフォームの利用も提案されている（理研 IMS: http://www.garuda-alliance.org ）。また、網羅的解析手法と進化実験、数理モデル解析を組み合わせた自動進化実験システム（理研 QBiC: http://www.qbic.riken.jp/mbd/index.html）の開発も行われている。しかし、人工知能とメタデータとの連携については、未だ明確なビジョンを持っている研究者は、理研内外を通じて研究論文数が少ないこと等を勘案するとまだ多くはないとみられる。

⑤ 生命科学おけるメタデータの活用

生命科学において、メタデータを記述する語彙の標準化については以前からオントロジー構築としてその努力がなされてきた。例えば、遺伝子の機能記述の共通語彙の策定を目指した Gene Ontology（GO）コンソーシアム（http://geneontology.org/）は 1998 年に、様々なバイオ医療分野の語彙策定を目指した The Open Biological and Biomedical Ontologies（OBO）コンソーシアム（http://www.obofoundry.org/）は 2001 年に、さらにバイオ系オントロジーの研究開発を目的とした The National Center for Biomedical Ontology（NCBO: http://www.bioontology.org/）が 2005 年に立ち上がり、それぞれオントロジー開発を推進している。後述のセマンティックウエブ技術の一つで、Web 上に存在するオントロジーデータのやり取りを行うための言語 Web Ontology Language（OWL）がよく知られるようになってからオントロジーは急速に発展を遂げ、現在では OBO/OWL をはじめ多様な分野を記述できるオントロジーが提案、公開されて

いる。このオントロジーの発展に呼応する形でライフ系データのセマンティックウエブ化が進んできた。

世界では 2006 年には既存の著名なデータベースをセマンティックウエブ形式に変換し公開するサイト Bio2RDF (http://bio2rdf.org/) の論文発表が行われ、2013 年には英国 EBI でも UniProt を含む 6 つのデータベースがセマンティックウエブに対応し公開された (https://www.ebi.ac.uk/rdf/platform)。ライフ系データベースに関するセマンティックウエブ研究コミュニティは医療や創薬分野を含めて発展してきており、セマンティックウエブのトップカンファレンス International Semantic Web Conference (ISWC: http://iswc2015.semanticweb.org/) でもライフサイエンスは主要なトピックの一つであり、また Semantic Web applications and tools for life sciences (SWAT4LS: http://www.swat4ls.org/) 等のライフサイエンスのアプリケーション実装に特化した国際会議も行われている。これらの国際会議では、近年欧米を中心に医療系の電子カルテや病名を統合するオントロジーの開発研究成果が多く発表されている。

日本では、2007 年に発足したライフサイエンス統合データベースセンター(DBCLS) において 2010 年ころよりコンテンツの有機的な活用のためにセマンティックウエブの活用が始まり、後の 2011 年に発足したバイオサイエンスデータベースセンター (NBDC) でもセマンティックウエブでのデータベース公開を支援するプログラムが発展的に進められている。

理研はこれらの活動に先んじて 2005 年ころからセマンティックウエブ技術を基礎としたデータベース構築研究が進められ、2009 年には理研が公開しているデータベースの統合化のために当該技術が採用された (理研プレスリリース: 理研のデータベース構築基盤の公開基準をセマンティックウエブに統一, http://www.riken.jp/pr/press/2009/20090331_2/)。2015 年からは、データベース作業部会が提言したデータ流通を第一義としたセマンティックウエブ準拠の理研メタデータベース (http://metadb.riken.jp) が、4 省庁統合データベースカタログ IntegBio (http://integbio.jp/ja/) や W3C のライフ医療系データベースのメタデータ HCLS Community Profile (http://www.w3.org/TR/hcls-dataset/) に準拠した形で公開されている。

Web を発明したティム・バーナーズ＝リーが 2005 年 Bit-IT World (http://www.bio-itworld.com/) の基調講演で、次世代の Web 技術であるセマンティックウエブについてはライフサイエンス分野がリーダシップを発揮するだろうと述べた。その予測通り、グローバルなデータ流通という場においてライフサイエンスの一つの情報インフラが形成され、発展してきている。

## ii. データベース技術

### ① 従来のデータベース: 関係データベース

データベースは、決められた形式に従ってデータを集めて管理する情報システムである。データベースとして集められたデータの検索や抽出を高速かつ正確に行えるよう、情報学ではデータ構造や管理方法を抽象化した理論であるデータモデルや、実用的な検索アルゴリズムの研究がすすめられてきた。現在最も普及しているデータベースは関係データベースと呼ばれるもので、集合論と述語論理に基づいて考案された関係モデルがその基礎を与えている。

② 生命科学系データの多様化、大規模化に伴うデータベース技術の変遷

関係モデルのデータは表として記述することができる。さらに表同士は関係と呼ばれるデータの組を利用することで、行や列を構成するデータを介して互いに連結することができる。表形式はバイオの実験研究者でもよく使われる形式であり、MySQL (https://www-jp.mysql.com/) や PostgreSQL (https://www.postgresql.jp/) など無料で利用できる実用的な実装が普及していることも相俟って、生命科学では関係データベースによるデータベース構築が行われてきた。しかし、ライフ系データの多様化に伴い、複数のデータベースからデータを収集し統合解析を行う需要が高まってきた。関係データベースは、一つのデータベース内で閉じた系を構成している。このため、複数のデータベースとの連携は、表データ形式や関係の相違、更にはデータ形式同一であっても必ずしもデータの意味が同じとならないことなどの要因が重なり、その実現には大きな困難を伴う。XML 形式への対応や myGrid (http://www.mygrid.org.uk/) などの WebService API の活用などの努力もなされてきたが、実装上のコストが大きいことなどの理由から、データ統合の大きな潮流にはならなかった。

データの大規模化に伴う汎用的なデータ処理の枠組みとして「ビッグデータ処理」が挙げられる。ビッグデータとは、これまでの関係データベースを含む従来の技術では処理しきれない膨大なデータのことである。蓄積され巨大化したデータに対し、短時間で処理を行えるような、またトレンド解析のようにリアルタイムにデータ処理結果を得るような新たな仕組みが求められた。NoSQL と呼ばれるデータベース技術はこの要求を満たすものとして活用が期待されている。関係データベースではデータ処理の排他制御や一貫性保証などのデータを正しく処理できる機能が実装されているが、NoSQL ではこのようなデータ一貫性の保証行わず、データ構造の単純化や分散並列処理の単純化によりデータ処理の高速化を優先させている。ライフサイエンスのデータ処理では、巨大ではあるが同種のデータの集合を扱うことが多く、すなわち問題 (処理データ) を単純に分割し互いに独立したプロセスによる並列処理を行えばよい場合が多い。具体的なデータベースシステムとして Hadoop プロジェクトの HBase が著名である。これは Google の大規模なデータを分散ストレージとして実現するデータベースシステムに触発され、オープンソースとして開発されたものである。生命科学用途での Hadoop の活用としては、ゲノムデータ解析に向いていることが知られてきており、例えば Contrail と呼ぶ大規模ゲノムアセンブリ処理系や、次世代シーケンサーデータの解析処理をクラウド上で実行するシステムの試作等が挙げられる。

データの大規模化に対応する技術については、上記のような汎用ビッグデータ処理基盤にとどまらず、生命科学に特化した技術提案も行われてきた。特に NGS によるデータの大規模化に対応するために、ゲノミクスに特化したクエリを高速に実現するための技術開発が行われた。具体的には、データの種類に応じた格納方法とインデックス方法により規定される
BAM (https://samtools.github.io/hts-specs/SAMv1.pdf)、
VCF (http://samtools.github.io/hts-specs/VCFv4.2.pdf)、
bigWig/bigBed (Kent et al. PMID:20639541) 等のファイルフォーマットが提案され、これをサポートするツール群が広汎に利用されている。

生物画像データに関しては、画像の管理・表示・解析を担うソフトウェアプラットフォームとして OMERO (https://www.openmicroscopy.org/site) が開発され、JCB Dataviewer (http://jcb-dataviewer.rupress.org/) などに採用されており、デファクトになりつつある。生物画像データは容量が非常に大きいため、研究機関内および機関間のやりとりには、それ相応のネットワーク帯域幅が必要になる。

③ グローバルな規模でデータ流通を促進するメタデータ技術

以上述べたように、データの種類に特化したデータ形式の標準化が進んでいるものの、データを統合的に解析するためのデータ流通技術については従来の枠組みでは難しい状況であった。データの統合化、流通促進に資する方策として、データベース統合をデータベースの実装のレイヤで実現するのでなく、データのレイヤで実現することが望まれるようになった。この方策の一つがメタデータ技術であり、特にグローバルなネットワークに対応する形でメタデータを記述、流通させる技術であるセマンティックウエブ技術 (http://www.w3.org/standards/semanticweb/) の活用に注目が集まってきた。

セマンティックウエブは、ウエブの発明者である World Wide Web Consortium (W3C, http://www.w3.org/) のティム・バーナーズ＝リーによって提唱された体系で、現在隆盛のグローバルなドキュメントのネットワーク (ウエブ) の発展形である。ウエブ上のドキュメントは HTML で記述される。HTML では文書の構造やドキュメント間のリンクを記述することは可能であるが、書かれているドキュメントの内容やその意味を記述するものではない。それに対し、セマンティックウエブはドキュメントやデータの意味を扱うことを、データ形式やインターフェイス、ツールを標準化することで実現することを目的としている。セマンティックウエブの核はデータをウエブ上で公開しデータ間の関係を意味が付けられたリンクを張ることでデータのネットワークを実現させる仕組みであり、ドキュメントを含むウエブ上のデータの共有や再利用が、人のみならず情報処理システムにおいても促進されることが期待される。

### iii. データ可視化

膨大なデータからデータの背後にある意味や規則を抽出する際、データ に含まれる現象や構造を人間が直感的に認識するための手段として、可視化 (visualization) は極めて重要な役割を担っている。また、多くの科学的分野のビッグデータ解析において、統計分析は知的発見のための重要な技術となっているが、データ解析におけるアンスコムの例のように、同じ統計量を持つ異なるデータの存在が示されているため、データを人間に認識させる可視化は、ビッグデータの正確な解析のために必須である。技術分野としては可視化には、インタラクティブなデータの 3 次元表示、粗視化、潜在変数発見などが含まれ、ビッグデータのインタラクティブな処理に対応するための計算処理の加速化も含まれる。一般的には、可視化や解析を行うためのツールとしては Gnuplot 、MATLAB、GrADS ParaView 等が広く用いられており、3 次元データの科学可視化に特化したアプリケーション ソフトウェア、特定のデータに特化したソフトウェアも数多く開発されている。しかし、大規模データからの知識獲得の加速を目的とした可視化ツールの開発は未だ始まったばかりである。本分野は海外では IEEE VIS などの主要な国際学会が存在するなど研究が盛んになっているが、国内での認知度は未だ低い。理研では京コンピュータを使ったシミュレーション結果 (AICS) や生命動態の解析データ (QBiC) を対象に可視化技術の開発が進められている。

## 3. 長期的な視野にたった理研データベース基盤に対する提言

### i. 日本の生命科学が直面する情報関連の課題／問題のまとめ

上記より、日本の生命科学のデータ基盤が直面する課題は、下記のようにまとめられる。

#### A. ビッグデータ活用のための情報インフラの整備

a. データ量増大に伴う大容量ストレージ整備、計算処理、データ転送処理の高速化

NGS データ、画像をはじめとして、ほぼ全ての分野でのデータ量増加が見込まれている。これらの実データを滞りなく保管、活用するためのインフラが必要である。

b. データのオープン化によるイノベーション促進

各分野とも分野外のデータあるいは知識の利用が必須となっている。また、理研のデータをオープン化することでのイノベーション促進も求められている。理研にどのようなデータ存在するのかを広く周知させることも必要である。

c. データやメタデータの統合・連結の記述方式の標準化

多岐にわたるデータや知識の相互利用のために、それらに辿り着くためのメタデータを体系的に相互連結し、かつ広く標準的に用いられるようにすることが必要である。

#### B. ビックデータからの知識抽出を可能とする技術

d. 人工知能的アプローチ

生命科学データは今後ますます増大、複雑化していき、データ解析には人工知能的アプローチが必須となっていくと考えられる。理研はそのための技術開発を積極的に行うべきである。データ活用基盤においても、人工知能から利用しやすくやるようなメタデータ整備、標準化が必要である。

e. 大規模データ解析に適した新たな統計解析方法論

従来の統計解析手法は、大規模、多項目データに適用すると false-negative が多くなりすぎる等の問題を抱えており、ビッグデータへの適用が難しいと言われている。世界的には、このような方法論を汎用化し、多くの解析で使えるようにする必要がある。国内、および理研のビッグデータ解析研究では、世界をリードする意味でも、効率的にデータの意味を抽出する方法論をいち早く採用していくべきである。

f. データ可視化、粗視化

人工知能的アプローチおよび統計解析的アプローチの原理的な問題点を解決する技術としてビッグデータ解析における可視化技術の重要性は今後、急速に高まっていくと予想される。可視化ソフトウェアのコレクション化やそれらが扱うデータフォーマットの統一化など、効率的に可視化ソフトワエアの開発と共有化が行われるような環境整備を、理研は積極的に進めるべきである。

g. 画像からの情報抽出、知識抽出

画像からの情報抽出は、定量的で再現性の高いデータをもたらし、センサーデータとともに、今後の生命科学ビッグデータの一翼を担うと考えられ、多くの研究分野で期待の高い技術である。生命科学分野では、フラッグシップ的なプロジェクトに特化した先進的なアルゴリズム開発と、個別研究に適用可能な汎用性の高いツール等の開発の両面が求められている。我が国の現状はこの両面において欧米に大きく遅れをとっている。生命科学分野の画像は多様であり、個別研究毎に適切な画像解析の方法を選別・開発する必要があるが、広いニーズに対してソリューションを提供する分野としての戦略が世界的にも定まっていない。そのため、我が国および理研がこの分野で世界をリードためには、広いニーズに対する戦略の選定が重要である。

C. データ活用によるイノベーション

h. オミックス情報の統合的理解による疾患、創薬、健康増進への直接的貢献

生命の統合的に理解によって、疾患、創薬、健康増進の技術が今後さらに発展すると考えられており、特に理研はこれらの技術への直接貢献が強く求められている。

本レポートが対象としているデータベース基盤は、上記のような状況を踏まえ、理研の各ライフサイエンスの研究部署に対し、より直接的な貢献および情報発信の推進補助できるよう構築整備することが望まれる。

## ii. データベース基盤への提言

理研の生命科学の研究活動は、理研が持つサイエンスの総合力や大規模施設を活かし、世界でもまれな多岐に渡る広範で大規模なものとなっている。いわば理研は世界のライフサイエンス活動の縮図とも言えるだろう。従って、上に述べた生命科学のデータ基盤の課題を、理研が世界に先立って解決していくことは、世界レベルでの研究リーダシップを執り続けていく上で極めて重要な施策と言える。

データ基盤は、おおまかに、先進的な研究を推進するデータ解析基盤と、データ共有、利活用を推進するデータベース基盤とに分けて考えることができる。本レポートが対象とするデータベース基盤の役割は、理研の成果を対外的にアピールをすることであり、また各研究、解析基盤との密接かつ効率的な連携を可能とすることが求められる。データベース基盤はさらに、① 各研究において生産される 実データの運用基盤(各研究の直接的サポート)と、② 理研の生産するデータの統合解析や利活用を最大化する、メタデータ運用基盤とに分類される。

① 実データ運用のサポート基盤について

実データ運用基盤の課題は、上記 a. データ量増大に伴う大容量ストレージ整備、計算処理、データ転送処理の高速化が相当する。現在運用中、あるいは運用開始された基盤としては、2015 年 4 月に開始された情報基盤センターの理研クラウドの実験サービスでは、情報基盤センターが管理運用している計算機を使用して、実用的なバイオインフォマティクスツールを備えた仮想計算機や多目的に使える仮想計算機をネットワークを介して提供している。また、2015 年 4 月より和光でサービスを開始した新スーパーコンピュータシステムである HOKUSAI GreatWave では、実験/シミュレーション/データ解析の融合に向けて、

機能の異なる 3 つのシステムを高速ネットワークで接続し、それらをあたかも 1 つの計算システムであるかのように利用できる環境を構築している。1 つ目は大容量データを格納するオンラインストレージシステムと階層型ストレージシステム、2 つ目は超並列計算システム、3 つ目は GPU や大容量メモリを搭載したアプリケーション演算システムである。システムの中核となるオンラインストレージシステムは、総実効容量 2.1 ペタバイト、総理論帯域は 190 ギガバイト/秒の性能を有しており、階層型ストレージ管理システムは 7.9 ペタバイトを装備している。超並列計算システムは 1 ペタフロップスの理論演算性能をもつ 1,080 ノード、34,560 コアのシステムであり、より演算の高速化を図るため、アプリケーション演算システムには、Tesla K20X を 4 枚/ノード搭載したサーバ 30 ノードと、1 テラバイトの大容量メモリを搭載したサーバを 2 ノード装備している
(http://www.riken.jp/pr/press/2015/20150403_1/)。
階層型ストレージシステムは、HOKUSAI-GW システムのユーザーのみならず、理研内全ての研究室を対象としたバックアップ・システム (HOKUSAI データ預かりサービス http://accc.riken.jp/data-depo/ 旧 D2S システム) にも利用しており、大容量な研究データなどを高速かつ安全にデータ転送し、バックアップする仕組みを導入している。さらに先進的な HPC システムとして世界トップの省電力性能を達成した SHOBU を導入し、いち早く最新の環境を用意することで、データ処理をさらに高速化するシステムの実現が期待できる。

このようなサービスを理研内研究者のニーズに応じて適切に提供することが、実データの扱いへの直接的なサポートとして重要である。特にデータ産出量がハードディスクの容量増加を上回っている現実がすでにあり、それに対する継続的な措置が必要である。また、特に画像等の大規模データのバックアップが、理研支所間でのネットワークを通すと上手くいかない等の問題もすでに起こっている。さらには、大規模災害発生等で拠点が甚大な被害を受けたとき、拠点に置かれた全データが失われることは、理研にとって大きな損失である。このような事態を防ぐため、データバックアップやネットワークを含めた総合的な対処が必要である。

② メタデータベース基盤について

上記の実データ基盤に加え、作業部会では、メタデータ基盤の重要性が増していくと考えている。勘案すべきは、すでに述べたとおり、データの統合解析がライフサイエンスの極めて重要な要となりつつあることである。しかしながら、データが大規模でヘテロであることがデータ解析の障害となっている。世界の縮図としての理研がこの課題を解決し、理研が産出するデータの流通促進とその解析プログラムへと橋渡しできるインターフェイスを備えた情報基盤を実現させることは、ライフサイエンス分野に大きな貢献となるであろう。理研では、既にセマンティックウエブ技術基盤のデータベース研究を行ってきた経緯があり、殊に昨今、ビッグデータ解析、人工知能、Graph500 で世界一を奪還した京コンピュータなど、理研の情報戦略とも連携させ、知識をグラフ構造で記述、流通させるメタデータの利点を活かした高度なデータ解析を発展させ、今後 5 年、10 年に渡って有機的な研究成果を生み出すことが求められる。その基盤として、世界のデータと有機的に繋がり、かつ人工知能が参照する知識の蓄積となるライフサイエンスデータのメタデータ化とそのデータベース構築が必要になる。

上記 i の「日本の生命科学が直面する情報関連の課題／問題のまとめ」で述べた課題や問題に関して、メタデータベースは多面的な解決を提供する。

b. データのオープン化によるイノベーション促進に関しては、データオープン化自体は各研究室で判断することが必要だが、イノベーションにつながるオープンデータの活用に関しては、適切かつ標準化されたメタデータの付与が不可欠であり、理研のデータに対するメタデータのワンストップショップとしてのメタデータベースを適切に運用することで、この課題に対して効率的に貢献することができる。

c. データやメタデータの統合・連結の記述方式の標準化に関しては、メタデータベースが直接的に貢献するところである。ただし、適切な標準化が継続して行われるよう、運用上の工夫が必要である。

d. 人工知能的アプローチに関しては、現代のメタデータの標準が機械可読な知識フォーマット（すなわちオントロジー）を用いて行われており、セマンティック Web が人工知能からの知識源として利用可能な機械可読性を実現する世界標準の人工知能技術の一つと位置付けられることからも分かる通り、理研の保有するデータを人工知能から利用可能とすることに貢献すると考えられる。人工知能技術は発達途上であり、先端研究としての盛衰がポイントとなるが、世界標準の技術を採用することで、比較的安定かつ、先端研究と一般社会貢献とのインターフェイスとなるメリットもある。

e. 大規模データ解析に適した新たな統計解析方法論、f. データ可視化、粗視化、g. 画像からの情報抽出、知識抽出に関しては、先端研究課題であるため、データベース基盤としては、直接的な貢献ではなく、むしろその成果に対して、どのようなメタデータを付加して、利活用度を高めるかということに貢献すると考えられる。

h. オミックス情報の統合的理解による疾患、創薬、健康増進への直接的貢献に関しては、各研究の課題ではあるが、理研の保持するデータを統合的にわかりやすく、かつ人工知能からも利用可能な形で提示するという意味では、そのためのデータ源を提供するという点で貢献すると考えられる。

また、セマンティックウエブをはじめとするメタデータ化技術が実用域に達した一方で、DNA 配列そのもの、高速処理のために高度にコード化されたデータなど、特に現在のセマンティックウエブ技術では、データベース化にそぐわないと考えられるデータも多く存在し、今後も技術発展によって生まれ続けると考えられる。しかしながら、データを統合的に扱う上では、（例えばデータセット単位でメタデータ付与する等）実データに何らかの形でメタデータを付与することは必須と考えられる。つまり、その時点で利用できるメタデータ化技術と実データとの最適な接点を常に見つけていくことが必要であり、そのこと自体も世界のデータ統合を推進するために極めて重要な技術であると考えられる。

現在、情報基盤センターが実験運用中の理研メタデータベースは、セマンティック Web の基盤である Resource Description Framework（RDF）のバックエンドデータベースとして、現時点で実用的な機能と処理速度を提供している。さらにデータ表示のフロントエンドとして、実験研究者に馴染みのあるシンプルな表形式を用いてわかりやすくデータを提示するなど、一般の生命科学研究者でも容易にメタデータと結びついたデータ公開を行うことができ、すでに世界的に先進的なメタデータベースとなっている。このことはデータベース作業部会の中間報告の進言を情報基盤センターが忠実に実装したことに他ならない。

一方で、理研で生産される緻密なデータに十分なメタデータを付加し、かつ、世界トップレベルの知識抽出基盤として運用しようとすると、現時点で流通する全ゲノムデータ、遺伝

子機能アノテーション等、内外からさらに膨大なデータを導入することになるため、容量不足や速度低下する状態に数年以内に到達することが懸念される。セマンティックウエブ技術では、インターネット上にデータを分散させ、それらに対してあたかも 1 つのデータソースに対して行うように検索する連合検索 (Federated query) が可能であるが、主にパフォーマンスの問題から、未だ広く活用される状況には至っていない。このパフォーマンスをさらに向上させて、真のデータ分散を実現し、無駄なデータ管理の負荷を軽減することも必要である。

このように、メタデータ記述 (データ表現) の方法論や技術はこれからも向上していくことが期待されており、技術発展の下で理研メタデータベースが陳腐化することなく、また収集されたデータ管理がおろそかにならないよう、継続運用のための技術開発や運用努力が必要になるであろう。

## 4. 結論

以上の議論から、作業部会は、人工知能等の情報技術の発展を見据え、5 年 10 年に渡る長期的視野に立って、データベース基盤について以下の各項を提言する。

### 1. 実データ基盤について

1.1. 大容量データインフラの整備として、情報基盤センターが、大容量、高速なストレージ、計算処理、データ転送処理等のインフラサービスを、理研内研究者のニーズに応じて、適切に提供すること。特に実データの容量の爆発的増加に対するサポートとして、産出されたデータすべてを保管することは不可能であるとしても、保管すべき部分については十分なストレージを担保すること。さらに、災害による貴重なデータの損失を避けるべく、地理的に離れた拠点間でのデータバックアップができるよう、ストレージやネットワーク等のインフラの強化を求める。また、実データとメタデータベースをより緊密に連携できるような総合的な対応を望む。

### 2. メタデータ基盤について

2.1. 今後ますます重要となるメタデータ運用を見据え、ライフサイエンスデータのメタデータ技術を基調として、理研が保有するデータに関するメタ情報のデータベース（メタデータベース）を情報基盤センターが運用することを提言する。具体的には、今年度実験運用が開始された理研メタデータベースの継続運用が適当であると判断する。また、上記実データとも関連して、今後理研データおよびメタデータのビッグデータ化に備え、大容量化、高速化を随時行っていくことを望む。

2.2. 2.1 に述べたメタデータベースについて、陳腐化を防ぐために、情報基盤センターは世界最高水準の情報技術を活用して主導をとって技術開発を進めるべきである。例えば、システム面では、上記の高効率連合検索（Federated query）の実現による、世界レベルでのメタデータの分散化、結果としてもたらされる理研でのメタデータ管理コストの最小化、データ面では、理研のデータを記述するオントロジーの整備や、既存のデータを世界標準のメタデータに変換する機能など、技術的研究課題解決を推進することを求める。

2.3. 2.1 に述べたメタデータベースに格納するメタデータ記述語彙に関して、今後、バイオインフォマティクスを含めた研究動向、及びデータ解析、人工知能研究との連携が行えるような充実化を図る体制を理研が整備すべきである。

喫緊には、以下の各項への対応を求める。

2.3.1. 情報学、生物学の研究者が一体となり全理研でデータベース活動を推進できるような仕組みの一つとしてデータベース基盤についての検討を行う組織を整備すること。これには、バイオインフォマティクス検討委員会で新たな部会を組織するか、あるいはデータベース作業部会が（分科会の組織も含め）継続的に検討するかの判断が必要であろう。

2.3.2. データベースの実現については、データベースの整備担当部署である情報基盤センターが、バイオインフォマティクスやウエブ技術の進展動向について随時調査研究を行いながら適切な技術を適用してデータの流通が最大となるような継続的な研究開発の努力を主体的に行っていくこと

2.3.3. メタデータベース基盤の利活用性向上のために、国内外のデータベースの専門家等の助言や評価を受ける体制を構築すること。例えば、定期的な評価が行えるよう、理研内外の研究者から構成される評価委員会等を組織すること

さらに将来的には、

2.3.4. オントロジー等の体系化されたメタデータに基づいて、多様な内容を含む大規模グラフデータからの知識発見（自動抽出、人工知能、数理モデルの活用、ディープラーニングなど）や、理研が自らの持つ研究データ、および外部データを最大活用して、世界における情報利活用を先導していくため、メタデータの開発と、理研内外の研究者との連携を図ること

## 3. その他、データベース基盤の対象ではない課題について

上記 3. i. に述べた、国内生命科学の直面する課題のうち、

d. 人工知能的アプローチ
e. 大規模データ解析に適した新たな統計解析方法論
f. データ可視化、粗視化
g. 画像からの情報抽出、知識抽出

に関しては、将来的にデータベース基盤として実現が望まれる課題であるが、それよりも先端研究として今般取り組むべき課題である。これらに関しては、別途作業部会を組むなどして検討することを提言する。

以上

## バイオインフォマティクス検討委員会 データベース作業部会

| | | |
|---|---|---|
| 部会長 | 桝屋 啓志 | バイオリソースセンター |
| 部会員 | 大浪 修一 | 生命システム研究センター |
| | 川路 英哉 | 予防医療・診断技術開発プログラム |
| | 二階堂 愛 | 情報基盤センター |
| | 野田 茂穂 | 情報基盤センター |
| 部会員 兼 事務局 | 小林 紀郎 | 情報基盤センター |