成果報告① **10月5日13:15—13:35**

生命科学におけるデータサイエンスの課題と PDB

中村 春木　（PDBj 代表、大阪大学蛋白質研究所　所長）

生命科学におけるデータサイエンスのゴールは、膨大な生命ビッグデータの解析から新規の知識体系を構築し、最終的に科学的智恵（自然界の原則・法則の発見・理解を含む問題解決能力）に結実させることと考えている。そのための必須な課題は、当然のことではあるが「データが無ければ問題解決はできない」こと、より正しくは「正確なデータが無ければ正しい問題解決はできない」という問題の解決である。データは正しいのか？最新なのか？複数のデータ間で矛盾はないのか？さらにはデータがない場合はどうするのか？という当たり前だが深刻な問題に行き当たる。利用者が個々のデータの信頼性を理解でき、最新のものが常に提供され、精度が常に検証される、という仕組みによって、初めて有用な知識体系が構築され、それを基に正しい問題解決が図られる [1, 2]。

本統合化推進プログラムによって支援を受けている日本蛋白質立体構造データバンク（PDBj: Protein Data Bank Japan, http://pdbj.org/）では、蛋白質、核酸、糖鎖などの生体高分子の立体構造情報を国際協力によって集積し、2016 年 6 月には 12 万件を超える構造データを公開している。PDBj は、国際的組織である wwPDB(worldwide PDB: http://wwpdb.org/)の一員として、共通した品質管理によるデータ登録作業とダウンロード・サイトの運営 [3-5]および独自の

サービス・ツールや二次データベースの提供、セマンティック化を行っている。

特に上記したデータサイエンスを推進するため、データ・アノテーションと検証法の継続的な改良、データ検証レポートの公開・配布、構造情報を裏付ける実験データの同時登録や登録者を特定するための ORCID（http://orcid.org/）を採用し、データの品質を保証している [5)]。一方、生体分子立体構造情報と他の異なる階層の情報を結ぶデータベースの統合的運用を図るため RDF 化を進め、wwPDB/RDF [6)] と NMR 化学シフトのデータベースである BMRB (http://www.bmrb.wisc.edu/)に対する BMRB/RDF [7)]を世界に先駆けて開発して国際標準とし、JST-NBDC の RDF ポータル（http://integbio.jp/rdf/）に統合した。さらに配列データとの強い連携のため、EBI の PDBe および UniProt と協力し SIFTS データの RDF 化も行った。

一方、クライオ電子顕微鏡による生体超分子複合体の構造解析の急進展に対応するため、巨大構造も高速に表示できる molmil ビューア [8)]や、PDB、EMDB、SASBDB の 3 つのデータベースに対して横断的に分子形状の同時比較・検索ができる Omokage 検索 [9)]の開発・公開など、科学の進展に合わせたデータベースおよびサービスの進化も続けている。

参考文献

1) 中村春木, 生物物理 53, 44-46, 2013.
2) 中村春木, データシェアリングシンポジウム, 2016 年 2 月 29 日, 東京 一橋講堂
3) Berman H, Henrick K, Nakamura H, Nature Structural Biology 10, 980, 2003.
4) Berman H, Henrick K, Nakamura H, Markley JL, Nucl. Acids Res. 35, D301-D303, 2007.
5) Berman H, Burley SK, Kleywegt GJ, Markley JL, Nakamura H, Velankar S, Curr. Opin. Struct. Biol. 40, 17–22, 2016.
6) Kinjo AR et al., Nucl. Acids Res. 40, D453-D460, 2012.
7) Yokochi M et al., J. Biomed. Semantics 7:16, 2016.
8) Bekker GJ, Nakamura H, Kinjo AR, J. Chemoinformatics 8:42, 2016.
9) Suzuki H, Kawabata T, Nakamura H, Bioinformatics 32, 619-620, 2016.

成果報告② **10月5日 13:35—13:55**

ゲノム・疾患・医薬品のネットワーク解析

金久　實　（京都大学化学研究所　特任教授）

本研究開発では、ゲノム情報を医療・産業等で有効利用するためのトランスレーショナルバイオインフォマティクスリソースとして KEGG MEDICUS の開発・提供を行ってきた。KEGG MEDICUS を構成する KEGG DISEASE データベースでは疾患と遺伝子・ゲノムとの関連、KEGG DRUG データベースでは標的分子や薬物代謝の情報などを、主に文献情報を基に蓄積し、また日本の医薬品添付文書に記載された医薬品相互作用に関する知識を体系化した。さらに現在は日本と米国の医薬品添付文書に記載された適応症をもとに、KEGG DISEASE エントリの追加・修正、及び医薬品と疾患の精密な対応づけを行っている。適応症には例えば、ALK 融合遺伝子陽性の非小細胞肺癌、EGFR 遺伝子変異陽性の非小細胞肺癌、CCR5 指向性 HIV-1 感染症、といった具合にゲノム関連情報で限定された病名が多く見られるようになってきた。個別化医療や希少疾病用医薬品指定制度の反映であり、今後は KEGG MEDICUS のようにゲノム・疾患・医薬品を統合的に理解できるリソースの重要性はますます高まっていくと考えている。

KEGG では分子間相互作用ネットワークを考慮する際に、KO (KEGG ORTHOLOGY)と呼ぶ機能オーソログで一般化した場合、ヒトの遺伝子やタンパク質に限定した場合、さらにはヒト遺伝子の変異等も区別した場合など、異なるレベルでの取り扱いができるようになっている。医薬品をこれらのネットワークに含めて解析するために、本研究では個々の医薬品を異なるレベルで一般化できるように DG (DRUG GROUP)と呼ぶグループ化を行った。これを用いて標的分子が関与するネットワークと薬物代謝に関するネットワークを同時に解析することで、薬剤耐性、薬物応答、薬物相互作用を理解できるようにしている。ここでは微生物ゲノムから抗微生物薬耐性(Antimicrobial resistance)を推定する試みと、複数医薬品を併用する際の安全性（禁忌・注意）を調べるために開発した医薬品相互作用チェックツールを紹介する。後者は現状では添付文書から抽出した医薬品や医薬品グループの相互作用情報のみに基づいているが、今後は例えば薬物代謝酵素遺伝子の多型情報のように、相互作用に影響を与えるゲノム関連情報も考慮できるようにしていく予定である。

がんのゲノム医療実装に向けた現状と課題：統合データベースへの期待

中釜　斉（国立がん研究センター　理事長）

近年、飛躍的に解析能力が向上したゲノムシークエンス技術を用いたがんゲノム解読の成果として、がん細胞では数千～数万にも及ぶゲノム変化が生じており、アミノ酸置換を伴う変異も数十～数百に及ぶことが分かってきた。このことは、がんはゲノム変化の観点からみても極めて多様な病態であることを意味しており、個々の症例に最適な医療（Precision Medicine）を提供するには、ゲノム変異と多様な臨床情報とを繋げるデータの集積が求められる。即ち、治療法選択に関わるがん（体細胞）のゲノムの構造・機能異常の情報や、薬剤反応性・副作用等の体質に関わる生殖細胞系列の遺伝情報と、経時的かつ詳細な臨床情報を統合したデータベース（DB）の構築が喫緊の課題となってくる。加えて、遺伝性乳がん・卵巣がん症候群（HBOC）に代表される家族性腫瘍も症例全体の数%～10 数%に存在することが想定されているものの、その遺伝的素因に関しては未だ十分に解明されておらず、日本人症例のゲノム情報、及び遺伝子型－表現型相関情報の集積も極めて不十分である。これらのがんの体細胞及び生殖細胞系列のゲノム異常及び多様性の生物学的及び臨床的な意義・有用性に関する日本人のデータを集積することが日本におけるゲノム医療の実装には不可欠であるが、現状では欧米の取り組みから大きく遅れをとっている。

国立がん研究センターでは、クリニカルシークエンスの体制を整え、個々の患者のがん病変における体細胞変異を二週間以内で診療現場に戻すなど、ゲノム情報に基づいた最適医療の実現に向けて取り組んでいる。Precision Medicine の実現には、臨床情報とゲノム・遺伝子解析情報が統合された DB の構築、疾患横断的なオールジャパン体制でのゲノム DB の整備が喫緊の課題である。今年度からスタートする「臨床ゲノム情報統合データベース整備事業」（AMED）では、①ゲノム解析 DB（raw データベース、RDB）、②各疾患領域グループ毎のゲノム医療研究開発のためのデータストレージ（DS）と疾患横断的な連携体制、③ゲノム医療リファレンス知識ベース（統合 DB）の、異なるタイプの DB 構築とこれらの DB 間の有機的な連携に基づく、産官学の広い範囲の研究者等とのデータシェアリングが求められている。ゲノム医療の実装における上記 3 つの特性を有する DB から構成される臨床ゲノム統合データベース構築の重要性とその意義・発展性等について概説する。

KNApSAcKデータベースを用いた昆虫・植物間化学的相互作用解析

小寺　正明　（東京工業大学生命理工学院　講師）

植物は植食性昆虫による食害を避けるため様々な防御物質を合成し、昆虫はそれに対抗する様々な仕組みを開発し多様化してきた。生態学は分子生物学よりも長い歴史を持ち、文献や図鑑などによる知識蓄積が豊富である。これらの知識の効果的な統合によりその利用価値を高めることが期待される。

本研究では、昆虫・植物間相互作用に関する種間関係・遺伝子・化合物の三者を記述したデータベースの構築を行なった。最初に、日本に棲息する鱗翅目昆虫の食草を様々な図鑑から網羅的に収集した。次に、植物化合物を中心とした大規模データベースである KNApSAcK[1)]とのリンク付けを行なった。また、新型シークエンサーにより得られた塩基配列データを蓄積した公共データレポジトリである DRA Sequence Read Archive[2)]を用いて、各昆虫種の RNA-seq データを取り出し、de novo アセンブルで予測されたタンパク質とのリンク付けも行なった。

以上のデータ統合により、昆虫・植物間相互作用に関する種間関係・遺伝子・化合物データからの知識抽出が可能となった。具体的には、相互作用ネットワークから鱗翅目昆虫の選択的食性と植物化合物との関係を解析した。続いて、植物化合物と昆虫化合物の構造比較を元に、食害時

に昆虫体内で起こる代謝の予測を行った。また草食性昆虫が植物を食害するのに必要だと考えられる種々のタンパク質の有無を推定し、食草選択との関係を解析した。

　生物種と遺伝子、生物種と化合物、生物種と別の生物種の関係を記述したデータベースはそれぞれ独立に存在するが、本研究では、遺伝子・化合物・種間関係の三者のデータ統合によって、昆虫・植物間相互作用における化合物および遺伝子の寄与について調べることを可能にした。昆虫はその多様性が研究対象としての魅力だが、その多様性ゆえに詳細な知識の蓄積がなかなか進んでいない。そのような分子生物学的知識の蓄積の乏しい昆虫に対するアプローチとしてデータ統合は非常に有効である。

参考文献

1) Afendi FM, Okada T, Yamazaki M, Hirai-Morita A, Nakamura Y, Nakamura K, Ikeda S, Takahashi H, Altaf-Ul-Amin M, Darusman LK, Saito K, Kanaya S. KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. Plant Cell Physiol. 2012 Feb;53(2):e1. doi: 10.1093/pcp/pcr165.

2) Nakazato T, Ohta T, Bono H. Experimental design-based functional mining and characterization of high-throughput sequencing data in the sequence read archive. PLoS One. 2013 Oct 22;8(10):e77910. doi: 10.1371/journal.pone.0077910.

ユーザー発表② **10月5日16:45―17:00**

日本人ゲノムデータベースの臨床ゲノム解析における利用

成戸 卓也　（徳島大学大学院医歯薬学研究部人類遺伝学分野　助教）

私たちは、遺伝的要因を背景に発症したと考えられる先天異常疾患を対象に、次世代シーケンサーを用いたゲノム配列や構造解析の探索を行っている。次世代シーケンサーは短期間に多くの先天異常疾患の原因遺伝子の解明に寄与し、臨床現場でも広く用いられようになっている。次世代シーケンサーを用いたエクソーム解析で得られるデータから目的の変異の絞り込みを行うのは困難であるため、各種データベースからマイナーアレル頻度（MAF）を参照している。国際データベース（1000 Genomes Project、NHLBI Grand Opportunity Exome Sequencing Project (ESP)）に加えて日本人特異的なデータベース（Integrative Japanese Genome Variation Database (iJGVD)、Human Genetic Variation Database (HGVD)）を用いることにより、日本人患者の変異候補をより絞り込めるようになってきた。

アミノ酸置換を伴う変異のタンパク質機能への影響の解釈には、タンパク質モデリング予測ソフトウェアや進化トレース法などの予測ソフトウェアツールがある。しかし予測であるので、基礎となるアミノ酸配列モチーフデータベースやタンパク質立体構造データベースの拡充が望まれる。今後、様々なデータベースが充実することにより原因変異かどうかの推定が容易になると思われる。

成果報告③ **10月5日17:00―17:20**

個別化医療に向けたヒトゲノムバリエーションデータベース

徳永 勝士　（東京大学大学院医学系研究科　教授）

SNP タイピング技術と次世代シーケンサ技術の技術革新により、ゲノムワイドな疾患関連多型・変異探索が可能となり、多くの疾患関連多型・変異が発見されつつある。我々のグループでは国内/アジアにおけるこれらのデータの散逸を防ぎ研究者間でデータ共有化するために、2007 年より GWAS-DB, CNV-DB を、2011 年度より Human variation DB, HLA-DB を構築し(https://gwas.biosciencedbc.jp/)、GWAS データ、NGS 変異データの提供を広く呼びかけると共にデータの預入れと再配布の運用を行ってきた。2014 年度からは NBDC がデータの預入れと再配布を実施し、その計算結果の可視化、及び、主要な DB の癌体細胞変異のドライバー変異やシグナルパスウェイと疾患関連変異の関係性の可視化を本 DB で引続き実施している。これらの DB においては GWAS のデータとともに、健常者や患者の多様な生殖細胞変異（SNV、長配列の挿入/欠失、構造多型）を収集対象とし、文献から抽出した疾患関連変異・臨床情報 30,000 以上のエントリーも登録しており、日本人/アジア人の変異と表現型（疾患、薬剤応答、ウィルス耐性）との関係の体系化を目指している。これらの DB を俯瞰することにより、異なる疾患に共有される疾患関連遺伝子や、疾患関連遺伝子ごとの疾患関連変異の分布の特徴等が明らかとなり、疾患の理解を深めるとともに、将来的には新規変異についての表現型への影響がある程度予測できる

ようになると期待される。本講演では、DB について概観するとともに、疾患関連変異の集団間差異や遺伝的多様性との関係等についても言及したい。

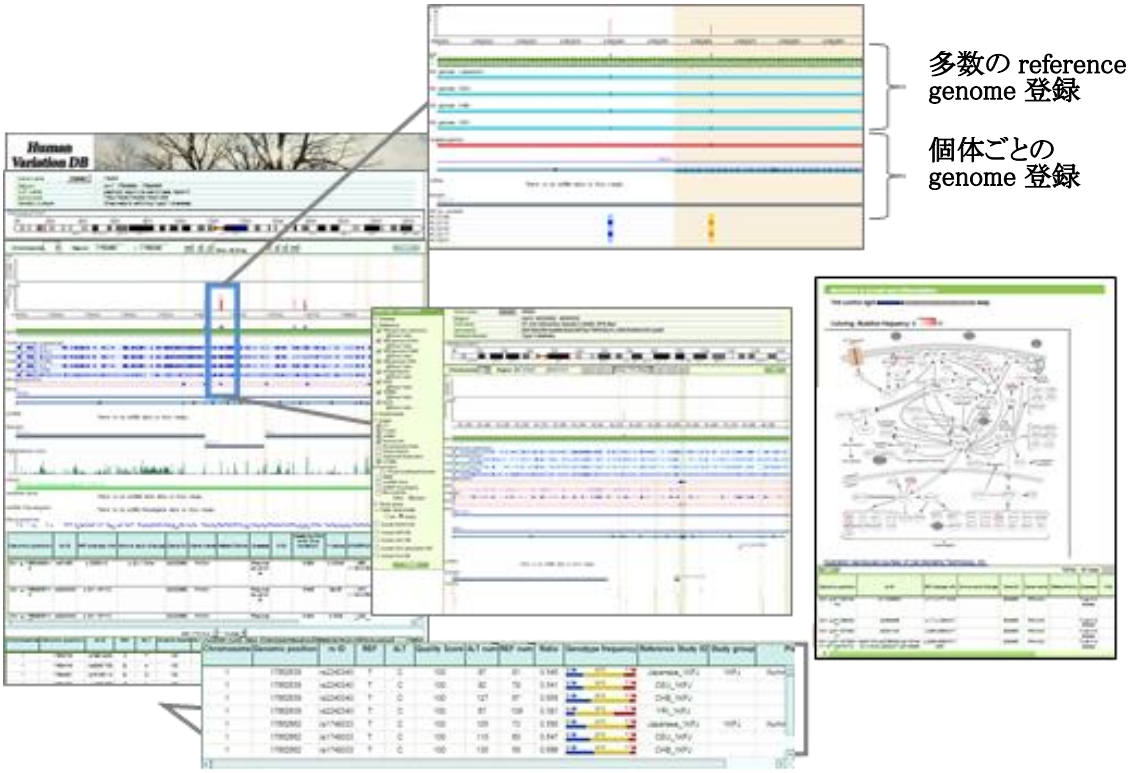


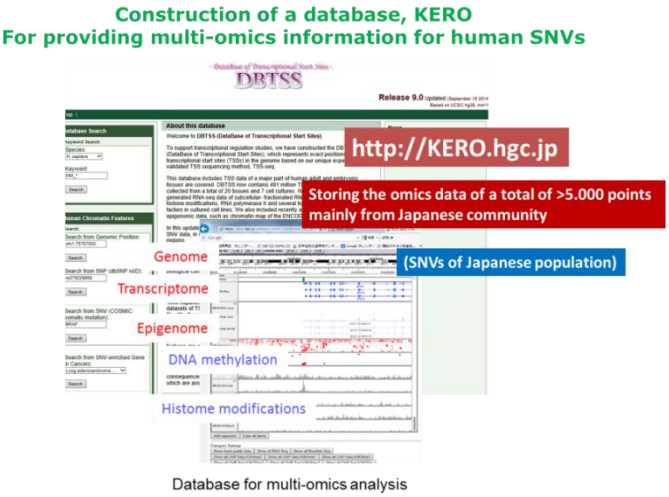
図　　Human Variation DB のスナップショット

成果報告④ **10月5日17:20—17:40**

疾患ヒトゲノム変異の生物学的機能注釈を目指した多階層オミクスデータの統合

菅野　純夫　（東京大学大学院新領域創成科学研究科　教授）

我々の研究グループでは、KERO-DB(Kashiwa Encyclopedia of Regulatory Omics Database)を開発し、ヒトのマルチオミクスデータの統合を試みている。本データベースは、特に、疾患に関連したヒトゲノム多型・変異に生物学的機能注釈を与えることを目的に、様々な目的で産出された、ゲノム変異位置、近傍におけるエピゲノム（ヒストン修飾、DNA メチル化パターン)、トランスクリプトーム情報（発現量、スプライスパターン）をヒトゲノム情報に統合している。データの統合は、現在、がんゲノム解析を志向したものを中心に行っているが、最終的には疾患の別を超えたデータの統合を目指す。現在は、ヒト培養細胞系あるいはマウスをはじめとする動物モデル系から得られたオミクスデータを中心にデータベースに収載している。これにより臨床検体で集積が乏しいエピゲノム情報を充実させ、同時に生物学的機能解析の実践の場としてのモデル系におけるオミクス情報を整備することを目指している。将来的にはヒト臨床検体から直接得られた多層オミクスデータを直接関連付ける予定である。



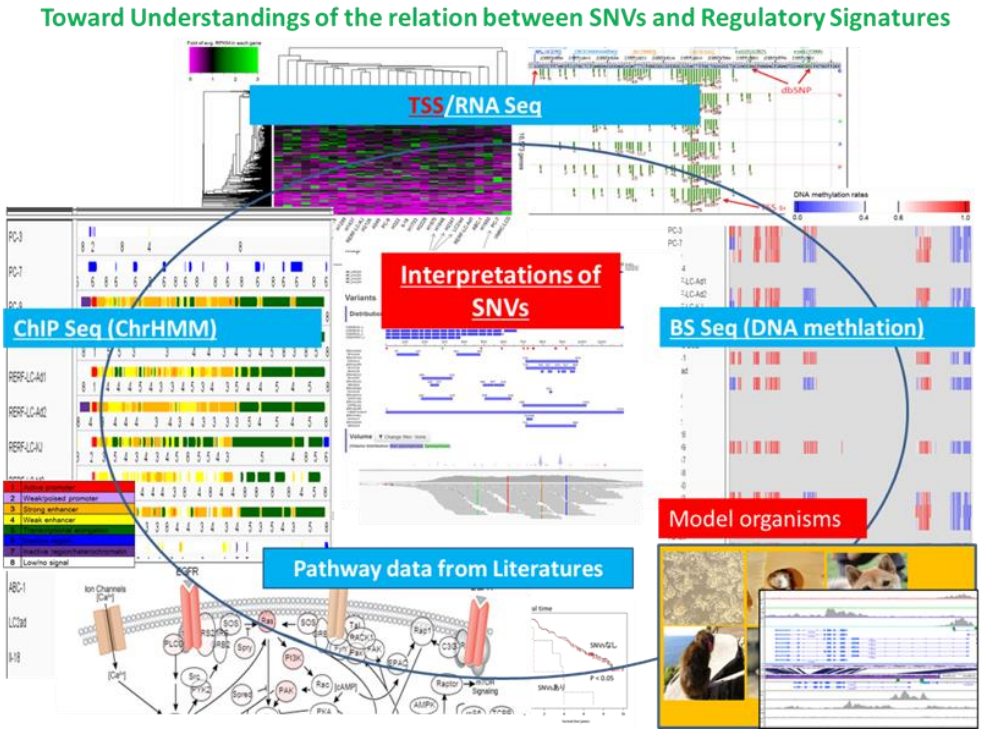
Database for multi-omics analysis

現在、公開されているデータの中でも、特に肺腺癌培養細胞の多層オミクスデータに特徴を持つ。同一のがん種に由来する培養細胞２６種について、全ゲノム配列、抑制性および活性性のヒストン修飾６種類、Methyl C Seq による DNA メチル化データ、転写開始点データ、トランスクリプトームデータ、miRNA データが体系的に記載されたものであり、各細胞のがんドライバーパターンについて、そのエピゲノム、トランスクリプトームの状態を横断的に検索することが可能なデータとなっている。他にも、マウス免疫担当細胞を種々の刺激源で刺激した詳細な経時計測系において、多階層オミクスデータを収集、データベースとして収載している（転写開始点デー

タ、ヒストン修飾等について計 400 計測点)。がん研究においても、がん免疫療法の急速な展開から、免疫担当細胞の多層オミクスの理解は大きな注目を集める分野となっている。本データベースでは、さらに多くのモデル生物種についての多層オミクスデータを文部科学省新学術領域「ゲノム支援」を中心としたゲノム解析支援活動を通じて収集している。現在、これらをモデルデータに、ヒト細胞系へのデータの写像を行ったデータベースの構築を行うべくデータを成形し、データベースとしての公開を逐次開始している段階である。現在までにマウスを中心に 10 系列のデータを公開しており、次年度にさらに拡張を行う予定である。

また、一方でゲノムデータについては、本データから参照する形で東北メディカルメガバンク、長浜コホート、日本ファーマコゲノミクスコンソーシアムにより産出された日本人多型データ約 5000 人分が無償、自由に閲覧可能となっている。

さらに、上記のモデルデータについての RDF 化によるデータ統合プラットフォームの整備を進めている。昨年度から着手した整備計画をさらに発展させる形で、格納されたデータについて、統合データベースの観点からの検索可能性を考慮した RDF 化によるデータフォーマット、検索の統一化を図っている。



参考文献

Suzuki A, Wakaguri H, Yamashita R, Kawano S, Tsuchihara K, Sugano S, Suzuki Y, Nakai K. DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. Nucleic Acids Res. 2015 (Database issue) 43 (D1): D1-D5.

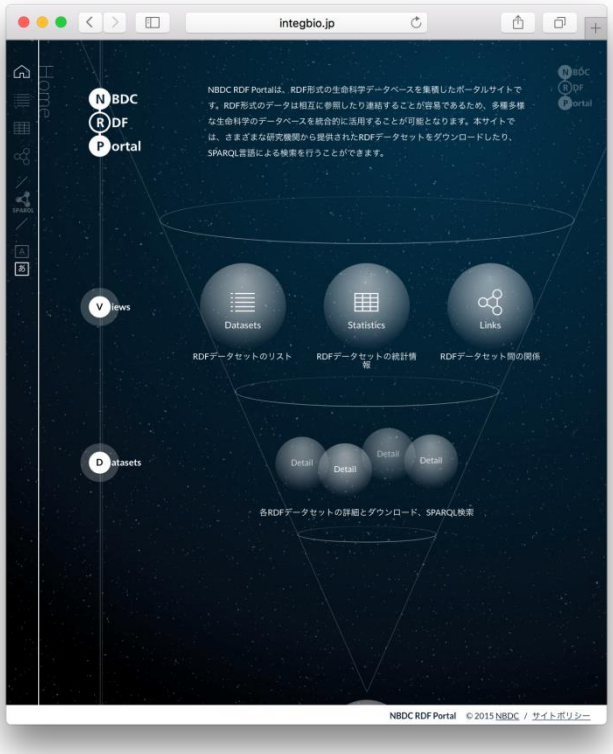
成果報告⑤ **10月6日10:00—10:20**

NBDC RDF ポータルが目指す、次世代データ統合利用基盤

川島　秀一
（情報・システム研究機構ライフサイエンス統合データベースセンター　特任助教）

Resource Data Framework (RDF) は、ウェブ技術を用いてデータを記述するための仕組みであり、World Wide Web Consortium (W3C) の標準規格として公開されている。RDF には、あらゆるものを Uniform Resource Identifier (URI)として記述することによる、記述対象のグローバルな一意性が担保されることや、オントロジーを利用することによる語彙の共通化が促進されること、オープンな規格であることなどの特徴があり、これらは、特に、多様で複雑な研究対象を扱う必要のある生命科学分野のデータを記述する上で大きな利点といえる。JST バイオサイエンスデータベースセンター（NBDC）と情報・システム研究機構 ライフサイエンス統合データベースセンター（DBCLS）は、特に、様々なデータを統合して利活用するという観点からこの利点を重要視し、データベースを開発している国内外の研究グループに対して RDF の採用を推奨してきた。また、データの利用者にとっては、国内で開発された RDF データを容易に検索・取得できるサービスが必要であると考え、2015 年 11 月より NBDC RDF ポータルと名付けた、RDF データの検索サービスの運営を開始した。本ポータル開発以前に、DBCLS では、毎月開催しているハッカソンイベント等を通じて蓄積してきた、質の高い RDF データを作成するための知見を、ガイドラインとしてまとめ公開してきた。本ポータルサイトに登録されている RDF データは、予

め本ガイドラインを踏まえてレビューされたものに限られており、全て一定以上の品質を備えている。本ポータルから取得できる RDF データを利用することで、異なるデータベース由来のデータを連携したアプリケーション開発や分野横断的なデータ解析等が、著しく促進されると期待される。さらに、コンピューターを利用したデータの解釈も一意に定まり、複雑な生命科学の問題に対して、近年大きな技術革新のみられる人工知能技術などを応用することも可能になると考えている。



～統合データベース、実りの時へ～ トーゴーの日

成果報告⑥ **10月6日10:20—10:40**

ゲノム・メタゲノム情報統合による微生物 DB の超高度化推進

黒川　顕　（情報・システム研究機構国立遺伝学研究所生命情報研究センター　教授）

ライフサイエンス研究において、知識の集積体であるデータベース（DB）は、知識の参照のみにとどまるものではなく、新たな研究分野を切り拓く上で欠く事のできない極めて重要な研究基盤である。ライフサイエンス研究の中でも、微生物研究は歴史も古く、これまで蓄積されたデータや知識は膨大かつ多岐にわたっている。さらに、昨今のゲノム科学の発展に伴い、ゲノムやメタゲノムなど圧倒的な量のデータが産出されており、これらを横断的にかつ簡便に利用できれば、新たな仮説や研究分野の創出がより容易になると期待される。

我々は NBDC 第 1 期統合化推進プログラムにて、国内外に散在する細菌の各種オミックス情報を広く収集し、遺伝子、ゲノム、環境の 3 つの軸に沿って遺伝子機能、分類学的情報、菌株保存情報、表現型情報などの知識を整理し、ゲノム情報を核としてセマンティックウェブ技術により統合した統合 DB「MicrobeDB.jp」を開発した。MicrobeDB.jp は、約 63 億トリプルで構成される Full RDF な DB であり、6 種類のオントロジー＆ボキャブラリおよび 180 種類の統合 DB 解析アプリケーション群（Stanza）を開発・実装した世界に類を見ない DB となっている。環境データを主軸の一つとして DB を統合した事により、微生物研究者のみならず、臨床医学、地球惑星科学など異分野の研究者からも利用されつつある。さらに、産業界においても、製薬、農業など

バイオ産業を筆頭に、住宅、化学、分析、商社などの多様な業種からも関心を持たれており、様々な発展的要望も寄せられるようになっている。

現在、MicrobeDB.jp を細菌のみならず真菌類、藻類を対象として拡張するとともに、持続可能なシステムの構築、Stanza の開発、さらにはユーザビリティの向上を徹底する事で、単語の検索や単なる統計量の羅列ではなく、大規模データからの新規知識発見を容易に実現する事が可能な DB システムの構築を目指している。MicrobeDB.jp を利用することで、分野特異的のみならず分野横断的な新興研究に資する新たな知見を抽出することが容易となり、ゲノム情報を核とした我国の微生物研究の発展に寄与できると考える。

成果報告⑦ **10月6日10:40—11:00**

PGDBj：幅広い用途に向けた植物ゲノム情報ポータルサイト

田畑　哲之　（かずさDNA研究所　所長）

ゲノム解析技術の普及によって、今や実験植物、実用植物を問わず多種多彩な植物種が研究対象とされ、そこで得られたゲノム関連情報は、ゲノム進化や有用遺伝子の同定や単離、分子育種等幅広い目的に利用されている。我々は、国内外に散在するさまざまな植物ゲノム関連情報を集約、整備したポータルサイトPlant Genome DataBase Japan (PGDBj, http://pgdbj.jp)を２０１３年８月から公開している。

２０１２年から２０１４年までは、基礎研究者を主な対象ユーザーとして、(1)オルソログデータベース、(2)リソースデータベース、(3) DNAマーカー・QTLデータベース、(4)横断検索システムを構築、提供した。この間、データベースに対するユーザーの要望が多様化したことを踏まえ、２０１５年以降は、それまでに構築したデータベースの拡張に加え、応用研究、特に育種研究分野に向けたデータの提供を強化した。例えば、上述のDNAマーカー・QTLデータベースでは主にマニュアルキュレーションによる情報の整備を進めてきたが、現在は同じデータセットに対して有用農業形質に着目したインターフェース「育種向けDNAマーカー検索メニュー」を構築し、データの活用頻度の向上に努めている。また、統合手法の中心となっているセマンティックウェブ技術を、(1) PGDBjコンテンツの情報と、(2)育種研究における検証例として作物の農業形

質情報のそれぞれに対し適用した。(1)では、PGDBj コンテンツが配列情報を介してゲノム上に集約できる点を利用し、RDF 構造に基づく高度な検索とゲノムブラウザによる結果の提示を可能にするシステムを構築中である。そして、(2)では、作物の系統関係や保有する農業形質の複雑な条件による情報検索ができるシステムを構築し、早期の公開を目指している。

ゲノム研究の進展による植物の基礎・応用研究の活性化が進む中、データベースに対するユーザーの要望も刻々と変化している。PGDBj は、このような用途の多様化にも柔軟に対応しつつ、ゲノム関連情報を有効に活用するための強固な研究基盤の整備に力を注いでいる。

成果報告⑧ **10月6日11:00―11:20**

糖鎖関連データベースの セマンティックウェブ対応と連携強化

成松　久　（産業技術総合研究所創薬基盤研究部門　総括研究主幹）

糖鎖は、遺伝子やタンパク質と同じように生体内に普遍的に存在し、 かつ構造の多様性に富む生体高分子である。糖鎖は酵素によって細胞内で生合成され、産生する細胞の分化やがん化あるいは炎症応答などの生理状態の変化にともなって構造が異なることが知られている。構造変化に伴う糖鎖分子の機能が生物学的な関心を集めるだけでなく、糖鎖構造の変化がバイオマーカーとして医療や研究ツールなどの応用分野での有用性も知られている。 このような糖鎖機能の解明を目指し、糖鎖分析技術が発展してきた。しかし現状でも質量分析による網羅的糖鎖構造解析技術では糖鎖構造を完全に確定できないため、単糖組成や N 型、O 型及び脂質などの基本構造に関する情報しか得られない。このような曖昧さを含む糖鎖構造と関連した各種のウェットのデータが産出されていたが、従来は独立したデータベースに格納されてきた。

これまでの統合化推進プログラムにおいてこの問題を解決すべく糖鎖領域におけるデータベースの統合化を進めてきた。第一の問題は糖鎖の表記の違いにより、同一の糖鎖構造をデータベース間で同定し、リンクすることであった。そこで、曖昧さを含む構造情報を誰が表記しても同じ文字列になるように WURCS を開発し、これを利用した国際糖鎖構造リポジトリ（GlyTouCan）を開発し公開してきた [1,2]。GlyTouCan を開発したことにより糖鎖構造に任意なアクセッション番号を割り当てることが可能となった。その結果、アクセッション番号を利用するだけで糖鎖構造が明確に指定することができるようになった。

なお、糖鎖関連データベース間の連携においては、セマンティックウェブ技術を用いて開発を進めた。まず、日本糖鎖科学コンソーシアムデータベース（JCGGDB）のコンテンツ（GGDB, GlycoProtDB や LfDB, PACOnto, GGDOnto など）をそれぞれ RDF 化した。そして新に実験デー

タが得られた糖タンパク質に結合している糖鎖の単糖組成情報を Glycoform データベースとしてセマンティックウェブ化した。このデータを GlycoProtDB と連携させた形でアジア糖鎖研究のコミュニティー（ACGG）の DB として公開 [3]した。

さらに、他の研究領域の研究者が糖鎖のデータへのアクセシビリティ向上を目指して開発を実施した。具体的には、統合化推進プログラムの蛋白質構造データバンクや生命と環境のフェノーム統合データベースとの連携を行った。また、NCBI の PubChem とも糖鎖構造の相互リンクを実施した。

また、海外の主要な糖鎖研究者と連携し、 糖鎖データ形式の標準仕様策定へ向けた活動を行なってきた。ACGG-DB 会議は本課題の活動としてアジア地域の学術交流を目的として主催してきた。また、本年度は日本で開催された MIRAGE のワーキンググループ会議および第 6 回ワーレンワークショップ国際会議において積極的に本プロジェクトの成果物の普及活動を行った。

本プロジェクト成果物を活用することでデータベースの開発者と利用者、学術論文出版社と論文読者の全てにおいて糖鎖情報共有の利便性が向上し、他分野との連携が容易になった。例えば、MIRAGE が提案した質量分析実験の報告ガイドラインに、GlyTouCan のアクセッション番号を推奨することにつながった。そして GlyTouCan においては、国内外の GlyTouCan Partner の約 10 プロジェクトにおよぶ連携が実現できた。これは国際的にセマンティックウェブ化を実施した結果であると言える。今後糖鎖科学分野の研究者だけでなく広く一般のライフサイエンス研究者の利用機会の向上が期待され、ひいては糖鎖科学の発展と普及に寄与できると期待している。

参考資料

1) GlyTouCan: http://glytoucan.org

2) Aoki-Kinoshita K, Agravat S, Aoki NP, Arpinar S, Cummings RD, Fujita A, Fujita N, Hart GM, Haslam SM, Kawasaki T, Matsubara M, Moreman KW, Okuda S, Pierce M, Ranzinger R, Shikanai T, Shinmachi D, Solovieva E, Suzuki Y, Tsuchiya S, Yamada I, York WS, Zaia J, Narimatsu H. GlyTouCan 1.0 - The international glycan structure repository. Nucleic Acids Res. 44(D1):D1237-D1242, 2016.

3) ACGG-DB: http://acgg.asia/acggdb.html

ユーザー発表③ **10月6日11:20—11:35**

統合データベースを活用した肝・胆道系の難治性疾患 “原発性胆汁性胆管炎”の疾患関連遺伝子解析

中村　稔　（長崎大学大学院医歯薬学総合研究科　教授）

原発性胆汁性胆管炎（PBC: primary biliary cholangitis）は、中年女性に好発する比較的稀な慢性胆汁うっ滞性肝疾患で、肝内小葉間胆管を標的とする自己免疫疾患と考えられています。一卵性双生児における発症一致率が 60%と高いことから発症には強い遺伝的素因が関係していることが示唆されていますが、未だ原因不明の疾患で、進行すると肝移植以外には救命方法がないために厚労省の難病にも指定されています。最近、欧米人を対象としたゲノムワイド関連解析（GWAS）や大規模なメタ解析により、30 以上の PBC 疾患感受性遺伝子が同定され、疾患発症経路の解析やサブ解析が精力的に開始されています。

我が国でも、国立病院機構や厚労省難治性の肝・胆道疾患に関する調査研究班参加施設を中心として多数の PBC 症例の DNA 検体と患者情報を収集し、GWAS 共同研究をすすめています。解析に不可欠な健常人対照の遺伝子データは、ライフサイエンスデータベース統合推進事業の一環として構築されているヒトゲノムバリエーションデータベースや東北メデイカル・メガバンク機構 （ToMMo） の Integrative Japanese Genome Variation Database、全ゲノムリファレンス

パネル（1JKPN）などを活用し、今までに欧米人とは異なる日本人に特徴的な PBC 疾患感受性遺伝子構造のあること（*TNFSF15*, *POU2AF1*, *PRKCB*, etc）を見出しています。本講演では、これらのデータベースを活用した我が国における PBC 疾患関連遺伝子解析の現状と今後の展望について紹介致します。

ユーザー発表④ **10月6日11:35—11:50**

公共 ChIP データを活用した 遺伝子制御ネットワーク構築と転写因子予測

川上　英良　（理化学研究所医科学イノベーションハブ推進プログラム　上級研究員）

細胞周期、ストレス応答、代謝といった細胞の多くの状態変化は遺伝子発現の結果として生じている。細胞は外的・内的な環境変化に応じて数千～数万の遺伝子の発現を変化させ組織特異的な細胞機能や恒常性を実現する。この遺伝子発現の制御は主に DNA の配列特異的に結合する転写因子によって行われている。制御に重要な転写因子を同定することは、疾患および正常細胞機能を解明する第一歩として大変重要である。

近年、転写因子をテーマとした多くの研究が行われているが、1000 種類以上存在すると言われる転写因子の活性を網羅的に計測するのはまだ非常に難しい。転写因子の活性は、そのタンパク質の量だけではなく、リン酸化・糖鎖付加・アセチル化といった修飾状態、複合体形成、細胞内輸送など様々な要因によって翻訳後調節を受けており、これらを毎回多くの転写因子について網羅的に測定するのは非現実的である。

一方、マイクロアレイや次世代シークエンシング技術の普及により遺伝子発現を網羅的に計測することは比較的容易になってきている。したがって、転写因子とその制御対象である遺伝子との間を何らかの関係性で繋ぐことができれば、計測が容易な遺伝子発現から計測が困難な転写因子の状態を予測することができると考えられる。

今回、私達は世界中のヒト・マウスの網羅的 ChIP データを 7000 以上集め、再解析を行うことで、転写因子ー遺伝子制御ネットワークを構築した。さらに、確率的な制御関係を考慮した weighted Parametric Gene Set Enrichment 解析（wPGSA）を導入することで、遺伝子発現からヒト・マウスともに 400 種類以上の転写因子の活性を極めて高い精度で予測することに成功した。

本発表では、膨大な公共データを統計的に活用する wPGSA の枠組みを概説するとともに、実際の遺伝子発現データへの適用と検証について紹介する。

参考文献

Eiryo Kawakami, Shinji Nakaoka, Tazro Ohta, and Hiroaki Kitano, (2016), Nucleic Acids Research, doi: 10.1093/nar/gkw355

http://nar.oxfordjournals.org/content/early/2016/04/30/nar.gkw355.full

Web アプリケーション

http://wpgsa.org/

データベース作りとデータ駆動科学

岡田　真人　（東京大学大学院新領域創成科学研究科　教授）

本講演では、スパースモデリングとデータ駆動科学に関して講演を行う[1]。より深く自然を知りたいという飽くなき探究心が、とどまることを知らない計測技術の向上をもたらし、大量の高次元観測データを日々生み続けている。「データ科学時代」における全ての科学分野に共通する重要な課題の一つは、データに隠された規則性を抽出するデータ解析の系統的技術の学理であるデータ駆動科学の構築である[1, 2]。

我々は、その方法論構築のためのキーテクノロジーが、近年、情報科学分野で大きな注目を集めているスパースモデリング(SpM)であると考えた[1]。スパースモデリングの基本的な考え方は(1)高次元データの説明変数が次元数よりも少ない(スパース(疎)である)と仮定し、(2)説明変数の個数がなるべく小さくなることと、データへの適合とを同時に要請することにより、(3)人手に頼らない自動的な説明変数の選択を可能にする枠組みである。

データ科学の有用性の根拠の一つとして、ある分野で用いられる有効なデータ解析手法が他の分野でも有効である例が報告されていることである。Reed は、天文学におけるデータ解析手法が、全く対象とスケールの異なる生命科学でも有効に働くことを報告している[3]。一方、同じように見えるデータでも、データ解析の目的や技術的な点から、データ解析手法が異なる場合がある[2]。我々は、このようなデータと解析手法の間の複雑な多対多の関係を取り扱うために、図 1 に示すデータ駆動科学の三つのレベルを提案した[2]。

これは、神経学者で計算論的神経科学とコンピュータビジョンに多大な影響を与えた David Marr が指摘した脳科学における三つのレベルに基づいている。データ駆動科学においては、Marr の指摘する計算理論のレベルと、表現・アルゴリズムのレベルの間に、モデリングのレベルが存在することにきづいた。これらの三つのレベルを新たに“データ駆動科学の三つのレベル”

と名づけ、データ駆動科学の学理の原点に位置付けた[2]。

最後に、講演のまとめとして、データ駆動科学の三つのレベルとデータベース作りの基本的な指針[4]との関係に関する私見を述べる。

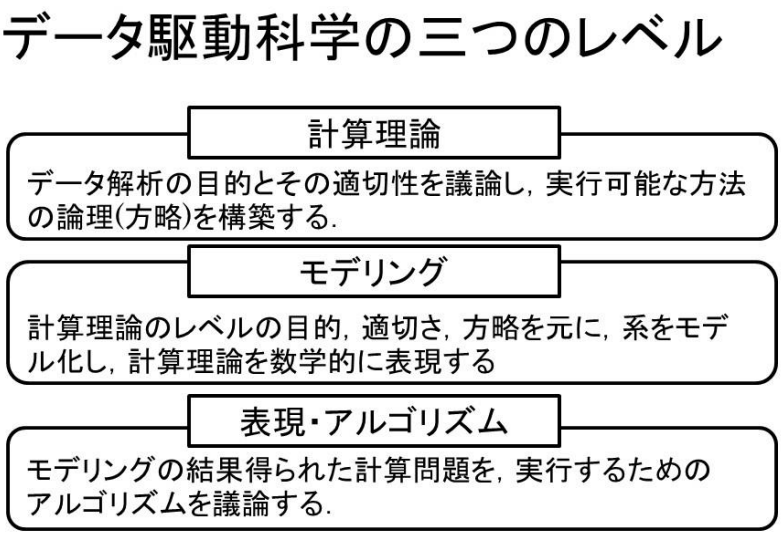


図１　データ駆動科学の三つのレベル

参考文献

[1] 文部科学省科学研究費補助金「新学術領域研究」平成 25 年度～29 年度 スパースモデリングの深化と高次元データ駆動科学の創成

http://sparse-modeling.jp/

[2] Igarashi, Nagata, Kuwatani, Omori, Nakanishi-Ohno and Okada, "Three Levels of Data-Driven Science," Journal of Physics: Conference Series, 699, 012001, 2016.

[3] Reed, "Is There an Astronomer in the House? ", Science, Vol. 331, pp. 696, 2011.

[4] バイオサイエンスデータベースセンター センター長挨拶　高木利久

http://biosciencedbc.jp/about-us/welcome-from-director

成果報告⑨ **10月6日15:55―16:15**

生物の表現型が研究コミュニティをつなぐ

桝屋　啓志
（理化学研究所バイオリソースセンターマウス表現型知識化研究開発ユニット　ユニットリーダー）

　生物は、遺伝や環境の結果として様々な性質＝『表現型』を示す。生命を理解することは、遺伝子と表現型との間にある仕組みを理解することでもあるので、研究成果として得られた表現型情報をデータベースを通じて共有することは、生命科学の発展に関わる重要な課題である。しかし、生物の示すありとあらゆる性質を対象とするため、ゲノムや分子のデータに比べてデータベース化の方法論の開発自体に時間がかかってきた。近年、オントロジーやセマンティックウェブの技術が発達し、標準化された語彙の相互関係として表現型を整理することで、多様なデータを標準化し国際的に共有することが可能になってきた。

　最近では、疾患情報体系化の面でも、表現型の概念が見直されつつある。例えば、疾患は様々な症状（＝表現型）を示すので、疾患や診断の情報を表現型のリストとして整理することができる。そうすることで、症状から見た疾患同士の関連性を情報学的に表現することも可能になり、診断のつかない疾患（未診断疾患）であっても、データとして示すことが可能である。さらには、生物横断的に表現型の類似性をマッピングすることで、生物種を超えた研究成果の共有を従来に比べて格段に効率的に行えるようになった。つまり現代では、「モデル生物」「疾患モデル動物」という考え方を、データベースの世界で計算処理を用いて表現していくことが可能になりつつある。

我々は、NBDC 統合化推進プログラム「生命と環境のフェノーム統合データベース」において、Resource Description Framework (RDF)を基盤としたセマンティックウェブ技術を用いた国内実験動物の表現型データ統合を行ってきた。遺伝的なバリエーションに従って示される表現型を、共通のスキーマを用いて整理統合することで、異なる生物のどんな表現型であっても、同じ検索方式で検索できるようにした。これらのデータの中には、国際マウス表現型解析コンソーシアム（International Mouse Phenotyping Consortium: IMPC）による網羅的表現型解析データ（約3200 系統、115 万データ、2.6 万アノテーション）も含まれ、「表現型ビッグデータ」が形成され始めている。これらのデータは、RDF データの公開基盤である理研メタデータベースを通じて公開され、画面での閲覧と全データダウンロード、SPARQL 言語での検索が可能である。ポータルサイト J-Phenome（http://jphenome.info）から、各 DB にアクセスできる。

技術面では、ウェブの世界における共通 API である RDF の採用が様々な面でメリットをもたらしている。多様な表現型データのデータベース化はもちろん、外部データや共通化ツールの利用による効率的なアプリケーション開発、異なるデータベース間でのデータの共通利用や物理的に離れたデータベースを横断的に検索することによる異分野データの連結、さらには、国内の表現型データを海外に向けて発信するためのデータベース連携にも発展している。表現型のデータを通じて、異なる研究コミュニティが連携していく基盤が構築されつつある。表現型データの世界でも、NBDC の目指す「フェデレーション型統合」の姿が見えてきた。

成果報告⑩ **10月6日16:15―16:35**

学術情報と社会をつなぐ代謝物と生物種のデータベース

有田　正規　（理化学研究所環境資源科学研究センター　チームリーダー）

　オープンサイエンスという言葉が第五期科学技術基本計画にも明記され、多くの人が科学研究の成果にアクセスできることが重要視されている。ここで同じく重要なのが、知識のわかりやすい伝えかたである。世の中にある代謝物の種類や、それらを作る生物の種類。代謝物を生物サンプルからどう検出するのか。その情報や知識を、メタボロミクスのデータベースという観点から紹介したい。本グループでは、大量メタボロミクスデータからの迅速な代謝物同定システム、エコロジーやヘルスケアに役立つデータベース、そして代表的な生物種に含まれる代謝物を概観できるウェブページなどを作成してきた。これらを結びつけることで、学術情報をどのように、わかりやすく発信できるのか、紹介したい。

～統合データベース、
実り
の時へ～
トーゴーの日

成果報告⑪ **10月6日16:35―16:55**

生命科学分野における質問応答システムの実現に向けて

金　進東

（情報・システム研究機構ライフサイエンス統合データベースセンター　特任准教授）

世界的に広く利用されている生命科学分野の学術論文検索サービス PubMed/MEDLINE は 1800 年頃以降に発表された 2600 万件を超える書誌情報を検索対象にしており、毎日平均 3000 件以上登録されている。このことから示唆される通り、生命科学は長い歴史を持ち、非常に多くの知見が蓄積され、そして今なお発展し続けている。これまでに得られた知見は主に文献として収められているほか、様々な切り口で多くのデータベースが構築されている。塩基配列情報やタンパク質情報、疾患情報などその数は 1600 をくだらない。生命科学という極めて広い研究領域において、生命の仕組みを解き明かすために必要な知識は増え続けており、様々なデータベースから効率よく必要な情報を取得しなくてはならない。

このような状況において複数のデータベースを横断的かつ合目的的に検索しやすくする基盤技術として、Resource Description Framework（RDF）というデータ表現方法が注目されている。各データは Uniform Resource Identifier（URI）を用いて識別され、互いに関係づけられることから、ウェブ上において曖昧なく情報を表現できるほか、機械によるアクセスと処理が容易になる。その結果、複数のデータベースから特定の値をキーとして必要な情報を効率的に収集できるなどの利点がある。このような基盤技術環境を実現するデータベースは Linked Data と呼ばれ、European Bioinformatics Institute（EBI）や National Center for Biotechnology Information

（NCBI）において提供されているデータベースもこれを採用し始めている。また、NBDC の RDF ポータルに集められている RDF データセットについても同様である。

　これらの技術的優位性を生命科学研究の発展のために発揮させるには、より多くの生命科学研究者がその有用性を実感できる形で提供されなくてはならないが、現状では必ずしも実現されているとはいえない。たとえば、RDF データベースから必要な情報を取得するためには SPARQL と呼ばれる問い合わせ言語を用いることが一般的であるが、これを使いこなすまでには時間がかかるため、なかなか普及していない。そこで DBCLS では別の方法でアクセスできる技術の開発に取り組んでいる。その一つが質問応答システム LODQA である。すなわち自然言語による質問を機械的に SPARQL に翻訳して各種データベースに問合せし、情報を収集して結果を提示するものである。自然言語は言うに及ばず、SPARQL も汎用的で表現力の高い言語であるから、ひとたび RDF データとして表現しさえすれば、すべてのデータに対して自然言語で問い合わせできるようになる。本発表ではこの LODQA を紹介しながら、それを構成する様々な関連技術を紹介する予定である。