セマンティックウェブ技術による理研データベース群の標準化と公開化

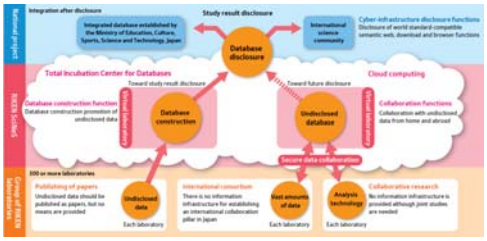
○土井考爾[1)], 国島直樹[2)], 横山茂之[3)], 豊田哲郎[1)]

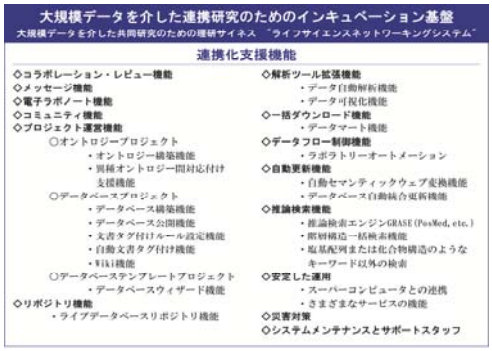
1)理化学研究所生命情報基盤研究部門, 2)理化学研究所放射光科学総合研究センター, 3)理化学研究所生命分子システム基盤研究領域

理研における研究活動によって、シロイヌナズナのオミックス情報、蛋白質構造データ及び付随する実験データが大量に蓄積されている。我々は、そうした情報のキュレーションやアノテーション、統合データベース化を進めている。また、本事業の基盤としてのバイオインフォマティクス用連携システム「理研サイネス」を活用し、データのアノテーションや標準化・共有化を進めてきた。公開データ利用のためのAPIとツール、データベースの検索機能などを提供し、理研外部者でもデータやプログラム公開ができるオープンな参加型システムの構築を図っている。

理研サイネスの特徴

サイネス(SciNeS/SciNetS)は、理研BASEが開発したデータベース構築・連携・公開化の共通基盤システムである。進行中の研究においてデータを研究グループ内のみで共有するところから、公開可能となった成果を既存学術情報と統合化して公開するまでの様々な段階をサポートする。





サイネスの重要な特徴の一つは、セマンティックウェブ技術を基盤としている点にあり、これは以下のような長所をもたらす。

1.「意味リンク」を活用して関連情報を結びつけ、DB横断的に関連情報の検索や閲覧ができる。

2.オントロジーを利用した用語の統一が容易に図れる上、様々な情報の間の関連性を機械可読な形で意味づけて記述できる。

3.個々の情報、およびそれらを結びつけるリンクについてURIといわれるIDが与えられ、さらにそれぞれについて「どういう意味を持つのか」がコンピューター処理可能な一定の書式に従って明示される。これにより、人力で行うよりも複雑で高度な検索やデータ分析が可能になる。

本シンポジウムでは、こうした特徴を生かして進められている植物およびタンパク質関係の統合データベース開発、および、データベース相互連携をさらに進めるための機能開発について報告する。

植物オミックス情報の注釈付けと公開化

理研のシロイヌナズナデータと世界のデータを統合化して公開

Resources from RIKEN
- **RIKEN RARGE Database**
- **RIKEN *Ds* transposon line**
- **RIKEN Activation tagging line**
- **RIKEN *Arabidopsis* Fox-hunting Line**
- **RIKEN *Arabidopsis* phenotype**
- **RIKEN *A. thaliana* gene family**
- **RIKEN *Arabidopsis* protein record**
- **RIKEN OmicBrowse – *Arabidopsis***

Resources opened by TAIR
- **Locus, PO annotation, GO annotation**

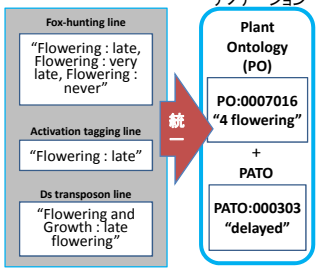
植物科学の分野で盛んに利用される実験植物であるシロイヌナズナに関して、理研では先進的な研究が数多く展開されてきた。こうして蓄積された多様な情報と、TAIRなどの研究機関が公開し広く利用されている公開データベースをサイネス上で統合化している(左図)。イネなど他の植物種のデータベース統合も進めている。

また、これまで整備が遅れていたフェノーム情報の統合に重点的に取り組んでいる。まず、各研究チームでばらばらに分類・定義されていた表現型情報をオントロジーを活用して標準化した。この新しい表現型アノテーションをデータベース化し、従来のデータベースとの間にリンクを張ることで、ある特定の表現型に関係している可能性のある遺伝子の情報をデータベース横断的に検索できる(右図)。現在、シロイヌナズナの表現型情報を扱う文献情報を収集してマニュアルキュレーションを進め、表現型のアノテーションと統合化の範囲拡大を図っている。また、哺乳類の統合データベース開発も進められており、共通の上位オントロジーを利用することにより、両者のフェノーム情報を統合する予定である。

植物オミックスデータ統合データベースは http://scines.org/db/plantからアクセス可能である。

フェノーム情報の統合

オントロジーで表現型情報を標準化



タンパク3000プロジェクトで解明された高等動植物等由来のタンパク質構造データ(X線による結晶構造解析)およびそれに付随する実験データのデータベース標準化

理研構造プロテオミクス研究推進本部(RIKEN Structural Genomics/Proteomics Initiative, RSGI)がタンパク3000プロジェクトで解明した高等動植物等由来のタンパク質構造解析データ(X線による結晶構造解析)と付随する実験データ(タンパク質試料の発現、精製、結晶化、結晶観察、回折実験等)を収集し、公開可能なデータに関しデータ間の紐付けを行った(理研サイネス: http://scines.org/db/ssbc)。これらの実験データはタンパク質試料ごとに相互に関連付けられているので、類似タンパク質の構造機能解析などの参考データとして役立つ。補完課題の最終年度である2010年度は新たに結晶観察画像50万件等を理研サイネスのデータベースに追加し、生命科学研究の発展に貢献することが期待される。最終的に統合データベースの一部として生命科学分野の幅広い研究に役立つ体制を確立する。

理化学研究所のデータベース統合化のための基盤構築

本事業において公開したデータを理研内外のクライアントサイドからウェブ経由でプログラム的に利用できるよう、APIとツールを構築して公開している。

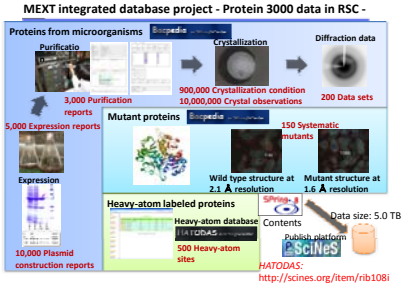
様々なプログラミング言語からアクセス可能な外部DBとの連携インタフェースとして、Semantic -JSONを開発した。利用者はSemantic -JSONを通してJSON@SciNeSと呼ぶサーバにアクセスしサイネス上の情報を利用できる。

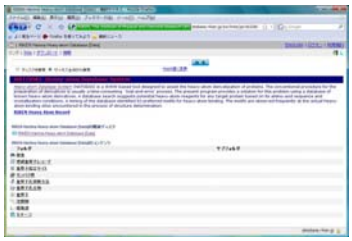
利用者は自分のワークスペースをサイネスに持つことができる。このワークスペースにはウェブブラウザからアクセスでき、サイネス上の情報を利用した様々なバイオインフォマティクス研究用プログラムの開発が可能である。

こうして作成されたプログラムと結果ファイルをサイネス上に蓄積し、公開化・再利用することができる。

理研播磨研究所における微生物由来の蛋白質構造データに付随する実験データ・重原子導入蛋白質に関わる情報のデータベース標準化

理研放射光科学総合研究センターにおけるタンパク質結晶構造解析データベースの内容は、1)進化及び生育環境のいずれの観点からも広い範囲をカバーする9種類の微生物に由来するタンパク質の実験情報(http://scines.org/item/rib220i)、2)タンパク質への変異導入がその結晶の品質に与える影響を研究するため系統的に設計した変異体タンパク質の実験情報(http://scines.org/item/rib220i)、3)タンパク質の重原子誘導体結晶調製に役立つ検索プログラム*HATODAS* (http://hatodas.harima.riken.jp)の基盤実験情報(http://scines.org/item/rib108i)であり、2009年7月に暫定公開された。





1)の微生物データは、類似タンパク質の構造解析などの参考資料として役立つ。2)の変異体データには、均一な条件で結晶化されるなど相互比較を行う上で有利な特長があり、ホモロジーモデリングの高精度化などに利用できる。3)の重原子データでは、重原子ラベルによる生体分子のイメージングなどへの応用が期待される。理研サイネスのデータベース統合は、データの再利用や自働処理化が容易なセマンティックウェブ形式を採用しており、一括ダウンロードや新たなデータベースの構築に適している。補完課題の最終年度である2010年度では、追加データの公開に加え、一般の研究者にとって使いやすい機能を追加し、最終的に統合データベースの一部として生命科学分野の幅広い研究に役立つ体制を確立する。

Licensed under a Creative Commons表示2.1日本 license (c) 2010 土井考爾, 国島直樹, 横山茂之, 豊田哲郎(理化学研究所生命情報基盤研究部門, 理化学研究所放射光科学総合研究センター, 理化学研究所生命分子システム基盤研究領域)