論文の内容の要旨

論文題目　　A study on standardization and interoperability of biological databases
（生命科学データベースの標準化と相互運用性に関する研究）

氏　　名　　片山　俊明

背景

生命科学分野では、1970 年代にタンパク質立体構造・アミノ酸配列・塩基配列といった主要な生体高分子のデータベース化がすでに開始されており、歴史的に研究データを公共のために無償で公開する文化が育まれてきた。これらは現在の基礎生物学からゲノム医科学まで、幅広い研究開発とその発展を支える基盤となっている。

1990 年代に国際的なゲノムプロジェクトが進展するとともに、大規模配列解析のためのワークフロー構築やそのデータベース運用を支える情報科学の発展と、GNU プロジェクトや Linux をはじめとするオープンソース運動が重なりあって、バイオインフォマティクス分野においても自由に利用改変が行えるソフトウェアの開発と共有が進んできた。

現在も生命科学のデータベースはその分量と種類において増加を続けており、バイオインフォマティクス研究ではこれらの膨大なデータを統合的に利用する必要性に迫られている。しかし、個々のデータベースは異なる形式や ID と語彙の体系を持っているほか、新しい技術とともに新規の概念やデータ形式が導入され続けてきている。

このため、必要なデータの検索から取得、データ形式の相互変換、ID の対応関係やデータの意味の整理といった前処理に多くの時間が必要であり、研究のボトルネックとなっている。これを効率化するには国際的な連携によるデータの標準化と相互運用性の向上が求められる。

本研究では、これらの諸問題の解決を目指し、実行環境に依存しないウェブ・サービスの開発と、セマンティック・ウェブによって様々なデータを統合したゲノムデータベースシステムの構築を行った。

データベースアクセスの標準化と相互運用性

ゲノム解析のワークフローを構築するため、2000 年代初頭からオープンソースのソフトウェアとして BioPerl、Biopython、BioJava など、プログラミング言語ごとにバイオインフォマティクス用ライブラリの開発が進展していた。私はゲノムとパスウェイ情報のデータベースである KEGG の構築に従事していたため、

ポストゲノム時代のデータ解析を見据え、Ruby 言語を用いた BioRuby ライブラリの開発を行ってきた。Ruby 言語では、オブジェクト指向によるパスウェイなど複雑なデータのモデル化と、スクリプト言語の特徴である迅速なプログラム開発を両立させることができる。一方で各言語のライブラリを用いるためには、それぞれのインストールとプログラム作成が必要であり、データベースの検索や、データ取得、データ形式の変換といった基本的な情報処理や、構築されたワークフローを他のコンピュータで実行するための環境構築には時間と手間がかかっていた。そこで本研究では、これらの処理をウェブ・サービス化する TogoWS を開発し、インストールの必要性とプログラミング言語や環境への依存性を解消した。

TogoWS は、米国国立生物工学情報センター(NCBI)、欧州バイオインフォマティクス研究所(EBI)、日本蛋白質構造データバンク(PDBj)、国立遺伝学研究所 DDBJ センター(DDBJ)、京都遺伝子ゲノム百科事典(KEGG)、米国カリフォルニア大学サンタクルーズ校(UCSC)の主要なデータベースに対応している。これらの各センターが提供する検索やデータ取得のための機能は統一されておらず、利用者はそれぞれの使い方に習熟する必要があった。さらに、取得される結果も XML や独自のデータ形式など様々で、ユーザーが必要な情報を抽出してワークフローを構築するには、返された情報からそれぞれに対応したプログラムを作成して解析（パース）する必要があることも課題となっていた。

TogoWS では、これらを統一的に扱うため、データの検索、取得、解析、変換において、全てのデータベースに共通の API を設計した。例えばデータベースエントリの取得においては、「http://togows.dbcls.jp」に続いて「/entry/DB/ID」の形式でデータベース名とエントリ ID を指定する。さらにエントリ中の部分要素を「/フィールド名」を付加することにより解析したデータを抽出、最後に出力形式を「.xml」や「.json」のように指定できる。これらは、TogoWS サーバ側に BioPerl や BioRuby の機能を持たせることで実現しており、利用者はプログラムを作成することなく情報を得られるようになった。また、これまで UCSC の提供するヒトゲノムを始めとした膨大なゲノムアノテーション情報を自在に取得するためには、UCSC の MySQL データベースに SQL 言語で問い合わせる必要があったが、TogoWS における Ruby UCSC API の採用により、REST API だけで容易にアクセスできるようになった。TogoWS のサービスは様々なバイオインフォマティクスのアプリケーションから 5 年以上にわたって安定的に利用され続けており、データベースアクセスの標準化と相互運用性に貢献してきた。

データベースコンテンツの標準化と相互運用性

TogoWS により主要なデータベースへのアクセスを共通化でき、データの取得

と加工を行うワークフローの構築は容易になった。しかし、データの持つ意味をふまえた統合的な利用には、データベースの記述内容そのものを標準化するとともに、データ間の有機的な繋がりを明示して相互運用性を高める必要があることが分かってきた。このため、主要なデータベース開発者が参集して議論と技術開発を行う、国際開発者会議 BioHackathon を開催することとなった。2008 年から 10 年以上にわたって開催してきた BioHackathon の中で、データの標準化と相互運用性を向上させるためにセマンティック・ウェブ技術の採用が提案された。

セマンティック・ウェブは World Wide Web (WWW)を作った Tim Berners-Lee 氏が提唱する、データの Web を構築するための標準規格である。セマンティック・ウェブでは、データを指し示す世界共通の ID として Uniform Resource Identifier (URI)を用いる。また、データの意味とデータ間の関係は、Web Ontology Language (OWL)による標準語彙（オントロジー）を用いて表現される。さらに、データのモデル化には Resource Description Framework (RDF)を採用し、主語、述語、目的語の組み合わせ（トリプル）で情報を記述する。最後に、RDF データの検索には SPARQL Protocol and RDF Query Language (SPARQL)検索言語を用いる。これらは WWW コンソーシアム(W3C)によって標準化されており、インターネット上に自由にアクセスできるデータベースを提供するための技術基盤となっている。これに基づき、生命科学における共通の URI とオントロジーを標準化し、各データベースの中身を RDF 化することで、多様なデータの統合を推進することになった。

本研究では、特に多様なデータベースからの情報統合が求められるゲノムのアノテーションにおいて、生物種、ゲノム、遺伝子、表現型、環境に関わるデータを RDF によって統合することで、新たなゲノムデータベース TogoGenome を構築した。

このために、まず BioHackathon における国際連携により、標準 URI として Identifiers.org の利用を推進するとともに、アノテーションの基盤となるゲノム座標系を表現するオントロジーFALDO を標準化した。続いて、国際塩基配列データベース(INSDC)の情報を表現するためのオントロジー、生物種タクソノミーを表現するためのオントロジー、微生物表現型のオントロジー、生育環境のオントロジーなどを国内外の研究者と共同開発した。これらをもとに、RefSeq のゲノム情報と UniProt のタンパク質アノテーションを中心に RDF によるデータ統合を行うとともに、表現型や環境情報のアノテーションを付加した。

さらに、これらの情報を SPARQL 言語で検索し、生物学的に意味のある機能単位ごとに可視化する TogoStanza を開発した。TogoGenome では、遺伝子や環境などのコンテクストごとに最適な TogoStanza を組み合わせて表示することで、

関連する情報をまとめたレポートページを構成している。また、各 TogoStanza はそれぞれ再利用が可能なため、国内で同時期に開発されてきた MicrobeDB.jp や CyanoBase など、複数のゲノム関連データベースで相互に利用することにより、開発を効率化することができた。TogoGenome の構築を通じて生命科学の主要なデータベースの RDF 化が進展し、これからのデータサイエンスを支えるデータベースコンテンツの標準化と相互運用性の向上を実現することができた。

結果と考察

膨大なデータを擁する生命科学データベースの統合的な利活用を実現するために、データベースへのアクセス方法及びデータベースの内容の標準化と相互運用性の向上を目指した研究とシステム構築を行った。このような標準化の実現は一機関だけで達成できるものではないため、国際連携を図りながら継続的に開発を行う必要があった。

また、データの RDF 化によって、これまでのデータベースでは困難だった、データの意味に基づく検索や解析を実現することができるようになってきた。例えば「海水と淡水に生息する藍藻の間で、環境応答に関連するヒスチジンキナーゼ遺伝子の組成を比較する」といった検索は、これまででであれば、多数のデータベースを駆使して、ゲノムの決まった藍藻を選択し、その生育環境を調べ、遺伝子セットを取得し、各遺伝子のアノテーションを確認して集計する、といった膨大な作業が必要とされたが、TogoGenome では解析プログラムを作成することなく実現できている。また、Togogenome では Pfam 機能ドメインの組成を、ヒト、マウス、魚類とホヤの間で比較することにより、脊椎動物以降に発達した適応免疫関連遺伝子を探索したり、ホヤにおいて酸素結合因子として鉄の代わりにバナジウムを利用している可能性のある機能ドメインを見出すといった比較ゲノム解析システムを実現することができた。

ウェブ・サービスによる分散データの統合的な利用から、セマンティック・ウェブによるデータの意味的統合を進めることで、これまでの技術では実現できなかった新しいデータ利用の可能性が見えてきた。今後は医科学分野への応用や、機械学習などによる高度な解析技術の研究を通じて、より幅広いデータの統合と、それに基づく新しいデータサイエンスの構築を目指したい。