論文の内容の要旨

応用生命工学専攻

平成 23 年度博士課程　入学

氏名　込山悠介

指導教員名　清水謙多郎

論文題目

Database Integration and Software Development for Bioinformatics using Computational Biosemantics

（計算バイオセマンティクスを用いたバイオインフォマティクスのための

データベース統合とソフトウェア開発）

背景

申請者は計算バイオセマンティクス（Computational Biosemantics）をセマンティックウェブ（Semantic Web）技術用いて、生命科学の知識をグローバルに統合化し、新しい生物学的な発見を得るためのデータマイニング手法と定義する。元々、バイオセマンティクスは哲学分野で定義されたと言葉であり、その意味は生命科学意味論である。それは生命現象の記号化によって生命の意味と価値を議論する学問分野である。その応用として生命科学分野の Gene Ontology が生まれた。オントロジー（Ontology）もまた哲学用語であり物事の存在とルールを定義する学問領域である。農学・創薬・医薬の各領域において生物情報工学の研究を行う上での問題点として、バイオインフォマティクス研究者が多数のデータベースを巡回して、データを集めるためにスクリプトのコーディングをしていたことが挙げられる。従来の関係データベースでは運用中にデータのスキーマを再設計することは危険とされ、困難をともなう。そして、データベースごとに規格が乱立しており、フォーマットも共通ではない。また、生命科学データベース研究者が新たにレコードの拡張を試みた場合、データベースのダンプデータの利用ライセンスが不明瞭なケースも多くある。一方、セマンティックウェブではベースとなる RDF （Resource Description Framework）を用いてグラフ構造で表現できる。グラフを結合して、そのネットワークを拡張することで、運用中のデータベースに新しい種類のレコードを容易に追加することができる。そして、URI（Uniform Resource Identifier）によってオリジナルのデータが何処のドメインに帰属するかが同定され、データの質が担保される。加えて、高い機械可読性を持つため

計算機ソフトウェアを柔軟に設計できる。LOD（Linked Open Data）の網を広げることでデータのライセンス問題もクリアすることが期待できる。本研究では、従来法の弱点を克服する目的で、セマンティックウェブを基盤とした生命科学データベースを開発することを提案する。バイオセマンティクスにより変革を遂げる近い将来のバイオインフォマティクスでは、RDF が標準フォーマットとなる。本研究の目的は、網羅的な分子間相互作用であるインタラクトーム（interactome）におけるバイオセマンティクスを実現するためのデータベース統合とツール開発の支援である。その意義は、データ自身の次世代化により知識循環プロセスの速度を上げ、情報生命科学者の研究開発の促進と、実験生命科学者の利便性を高めることにある。

対象と手法

我々はどのように領域特有の LOD を構築すればよいか？本研究ではインタラクトミクスの機械学習予測ツールの開発のために、RDF と OWL（Web Ontology Language）をデータセットとして利用する技術を紹介する。さらに、インタラクトーム LOD をグラフデータベースへ実装し、SPARQL（SPARQL Protocol And RDF Query Language）を使用した検索例を考案する。これらはセマンティックウェブの要素技術である。我々はインタラクトームの LOD の作成に PDB（Protein Data Bank）、PDB Ligand、UniProt（Universal Protein Resource）の RDF データを利用する。これに独自に RDF 化した EBI（European Bioinformatics Institute）SIFTS（Structure integration with function, taxonomy and sequence）のデータセットを加える。さらに、Saad らの先行研究において Octree を用いて計算された PDB 中のタンパク質–リガンド間の原子間距離を RDF 化する。本研究において、我々は大きなサイズの RDF シリアライズには Ruby、Python、Perl などの RDF ライブラリを使用する。中程度のサイズのデータに対しては Open Refine（旧 Google Refine）でスキーマを設計とシリアライズを実行する。フォーマットの変換には Redland RDF Libraries、rdf2rdf や ConvRDF が有用である。作成する LOD のコンテンツは、PDB 中のタンパク質–リガンド間の原子間距離の情報、チロシンキナーゼ（Tyrosine Kinase）と相互作用を持つリガンドタンパク質との組合せの情報、アミノ酸残基レベルでの糖（炭水化物）結合タンパク質部位の情報、SIFTS に基づくタンパク質鎖レベルでの他主要データベースへの ID マッピングの情報である。我々はこれらの情報について RDF スキーマモデルを設計し、Allegro Graph 4.11 と Virtuoso 7 の RDF ストア中へ実装する。そして、全成果をウェブからアクセス可能な SPARQL エンドポイントで即時公開する。これらのデータベースでエントリー数が 10 億件（トリプル）を超える RDF を操作するためには、約 500 GB から 1 TB の主記憶装置を持つ高性能サーバーが必要である。上述のインタラクトーム LOD をバックエンドのデータベースとする。そして、WWW（World Wide Web）上でプログラムアクセスできる SPARQL エンドポイントを一つのウェブサイト上に集約し、ドキュメントを付加する。EBI RDF Platform によく似たサービスの提供である。これらのエンドポ

イントを使い、いくつかのタンパク質–リガンド間相互作用予測ツールを短期的に開発する。これはインタラクトーム LOD を用いた機械学習の研究開発支援という立場である。開発された予測ツールはウェブアプリケーションのワークフローのモジュールとして利用できる。このワークフローのウェブフレームワークには Galaxy を使用する。また、アプリケーション開発のために我々は予めイメージコンテンツを準備する必要がある。申請者はオープンライセンスで利用できるイメージライブラリを構築する。

結果と考察

分子間相互作用のための統合グラフデータベースが開発された。それは UTProt RDF Platform と命名された。その内分けは PDB と PDB Ligand を基にした PLBSP (Protein Ligand Binding Site Pair database)、UniProt と EBI SIFTS から作成された RDF-SIFTS である。PLBSP と RDF-SIFTS は２つのコアデータベースであり、それを基にデベロッパーが目的別に軽量化した SBP（Sugar Binding Site database）や PLBSP residue などのサブセット LOD が誕生した。2013 年 12 月現在の UTProt の総トリプル数は 300 億トリプルである。**表 1** に各データの詳細を示す。

表１　UTProt RDF Platform で統合化された RDF のトリプル（エントリー）数の内訳。

UTProt database name	triples [#]
PLBSP: Protein–Ligand Bindig Site Pair (with atom coordinate)	1,337,648,853
PLBSP2PDB	52,582,582
TKIP: Tyrosine Kinase Interaction Pair	440,806
SBP: Sugar Binding Pair	20,202,647
RDF-SIFTS version 2.0	10,527,560
name of inclusive public database	**triples [#]**
UniProt 2011	3,104,363,044
PDB 2013 (with atom coordinate)	25,259,162,986
PDB 2011 (no atom coordinate)	612,736,982
PDB Ligand	28,040,694
Total	30,425,706,154

次に SPARQL を用いた統合データベースの検索事例を挙げる。申請者は PLBSP と RDF-SIFTS と外部の公共データベースの SPARQL エンドポイントを組み合わせ、約 80 の SPARQL クエリを考案した。これらのデータベースは糖結合タンパク質の部位予測ツールの作成を支援した。開発されたプログラムはパイプラインにおいてモジュール化されている。そのパイプラインはウェブ上で利用できるワークフローシステム UTProt Galaxy として公開された。**図 1** に Galaxy でのワークフローを作成する画面で例示する。

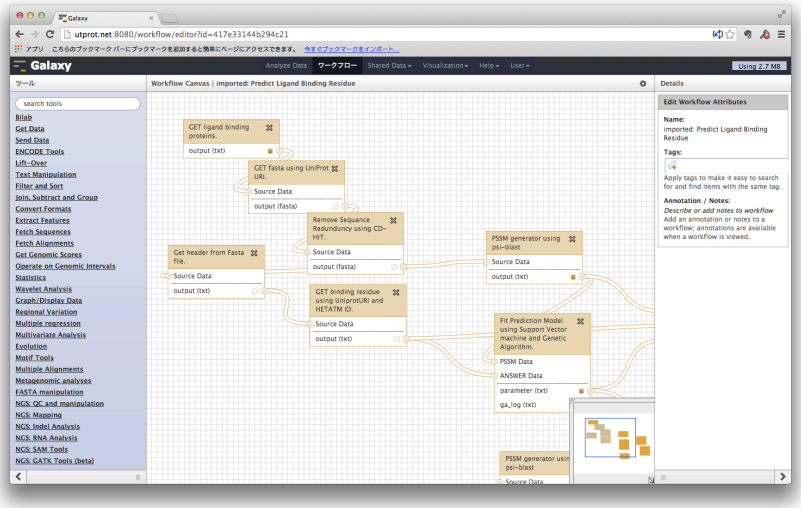

図 1　UTProt Galaxy におけるワークフローの実行画面。ウェブブラウザからアクセスしてグラフィカルユーザーインターフェース上で予測ツールを利用することができる。

UTProt Image ではアプリケーションを開発する際に必要となるグラフィックスにオープンライセンスを与えて公開している。約 48 万件のタンパク質とリガンドのイメージが登録されている。そのうち 40 万件は HPC（High Performance Computing）によって Ligplot で並列計算したタンパク質–リガンド結合部位の 2D グラフィックスである。これら一連のインタラクトームのためのツールとデータベースを統合したポータルサイトが UTProt である。

結論

申請者はバイオセマンティクスの手法を用いることで、バイオインフォマティクス研究者がウェブ上に散在した生命科学データベースを巡回しなくても、必要な情報を入手できる RDF プラットフォームを開発した。それらの基盤データベースを拡張し機械学習による分子間相互作用予測ツール開発の支援を行った。その結果、アプリケーションの開発期間を大幅に短縮することができた。予測ツールはウェブ上でワークフローのモジュールとして実行できる。データベース開発者向けにオープンライセンスが付与させたイメージ 48 万件を公開した。ライセンスは全て『Creative Commons 表示 (CC-BY)』で運用する。いずれもウェブアプリケーションとして公開しているため、バイオインフォマティクスの非専門家にとっても利用しやすい実装といえる。セマンティックウェブは将来的にバイオインフォマティクスの基盤になり、その上で二次データベースやアプリケーションが出現し、WWW を通じたさらなる知識発見を行うのが計算バイオセマンティクスである。UTProt のサービスはウェブブラウザから http://utprot.net へアクセスすることで利用できる。