

本論文は、現在の生命科学データベース開発の現状と問題点、その解決のための情報学の手法について述べたものである。はじめに、農学・創薬・医薬などの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解析や手法開発に必要なデータセットを複数のデータベースから時間と手間をかけて収集している点が挙げられている。また、データベース開発者にとって生命科学データベースはデータベースごとに規格や設計図が乱立しており、ファイル形式も共通ではないことがデータベース統合の障壁となっている。本研究の目的は、網羅的な分子間相互作用(インタラクトーム) 研究におけるデータベース統合とツール開発の支援としている。その意義は、知識循環のサイクルを短縮しバイオインフォマティクス研究の促進および実験生命科学者の利便性を高めることにあるといえる。本研究の手法として人工知能の一分野であるセマンティックウェブ(メタデータによりデータの意味を参照できるウェブ) とオントロジー(記述論理を用いて概念を階層的に表現する手法) をインタラクトームに適用している。セマンティックウェブではグラフ構造を持つデータ(リソース) からデータモデル(スキーマ) を設計する。データモデルの標準規約が RDF (Resource Description Framework) であり、オントロジーの記述様式の 하나가 OWL (Web Ontology Language) である。複数のデータおよびデータベースをグラフに変換し、結合することでデータのネットワークを形成する。また、RDF では URI (Uniform Resource Identifier) により、リソースがウェブ上でどのドメインに帰属するかを同定することができる。これは公共データベースのリソースや論文をとまなうオントロジーから統制語彙を使うことで、その質が担保されることを意味する。さらに、RDF は高い機械可読性を持つため、ウェブを介したソフトウェアを柔軟に設計することができる。RDF とオントロジーを組み合わせた高度な知識ベースの推論が可能なグラフデータを Linked Data とよび、その中でもウェブ上でオープンアクセス可能なデータについて特に Linked Open Data (LOD) という。本研究では、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している。申請者はセマンティックウェブ技術を用いて、生命科学の集合知を統合し、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を計算バイオセマンティクス (Computational Biosemantics) と定義している。

第1章は序論で、本研究の背景、目的を述べている。本論文は全4章から構成されている。

序論に続き、第2章では具体的なデータベースシステムの構築手法について述べられている。本章ではインタラクトームにおける機械学習を用いた予測ツールの開発のために、RDF と OWL を利用したプログラミング技法を紹介している。さらに、実装したインタラクトームのデータベースに対して、グラフデータベースのクエリ言語である SPARQL (SPARQL Protocol And RDF Query Language) を使用した検索の実例を考案している。次に、本研究で用いた公共データベースや新たに作成したデータセットについて述べられている。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。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ている。さらに、Gul 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 化している。本研究において、大きなサイズの RDF シリアライズには Ruby, Python, Perl などのプログラミングの RDF ライブラリを使用している。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している。フォーマットの変換には Redland RDF Libraries, rdf2rdf や ConvRDF が有用とされている。作成した 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。これらのデータについて RDF スキーマモデルを設計し、インタラクトーム LOD としてデータセット化している。統合されたインタラクトーム LOD をグラフ構造用のデータベース AllegroGraph 4.11 と Virtuoso 7 中へ実装している。データベースはウェブからアクセス可能な SPARQL

エンドポイントとして API (Application Program Interface) を公開している。応用として前述のインタラクトームのデータベースをバックエンドとして、ウェブアプリケーションを開発している。2章後半ではその手法について述べられている。SPARQL エンドポイントを一つのウェブサイト上に集約し、サンプルクエリなどを含めたドキュメントを付加している。これらのエンドポイントを使い、複数のタンパク質-リガンド間相互作用予測ツールが短期的に開発された。実際に平成 25 年度の JST NBDC の統合データ解析トライアルにおいて、4ヶ月間の研究プロジェクトでは基盤技術として採用された。加えて、開発された予測ツールをエンドユーザー (バイオインフォマティクス研究者、実験研究者) が利用しやすいようにウェブアプリケーション化している。ウェブアプリケーションのワークフローには Galaxy を使用している。

第 3 章では開発された成果物であるデータベースとソフトウェアについて述べられている。本研究ではセマンティックウェブをベースにしたインタラクトームの統合データベースが開発された。それは UTProt RDF Platform と命名された。インタラクトーム LOD は PDB と PDB Ligand の RDF をベースにした 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 UTProt RDF Platform の2つのコアデータベースである。それをもとに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 が作られた。2014 年 2 月現在の UTProt RDF Platform のエントリー総数は 300 億トリプルと報告された。次に SPARQL を用いた統合データベースの検索事例を挙げている。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 クエリを考案している。これらの要素技術は、糖結合タンパク質の部位予測ツールの作成を支援した。開発されたツールはウェブアプリケーションのワークフローとしてモジュール化されており、UTProt Galaxy として公開された。これら一連のインタラクトームのためのデータベースとツールを一元化したポータルサイトを UTProt と命名している。

第 4 章では結言として、本研究の成果が農学生命科学における知識循環にとって有用であることを述べている。申請者は計算バイオセマンティクスの手法を用いることで、インタラクトームのデータベースを統合し、目的に応じたデータセットを再編纂できる RDF 版データプラットフォームが開発された。その基盤データベースを拡張することで、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ている。その結果、予測ツール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アプリケーションのワークフローのモジュールとして実行可能としている。当初、バイオインフォマティクス研究者に向けた開発支援として研究が進められた。しかし、ウェブアプリケーションでのワークフローの実装により予測ツールを GUI (Graphical User Interface) 上で連結して実行できることから、実験研究者にとっても利用しやすい成果に達したといえる。研究成果である UTProt の各サービスへは、ウェブブラウザからポータルサイト <http://utprot.net> へアクセスすることで利用できる。

申請者が提唱する計算バイオセマンティクスとは、セマンティックウェブ技術を通じた生命科学データベース開発と、それを活用したデータマイニングによる知識発見である。大量の実験データから生物学的な意味を見出すためには、解析や検証に活用できるデータベースの整備が今後必要であるが、本研究成果は実験研究者とバイオインフォマティクス研究者の知識循環を、計算バイオセマンティクスによって円滑にした一つの実例といえ、農学生命科学において学術上、応用上貢献したといえる。よって、審査委員一同は、本論文が博士 (農学) の学位論文として価値あるものと認めた。