1G3-4

# セマンティック・ウェブを利用した遺伝子発現データの解析プラットホーム

## Interactive platform for gene expression analysis based on Semantic Web technologies

片山 俊明[*1] 菊池 正隆[*2] 荻島 創一[*3]
Toshiaki Katayama Masataka Kikuchi Soichi Ogishima

[*1] ライフサイエンス統合データベースセンター
Database Center for Life Science

[*2] 新潟大学 脳研究所/NEDO
Brain Research Institute, Niigata University

[*1] 東北大学 メディカル・メガバンク機構
Tohoku Medical Megabank Organization, Tohoku University

We developed an application which integrates various biological datasets and provides a faceted query interface for interactive analysis of gene expression data. In the course of gene expression analysis, it is required to interpret data by referencing knowledge bases of genetics, pathways, diseases and drugs. However, because those external resources are often stored in distributed databases in various formats, it is hard for biomedical scientists to use them in combination. Semantic Web technologies are suitable for integration of those heterogeneous datasets using Resource Description Framework (RDF) and providing a faceted search interface. In this work we applied this platform to the gene expression analysis of Alzheimer's disease.

## 1. はじめに

医学生物学研究者が手元にもつ遺伝子発現データを解釈するためには、様々なデータベースを参照して、データドリブンに仮説を検討する必要があるが、このためのインタラクティブなプラットホームはほとんどない。本研究では、遺伝学、パスウェイ、疾患・医薬品、文献などの様々な知識の整備が進んでいるアルツハイマー病について、遺伝学、パスウェイ、疾患・医薬品、文献などの様々な外部知識を参照しながら仮説を検討するインタラクティブなプラットホームとして Linked Open Alzheimer's Disease (LOAD) を開発したので、これを報告する。

## 2. データ統合と解析環境

### 2.1 セマンティック・ウェブによるデータ統合

アルツハイマー病の遺伝子発現解析では、遺伝学、疾患・医薬品、文献などの外部知識にくわえて、遺伝学における疾患感受性遺伝子のリストである AlzGene[1] [Bertram 07] やパスウェイマップである AlzPathway[2] [Mizuno 12] などのアルツハイマー病に特有のリソースを参照しながら、どの遺伝子が疾患の発症 (pathogenesis) や進行 (progress) などに関係があるという仮説を検討する必要がある。しかし、通常これらの外部知識は個々のデータベースに格納されているため、組み合わせて参照することが困難である。また、参照したい外部知識は CSV やパースの必要な DB エントリなどの様々な形式で公開されており、これらを統合的に再利用することは医学生物学研究者には容易ではない。近年、ライフサイエンスの分野においてもこのような異種多様なデータの統合においてセマンティック・ウェブ技術の利用が普及している[Katayama 14]。セマンティック・ウェブでは検索対象となる様々なデータを RDF に変換することで容易に追加でき、データ間で関連する情報は URI を用いて結合されるため、生命科学の多様なデータを統合化するのに有効である。

### 2.2 既存の解析プラットホームとの比較

医学生物学研究者が遺伝子発現データを解析する際には、R[3] の BioConductor[4] や Spotfire[5] などのソフトウェアが利用されることが多い。BioConductor では、遺伝子アノテーションの機能や様々な統計解析ライブラリが開発されているが、これらを組み合わせて利用するためにはプログラミングが必要でインタラクティブな試行錯誤は簡単ではない。Spotfire ではグラフィカルなインターフェイスが提供されており、外部のアノテーションデータを取り込んだインタラクティブな解析が可能であるが、商用であるため医学生物学研究者が自由に利用できるサービスとして提供することは難しい。

本研究で開発した LOAD[6] は、インタラクティブな操作でウェブサイトから自由に使えるため、遺伝子発現データの解析を容易に試行できる。さらにセマンティック・ウェブにより、遺伝子のゲノム上の位置、関連するパスウェイ、医薬品など、複数の条件を組み合わせて関連データを絞り込むファセット検索を容易に行うことができる。また、絞り込んだ遺伝子から関連情報へのリンクを辿れる点はウェブアプリケーションのメリットといえる。

連絡先：片山俊明，大学共同利用機関法人 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，〒277-0871 千葉県柏市若柴 178-4-4 東京大学柏の葉駅前研究棟 6 階，電話 04-7135-5508，Fax 04-7135-5534，ktym@dbcls.jp

[1] http://www.alzgene.org/
[2] http://alzpathway.org/
[3] http://www.r-project.org/
[4] http://www.bioconductor.org/
[5] http://spotfire.tibco.com/
[6] http://load.sysmedbio.org/

## 3. 生物学データの RDF 化

システム構築にあたり、遺伝子発現解析に広く使われている Affymetrix 社の GeneChip のプローブとアノテーションの情報、UniProt のタンパク質機能アノテーション、DrugBank の医薬品とタンパク質の相互作用データ、OMIM の遺伝子と疾病の関係データ、AlzPathway の分子パスウェイデータを RDF 化した。

### 3.1 プローブ ID と遺伝子 ID

遺伝子発現解析実験に使われるマイクロアレイには、プローブとよばれる数十万もの短い配列断片が載っている。各遺伝子に対し複数のプローブが対応するため、発現データで使われるプローブ ID と遺伝子 ID の対応を取る必要がある。LOAD では、これを RDF のリンク情報として持たせ、発現量の増減分布はプローブで選択し、残りのファセット検索では対応する遺伝子 ID での絞り込みを行うこととした。

### 3.2 タンパク質の機能アノテーション

Swiss Institute of Bioinformatics ではタンパク質のアミノ酸配列とその機能アノテーションのデータベース UniProt の作成とその RDF 化を以前から進めてきており、とくに外部データベースとのリンク情報はファセット検索の構築に有用である。とくに遺伝子オントロジー(GO)、立体構造(PDB)、機能ドメイン(Pfam)、医薬品(DrugBank)、希少疾患(Orphanet)、遺伝病(OMIM)等との関係情報は、様々な外部データベースの情報を LOAD に追加する際に RDF のリンクを束ねるハブとして機能している。

### 3.3 医薬品や疾患情報

UniProt からリンクされているもののうち、医薬品の薬理とターゲットとなるタンパク質などのデータベース DrugBank からは、リンクされた ID に対応する分子名などの情報を抽出して RDF 化した。同様に遺伝子疾患については OMIM データベースから関連情報を RDF 化して利用した。

## 4. ウェブアプリケーションの実装

### 4.1 トリプルストア

作成した RDF と UniProt の RDF は、無料でオープンソース版が利用できること、SPARQL 1.1 をサポートしていること、RDF データのインポートが比較的高速であることから Virtuoso 7 に格納してエンドポイントを構築した。

### 4.2 パスウェイ表示

アルツハイマー病の分子パスウェイ AlzPathway は Systems Biology Markup Language (SBML) に準拠した XML 形式のデータが利用できるため、ここから各分子の座標情報などを抽出した。パスウェイの画像については CellDesigner [Funahashi 08] を利用して出力し、Google Map Image Cutter で処理し Google Map API を利用して表示した。

### 4.3 ウェブアプリケーション

ユーザはデフォルトのデータセットもしくはユーザのアップロードした遺伝子発現データをもとにインタラクティブな解析を行う。サーバ側のアプリケーションは Node.js で実装し、ウェブインターフェイスから受け取った情報を SPARQL クエリに変換、内部処理した上でクライアント側に反映している。データ解析の試行錯誤が容易になるよう、Ajax を利用した画面遷移のないアプリケーションとした(図 1)。

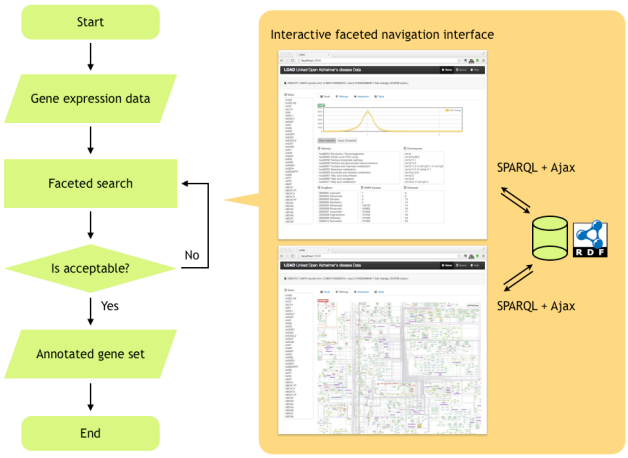


図1: LOAD のセマンティック・ウェブによるインタラクティブなファセット検索を利用した遺伝子発現解析のワークフロー

### 4.4 デフォルトのデータセット

手元に遺伝子発現データを持たないユーザも公共データを利用した解析ができるよう、NCBI の GEO データベースからアルツハイマー病に関連する公開データセットとして

- GSE4757 Neurofibrillary tangles
- GSE16759 Parietal lobe cortex
- GSE18309 Peripheral Blood Mononuclear Cells
- GSE28146 Hippocampus of incipient patients
- GSE29652 Astrocyte (ApoE genotype)

については選択するだけで利用できるようにした。

## 5. まとめ

セマンティック・ウェブを利用した遺伝子発現データ解析のためのプラットホームを構築し、多様なデータの統合に RDF が有効であること、ファセット検索と SPARQL の親和性、インタラクティブなウェブアプリケーションでの実用性などを確認した。ここ数年の RDF によるデータ公開の流れと、トリプルストアの性能向上により、生命科学者にも利用できるアプリケーションがセマンティック・ウェブを元に構築できるようになってきたといえる。一方でまだ RDF 化がなされていないデータベースも多いため、より広い分野をカバーするサービス構築のためにはオントロジー整備などを含めた継続的な基盤整備が必要である。

## 参考文献

[Bertram 07] Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE: Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database, *Nature Genetics*, Vol. 39, No. 1, pp. 17-23, 2007.

[Mizuno 12] Mizuno S, Iijima R, Ogishima S, Kikuchi M, Matsuoka Y, Ghosh S, Miyamoto T, Miyashita A, Kuwano R, Tanaka H: AlzPathway: a comprehensive map of signaling pathways of Alzheimer's disease, *BMC System Biology*, 6:52, 2012.

[Katayama 14] Katayama T, *et al.*: BioHackathon series in 2011 and 2012: penetration of ontology and Linked Data in life science domains. *Journal of Biomedical Semantics*, 5:5, 2014.

[Funahashi 08] Funahashi A, Matsuoka Y, Jouraku A, Morohashi M, Kikuchi N, Kitano H: CellDesigner 3.5: A Versatile Modeling Tool for Biochemical Networks, *Proc. IEEE*, Vol. 96, No. 8, pp.1254-1265, 2008.