1N5-OS-10c-4

微生物表現型オントロジーおよびLODの開発

Developing an ontology and LOD for describing phenotypes of microbes

川島 秀一[*1]
Shuichi Kawashima

岡本 忍[*1]
Shinobu Okamoto

[*1] 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
Database Center for Life Science, Research Organization of Information and Systems

In biomedical research field, diverse and large amount of data has been produced, which has been resulted in lots of databases in the field. Database Center for Life Science in Japan has focused on building infrastructure to enable an effective use of such data in an integrated way by using Semantic Web technology. As part of the efforts, we have been developing Microbial Phenotype Ontology, which describes the relationships between microbial species and corresponding phenotypes, and a linked open data (LOD) dataset based on it. This paper presents the newly developed resources and discusses how to utilize them in integration with other life science-related LODs that already exist.

1. はじめに

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（以下 DBCLS）は、生命科学分野にこれまで蓄積されてきた大量のデータに対して、メタデータの付与、フォーマットや用語の統一等を行うことで再構築し、利用者がより効率的にデータを活用できる環境の構築を目指している。具体的には、既存データの Linked Open Data（LOD）化や、そのために必要なオントロジーの整備といった、セマンティックウェブ関連技術を活用することで、この目標を実現しようとしている [山口 2010]。

ただし、生命科学分野には、既にデータベース化されている情報だけでも膨大に存在しており、それらを網羅的に扱うことは困難である。例えば、Nucleic Acids Research 誌は、毎年 1 月に生命科学系データベースの論文を紹介する号を出しており、さらにこれまで紹介されたデータベースについてポータルサイトを提供しているが、2013 年の時点で 1,512 もの データベースが掲載されている [Fernández-Suárez 2013]。このため、DBCLS では、まずゲノム配列情報を RDF 化し、それを中心として、適宜、各種関連データを LOD 化していくことで、データの統合化を行なっていくという戦略をとっている[*1]。

本稿では、ゲノム配列を中心とした生物学データの統合の一貫として我々が取り組んでいる、微生物の表現型データの LOD 化に関して報告したい。また、その際に必要となるオントロジーの開発についても紹介する。

2. 微生物の表現型

2.1 ゲノム配列を中心とした生命科学データの LOD 化

DBLCS が生命科学データの LOD 化によるデータ統合において、ゲノム配列情報を中心にすえたのにはいくつかの理由がある。まず、タンパク質配列情報については、すでに網羅的なタンパク質データベースである UniProt が、RDF 化されたデータとして公開されている [The UniProt Consortium 2013] 。一方で、相補的な関係にあるゲノム配列情報に関しては LOD 化された情報を公開している研究機関がまだないため、ゲノム配列情報 LOD を提供することは価値があると考えられる。次に、特定の生物種を記述する情報として NCBI が提供しているタクソノミー ID が広く利用されているが、これはゲノム配列情報と一対一で対応がつけられる。タクソノミーID は、URI も提供されていることから、生物種に関わる情報を LOD 化する際に都合がいい。さらに、ゲノム配列には、生命現象を担う遺伝子およびタンパク質や機能性 RNA 等の遺伝子転写産物（以下まとめて遺伝子と呼ぶ）がコードされているが、生命科学分野の研究成果は、論文やデータベースというかたちで遺伝子に紐付けられることが多く、それらの情報は、遺伝子の ID や、ゲノム配列の座標を用いることで、ゲノム配列上に集約することが可能である。すでに、ゲノム配列上に登場する遺伝子群を記述するオントロジーとしては Sequence Ontology [Eilbeck 2005] が利用でき、また、ゲノム上の座標を記述するための、FLADO オントロジー[*2] が BioHackathon 2012[*3] で提案され開発中である。このように、ゲノム配列 LOD を構築する環境は整いつつあり、現在、DBCLS で RefSeq データベースを基にドラフトバージョンの LOD を開発している。

2.2 微生物データの意義

DBCLS では、前述のようにゲノム配列データを中心に質の異なる様々な研究結果のデータ統合を目指している。本研究では、実証実験として、細菌を中心とした微生物に焦点をしぼって、データの LOD 化、オントロジー開発を行なった。

微生物は、地球上のあらゆる場所に生育しており、その多様な生きざまを通して、すべての生物に必須要素である炭素や窒素、およびリンの地球規模での循環に関わっている [Stanier 1986]。例えば、植物の根に共生し土壌肥沃を維持したり、土壌中の有害物質を浄化する微生物が多く存在する。また、人間との直接的な関わりも深く、醸造やチーズ、パン、醤油など、食物の生産過程において、数千年にわたって微生物が利用されてきた。一方で、微生物の引き起こす食品腐敗や食中毒やコレラ菌、天然痘、マラリア原虫のような微生物に起因する病気に苦しめられてきた歴史がある。しかし、近年は抗生物質や医薬品として利用できるポリペプチドの生産や、バイオテクノロジーの進歩

連絡先：川島 秀一
ライフサイエンス統合データベースセンター，〒113-0032 東京都文京区弥生 2-11-16 東京大学工学部 12 号館 5 階，kwsm@dbcls.rois.ac.jp

*1 http://ceur-ws.org/Vol-952/paper_42.pdf
*2 https://github.com/JervenBolleman/FALDO
*3 http://2012.biohackathon.org/

にともなって、アセトンや酢酸などの化学製品合成が可能になっている。また、微生物は、取り扱いが容易であり、他の高等生物とくらべて倫理上の問題が少ないことから、生化学や遺伝学のモデル生物として広く用いられ、生命科学研究の発展に大きく寄与している。このように微生物は、環境、医薬、食品、技術、研究の分野で重要な役割をはたしており、過去の知識の蓄積と近年目覚ましい進歩をとげるゲノム科学分野の情報を結びつけることによって、さらなる生命科学の発展と、産業への応用が期待できる。

2.3 微生物における遺伝子型と表現型

微生物個体から得られる情報には、ある生物個体が持つ遺伝子の構成を表す遺伝子型(genotype)とある生物のもつ遺伝子型が性質として表現された表現型(phenotype)がある(図 1)。遺伝子型から表現型を説明できるようになることは、微生物学の大きな目標の一つである、近年、計算機により遺伝子型と表現型のシステマティックな比較をすることで、生物学的な仕組みを理解する手がかりを得ようとする、多くの先行研究が存在する[Henry 2011, Nichols 2011, Karr 2012] 。

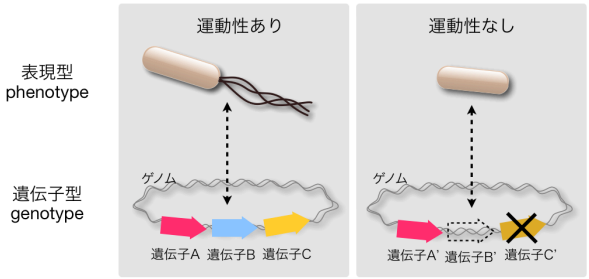

図 1 遺伝子型と表現型の関係

現在、ゲノム配列を解読する技術の進歩により、生物個体の遺伝子型の情報(ゲノム情報)を解読するコストが低下したために、遺伝子型情報については、容易に入手する事ができるようになった [Metzker 2010]。その一方で、微生物の表現型の情報は、過去一世紀以上の膨大な蓄積があるにもかかわらず、主として文献に記載されている状態であり、電子化が進んでおらず、また一貫した表記方法で記載されていないこと等から、これまで容易に利用することが出来なかった。そこで我々は、既存の微生物の表現型データの代表例としてゲノムプロジェクトデータベース Genomes OnLine Database (GOLD) [Pagani 2012] のデータを LOD として整備することを始めた。これにより、分散した遺伝子型データと表現型データに対して、横断的な検索が行えるようになり、また新しい解析アプリケーションを作成することが可能になることを示した。

3. 微生物表現型オントロジー

3.1 データの選定

本研究で対象とする微生物表現型のデータは、様々な書籍や文献、データベースに記述があり、多岐にわたっている。本研究では、ゲノム配列情報と表現型の情報をリンクするための有効性を考慮して GOLD データベース の微生物データを選定した。GOLD[*4] は、ゲノム配列情報が利用可能な生物に関して、多岐にわたるゲノム関連情報が記載されたデータベースである。それら情報の中には、多くの表現型情報も含まれており、電子的に利用できる情報としては、特に網羅性が高い。GOLD に記載されている表現型関連の項目は、細胞の形、細胞の配置、酸素要求性(特定の微生物が好気的または嫌気的か)、運動性、pH、生育温度(生育に適した温度の上限および下限)、生育温度嗜好性(通常の温度が生育に適しているのか、好熱菌または好冷菌か)、塩濃度嗜好性(塩濃度が高い環境で増殖できるか否か)、栄養的分類、色、グラム染色性、血清型、病原性情報、生育環境、各種化合物の代謝能力等、多岐に渡る。このうち、生育環境および病原性については、それぞれ多くの語彙が存在し、別プロジェクトで個別のオントロジーが構築されているために、今回の対象からははずした。さらに、現状では、表記ゆれ等のない項目に限って語彙のオントロジー化、および LOD 化を行った。

3.2 微生物表現型オントロジー

前述のような観点から、微生物表現型オントロジー(MPO: Microbes Phenotype Ontology)の構築を開始した。

MPO を構築するにあたって、我々は、ゲノム情報 LOD につなげることができる微生物表現型の LOD を速やかに作成することができるかどうか、ということを重視した。すなわち、オントロジーを構築するにあたって、対象分野における概念に対し、適切な統制語彙を割り当てていくことになるが、その際、すでに電子化されているデータで使われている語彙を採用することにした。例えば、微生物表現型には、細胞の配置(Cell arrangement)という概念がある。これは細菌の形態を表す特徴のなかで、細胞個体どうしの配置のパターンを分類したものである。例えば、球形の細菌を球菌(coccus)、細長い棒状の細菌を桿菌(bacillus)と呼ぶが、これが 2 つつながった配置は、それぞれ Diplococcus および Diplobacillus と呼ばれる。これらは学術用語として定義されている語彙であるが、GOLD のデータでは、もう少しカジュアルに、2 細胞がつながったことのみ注目して、Pair との記述がされている。そこで、この場合、GOLD から容易に LOD を生成できるように、Pair という語彙を MPO に採用し、さらにその下位に Diplococcus と Diplobacillus を配置することで、検索等の際に、それら学術的な語彙も利用可能となるように配慮した(図 2)。

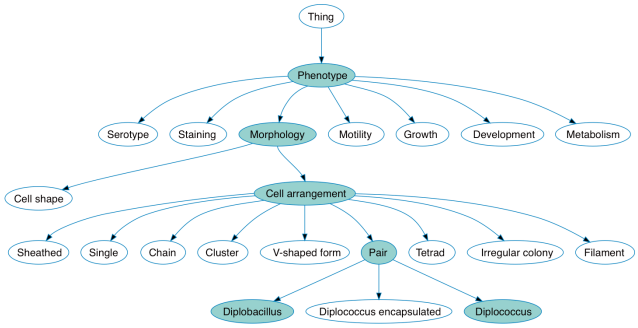

図 2. 微生物表現型オントロジーの一部

このようなポリシーで、GOLD データベースから微生物表現型の LOD を構築するのに、必要性の高い語彙を中心に、現在 158 語彙を微生物表現型オントロジーとして構築した。オントロ

*4 http://www.genomesonline.org/

*5 http://bioportal.bioontology.org/

ジー構築には、OWL エディターである TopBraid ソフトウェアを利用した。構築したオントロジーは、随時オントロジーのポータルサイトである、BioPortal[*5] [Whetzel PL 2011] への登録を行なっている。

各クラスとプロパティには、MPO_00001 といった機械的な ID をふり、rdfs:label に、人間が読むための記述を定義した。これにより、ID さえ変更しなければ、一度このオントロジーを利用して LOD を構築したあとに、ラベルの記述を変更しても LOD 自体は変更する必要がない。また、ラベルには、使用言語の指定を行い、できるだけ日本語でのラベル記述を行った。英語のラベルは必須であるが、これにより、将来、アプリケーションの日本語化もやりやすくなると考えている。

3.3 微生物表現型 LOD の構築

構築したオントロジーを用いて、主語を生物種のタクソノミーID、目的語を各種表現型とした LOD の生成を行った。表現型情報については、GOLD データベースの Organism Metadata セクションに記述されている表現型情報のうち、3.1 に記述した項目について LOD 化を行った。その結果、107,933 トリプルを生成した。表 1 に、表現型の項目別にトリプル数を示す。

表 1　構築した LOD の微生物表現型別トリプル数

表現型	トリプル数
細胞の形	18,036
細胞の配列	14,724
酸素要求性	17,946
運動性	16,782
pH	2,106
生育温度嗜好性	20,106
生育温度	9,882
塩濃度嗜好性	660
栄養的分類	7,686

次に、海洋性細菌 *Rhodopirellula baltica* SH28 (Taxonomy ID 993517) の細胞の形を MPO を用いて表現した LOD の例を示す（図 3）。細胞の形を英語表記の”Rod”@en だけでなく、日本語表記の”桿菌@ja”としても記述している。また、知識体系を共有、リンク付けするための一般的なデータ・モデルである SKOS（Simple Knowledge Organization System）オントロジーを使って、Rod の別表現である Rod-shaped と Bacillus も記載している。

```
<http://www.ncbi.nlm.nih.gov/taxonomy/993517> mpo:MPO_10001 mpo:MPO_01015 .
mpo:MPO_01015
    rdf:type owl:Class ;
    rdfs:label "桿菌"@ja, "Rod"@en ;
    skos:altLabel "Rod-shaped"@en, "Bacillus"@en .
mpo:MPO_10001
    rdfs:type owl:ObjectProperty ;
    rdfs:label "cell shape"@en ;
    rdfs:range mpo:MPO_01001 .
```

図 3 微生物表現型 LOD の例

4. 考察

4.1 アプリケーション例

今回構築した LOD により、DBCLS ですでに構築しているゲノム情報 LOD および UniProt が提供するタンパク質情報 LOD を、微生物表現型の観点から、検索できるようになった。現在、我々はこれらの LOD を活用した、様々な解析アプリケーションの開発を進めている。

図 4 は、そのようなアプリケーションの一例である。この散布図でプロットされた各点は、それぞれ別の微生物を表しており、横軸は各細菌のゲノムにコードされた遺伝子の数、縦軸は各細菌のゲノムのサイズを表している。縦軸と横軸は、スイッチによって他のゲノム統計情報や生育温度、pH の値に切り替えることができる。またデータカテゴリーメニューから、今回作成した表現型 LOD を使って個々の細菌を色分け表示することが可能である。図 4 の例では、各微生物を表すスポットが生育温度嗜好性により色分けされている。例えば、赤で表された超好熱菌（至適生育温度が 80℃以上の細菌）は、全体的な傾向としてゲノムサイズ、および遺伝子数が小さい傾向にあることが読み取れる。

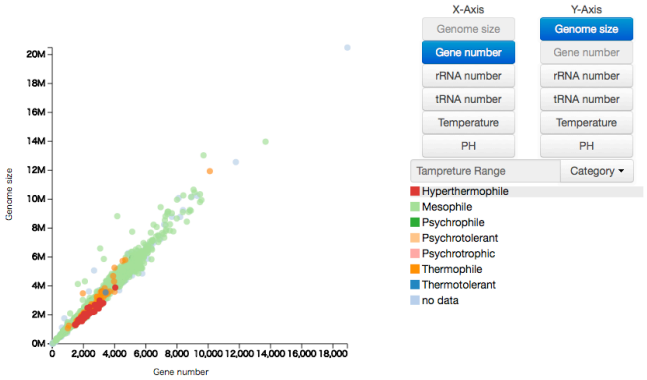

図 4　微生物表現型 LOD を利用したアプリケーションの例

4.2 微生物における表現型と遺伝子型の関連

近年、新型シーケンサー等の各種技術革新により、大量の遺伝子型情報が得られるようになり、ゲノム規模の遺伝子型情報からボトムアップに表現型と対応付ける試みが活発に行われている。微生物の表現型には、そのしくみが、ひとつ～少数の遺伝子で説明できるものもあり、そのような場合には、遺伝子型側からのボトムアップのアプローチが、表現型の理解に有効である。また、それ以上の複数の遺伝子が関与する表現型については、代謝パスウェイなど、既に遺伝子ネットワークとして生命システム情報が整理されているデータベースを利用して遺伝子機能別のグルーピングを行なうことで、表現型の理解がすすむ。一方で、今回我々が対象としているマクロなレベルの表現型情報は、複雑な作用機序の結果として観察されており、すぐに遺伝子レベルの情報と関連付付ける事は困難である。しかし、最終的にはこれらの表現型をも遺伝子型情報から理解することが、生物学の大きな目標でもある。このように、遺伝子型と表現形を結びつけて微生物の包括的な理解をするためには、遺伝子型からのボトムアップと、表現型からのトップダウンの両方の手法が欠かせないと考えられている。

今回報告したような、マクロなレベルの表現型を LOD 化することで、ゲノム情報の LOD と統合して、今までにない視座での検索が容易になり、表現型情報から遺伝子型情報に迫る、トップダウン的な解析の一助となることを期待している。

5. まとめ

本稿では、微生物の表現型情報に関して、ゲノム情報 LOD と統合した利用を想定した、微生物表現型 LOD の構築について報告した。表現型のデータとしては、GOLD データベースに記載された情報を利用した。また、LOD 化に必要なオントロジ

ーを新規に構築した。さらに、ゲノム情報 LOD と表現型 LOD を活用したアプリケーションの一例を示した。今後、教科書や、論文等の、GOLD 以外の情報源から表現型情報を収集することで、さらに幅広い解析を可能とする表現型 LOD の構築を目指していく。

謝辞

本研究は、DBCLS のメンバーおよび、JST NBDC の統合化推進プログラム「ゲノム・メタゲノム情報を基盤とした微生物 DB の統合」(代表 黒川顕東工大教授)メンバーの方との共同作業のなかでなされたものです。関係者の皆様に感謝の意を表します。

参考文献

[山口 2010] 山口敦子, 片山俊明: 我が国のデータベース構築・統合戦略(第 2 回) データベースを統合利用するための基盤としてのセマンティックウェブ技術, 細胞工学 Vol. 30, No.11, pp. 1210-1215, 学研メディカル秀潤社, 2011.

[Eilbeck 2005] Eilbeck, K. et al.: The Sequence Ontology: a tool for the unification of genome annotations, Genome Biology, Vol.6, R44, 2005.

[Fernández-Suárez 2013] Fernández-Suárez, X.M. and Galperin, M.Y.: The 2013 Nucleic Acids Research Database Issue and the online molecular biology database collection, Nucleic Acids Research Vol. 41(Database issue), pp D1-D7, 2013.

[Henry 2011] Henry, C.S. *et al*.: Connecting genotype to phenotype in the era of high-throughput sequencing , Biochimica et Biophisica Acta Vol. 1810, pp. 967-977, 2011.

[Karr 2012] Karr, J.R. *et al*.: A whole-cell computational model predicts phenotype from genotype, Cell Vol. 150, pp. 389-401, 2012.

[Metzker 2010] Metzker M.L.: Sequencing technologies – the next generation. Nature Review Genetics Vol. 11, pp 31-46, 2010.

[Nichols 2011] Nichols, R.J. et al.: Phenotypic Landscape of a Bacterial Cell, Cell Vol.144, pp. 143-156, 2011.

[Pagani 2012] Pagani, I. *et al*.: The Genomes OnLine Database (GOLD) v.4: status of genomic and metagenomic projects and their associated metadata, Nucleic Acids Research Vol. 40(Database issue), pp D571-579, 2012.

[Stanier 1986] Stanier, R.Y. *et al.*: The Microbial World Fifth edition, Prentice Hall College Div, 1986.

[The UniProt Consortium 2013] The UniProt Consortium: Update on activities at the Universal Protein Resource (UniProt) in 2013, Nucleic Acids Research Vol. 41(Database issue), pp D43-D47, 2013.

[Whetzel PL 2011] Whetzel, P.L. *et al*.: BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications, Nucleic Acids Research, Vol.39(Web server issue), pp 541-545, 2011.