1G3-3

セマンティック Web 技術を用いた、生物表現型統合データベース

The integrated phenome database with semantic web technology

桝屋 啓志[*1] Hiroshi Masuya　高月 照江[*1] Terue Takatsuki　斎藤 実香子[*1] Mikako Saito　高山 英起[*1] Eiki Takayama　吉田 有子[*2] Yuko Yoshida

蒔田 由布子[*3] Yuko Makita　望月 芳樹[*4] Yoshiki Mochizuki　土井 考爾[*2] Koji Doi　小林 紀郎[*2] Norio Kobayashi　豊田 哲郎[*2] Tetsuro Toyoda

[*1] 理化学研究所 バイオリソースセンター RIKEN BioResource Center

[*2] 理化学研究所 情報基盤センター Advanced Center for Computing and Communication, RIKEN

[*3] 理化学研究所 環境資源科学研究センター RIKEN Center for Sustainable Resource Science

[*4] 理化学研究所 統合生命医科学センター RIKEN Center for Integrative Medical Sciences

The phenome is the set of phenotypes correlated each other through common molecular pathways underlying, evolutional relationships of molecules or biological interaction between environments. Understanding of phenome or individual phenotypes is one of the fundamental issues to promote innovation in the biomedical science. We have tried to integrate variety of phenotype data using semantic web technologies and RDF data format. We designed scheme of phenotypic data based on the study of upper ontology, and also designed interrelationships among different data, gene, organisms and molecules. Using these linked data, we developed applications to propose “recommendations” of biological materials, which show similar phenotype to selected material by users. In this paper, we report application of semantic web technology for improvement of access to biological resources.

1. はじめに

Phenotype（表現型）とは遺伝因子と環境因子の相互作用によって現れる表現形質の変化として定義される。近年、このような表現型の総体を示す概念として「フェノーム」が使われるようになっている、これは、生物の表現型が個々に独立して現れるのではなく、相互に深く関係している事を意味している。例えば、1個体の生物が示す様々な表現型は、その基盤となる分子メカニズムを通して互いに関係しており、生物種間では、分子の進化的な関係を通して、互いに関連している。さらに、生態系においては、生物同士の相互作用や、環境応答を通して関係している。

遺伝子の多型が薬物応答の個人差を生み出すように、遺伝因子と環境因子の組み合わせで表現型の違いを理解することは今後のライフサイエンスの重要課題である。また、近年ニーズが高まっている生命原理応用によるイノベーションを加速するには、様々な生物で得られた研究成果情報を、種の違いを踏まえながら統合する必要がある。

従来、生物種横断的な情報統合は分子レベルでは広く行われてきたが、より高度な情報である表現型に関しては、標準化や統合が困難なために公開されても各生物種の研究コミュニティ内で限定的にしか用いられず、新たなイノベーションの障壁となっている。そこで我々は、表現型とその関連データの研究分野の垣根を越えた情報共有を、分子データと連携可能なかたちで実現させることを最終目的として、セマンティック Web 技術を用いた生物フェノームの統合データベースの構築に取り組んでいる。本報告では、RDF に基づいた表現型の基本スキーマの作成と、RDF のデータリンクを用いた、表現型情報利用アプリケーションについて紹介する。

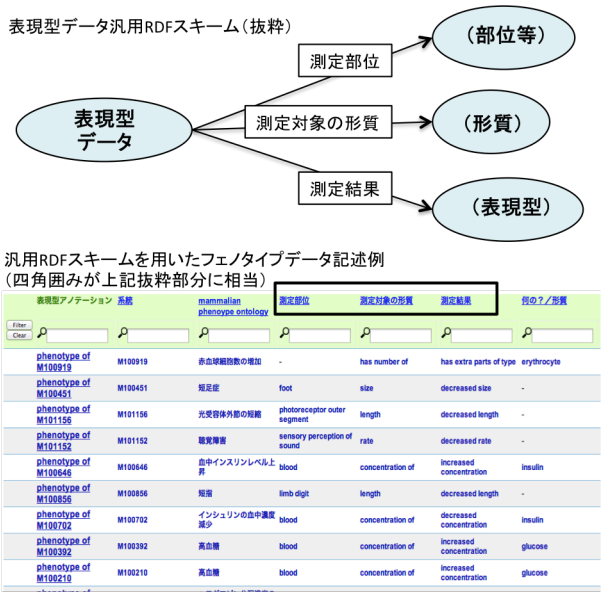
連絡先：桝屋啓志，理化学研究所バイオリソースセンター，茨城県つくば高野台 3-1-1，hmasuya@brc.riken.jp

2. データおよびアプリケーションの概要

本研究では、OWL に準じた RDF スキーマの編集および、RDF インスタンスデータの管理、データの可視化の統合環境として、サイネス[SciNets, 桝屋 2010-2]を使用している。

3. 表現型の RDF スキーマデザインとデータの作成

まず、RDF 技術を用いた表現型の体系的なデータ記述を行うために、生物の表現型を網羅的に記述できる汎用フォーマットのデザインを行なった。RDF は、「主語」「述語」「目的語」の3つ組を基本構造としているが、表現型データを主語として、表現型を示す主体である測定部位、生物学的特徴の分類である測定対象の形質、各形質において生物が具体的に示す特徴（定性値あるいは定量値）である測定結果という3種類の述語を持ち、それぞれの目的語に、国内外で使用されているものに共通のオントロジー、またはデータベース値を代入するよう設計した（図1）。これは、我々が以前、上位オントロジーYAMATO を基盤に提唱した、表現型データ概念とほぼ同型である [桝屋 2010-1, 2011, 2013]。ただし、現状のデータ表示のインターフェースを考慮して、RDF および OWL のデータ形式では、オブジェクト間のリンク構造が複雑になってしまうロール概念を省略している。



表現型アノテーション 系統	mammalian phenotype ontology	測定部位	測定対象の形質	測定結果	何の？／形質
phenotype of M100919	M100919	赤血球細胞数の増加	-	has number of	has extra parts of type erythrocyte
phenotype of M100451	M100451	短足症	foot	size	decreased size -
phenotype of M101156	M101156	光受容体外節の短縮	photoreceptor outer segment	length	decreased length -
phenotype of M101152	M101152	聴覚障害	sensory perception of sound	rate	decreased rate -
phenotype of M100646	M100646	血中インスリンレベル上昇	blood	concentration of	increased concentration insulin
phenotype of M100856	M100856	短指	limb digit	length	decreased length -
phenotype of M100702	M100702	インシュリンの血中濃度減少	blood	concentration of	decreased concentration insulin
phenotype of M100392	M100392	高血糖	blood	concentration of	increased concentration glucose
phenotype of M100210	M100210	高血糖	blood	concentration of	increased concentration glucose

図 1　フェノタイプを記述する汎用 RDF スキーマ

次に、このフォーマットに準拠して、国内でオープンに利用できる研究材料（マウス、培養細胞、微生物）および、これらのバイオリソースの表現型情報を、元のデータベース内の自然言語記述、既存のアノテーション情報を通じて収集した。マウス系統約 5000 件、細胞株約 3600 件、微生物株約 15000 件に加え、表現型データ約 18000 件を上記 RDF フォーマットに落とし込むとともに、月ごとに最新情報に更新できる体制とした。表現型データの測定対象の形質および、測定結果のプロパティの目的語となるオントロジーは Phenotypic Quality オントロジー(PATO) [Gkoutos 05]の語彙を用いた。また、マウスでは、哺乳類特有の表現型を示すオントロジー、Mammalian Phenotype Ontology (MP) [MP]が提供されているため、これとのマッピングも示すようにした。

4. 表現型情報利用のためのインターフェイス

上記のフォーマットにより記述されたデータは、図1に示すような表形式、およびカード型のインターフェースによって示されるようにした。表の各カラムに付加されたフィルター機能により、表現型が現れた部位、形質などによる簡易な絞り込みができる。

図 2　表現型類似性を表示するインターフェース

さらに、体系立った RDF のリンクを用いて、関連する表現型の情報を自動検索するインターフェース（お勧め機能）も提供される。これを用いて、実験材料を検索する生物学者が、Amazon オンラインショップを操作するような感覚で、関連する表現型を示す材料を次々と閲覧し、自分の研究にふさわしい材料を選びやすくするようにした（図2）。

5. 考察と展望

生命科学は個別知識の集大成であり、膨大な知識のネットワークで構築されている。特に現代では、生命科学知識は膨大化の一途を辿っており、先鋭化した専門知識のみでは、生命現象を解き明かすことは困難となっている。研究者は、深い専門知識と同時に、常に専門外の知識について、出来うる限り広く深い情報収集を行なうことが求められており、それを可能にする情報インフラの構築が、バイオインフォマティクスにおける大きな課題となっている。本研究では、研究分野の垣根を超えた表現型情報の共有を目指し、表現型に普遍的な RDF データ形式を設計し、それを用いた横断的なデータベース作製を試みるとともに、RDF のリンクを生かした表現型情報利用のインターフェース開発を行なった。

フォーマットに関しては、表現型データの多様性が、3つのフィールドそれぞれのオントロジーの分類体系によって吸収されるため、自由度の高い記述が可能であり、様々な生物の多様な表現型情報を格納することができた。また、従来の MP オントロジーでは難しかった、部位や形質による検索を容易に実現する事ができた。この機能を用いることにより、例えば、マウスの骨格形成において、上腕骨や大腿骨等の長骨において、長軸方向の成長に異常をきたす突然変異と、径を太くする方向の成長が異常になる突然変異とをそれぞれ区別して整理することができる。これは、それぞれの変異の原因遺伝子の機能を推測することにつながる。

また、図2に示す表現型類似性を示すインターフェースは、表現型情報を通して、研究者に複数の可能性を同時に提示して、新たな「気付き」を誘導するところに特徴がある。例えば、研究材料としてのノックアウトマウスの選択では、検索にヒットしたマウスと同時に、表現型の観点から他の候補マウスを提示する事で、研究に最もふさわしい材料の選択において、ユーザー自身が「新たな可能性」について、比較検討することができる。今後、オントロジーの is_a 関係を用いた推移的推論を加味して類似性を処理する機能を実装することで、研究者が用いる表現型概念に即した分類関係を示す事が可能になると考えられる。

今後対象とする生物種を拡張することで、ヒトの疾患研究のモデルとして活用できる実験動物を検索するなど、より実用に即したデータベースの構築が期待される。我々は、ヒト疾患と、マウス・ラットの表現型の関係性についても、オントロジーによる関連データを作成しつつあるため[桝屋 2013]、これを利用する事も可能である。

一方で、微生物と哺乳類の表現型情報は、同じ形式で記述できるものの、注意が必要な点も示唆された。哺乳類表現型は、一般に、種内の遺伝的バリエーションである遺伝子型と対応すると考えられ、マウスの高血糖、ラットの高血糖といった表現型は、ヒトの糖尿病に相当するもの考えられる。ここでは、血糖という形質の「高濃度」という定性値（順序尺度）は、暗黙的に「正常よりも高濃度である」という意味が含まれる。これに対して、ほとんどの微生物における表現型は、生物種間の区別に用いられ、「正常に対して」という意味合いを含まない。

今回の微生物の表現型データには、比較対照を必要とする順序尺度は含まれず、名義尺度のみであった。しかしながら、微生物においても実験用大腸菌など、遺伝子操作が可能なモデル生物として用いられるものでは、哺乳類と同様に「正常と比較して」という意味合いの表現型が用いられる。今後、研究分野や、生物種に横断的な表現型情報統合に向けて本データベースを拡張するためには、ロール概念を用いたコンテキスト依存定性値の統合[桝屋 2013]や、それによって複雑化するデータ形式のビュワーなど、さらなる工夫が必要になると考えられる。

参考文献

[BioLOD] http://biolod.org

[Gkoutos 05] Gkoutos GV, Green EC, Mallon AM, Hancock JM, Davidson D: Using ontologies to describe mouse phenotypes, Genome Biol, 6, R8. (2005)

[MP] ftp://ftp.informatics.jax.org/pub/reports/MPheno_OBO.ontology

[SciNets] http://database.riken.jp/

[太田 2011] 太田 衛, 古崎 晃司, 溝口 理一郎:実践的なオントロジー開発に向けたオントロジー構築・利用環境「法造」□拡張 ― 理論編 ―人工知能学会論文誌, Vol.26 No.2,pp.387-402, (2011)

[桝屋 2010-1] 桝屋啓志, 田中信彦, 脇和規, 櫛田達矢, 古崎晃司, 溝口 理一郎:上位オントロジーに基づく生物表現型 データ記述の考察, 第 24 回人工知能学会全国大会予稿集, 1B5-4 (2010)

[桝屋 2010-2] Masuya H., Makita Y., Kobayashi N., Nishikata K., Yoshida Y., Mochizuki Y., Doi K., Takatsuki T., Waki K., Tanaka N., Ishii M., Matsushima A., Takahashi S., Hijikata A., Kozaki K., Furuichi T., Kawaji H., Wakana S., Nakamura Y., Yoshiki A., Murata T., Fukami-Kobayashi K., Mohan S., Ohara O., Hayashizaki Y., Mizoguchi R., Obata Y., Toyoda T.: The RIKEN integrated database of mammals, Nucleic Acids Res. 39, D861-D870, (2010).

[桝屋 2011] Masuya H., Gkoutos G.V., Tanaka N, Waki K, Okuda Y, Kushida T., Kobayashi N, Doi K, Kozaki K, Hoehndorf R., Wakana S, Toyoda T., and Mizoguchi R.: An Advanced Strategy for Integration of Biological Measurement Data, Proc. of 2nd International Conference on Biomedical Ontology (ICBO2011) ,pp.79-86 (2011).

[桝屋 2013] 桝屋啓志,古崎晃司,大江 和彦,溝口理一郎コンテキストに依存した定性値を扱う生物表現型統合データベースの試作, 第 27 回人工知能学会全国大会予稿集, 3I1-2 (2013)