1G5-4

生物表現型情報と、疾患情報をつなげるデータベース

Development of the database to show relationships between biological phenotypes and diseases

桝屋 啓志[*1] Hiroshi Masuya
高山 英紀[1] Eiki Takayama
古崎 晃司[2] Kouji Kozaki
今井 健[3] Ken Imai
大江 和彦[3] Kazuhiko Ohe
溝口 理一郎[4] Riichiro Muzoguchi

[*1] 理研バイオリソースセンター RIKEN BioResource Center
[*2] 大阪大学産業科学研究所 The Institute of Scientific and Industrial Research (ISIR), Osaka University
[*3] 東京大学大学院医学系研究科 Graduate School of Medial and Faculty of Medicine, The University of Tokyo
[*4] 北陸先端科学技術大学院大学 Japan Advanced Institute of Science and Technology

For the data-driven science in the biomedical study field, ontology-based data description and visualization of biological measurement data capturing the phenotypes of organisms represent a broad range of variations is one of the most important issues. With the aim of integrating measurement data across various biological experiments, we developed a Web-based database fully based on an upper ontology, Yet Another More Advanced Top-Level Ontology. In this database, all the metadata was described directly on the ontology. A software application parsed the ontology to represent the measurement data in a spreadsheet style and provided functions for the conversion of qualitative data into quantitative data represents higher abnormal values, normal values and lower abnormal values. Furthermore, the application enables retrieval of related disease defined by Clinical Medical Ontology. This study provided a concrete example of algorisms to show phenotype-disease association via relationship between measurements data and small abnormal state of diseases, and context-dependent visualization of graph data described in the top-level ontology-based database.

1. はじめに

現代の生命科学は、生物を分子の部品で構成された複雑なシステムと捉えている。そのシステムの設計図にあたる遺伝子の塩基配列情報を、コンピュータを通じて研究コミュニティ全体で公開／共有することにより、生命科学は飛躍的に発展してきた.

しかし、今後の生命システム全体の理解のためには、設計図だけでなく、設計図によって構築された生物体の構造や機能、表現型など、さらに高いレベルの情報を知識基盤として共有していく必要がある。

表現型（Phenotype）とは、生物が遺伝因子や環境因子の結果として示す形質、あるいはその特性である。多くの生命科学研究において、表現型は、観察や実験の「結果」として記述される。このような結果を広く共有するための技術を確立することは、生命科学の情報知識基盤を形成する上で、極めて重要な課題である。

バイオインフォマティクス分野では、生物種に特有な表現型オントロジーが多く作成されており、これらによってアノテーションされた多くの情報がある。さらに、これらのオントロジー語彙を、生物種特有な語彙ではなく一般的な性質語彙として作成された Phenotypic Quality オントロジー(PATO) [Gkoutos 05]へマッピングすることで、生物種横断的な表現型の相同性（ヒトの疾患である頭蓋骨癒合症と、マウスの表現型である縫合線の閉塞の相同性など）を推論することが可能となっている[Hoehndorf 2011], [Köhler 2013]。

しかしながら、表現型データの共有や利活用という視点では、解決すべき課題が残っている。例えば、表現型データには、多くの「測定データ」が含まれ、この測定のコンテキストが極めて多様であることが挙げられる。まず、数値データが生命の特性として扱われるためには、「大きい／小さい」「高い／低い」などの定性値化が必要となるが、一般に、定性値には様々なコンテキストがある。コレステロール値の測定を例にとると、1)1 匹のマウスのコレステロール値の経時的な変化として、2つの時点を比較して片方が高い、2) 単にラット個体 X とマウス個体 Y のコレステロール値を絶対的に比較した場合に、マウス Y の方が、コレステロール値が高い、あるいは、3)マウスやラットそれぞれについて、「正常」と見立てたコントロールと比較した際にコレステロール値が高い、など、コンテキストに従って、定性値は異なる意味を持つ[桝屋 2010, 2011]。

また、測定データは、しばしば表現型や疾患が示す性質のひとつでしかない。例えば、I 型糖尿病にとって、「血糖値が高い」ことは、メインの病態の1つではあるが、疾患全体を示す概念ではない。疾患モデルとして用いられる動物の表現型データと、ヒトの疾患の情報をつなげて広く利活用するためには、このようなギャップを解決する必要がある。

我々は、生物の表現型情報を、セマンティックウェブ上でより効果的に利活用できるようにするための、オントロジーの基盤技術について研究している。以前の研究において、上位オントロジー Yet Another More Advanced Top-level Ontology (YAMATO) [溝口 2009, YAMATO]のフレームワークを用いて、PATO の概念を、コンテキスト依存の定性値として再定義し

連絡先：桝屋啓志，理化学研究所バイオリソースセンター，茨城県つくば市高野台 3-1-1，029-836-9018，

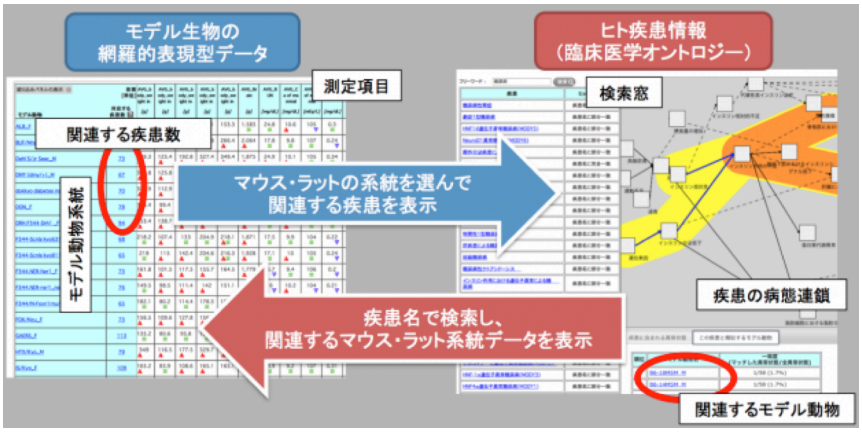

図 1　Web アプリケーションの画面構成

た参照オントロジー、PATO2YAMATO を作成し [桝屋　2010, 2011]、さらに、このオントロジーに基づいて、多様な測定データを一貫した形式で記述し、コンテキストに従って、生物種としての正常／異常に基づく定性値化を行うプログラムを試作した。また、同じく YAMATO のフレームワークに準拠して作成された臨床医学オントロジー[大江 2009, 溝口 2011]を用い、双方での定性値記述の同等性を定義することで、マウス、ラットの表現型解析の測定値データから、その表現型を、病態として含む疾患をリストアップすることも可能にした[桝屋　2013]。

本研究では、これらの研究に基づいて、哺乳類の測定データを統合して表示し、かつ関連するヒト疾患を推論して示す、実用的な Web アプリケーションの開発を行った。

2.　データの概要

本研究では、表現型データとして、マウス、ラットにおける下記の2つの公開データベース 1)京都大学・NBRP ラットデータベース[NBRP ラット]の、雄 173 系統 雌 44 系統。2) 国立遺伝学研究所・マウス表現型データベース[NIG Mouse DB]の、雄 30 系統、雌 29 系統の表現型データ（合計 8171 データポイント）を使用した。また、表現型と関連させる疾患オントロジーデータとして、臨床医学オントロジー[大江 2009, 溝口 2011]および、これを RDF 化して公開している疾患連鎖 LOD[疾患連鎖 LOD]を使用した。概要を以下に示す。

2.1 表現型データ

全ての表現型データは、PATO2YAMATO における概念定義に従って、法造オントロジーファイルに直接記述した。詳細は、[桝屋　2013] に示した通りである。概要としては測定対象である実体（Entity:マウスやラットの部位や器官組織）、性質タイプ（Attribute: 各計測のパラメータである形質）、値（Value: 測定値）を、オントロジー概念のインスタンスとして記述し、さらに、実体が特定の性質を持つ事を記述する Entity-Attribute-Value3つ組形式の表現のインスタンスとして個々のデータを記述した。

2.2 疾患−表現型マッピングデータ

疾患モデル生物であるマウスやラットでは、「正常に比べて異常に高い／低い値」すなわち異常値は、疾患の症状と直結して考えられる。例えば、血糖値の異常は、比較対照が正常と見なせる集団であれば、人間の血液検査の血糖値異常とほぼ同等に扱われる。従って、本研究では、生物コンテキストの下で「異常」と判断される値と、臨床医学オントロジーにおける異常状態について、同等性を示す対応を法造ファイル内で定義した。

2.3 疾患情報

マウス／ラットの表現型と対比させるための疾患情報として、臨床医学オントロジーを用いた。このオントロジーは、1)疾患に含まれる部分病態が「異常状態」として、その連鎖と共に詳しく記述されている。2) YAMATO のフレームワークに準拠して構築されおり、疾患の要素である「異常状態」は、PATO2YAMATO で用いられている「定性値」概念との相互関係を明確に示すことができる[溝口　2009, 山縣　2012]。本研究

図 2　「モデル動物から調べる」画面と、絞り込みパネル

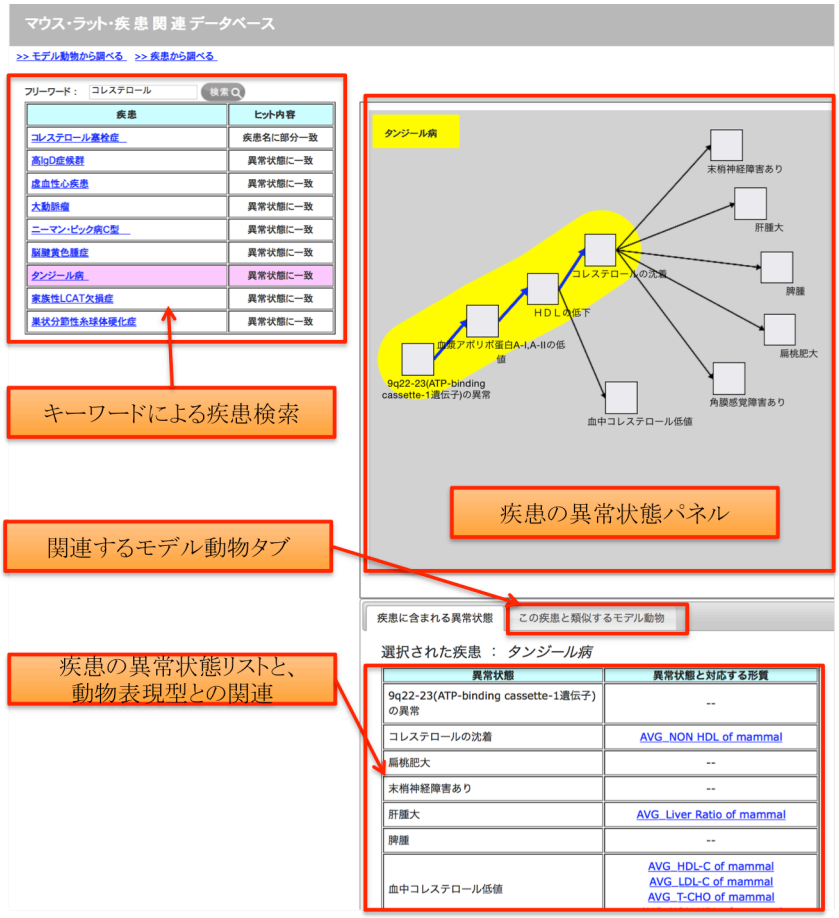

図 3　「疾患から調べる」画面

では、疾患データ自体はシステム内に持たず、上記の疾患ー表現型マッピングデータを介して、すでに公開されている疾患 LOD を参照するようにした。

3. アプリケーションの概要

本ソフトウェアプログラムは、上記のデータを閲覧できる Web アプリケーションとして、Java 7 および、Apache Tomcat 7.0 を用いて開発した。なお、このアプリケーションおよび上記データを用いて作成したデータベースは、http://kb2.riken.jp にて公開している。

3.1「モデル動物から調べる」機能

(1) 性質データの表形式表示

本アプリケーションは、「モデル動物から調べる」と、「疾患から調べる」の2つの画面から利用できる（図 1）。「モデル動物から調べる」画面（図 2）では、オントロジーとして定義された EAV 形式の測定データを、表形式で表示する。この表では、水平方向に性質タイプ、垂直方向に測定対象の動物グループが示される。本アプリケーションでは、哺乳類（本データベースではマウスとラットのみ）の表現型をひとつの表インターフェースで表示することを前提としたので、それに従い、水平方向に並ぶ性質タイプは、哺乳類のレベルで一般化したものが並んでいる。例えば、マウスという生物における血中コレステロール量と、ラットという別の生物におけるコレステロール量は、判定基準が異なる別の性質タイプであるが、一般化すれば、哺乳類の血中コレステロール量として同一視可能である。本研究のデータでは、上位概念である「哺乳類の血中コレステロール量」が定義してある[桝屋 2013]ので、表のコンテキストに従って哺乳類レベルの性質タイプを列挙し、マウスとラットの血中コレステロール量を同一視して1列のカラムに収めて表示する。

また、表示する値は、定量値である数値データを示すとともに、[桝屋　2013]で報告した通り、定量値から定性値の変換機能によって、性質タイプを共有する定性値の定義を参照して、閾値との比較を行い、定性値へと変換した値を表示する。

(2) 各種絞り込み機能

データ形式が包括的なオントロジーであることを利用して、上記表データは絞り込みパネル（図 1）のツリーを辿るかたちで絞り込みが可能である。

1) 部位の絞り込み：測定対象である実体（Entity）の部位を部位オントロジーで絞り込む

2)生物種の絞り込み： 測定対象である実体（Entity）の部位がどの生物種かにより、生物種オントロジーから絞り込む

3) 値の種類で絞り込み： 値（Value）のオントロジーからの絞り込み。定量値、定性値、順序値等を絞り込める。

4) 順序尺度のコンテキストの絞り込み：上述のように、 値（Value）に対して定義したコンテキストのツリーで絞り込むことができる。

(3) 一般化（上位概念への遷移）による推論機能

実験動物の活用のひとつとして、ヒトでは実験解析が難しい疾患メカニズムを解析するためのモデル系として用いることが挙げられる。このような疾患モデル動物は、ヒト疾患と相同と”考えられる“表現型を示す動物が用いられる。モデルとして成立するかどうかは解析による検証が必要だが、情報技術への期待としては、疾患モデルとして利用出来る可能性のある生物を、データから提示し、気づきを与えることが挙げられる。

我々は、[桝屋　2013]で報告した性質タイプの統合アルゴリズム、疾患との関連性推論機能を利用して、哺乳動物表現型データから、関連性のある疾患を提示する機能を実装した。上記の通り、表インターフェースでは、マウス、ラットの各表現型を「哺乳類としての異常値／正常値」として定性値化して示している。この定性値は、ヒトを対象として特殊化すると、ヒトとしての異常値／正常値すなわち、臨床医学オントロジーで言う所の、異常状態に容易にマッピング可能なデータとなる。本データベースでは、哺乳類の異常値を、上記の方法で、臨床医学オントロジーの異常状態とマップした上で、1系統の動物が示す異常値が、異常状態としていくつの疾患に含まれるかを計算する。

(4) 疾患閲覧画面へのジャンプ

上記の疾患数をクリックすると、その動物が示す異常値と同等の異常状態を含む疾患のリストが表示される（図3）。この画面では、疾患連鎖 LOD の API を用いて、それぞれの疾患がどのような異常状態の連鎖を持つか、および、どの異常状態が、そ

の動物の異常値と同等であるかを閲覧できる。また、ひとつの疾患を選んだ時点で、その疾患と関連する異常値を示す動物のリストも表示される。

3.2 「疾患から調べる」機能

上記の疾患閲覧画面は、疾患連鎖 LOD の API を用いて、キーワードによる疾患検索なども行える。また、疾患と関連する異常値を示す動物のリストをクリックすることで、動物表現型の表インターフェースが開き、該当する動物のデータのみを絞り込んで表示する。この画面を最初に開くことで、上記とは逆方向に画面が遷移し、疾患から、関連する動物を調べるという逆方向の検索が行える(図 1)。

4. 考察と今後の展望

4.1 他の表現型データベースとの違い

PATO を基盤にして、生物横断的に表現型の関連性を示すシステムには、PhenomeNet [Hoehndorf 2011]、UberPheno [Köhler 2013]等が知られている。これらのシステムと比べて、我々のシステムには、以下のような特徴がある。1) 性質値の違いを扱える。PATO2YAMATO オントロジーでは、PATO の性質一階層モデル分類を、YAMATO の性質タイプ／値の二階層分類にマッピングすることで、値の違い、すなわち、いわゆる定量値である比例尺度、間隔尺度、定性値である順序尺度、名義尺度(カテゴリ値)の違いを扱うことができる(尺度水準 [Stevens 1946])。これにより、定量−定性値変換を、性質タイプを変更せずに行うことができる。2) 定性値のコンテキストを区別できる。生物種の違い、実験環境の違い、値の解釈や視点の違いなど、順序尺度値には様々なコンテキストが存在する。その違いを系統立てて記述でき、推論処理に用いることができる。3) 疾患の詳細病態データを用いることができる。臨床医学オントロジーにおける異常状態連鎖のデータを用いることで、他の疾患オントロジーには無い、疾患を構成する細かな病態(異常状態)と、測定値とのマッピングができる。一般に測定値は疾患の示す病態の一端にしか一致しないため、測定値は疾患そのものよりも異常状態にマッピングしやすくなるとともに、動物の検査データに含まれる異常値のうち、何個が疾患の詳細病態に一致するのか、といった“一致度”の概念を持ち込むことが可能であり、測定データに基づくより客観的な関係性提示が可能である。

これらの特徴は、様々な目的をもって行なわれる生物の計測データを整理統合し、それぞれの違いと同等性の情報をできるだけ劣化させず、かつ、出来うる限りシンプルに体系化して、データベース化するために重要な技術であると考えている。

4.2 表インターフェースとコンテキスト

法造を用いたデータ記述において最も有利なことの一つは、ロール概念を用いることで、ひとつの概念が様々なコンテキスト(場面や状況、視点等)において、元の概念の意味を保ちながら、異なる役割を演じて特殊化されることを明確に記述できることである。例えば、「コレステロール値が高い」という定性値は、1節に述べたように、1)経時的な変化としてコレステロール値が高い、2)単に他と比較してコレステロール値が高い、あるいは、3)哺乳類としてみたときに「正常」と比べ(異常に)コレステロール値が高い、など、様々に特殊化可能であり、法造ではこれらの相互関係を、系統立ててモデル化可能である。

本研究の表インターフェースでは、法造のオントロジーデータから、特定のコンテキストの性質値のみを選んで表示する。これは、特定のコンテキスト、目的、あるいは視点に基づいて、一部のデータのみを表示するような、いわゆる「view」であり、データ可視化の自動化にとって重要な課題であると考えている。本研究では、未だプリミティブなレベルにすぎないが、今後さらに検討を深め、科学データの統合化技術の一端を確立していきたいと考えている。

謝辞

本研究を行うにあたり、真下知士先生、高田豊行先生、若菜茂晴先生よりラットおよびマウスの表現型特性データの提供いただきました。ここに感謝の意を表します。また、本研究は JSPS 科研費 23300161 の助成を受けたものです。

参考文献

[Köhler 2013] Köhler S, Doelken SC, Ruef BJ, Bauer S, Washington N, Westerfield M, Gkoutos G, Schofield P, Smedley D, Lewis SE, Robinson PN, Mungall CJ.: Construction and accessibility of a cross-species phenotype ontology along with gene annotations for biomedical research., Version 2. F1000Res. 2013 Feb 1 (doi: 10.12688/f1000research.2-30.v2. eCollection 2013) (2013).

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

[NBRP ラット] http://www.anim.med.kyoto-u.ac.jp/nbr/

[NIG Mouse DB] http://molossinus.lab.nig.ac.jp/phenotype/

[Hoehndorf 2011] Hoehndorf, R., Schofield, P.N. and Gkoutos G.V.: PhenomeNET: a whole-phenome approach to disease gene discovery. Nucleid Acids Research, 39, e119 (2011)

[Hozo API] http://www.hozo.jp/hozo/

[Stevens 1946] S. S. Stevens: On the Theory of Scales of Measurement, Science 103: 677-680 (1946)

[YAMATO] http://www.ei.sanken.osaka-u.ac.jp/hozo/onto_library/upperOnto.htm

[大江 2009] 大江和彦: 病名用語の標準化と臨床医学オントロジーの開発, 情報管理, Vol. 52 , No. 12 p.701-709. (2009)

[桝屋 2010] 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 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)

[溝口 2009] Mizoguchi, R.: Yet Another Top-level Ontology: YATO, Proc. of the Second Interdisciplinary Ontology Meeting, pp.91-101, (2009)

[溝口 2012] Mizoguchi R., Kozaki K., Kou H., Yamagata Y, Imai T, Waki K, Ohe K.: River Flow Model of Diseases, Proc. of 2nd International Conference on Biomedical Ontology (ICBO2011),pp.63-70 (2011)