3I1-2

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

Trial to develop a database of context-dependent phenotype data

桝屋 啓志[*] Hiroshi Masuya
古崎 晃司[*2] Kouji Kozaki
大江 和彦[*3] Kazuhiko Ohe
溝口 理一郎[*4] Riichiro Mizoguchi

[*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

Biological measurement data capturing the phenotypes of organisms represent a broad range of variations in metadata, control data and contexts for the interpretation of biological functions. With the aim of integrating measurement data across various biological experiments, we attempted to develop a trial version of a 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 tabular form and provided functions for the conversion of qualitative data into quantitative data, which were the results of interpretations in specific experimental contexts. Furthermore, the application enables retrieval of related disease defined by Clinical Medical Ontology. This study provided a concrete example of a top-level ontology-based database that could be used as an integrated database of biological measurements to represent phenomes across biological species and experimental contexts.

1. はじめに

マウスおよびラットは、実験解析が可能な疾患モデル動物として極めて重要なバイオリソースである。今後、ゲノム情報や大規模解析技術の向上によって、これらのリソースの機能解析はますます進展していくと考えられ、疾患や創薬研究に対して大きく貢献することが期待されている。しかし、ニーズに十分に応えるためには、情報面の整備、つまり、多種多様な情報を整理統合し、互いの関連性が明らかになるようなかたちで研究コミュニティへ供給することが必須である。

生物の表現型解析における測定項目は極めて多種、多様である。これらの情報をコンピュータを用いて整理統合するには、測定対象となる部位や、生命現象、種間の違い、さらには、各実験によるコントロール値の違い、実験結果を解釈する考え方の違いなどを、あらかじめ与えられた知識情報に基づいて自動的に処理する必要がある。

例えば、「血糖値が高い」という定性値は、血液のグルコース濃度という性質タイプの値という意味だけでなく、「高い」ということがどのような意味を持つかを考慮する必要がある。具体的には、1)1 匹のマウスの血糖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 の血糖値を絶対的に比較した場合に、マウス Y の方が、血糖値が高い、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血糖値が高い、など「血糖値が高い」という定性値に様々な文脈(コンテキスト)が存在する。それぞれ、血糖値という性質について、何らかの値と比較して「高い」ことは共通しているが、比較対照や持っている意味が異なっている。3)の例の場合では、コントロール群を「正常」であると見立てたことで、実験区のマウスが「マウスとしては異常に血糖値が高い」ということになる。この場合は、糖尿病の主要病態としての高血糖と極めて近い意味を持つが、1)および 2)の例の場合は、直ちに疾患との関係が導かれる訳ではない。このようなコンテキスト依存の値の同等性や相違性を、機械可読な方式で記述する必要がある。

以前、我々は、上位オントロジーYet Another More Advanced Top-level Ontology (YAMATO) [溝口 2010, YAMATO]のフレームワークに、生物学分野で表現型記述に用いられる Phenotypic Quality オントロジー(PATO) [Gkoutos 05]の概念をインポートし、コンテキスト依存の定性値として再定義した参照オントロジー、PATO2YAMATO を作成した[桝屋 2010-1, 2011]。このオントロジーを用いると、例えば、「大きなアリ」と「小さなゾウ」を比較した場合、いかに大きなアリであっても、小さなゾウより小さい、といったことを、比較のコンテキストを区別することによって、記述できる。つまり、「生物種内での正常範囲と比べた大小比較のコンテキスト」、と「単純な大小比較のコンテキスト」の分類関係が定義されており、かつ、それぞれのコンテキストにおける「大きい」、および「小さい」値を区別することができる。そして、「アリとしては大きい」「ゾウとしては小さい」ことを、それぞれの生物種内での大小の値として定義しながら、かつ、それぞれの大小値が、単純比較のコンテキストに移行した場合は、「ゾウとしては小さい」という値は大きいという役割(ロール)、「アリとしては大きい」値は小さいというロールを担うことによって、小さなゾウより小さい」ことを示す事ができる。

本研究では、PATO2YAMATO オントロジーに基づいた表現型データベース、および推論システムを試作した。このシステムでは、多様な測定データを一貫した形式で記述する事ができる。また、オントロジー内に生物種分類を記述することで、種の違いを明示しながら、定量値から定性値への自動変換と、生物種としての正常／異常を区別する事ができる。

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

さらに、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ができる。

このしくみは、将来的に、モデル生物の表現型データを統合し、ヒト疾患情報データとの自動的な関連づけを行うシステムに発展させていくことができると考えている。

2. データおよびアプリケーション機能の概要

本ソフトウェアプログラムの機能は、表現型計測データ、および、そのメタデータ(生物種情報、実験コンテキスト、性質タイプや定性値の定義など)が、法造形式のオントロジーファイルで、かつ、PATO2YAMATO で規定されたセマンティクスで記述されていることを前提として設計されている。Java を開発言語とし、法造 API[Hozo API]を用いて開発を行った。

2.1 表現型解析測定データの記述

本研究では、マウス、ラットにおける下記の3つの公開データベースからデータをインポートした。1) 理研バイオリソースセンター日本マウスクリニック[JMC]。2) 京都大学・NBRP ラットデータベース[NBRP ラット]。3) 国立遺伝学研究所・マウス表現型データベース[NIG Mouse DB]。なお、インポートしたデータは、現状では部分的である。

全ての表現型データは、PATO2YAMATO における概念定義に従って、オントロジーファイルに直接定義を行った。YAMATO において、測定データは、特定の「形式」を用いて、性質を記述する「性質表現」という概念に分類される。PATO2YAMATO では、この記述形式として、測定対象である実体 (Entity: 以下 E), 性質タイプ (Attribute: 以下 A)、値 (Value: 以下 V) の三つ組みによって、実体が特定の性質を持つ事を記述する EAV 形式を採用しており、これに従って個々のデータを記述した(図1A 画面中に例を示す)。

(1) 性質タイプ

測定データの A には、各計測のパラメータである形質を代入する。形質は、基本概念としての性質タイプが、測定対象物 E のコンテキストにおいて、形質という役割(ロール)を演じている状態(ロールホルダー)として定義される。

また、本研究では、統計によって算出される平均などの特殊な性質タイプを、PATO2YAMATO 上に新たに定義した。平均等の統計は、グループとしての測定対象にのみ存在することから、グループのコンテキストでのみ成立するコンテキスト依存の概念である。従って、例えば平均体重は、「通常の性質タイプである体重が、グループのコンテキストにおいて、平均体重ロールを演じるロールホルダー」として定義できる。ロールホルダーと、そのプレイヤーとなる概念との間には、コンテキストの下で成立する継承関係が成立するため[太田 2011]、平均体重とは、体重の一種として推移的推論を行なう事が可能になる。

(2) 値

測定データの V フィールドには、計測値である定量値を代入した。また、YAMATO では、定性値を定義する際に、定量値を閾値として参照することができるので、これに従って、コンテキスト毎に、コントロール値を参照する定性値を定義した。例えば、遺伝学研究所の計測では、C57BL/6 系統で得られる値を、マウスにおける正常と見立てて実験を行っている。よって、「マウス種内での正常範囲と比べた大小比較のコンテキスト」において定義される定性値が「遺伝学研究所における実験」をコンテキストにおいても継承され、実際には C57BL/6J の値を閾値として定義されていることを記述した。

2.2 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、下記のような特徴がある。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。これによって、原因と途中経過を含めた一連の状態変化の連鎖と, それにより引き起こされている結果状態との総体が疾患であると定義されている[大江 2009]。2) PATO2YAMATO と同様に、YAMATO のフレームワークに準拠して構築されている。特に、疾患の要素である「異常状態」は、「特性」として定義されており、PATO2YAMATO で用いられている「定性値」概念との相互関係が明確に定義されている[溝口 2010, 山縣 2012]。

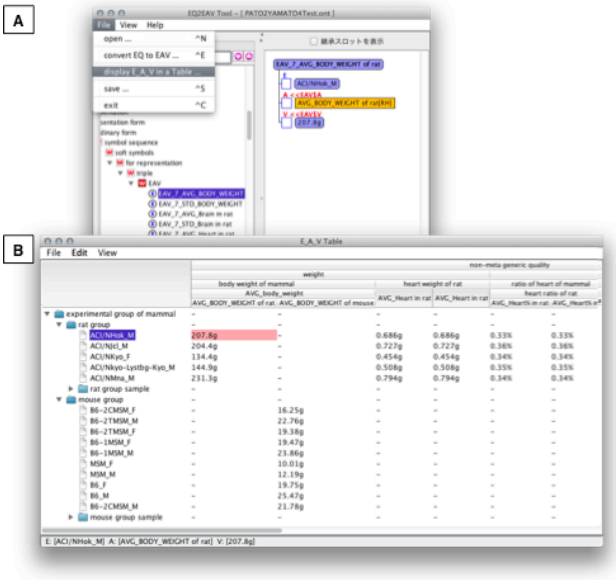

図1 本アプリケーションの画面。A: オントロジーのツリーと概念構造の図示。EAV による測定データを表示している。B: 複数測定データの表形式での表示。

2.3 異常値と異常状態の関係性の定義

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、同等性を示す対応表を、約 50 の異常状態について試験的に作成した。

2.4 アプリケーションの機能の概要

本ソフトウェアプログラムは、オントロジーとして定義された EAV 形式の測定データを、表形式で表示するようにした。この表では、水平方向に全ての性質タイプ、垂直方向に測定対象の動物グループが、その分類に従って細分化して示される(図1B)。デフォルトでは、計測データは、全てのコンテキストを区別した形で示される。例えば、マウスとラット、さらにそれぞれの雌

雄の平均体重は同列として比較できる測定パラメーター(性質タイプ)ではないので、別々のカラムに表示される。

(1) 定量値から定性値の変換機能

各定量値は、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、定性値へと変換することができる。この定性値化機能では、コンテキストを参照して、異なる定性値化が可能である。本アプリケーションでは、ユーザーが「遺伝学研究所における実験」の値といった、ひとつのコンテキストを選択することによって、測定パラメータ毎に一括した定性値変換を行なう事ができる。

(2) 性質タイプの統合機能

上記表において、性質タイプは、分類されて表示されている。この分類に従って、同じ種類の性質タイプを、上位概念としてマージする事が可能である。つまり、マウス平均体重は、マウス体重として扱う事ができる。さらに、マウス体重とラット体重は、上位概念の哺乳類体重として同等に扱う事ができ、マージすることができる(図2A および B)。

ただし、すでに定性値変換を行っている場合には、マージする対象の定性値が同じコンテキストで定義されており、かつ、上位概念での統合を行なって整合性があることが条件であるので、このアプリケーションは、定性値と性質タイプの上位概念を辿る事で、この整合性をチェックすることができる。例えば、マウスとしての異常値と、ラットとしての異常値は「哺乳類としての異常値」と見なせるために統合可能であるが、マウスとしての異常値と、マウスとラットを比較した値とは、上位性質タイプでの統合と整合性が無いために却下される。

(3) 疾患との関連性推論機能

マウスあるいは、ラットとしての異常値、すなわち生物コンテキストの下で「異常」と判断される値は、臨床医学オントロジーの異常状態と同等と見なすという対応テーブルを参照して、各異常値から、対応する異常状態を含む疾患を検索し表示する事ができる(図2C)。この機能により、当アプリケーションのユーザーは、詳しい医学知識を持たずとも、マウスおよびラットの測定データから、それと関連する疾患とその対応の根拠の情報を簡単に得る事ができる。

3. 考察と今後の課題

生物機能は分子間相互作用のネットワークで成り立っており、生命科学は個別知識の集大成であるとともに、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、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であり、研究者は、深い専門知識と同時に、常に専門外の知識に目を光らせ、出来うる限り広く深い情報収集を行なうことが成功の鍵となっている。そのため、従来文献として蓄積されてきた知識をいかにデータベース化し、いかに分野横断的に効率的に共有するかが、黎明期から変わらないバイオインフォマティクスの重要課題の一つとなっている。オントロジーを用いた知識の体系化、大規模データベース化は、それを解決する手段の一つとして大きな期待を背負っている。

本研究では、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ことを目指している。我々は、この目的に対する有望な方法論として、上位オントロジーYAMATO に準拠して、実験生物学の世界を記述し、測定結果をデータ、関連する概念をメタデータとして扱ったデータベースアプリケーションを試作した。このアプリケーションでは、様々な測定データを一貫した形式で格納でき、それぞれのメタデータに応じて分類/統合が可能で、さらには、臨床医学オントロジーを参照することによって、マウス、ラットにおける測定データと、その結果に直接関連する疾患を示すことを可能にした。ヒト疾患モデル生物としてマウスやラットを研究する場合に、疾患に関する知識は必須ではあるものの、その膨大な知識をフォローする事は必ずしも容易ではない。また、現代の遺伝子を中心とした研究では、研究者が予想しない生命機能にぶつかることは極めて多く、膨大な疾患と病態の知識を効率的に、マウス/ラットの研究者が利用するための研究サポートシステムとしては、極めて有効であると考えられる。

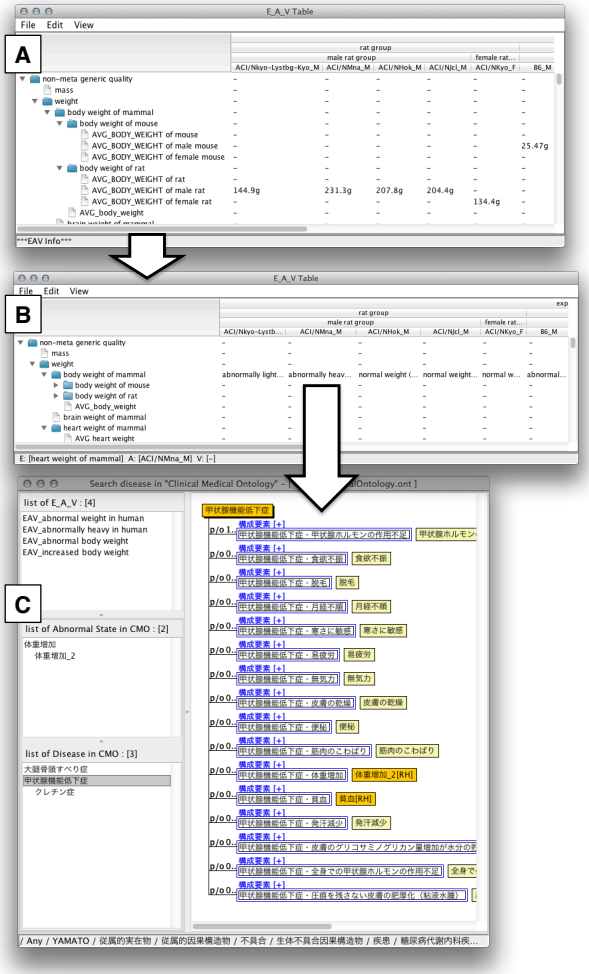

図2 定量値の定性値化と、関連疾患検索機能。A: 定量値表示画面。性質タイプ統合の準備として、表の行列を入れ替えて表示している。B: 定性値化と性質タイプの統合後。C: 定量値を一つ選び(体重の増加)、それに関連する疾患の検索。甲状腺機能低下症に、体重増加が含まれる。

本研究の大きな特徴は、表現型解析における様々なメタデータを、アドホックなパラメータではなく、オントロジーそのものとして記述した点にある。表現型解析は様々なコンテキストで行われる。それによって用いられるコントロールが異なるし、測定対象となる部位や、生命現象、種間の違い、さらには、各実験によるコントロール値の違い、実験結果を解釈する考え方の違いなど、区別すべき違いは多種多様である。本アプリケーションでは、

それらを個別に定義するのではなく、PATO2YAMATO オントロジー上で記述された実験世界そのものを参照して、解釈するようにしている。これによって、柔軟かつ一貫性を持った表現力を得た。さらに、オントロジーによる定義は、一度定義した条件(メタデータ)の持続的運用や標準化を可能にして、再利用に貢献すると考えられる。また、データレベルだけでなく、プログラムレベルでも、推移的推論等、推論のアルゴリズムの再利用に貢献し、結果的に効率的で持続的運用可能なシステム構築につながると考えられる。

ただし、異常値と、疾患の異常状態の対応など、一部でアドホックな実装を行っている部分もある。今後は、マウス、ラットの統合と同様に、E, A, V それぞれの同等性を評価し、臨床医学オントロジーで記述される疾患との関係を推論できるように改良を行う予定である。

一方、本アプリケーションは、データ入力や閲覧のインターフェースという点では、極めて未熟であることが大きな課題である。例えば、全てのデータをオントロジーとして記述する事は、極めて煩雑な作業であった。これは、YAMATO オントロジーの表現力と普遍性を考えれば当然とも言えるが、特定の用途では暗黙とされている情報を逐一明示することが求められることが、煩雑さに繋がっていると考えられる。

我々は、このような問題を解決するのはアプリケーションの役目であると考えている。今後は、ユーザーのドメインを絞り、入力データの種類を限定することによって、効率的なデータ入力を行なうことができるインターフェースの開発が必要と考えられる。さらに広い分野で連携するためには、ドメイン毎に異なるインターフェースを用意したり、想定外の用途でのみオントロジーエディターを併用するような、効率と汎用性を確保した入力ワークフローの設計が必要になると考えられる。

データの表示ではオントロジーに一般的なグラフ型、あるいはツリー型の表現だけではなく、適宜表形式を用いることの有効性が確認できた。表形式は、大量のデータの閲覧だけではなく、バルクでのデータ入力など、様々な場面で有効である。同じくオントロジー型のデータであるセマンティック Web では、表形式を用いた、理解しやすいインターフェースを用いたデータベース型アプリケーションが存在する[桝屋 2010-2]。

また、多くの場面で、表示する情報を適宜間引く必要性も感じられた。特に、値の概念ラベルは、「異常に軽い、ただし ACI/MKyo 系統をコントロールとして用いている」など、類似の概念との区別のために長い名称をつけることが、却って使用者の理解を妨げることがあった。システムとして下位概念の区別が必要でも、上位概念として丸められる場合は、必ずしもユーザーが異なる概念として認識していないことがあった。本システムで表示される概念は殆ど全てがロールホルダーとして定義されている。法造では、ロールホルダーのラベル定義は必須ではないので、必要な画面でのみ、コンテキストとクラス制約からラベルを合成した方が良いと考えられる場面もあった。

以上、本アプリケーションは、YAMATO オントロジーに準拠することにより、データ記述の普遍性および再利用性に関する大きなポテンシャルを持っていると考えられるが、そのポテンシャルを最大限発揮するためには、ドメインの視点に立ったデータ入力のサポート(インターフェース)も合わせて必要であると考えられる。

4. まとめ

本研究は、上位オントロジーに基づくデータ記述の標準化と知識表現が、表現型統合という極めて雑多なデータの統合に貢献しうるポテンシャルを持つ事を示した。今後、コンテキスト情報をより詳しく抽出し、データの統合対象を自動的に選ぶ事、データの入出力や閲覧における、理解しやすい／使いやすい GUI の追求、大規模データへの対応などの課題を解決する事で、多様な表現型の統合を可能とする新たな統合データベースとすることが可能になると考えている。

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を提供いただきました。また、太田衛様には、原稿に関するアドバイスをいただきました。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

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

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

[JMC] http://www.brc.riken.jp/lab/jmc/mouse_clinic/

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

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

[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)

[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」の拡張 ― 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)

[溝口 2010] Mizoguchi, R.: YAMATO: Yet Another More Advanced Top-level Ontology, The Sixth Australasian Ontology Workshop, pp.1-16, (2010)

[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)

[山縣 2012] 山縣友紀, 国府裕子, 古崎晃司, 今井 健, 大江 和彦, 溝口 理一郎: 異常状態オントロジーとその応用, 第26回人工知能学会全国大会予稿集, 1I2-R-4-3 (2012)