

平成 24 年度 研究開発実施報告書

ライフサイエンスデータベース統合推進事業「統合化推進プログラム」
研究代表者

豊田哲郎

独立行政法人理化学研究所生命情報基盤研究部門・部門長

生命と環境のフェノーム統合データベース

§1. 研究実施体制

(1) 豊田グループ

- ① 研究代表者 豊田 哲郎 (理化学研究所生命情報基盤研究部門・部門長)
- ② 研究項目
フェノーム統合化・プロパティ標準化・先端計測データ統合化・フェノーム利用・
ワークフロー開発

(2) 榊屋グループ

- ①主たる共同研究者: 榊屋 啓志 (理化学研究所バイオリソースセンター・マウス表
現型知識化研究開発ユニット・ユニットリーダー)
- ②研究項目
フェノタイプ記述子の体系化: 識別子体系化と評価
フェノーム統合化: バイオリソースフェノーム

§2. 研究実施内容

(文中に番号がある場合は(4-1)に対応する)

課題1「フェノタイプ記述子の体系化」

プロパティ標準化と評価 (豊田グループ)

RDF での体系的なデータ記述を行うためのプロパティを国内外で使用されているものに共通化させていくことで、データ統合化の基礎となるプロパティ標準化を行うことを目的に、URL の標準化を進めた。具体的には、まず、NBDC で提供している purl.jp システムに合



わせて、我々の biolod.org の URL をマッピングし直し、ダウンロードできるすべての RDF ファイルを purl.jp に変換して提供した。また、システムバイオロジーのポータルサイト BioGateway の開発グループによって提供されている標準プロパティセットである biorel を用いて、これまでに開発したデータベース中のプロパティ定義の標準化を進めた。biorel はオミックスデータに対応する標準プロパティセットとしては今のところ最も内容が充実しており、[Cell Cycle Ontology \(CCO\)](http://Cell Cycle Ontology (CCO))、[Gene Expression Knowledge Base \(GeXKB\)](http://Gene Expression Knowledge Base (GeXKB))などの公開サイトで使われている。現時点で、BRC 細胞リソースデータベース、BRC 表現型アノテーションデータベース、文献キュレーションによる植物フェノームデータベースについて標準化作業を完了した。さらに、作成したデータの評価手法として、アクセス解析を体系的に行えるようにすることで各データベースのランキングを計算し、サイト上で閲覧可能にした。その結果、バイオリソース系のデータのアクセスが上位にあることがわかり、成果がユーザから多く利用されている傾向がみられた(図 1)。アクセスログは来年度から NBDC の要請に応じて提供していける体制が整った。

全データベースの週間アクセスランキング
トップ50位 2013-04-23 ~ 2013-04-29

週間ランキング	データベース名	トレンド	機関
1	研究者	unch. →	BASE
2	細胞株(理研バイオリソースセンター)	unch. →	BRC
3	NCBI 分類学	unch. →	NCBI
4	科学誌一覧	unch. →	NCBI
5	マウス系統(理研バイオリソースセンター)	unch. →	BRC
6	理研 組織図	unch. →	
7	PDFファイルリポジトリ	unch. →	BASE
8	Mouse MGI Gene	up ↗	MGI
9	RIKENBASE	down ↘	BASE
10	OMIM	unch. →	NCBI
11	Protein Data Bank	up ↗	RCSB
12	MGI Alleles	up ↗	MGI
13	InterPro	up ↗	EBI
14	A.thaliana locus	down ↘	TAIR
15	細胞特性データ	up ↗	BRC
16	文献リポジトリ	down ↘	
17	mammalian phenotype ontology	up ↗	OBO
18	RIKEN Press Releases	down ↘	RIKEN

図 1:アクセスランキングの表示例

識別子体系化と評価 (榊屋グループ)

RDF での体系的なデータ記述を行うためのオントロジーや、インスタンスを国内外で使用されているものに共通化させていくことで、データ統合化の基礎となる識別子の体系化を行うことを目的として、昨年度は概念定義および、データ作成のワークフローを確立した。平成 24 年度は、このワークフローを用いて、特にマウスを対象として、実際的な表現型データ作成を行なった。各データは、OBO 提案の改良型である、部位、形質、表現型というそれぞれのフィールドに、オントロジー、あるいは、データベース値を、オントロジーの

下位概念と位置づけた上で、代入することによって、記述される（図2）。リソースデータベース、および関連論文を参照するキュレーション作業を行い、「識別子体系化作業」として、マウスの表現型データ約1000件、細胞の特性データ約5300件を作成した。今後、同様のワークフローを用いて、微生物特性のデータ作成を行う予定である。

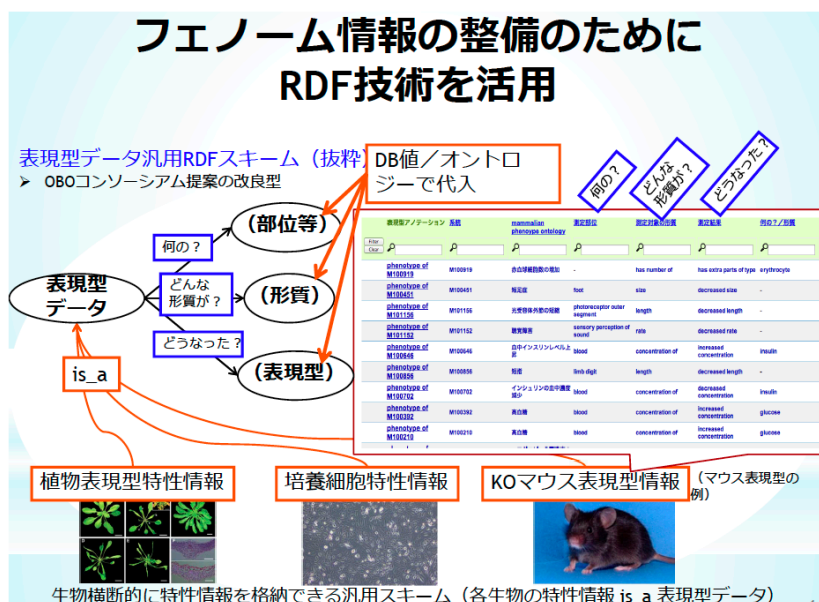


図2: RDFでのフェノームデータ記述の概要

課題2「フェノーム統合化」

バイオリソースフェノーム (梶屋グループ)

バイオリソースに関連付けられるフェノーム情報について、上記のフェノタイプ記述子に対応付けながら情報の収集と整理を行い、マウス系統、細胞株、微生物株、植物株などの表現型情報や有用性情報を統合化する目的で、平成24年度は、バイオリソースデータ収集プラットフォームを培養細胞についても拡張した、「基盤データ収集のための、バイオリソースデータ収集プラットフォーム (細胞株) 構築」を作成し、これを用いて、「細胞株リソースデータ収集作業」として、約3600株の情報を収集、データベース化し、月ごとに最新情報を配信している。マウスにおいても月ごとのデータ更新を続行するとともに、識別子体系化作業において作成した表現型データとリソースデータの結合を行い、下記に述べるフェノーム利用のために開発したインターフェース (画面) を用いて、配信を行なっている。

先端計測データフェノーム (豊田グループ)

シロイヌナズナに関して、以下の2種類のデータセットを最新の情報に更新したうえで、後述のSWASによる推論検索に使えるように再編集した。



1. 文献キュレーションにより収集したシロイヌナズナ表現型情報 824 件。文献上の表現型観察情報を個々に解釈し、それぞれオントロジーで標準化するとともに関連遺伝子へのリンク等を整備したもの。

2. 理研内で開発されたシロイヌナズナ変異株 3763 種類の表現型情報 14631 件

この統合化により、表現型関連のキーワードから、その表現型への関与が考えられる遺伝子の候補やその表現型の解析に利用可能な変異株の候補などの推論検索が実現した。

また、マウスに関して新規 874 件、培養細胞に関して 新規 5000 件のフェノームデータも追加され、SWAS により表現型から関連するマウス系統等を検索できるようになった(図 3)。

これらを踏まえ、平成 25 年度はフェノーム情報とバイオリソース情報等との統合、フェノーム情報を出発点とした各種情報の検索システムの整備をさらに進め、フェノーム情報の利用価値を高める。

The screenshot shows the PosMed Positional Medline search results for the keyword 'flowering' in Arabidopsis. The search returned 269 hits in 0.297 seconds. The results are displayed in a table with columns for rank, mutant name, gene IDs, document count, and links to RIKEN and SciNetS. The top results include:

Rank	Mutant Name	Gene IDs	Docs	Hits	Phenotype
1.	F24649, Mutant of FOX hunting; rosette l...	F24649, Mutant of FOX hunting; rosette L...	1 doc	1 hit	Flowering
2.	Z032806, Mutant of Activation tagging li...	Z032806, Mutant of Activation tagging li...	1 doc	1 hit	Flowering
3.	Z031044, Mutant of Activation tagging li...	Z031044, Mutant of Activation tagging li...	1 doc	1 hit	Flowering
4.	F23728, Mutant of FOX hunting; seed:incr...	F23728, Mutant of FOX hunting; seed:incr...	1 doc	1 hit	Flowering
5.	F12202, Mutant of FOX hunting; rosette l...	F12202, Mutant of FOX hunting; rosette L...	1 doc	1 hit	Flowering
6.	Z100145, Mutant of Activation tagging li...	Z100145, Mutant of Activation tagging li...	1 doc	1 hit	Flowering
7.	11-2516-1, Mutant of DS transposon line;...	11-2516-1, Mutant of DS transposon line;...	1 doc	1 hit	Flowering

図 3 : SWAS 検索の例



課題3「フェノーム利用ワークフロー開発」




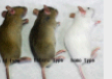
RDF で記述されたマウス表現型データを分かりやすく配信するためのインターフェース開発を行った。RDF のデータは関連情報が網羅的につながったネットワークを構成しているが、まずはシンプルに表を用いたカード型の画面によって最低限の情報を見せておき、関連情報ボタンをクリックすることにより、関連情報を次々と開いて参照できるようにした。また、RDF のリンクを用いて、関連情報を自動検索するインターフェースも開発した。これを用いて、1つのマウス系統から関連する表現型を示す、Amazon のオンラインショップを模した「お勧め機能」を作成した(図4)。

**RDFの繋がりを機械的に探索し、
「お勧めマウス」を自動提示**

RDFのリンク辿る事で、類似の表現型を示すマウスを自動収集。
“Amazon” 風のインターフェースで、より目的に合ったリソースに誘導

RDFのリンク (抜粋)

同じ表現型を示すマウス系統 (56) (各系統の示す表現型の一致度が高い順番に提示)

<p>7個のmammalian phenotype ontologyが一致 C57BL/6-Wf/+</p>  <p>共通の mammalian phenotype...</p> <ul style="list-style-type: none"> 肥満細胞数の減少 薄い毛色 腹部の白斑 大球性貧血 赤血球数の減少 	<p>4個のmammalian phenotype ontologyが一致 C57BL/6-Wsh/Wsh</p>  <p>共通の mammalian phenotype...</p> <ul style="list-style-type: none"> 肥満細胞数の減少 赤毛色異常 赤血球数の減少 可変性体部斑点 	<p>3個のmammalian phenotype ontologyが一致 C57BL/6-KitW-37J</p>  <p>共通の mammalian phenotype...</p> <ul style="list-style-type: none"> 肥満細胞数の減少 薄い毛色 白色スポット 	<p>C3.Cg-Kitl<SI-pan&g...</p>  <p>共通の mammalian phenotype...</p> <ul style="list-style-type: none"> 薄い毛色 大球性貧血 白色スポット
---	--	--	--

類似度高 ← → 類似度低

図4：お勧め機能の概要

また、以前に開発した、統計的なセマンティック Web データの検索システム PosMed を統合して、Semantic Web の特性を活用した大規模相関解析を実現し、これを SWAS (Semantic-Web Association Study) と名付けた。現在、ゲノム領域を対象として遺伝情報の相違を検定して病気の原因遺伝子等を見つけ出す GWAS という手法が盛んに利用されているが、SWAS ではこの発想をセマンティックウェブデータに適用し (図5)、研究対象の表現型などのキーワードと相関の高いデータ群を検定により見つけ出す。本システムは、以前からセマンティックウェブベースのデータの検索問合せに利用されている SPARQL と比較して、高速性、セキュリティへの配慮などの点で優位性がある。

フェノームを起点として多様なオミックスデータを横断的に検索するシステムの整備の必要性は研究コミュニティでも指摘されており^{*1)}、SWAS 解析システムの整備はそのニーズ

を実現するものと考えている。

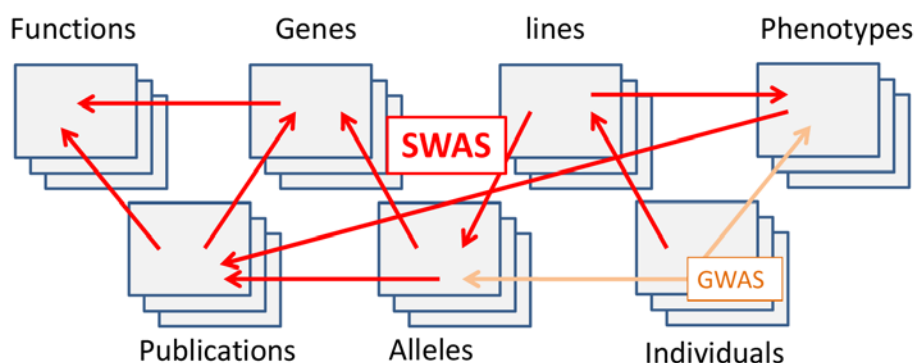


図 5 : SWAS 解析の概念図

§ 3. 成果発表等

(3-1) 原著論文発表

- ① 発行済論文数(国内(和文) 0 件、国際(欧文) 1 件) :
- ② 未発行論文数(“accepted”、“in press”等)(国内(和文) 0 件、国際 (欧文)0 件)
- ③ 論文詳細情報

* 1 . Baerenfaller K, Bastow R, Beynon J, Brady S, Brendel V, Donaldson S, Forster M, Gifford D, Grotewold E, Gutierrez R, Huala E, Jaiswal P, Joshi H, Kersey P, Liu L, Loraine A, Lyons E, May S, Mayer K, MacLean D, Meyers B, Mueller L, Muller R, Muller H.M, Ouellette F, Pires J.C, Provart N, Staiger D, Stanzione D, Taylor J, Taylor C, Town, C.J. Toyoda T, Vaughn M, Walsh S, Ware D, Weckwerth W. ‘Taking the Next Step: Building an Arabidopsis Information Portal’ Plant Cell 24: 2248-56(2012) (DOI: 10.1105/tpc.112.100669)

シロイヌナズナの研究コミュニティにおいて蓄積された各種オミックス情報を今後どう収集し、役立てているべきか、多方面からの検討が行われている。その中で、フェノーム情報から関連各種情報を検索するシステムが重要であり、特に世界規模の食糧・エネルギー問題に対処するためにも表現型-遺伝情報間の関連解析から遺伝子の役割を解明することが重要と指摘している。そのような課題に対処しうる実例の一つとして理研 BASE で過去に開発した PosMed システムにも言及している。

(3-2) データベースおよびウェブツールの構築と公開

公開中のデータベース・ウェブツール等



別紙を参照

(3-3) その他の著作物(総説、書籍など)

1. 豊田哲郎 “SciNetS: セマンティックウェブ技術を活用した創薬のための情報基盤”
SNR News No.23 2-5 (2012)
2. Hiroshi Masuya “Roles and Applications of Biomedical Ontologies in Experimental Animal Science”, *Experimental animals* vol. 61, No. 4, 365-373, 2012 (DOI: 10.1538/expanim.61.365)
3. 土井考爾・榎屋啓志・豊田哲郎、フェノーム・バイオリソースを中心とした統合データベース開発の現状、日本バイオインフォマティクス学会ニューズレター第 26 号 (2013 年 4 月) 掲載予定 (投稿済み)

(3-4) 国際学会および国内学会発表

- ① 招待講演 (国内 4件、国際 3件)

〈国内〉

1. 榎屋啓志、“マウス表現型の情報統合へ向けた研究”、第 26 回モロシヌス研究会、東京、6 月 15 日
2. 豊田哲郎、“公共データの RDF 化と公開を簡単に行う方法” 第 1 回 LOD チャレンジデー「Linked Open Data 入門」、東京、8 月 25 日
3. 豊田哲郎、“公共データの RDF 化と公開を簡単に行う方法” 第 2 回 LOD チャレンジデー「Linked Open Data 入門」、名古屋、9 月 1 日
4. 榎屋啓志、上-中位オントロジーからつくる表現型統合データベース、ライフサイエンス統合データベースセンター勉強会、東京、2 月 18 日

〈国際〉

1. Masuya H, “MUSDB: LIMS for Japan Mouse Clinic”, IMPC Informatics Workshop, Washington DC, 5月 25 日
2. Masuya H, Development of information technologies for the integration of mouse phenotype data, Infrafrontier/IMPC Korea Meeting, Jeju Korea , 9 月 27 日
- *3 . Masuya H , Internationally Cooperative Development of the Informatics



Infrastructure for IMPC、IMPC International Symposium、Tokyo、9月28日

② 口頭講演 (国内 3件、国際 3件)

〈国内〉

1. 梶屋啓志、田中信彦、脇和規、高月照江、齋藤実香子、小幡裕一、理研 BRC IMPC (国際マウス表現型解析コンソーシアム) 参画計画3 -表現型解析情報の公開と統合-、日本実験動物科学・技術 九州、別府、5月24日
2. 梶屋啓志、溝口理一郎、遺伝子のオントロジー、2012年度 人工知能学会 全国大会(第26回)、山口、6月5日
3. 梶屋啓志、溝口理一郎、遺伝子の意味を情報モデルとして記述する。日本遺伝学会第84回大会、福岡、9月24日

〈国際〉

1. Masuya H and Mizoguchi, An Ontology of Gene, 3rd International Conference on Biomedical Ontology (ICBO) 、Graz、9月23日
2. Masuya H, “A trial of the development of Linked Open Data (LOD) of bioresources”, 4th ANRRC International Meeting (ANRRC 2012) Jeju Korea 10月18日
3. Masuya H, Takatsuki T, Makita Y, Yoshida Y, Mochizuki Y, Kobayashi N, Yoshiki A, Nakamura Y, Toyoda T, Obata Y、Development of Linked Open Data for Bioresources, 2nd Joint International Semantic Technology Conference (JIST2012) 、Nara、12月3日

③ ポスター発表 (国内 5件、国際 0件)

〈国内〉

1. 田中信彦・茂木浩未・鈴木智広・金田秀貴・三浦郁生・山田郁子・古瀬民生・小林喜美男・土岐秀明・井上麻紀・美野輪治・野田哲生・若菜茂晴・梶屋啓志、“マウス表現型の類似度解析ツールの開発:ワークフローの検討”、日本実験動物科学・技術 九州 2012、別府、5月25日
2. 土井考爾・西方公郎・下山紗代子・梶屋啓志・豊田哲郎、BioLOD:フェノームおよびバイオリソース情報の統合と共有、トーゴの日シンポジウム 2012、10月5日
3. 梶屋啓志・高月照江・齋藤実香子・松庫義弘・蒔田由布子・吉田有子・望月芳樹・小林紀郎・吉木淳・中村幸夫・豊田哲郎・小幡 裕一、哺乳類バイオリソースの Linked Open Data データベースのアップデート、第35回日本分子生物学会年会、福岡、12月12日
4. 田中信彦・茂木浩未・鈴木智広・金田秀貴・三浦郁生・山田郁子・古瀬民生・小林喜美男・土岐秀明・井上麻紀・美野輪治・野田哲生・若菜茂晴・梶屋啓志、網羅的マウス表現型解析デー

タを用いたデータマイニング:ワークフローの検討、第 35 回日本分子生物学会年会、福岡、12 月 13 日

5. 土井考爾・下山紗代子・西方公郎・高月照江、BioLOD (Biological Linked Open Databases) を使ったフェノタイプ情報の共有と検索、第 35 回日本分子生物学会年会、福岡、12 月 13 日

〈国際〉

該当なし。

(3-5) 知財出願

該当なし。

(3-6) 受賞・報道等

該当なし。

§ 4. 研究期間中に主催した活動(主催したワークショップ等)

該当なし。

