

背景と目的

データリポジトリの意義は、例えばメタ解析、即ち多数のデータの解析結果から全体の傾向を見出す解析のような、データを再利用する研究に資することである。公開データから各オミクスのメタ解析を行い、その結果を統合する「疑似トランスオミクス解析」研究は今後の増加が予想され、それに備えることはまた、データリポジトリの今後の課題と考えられる。

しかし実際には、メタデータの項目・用いられている統制語彙などはオミクスによって異なり統一されていない。メタデータの内容を規定する基準は各分野に存在するにもかかわらず、具体的な語彙の使用については規程が存在しないため、同一概念に対して使用されている語彙がオミクス分野によって異なる場合さえある。

我々は今までに、プロテオーム・データリポジトリ & データベース jPOST 及びメタボローム・データ解析パイプライン Shin-MassBank の2つのプロジェクトに於いてデータリポジトリを開発し、メタデータの処理を実装してきた。一般にプロテオームとメタボロームは、データ処理のプロセスが質量分析の応用分野の中では最も類似しているとされるが、実際に jPOST のメタデータをベースに Shin-MassBank のリポジトリ MB-POST 用のメタデータを作成する際に比較したところ、データの種類・データの構造・実際に用いられる語彙 など多くの点に差異が見られた。

本発表では、将来的なりポジトリ・データ解析の統合を見据えて、これら両分野のメタデータの対応づけを開始した。

プロテオーム対メタボローム：メタデータ比較（1）

メタデータの構成

・最少情報基準のあるプロテオミクスでは、**生データファイル毎**にそのデータの属性（試料の内容や行った前処理など）を記述する**SDRF** (Sample Data Relation Format) 情報が用いられる。

1	source name	characteristics[organism]	characteristics[organism part]	characteristics[cell line]	characteristics[ancestry category]	
2	Sample 1	Homo sapiens	not applicable	HeLa	Black	
3	Sample 10	Homo sapiens	not applicable			
4	Sample 11	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120315QEx2_RS1_50nl-min_10ngHeLa_3h_01.raw	1
5	Sample 12	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120327QEx2_RS1_50nl-min_0k1HeLa_14h_01.raw	1
6	Sample 13	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120332QEx2_RS1_20nl-min_0k1HeLa_10h_01.raw	1
7	Sample 14	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120315QEx2_RS1_50nl-min_10ngHeLa_3h_02.raw	1
8	Sample 15	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120323QEx2_RS1_20nl-min_0k1HeLa_5h_01.raw	1
9	Sample 16	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120323QEx2_RS1_20nl-min_0k1HeLa_3h_01.raw	1
10	Sample 17	Homo sapiens	not applicable	proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120315QEx2_RS1_20nl-min_10HeLa_8h_01.raw	1
				proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120330QEx2_RS1_50nl-min_10ngHeLa_14h_01.raw	1
				proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120315QEx2_RS1_50nl-min_10ngHeLa_3h_01.raw	1
				proteomic profiling by mass spectrometry	https://ftp.pride.ebi.ac.uk/pride/data/archive/2014/07/PXD000396/120315QEx2_RS1_50nl-min_10ngHeLa_3h_01.raw	1

PXD000396

<https://github.com/bigbio/proteomics-sample-metadata/blob/8297770171ce449694d92a3df65d24fd327695b8/annotated-projects/PXD000396/PXD000396.sdrf.tsv>

ファイル名が記載されている

・これに対してメタボロミクスでは、ISA (Investigation, Study and Assay) データモデルに基づいたメタデータが作成されており、SDRFに類似する情報としては"Study Design"情報が用いられるが、ファイル名は必須ではなく、記載されていないことが多い。

Source Name	Characteristics[Organism]	Term Source REF	Term Accession Number	Characteristics[Organism part]	Term Source REF	Term Accession Nun
Wild Type	Escherichia coli	NCBITaxon	http://pubt.obolibrary.org/obo/NCBITaxon_562	Cell	SLSO	http://www.ebi.ac.uk/ontology/term/SLSO
V610F	Escherichia coli	NCBITaxon	http://pubt.obolibrary.org/obo/NCBITaxon_562	Cell		
Wild Type	Escherichia coli	NCBITaxon	http://pubt.obolibrary.org/obo/NCBITaxon_562	membrane		
V610F	Escherichia coli	NCBITaxon	http://pubt.obolibrary.org/obo/NCBITaxon_562	membrane		

https://www.ebi.ac.uk/metabolights/ws/studies/MTBLS13050/download?file=s_MTBLS13050.txt

MetaboLights: MTBLS13050

・Study Designで示されるのは基本的に「試料についての属性」であるが、一つの試料が必ず一つのファイルに記録されるわけではないので、このデータを再利用（再解析）するためには、他のメタデータと突き合わせる必要が最低限必要になる。

・この意味で、データ再利用には追加の処理が必要になる。

Metabolomics Workbench: ST004161

<https://www.metabolomicsworkbench.org/data/DRCCMetadata.php?Mode=Study&DataMode=FactorsData&StudyID=ST004161&StudyType=MS&ResultType=1#DataTab>

label	n1	n2	n3
Homo sapiens	Caenorhabditis elegans	chimpanzee	Aedes aegypti
human	dog	horse	apple
Pan troglodytes	Pan paniscus	Bufo gargarizans	Theropithecus gelada
chimpanzee	cheetah	Plasmodium malariae	bonobo
Pan paniscus	Pan troglodytes	Panulirus ornatus	Paniscus halli
bonobo	dingo	chimpanzee	cheetah
Gorilla gorilla gorilla	Marmota marmota marmota	Bison bison bison	Conger conger
western lowland gorilla	western yellowjacket	western mosquitofish	western predatory mite

Ilama

label	n1	n2	n3
Homo sapiens	Hippocampus comes	Hydra vulgaris	Hyaena hyaena
human	rabbit	rat	pig
Pan troglodytes	Pan paniscus	Theropithecus gelada	Panthera pardus
chimpanzee	chum salmon	chicken	cherry salmon
Pan paniscus	Panthera uncia	Panthera pardus	Panthera leo
bonobo	dingo	wapiti	biaby
Gorilla gorilla gorilla	Gekko japonicus	Hydra vulgaris	Aquarana catesbeiana
western lowland gorilla	eastern brown snake	western painted turtle	northern sea otter

背景：生データの公開

・測定データ（生データ）を公開する、という制度はヒトゲノム計画のデータを対象に開始され、順次他のオミクスに拡大した **[1]**。

対象：

最少情報基準
(最少報告ガイドライン)：

1996

2001

2003

2007

2008

2009

2012

バミューダ原則

フォート・ローダーデール協定

アムステルダム原則

トロント協定

HGP (Human Genome Project)

各種ゲノム計画

プロテオーム

大規模測定、
メタボロームや
コホート研究

MIAME (マイクロアレイ)

MIAPE (プロテオミクス)

MIRAGE (グライコミクス)

・この結果、「データは解析されなくても、存在している（産生された）だけで価値がある」という思想が一般化し、データの再利用（プロテオームなど FDR 制御に Target-Decoy 法を使っている場合はデータの再解析）が普及した。

・再利用のためにはデータの属性を説明する**メタデータ** (metadata) が必須になり、関係学会によって順次、「最低でもこれだけの情報は必要」という「**最少情報基準** (minimum information standard)」が決定された。但し、例えばメタボローム分野では Metabolome Society が対応できなかったため、メタボロミクスの最少情報基準は存在していない。

・このため、特にメタボロミクスでは「基準に寄せる」形で用語を標準化することが難しい。

プロテオーム対メタボローム：メタデータ比較（2）

メタデータの項目

・Shin-MassBankのリポジトリMB-POST構築時には、先行するMetaboLights/Metabolomics Workbench/MetaboBankのメタデータを調査し、それをjPOST（正確にはデータジャーナルJPDM **[2]**）のメタデータにマップし、更にメタボローム研究者が吟味することによって項目を決定した。

・右パネルに示すように、プロテオミクスでは測定方法・ソフトウェア解析の段取りが固まっているのに対し、メタボロミクスでは質量分析の測定方法がより複雑である。

・例えばプロテオミクスでは陽イオンモードでの測定のみであるのが通常であるが、メタボロミクスでは陰イオンモードの測定も行うことが多い

・またメタボロミクスでは情報解析の定番方法論がないため、自由記述の比率が高い。

・これらの差異については各分野の測定・解析に基づく必然性があるため、無理な用語の統一よりも、項目を増加させてカバーする（例えばプロテオミクスでは「ionization」に「ESI」、「polarity」に「positive」、「collision energy」欄は空白、など）のが適切と思われる。

プロテオームとメタボロームのメタデータで
対応するカテゴリにおいてそのーム特有の内容

JPOST
Repository Database

Japan Proteome Standard
Repository/Database

Fractionation
細胞・タンパク質
ペプチドの分離

Enzyme_mod
(消化酵素・
翻訳後修飾の設定)

MS_mode
Quantification platform
Plex

Software_setting
(より詳細)

Shin-MassBank
Mass spec data processing pipeline

Preparation
誘導体化
内部標準の有無・種類
(対応カテゴリなし)

Analytical_condition
ionization
polarity
collision energy

Software_setting
(自由記述多し)

プロテオーム対メタボローム：メタデータ比較（3）

メタデータの語彙

・項目同士に対応がついた場合でも、用いられている語彙自体が異なることがあり、その対応づけが必要になる。現状では、LLM (Large Language Model) を用いるのが最も適当と考えられる。

・予備調査として、「生物種名と学名の対応」について、各種の LLM が持つ情報について調査した。

・LLM: multilingual-e5-large (e5), all-MiniLM-L6-v2 (l6), Meta-Llama-3.1-8B-Instruct (llama)

各 term を LLM で embedding vector に変換、各ベクトル間の (1-cosine 類似度) を計算、特徴空間内での term 間距離を得る。

なお llama はプロンプトからの質問「humanの学名は何ですか？」
「学名が Homo sapiens の生物種は何ですか？」に正しく回答している。

・右表に示すように、単語間の「距離」を単純に比較する場合に、一般名と学名をタイトに結びつけている LLM は、調査したものにはなかった。但し l6 は「ヒト→学名」と「ウェスタン・ローランド・ゴリラとその学名の双方向」を正しく第 1 位にしており、一般名と学名の対応づけを目指す追加学習には適切かもしれない。

各種言語モデル中の生物種名・学名の類似度の例

e5

label	n1	n2	n3
Homo sapiens	Caenorhabditis elegans	chimpanzee	Aedes aegypti
human	dog	horse	apple
Pan troglodytes	Pan paniscus	Bufo gargarizans	Theropithecus gelada
chimpanzee	cheetah	Plasmodium malariae	bonobo
Pan paniscus	Pan troglodytes	Panulirus ornatus	Paniscus halli
bonobo	dingo	chimpanzee	cheetah
Gorilla gorilla gorilla	Marmota marmota marmota	Bison bison bison	Conger conger
western lowland gorilla	western yellowjacket	western mosquitofish	western predatory mite

l6

label	n1	n2	n3
Homo sapiens	chimpanzee	human	Bornean orangutan
human	Homo sapiens	dog	chimpanzee
Pan troglodytes	Gorilla gorilla gorilla	Atta cephalotes	Trachymyrmex cornetzi
chimpanzee	Pan troglodytes	pig-tailed macaque	golden snub-nosed monkey
Pan paniscus	Pan troglodytes	Panulirus ornatus	Suncus etruscus
bonobo	Bolivian squirrel monkey	Gorilla gorilla gorilla	rhesus monkey
Gorilla gorilla gorilla	western lowland gorilla	chimpanzee	green monkey
western lowland gorilla	Gorilla gorilla gorilla	Plasmodium sp. gorilla clade G2	African savanna elephant

llama

label	n1	n2	n3
Homo sapiens	Hippocampus comes	Hydra vulgaris	Hyaena hyaena
human	rabbit	rat	pig
Pan troglodytes	Pan paniscus	Theropithecus gelada	Panthera pardus
chimpanzee	chum salmon	chicken	cherry salmon
Pan paniscus	Panthera uncia	Panthera pardus	Panthera leo
bonobo	dingo	wapiti	biaby
Gorilla gorilla gorilla	Gekko japonicus	Hydra vulgaris	Aquarana catesbeiana
western lowland gorilla	eastern brown snake	western painted turtle	northern sea otter

現時点でのまとめ

・データの統一的な取り扱いやトランスオミクス解析のために、メタデータの対応づけは必須である。そのためには最少情報基準の突き合わせが効率的であるが、基準のない分野のためには“自前”で対応づけを検討する必要がある。

・基準のないメタボロミクスなどの場合、リポジトリ毎にメタデータを調査し、どのファイルにどの項目が含まれるか調査する必要がある。更に分野特有のメタデータについては、それらを網羅してメタデータ項目の合併集合を作成する必要がある。

・最も困難があるのは語彙自体の対応づけである。これは「全場合に対応できる網羅的なオントロジー」の方向よりも、文脈に応じて詳細に意味を指定する方が現実的と考えられ、そのためには LLM を用いるのが有効と考えられる。但し LLM はモデルによって内部に持つ情報の構造化に差があると考えられるため、最適なモデルを選び、その上で目的に特化した追加学習を行うことが必要と考えられる。

・今回の簡単な予備調査でも、プロンプトからの質問への回答が、モデル内部の構造化情報をそのまま出力しているのではないことが強く示唆された。適切なモデルの選択・適切な追加学習・適切なプロンプトエンジニアリングの組み合わせが、メタデータの対応づけという問題解決に必要であると考えられる。

謝辞

jPOST 及び Shin-MassBank: MB-POST の研究・開発は、科学技術振興機構 (JST)・NBDC 事業推進室による統合化推進プログラム予算によって実施した (課題番号: JPMJND2304, JPMJND2305)。JPDM の発行には、日本学術振興会の科学研究費補助金・研究成果公開促進費 (国際情報発信強化 (B)) の支援を受けた (採択課題番号: 21HP2004)。

また本研究の実施には、日本学術振興会の科学研究費基金・基盤研究 (B) の支援を受けた (課題番号: 25K03216)。

科研費
KAKENHI

文献

[1] 吉沢明康・小林大樹・河野信, *Proteome Letters*, **10**, 11-26 (2025)

[2] 吉沢明康 他, トーゴーの日シンポジウム2024、ポスター16 (2024)

Oct. 20, 2025 品川ザ・グランドホール, 東京

トーゴーの日シンポジウム 2025

Licensed under a Creative Commons 表示4.0国際ライセンス
©2025 吉沢 明康 (新潟大学)

