

jPOST: 「コミュニティ・ベース論文調査」によるメタデータキュレーションの加速



○吉沢明康¹⁾、守屋勇樹²⁾、小林大樹³⁾、張智翔⁴⁾、奥田修二郎⁵⁾、田畑剛⁶⁾⁷⁾、河野信¹⁾²⁾、幡野敦³⁾、高見知代³⁾、松本雅記³⁾、山ノ内祥訓⁸⁾、荒木令江⁴⁾、岩崎未央⁷⁾、杉山直幸⁶⁾、福島敦史⁹⁾¹⁰⁾、田中聡¹¹⁾、五斗進²⁾、石濱泰⁶⁾

1)富山国際大学 現代社会学部 2)情報・システム研究機構 ライフサイエンス統合データベースセンター 3)新潟大学 大学院医歯学総合研究科
4)熊本大学 大学院生命科学研究部 5)新潟大学 医学部メディカルAIセンター 6)京都大学 大学院薬学研究科 7)京都大学 iPS細胞研究所
8)熊本大学 医学部附属病院 9)京都府立大学 大学院生命環境科学研究科 10)理化学研究所 環境資源科学研究センター 11)Trans-IT

プロテオーム統合データベースjPOSTでは計画開始以後、以下のような開発を行ってきた：

(i) 高速ファイル転送とユーザーフレンドリーなインターフェースを実装した**リポジトリ**の開発（トーゴの日シンポジウム・ポスター・2016年44, 2017年57, 2020年53）

(ii) データの**再解析**プロトコルの開発（同2017年59）

(iii) **RDFデータベース**の構築と多数のリソースとの連携（同2017年60, 2019年35）

このうち(ii)の効果については、COVID-19研究の公開データを再解析することによって同定ペプチド数が65%増加するなどの効果を確認している（同2021年35）。



この**再解析**には正確なメタデータが不可欠である。jPOSTのメタデータはMAGE-TABのIDFをカバーするが、SDRF相当情報を得るには人間による追加のキュレーションが必要であり（同2017年58, 2021年36）、再解析にもこの情報が必要である。このためにデータジャーナルJPDMの創刊による情報収集も開始している（同2019年35, 2020年53）が、現時点では依然として、多くのデータセットについては論文からの情報抽出が必要である。

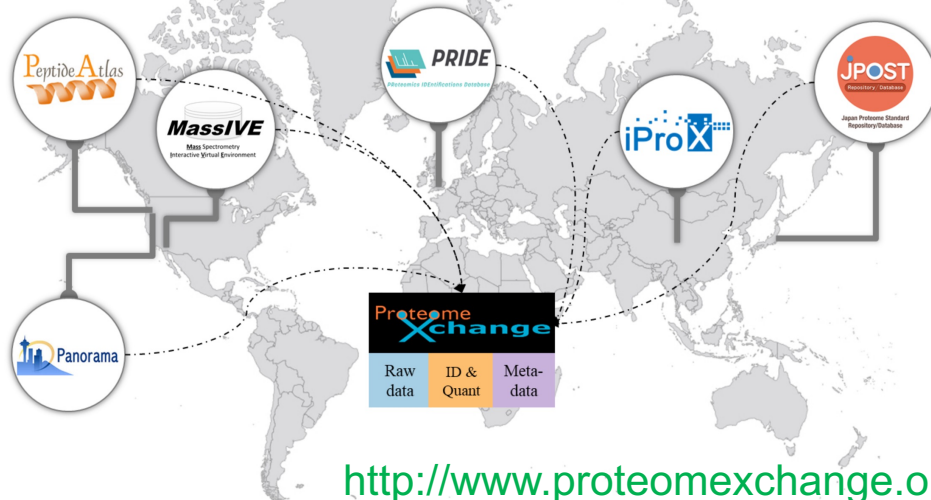
そこで (a)再解析のためのキュレーション (b)キュレーション半自動化のためのトレーニングデータセット構築 (c)学生・若手研究者のトレーニング を目的とした、「論文の下読み・情報抽出によるメタデータ作成補助」企画を開始し、日本プロテオーム学会を通じて担当者を募集、メタデータの作成を加速した。本発表ではこの取り組みについて発表する。



jPOSTの目的

データの公開 — プロテオミクスの場合

- 結果だけでなく生データも公開
 - データ解析に改善の余地
 - 学術ジャーナルの主導
- ProteomeXchange
 - データの収集方法について一定の基準を満たしたリポジトリの連合



<http://www.proteomexchange.org/>

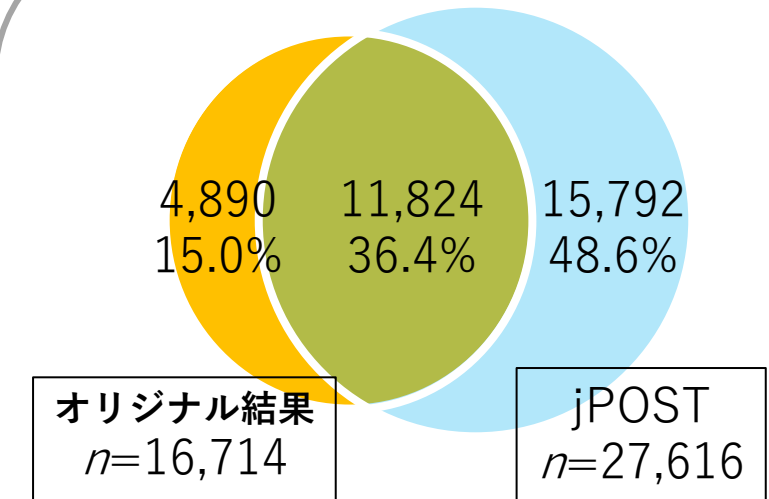
- ProteomeXchangeのメンバーとして、世界的なデータの蓄積と公開の負担の一翼を担う
 - 公開にあたっての基準はすべて、独自のものではなく**世界標準のもの**を採用する
 - 特にProteomeXchangeおよびHUPO-PSI
- 質量分析データから情報を“限界まで搾り取る”ため、解析方法を改良し（これは生データ蓄積の目的の一つである）、その結果をデータベースの形で公開する
 - 解析方法の改良は、データベース構築の方法論の一部であり目的の一つでもある



jPOSTdbの特徴

- UniProtのようなタンパク質データベースとは異なる
 - 共発現情報が重要
- 基本的にはプロジェクト（生データを登録するときの単位）ごとの結果を表示する
- **オリジナルの解析（同定）結果は用いない**
 - オリジナルの解析結果は、たとえば論文査読のために必要
 - jPOSTでその生データを**再解析した結果**を用いる
 - データの品質を一定にするため

UniScoreを用いた再解析の例



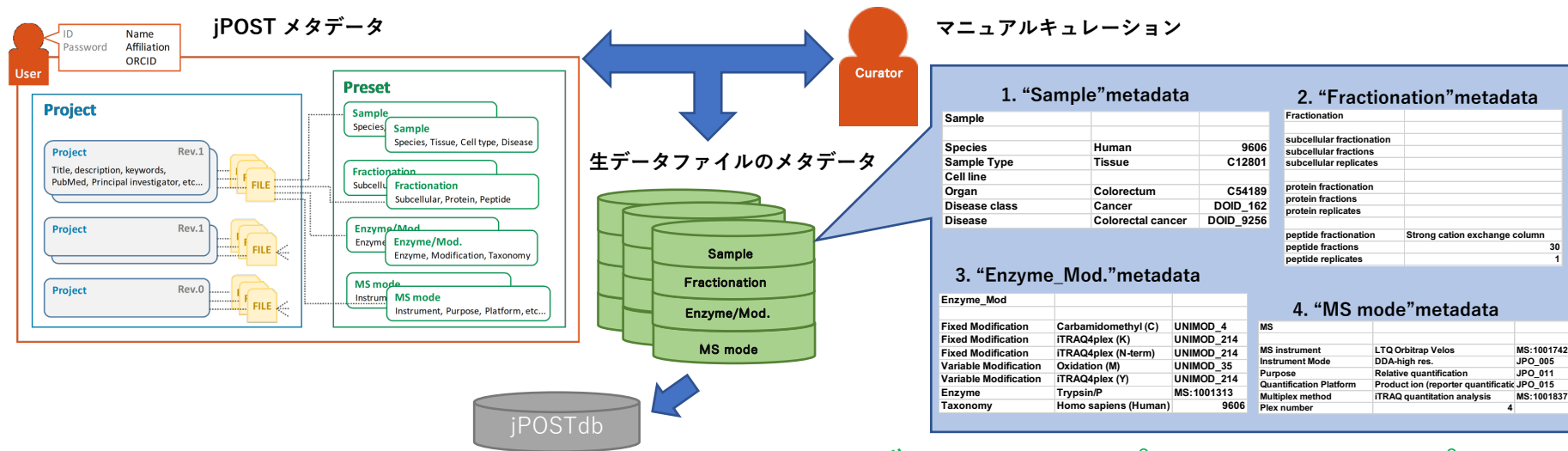
- 同定ペプチド数が65%程度増加（共にFDR < 1%）

トーゴーの日シンポジウム2021
ポスター35



メタデータの利用

- データベースは**データの再利用**が前提
- 質量分析の生データの**再利用にはメタデータが必要**
 - e.g.* variable modification (PTM, isobaric tag, etc.)
 - 再解析はいわば「(メタデータと) データを利用する**最初の例**」



トーゴーの日シンポジウム2017・ポスター58



メタデータの種類

マイクロアレイ (トランスクリプトミクス)

プロテオミクス

ガイドライン

フォーマット (具体的な項目)

MIAME (Minimum Information About a Microarray Experiment)

MIAPE (Minimum Information About a Proteomics Experiment)

MAGE-TAB (MicroArray Gene Expression TABular)

MAGE-TAB for Proteomics

IDF (Investigation Description Format)

SDRF (Sample and Data Relationship Format)

ProteomeXchange
メタデータ
(PRIDEなど)

SDRF-Proteomics



実験全体を記述 (実験デザインやプロトコル、文献情報など)

試料・測定機器の属性、実験の流れを示し、試料とデータファイルの関係を記述

jPOSTメタデータ

- フルに入力されていればSDRF相当の内容までカバーされ得るが、情報が足りないことが多い



メタデータの整備

- リポジトリに登録された情報では不十分なことが多い
 - 試料とデータファイルの関係
 - Technical replicate (特に injection replicate と呼ばれるもの) の処理を変えるため
 - その他
 - e.g. 分画・modification情報など

- 将来的に機械学習などで情報を抽出する(できる)としても、そのトレーニングデータの蓄積が必要
- 現時点では人間によるcurationが不可欠
 - リソース的に、jPOSTメンバーのみでcuration全工程を担当するのは困難
 - (“sustainableな開発を行うための最後のmissing piece”)
- **メタデータの補完・改善**
 - A) データ論文 (データ生産者に対するインセンティブ)
 - B) 論文調査によるメタデータの補完



A) データ論文による補完

Sample Article

Sample Article

Data Descriptor

<https://dx.doi.org/10.14888/jpdm.2020.xxxx>



Data for proteomic analysis of DNA-binding proteins

Taro Suzuki¹, Jack Smith¹, Sun-Moon Lake², Andrew Baker^{1*}

1. Laboratory of Proteome Research, National Institute of Biomedical Research, Osaka, Japan
2. Graduate School of Science, University of Proteomics, New York, NY, USA

ORCID: Andrew Baker: <https://orcid.org/0000-000X-XXXX-XXXX>

*Corresponding author.
E-mail address: abaker@nib.go.jp

© 2019 The Authors. This is an open access article under the CC BY license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords

Cell line, DNA-binding, Transcription factor

Dataset summary

Specific subject area	DNA-binding proteins and mass spectrometry
Data acquisition	Data-dependent acquisition acquired on Q-Exactive (Thermo)
Dataset repository	JPOST
Dataset identifiers	JPST100999

Abstract

Interactions between DNA and DNA-binding proteins are required for most cellular processes. Thus, it is clearly important to identify and quantify these interactions for understanding basic cellular mechanisms. We used an oligonucleotide probe to enrich specific DNA-binding proteins. As a result, we identified 512 proteins, including 120 novel DNA-binding protein candidates. The data accompanying this paper have been deposited to JPOST with identifier JPST100999.

1. Materials and Methods

In the present work, we provide the DNA-binding protein catalog obtained by proteomics experiments using the affinity purification with the oligonucleotide probe followed by LC/MS/MS analysis [1].

1.1. Samples

HeLa-S3 cells were grown in DMEM with 10% fetal bovine serum plus antibiotics in 10% CO₂ at 37 °C. For SILAC labeling, HeLa-S3 cells were cultured in DMEM supplemented with 10% dialyzed fetal bovine serum and either 28.0 mg/L normal isotopic abundance arginine and 48.7 mg/L normal isotopic abundance lysine (Light) or 28.0 mg/L arginine with six ¹³C and four ¹⁵N atoms and 48.7 mg/L lysine with six ¹³C and two ¹⁵N atoms (Heavy) [2]. Labeling efficiency was confirmed after five passages.

1.2. Sample pretreatment for MS analysis

Cell lysate was diluted with reaction buffer (20 mM HEPES pH 7.5, 150 mM NaCl, 0.1% Triton X100) to a final protein concentration of 4 mg/mL. Lysate was added to 250 µL of lysis buffer at a final concentration of 20 mM. Affinity purification of DNA-binding proteins was carried out at different concentrations of the oligonucleotide probe (1 nM – 10 µM) with SILAC-labeled "Light" lysate at room temperature with gentle shaking for 10 min. The control samples were prepared in the same way using the control probe with SILAC-labeled "Heavy" lysate. After the incubation, these samples were mixed, denatured by 5 M urea, reduced with DTT (5 mM final concentration), and alkylated with iodoacetamide (20 mM final concentration). After the alkylation step, the solution was substituted by digestion buffer by gel filtration followed

Received 19 June 2009; Received in revised form 7 October 2009; Accepted 15 October 2009; Published 26 October 2009

by digestion at 37 °C overnight with sequencing-grade trypsin at an enzyme/substrate ratio of 1:100. Peptides were dissolved in 50 µL of 2 M urea and 1% TFA and desalted using StageTips [3]. Each experiment was performed in duplicate.

1.3. MS analysis

All nanoflow LC/MS/MS experiments were performed on a Q Exactive mass spectrometer (Thermo Scientific). Data were acquired in data-dependent mode using Xcalibur software. The precursor ion scan MS spectra (m/z 350–1800) were acquired in the Orbitrap with 70,000 resolution after accumulation of ions to a 3 × 10⁴ target value. The twelve most intensive ions were isolated and fragmented in the collision cell by higher-energy collisional dissociation (HCD) with a maximum injection time of 120 ms and 35,000 resolution. In data-dependent LC/MS/MS experiments dynamic exclusion was used with 20-s exclusion duration. Typical mass spectrometric conditions were as follows: spray voltage, 2 kV; no sheath and auxiliary gas flow; heated capillary temperature, 250 °C; normalized HCD collision energy 25%. The MS/MS on selection threshold was set to 5 × 10³ counts. A 3.0 Da isolation width was chosen.

1.4. Data analysis

All raw files were processed by MaxQuant software suite (version 1.3.0.5) supported by the Andromeda search engine for peptide identification [4]. MaxQuant was used to score peptides for identification based on a search with an initial allowed mass deviation of the precursor ion of up to 7 ppm. The allowed fragment mass deviation was 20 ppm. Peak lists were searched against a concatenated forward and reversed version of the UniProt human database (release 2011.11) combined with 262 common contaminants was performed using the Andromeda search engine. Enzyme specificity was set as C-terminal to Arg and Lys with allowed cleavage proline bonds and a maximum of three missed cleavages. Carbamidomethylation of cysteine was set as a fixed modification and methionine oxidation as a variable modification. Peptides and proteins were accepted with a false discovery rate of 1.0%.

2. Data description

In the present work, 10 LC/MS/MS runs were conducted for 5 different samples (probe conc: 1 nM, 10 nM, 100 nM, 1 µM and 10 µM) with duplicate injections. The SILAC light-labeled lysate was used for the control, while the heavy-labeled lysate was used for the samples with the oligonucleotide probe. The details on the relationship between samples, the peak list files, MS raw files and the result files are described in Supplementary Table 1. As a result, 512 proteins including 120 novel DNA-binding protein candidates were identified.

Acknowledgments

This work was supported by Grants-in-Aid 87878787 to T. S. and 78787878 and 19191919 to A. B. from the Ministry of Science of Japan. We thank Jiro Sato for advice on MS analysis.

References

- [1] Suzuki, T.; Baker, A., *J. Proteome Res.* **2014**, *13* (2), 5461–70. DOI:10.1021/pr5012685.
- [2] Ong, S.; Blagoev, B.; Kratchmarova, I.; Kristensen, D. B.; Sørensen, H.; Pandey, A.; Mann, M., *Mol. Cell. Proteomics* **2002**, *1*, 376–86. DOI:10.1074/mcp.m200025-mcp2002.
- [3] Rappaport, J.; Mann, M.; Ishihama, Y., *Nat. Protoc.* **2007**, *2* (8), 1896–1906. DOI:10.1038/nprot.2007.261.
- [4] Cox, J.; Neuhauser, N.; Michalski, A.; Schelteme, R. A.; Olsen, J. V.; Mann, M., *J. Proteome Res.* **2011**, *10*, 1794–1805. DOI:10.1021/pr101065j.

Supporting Information

Supporting information is available online at <http://www.jpdm.jp/jpdm/>.

メタデータを記述するExcelファイル

	A	B	C	D	E	F	G	H	I
1	File	File Type							
2	20190930_AAS_oligo1nM01.raw	raw							
3	20190930_AAS_oligo1nM02.raw	raw							
4	20190930_AAS_oligo10nM01.raw	raw							
5	20190930_AAS_oligo10nM02.raw	raw							
6	20190930_AAS_oligo100nM01.raw	raw							
7	20190930_AAS_oligo100nM02.raw	raw							
8	20190930_AAS_oligo1uM01.raw	raw							
9	20190930_AAS_oligo1uM02.raw	raw							
10	20190930_AAS_oligo10uM01.raw	raw							
11	20190930_AAS_oligo10uM02.raw	raw							
12	msms.txt	peak							
13	msmsScans.txt	peak							
14	msScans.txt	peak							
15	20190930_AAS_oligo_DesBiotin(K)Sites.txt	result							
16	20190930_AAS_oligo_evidence.txt	result							
17	20190930_AAS_oligo_experimentalDesignTemplate.txt	result							
18	20190930_AAS_oligo_modificationSpecificPeptides.txt	result							
19	20190930_AAS_oligo_parameters.txt	result							
20	20190930_AAS_oligo_peptides.txt	result							
21	20190930_AAS_oligo_proteinGroups.txt	result							
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									
33									
34									
35									
36									
37									
38									
39									
40									
41									

- ショート論文
- 再解析を可能にする詳細なメタデータ

トーゴーの日シンポジウム2019・ポスター35,
同2020・ポスター53



A) JPDM data descriptor論文

- より**詳細・正確なメタデータ**
(Excel形式) をsupplementary
として(必ず)提出する



手作業での入力

- Sample 詳細
- 実験の条件
- 機器の詳細
- Fileの対応関係



- 投稿数増加のための広報を行っているが、まだ不十分
- (今後に期待)



<https://repository.jpostdb.org/jpdm-excel/>

jPOSTrepoに入力・公開済みメタデータをExcelファイルとして出力する機能



jPOSTリポジトリからExcelファイルを自動生成(・修正)

トーゴーの日シンポジウム2019・ポスター35,
同2020・ポスター53



B) 論文調査による補完

- 研究論文の記述内容から抽出
 - 通常の学術論文の場合
 - 論文をきちんと読解する必要 (次頁)
 - (将来的には機械学習で自動処理?)
 - どのファイルがどの試料か
 - 外部curatorが必要

A	B	C	D
1	メタデータ調査	プロジェクト名: JPST00XXXX	
2	論文およびJPOSTメタデータから、以下の項目について調査する (記述形式は任意。これらの内容を含んでいれば、このフォーマットで回答しなくてもよい。例えば文章による記述でもよい)	調査結果	根拠となる記述などをコピーする (根拠としては論文を優先し、JPOSTメタデータは根拠には極力用いないようにする)
3	(JPOSTメタデータについて)		
4			
5	試料作成について		
6			
7	試料 [JPOSTのメタデータとは項目が異なる]		
8	(記入例) 生物種		
9	試料タイプ (tissue, cell line, primary cell, それ以外) *		
10	どの臓器由来か、細胞株の場合は株名称 *		
11	疾病由来試料の場合は疾病名 *		
12	他・注意事項		
13	* ... 現在のところ、哺乳類培養細胞や疾病関連の場合にのみ対応しているため、それ以外の試料の場合は、これらの項目には記入不要。臓器などの名称は、↓のフォームから採用する。		
14	こちらの例のフォームに直接入力してもよい		
15			
16	分画		
17	(記入例) subcellular (オルガネラ)、タンパク質、ペプチドの各段階について、どのような手段で何個のフラクションに分画したか。またレプリケートは何個か		
18	他・注意事項		
19			
20	ラベル化定量している場合		
21	(記入例) TMTのチャンネル情報・SILACのラベル情報など (各生データファイルとの対応があればなおよい)		
22	他・注意事項		
23			
24	測定について		



B) 論文調査による補完

- トレーニングデータをよほど充実させないと、機械学習で判定はまだ無理

実例1

「Mitotic chromosomes were isolated after mixing equal amounts of WT cells cultured with lysine-8 and arginine-10 (heavy) and BAZ1B KO or BAZ1A/B double-KO cells cultured with lysine-0 and arginine-0 (light) (supplemental Fig. S5A), and their proteomes were analyzed by MS.」

【論文本文】



メタデータ「目的」：「相対定量」 (Relative quantification)

メタデータ「定量手法名」：「Product ion (reporter quantification) by SILAC (lysine-8 and arginine-10 [heavy] or lysine-0 and arginine-0 [light])」

実例2

「The digested peptides or enriched phosphopeptides were analyzed on a Q Exactive Hybrid Quadrupole-Orbitrap Mass Spectrometer or an Orbitrap Fusion Tribrid Mass Spectrometer.」 【論文本文】



ところがjPOSTには「LTQ Orbitrap Velos」と登録



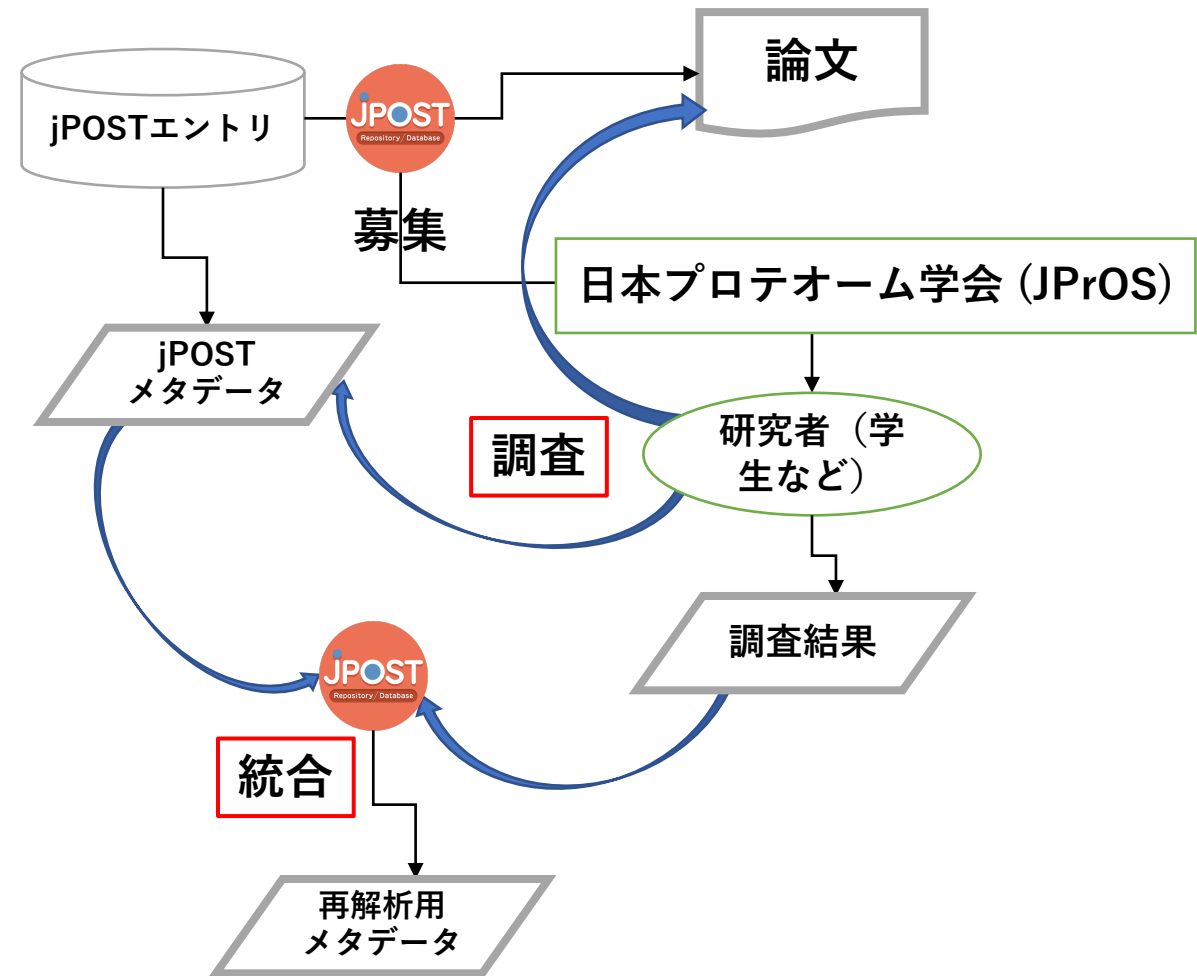
生データファイルを調査すると「LTQ Orbitrap Velos」と記録されていた



B) 論文調査 (curation) プロジェクト

• 論文調査 (curation) プロジェクト

- 日本プロテオーム学会 (JPrOS) を通じてcurator (謝金あり) を募集
- 若手を想定するが年齢制限はしない
- 2021年度応募者実績：
18名 (教員3名、D学生3名、M学生4名、学部生8名)



B) 論文調査 (curation) プロジェクト



- jPOST Literature Survey
- 日本語トップページ

クイックチャート

調査の進め方

調査進行表

担当論文報告票

調査結果提出

口座登録

各種ファイル

Top page in English



このページは何？

このページへのアクセス

この調査を初めて担当する人は:

jPOST Literature Survey

- 日本語トップページ

クイックチャート

調査の進め方

調査進行表

担当論文報告票

調査結果提出

口座登録

各種ファイル

Top page in English

作り直す必要があります。

○ 特に、各生データファイルが「どのような処理・測定をした試料由来なのか」の情報が重要です。

調査の進め方

○ 「論文プール」に置かれた論文を読み、回答用の Excel ブック「[metadata_extraction.xlsx](#)」の「調査項目」タブ B 列に書かれた各項目（下図赤矢印）の内容を抽出してください。

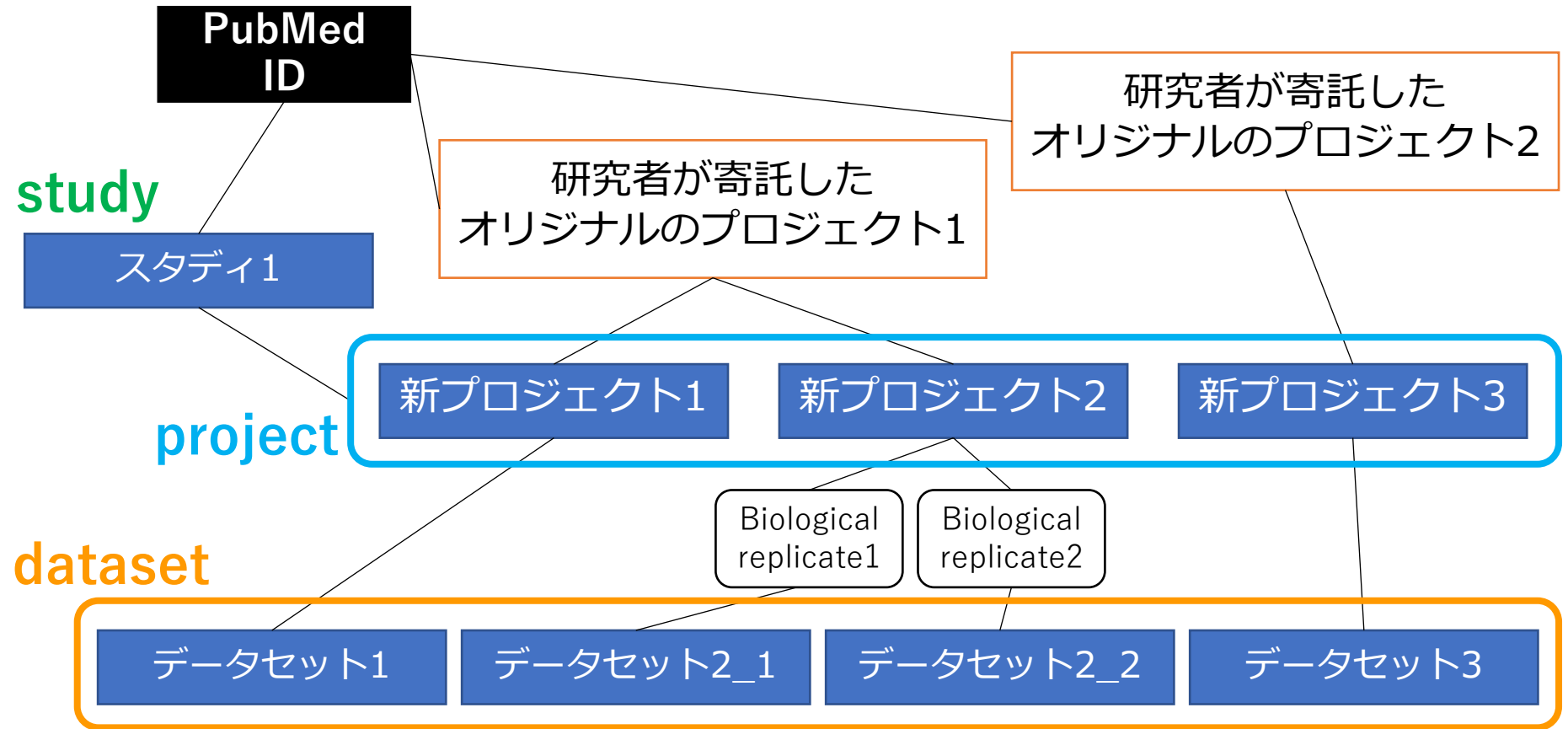
メタデータ調査			
A1			
1	メタデータ調査	プロジェクト名: JPST00XXXX	
2	論文およびJPOSTメタデータから、以下の項目について調査する（記述形式は、これらの内容を含んでいれば、このフォーマットで回答しなくてもよい。例えば文章による記述でも可（JPOSTメタデータについて）	調査結果	根拠となる記述などを示し、JPOSTメタデータと
3			
4			
5	試料作成について		
6			
7	試料 [JPOSTのメタデータとは項目が異なる]		
8	(記入例) 生物種		
9	試料タイプ (tissue, cell line, primary cell, それ以外) *		
10	どの臓器由来か、細胞株の場合は株名称 *		
11	疾病由来試料の場合は疾病名 *		
12	他・注意事項		
	* … 現在のところ、哺乳類培養細胞や疾病関連の場合にのみ対応しているため、それ以外の試料の場		

• 2022年度応募者募集中



jPOSTdbのデータセット（最新版）

- 定常的な解析がようやく軌道に乗った
- オリジナルのデータセット・メタデータに応じて分割したデータセットなどを右図のように **study**, **project**, **dataset** という用語で分類した
- 2022年9月末日現在で 37 study、561 project、657 dataset を収録（増加中）



まとめ

- jPOSTは

- I. ProteomeXchangeの一員としてデータ蓄積・公開
- II. 情報を限界まで抽出するために再解析手法を確立
- III. 再解析には詳細なメタデータが必要
 - I. メタデータの入力にはインセンティブが必要（ほとんどの場合、内容不足）
 - II. インセンティブを上げる手段…データジャーナルの発刊
 - III. 既存のメタデータを補完する方法として「論文の内容を調査しメタデータを補完修正する」ことも考えられる
- IV. メタデータを補完修正するために、論文調査プロジェクトを、日本プロテオーム学会の協力の下開始した
- V. 今後のsustainableなデータ追加を目指している



本研究・開発は科学技術振興機構(JST)・NBDC事業推進部による統合化推進プログラム予算によって実施した。