LLM を用いた BioSample データベース メタデータの品質向上

◯池田秀也 [1,2]、守屋勇樹 [1]、川島秀一 [1]、坊農秀雅 [1,2,3]、鄒兆南 [4]、沖真弥 [4]、大田達郎 [1,5,6]

1 情報・システム研究機構データサイエンス共同利用基盤施設ライフサイエンス統合データベースセンター、2 広島大学大学院統合生命科学研究科、
3 広島大学ゲノム編集イノベーションセンター、4 熊本大学生命資源研究・支援センター、5 千葉大学大学院医学研究院人工知能（AI）医学、6 千葉大学国際高等研究基幹

BioSample は、実験に用いられた生物学的サンプルのデータベースであり、サンプルの性質を記述したメタデータを蓄積している。公共実験データを再利用したい研究者は、興味のあるサンプルや実験の特徴の情報を使って BioSample を検索することができる。しかし、メタデータの記法の多くは投稿者の裁量に委ねられているため、同一の実験条件であっても投稿者によって異なる記述がされており、データの再利用性を低下させる要因となっている。これまでに、メタデータをオントロジーにマッピングすることで検索性を向上させる試みがなされてきたが、事前に定めたルールベースで行う手法では膨大な表記パターンをカバーしきれていなかった。そこで我々は、大規模言語モデル (LLM) を用いてメタデータを解釈し、オントロジーにマッピングするべき文字列を抽出することを試みている。

BioSample の課題

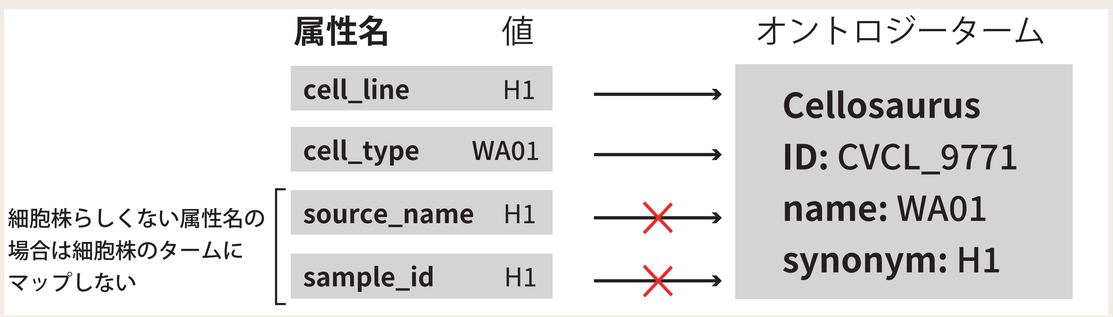
属性名とその値のペアの形でサンプルメタデータを記述
- 属性名が統一されておらず検索・管理しにくい
- 同じものを表すのにシノニムが使われており検索しにくい
- 同じ文字列でも違うことを意味しているかもしれない

4000 万レコード以上あり、手動での網羅的なキュレーションは困難

sample name	iPSC_1390G3
cell line	1390G3-526
cell type	iPSC
sex	female

source name	Induced pluripotent stem cell
biomaterial provider	parental cell line from Coriell
tissue	Induced pluripotent stem cell
derived from cell line	NA19193

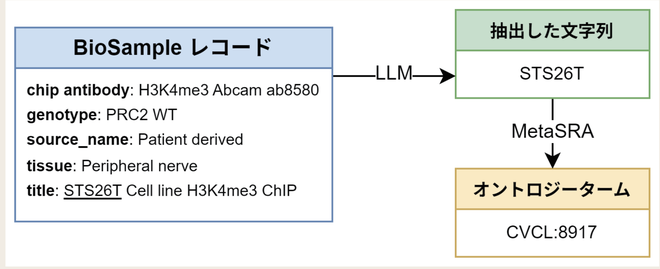
既存手法 MetaSRA [1]: ルールベースでオントロジーにマッピング
- 表記揺れを吸収し検索性を向上
- 属性名を限定することでミスマップを軽減



課題 : 事前に許容した属性の値しか使えない
→ **LLM による改善の可能性**

方針

- 目的のカテゴリー (細胞株など) を表す文字列の抽出を LLM で行う
- オントロジーへのマッピングは、ハルシネーションを避けるため従来法で行う



実行環境

- GPU: NVIDIA RTX 6000 Ada (memory 48 GB)
- model: Llama-3.1-8B-Instruct-fp16
  → 細胞株名・遺伝子名抽出タスクのプロンプトで、
  1200 ~ 1400 samples/h 程度の処理速度
- プロンプトなど `https://github.com/sh-ikeda/bsllmner/`
  Think-step-by-step 法を使用
  … プロンプトに “Think step by step.” というフレーズを加えると、
  解答に至るまでの経緯が同時に出力され、精度が上がる

細胞株名の抽出

ChIP-Atlas [2] のマニュアルキュレーションの成果を利用し、評価用のテストセットを作成

BioSample ID	BioSample Attributes	抽出するべき文字列	マップするべきオントロジーターム
SAMN13478071	**chip antibody**: H3K4me3 Abcam ab8580 **genotype**: PRC2 WT **source_name**: Patient derived **tissue**: Peripheral nerve **title**: STS26T Cell line H3K4me3 ChIP	STS26T	CVCL:8917
SAMN09917808	**cell strain**: SMMC-7721 **chip antibody**: H3K27ac (Active Motif, 39133, lot 31814008) **source_name**: epatocellular carcinoma **title**: SMMC-7721_H3K27ac_ChIPSeq_DMSO	SMMC-7721	CVCL:0534
SAMN02469158	**antibody**: anti p53 mouse monoclonal (DO-1) Sigma **condition**: pAPO **factor**: p53 **source_name**: diploid fibroblast **title**: Apoptosis IMR90 p53 r3	-	-

複数のプロンプトで精度を比較

	① 基本	② 分化した幹細胞株に注意	③幹細胞株そのものなら抽出することを強調
細胞株の説明	A cell line is a group of cells that are genetically identical and have been cultured in a laboratory setting. For example, HeLa, Jurkat, HEK293, etc. are names of commonly used cell lines.	A cell line is a group of cells that are genetically identical and have been cultured in a laboratory setting. For example, HeLa, Jurkat, HEK293, etc. are names of commonly used cell lines.	A cell line is a group of cells that are genetically identical and have been cultured in a laboratory setting. For example, HeLa, Jurkat, HEK293, etc. are names of commonly used cell lines.
入力の説明とタスクの指示	I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be a cell line, extract the cell line name from the input data.	I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be a cell line, extract the cell line name from the input data.	I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be a cell line, extract the cell line name from the input data.
出力形式の指定	Your output must be JSON format, like {"cell_line": "NAME"} . "NAME" is just a place holder. Replace this with a string you extract. When input sample data is not of a cell line, you are not supposed to extract any text from input. If you can not find a cell line name in input, your output is like {"cell_line": "None"} .	Your output must be JSON format, like {"cell_line": "NAME"} . "NAME" is just a place holder. Replace this with a string you extract. When input sample data is not of a cell line, you are not supposed to extract any text from input. If you can not find a cell line name in input, your output is like {"cell_line": "None"} . Note that some samples are cells differentiated from a stem cell line. In this case, the stem cell line name is mentioned in the metadata, but the sample is not the cell line itself. Therefore, the output must be {"cell_line": "None"} .	Your output must be in JSON format, like {"cell_line": "NAME"} . "NAME" is just a placeholder. Replace it with a string you have extracted. If the input sample data is not of a cell line, you should not extract any text from the input and your output must be like {"cell_line": "None"} . Note that some samples are cells differentiated from a stem cell line. In this case, the stem cell line name is mentioned in the metadata, but the sample is not the cell line itself. Therefore, the output must be {"cell_line": "None"} . Of course, if the sample is considered to be the stem cell line itself, extract the stem cell line name and include it in your output.

※ ② では出力が慎重になる傾向が見られたため、③ も試行した

従来法 MetaSRA のみで行う場合と比較して、precision や recall が改善された

method	precision	recall
direct MetaSRA	0.76	0.63
direct MetaSRA (属性名制限なし)	0.62	0.74
① → MetaSRA	0.87	**0.86**
② → MetaSRA	**0.93**	0.75
③ → MetaSRA	**0.92**	0.81

遺伝子名の抽出

ノックアウト (KO) あるいはノックダウン (KD) された遺伝子を抽出し、抽出した結果をマニュアルで評価

KO, KD の説明	A gene knockout (KO), also known as a gene deletion, involves completely eliminating the expression of a target gene by replacing it with a non-functional version, usually through homologous recombination in cells or animals. This results in a complete loss of the gene's function. Meanwhile, a gene knockdown (KD), also known as RNA interference (RNAi), involves reducing the expression of a target gene without completely eliminating it. KD is achieved by introducing small RNA molecules, siRNA or shRNA, that specifically bind to and degrade the messenger RNA (mRNA) of the target gene.
入力の説明とタスクの指示	I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to be had a gene knocked-out or knocked-down, extract the gene name from the input data.
出力形式の指定	Your output must be JSON format, like **{"knockout": ["NAME"], "knockdown": ["NAME"]}** . "NAME" is just a place holder. Replace this with the gene name you extract. When input sample data is not considered to be had a gene knocked-out or knocked-down, the value of "knockout" or "knockdown" of your output JSON must be an empty list. Note that multiple genes can be knocked out in one sample. In that case, include all of them in the list of the output JSON. For example, if you found "PRNP" and "MSTN" as knocked out genes, the value of the "knockout" attribute must be ["PRNP", "MSTN"].

→ 3723 サンプル中、473 サンプルで KO か KD 遺伝子を抽出、
うち 76% (358 サンプル) が正解
→ KO/KD 以外の発現調整手法が使われているような場合も、強引に
KO か KD として解釈して出力するケースが見られた
(例) dTAG を KD として出力
→ 手法を同時に出力するプロンプトに変更し、再評価中

展望

精度の改善
- 同名の異なる細胞株がオントロジーに存在するとき、いずれを採用するのが適切であるかの判定を LLM で行う

結果の提供
- オントロジーにマッピングした結果を RDF などで提供
- ChIP-Atlas などのアプリケーションでのサンプル検索にマッピング結果を利用

[1] Bernstein MN, Doan A, Dewey CN. MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. Bioinformatics. 2017 Sep 15;33 (18):2914–23.
[2] Zou Z, Ohta T, Oki S. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. Nucleic Acids Res. 2024 Jul 5;52(W1):W45–53.