大規模言語モデルを用いた BioSample データベースメタデータの品質向上

◯池田秀也 [1]、守屋勇樹 [1]、川島秀一 [1]、片山俊明 [1]、坊農秀雅 [1,2,3]、末竹裕貴 [4]、鄒兆南 [5]、沖真弥 [5]、大田達郎 [1,6,7]

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 株式会社 Sator、5 熊本大学生命資源研究・支援センター、6 千葉大学大学院医学研究院人工知能（AI）医学、7 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。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正確性に限界があった。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た。マニュアルキュレーションの結果を利用した評価の結果、LLM による抽出によって、従来のルールベースの手法と比較して精度と再現率を高めることができることを確認した。

BioSample レコードは 4500 万件を超えるため、効率的に処理するためには実行環境やプロンプトに工夫が必要となる。本発表ではそれらについても報告し、LLM によるキュレーションの結果を大規模データベースの利便性向上につなげるまでの道筋について議論する。

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
- 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
- 属性名が統一されておらず管理しにくい

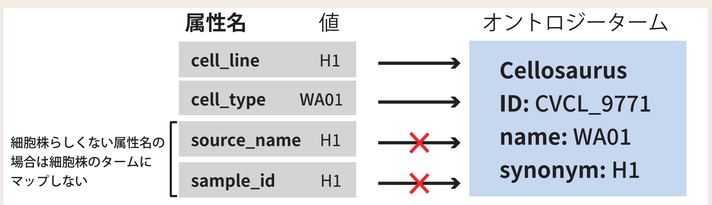
sample name	iPSC_1390G3
cell line	1390G3-526
cell type	iPSC
sex	female

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

属性名	サンプル数
cell line	4734
source_name	4360
cell type	1043
cell_line	635
well	263
isolate	192
strain	160
tissue	77
genotype	59
treatment	54

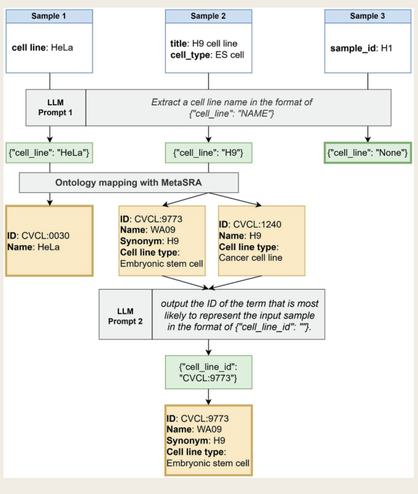
同じ文字列でも違うことを意味しているかもしれない
← 値に “H1” という文字列を含む属性の属性名 (一部)。“H1” は細胞株名、ヒストン、サンプル ID などを表す可能性がある

5000 万レコード近くあり、手動での網羅的なキュレーションは困難
既存手法 MetaSRA [1]: 文字列一致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



課題 : ミスマップを軽減するため、事前に許容した属性の値しか使っていない
→ **LLM による改善の可能性**

方針



サンプル情報を入力

目的のカテゴリーを表す文字列を LLM で抽出

文字列一致ベースでオントロジーマップ

同名の候補が得られた場合、妥当なものを LLM で選択

正解セット作成

ChIP-Atlas [2] のマニュアルキュレーションの成果を利用し、評価用の正解セットを作成
細胞株名を対象、オントロジーは Cellosaurus を使用
600 サンプルの正解セットを作成 (322 細胞株、278 非細胞株)

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	-	-

プロンプト

抽出のプロンプト

```
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.

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.

Your output must be JSON format, like {"cell_line": "NAME"} .
"NAME" is just a placeholder. 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"} .
Are you ready?
```

候補選択のプロンプト

```
I searched an ontology for the cell line, "{{cell_line}}".
I have found multiple terms which may represent the sample. Below are the annotations for each term. For each term, compare it with the input JSON of the sample and show your confidence score (a value between 0-1) about to what extent the entry represents the sample. In the comparison, consider the information such as:
 - Whether the term has a name or a synonym exactly matches the extracted cell line name, "{{cell_line}}".
 - Whether the term has disease or cell line type information which matches sample information.

Based on the confidence score, output the ID of the term that is most likely to represent the input sample in the format of {"cell_line_id": "<ID>"}.
If it is not clear which one is most likely from the given information, output {"cell_line_id": "not unique"}.
```

細胞株名の抽出による性能評価

従来法 MetaSRA と、LLM で細胞株名を抽出した結果を MetaSRA でマッピングする方法とで評価
モデルは Llama 3.1:70B Q4_0
LLM による抽出で従来法より良い結果が得られた

Pipeline	Accuracy	Coverage
MetaSRA	0.819	0.837
LLM-assisted	**0.929**	**0.933**

誤答の例
- 幹細胞株に由来する、分化した細胞
- 抽出するべき細胞株名を LLM が見逃し
- オントロジーでシノニムとして登録されていない表記
- オントロジーに登録のない細胞株

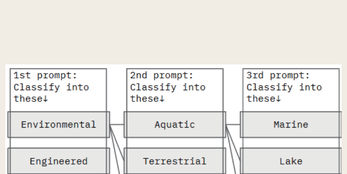
その他の試行例

- 実験的に操作された遺伝子とその手法の抽出

```
There are several experimental methods to modulate gene expression.

 Gene knockout (KO), also known as 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, 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.

 Gene overexpression refers to the process of increasing the expression of a specific gene beyond its normal levels in a cell. This is achieved by transfection of a plasmid carrying the gene of interest, transduction of viruses carrying the gene of interest, etc.

 I will input json formatted metadata of a sample for a biological experiment. If the sample is considered to have genes whose expression is experimentally modulated, extract the gene names from the input data and specify the modulation method.

 Your output must be in JSON format, like [{"gene": "GENE_NAME", "method": "METHOD_NAME"}] .
 "GENE_NAME" and "METHOD_NAME" are placeholders. Replace them with the gene name you extract and the modulation method name you specify, respectively.
 If the modulation method is either gene knockout, gene knockdown, or gene overexpression, the value of the "method" attribute must be "knockout", "knockdown", and "overexpression", respectively. Otherwise, the value of the "method" attribute must be the method name found in the input data.

 If the input sample data is not considered to have genes whose expression is modulated, your output JSON must be an empty list (namely, '[]').
 Note that multiple genes can be modulated in one sample. In this case, be sure to include all of them in the list of the output JSON. For example, if you find "PRNP" and "MSTN" as knocked out genes, your output must be [{"gene": "PRNP", "method": "knockout"}, {"gene": "MSTN", "method": "knockout"}].
 Note also that multiple gene modulation methods can be used for one sample. For example, you may find "ARID1A" as a knocked-out gene and "CHAF1A" as a gene treated with dTAG. In this case, your output must be [{"gene": "ARID1A", "method": "knockout"}, {"gene": "CHAF1A", "method": "dTAG"}].
```

→ 3723 サンプルを用いた評価で、8 割程度の正答率

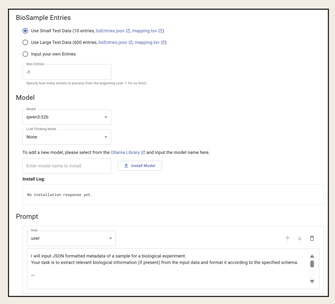
- 細菌分離源の段階的分類



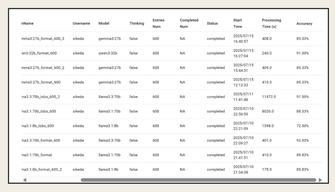
BacDive [3] のマニュアルキュレーションの再現を試みた
→ カテゴリによっては 0.9 程度の精度・再現率が出るが、"Host Body-Site" と "Host Body Product" の区別などは難しい

高速化のための実行条件検討

数千万件規模のデータを効率的に処理するための条件を検討



←検討用にブラウザから操作する GUI を作成
プロンプトやモデルなど条件を変えて繰り返し実行可能



← 実行速度や、正解セットを用いた評価結果をテーブルで一覧

プロンプトの検討
- Think-step-by-step 法 (思考過程を出力させる) は細胞株抽出のタスクにはあまり効果的でない
  - 出力トークン数が増え遅くなるデメリットが大きい
- 出力フォーマットを JSON スキーマで指定するのは効果的
  - 出力をコンパクト化し、処理時間を軽減。精度も維持

手法	モデル	入力数	実行時間 (秒)	正答率
TSBS	qwen3:32b	600	1288	89.00%
Format	qwen3:32b	600	93	91.33%

モデルの検討
→ qwen3:32b が軽量かつ十分な精度

※ 上表との差異は並列度の違いのため

モデル	入力数	実行時間 (秒)	正答率
phi4:14b	600	**127**	86.67%
llama3.1:8b	600	175	85.83%
deepseek-r1:32b	600	229	85.67%
qwen3:32b	600	**240**	**91.00%**
llama3.3:70b	600	401	**92.00%**
gemma3:27b	600	408	85.33%
llama3.1:70b	600	410	89.83%

Ollama のパラメータの検討

パラメータ	設定値	備考
OLLAMA_NUM_PARALLEL	16	並列化。これ以上大きくすると CPU 側のメモリが枯渇し、かえって遅くなる
OLLAMA_KV_CACHE_TYPE	q8_0	同じ予想をするときに使用されるキャッシュの設定。あまり効かず
OLLAMA_FLASH_ATTENTION	1	VRAM の使用量を抑える。あまり効かず
CUDA_VISIBLE_DEVICES	0,1	搭載している 2 GPU を両方使うための設定

→ 当初 400 samples/h くらいだったのが、
20000 samples/h くらいで処理できるようになった

論文

JOURNAL ARTICLE

Extraction of biological terms using large language models enhances the usability of metadata in the BioSample database

Shuya Ikeda, Zhaonan Zou, Hidemasa Bono, Yuki Moriya, Shuichi Kawashima, Toshiaki Katayama, Shinya Oki, Tazro Ohta

GigaScience, Volume 14, 2025, giaf070, https://doi.org/10.1093/gigascience/giaf070
Published: 23 June 2025 **Article history**

展望

- 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
- 拡張先を募集中
  - 今まではヒトのデータを中心に取り組んでいた。他の適用先を検討し段階的に拡張したい

[1] Bernstein MN et al. (2017). MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive.
[2] Zou Z et al. (2024). ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data.
[3] Schober I et al. (2025), BacDive in 2025: the core database for prokaryotic strain data.