大規模言語モデルを用いたサンプルメタデータの自動キュレーション

ライフサイエンス統合データベースセンター (DBCLS) 特任研究員
池田秀也

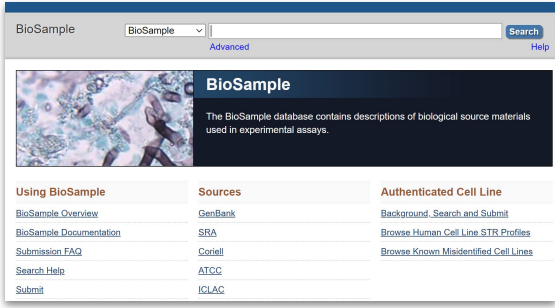
概要

- 実験サンプルデータベース BioSample は、投稿者による自由記述が許されており検索性が低い
- サンプルの属性を表すオントロジータームを割り当てる (= マップする) ことで表記揺れを解消し、検索性を向上させることができる
- 従来のルールベースアプローチでは、記述の中から何をオントロジーマップするべきかの判断が難しかった
- → 大規模言語モデル (LLM) により、記述中に出現する用語を意味的に判断し、オントロジーマップするべき語を抽出させることで、精度とカバー率の両立を実現できることを評価した
- **オントロジーマップを行うことにより、データがオントロジーで整理された情報につながり、ナレッジグラフの品質を高める**

背景: 公共データリポジトリによる実験データ共有の恩恵

- 実験技術の発展により、大量の実験データが産生されるように
- データを共有するための公共データリポジトリが開発される
  - Sequence Read Archive (SRA): 核酸配列情報
  - Gene Expression Omnibus (GEO): 遺伝子発現情報
  - :
  - 論文投稿時にリポジトリでの公開が必須
- 投稿された公共データは他の研究者が活用できるように

背景: BioSample データベース



- SRA や GEO などにデータを登録するときに、サンプルの情報を BioSample に登録する
  - メタデータ = 実験データについてのデータ
- サンプルの情報を、属性名と値のペアの形で記述
- 2025 年 6 月現在 4500 万件以上登録

属性名	値
age	3-5 days old
sex	female
strain	Harwich
temperature	24 degrees C
tissue	ovary

https://www.ncbi.nlm.nih.gov/biosample/SAMN00002080

課題: サンプル情報の不備が
データの再利用性を下げている

- あらゆる実験のパターンについて記述方法を事前に定めておくことは不可能なため、サンプル情報は自由記述が許されている
- 結果として同様の内容であっても登録者ごとに異なる記法で書かれている
    - 興味のある条件のサンプルを一括で検索するのが難しい
    - 実験間の条件を比較整理するのが難しい

潜在的に有用なデータが十分に活用されていない状況を生んでいる

実験データ共有により研究を良くしようという目的の妨げに

課題の具体例

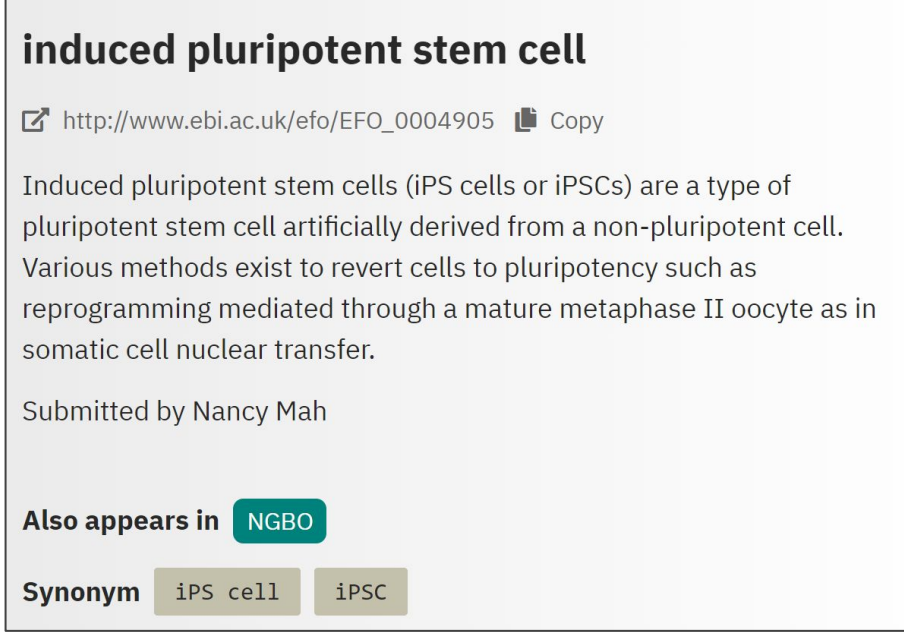
- 同義語
  - 同一のものを指す複数の呼称がある
  - “induced pluripotent stem cell“ で検索しても、
    登録者が “iPSC” と記述している場合は取得できない
- 属性名の不揃い
  - 同じ意味のことがどの属性に書かれているかは
    登録者依存

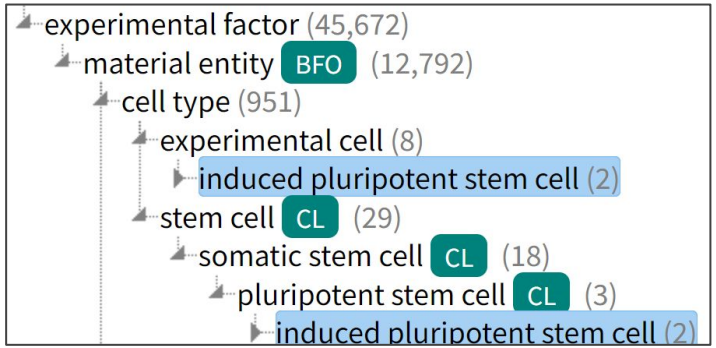
sample name	iPSC_1390G3
cell line	1390G3-526
cell type	iPSC
sex	female

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

課題の解決策:
オントロジーの利用

- 同一のものを指すのに複数の呼称がある
  → オントロジーを用いることで解決可能
- **同一概念は同一のオントロジータームで表現される**ので、オントロジータームによる検索で、該当するものを一括で取得できるようになる。
- 階層構造も検索に利用できる。上位概念で検索することで下位概念に属するものをまとめて取得。
- 実際には BioSample の登録時にオントロジーはあまり使われていない





既存研究: MetaSRA

- MetaSRA (Bernstein et al., 2017): BioSample のメタデータのうち、組織、疾患、細胞タイプ、細胞株などを自動でオントロジーにマップし、オントロジーによる検索を可能に
- 入力文字列を単語列に分解してオントロジーに曖昧一致検索
- 使用する属性を限定するなど、事前に定めたルールで精度を改善
  - “cell line”: “H1” -> H1 細胞株のタームにマップする
  - “sample ID”: “H1” -> H1 細胞株のタームにマップしない

既存研究の課題

事前に定めたルールでは誤りを除ききれない

- “cell line”, “cell type” 以外にも、細胞株を表すのに使われている属性名は多数存在
  - 無条件にこれらの属性名を許容すると誤ったマッピングが増える
    - “source name”: “H1” → cell line? sample ID?
  - 許容してよい属性名を単純に事前定義するのでは対応しきれない

Attribute name	# of Projects	# of Samples
cell line	667	11547
cell_line	435	96516
source_name	338	4972
isolate	88	2641
cell type	33	326
strain	20	1104
cell_type	7	507
tissue	6	442

属性値が “HEK293T” である属性の属性名(一部)

既存研究の課題

事前に定めたルールでは誤りを除ききれない

- “長文中に出現する語が、サンプルとどのような関係にあるのか判断が困難
  - 例) For DNA extraction, ten **seeds** of one genotype were grown in the greenhouse and a single, approximately 10 cm **leaf** was harvested from a 10-days-old seedling.

研究目的: LLM を活用したサンプル情報の品質の改善

- ChatGPT の登場以降、Large Language Models (LLMs) が注目されている
  - 大規模なデータセットを用いた学習に基づく、高度な文脈把握性能
  - 学術論文などから情報を抽出し構造化するなどの応用が報告されている
- BioSample の課題の解決に役立てられる
  - MetaSRA のような従来法で必要だった事前のルール設定に基づかずに必要十分な情報を抽出
  - 抽出された文字列を、文字列マッチでオントロジーへマッピング

手法: LLM 実行環境

- 最終的に数千万件規模のレコードを処理したい。ChatGPT などの商用のサービスに依存するのは非現実的
  - → Llama 3.1:70B Q4_0 を使用
    - Meta 社開発のオープンソース LLM
    - 試行時時点では、オープンソース LLM の中では性能が特に優れているとされていた
      - 2025 年 6 月時点で、同規模以下の良いモデルは、Llama3.3:70B, DeepSeek-R1:70B, Gemma 3:27B, Qwen3:32B etc.
- Ollama
  - ローカル LLM を簡単に実行できるプラットフォーム
- NVIDIA RTX 6000 Ada GPU (48 GB VRAM)
  - より軽量のモデルならば、一般的なゲーミング PC (VRAM 6 GB 程度) で十分動く

結果: 細胞株名抽出とオントロジーマッピングの評価のための正解セット作成

- エピゲノム統合データベース ChIP-Atlas でマニュアルキュレーションされた結果を利用して正解セットを作成
  - ChIP-Atlas では、発生生物学の専門家のマニュアル作業により、サンプルの由来する組織や細胞株名を独自の統一語彙に名寄せしている
  - 性能評価用のセットでは対象を細胞株名に絞った。オントロジーは Cellosaurus を使用
    - 「あるサンプルがその細胞株であるかどうか」はイエス・ノーで確実に答えられるはずなので、曖昧性がなく定量的な評価には適切なため
      - 組織名だと、「右心房」と記述されているときに「心房」を抽出するのは誤りなのか？といった曖昧性がある

結果: 細胞株名抽出とオントロジーマッピングの評価のための正解セット作成

- 600 サンプルの正解セットを作成した
  - 同じプロジェクトに属さない、かつ ChIP-Atlas での分類が異なるもの
  - 322 は細胞株。278 は細胞株ではなく、何も抽出しないことが正解。

LLM による細胞株名抽出: プロンプト

細胞株の定義を与えたうえで、JSON 形式で入力される BioSample データから細胞株名を抽出し、{“cell_line”: “NAME”} の形式で出力することを指示

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 place holder. Replace this with a string you extract.

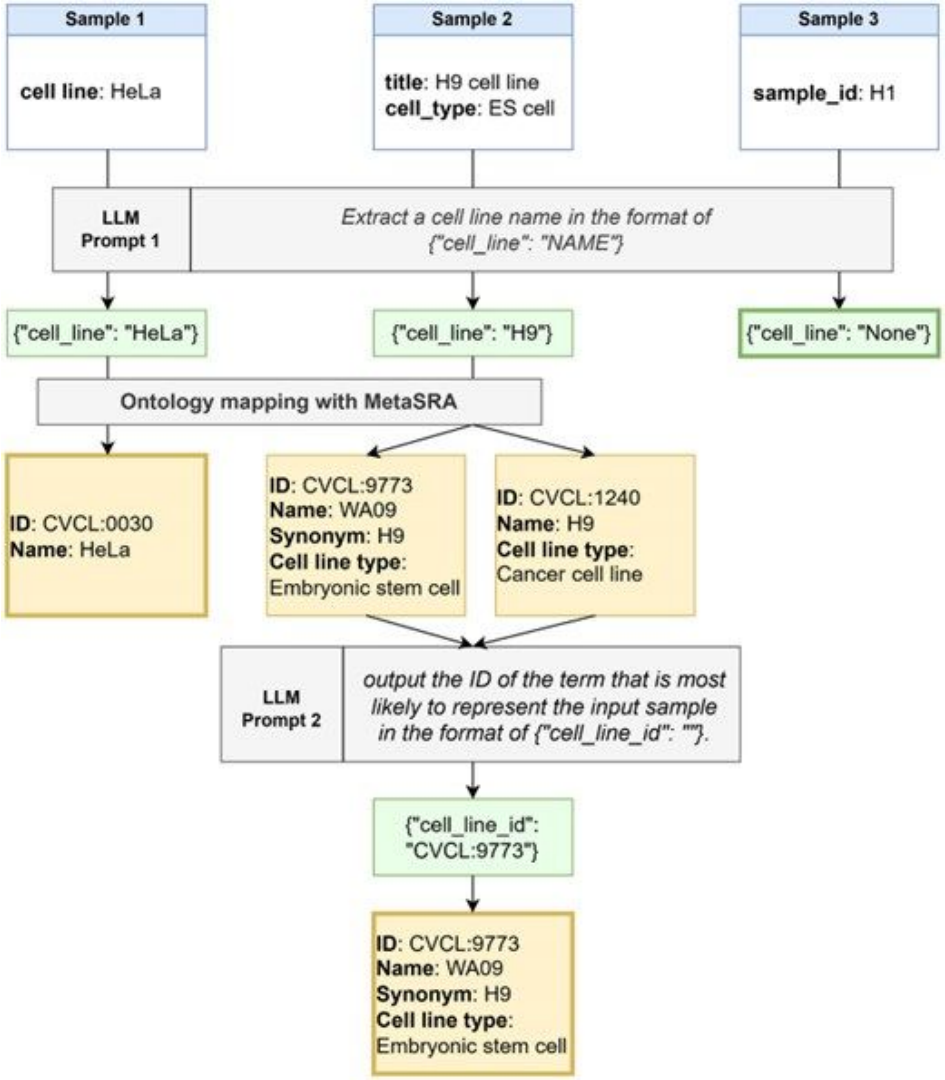
When input sample data is not of a cell line, you are not supposed to extract any text from input.
If you can not find a cell line name in input, your output is like {"cell_line": "None"} .
Are you ready?

LLM による細胞株名選択: プロンプト

抽出された文字列をオントロジーにマッピングした結果、同名の複数の細胞株が得られてしまうことがある。
LLM にオントロジーの情報を与え、BioSample の記述と照合することで判定させたい。

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"}.



Ikeda et al., 2025, Fig. 2

結果: 正解セットを用いた性能評価

- 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
- 従来法を改善した結果が得られ、LLM による抽出の有用性を確認
  - LLM は、非細胞株文字列を細胞株名として誤認することを少なく抑えつつ、細胞株名を表す文字列を的確に認識する

Pipeline	Accuracy	Coverage
MetaSRA	0.818923	0.836667
LLM-assisted	0.92869	0.933333

誤答パターン

- Derivation: 幹細胞株に由来する、分化した細胞の場合。細胞株名が記述されているが、その細胞株そのものではないので(今回は)抽出しないのが正解
- Overlook: LLM が、抽出するべき細胞株名を見逃した
- Non-canonical name: BioSample の記述に使われている細胞株名が、オントロジー側でシノニムとして認められていないなど
- Selection failure: 複数候補からの選択を間違えた
- Wrong extraction: 正答となるマッピングが可能な文字列を抽出できていない。”Jurkat E6-1 cell line” から “Jurkat” だけ抽出した場合など。
- Ontology insufficiency: 抽出した文字列は細胞株名だが、その名前を持つ Cellosaurus タームは同名の他人しかない

Derivation	12
Overlook	8
Non-canonical name	8
Selection failure	7
Wrong extraction	4
Ontology insufficiency	2
Total	41

応用例: 遺伝子名の抽出

- 細胞株名抽出タスクで、LLM を用いた概念抽出の有用性を確認した
- 既存のマニュアルキュレーションが及んでいない領域に適用し、アプリケーションの利便性向上を目指す
- → ChIP-Atlas でマニュアルキュレーションされていない、実験的に発現を調整された遺伝子名とその調整手法の抽出を実行
    - 表記揺れが大きく、アルゴリズミックな抽出が難しい分野である
        - ノックアウトを表す文字列: knockout, KO, -/-, deletion など
        - ノックダウンを表す文字列: knockdown, KD, shRNA, siRNA, si(Target gene name), RNAi など

LLM による遺伝子名抽出: プロンプト

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 trasfection 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"}].

Are you ready?

LLM による遺伝子名抽出: プロンプト

- 遺伝子ノックアウト、ノックダウン、過剰発現の定義を与える
- JSON 形式で入力される BioSample データから遺伝子名と手法名を抽出し、

```
[
  {"gene": "ARID1A", "method": "knockout"},
  {"gene": "CHAF1A", "method": "dTAG"}
]
```

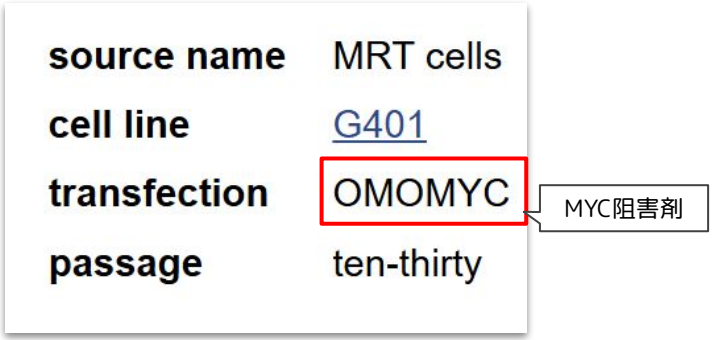
  のような形式で出力することを指示
  - 複数の遺伝子が対象となっている場合があるので配列型で出力する
  - 遺伝子名と手法名の組の形で抽出する
  - knockout, knockdown, overexpression についてはこれらの語に名寄せする
  - それ以外の場合は書いてある通りに抽出する

LLM による遺伝子名抽出: 結果

- 3723 サンプル中 600 サンプルから一つ以上の遺伝子を抽出
- 一つ一つ目視で正誤を判定
  - 既存のマニュアルキュレーション結果が存在せず正解セットの構築は難しいため、カバレッジは計算せず
  - 今回のプロンプトでは考慮に入れておらず正解が未定義であるものは評価の対象外
    - 融合遺伝子など
- 遺伝子名と手法名の組の抽出の正答率は 465/579 = 0.803
  - 遺伝子名抽出の正答率は 542/592 = 0.916
  - 手法名抽出の正答率は 497/587 = 0.847

LLM による遺伝子名抽出: 結果

- 今回のプロンプトでは対応が難しい複雑なパターン
  - 阻害剤だけが言及されており、発現が影響を受けている遺伝子名が直接記述されていない
  - 内在性の遺伝子が KO されているが、外部から同じ遺伝子に変異を加えたものやエピトープタグを付けたものが導入されている
    - 確かにその遺伝子は KO されているが、結果として発現は残っているので、ユーザーが「その遺伝子が KO されたサンプル」を検索したときにこれが取れてしまうのは好ましくない結果につながる。



まとめ・展望

- 細胞株名の抽出タスクによる評価で、BioSample のメタデータ品質改善における LLM の有効性を確認
- 単純には対応が難しい複雑なケースの存在も明らかになった
  - 分化した細胞株など
  - プロンプトの工夫で対応できるケースもある
  - モデルが改良されることで改善される可能性も高い
- テストを行った環境での実行速度は ~400 件/h
  - 全 BioSample レコードは 4500 万件。システムの高速化を図り、大規模に実行した結果を提供したい
    - RDF で提供
    - 検索する UI を提供

- 論文
  - Extraction of biological terms using large language models enhances the usability of metadata in the BioSample database
- ソースコード
  - LLM curation: https://github.com/sh-ikeda/bsllmner
  - Ontology mapping: https://github.com/sh-ikeda/MetaSRA-pipeline
- 評価用正解セット
  - https://zenodo.org/records/14881142