

生命科学分野におけるナレッジグラフとオントロジー

AJACS 「生命科学分野におけるナレッジグラフとオントロジーを知って・学んで・使う」

大田達郎

千葉大学 国際高等研究基幹/大学院医学研究院 人工知能医学/千葉大学データサイエンスコア

github.com/inutano

2025-06-26

今日の講習会のテーマ

- ナレッジグラフとオントロジーを活用した
生命科学データの整理・統合・再利用
- 「生命科学データを高度に扱う」ための第一歩

なぜ今ナレッジグラフなのか？

背景: 生命科学データの特殊性

- ゲノム・プロテオーム・メタボローム・イメージング・電子カルテ・文献・臨床試験.....
→ データの形式・粒度・意味がバラバラで統合が困難
- 例：
 - **PubMed**：3,800万件超の文献データ、キーワードや抄録を含む非構造テキスト
 - **GenBank**：ゲノム・cDNA 配列と注釈情報（配列 + 機能情報）
 - **UniProt**：タンパク質配列、構造、機能、発現、翻訳後修飾などを統合
 - **ChEMBL**：化合物-標的-薬理活性などの関係データベース
- 量も多く、種類も多い。意味の接着剤が必要

ナレッジグラフへの注目

- データの量が多いだけでは意味がない
→ 「どうつながっているか」がわからなければ活用できない
- AIやデータサイエンスが進展する一方で、
学習・解析の精度は構造化された知識の整備に依存
- ChatGPTなどの大規模言語モデルも、
高品質な構造化データとの統合が鍵に

FAIR原則

- データの持続的利活用を支える国際的な原則
- 2016年に発表 (Wilkinson et al., *Scientific Data*, 2016)
- 4つの基本理念：

項目	意味
Findable	メタデータとデータは識別可能で検索可能である
Accessible	明確なアクセス手段で誰でも到達できる
Interoperable	他のデータやツールと組み合わせられる形式で記述されている
Reusable	十分な文脈情報とライセンスにより再利用可能である

- ナレッジグラフとオントロジーは、FAIR原則を技術的に支える鍵である

本日の講演内容

1. 生命科学分野における知識グラフとRDFポータル (川島さん)
公開データを“つなぐ”基盤の基本と活用法
2. **TogID** による生命科学データベースの横断利用 (池田さん)
異なるIDを自動で橋渡し
3. 生命科学分野におけるオントロジーの活用事例: **NANDO** と **PubCaseFinder** (藤原さん)
疾患・症状とデータをつなげる事例紹介
4. 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 (池田さん)
AIで自由記述メタデータを整える最前線

本日のゴール

- ナレッジグラフやオントロジーの役割を知る
- 実際に使えるツールを知る
- 「自分の研究にどう使えるか」のヒントを持ち帰る
- 生命科学データベースという研究分野の現場を覗き込む