

2025/1/30

ライフサイエンスにおける オープンアクセスの歴史

川島秀一 (ライフサイエンス統合データベースセンター)

オープンアクセスの重要性

- ・ オープンアクセス（OA: Open Access）とは、学術研究の成果を、無料で誰でもアクセスできるようにする公開モデル
- ・ ライフサイエンス分野では、オープンアクセスが特に重要
 - ・ 研究の加速とイノベーションの促進
 - ・ 例) ヒトゲノムプロジェクト、AlphaFold、
- ・ 疾病研究・公衆衛生への貢献
 - ・ SARS-CoV-2ウイルスのゲノム配列が、INSDCから公開されたことで、短期間のワクチン開発が実現した
- ・ 再現性の向上と科学の透明性
- ・ 産業・社会への応用
- ・ 論文を発表する際には、配列データはINSDCに、タンパク質構造データはPDBに、など、公共データベースへの登録が義務付けされている場合もよくある

生命科学分野におけるオープンデータの歴史

年	データベース／出来事	内容	組織	メディア
1879	Index Medicus	文献	軍医総監局図書館	書籍
1964	MEDLARS	文献	NLM	ファイル
1965	Atlas of Protein Sequence and Structure	タンパク配列	ジョージタウン大学	書籍
1971	MEDLINE	文献	NLM	ファイル
1971	PDB	タンパク構造	ブルックヘブン国立研究所	磁気テープ
1982	GenBank	DNA配列	ロス・アラモス国立研究所	磁気テープ (1989～インターネット)
1984	PIR	タンパク配列	ジョージタウン大学	
1986	SWISS-PROT	タンパク配列	ジュネーブ大学	磁気テープ
1990	ヒトゲノム計画 発足			
1995	KEGG	パスウェイ	京都大学	WWW
2000	クリントン・プリア声明			
2002	UniProt	タンパク配列	SIB	WWW
2004	PubChem	化学化合物	NCBI	WWW
2008	1000ゲノム計画 発足			
2022	ヒトゲノムの完全解読(T2T)			

生命科学分野におけるデータベース

- 大量のデータベースが公開されている
(2024/3現在)
- NAR誌のデータベースカタログ：2,105
- NBDC integbio DBカタログ：2,566
- Dabase Commons：2,389
- FAIR sharing.org：2,102
- Bioinformatics誌：313 論文
- Database誌：1,549 論文
- Scientific data：3,921 データセット
- Giga Dataに：2,462 データセット
- データベースの種類
 - NAR誌データベースカタログの分類
 - 15カテゴリ40サブカテゴリ
- ゲノムプロジェクトの数
 - 生物種数で、506,976 (GOLDデータベース)

ヒトゲノム計画とオープンデータの黎明期

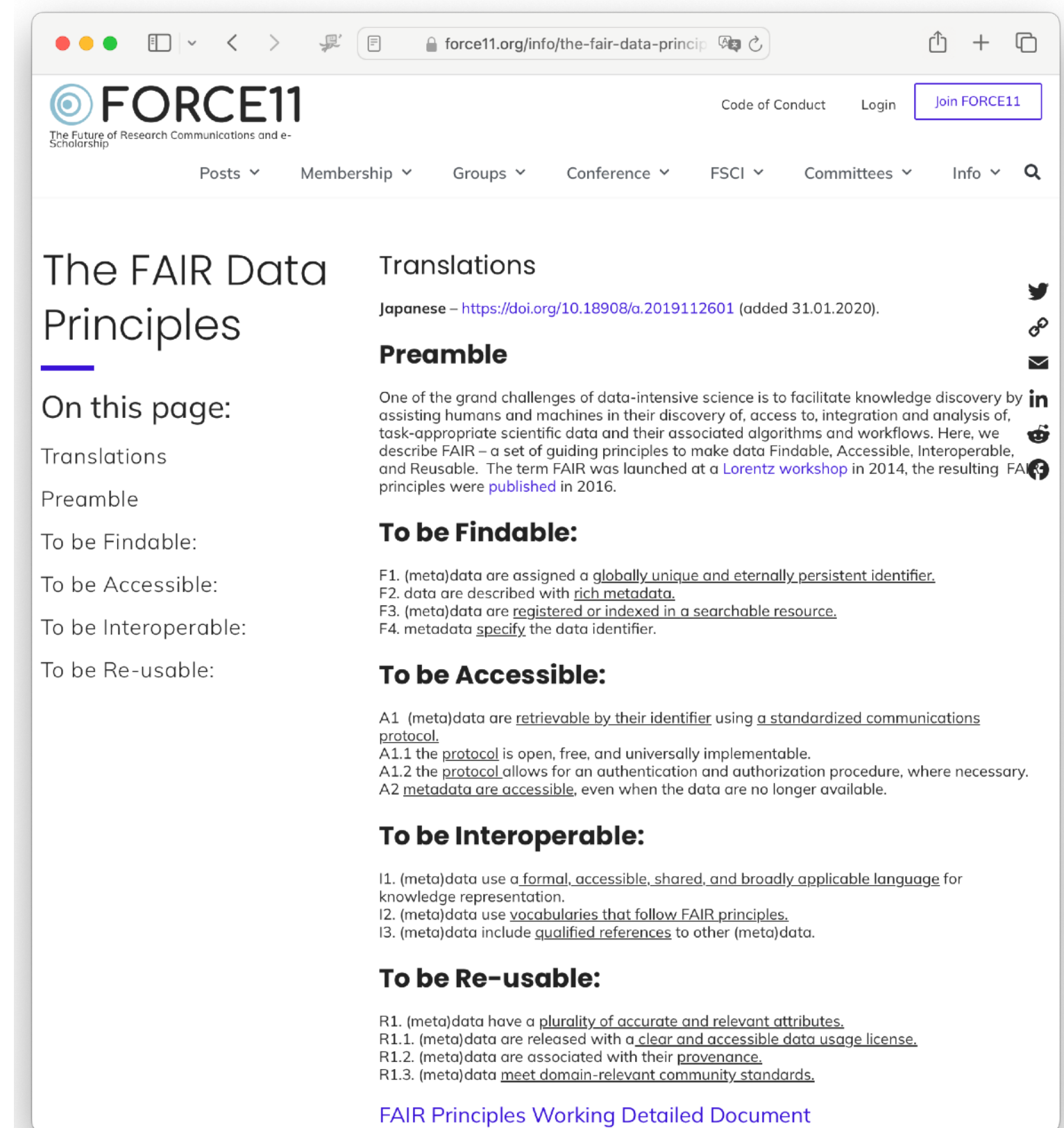
- ・ ヒトゲノム計画の概要と目標
 - ・ ヒトの全ゲノム（約30億塩基対）を解読することを目的とした国際的な研究プロジェクト
- ・ Celera Genomics vs 公的プロジェクト（特許戦争）
- ・ バミューダ原則: ヒトゲノムプロジェクトにおけるデータ共有の基礎
- ・ データの迅速な公開と国際協力

バミューダ原則

- ・ ゲノムデータは、速やかに公開されるべき
 - ・ シーケンシングされたDNA配列は、24時間以内に公開データベース（GenBank, EMBL, DDBJ）に登録することが推奨された。
 - ・ 研究者や企業によるデータの独占を防ぎ、オープンサイエンスを促進。
- ・ すべての研究者がデータを自由に利用できるべき
 - ・ ゲノム配列データは、誰でも無料でアクセス可能にすることが原則。
- ・ データ共有が科学の進歩を加速する
 - ・ 競争よりも国際協力を優先し、ゲノム情報を多くの研究者が利用できる環境を整える。

FAIR原則

- データ集約型科学を推進するために必要と考えられる原則をまとめたもの。
- FAIRは次の4原則の頭文字
 - Findable（見つけられる）
 - Accessible（アクセスできる）
 - Interoperable（相互運用できる）
 - Reusable（再利用できる）
- 例えば、Horizon Europe（欧州の科学研究資金助成計画：予算規模は2021年～2027年で955億€）では、研究データの管理と公開に関する要件として、FAIR原則に従うことが明確に規定されている。
- European Open Science Cloud (EOSC) は、研究データの保存、管理、共有、再利用を促進するためのプラットフォームであるが、FAIR原則の実践を支える役割を担っている。



国際社会における対立とオープンデータの課題

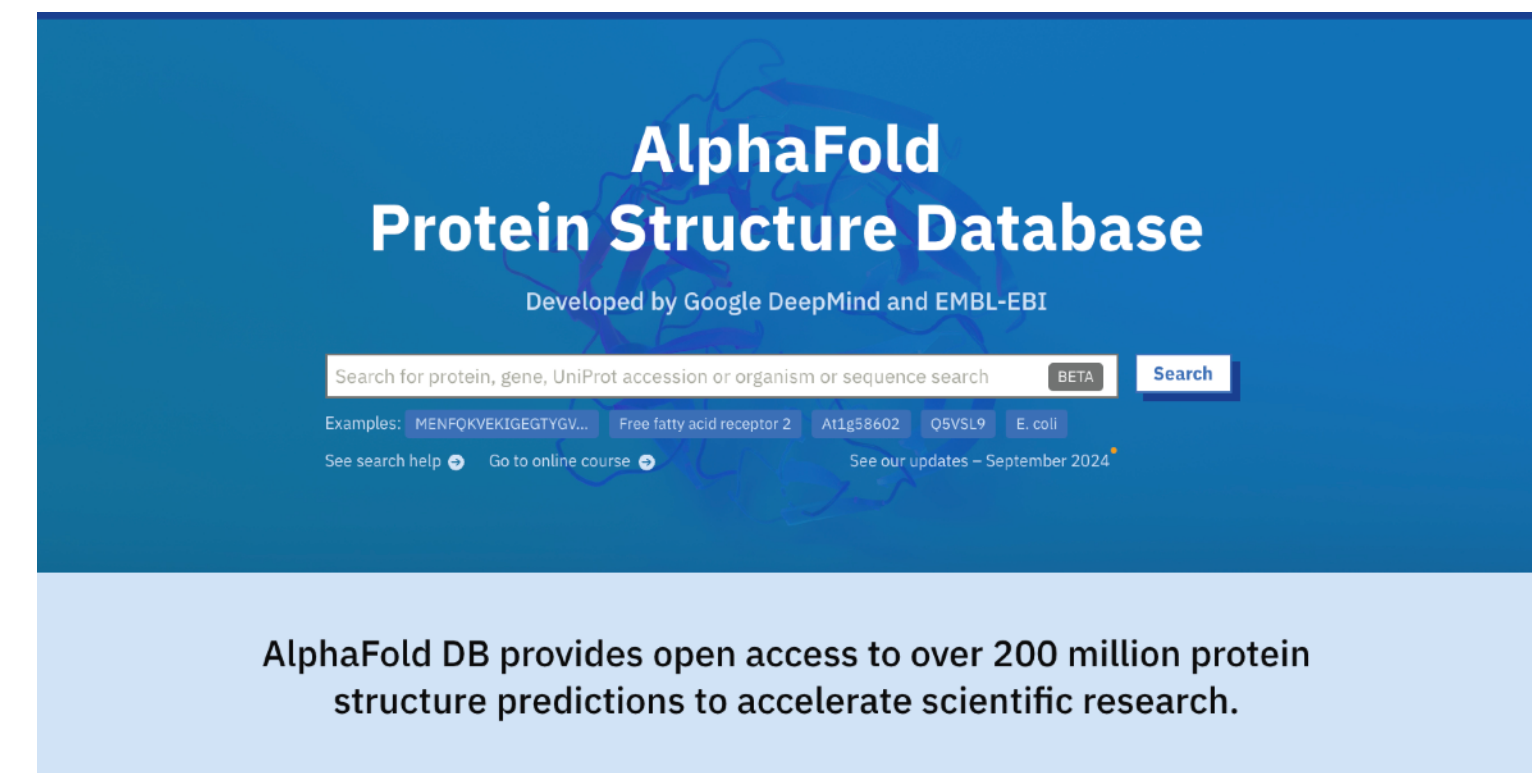
- ・ 生物多様性条約（CBD） 締約国会議（COP）とDSI問題
 - ・ 生物多様性条約（CBD）は、地球上の生物多様性を保全し、その持続可能な利用を促進するための国際条約
 - ・ DSI（Digital Sequence Information）は、生物から取得された遺伝情報をデジタル化したデータ
 - ・ ゲノム配列・遺伝子配列・タンパク質配列・メタゲノムデータ等
- ・ 遺伝資源は採集した国の許可が必要だが、デジタル情報となった遺伝情報はオープンにアクセスできるため、どの国も研究や商業利用が可能
- ・ 途上国はDSIの利用による利益を公平に享受できていない
 - ・ 南北問題となっている

国際社会における対立とオープンデータの課題

立場	主な主張
先進国（米国、欧州、日本など）	DSIは研究とイノベーションのために自由にアクセス可能であるべき。 制限を加えると、ワクチン開発や医療研究に悪影響を及ぼす。 DSIを物理的な遺伝資源と同様に扱うのは非現実的。
途上国（アフリカ、南米、アジアの一部）	生物多様性が豊富な国は、DSIの商業利用から利益を受けべき。 先進国の企業や研究機関は無料でDSIを利用し、新薬や農業技術を開発しているが、原産国には利益が還元されていない。 DSI利用者は利益を共有する仕組みを作るべき。

AIとビッグデータがもたらす可能性

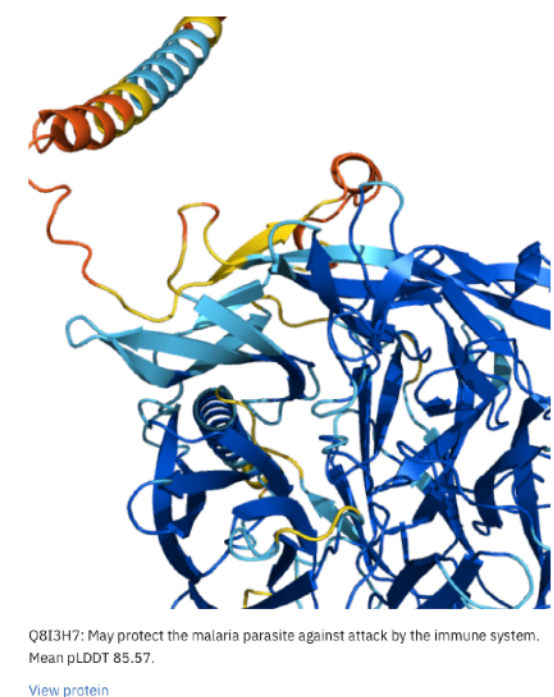
- AlphaFold: 人工知能（AI）を用いてタンパク質の立体構造を高精度で予測するモデル
- タンパク質の立体構造を決定するには、従来、X線結晶構造解析、NMRなど非常に時間がかかる実験手法に依存していた。
- AI技術により計算機で高精度に立体構造を予測できるようになった
 - 開発者の Demis Hassabis と John Jumper は、2024年ノーベル賞を受賞
- AlphaFoldの成功には様々な理由があげられるが、PDBやUniProtなど、科学者が何十年もの間、**高品質なタンパク質データを整備し、オープンに公開**していたことは不可欠の要素である
- 同様に、創薬分野、パーソナライズ医療、感染症の予測とワクチン開発など様々なAI応用が期待される分野で、高品質なオープンデータが必要になる



Background

AlphaFold is an AI system developed by Google DeepMind that predicts a protein's 3D structure from its amino acid sequence. It regularly achieves accuracy competitive with experiment.

Google DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) have partnered to create AlphaFold DB to make these predictions freely available to the scientific community. The latest database release contains over 200 million entries, providing broad coverage of UniProt (the standard repository of protein sequences and annotations). We provide individual downloads for the human proteome and for the proteomes of 47 other key organisms important in research and global health. We also provide a download for the manually curated subset of UniProt (Swiss-Prot).



まとめ

- ・ ライフサイエンスにおけるオープンアクセスの重要性
 - ・ 数多くの公共データベースに支えられている
- ・ データ共有を促進する原則
 - ・ バミューダ原則、FAIR原則など
- ・ 一方で、オープンアクセスと対立するような問題もある
 - ・ DSI（デジタル配列情報）問題とCOP会議
 - ・ 特許問題
- ・ 今後のAI応用の観点からも、今後も高品質なオープンデータは必要不可欠