82. 生命科学研究データ再利用性担保プラットフォーム構築

水口　賢司

大阪大学 蛋白質研究所 計算生物学研究室

Key words：FAIR 原則，データ共有，データ再利用性担保，データウェアハウス，機械学習

緒　言

生命科学分野では、研究成果であるデータを適切にオープンにすることが求められている。このような動きは、オープンサイエンス推進における国内外の議論でも取り上げられ、FAIR 原則である「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）」がしばしば言及される［1］。

単一の実験結果であれば、データを公共リポジトリに登録することが有効である。タンパク質立体構造（worldwide Protein Data Bank）や塩基配列（DDBJ、NCBI、ENA／EBI）など、古くから確立された国際協調によるデータベース構築が、最近はプロテオームデータなど他の分野にも広がりつつある。しかし、近年の生命科学研究においては、遺伝子機能とタンパク質間相互作用など、異なった複数種類のデータを組み合わせて解析する事例が増えている。FAIR 原則に則って、研究データの再利用性を担保し共有・活用を促進するためには、異種の生物学データを統合して、共有・活用できる形に整理する必要がある。その作業には、一般に多大な労力が求められるという問題があった。

一方、こうして整理されたデータから、新たな解析目的に応じて柔軟に部分データを再構築できれば、研究データの再利用性を超えた利活用が期待できる。様々な研究分野で人工知能（artificial intelligence：AI）の活用が進んでいるが、生命科学分野における AI・機械学習の応用においては、学習（正解）データ作成のコストの大きさが深刻な課題と認識されている。先に述べた複数種類の研究成果データが統合・公開されたシステムから、コンピュータ解析に適した形式で必要な部分データを自由に生成できれば、機械学習モデル構築のボトルネックが解消され、生命科学分野のデータ駆動型研究に変革がもたらされると考えられる。

私たちは 10 年以上にわたり、創薬支援データウェアハウス TargetMine の研究開発を続けてきた［2，3］。TargetMine は、30 以上の公共データベースから得られる遺伝子、タンパク質、化合物、疾患などのデータを統合し、各項目間の関連性を効率的に探索する機能を提供する。現在では、創薬支援だけでなく、基礎生物学分野のオミックスデータ解析などにも広く利用されている。本研究では、TargetMine の開発経験を生かして、生命科学分野の研究データ再利用性担保及び学習データ自動生成プラットフォームを構築することを目的とした。異種データの組み合わせを取得する検索機能を実装し、カスタマイズした多種類のデータの統合とオンライン共有を実現するシステムを開発した。

具体例として、アミロイド前駆体タンパク質（amyloid-beta precursor protein：APP）の産生メカニズム解明を目的とした網羅的遺伝子ノックダウンとタンパク質間相互作用データを組み合わせたデータセット［4］を用いて、上記の実現可能性を証明した。また、一般のユーザーが自分自身のデータを使用して同様のデータウェアハウスを構築するための詳細なプロトコルの公開を予定している。

方法および結果

1. プラットフォームの定義

従来、データの共有は、一連のファイルまたはデータベースの単純な交換であると考えられてきた。しかし、生命科学研究で FAIR 原則に適切に従うためには、基礎となるデータにアクセスして再利用可能にするツールや方法の共有も考慮する必要がある。

本報告では、アミロイド前駆体タンパク質（APP）の産生メカニズム解明を目的とした網羅的遺伝子ノックダウンとタンパク質間相互作用（protein-protein interaction：PPI）データとを組み合わせたデータセット（以降 APP データセットと呼ぶ）を用いて、方法と結果を説明する。このデータセットの場合、実験から得られた情報は、PPI ネットワークと、一連の siRNA ノックダウン下における 4 種類のアミロイド生成ペプチド産生量、という 2 つのテーブル（ファイル）から成る［4］。図 1 に、APP データセットの一部を示す。a）は PPI ネットワークを形成する遺伝子ペアのリストであり、各行は、6927、55281 などの ID で表される 2 つの遺伝子が相互作用してネットワークのエッジを形成することを示している。b）は、カラム A に示される各遺伝子をノックダウンした際に細胞が産生する 4 種類のアミロイド生成ペプチドの測定値が、それぞれカラム C、D、E、F に示されている。ここで、a）と b）に含まれる全ての遺伝子のリストは必ずしも一致しない。

```
1   6927,55281,HNF1A,TMEM140
2   164,8906,AP1G1,AP1G2
3   3689,11082,ITGB2,ESM1
4   57326,151871,PBXIP1,DPPA2
5   5336,23236,PLCG2,PLCB1
6   9146,57596,HGS,BEGAIN
7   842,317,CASP9,APAF1
8   2203,3837,FBP1,KPNB1
9   8658,25913,TNKS,POT1
10  5985,5981,RFC5,RFC1
```

a) PPI ネットワークデータ

	A	B	C	D	E	F
1	GENEID	VIABILITY	AB40	AB42	SAPPA	SAPPB
2	1	-0.293511677	0.509799	0.063421	1.624922	1.312808
3	2	-0.475129933	0.757599	1.043317	0.26648	0.491344
4	9	2.377759282	-1.213444	-1.330486	-0.458607	-1.471134
5	10	-0.679732748	-0.415497	-0.230574	0.51956	-0.136712
6	12	0.505662562	0.191584	0.569691	1.248911	1.476637
7	13	3.0305155	-3.108676	-3.8627	-0.780368	-1.64174
8	14	4.230880349	-3.004392	-2.810236	-1.218654	-2.155312
9	16	0.754129753	0.650664	1.335904	0.814069	1.183435
10	18	-0.589396952	0.717946	1.22595	1.328739	1.237791

b) siRNAデータ

図 1.　APP データセットを構成する 2 つの異なるファイル

a）　PPI ネットワークを形成する相互作用のリスト。

b）　与えられた遺伝子に対する異なる siRNA スクリーン値のリスト。

このデータセットを所有して統合的なアクセスを提供したいユーザーと、データを照会し、すべてまたは一部のデータを抽出したいユーザーの両方を考慮し、データ（ファイル）とデータを照会するために使用するツール（ソフトウェア）の両方を共有できるプラットフォームを開発した。

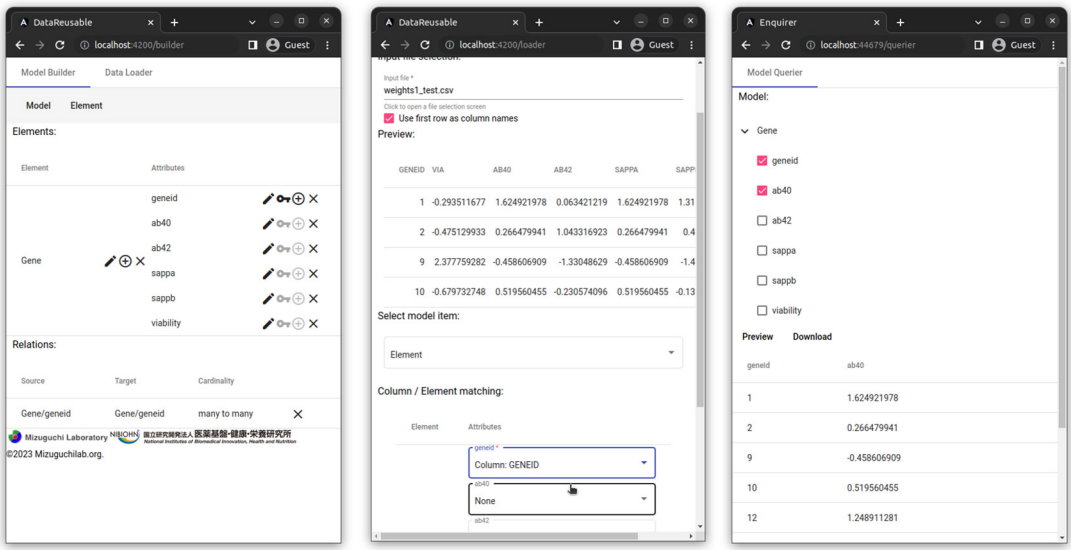
2. プラットフォームアーキテクチャ

プラットフォームの最初の構成要素は、共有するすべてのデータの共同リポジトリとして機能するデータベースである。APP データセットについては、このデータベースに PPI や siRNA データを保存した。

本データ共有プラットフォーム上に、データ提供者がどのような種類の生物学的データを共有するかを規定し（これをモデル構築と呼ぶ）、対応するデータをアップロードするためのユーザーインターフェースを構築した。例えば、図 2a は、APP データセットにおいて、遺伝子（Gene）という要素が、ID や siRNA ノックダウンの測定値などの属性を持つと定義することによって、他のデータ（ファイル）との紐付けが可能になる様子を示している。一方、PPI データは、遺伝子間の関係として定義される。このようなモデル構築・保存後に、実際の値がデータベースに読み込まれる（図 2b）。

データ照会者には、別のインターフェースを用意した。図 2c に示すように、このインターフェースは、ダウンロードするデータの形を柔軟に選択できることを重視した。例えば、あるユーザーが 1 つの測定値（AB40）だけに興味を持った場合、AB40 の産生量に影響を与える遺伝子とそれらが形成する PPI ネットワークのデータのみを抽出することができる。

最後のモジュールは、それぞれのユーザーインターフェースと基礎となるデータベースを結びつけるために使用される。以下、このモジュールを接続インターフェース、または Application Programing Interface（API）と呼ぶ。



a) データ提供者インターフェイス：モデル保存　b) データ提供者インターフェイス：データアップロード　c) データ照会者インターフェイス

図 2. データ提供者と照会者用アプリケーションのスクリーンショット

a) 提供者アプリケーションに含まれるモデル構築用ビュー。

b) 提供者アプリケーションに含まれるデータアップロード用ビュー。

c) 照会者アプリケーションに含まれるデータ検索のためのビュー。

3. 実装

以下では、本プラットフォームの技術的な実装の詳細について説明する。

データ提供者と照会者のインターフェースは、Angular フレームワークを用いた Web アプリケーションとして実装した。プラットフォームが使用する統合データベースの管理には PostgreSQL を使用した。最後に、接続インターフェース（API）は、Node.js フレームワークを使用して実装した。

図 3 は、すべてのモジュールが連携している様子を示している。

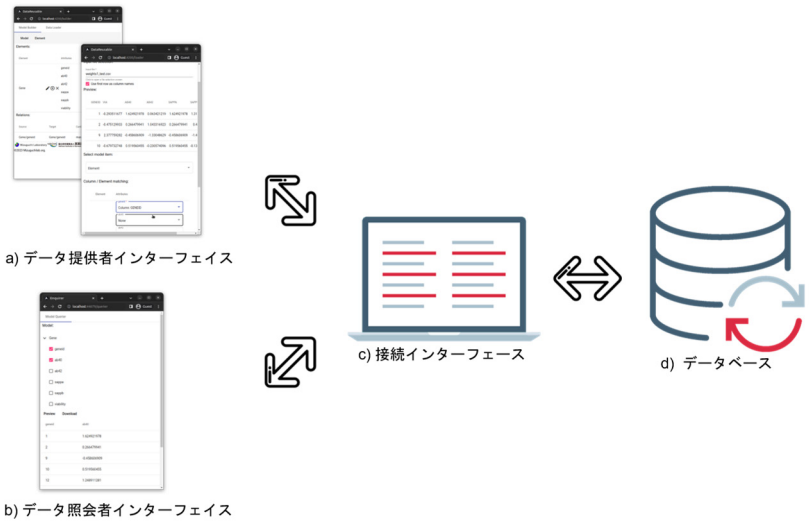


図 3. データ利用性担保プラットフォームアーキテクチャ

データ提供者であるユーザーとデータ照会者であるユーザーに特化した異なるアプリケーションやデータベースサービスを組み合わせることで、統合的な共有機能を提供する。

考　察

これまで、データを共有する場合、データ提供者は、データをプレーンテキストや表計算ソフト（Microsoft Excel など）で提供することが多かった。しかし、共有されるデータが複数の種類を含み、かつ、それらの種類間に何らかの関係がある場合、ユーザーは内容を再構成する必要があり、データ再利用のハードルが高くなる。データをリレーショナルデータベースに変換することで、データ共有を統合的な形にすることができるが、データベースを操作するためには専門的な知識を必要とする。Microsoft Access や Claris FileMaker などの市販ソフトウェアを使えば、少ない労力でデータベースを作成することができるが、これらのソフトウェアはアプリケーションを開発するためのもので、データを共有するためのものではない。この問題を解決するために、本研究では、生物学データをオンラインで共有する汎用的な方法を提案し、実証した。

今回の実装では、技術的な理由から、より網羅的なデータ統合や高度なデータクエリへの対応が可能な InterMine［5］フレームワークを組み込むことはできなかったが、今回提案したデータ共有のコンセプトに InterMine を導入することは可能と考えている。データ提供者がフロントエンドアプリケーションでデータモデルを定義すれば、その情報を用いて InterMine フレームワークがバックエンドと照会者アプリケーションを生成するという枠組みが想定される。さらに、各種データの分布を可視化する機能についても実装を計画している。

将来的には、本プラットフォームを通してデータ提供者が Docker［6］イメージを作成し、公開できる機能の付加を予定している。これにより、共有データに興味を持つユーザーは、Docker をインストールするだけで、簡単にデータの全部または一部を利用することができる。さらに、プラットフォームを構成する各モジュールのソースコードは、GitHub 上で MIT License のもとの公開を予定している。

共同研究者・謝辞

本研究の共同研究者は、医薬基盤・健康・栄養研究所、AI 健康・医薬研究センターの陳怡安及びロドルフォ・アジェンデスである。

文　献

1） Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Mar 15;3:160018. doi: 10.1038/sdata.2016.18. Erratum in: Sci Data. 2019 Mar 19;6(1):6. PMID: 26978244; PMCID: PMC4792175.

2） Chen YA, Tripathi LP, Mizuguchi K. TargetMine, an integrated data warehouse for candidate gene prioritisation and target discovery. PLoS One. 2011 Mar 8;6(3):e17844. doi: 10.1371/journal.pone.0017844. PMID: 21408081; PMCID: PMC3050930.

3） Chen YA, Allendes Osorio RS, Mizuguchi K. TargetMine 2022: a new vision into drug target analysis. Bioinformatics. 2022 Sep 15;38(18):4454-4456. doi: 10.1093/bioinformatics/btac507. PMID: 35894632; PMCID: PMC9477527.

4) Camargo LM, Zhang XD, Loerch P, Caceres RM, Marine SD, Uva P, Ferrer M, de Rinaldis E, Stone DJ, Majercak J, Ray WJ, Chen YA, Shearman MS, Mizuguchi K. Pathway-based analysis of genome-wide siRNA screens reveals the regulatory landscape of APP processing. PLoS One. 2015 Feb 27;10(2):e0115369. doi: 10.1371/journal.pone.0115369. Erratum in: PLoS One. 2015;10(6):e0129641. PMID: 25723573; PMCID: PMC4344212.

5) Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, Lyne M, Lyne R, Kalderimis A, Rutherford K, Stepan R, Sullivan J, Wakeling M, Watkins X, Micklem G. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. 2012 Dec 1;28(23):3163-5. doi: 10.1093/bioinformatics/bts577. Epub 2012 Sep 27. PMID: 23023984; PMCID: PMC3516146.

6) https://www.docker.com/