糖鎖関連物質のリポジトリ開発

Development of data repositories for glycan-related resources

15D5603　　高橋悠志　　指導教員　木下フローラ聖子

SYNOPSIS

The importance of large-scale omics analyses has been growing more and more to understand the precise interactions among biological molecules. To perform omics-wide analyses, sharing experimental data from researchers is extremely valuable. Therefore, many public data repositories have been developed in the life sciences, including PRIDE and PeptideAtlas in proteomics and GlyTouCan, GlycoPOST and UniCarb-DR in glycomics, to promote sharing and accumulating data from researchers under the FAIR (Findability, Accessibility, Interoperability, and Re-usability) data principles. However, one of the missing pieces is the accumulation of glycoconjugate data, including glycopeptides, glycoproteins, glycolipids, and glycosides. GlyTouCan, for example, assigns unique identifiers for glycan structures, including monosaccharide compositions. To assign such unique identifiers to glycoconjugates, we have developed a novel data repository called GlyComb. Researchers can currently register glycopeptides and glycoproteins by specifying an amino acid sequence or UniProt ID, glycosylation sites, and glycan information to GlyComb in TSV format. Because GlyComb is built on top of Semantic Web technologies, users will be able to collect the specified glycoconjugate-associated information easily using the SPARQL query language. In addition, existing glycomics data repositories required further collaboration enhancements in order to optimize user convenience.　In particular, UniCarb-DR and GlycoPOST are both similar repositories for glycomics experimental data, but they were developed and maintained independently. Therefore, in this study, in addition to developing a novel data repository for glycoconjugate data, and I have enhanced the integration of UniCarb-DR and GlycoPOST for integrating the results of glycomics mass spectrometry experiments.　As a result, these developments enable glycoscientists to more easily submit glycoconjugate data and mass spectrometry data in the glycosciences into the worldwide informatics infrastructure.

Keywords: Glycan, Glycoconjugate, Glycoinformatics, Glycomics, Data repository

1. 背景

生体内におけるタンパク質や糖鎖、脂質といった様々な生体分子同士の相互作用を正確に理解するため、大規模なマルチオミックス解析の重要性が高まっている。このようなオミックス解析を促進・加速させていくためには、世界中の研究者同士で質量分析実験や液体クロマトグラフィー実験などから得られた生データを含む様々な実験結果・解析結果を共有し、これらをコンピュータによる再解析の対象としていくことが必要不可欠である。この目的のため、近年ライフサイエンス分野では各々の研究者が自身の論文を執筆・発表する際にその研究の過程で得られた実験データをインターネット上にアップロードおよび公開するための情報基盤として多数の公開データリポジトリが開発されてきた。これらの中には、プロテオミクスにおける PRIDE [1] や PeptideAtlas [2]、糖鎖構造に一意な識別子を与える GlyTouCan [3]、グライコミクスにおける GlycoPOST [4]および UniCarb-DR [5] が含まれている。各データリポジトリに対して投稿されたデータにはそれぞれ一意な識別子が割り当てられ、研究者は自身の論文中でその識別子を示すことによって読者に対して自身の研究データを一意に示すことができる。これらのデータリポジトリは研究者から投稿された貴重な実験結果・解析結果を FAIR（Findability, Accessibility, Interoperability, and Re-usability）データ原則に則って研究者間で共有・蓄積していくことを目指している。

さらに、蓄積されたこれらの研究データをコンピュータ上での再解析の対象とするためには、使用したサンプルの調整方法や試薬の種類、濃度、反応時間や、データ測定に使用した分析計のセッティングなどといった様々な実験条件についての情報も考慮する必要がある。このような定量的・定性的な実験報告のためのガイドラインとして、プロテオミクスでは The Minimum Information About a Proteomics Experiment (MIAPE) [6]、グライコミクスにおいては Minimum Information Required for A Glycomics Experiment (MIRAGE) [7] と呼ばれる標準ガイドラインが提案されている。これらのガイドラインにはそれぞれの研究分野ごとにサンプルの調整方法や質量分析実験、液体クロマトグラフィー実験やキャピラリー電気泳動実験などの報告の際に共に報告が必要となる最小限の実験情報がまとめられている。

糖鎖科学研究分野においては、GlyCosmos [8] プロジェクトのもとで糖鎖構造情報のための国際的なデータリポジトリである GlyTouCan を始めとしていくつかのデータリポジトリが開発・公開されてきた。GlycoPOST はグライコミクスにおける質量分析実験から得られた生データを MIRAGE ガイドラインに則って投稿するためのデータリポジトリであり、UniCarb-DR は GlycoWorkbench [9] と呼ばれるグライコミクス質量分析実験の同定補助ソフトウェアを用いてアノテーション付けされた質量分析実験同定結果のためのデータリポジトリである。特に、質量分析実験から得られた生データをそのまま投稿できる GlycoPOST が開発されたことで、グライコミク

スにおける質量分析実験結果に対して世界中の研究者がいつでも再解析を行えるようになった。これらのデータリポジトリが開発されたことで糖鎖科学研究から得られる多くのデータを蓄積・共有していくことが可能となったが、糖鎖の機能は単独に働くのではなく、修飾している分子との関係によって発揮されていることがわかっている。これに対し、糖ペプチド、糖タンパク質、糖脂質、配糖体などの複合糖質の情報をまとめて捉えられるための基盤は存在しなかった。つまり、GlyTouCan のように複合糖質に対して識別子を割り振るシステムがこれまで存在していなかった。また、類似したデータリポジトリが複数開発されてきたことで、利用する研究者の利便性を最大化するためにそれぞれのデータリポジトリ間の連携の強化が課題となっていた。

2. 目的

本研究では、糖鎖科学関連物質のためのリポジトリ開発として既存のデータリポジトリでは情報の蓄積ができなかった複合糖質情報の蓄積を促進するための新たなデータリポジトリ GlyComb の開発と、既存のデータリポジトリ間の連携強化のための第一歩として GlycoPOST と UniCarb-DR という 2 つのデータリポジトリ間の連携強化のための改良を行った。これにより、これらのリソースが糖鎖構造のみならず糖ペプチドや糖タンパク質などの様々な複合糖質情報も蓄積できる、糖鎖科学研究のための包括的なデータリポジトリシステムとして機能できるようになることを目指した。

3. 方法・結果

複合糖質データリポジトリ GlyComb の開発

本研究で開発した GlyComb は複合糖質のためのデータリポジトリであり、現在は糖ペプチドおよび糖タンパク質のエントリのみを登録することができる。糖鎖構造情報のためのデータリポジトリである GlyTouCan は単糖情報や結合情報に曖昧性を含む任意の糖鎖構造に対して一意な識別子である GlyTouCan ID を与えることができるが、GlyComb は同様にして任意の複合糖質エントリに対して一意な識別子である GlyComb ID を与えることができる。このような識別子は異なる様々なデータベース中のエントリ間をリンクすることができるため、非常に有用である。図 1 には特定のペプチド配列中のある Asn 残基に対し、複数個の糖鎖構造の修飾が考えられるような糖ペプチドエントリの例を Symbol Nomenclature for Glycans（SNFG）[10] 単糖シンボル表記法を用いて示した。GlyComb ではこのような複合糖質情報を TSV 形式の文字列として登録することができる。糖ペプチドエントリの場合、アミノ酸配列、糖鎖修飾残基番号、糖鎖構造をタブ区切りで記述する。図 2 に図 1 に示した糖ペプチドエントリの TSV 文字列表現を示した。糖鎖構造の指定には図 2 中に示されているような組成記法の他に

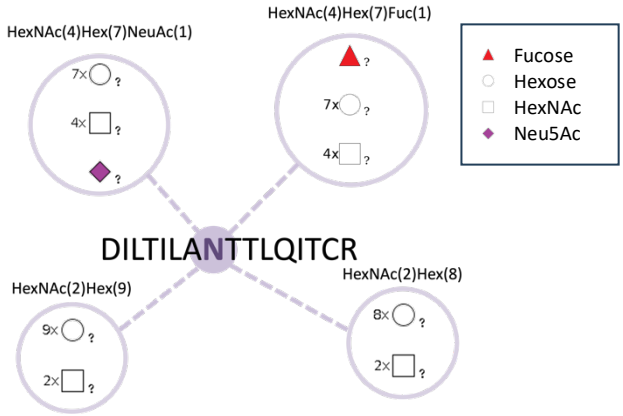

図 1. Asn 残基（N）に 4 種類の糖鎖構造のうちのいずれかの修飾を取りうる糖ペプチドの例

DILTILANTTLQITCR	TAB	8	TAB	HexNAc(4)Hex(7)NeuAc(1)
DILTILANTTLQITCR	TAB	8	TAB	HexNAc(4)Hex(7)Fuc(1)
DILTILANTTLQITCR	TAB	8	TAB	HexNAc(2)Hex(8)
DILTILANTTLQITCR	TAB	8	TAB	HexNAc(2)Hex(9)

図 2. GlyComb に入力できる TSV 形式の文字列で表現された糖ペプチドエントリの例

GlyTouCan ID を用いることができる。

一般的に、このような糖ペプチドエントリ情報は LC-MS/MS を用いたグライコプロテオミクス実験の同定結果として得られる。一回のグライコプロテオミクス実験の結果、通常は 10,000 以上の Peptide-Spectrum Match (PSM) が得られるが、得られた PSM データのフォーマットは解析に用いたソフトウェアごとに独自の形式となっており、研究者自身にこの同定結果を GlyComb が認識できる TSV 形式の文字列として手作業で変換してもらうのは多大な労力を要求してしまうこととなる。この負担を軽減するため、GlyComb ではグライコプロテオミクス実験結果の解析に幅広く使われているソフトウェアの 1 つである PMI-Byonic [11] が生成する解析結果のサマリー Excel ファイル中から GlyComb 用の TSV 文字列を自動的に抽出するための変換ユーティリティをインターネット上で提供している。この変換ユーティリティを用いて、現在 PRIDE データリポジトリ上で公開されている 1,465 個の PMI-Byonic 由来のサマリー Excel ファイル中から 95,125 件の糖ペプチドエントリと 24,831 件の糖タンパク質エントリを抽出して GlyComb に登録した。登録した糖ペプチドエントリのうち、56,788 件のエントリには *N*-型糖鎖修飾が含まれており、38,376 件のエントリには *O*-型糖鎖修飾が含まれていた。同様にして、プロテオミクス実験解析ソフトウェアおよびデータベースのスイートである ProteinProspector [12] に含まれている MS-Viewer [13] リポジトリ中からは 3,189 件の糖ペプチドエントリと 1,096 件の糖タンパク質エントリを抽出し、GlyComb へと登録した。この糖ペプチドエントリのうち、1,032 件には *N*-型糖鎖修飾が含まれており、1,911 件には *O*-型糖鎖修飾が含まれていた。GlyComb に登録されたこれらの情報は GlycoCoO [14] と呼ばれる複合糖質のためのオントロジーを用いて Resource Description

Framework (RDF) データとしてデータベースに保存され、SPARQL に代表されるセマンティックウェブ技術を用いて他のデータベースやデータリポジトリと一緒に検索に利用することができるようになっている。GlyComb は現在 GlyCosmos プロジェクトのポータルサイトから利用することができるようになっている。

GlycoPOST と UniCarb-DR の連携強化

GlycoPOST と UniCarb-DR はどちらも GlyCosmos プロジェクトのポータルサイト上からアクセスできるデータリポジトリであり、いずれもグライコミクスにおける質量分析実験の結果を MIRAGE ガイドラインに従って投稿することができる。両者の違いは、GlycoPOST が質量分析実験の生データを含む任意の種類のファイルをアップロードすることができるのに対し、UniCarb-DR は GlycoWorkbench ソフトウェアを用いてアノテーション付けされた質量分析実験の結果のみを受け付けるという点である。そのため、GlycoPOST は様々な種類の実験結果データを保管できる一方で投稿された内容を可視化するための機能は一切持っていない。これに対し、UniCarb-DR にはアノテーション付けされた MS/MS 実験のスペクトルデータを可視化するための機能や強力な検索機能が搭載されている。また、GlycoPOST が論文の出版前に投稿内容を査読者だけに限定公開できる Embargo 機能を持っているのに対し、UniCarb-DR には Embargo 機能は搭載されていないため、論文執筆の際に利用しづらいという問題があった。これらの両者の特徴を踏まえ、利用者がこれらの相補的なデータリポジトリの利点をこれまで以上に引き出せるよう、これらのデータリポジトリ間の連携の強化を図った。まず、ユーザーからのこれら 2 つのデータリポジトリへのデータ投稿を統一するために、データ投稿システムを GlycoPOST のデータ投稿システムへと一本化して UniCarb-DR のデータ投稿ページを廃止することとした。GlycoPOST では Embargo 期間が終了して一般公開されたデータは API を用いて取得することができるようになっているため、UniCarb-DR 側では定期的にバッチ処理を実行して GlycoPOST の API を呼び出すことで新たに GlycoPOST 上で公開されたデータ中に GlycoWorkbench ファイルが含まれている場合にはそれらを自動的に取得し、UniCarb-DR へと登録を行うように変更した。これにより、UniCarb-DR 自体は Embargo 機能を持っていないが Embargo 期間が終了して一般公開されたデータ中に含まれている GlycoWorkbench の内容は自動的に UniCarb-DR に登録され、可視化することが可能となった。このとき、GlycoPOST と UniCarb-DR に登録されたそれぞれのエントリ同士は自動的に相互参照するようにした。なお、GlycoPOST の改良も同時に行い、MiniCarb-Viewer と呼ばれる Embargo 期間中にも投稿された GlycoWorkbench ファイルの内容を可視化することができる機能も実装した。図 3 に MiniCarb-Viewer を用いて Embargo 期間中に投稿された GlycoWorkbench ファイルの可視化を行う際の様子を示す。

これらの改良に加え、データ投稿システムの一本化を行うのに当たり、GlycoPOST では従来液体クロマトグラフィー実験に関する実験情報を登録できなかったのに対し、UniCarb-DR では限定的ながら HPLC 実験の情報を他の MIRAGE ガイドライン情報と共に登録することができた。このまま UniCarb-DR のデータ投稿システムを廃止すると液体クロマトグラフィー実験の実験情報が一切これらのデータリポジトリに登録できなくなってしまうという問題があったため、GlycoPOST の拡張を行い、MIRAGE の液体クロマトグラフィーガイドラインに従って液体クロマトグラフィー実験情報も GlycoPOST に登録できるように実装を行った。これにより、従来 UniCarb-DR に登録できていた以上の実験情報を GlycoPOST に登録することが可能となった。

4. 考察・課題

本研究では複合糖質情報のための新たなデータリポジトリの開発とグライコミクスにおける質量分析実験結果のための 2 つのデータリポジトリの連携強化を行った。GlyTouCan および GlycoPOST は特に現在のグライコミクス研究の基盤として世界中から利用されているデータリポジトリであるが、本研究によってこのうちの GlycoPOST の利便性向上に寄与することができたと考えられる。これら既存のデータリポジトリに加え、複合糖質情報を蓄積する GlyComb を活用することで、今後グライコミクスのみならずグライコプロテオミクス情報をも蓄積していくためのデータリポジトリ基盤が整えられたと考えられる。今後、これらのデータリポジトリ群の連携を更に強化していくためのパートナーシップ機能の実装を行っていく予定である。これにより、例えば GlycoPOST に投稿された GlycoWorkbench ファイル中に記述されていた糖鎖構造に対して対応する GlyTouCan ID が存在しない場合に自動的に GlyTouCan 上で新たな糖鎖構造の登録が行われて GlyTouCan ID の発行が行えるようになるほか、GlycoPOST 中に投稿された実験結果中に含まれて

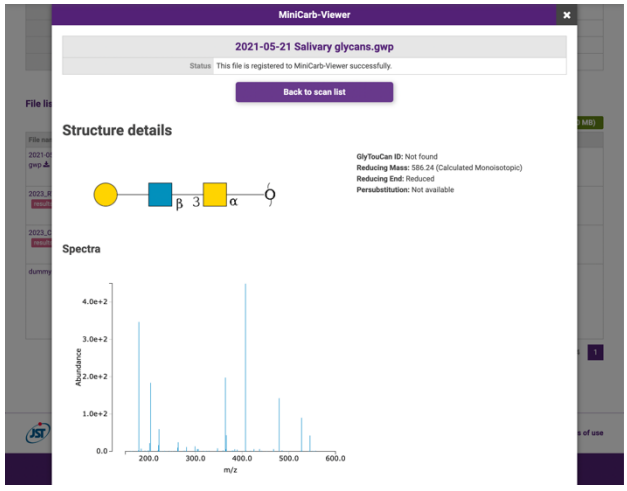

図 3. MiniCarb-Viewer を用いた Embargo 中の投稿中に含まれている GlycoWorkbench ファイルの可視化

いる糖ペプチド・糖タンパク質情報が自動的に GlyComb に登録することができるようになるといった様々なリポジトリ間連携が可能になると考えられる。また、 GlyComb では糖ペプチド・糖タンパク質以外の複合糖質情報である糖脂質や配糖体情報の登録にも順次対応していく予定である。

5. 展望

本研究において複合糖質情報に対して一意な識別子を与えるデータリポジトリである GlyComb が開発されたことによって、生体内パスウェイ中に現れる今まで識別子が割り当てられていなかった糖鎖関連物質に対しても識別子を与えることが可能となった。単独の糖鎖構造やペプチド、タンパク質、脂質ではなく、糖鎖修飾を受けた特定の生体分子情報を一意に示すための基盤ができたことで、今後セマンティックウェブ技術を用いて様々なバイオインフォマティクスリソース間で複合糖質情報を含めた情報の統合が進むだろう。これにより、例えば特定の複合糖質分子が発症に関わっている疾患を探し出すこともこれまで以上に容易になることが期待できる。また、GlycoPOST に投稿された実験結果中に含まれている複合糖質分子に対して識別子を与えることができれば、セマンティックウェブ技術を用いてその複合糖質分子が含まれている実験結果データだけをデータリポジトリ中から取り出すことも可能になるだろう。これにより、糖鎖科学研究のさらなる発展を効率的に促進することができると考えられる。本研究によって GlycoPOST と UniCarb-DR の間で密に連携が行われるようになったことで、既に特定の MS スペクトルデータと類似した MS スペクトルデータを含む質量分析実験の結果を探すこともできるようになった。今後はこのような糖鎖科学研究から得られた様々な実験データの統合と再解析が進んでいくだろう。

タンパク質や脂質といった生体内分子の多くは糖鎖修飾を受けており、これらの分子同士の正確な振る舞いを理解するためには糖鎖修飾情報も含めた解析が不可欠である。そのため、グライコミクスおよびグライコプロテオミクス分野の情報を蓄積することができるデータリポジトリは今後のマルチオミックス解析において欠かせないものとなるだろう。

6. 参考文献

[1] Perez-Riverol, Yasset, et al. "The PRIDE database resources in 2022: a hub for mass spectrometry-based proteomics evidences." Nucleic acids research 50.D1 (2022): D543-D552.

[2] Van Wijk, Klaas J., et al. "The Arabidopsis PeptideAtlas: harnessing worldwide proteomics data to create a comprehensive community proteomics resource." The Plant Cell 33.11 (2021): 3421-3453.

[3] Fujita, Akihiro, et al. "The international glycan repository GlyTouCan version 3.0." Nucleic acids research 49.D1 (2021): D1529-D1533.

[4] Watanabe, Yu, et al. "GlycoPOST realizes FAIR principles for glycomics mass spectrometry data." Nucleic acids research 49.D1 (2021): D1523-D1528.

[5] Rojas-Macias, Miguel A., et al. "Towards a standardized bioinformatics infrastructure for N-and O-glycomics." Nature communications 10.1 (2019): 3275.

[6] Taylor, Chris F., et al. "The minimum information about a proteomics experiment (MIAPE)." Nature biotechnology 25.8 (2007): 887-893.

[7] York, William S., et al. "MIRAGE: the minimum information required for a glycomics experiment." Glycobiology 24.5 (2014): 402-406.

[8] Yamada, Issaku, et al. "The GlyCosmos Portal: a unified and comprehensive web resource for the glycosciences." Nature Methods 17.7 (2020): 649-650.

[9] Ceroni, Alessio, et al. "GlycoWorkbench: a tool for the computer-assisted annotation of mass spectra of glycans." Journal of proteome research 7.4 (2008): 1650-1659.

[10] Varki, A. et al. Symbol nomenclature for graphical representations of glycans. Glycobiology 25, 1323–1324 (2015).

[11] Bern, Marshall, Yong J. Kil, and Christopher Becker. "Byonic: advanced peptide and protein identification software." Current protocols in bioinformatics 40.1 (2012): 13-20.

[12] Bi, Yang, et al. "SPINDLY mediates O-fucosylation of hundreds of proteins and sugar-dependent growth in Arabidopsis." The Plant Cell 35.5 (2023): 1318-1333.

[13] Baker, Peter R., and Robert J. Chalkley. "MS-viewer: a web-based spectral viewer for proteomics results." Molecular & Cellular Proteomics 13.5 (2014): 1392-1396.

[14] Yamada, Issaku, et al. "The glycoconjugate ontology (GlycoCoO) for standardizing the annotation of glycoconjugate data and its application." Glycobiology 31.7 (2021): 741-750.