第 3 回 SPARC Japan セミナー2019

「実践 研究データ管理」

糖鎖科学における研究データ管理

山田 一作
（野口研究所）

講演要旨

糖鎖科学コミュニティーにおいて塩基配列やタンパク質のリポジトリは利用できたが糖鎖構造のリポジトリはなかった。2013年に中国・大連において開催された国際会議において、糖鎖科学研究において糖鎖構造の明確化について議論された。この会議の結果、糖鎖構造のリポジトリを構築し糖鎖構造に固有のアクセッション番号を付与することが合意された。そし 2014 年 8 月に国際糖鎖構造リポジトリ GlyTouCan が公開された。また、糖鎖科学では質量分析法により糖鎖構造を解析するが、その際に生成するデータも重要であり、これらのリポジトリとして UniCarb-DR および GlycoPOST が開発された。UniCarb-DR は質量分析データを解析したピークリストなどのリポジトリであり、GlycoPOST は、質量分析の実験法や生データのためのリポジトリである。本発表ではこれらのリポジトリの開発の経緯なども含めて紹介したい。

山田 一作

公益財団法人野口研究所インフォマティクスプロジェクトリーダー。1997年に東京都立大大学院工学研究科にて博士号取得。2002年から野口研究所研究員、2006年から同所にて糖鎖科学研究を開始し、糖鎖構造表記法、オントロジー、データベースなど糖鎖インフォマティクス研究に従事している。

自己紹介＋α

今日は、糖鎖科学の領域の研究データをどうやって扱っているのかという糖鎖の領域のお話で、図書館の方たちと少し違うのですが、図書館の方や他の領域の方の意見を聞きながら、より良いものにしていきたいと考えています。

私は学生のときから有機化学を専攻していて、野口研究所に入って糖鎖科学を始めました。有機化学の中にも糖鎖構造はありますが、糖鎖科学は少し特殊というか、私としては取っ付きにくかった領域で、すごく難しいものだと考えていました。しかし、有機化学を学んでいたので化学構造には馴染みがあり、それを中心に糖鎖科学を見ていくと、難しいのですが実は面白い、やりがいがあるものだと思いまして、結局今も何とかこの分野を良くしていきたいという思いで続けています。

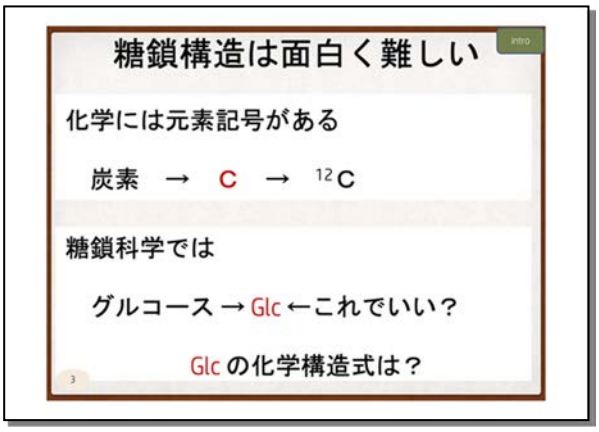
難しいという一つには、化学が決して簡単なわけではないということもあるのですが、例えば化学だと炭素を表すのに元素記号で C と書けば間違いなく炭素だという認識があります（図 1）。肩に 12 と書くと同位体表記があるという共通認識があります。それは世界共通だと思います。そういう世界でずっとやってきたのですが、では糖鎖科学ではどうなっているかというと、糖鎖の化学の人たちは元素記号を使って化学構造式を書いたりします。ただ、サイエンスとなると、化学以外にも生物や医学などさまざまな人が絡んでき

ます。そういう人たちが集まってグルコースといったときに、国際純正・応用化学連合（IUPAC）では Glc と書くと決まっているのですが、これでいいのか。駄目かといわれると「いいんじゃない？」という答えが大体返ってくるのですが、化学から見ると Glc の化学構造式はどのようなものなのかが気になります。

Glc はグルコースだと考えると、図 2 は欧州バイオインフォマティクス研究所（EBI）の ChEBI というデータベースですが、丸で囲ったところにグルコースがあります。化学構造式で書けるものに関しては四角の左側に化学構造が記載してあります。しかし、化学構造式で書けないもの、非常に書きにくいものに関しては構造の記載がありません。

グルコースといった場合に、非常に曖昧な部分が含まれていて、その部分を明確に表すことができません。グルコースには D 体と L 体という鏡像異性体があります。グルコースといったときに、生物の方などは生体内に一般的にあるような D-glucose を思い浮かべますが、D-glucose も化学構造式が載っていません。それは、グルコースにも、環状になっていたり、直鎖状になっていたり、さまざまなパターンがあるからです。そういうところが明確に記載されていないと、さらにグルコースの細かいデータを付加していかないと情報が正確ではないことになります。そんなところが私としては非常に面白いです。

今日の内容は、最初に、糖鎖とはどのようなものなのか。その後に、糖鎖科学における研究データとはどのようなものか。そして、私自身が構造に興味を持っているということもあり、糖鎖の構造と、糖鎖科学における標準化・共通化の取り組み、そして、それらを使って糖鎖科学でデータを保管するリポジトリを作っているので、その利用を含めてご紹介したいと思います。

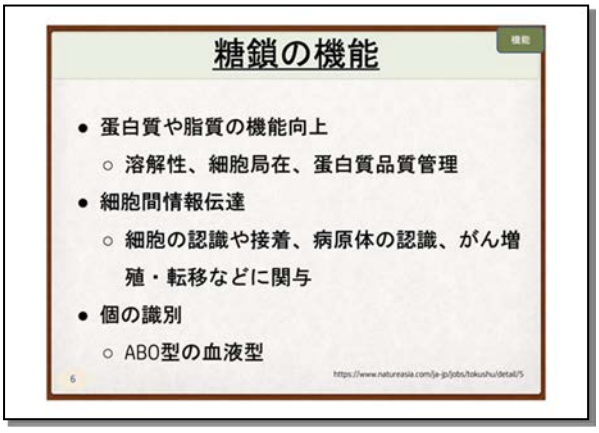


（図 1）

（図 2）

糖鎖の機能

図 3 は糖鎖の機能の一例です。糖鎖は単体で機能することもありますが、蛋白質や脂質と複合体を形成して機能することも知られています。蛋白質や脂質は水溶性ではないものが多いのですが、そこに糖鎖という水溶性が高いものが付加すると溶解性が増して、その付き方によって機能が変わっていきます。また、蛋白質に糖鎖が付くのですが、その糖鎖の構造が変わることによって、蛋白質の品質管理を行っています。



（図 3）

二つ目の機能は、細胞間の情報伝達に使われているということです。糖鎖は細胞の表面に非常に多く存在し、細胞間の認識や接着に関わっています。病原体の認識や、がんによって糖鎖の構造が分かるといったことがよく知られています。

三つ目の機能は個の識別です。糖の話をするときによく出てくるのが ABO の血液型の話です。この血液型は糖鎖の構造が異なることによって型が決まっているので、私から見ると糖鎖の構造が非常に重要だということになります。

糖鎖の分野における研究の手法を全て網羅することはなかなか難しいですが、例えば図 4 のようなものがあります。合成で糖鎖を作るとなると、化学合成や酵素合成という方法が行われています。あとは分析です。例えば糖鎖の構造を解析したり、各種の相互作用や機能を見ていく、そして糖鎖の局在などをイメージングするということが行われています。

糖鎖の研究では、どのような糖がどのような機能になっているのか、どういう相互作用があるのかということが重要になってきて、一つの鍵として糖鎖構造というものが考えられます。

糖鎖の構造

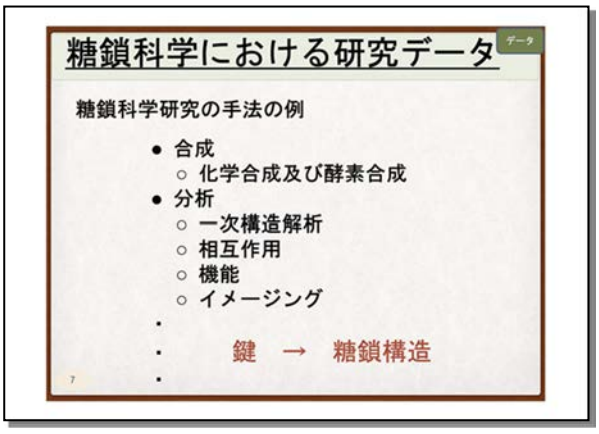
合成や分析などいろいろな研究があるのですが、糖鎖の構造とその構造を決めるための分析データが、糖鎖研究の中でも重要な位置にあります。それ以外も非常に重要なのですが、今回は構造にフォーカスしたお話をしたいと思います。

先ほど糖鎖の構造データということで、Glc（グルコース）の話をしました。その文字列の表記方法としては、図 5 の IUPAC の下に長い文字が書いてありますが、このような表記法がよく使われます。それ以外に、この構造の糖鎖はこういう名前ですよという名称がいろいろ付いています。

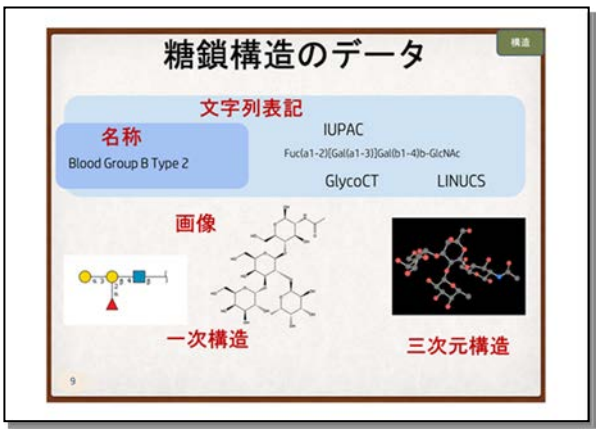
左下のカラーのものは単糖のシンボルが複数つながって糖鎖になっていますし、真ん中の画像のような元素記号を使った化学構造式や、シミュレーションで使われる三次元構造など、さまざまな構造データがあります。こういうものが論文などいろいろなところにばらばらに存在しています。全て一つの糖のことを表しているのですが、分野によって、自分たちに一番分かりやすいように書き表し方が違っています。しかし、データを見る上では、どれかを見たときに他のデータも見えるようにならないといけないということがあり、これらのデータをうまく扱っていけるような仕組みが必要です。

糖鎖科学における標準化

そういった中で、糖鎖科学における標準化、共通の言葉やルールをつくっていこうという動きがあります（図 6）。JIS Z 8002:2006（標準化及び関連活動－一般的な用語）における標準化の定義には「最適な秩序を得ることを目的として」と書いてあります。標準化は、相互理解や互換性の確保、品質、正確な情報という点



（図 4）



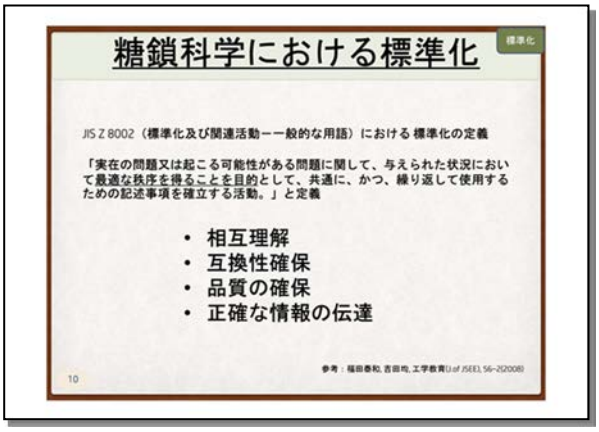
（図 5）

で非常に重要であり、糖鎖の分野でもそういうところを意識した取り組みを行っています。その取り組みの中で、糖鎖構造の記号について、それから Glc などの文字列表記について、そして、糖鎖のガイドラインづくりや、糖鎖構造に ID を付けてひも付けることで全てが同じであることを示す仕組みづくりに取り組んできています。

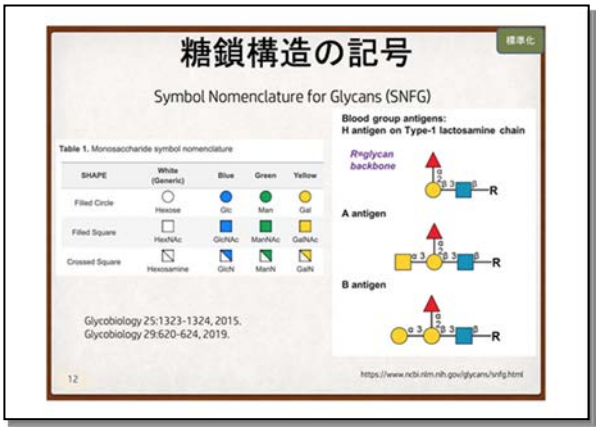
図 7 は糖鎖構造で使う記号で、糖鎖と単糖です。これを糖鎖の領域でみんなで使っていきましょうという流れになっています。これは日本だけではなくアメリカの国立生物工学情報センター（NCBI）や国立衛生研究所（NIH）などの人たちが主導しているのですが、それにいろいろな形で世界の研究者が関わり、いろいろなルールを決めていき、そしてワーキンググループで、この記号についてどうするかという議論が行われています。その中で、最初に 2015 年に、このように糖鎖の記号を表しましょうという論文が出ました。その後もアップデートされるごとに論文を出すと同時に、研究者たちがインターネットでフリーで見られるような形で情報を提供しています。

これだけ見ると、糖の領域はこのように書いてあればみんな分かるではないかというイメージを持つかもしれません。しかし、私が 14、15 年前に糖の領域の学会に行ったときには、この記号がポスターやスライドによってみんなばらばらだったのです。このグルコースは青で書いてありますが、赤の丸がグルコースという脚注があるポスターなども何割かありました。私自身、その構造に興味を持っているので、一時期、学会のポスターや講演で糖鎖の研究者の人たちがどんな記号を使っているのか、SNFG の前には CFG、オックスフォード、IUPAC などいろいろな記号があるので、それぞれ統計を出していたのですが、非常にばらばらでした。それがだんだん統一化されてきています。普及啓蒙活動が功を奏していると思うのですが、昨年参加した学会では、かなりの人たちがこれに従っている、またはこれに近い形で書くようになっていました。

次は文字列表記です。IUPAC などで糖鎖を表現するときがあるのですが、糖鎖の構造のデータ管理をするときに、やはりコンピューターを使って糖鎖の構造を表せないとなかなか管理が難しいです。そういうときに、国際糖鎖構造リポジトリを作ることになったのですが、そのときに使える糖鎖構造を表すユニークな文字列が必要だということで開発が始まったのが Web3 Unique Representation of Carbohydrate Structures（WURCS）です。ユニーク性があるので、同一の糖鎖構造かどうかが、この文字列を使うことによって判断できるようになってきています。糖鎖の構造は、分岐構造、繰り返し構造、環状構造、分岐とリピートがあるなど非常に複雑です（図 8）。化学構造と化合物は、ある意味似ているものがあります。既に DDBJ という塩基配列のリポジトリなどでは配列で書けるので、そういうものは比較的やりやすかったのですが、そういう方法が糖鎖では取れなかったので開発したのが WURCS です。



（図 6）



（図 7）

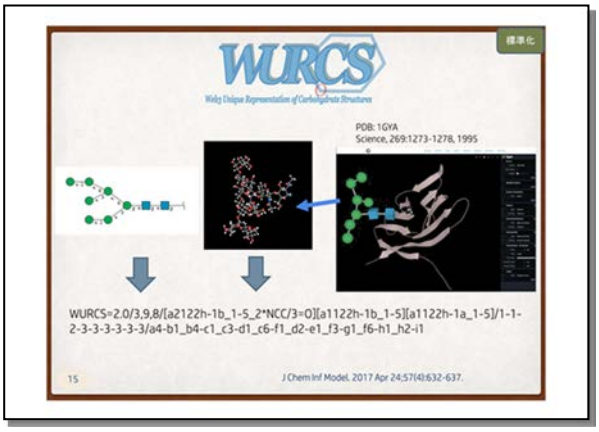
一例をお示しします。図 9 は Protein Data Bank（PDB）です。蛋白質の立体構造のリポジトリですが、蛋白質と三次元の SNFG という表記方法で書いたものがあると、この蛋白質に糖鎖が付いていて、この糖鎖が何なのだろうと取り出したときに、糖鎖の SNFG 記号で表すことができ、原子レベルで三次元で表すことができます。WURCS という表記法に変換することで同じものだと確かめることができ、他のエントリーなどで同じ糖鎖の構造があれば、同じ文字列になることが保証できるというものです。

これまで標準化・共通化の話をしてきました。少し毛色が違いますが、ガイドラインが糖鎖科学の領域で作られてきました。Minimum Information Required for A Glycomics Experiment の頭文字を取って MIRAGE というプロジェクトなのですが、目的としては、文献の糖鎖関連データの品質を向上させましょうというものです。実験をどのようにするかというものではなくて、実験結果を正しく解釈して再現できるようにする。論文化されたデータを誰でもきちんと再現できるようにする。そのためにはどんな情報が必要かということを、MIRAGE のグループに世界中の研究者たちが参加して議論し、いろいろなガイドラインが作られています。そのガイドラインに従って論文やデータベースを利用していきましょうという動きがあります。

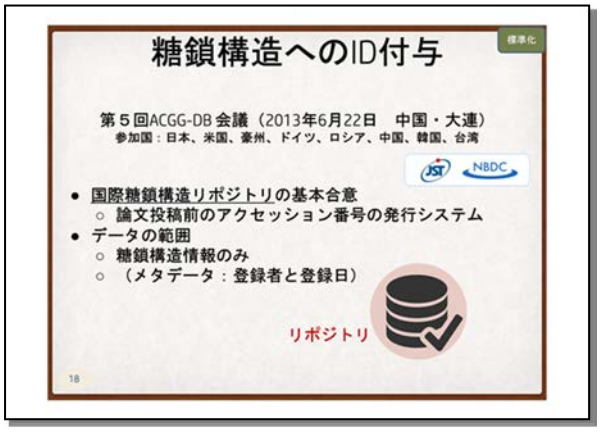
MIRAGE のガイドラインは糖鎖の実験のガイドラインで、サンプルの調整方法、質量分析の実験、マイクロアレイの実験、液体クロマトグラフィの実験などについて定めています。来月にもまた MIRAGE の会議が開かれて、新しいガイドラインについて検討されます。このように必要に応じてガイドラインを作って、それを領域で利用していこうという動きがあります。

図 10 は、糖鎖構造に対して ID を付加してあげようという話です。糖鎖構造にはたくさん書き方がありますが、それに対して ID を付けてあげると、全部同じものだということが分かります。そのために WURCS という文字列を作ってきましたが、その基になったのは 2013 年に中国の大連で開催された ACGG-DB の会議でした。日本をはじめアメリカ、オーストラリア、ドイツ、ロシア、中国、韓国、台湾の人たちが参加し、糖鎖のデータベースやリポジトリ、アクセッション番号が必要だろうという議論がなされました。それ以前にもこういう議論はされていましたが、この会議において、具体的に国際的なリポジトリを 1 個作

（図 8）



（図 9）



（図 10）

ってみんなで利用していきましょうという合意が得られました。データの範囲としては、糖鎖の構造と登録者、登録日ぐらいのシンプルなリポジトリを作り、それを利用しようということになりました。

糖鎖科学のリポジトリ

研究者が実験をして糖鎖の構造などのデータを登録すると、リポジトリはその構造を見て適切なアクセッション番号を研究者に発行します。研究者は、このアクセッション番号を論文に記載することにより、どんな糖鎖を用いたかを明確にします。一方で、それ以外の研究者たちは、論文を読んで、そこにアクセッション番号が記載されていると、このアクセッション番号を基にリポジトリのデータを見ます。論文には自分に馴染みのない構造で書いてあったとしても、リポジトリを見ると、自分にとって分かりやすい表記がされています。そうすることで、どんなものを扱っているかがよく分かるという仕組みです。

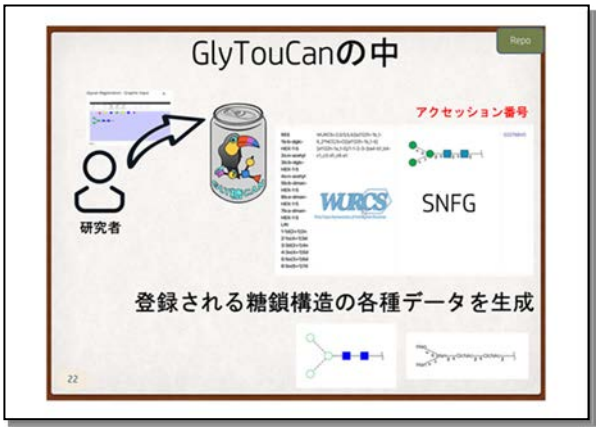
このようなリポジトリとして、GlyTouCan という糖鎖構造のリポジトリを作りました。それから、UniCarb-DR と GlycoPOST という二つのリポジトリがあるのですが、これは最初の頃にお話しした糖鎖構造を決めるための質量分析に関するデータを保存するためのリポジトリです。これらのものを、標準化を利用して作っています。

まず GlyTouCan に関してです（図 11）。これは国際糖鎖構造リポジトリといって、世界で唯一の糖鎖構造に対するリポジトリです。世界中の研究者が自由に登録・利用できるようにしています。研究者には糖鎖構造を書くためのウェブツールを提供しています。それを使ってここでお絵描きをして登録ボタンを押すと、GlyTouCan のシステムが働いて、アクセッション番号を研究者に返すという形になっています。先ほどの WURCS という表記法や GlycoCT、IUPAC など複数の構造でデータを入力しても、同じようにアクセッション番号を返すというシステムです。

図 12 は GlyTouCan の中身です。先ほど言ったように研究者がお絵描きをして、その構造を GlyTouCan の中に投げ込みます。例えば GlycoCT という表記方法のものを GlyTouCan のシステムで WURCS という表記方法に変換したり、SNFG 記号のイメージを作ったり、オックスフォード、IUPAC というさまざまな表記方法で糖鎖構造を表現します。あとは WURCS という文字列が GlyTouCan の中にあるかどうかを検索して、もしあれば、その構造は既に登録されているので、そのアクセッション番号を研究者の方に返します。なければ、新しいアクセッション番号を発行して研究者に返すという形になっています。

図 13 は、UniCarb-DR というアノテーションした質量分析データのリポジトリです。MIRAGE のガイドラインに、サンプル調整の方法、どういう条件で測定したか、スペクトルのピークデータなど、アノテーション済みのデータを登録することができます。この MIRAGE のガイドラインに沿ったデータをウェブの

（図 11）



（図 12）

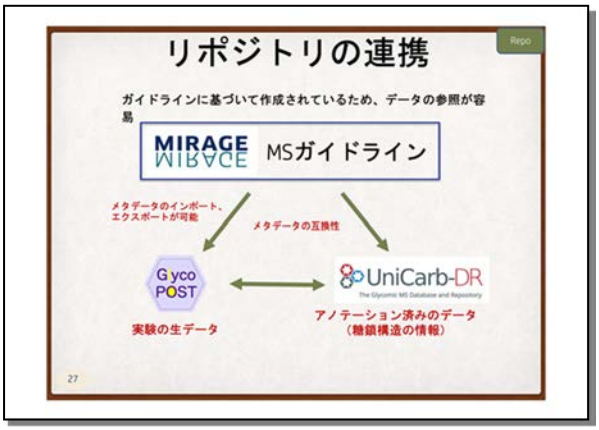
インタフェースで書き、それをエクセル形式でエクスポートすることができます。

図 14 は GycoPOST というリポジトリです。こちらも質量分析データのリポジトリで、MIRAGE のガイドラインに従ってメタデータ等を作っていき、UniCarb-DR で作られたエクセルデータをインポートして利用することもできます。違いは、測定したときに機械から出てきた生データです（図 15）。UniCarb-DR は、人がアノテーションしたデータを入れるリポジトリですが、GycoPOST はアノテーションされていない、生で機械から出てきたデータをそのまま入れているリポジトリです。これらは、MIRAGE のガイドラインに従ってそれぞれアノテーションされたデータと生のデータが作られているので、そのメタデータには互換性があり、それぞれのデータを参照することができます。

次に、分析データのリポジトリです（図 16）。アノテーションが付いているデータを、UniCarb-DR というリポジトリから UniCarb-DB というデータベースにデータをインポートするときに、糖鎖構造に GlyTouCan のアクセッション番号を付けて UniCarb-DB の方に持ってきてデータベースを整理します。UniCarb-DB の方では GlyTouCan のアクセッション番号が付いているので、糖鎖の研究者たちはどんな糖鎖構造なのかがよく分かるようになっています。

図 17 のように、GlyTouCan という糖鎖構造のリポジトリを作って、それを GlyCosmos 、Glycomics@ExPASy 、GlyGen という、日本、スイス、アメリカのプロジェクトが共通して利用していきましょうということで、糖鎖領域の三つのプロジェクトで GlySpace Alliance をつくり、データの共通利用やライセンスについてみんなで協力していく形になっていま

（図 13）



（図 15）

（図 14）

（図 16）

す。それ以外にも、糖のデータベースや、先ほどの PDB、PubChem Compound のデータベースなどに対しても GlyTouCan のアクセッション番号を付けることによって、いろいろな領域において糖鎖の構造がどのようなものかが分かるような取り組みをしています。

まとめ

糖鎖科学におけるデータ管理のために、標準化・共通化をして、糖鎖の記号と文字列表記を決めてアクセッション番号を付け、品質向上のためにガイドラインづくりを行ってきました。これらの標準を使って、GlyTouCan、UniCarb-DR、GycoPOST の各リポジトリを運用・連携しています。これらの整備によって、リポジトリとデータベースを利用した正確な研究データが利用可能になっています。

これらのプロジェクトは、GlyTouCan に関しても現在進行形で研究開発しており、まだまだいろいろな問題があります。ユーザのサポートをしなければいけないなど、運用でも大変なところはありますが、糖鎖の研究者に役立つような環境を提供していきたいということで、私だけではなく、日本だけではなく、世界の研究者が協力して糖鎖領域のデータを良くする取り組みをしています。

（図 17）

●**フロア 1**　高エネルギー加速器研究機構の職員です。物理屋からすると、とてつもなくたくさんのバリエーションがあって気が遠くなるような話ですが、糖鎖以外にもさまざまな物質が世の中にはあると思います。それらの標準化を世界的に統括するような仕組みはあるのでしょうか。例えば物理だと、国際純粋・応用物理学連合（IUPAP）という国際的な委員会のような組織が大きなことを決めていくのですが、化学の世界でも、例えば幾つか流儀があったときに統括したり、足りないところを補ったりするような、仕切る組織があるのでしょうか。

●**山田**　まず化学の方では、IUPAC で元素記号のルールが明確に決められており、糖鎖の方では国際生化学・分子生物学連合（IUBMB）と IUPAC が連携して命名法のガイドライン的なものを出していますが、それだけでは対応できないものが結構あります。あまり悪口を言っては良くないですが、お医者さんで、化学構造式で書いても分からない人、名前で書いても分からない人、分野ごとに自分たちが分かればいい、それで研究が済むという領域の人たちは、それ以外のことは必要ないけれども、論文を読んだときにそれが何なのかが分からなければいけないので、それをどうするか。糖鎖の中でそれができていないということで、大連で開催されたミーティングでそれを明確にしようという動きがありました。

糖鎖の領域には幾つか学会がありますが、学会主導でガイドラインを作るというよりも、ある程度コミュニティが主導になって、「これが必要だ」という動きになる場合が多いです。