糖鎖構造情報学の基盤強化のための糖鎖リポジトリと単糖置換行列の開発

Development of a glycan repository and monosaccharide substitution matrix to strengthen the foundation of glycan informatics

13D5605　藤田晶大　　指導教員　木下聖子

SYNOPSIS

In glycoinformatics, GlyTouCan, the international glycan structure repository, has been instrumental to properly manage the increasing numbers of glycan structures being researched in glycobiology. We have performed a major update on the GlyTouCan batch system to ensure consistency in registering its structures. In GlyTouCan, in order to search for glycan structures, currently only exact match searches have been implemented. However, the similarity of glycans could be better evaluated by creating a more versatile monosaccharide substitution matrix. Therefore, we have created a monosaccharide substitution matrix that takes into consideration not only their 3D structures but also their physicochemical properties. In addition to searching for glycan structures, this is the first attempt to quantify the similarity of monosaccharides. We have also evaluated this substitution matrix on the human data in GlyTouCan, thus demonstrating its validity and applicability.

Keywords: Glycan, Glycan repository, Glycoinformatics, Monosaccharide substitution matrix

背景・目的

糖鎖はDNAやタンパク質に次ぐ重要な生体高分子として知られており、免疫反応や細胞間認識に置いて重要な役割を担っている。一方で糖鎖構造は多様性を持つことが知られており、糖鎖の機能解明のためにより深い糖鎖構造の理解が不可欠である。そこで、国際構造糖鎖リポジトリであるGlyTouCan[1]が開発され、糖鎖構造データを管理できるシステムができた。GlyTouCanはすでに12万もの糖鎖構造に対し、アクセッション番号を割り当て、管理している。

GlyTouCanは幅広い糖鎖を記述可能なWURCS[2]形式を用いて糖鎖構造を管理しているが、WURCSは仕様・実装レベルでバージョンアップが続いており、古い仕様やソフトウェアに基づいた構造情報の整理などが必要となっていた。GlyTouCanでは登録時に即座にアクセッションナンバーを割り当てていたが、バージョン違いのソフトウェアで生成したWURCSの場合は同じ構造でも別のアクセッションナンバーが発行されていたため、今までよりも重複に注意して登録するようにチェックする必要があった。さらに、他の糖鎖データベースとの関連性を管理するためのシステムも同様に更新する必要性があった。

また、GlyTouCanはすべての糖鎖構造を一元的に管理することを目指しており、GlyTouCanのデータが糖鎖構造情報学の基盤になるため、GlyTouCanのデータに対して、類似検索を行うなどの糖鎖解析の基礎づくりが必要であった。GlyTouCanを含め糖鎖生物学において、単糖をシンボルとして扱い、糖鎖を木構造として記述する。この木構造の検索や比較を行う際に、一致や部分一致を評価することは容易いであるが、類似性を評価することは難しい。糖鎖を検索する方法にペアワイズアラインメントのスコアを算出するKCaM[3]がある。KCaMは類似構造検索に用いることができ、KCaMを用いて糖結合のスコア行列[4]の概念が提案されていた。このスコア行列は二糖とその間の結合をまとめて最小単位として扱い、BLOSUMアルゴリズム[5]をベースに開発された。しかし、二糖の組み合わせが膨大となり、実用的な利用は難しかった。本論文ではより幅広い糖鎖に対応するため、単糖の立体構造情報に主に注目し、物理化学的な見地からも単糖の類似度を決定し、数値化することが必要と考え、単糖の

$$V_{rmsd(s1,s2)} < F_n = 0.1 \times log10 \times \{ n \times (n-1) \times (n-2) \times 0.2 \} + 1 \quad \cdots (1)$$

$$Distance_{(s1,s2,n,m)} = V_{rmsd(s1,s2)} + \sum_{i=n+1}^{m} F_i \quad \cdots (2)$$

図1：Vは分子s1とs2間の最大共通部分の重ね合わせた際のRMSDで、nは原子数とした。 （1）式を満たすことを必要条件とする。 立体構造間の距離Distanceは（2）式で与えられる。 （1）の右式で閾値となっているF_nを$n+1$からmまで足し合わせることによって最大共通部分が大きければ大きいほど距離が近いように定義している。mは全ての単糖立体構造のバリエーションを重ね合わせた中で最も大きい原子の数を示す。

立体構造比較を算出する高速なアルゴリズムを考案した。単糖の立体構造比較、単糖の官能基の比較、単糖の親水性の比較を行うことで単糖間の類似性が数値化できると考えた。

本研究ではこれらの特徴を含めた新規な単糖置換行列を算出した。また、これを用いてGlyTouCanに登録されているヒトの糖鎖データを解析し、その有用性を示した。本研究は、糖鎖情報科学の基盤となる糖鎖リポジトリの正しい運用と糖鎖の最小単位となる単糖の特性を生かした糖鎖構造検索の手法を改良し、今後の糖鎖構造における解析基盤に大いに貢献するものである。

方法

GlyTouCanに登録されている構造情報を最新状態に更新するため、GlyTouCanシステムに組み込まれているWURCSを扱うライブラリの更新を行った。同時にすでに登録されていたWURCSに対してもライブラリを用いて新しいWURCSへ再変換を行った。また、古いWURCSが登録される可能性があったため、即時登録から登録申請の方式へ変更し、新しい登録システムの開発を行った。新しい登録システムでは申請された糖鎖構造をそのまま保存し、WURCSへ変換し正規化したものを登録する。これらの処理を定期的に行うバッチプログラムも開発した。ユーザーはその処理状況を確認できるようUser Submissions Pageも用意した。

すでに登録されていた糖鎖構造情報も新しいWURCSに更新するため、古い情報や新しいバージョンのライブラリで間違っているとわかった構造やすでに重複して登録されていたエントリーなどの整理を行った。 必要でなくなったエントリーは削除ではなくアーカイブとし、アーカイブしたエントリーは公表した。また、国際的にエントリーの管理を行うため、管理者に近いユーザーとして連携するデータベースの機関をパートナーとする、パートナーシステムを開発した。パートナーはAPIを通して所属データベースの情報をメタデータとして登録できる。

また、GlyTouCanデータに検索を行う際、現状では一致検索、部分一致検索のみが実装されており、類似構造検索ツールが存在しなかった。 類似構造検索は二つの糖鎖構造の比較の問題に還元できるため、これを解決する必要があった。糖鎖構造間の比較を行うため、糖鎖の最小単位である単糖の類似度を算出する。そこで、代表的な単糖を取りまとめたSymbol Nomenclature for Glycans (SNFG)[6]に列挙されている単糖に着目することとした。 SNFGの単糖情報にはアノマー情報が含まれないためアノマー情報を付加し、１１８の単糖情報を収集した。また、対応する単糖の立体構造情報をwww.glycosciences.de[7]より取得した。単糖の構造は柔軟であり、多くの立体構造を取り得ると考えられる。そこで、glycosciences.deから得たある瞬間の立体構造のみならず、取り得る立体構造情報を算出するため、糖鎖の立体構造予測ソフトShape[8]を利用した。その結果、１１８の単糖それぞれの立体構造の予測結果を得た。 この段階で一単糖あたり数千のpdbファイルを得た。これらの立体構造にはほぼ同じ構造のものが含まれていたため、クラスタリングによってグループ分けを行い、グループの代表を選ぶことで単糖あたりの立体構造の正規化を行った。結果として単糖あたり１０～２００の立体構造を得た。この段階で、得られた単糖の立体構造が間違っていないか、環構造のイス型の異性体を含め確認を行った。

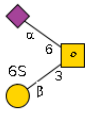
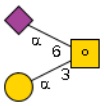
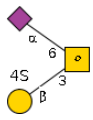
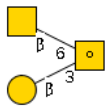
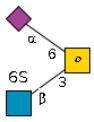
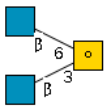
次に、単糖の立体構造の比較のために、得られた立体構造データの重ね合わせを行うこととし、そのためにKabschアルゴリズム[9]を用いた。また、立体構造間で重ね合わせを再帰的に繰り返し最大共通部分を探索するライブラリを開発し、すべての単糖間で立体構造間の比較を行った。最大共通部分を抽出し、最大共通部分のRMSDを求めた結果、全単糖対の立体構造間のRMSD値を得た。

RMSD値は重ね合わせた構造の距離として考えることができるが、重ね合わせた原子の数が多ければ多いほど、より近い単糖であると考えることができる。そこで、図１に示すように原子数が多いものほど距離が近くなるようにRMSDと原子数を元に立体構造間の距離(Distance)を定義して計算した 。これにより、RMSD値をDistanceに変換し、さらに、各単糖のすべての立体構造間のDistance郡の平均を取ることで単糖間距離に変換した。これによって単糖の立体構造から単糖間の距離行列を得ることができた。

さらに、単糖の分子量、水酸基、アミノ基、カルボキシル基を数え上げ、それぞれの特性の差を計算した簡易的な行列も作成した。この行列とDistanceの距離行列を合算して一つの距離行列にまとめ、距離を類似度に変換し、最終的な単糖置換行列を算出した。

表１：GlyTouCanから得られるヒトの糖鎖構造データを対象にG05768VSをアライメントし、アライメントスコアで並べ替えた結果である。KCaM-1.0は従来のKCaMの結果であり、KCaM-1.1は改良したKCaMの結果を示している。

Neu5Gc　GlcNAc
Neu5Ac　Gal
GalNAc　GalNAc-ol

QUERY		G05768VS	
KCaM-1.0		KCaM-1.1	
			
G75947FL	88.3	G75947FL	87
			
G05235KE	85	G54953IZ	70.7
			
G29386CS	76.7	G05235KE	69.7
			
G54953IZ	76.7	G65285QO	69.3
			
G65285QO	76.7	G65790DQ	69.3
			
G65447BW	76.7	G45129XM	67
			
G93802CF	76.7	G61833XV	65.3

結果・考察

新たなGlyTouCan登録システムを開発し適応させた。申請された糖鎖構造が古いものであればWURCSが異なることがあるが、同じ構造を指すWURCSを登録できる。システムが申請された糖鎖構造を糖鎖構造と認識できない場合は登録保留とし、GlyTouCanのバッチシステムがアップデートした際に改めて処理を行う。バッチシステムのアップデートはサーバの再起動を伴わず、柔軟に行えるようになり、問題が発生した場合に直ちに対応が取れるようになった。パートナーはエントリーに関連データベース情報を追加できるようになった。

作成した単糖置換行列の有用性を検証するため、KCaMに単糖置換行列を導入し、アライメントスコアの変化を調べた。つまり、単糖置換行列を用いることで糖鎖を検索する際に今までより優位に糖鎖を検索できるかを検証した。既存のKCaMでは単糖の一致スコアは0または1で判断されたが、一致しない場合でも単糖置換行列より類似スコアとして0から1までの値で計算するように再開発を行った。

表1に従来のKCaM-1.0と、単糖置換行列を導入したKCaM-1.1の比較結果を示した。どちらのバージョンでもG75947FLが1位になったが、KCaM-1.1では置換された単糖が異なれば検索結果のスコアが変動することが確認できた。KCaM-1.0では、例えば76.7のスコ

アが異なる構造データでも同じスコアになることがあるが、KCaM-1.1では置換された単糖が異なる場合、置換スコアが新たに計算されたことによって細かく順位づけされた。

表２は表１で示したスコアの根拠となる単糖置換行列の数値を示したものである。基本的にa-Neu5Gcは非常に高い類似度を示すa-Neu5Acに置換される。表１の二位の結果が特徴的であり、KCaM-1.0では、b-GalNAcがa-GalNAcに置換されたG05235KEに高いスコアが付いているが、これは単糖を"GalNAc"として認識し一致したものとして、結合情報としてのアノマー情報が不一致したものとして計算されるためである。一方、KCaM-1.1ではb-GalNAcがb-GlcNAcに置換されたG54953IZが二位に順位づけされた。これは表２に示すようにb-GalNAcとb-GlcNAcの類似度が高いからである。

表２：表１で示した結果の中で使用された単糖置換行列の値である。

置換単糖対		類似度
a-Neu5Gc	a-Neu5Ac	0.73056
b-GalNAc	b-GlcNAc	0.58792
b-GalNAc	a-GalNAc	0.55319
b-GalNAc	b-Gal	0.53561
b-GalNAc	a-Gal	0.53560

表２で示した単糖対のうち、b-GalNAcとa-GalNAcの違いは一位の炭素に結合する水酸基の向きが異なることであり、b-GalNAcとb-GlcNAcの違いは四位の炭素に結合する水酸基の向きが異なるためである。不一致になる原子数は同数なので同程度の違いかと考えられるが、立体構造を元に類似度を算出する今回の手法によって微妙な差を算出できる。検索結果としてもアノマー情報が変わると結合情報が変わり、二糖構造の二面角が変わるため全体の構造も大きく変化する。構造学的により正しい順位となっていることがわかる。

総括

増え続ける糖鎖構造に対しより強固なGlyTouCanの登録システムとWURCSの仕様と実装のアップデートに柔軟に対応できる管理システムになったことで登録された糖鎖構造を最新のWURCSで表記し、表記のエラーをWURCSレベルで検知できるようになった。今後は未知の問題が発生した場合も柔軟に対応することができることが見込まれる。合わせて、単糖の立体構造と物理化学的な特性に基づいた単糖置換行列の開発を行った。KCaMによる応用例を示し、機械的によく似た構造の順位付けが可能になることを示した。

今後はSNFG以外の単糖にも対応することで他生物種間で類似の糖鎖グループがないかを調査することや、糖鎖のビッグデータに対して糖鎖間の類似性スコアを計算し、クラスタリング手法と組み合わせて糖鎖の分類などができる。また、マルチプルアライメントを行うMCAW[10]など他の糖鎖構造分析ツールに応用することでKCaMで分類を行った場合と比較検討などをして構造情報をベースとした分類を可能にする。今後、リン酸など単糖の修飾へ対応することで比較が高精度で行えるようになると考えられる。

今までの糖鎖構造情報学において示されていたスコア行列は二糖構造に限定されていたもので実用は難しかったが、単糖にアノマー情報を加え単糖間の類似性を定義することで比較的容易に構造の類似性を計算できるようになった。糖鎖構造の類似性を測定する際に決まった方法は今までなかったが、形式化した単糖置換行列を用いることで、異なるツールやアルゴリズムを用いた比較でも同じ条件で精度の比較が行えるようになる。将来的に単糖置換行列は対応する単糖を増やし、数値を改善していくことでGlyTouCanに蓄積され続ける糖鎖構造のグライコーム解析への道筋となる。

参考文献

[1] Matsubara M, Aoki-Kinoshita KF, Aoki NP, Yamada I, Narimatsu H. J Chem Inf Model. 57, 632-637, 2017.

[2] Fujita A, Aoki NP, Daisuke S, Matsubara M, Tsuchiya S, Shiota M, Ono T, Yamada I, Aoki-Kinoshita KF. Nucleic Acids Research. 49, D1529-D1533, 2020.

[3] Aoki KF, Yamaguchi A, Ueda N, Akutsu T, Mamitsuka H, Goto S, Kanehisa M. Nucleic Acids Research. 32, 267-72, 2004.

[4] Aoki AF, Mamitsuka H, Akutsu T, Kanehisa M, Bioinformatics, 21(8), 1457–1463, 2004.

[5] Henikoff S, Henikoff JG, Proc Natl Acad Sci USA, 89(22), 10915-10919, 1992.

[6] https://www.ncbi.nlm.nih.gov/glycans/snfg.html

[7] Lütteke T, Bohne-Lang A, Loss A, Goetz T, Frank M, von der Lieth CW. Glycobiology, 16(5), 71R-81R, 2006.

[8] Rosen J, Miguet L, Perez P. Journal of Cheminformatics, 1(1), 1-16, 2009.

[9] Kabsch W. Acta Crystallographica A34, 827–828, 1976.

[10] Hosoda M, Takahashi Y, Shiota M, Shinmachi D, Inomoto R, Hashimoto S, Aoki-Kinoshita KF. Carbohydr Res. 464, 44-56, 2018