博士学位請求論文

論文題目　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

専 攻 名　　生命情報工学

学籍番号　　13D5605

氏　　名　　藤田　晶大

指導教員　　木下　聖子

創 価 大 学 大 学 院
工 学 研 究 科

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

２０２２年２月

藤田 晶大

目次

第 1 章　はじめに 7

第 2 章　背景 9

2.1 糖鎖構造 9

2.1.1 単糖構造 10

2.1.2 グリコシド結合 13

2.1.3 糖鎖構造の比較 14

2.1.4 糖鎖構造形式 15

KEGG Chemical Function（KCF） 15

GlycoCT 16

LInear Notation for Unique description of Carbohydrate Sequences（LINUCS） 16

Web3 Unique Representation of Carbohydrate Structures（WURCS） 17

2.2 構造データベースやリポジトリ 17

2.2.1 GenBank 17

2.2.2 PDB 17

2.2.3 GlyTouCan（国際糖鎖構造リポジトリ） 18

2.3 セマンティック・ウェブ 19

2.3.1 RDF 20

2.3.2 オントロジー 21

第 3 章　GlyTouCan の継続的開発 23

3.1 背景 23
3.2 方法 25
3.3 結果 26
3.3.1 ユーザー登録 26
3.3.2 糖鎖構造情報の登録 26
3.3.3 登録状況の確認 27
3.3.4 API による登録 27
3.3.5 パートナープログラム 28
3.3.6 アーカイブ処理 29
3.4 本章の考察 29

第 4 章 発展的な糖鎖構造比較 33

4.1 方法 33
4.1.1 単糖データ 33
4.1.2 ソフトウェア 34
4.1.3 ハードウェア 36
4.1.4 手順 36
4.1.5 単糖置換行列を用いた検索 37
KCaM への実装 37
4.2 結果 38
4.2.1 データ 38
4.2.2 G05768VS の検索について 39
4.2.3 G71832QJ の検索について 42
4.2.4 G00054MO の検索について 46
4.2.5 まとめ 49
4.3 本章の考察 49

第 5 章 全体の考察 51

5.1 まとめ 51
5.2 今後の課題 51
5.3 展望 53

参考文献 57

付録 63

1.4 データ 63
1.4.1 用いた単糖のリスト 63
1.4.2 用いた単糖データ 63
1.4.3 作成した単糖置換行列 63
1.4.4 検索に用いたヒト糖鎖構造データ 63
1.5 単糖置換行列の作成に必要なソースコード 64

第 1 章

はじめに

バイオインフォマティクスは生体高分子を情報工学によって読み解く試みとして、核酸やタンパク質を始めに発達した。この発達の大きな要因は、核酸やタンパク質は化学物質であり、原子をベースとした分子構造として書くことができる一方で、セントラルドグマを貫く遺伝暗号として、塩基配列とアミノ酸配列として構造を直鎖のテキストとして略記できることが大きい。文字数の限られた直鎖のテキストは情報工学との相性が良かった。テキストマイニングとして知られたアルゴリズムを応用できたからである。そして、核酸はらせん構造であり、タンパク質の立体構造は多様であるが、そのアミノ酸配列と相関があることが分かっている。このことは遺伝子やタンパク質の高速で詳細な比較を可能にした。それぞれの遺伝子やタンパク質の特異性は他の遺伝子やタンパク質とどれだけ異なっているかを計算することで把握されるので、新規に発見されたものが既存の生体高分子にあるか、ないか、どれだけ似ているのかを即座に分析できるのはこれらのおかげである。しかし、重要な生体高分子は核酸とタンパク質のみならず、糖や脂質などが存在しており、バイオインフォマティクスの課題はポストゲノムとよばれる時代が始まり、糖鎖と脂質が新たなターゲットとなった。特に糖は、エネルギー産生の中間体という役割だけではなく、免疫反応や細胞間認識など多岐にわたる。加えて、多くの生体高分子と細胞が、単糖がグリコシド結合で連なる糖鎖を持っていることからも生体の複雑さと糖との関連が伺える。糖鎖の配列は単糖をノード、グリコシド結合をエッジとした木構造を略記することにより、木構造を配列情報として一般に広く記述されている。

核酸やタンパク質のように糖鎖でも高速に比較を行うには、核酸やタンパク質と比較するとそ

の配列情報の取り扱いに未熟な点があるため改善の余地がある。糖鎖構造を構成する要素は単糖とグリコシド結合である。単糖は数多くの種類があり、これらを特徴づけた数量としての表現が求められているし、水酸基が脱水縮合して行われるグリコシド結合には不斉炭素が絡んだ立体化学と結合位置によって大まかな様式が決まり、単糖の種類によっても微妙に変化する。加えて、単糖はグリコシド結合を形成するための水酸基を複数持つため糖鎖は枝分かれ構造を持つ。このように糖鎖構造は多様であり、糖鎖は多機能な情報分子として働く。

これらの糖鎖配列の特徴にあわせて糖鎖情報をあつかう手法がグライコインフォマティクスとして発展した。糖鎖構造データを分析・解析するツールや、糖鎖構造を表記するための様々な形式、糖鎖を管理するためのデータベースが生まれた。様々な糖鎖構造を表記できる形式が作り出され、糖鎖構造配列の表記法は成熟してきているし、糖鎖構造の比較のためのアルゴリズムはいくつか提案されている。しかし、糖鎖構造配列向けの分析手法は考案されているが、糖鎖分子の特異性をデータとして反映させることは少なかった。木構造を構成する要素のうち、単糖間の比較、グリコシド結合間の比較については糖鎖に最適化されているとは言い難く改善の余地がある。

新しい糖鎖構造を分析する際、即座に基本的な性質を知るためには核酸やタンパク質と同様に既存の構造に対して速やかな比較ができる必要がある。

本論文では既存の糖鎖データを管理し、データベース間で横断して使用可能性を広げる国際糖鎖構造リポジトリである GlyTouCan の継続的開発と増え続ける糖鎖構造に対して効果的な比較法の基礎を担う、単糖置換行列の作成とその効果について述べる。

開発したプログラムのソースコードや結果として得られたデータのサイズが大きいのでホスティングサービスの gitlab のリンクを参照することにする。簡便なアクセスのため、参照した URL のリンクを集めたページを用意した。`https://research-assets.gitlab.io/thesis-defense/sourcecode/`

第 2 章

背景

2.1　糖鎖構造

糖鎖構造はいくつかの単糖が鎖状にグリコシド結合でつながった構造を持つものが一般的に知られている。糖鎖を原子レベルの分子構造ではなく、単糖を最小ユニットとした概念図として書く場合、図 2.1 のように糖鎖構造を構成する単糖をノード、グリコシド結合をエッジとした木構造として図示される。

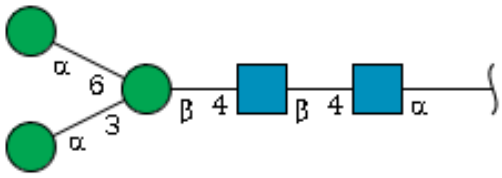


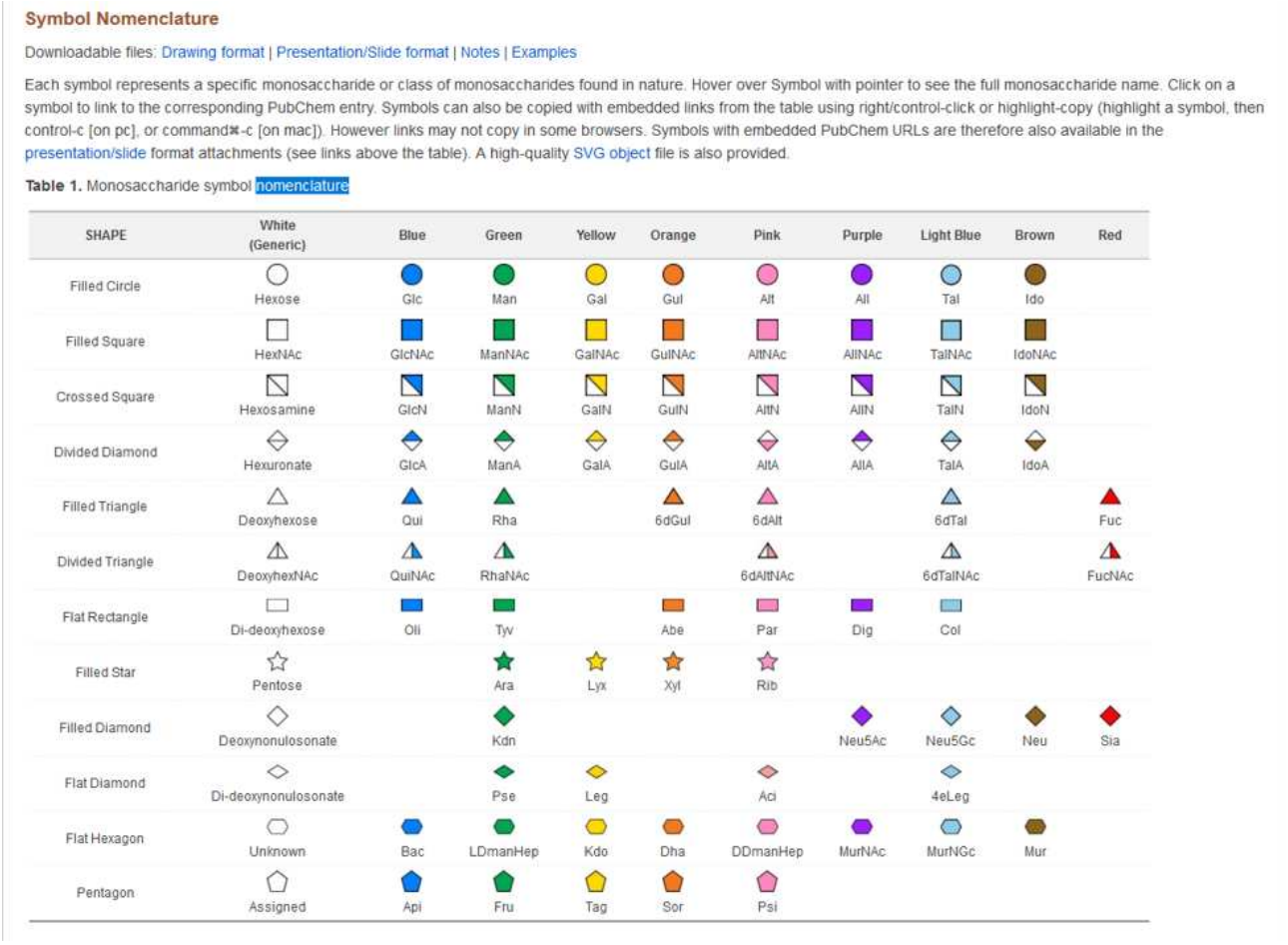
図 2.1　糖鎖構造を図示した例、ノードの画像（シンボル）は単糖を意味しており、この画像では SNFG によって定められているシンボルに従っている。グリコシド結合はシンボル同士をつなぐエッジとして書かれ、エッジに結合情報が文字情報で書かれている。

図中で用いられる単糖のシンボルは Symbol Nomenclature for Glycans [1] (SNFG) によって主要な単糖について定義されており、これを用いることが一般的である。糖鎖構造は DNA 配列やアミノ酸配列と異なり木構造の配列表現になるため、単一の表現が難しく表現のための形式が増えた。また、DNA 配列やアミノ酸配列と比較して最小ユニットの単糖の数が多く、その表記ゆれなども

発生し、データベース間での糖鎖の比較や、情報の統合、構造情報の対応づけが困難になっていた。本論文では、” 糖鎖構造” といった場合には実際の糖鎖分子の構造を指し、ある形式によって表記された文字列を” 糖鎖構造配列” と呼び区別をする。

2.1.1 単糖構造

単糖情報は単糖データベースである MonosaccharideDB[2] によって 776 が管理されていが、SNFG で管理される単糖のシンボルは図 2.2 のように形と色をもちいて炭素数と修飾基によって大まかに分類される。ペントース、ヘキソース、ヘキソサミン、デオキシヘキソース、ウロン酸、シアル酸である。



Symbol Nomenclature

Downloadable files: Drawing format | Presentation/Slide format | Notes | Examples

Each symbol represents a specific monosaccharide or class of monosaccharides found in nature. Hover over Symbol with pointer to see the full monosaccharide name. Click on a symbol to link to the corresponding PubChem entry. Symbols can also be copied with embedded links from the table using right/control-click or highlight-copy (highlight a symbol, then control-c [on pc], or command-c [on mac]). However links may not copy in some browsers. Symbols with embedded PubChem URLs are therefore also available in the presentation/slide format attachments (see links above the table). A high-quality SVG object file is also provided.

Table 1. Monosaccharide symbol nomenclature

SHAPE	White (Generic)	Blue	Green	Yellow	Orange	Pink	Purple	Light Blue	Brown	Red
Filled Circle	Hexose	Glc	Man	Gal	Gul	Alt	All	Tal	Ido	
Filled Square	HexNAc	GlcNAc	ManNAc	GalNAc	GulNAc	AltNAc	AllNAc	TalNAc	IdoNAc	
Crossed Square	Hexosamine	GlcN	ManN	GalN	GulN	AltN	AllN	TalN	IdoN	
Divided Diamond	Hexuronate	GlcA	ManA	GalA	GulA	AltA	AllA	TalA	IdoA	
Filled Triangle	Deoxyhexose	Qui	Rha		6dGul	6dAlt		6dTal		Fuc
Divided Triangle	DeoxyhexNAc	QuiNAc	RhaNAc			6dAltNAc		6dTalNAc		FucNAc
Flat Rectangle	Di-deoxyhexose	Oli	Tyv		Abe	Par	Dig	Col		
Filled Star	Pentose		Ara	Lyx	Xyl	Rib				
Filled Diamond	Deoxynonulosonate		Kdn				Neu5Ac	Neu5Gc	Neu	Sia
Flat Diamond	Di-deoxynonulosonate		Pse	Leg		Aci		4eLeg		
Flat Hexagon	Unknown	Bac	LDmanHep	Kdo	Dha	DDmanHep	MurNAc	MurNGc	Mur	
Pentagon	Assigned	Api	Fru	Tag	Sor	Psi				

図 2.2 SNFG により単糖名に対して定義されるシンボル、構造的な特徴から色と形で分類されており、一目で把握しやすいように工夫されている

単糖の異性体は最も番号が大きい不斉炭素原子によって決定され、脊椎動物において、フコー

ス、イズロン酸を除き、D 型が天然に存在する。単糖は溶液中で遊離した状態では環状構造と線形構造が平衡状態にあるが、糖鎖構造中の単糖は、通常、環状構造のみを考えれば良い。環化にあたっては線形構造の末端がヘミアセタール構造を取ることにより環状化する。この構造変化でできた一位のヒドロキシル基が糖鎖形成時のグリコシル結合に使われる。環状構造を取る時、1 位の炭素が不斉中心となり、エピマー化がおきる。この区別のため糖鎖中の単糖の名称の前には α-、β-をつける。また、単糖は不斉炭素に基づく異性体が存在している。例えば、グルコースとガラクトースはジアステレオマーの一種であり、数ある低分子の中で化学構造的に良く似た分子であるが、生体内では異なる役割を果たす。そして、通常単糖が環状構造をとるとき、五員環は柔軟な配座をとり、六員環の構造はイス型を占めると考えてよい。ただし、六員環の構造は 38 個のバリエーションに多様に分類 [3] されており、酵素反応中の遷移状態であるとか、硫酸修飾が過度になった場合などはイス型から外れると考えられている [4]。このことは、例えば GAG に分類される糖鎖の立体構造を考えるとき、常以上に単糖の立体構造の検討が必要なことを示唆している。イス型の単糖を前提に単糖の立体構造を考える場合でも二種の 1C4、4C1 と呼ばれるイス型が存在しており、どちらのイス型がよりエネルギー的に安定しているかを検討する必要がある。

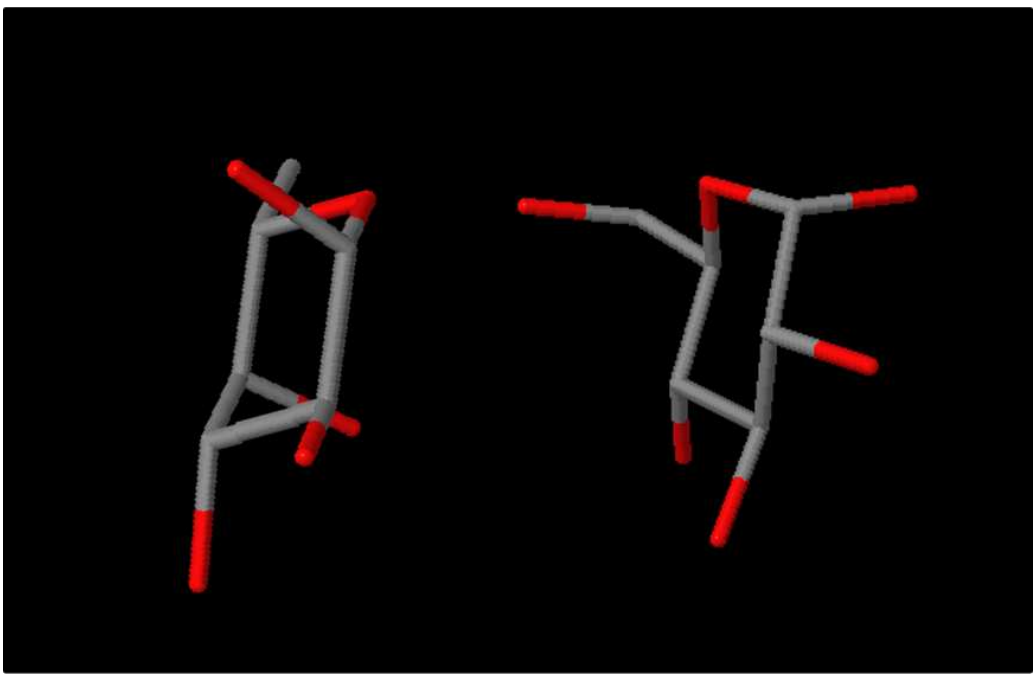


図 2.3　左図の 1C4 のイス型を持つ α−L-Fucp と右図の 4C1 のイス型を持つ α−D-Glcp の立体構造、イス型が表裏の関係に回転しているのがわかる

我々は普段意識せずに単糖を名称で概念的に識別しているが、その環化した立体構造を思い浮かべることは困難になる。

以上は立体構造の観点から見た単糖の特徴について述べたが、単糖同士がどれだけ似ているのかという問題は、一方的な見地から観察するのではなく、多角的な指標から判断する必要がある。それに、糖鎖情報学では立体構造の座標を直接用いることよりも抽象化されたスカラーの量を用いることのほうが情報学的に応用する際にハードルが低いため、望ましいと言える。

図 2.4 に示すように、単糖の物理化学的な特徴量は PubChem[5] のエントリーページに”Chemical and Physical Properties” として示される。これらはケモインフォマティックスのツールによりコンピュータによって計算されることが一般的であるが、不斉炭素中心を考慮するものが少なく、単糖について用いると分子量が同じ単糖は同じ特徴量となってしまう。

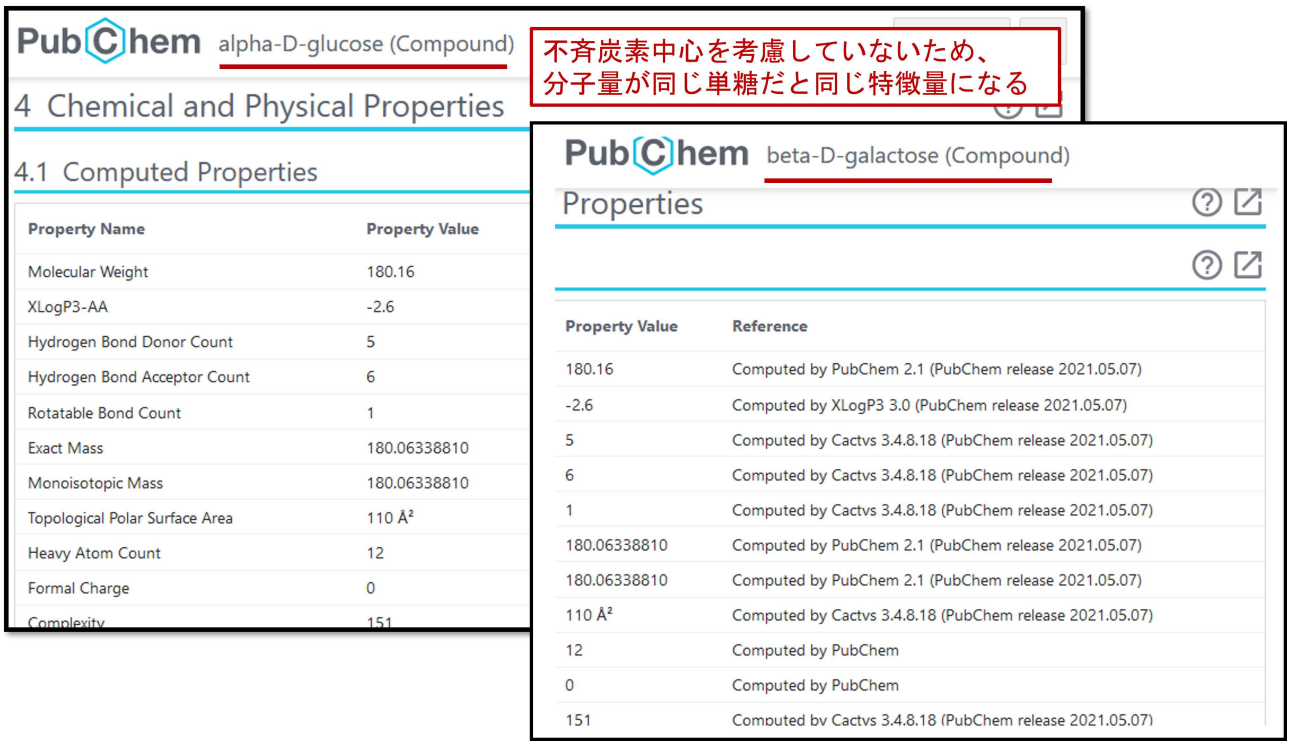


図 2.4　異なる単糖であるのに全く同じ特徴量を提示する PubChem の”Chemical and Physical Properties”

よって、単糖を特徴づける数値を得るためには情報化学を単純に応用するのみならず、単糖独自のものさしが必要になる。現状では、単糖を構造上の特徴で区別することは可能だが、構造上の特徴を数量化し、合理的に比較することは難しい。このことは単糖を識別する決定的な数値情報は立体構造の座標情報ということを示唆している。

2.1.2 グリコシド結合

糖鎖構造を構成するグリコシド結合は一般に O-グリコシド結合と呼ばれる脱水縮合を伴う様式である。グリコシド結合を中心に単糖同士はある程度回転する。その回転の程度は二面角として計算し、二次元の数値化することができる。この二面角のプロットはラマチャンドランマップとして知られ、ペプチドの立体構造解析でよく知られている。

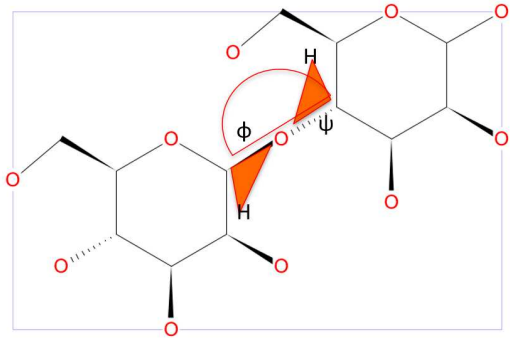


図 2.5 水素を用いる二面角の定義

二面角の定義はグリコシド結合を中心に水素を使わない $C_{a+1} - C_a - O - C_b - C_{b+1}$ と水素を含む $H - C - O - C - H$ があるが、これは実験的に立体構造が決まるとき、結晶構造解析の場合は水素が見ない場合が多く、核磁気共鳴法（NMR 法）による場合は水素がよく見えるからである。図 2.5 に示すように二面角はそれぞれ ϕ と ψ と呼ぶことになっている。

これらの情報を計算によって求め、まとめたデータベースが Glycosciences.de によって GlycoMaps Database [6] として公開されている。ラマチャンドランマップの詳細を図 2.6 のエントリーを閲覧することができる。ただし、単糖の組み合わせの数だけ存在するため計算された単糖ペアは限定的である。

二単糖間の立体構造のバリエーションを示すラマチャンドランマップは単糖のペアと結合位置によってエネルギー的に安定している位置が異なるため、グリコシド結合の立体構造は結合様式以上に多様であるということが言える。

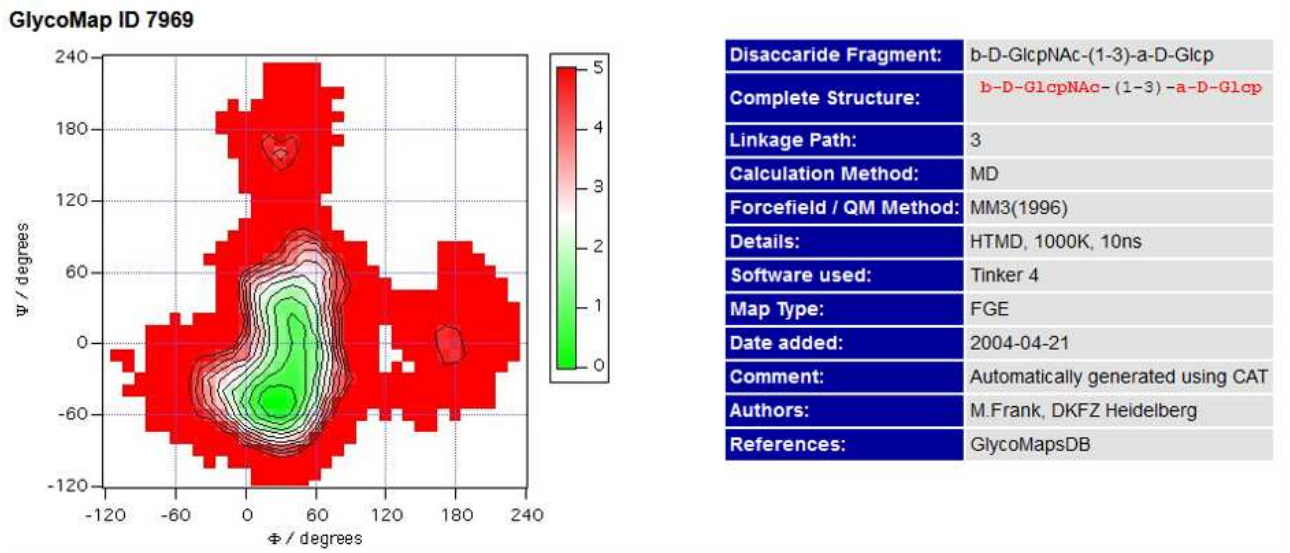


図 2.6 GlycoMaps Database で公開されているラマチャンドランマップの一例

2.1.3 糖鎖構造の比較

糖鎖構造は木構造の配列として書けるので配列情報をアライメントすることにより、情報学的に糖鎖の比較が可能になる。アライメントスコアによって類似性を評価する事ができるが、糖鎖が木構造配列であることはその比較にも DNA 配列やアミノ酸配列と異なる新たな方法論が必要なことを意味する。糖鎖構造のペアワイズアラインメントのために KCaM[7] が開発されており、類似構造検索に用いるために糖結合のスコア行列 [8] 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 [9] 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。この先行研究における重要な成果は糖鎖構造を糖鎖構造配列に書いたり、絵に描くだけではなく、糖鎖構造の木構造をデータ構造に落とし込み情報工学的手法を応用できること示した点にある。つまり、KCaM のアルゴリズムは糖鎖構造向けに整えられたソフトウェアであったが、単糖と単糖の比較、結合と結合の比較、に関してその重みづけにおいて改善の余地があることが分かっていた。

2.1.4 糖鎖構造形式

多くの糖鎖構造形式があるが、ここでは本論文に関わる形式のみを紹介する。また、本論文では糖鎖構造形式とは木構造で記述される形式とし、一般の化合物として表記できるものとは区別する。

KEGG Chemical Function（KCF）

KCF [10] には 2 つのバリエーションがあり、化学構造を記述するものと糖鎖を記述するために拡張したものがある。

```
ENTRY    G00005    Glycan
NODE     6
         1  PP-Dol   15   1
         2  GlcNAc    8   1
         3  GlcNAc    0   1
         4  Man      -9   1
         5  Man     -16   7
         6  Man     -16  -6
EDGE     5
         1   2:a1   1
         2   3:b1   2:4
         3   4:b1   3:4
         4   5:a1   4:6
         5   6:a1   4:3
///
```

図 2.7 KCF による糖鎖構造配列の例

図 2.7 に例として示す。KCF 形式では NODE ブロックと EDGE ブロックに分けて記述し、NODE 名に当たる単糖名は自由に書くことができる。画像を表示するための座標が NODE ブロックに書かれており、座標上に木構造を書いて表示する前提を感じることができる。

GlycoCT

GlycoCT [11] は KCF 形式と同じく複数行で書き表し、ノード情報である単糖情報を RES、エッジにあたる結合情報を LIN として分けて書く。図 2.8 に GlycoCT の例を示す。LIN では単なるグリコシド結合以外にも対応するようアルファベットを用いて結合様式に情報を付加している。また単糖情報は MonosaccharideDB が管理する表記と同等の表記を用いる。

```
RES
1b:a-dglc-HEX-1:5
2s:n-acetyl
3b:b-dglc-HEX-1:5
4s:n-acetyl
5b:b-dman-HEX-1:5
6b:a-dman-HEX-1:5
7b:a-dman-HEX-1:5
LIN
1:1d(2+1)2n
2:1o(4+1)3d
3:3d(2+1)4n
4:3o(4+1)5d
5:5o(3+1)6d
6:5o(6+1)7d
```

図 2.8 GlycoCT による糖鎖構造配列の例

LInear Notation for Unique description of Carbohydrate Sequences（LINUCS)

LINUCS [12] は Glycosciences.de によって開発された形式で、単糖の表記法は IUPAC のルールを用いる。図 2.9 に表記の例を示す。糖鎖の複雑な分岐構造を一意に表記するために、かっこの入れ子構造によって結合情報を階層的に表記する。

```
[][PP-Dol]{[(0+1)][a-GlcNAc]{[(4+1)][b-GlcNAc]{[(4+1)][b-Man]{[(6+1)][a-
Man]{}[(3+1)][a-Man]{}}}}}
```

図 2.9 LINUCS による糖鎖構造配列の例

Web3 Unique Representation of Carbohydrate Structures（WURCS）

WURCS [13] は GlyTouCan プロジェクトで発案された形式で、Web 上でユニークな文字列として扱うことを前提に開発された。加えて、単糖名を IUPAC 名や慣習による文字列に依存せず、単糖の化学構造を記述することで明確に構造を指す。この時、Stereocode を拡張した ResidueCode を用いる。図 2.10 に WURCS による糖鎖構造配列を示す。

```
WURCS=2.0/4,5,4/[a2122h-1a_1-5_2*NCC/3=O][a2122h-1b_1-5_2*NCC/3=O]
[a1122h-1b_1-5][a1122h-1a_1-5]/1-2-3-4-4/a4-b1_b4-c1_c3-d1_c6-e1
```

図 2.10　WURCS による糖鎖構造配列の例

2.2　構造データベースやリポジトリ

ここでは塩基配列情報とタンパク質がどのように管理されているのかを概観し、続いて国際糖鎖構造リポジトリについて述べる

2.2.1　GenBank

GenBank[14] は National Center for Biotechnology Information (NCBI) が提供する塩基配列を蓄積し提供するデータベースである。GenBank は European Bioinformatics Institute（EBI）が提供する European Molecular Biology Laboratory（EMBL）や日本 DNA データバンク（DDBJ; DNA Data Bank of Japan）と密接に連携し、相互にデータを共有している。これらの三大国際 DNA データバンクの枠組みは International Nucleotide Sequence Database Collaboration (INSDC) として知られている。図 2.11 に示すように、塩基配列に対応するアミノ酸配列も保管されており、BLAST[15, 16] による相同性検索が可能になっている。

2.2.2　PDB

Protein Data Bank (PDB) [17] はタンパク質、核酸、糖鎖などの生体高分子の立体構造情報を蓄積しているデータベースで、日米欧の国際的な協力のもと成立している。それぞれ、米国の Protein

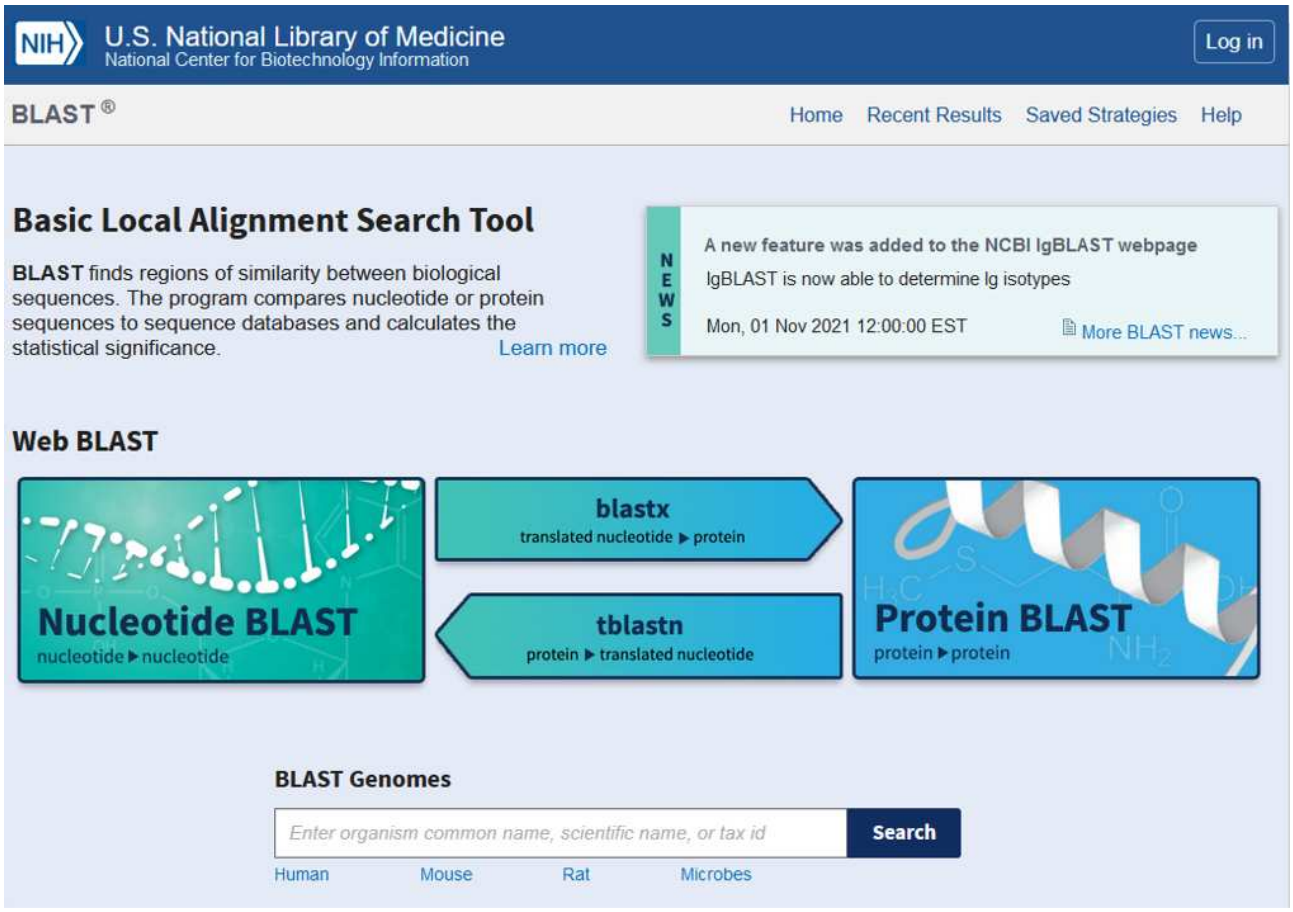


図 2.11　NCBI が提供する BLAST による相同性検索の Web インターフェース

Data Bank (RCSB PDB)、欧州の Protein Data Bank in Europe (PDBe)、日本の Protein Data Bank Japan (PDBj) である。図 2.12 に示すように PDB では通常の絞り込み検索のほかに”Sequence Navigator” が用意されており、BLAST による相同性検索とともに同じ配列部分の立体構造の重ね合わせも行うことができる。

また、タンパク質の立体構造を論文上に公開する際、PDB に登録することが必須となっており、これによって文献上の情報から最新の構造情報へアクセスすることが容易になっている。ちなみに糖鎖の立体構造情報データは 100 もないことが報告 [18] されている。

2.2.3　GlyTouCan（国際糖鎖構造リポジトリ）

糖鎖構造情報が増加の一途をたどる一方で情報が各データベースに分散し、論文上や、会議で議論する上でどの糖鎖構造を指しているかなど、糖鎖構造を扱う上で問題があることが指摘されていた。2013 年、中国・大連において開催された ACGG − DB 会議 [19] において、すべての糖

図 2.12　PDBj が提供する Sequence Navigator による相動性検索の Web インターフェース

鎖構造に固有のアクセッションナンバーを付与することが合意され、2014 年に国際糖鎖構造リポジトリとして GlyTouCan[20, 21] が開発・公開された。図 2.13 に示す GlyTouCan は Genbank や PDB と同様に広く公共のために無料で使うことができる。

糖鎖構造を GlyTouCan に対し問い合わせる際、現状は一致や部分一致に対応している。相同性検索のような、配列類似性を効果的に評価することは現状は難しいが、一括管理した糖鎖構造のうち似ているものを探すということは糖鎖情報学における喫緊の課題である。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が、今後も糖鎖構造を一元的に管理することを目指しているため、まだまだ増え続けることが予想される。GlyTouCan のデータが糖鎖構造情報学の基盤になるため、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。

2.3　セマンティック・ウェブ

セマンティック・ウェブ [22] は World Wide Web で用いられる技術の標準化団体である World Wide Web Consortium (W3C) によって提唱され、標準化が進んでいるプロジェクトである。今ま

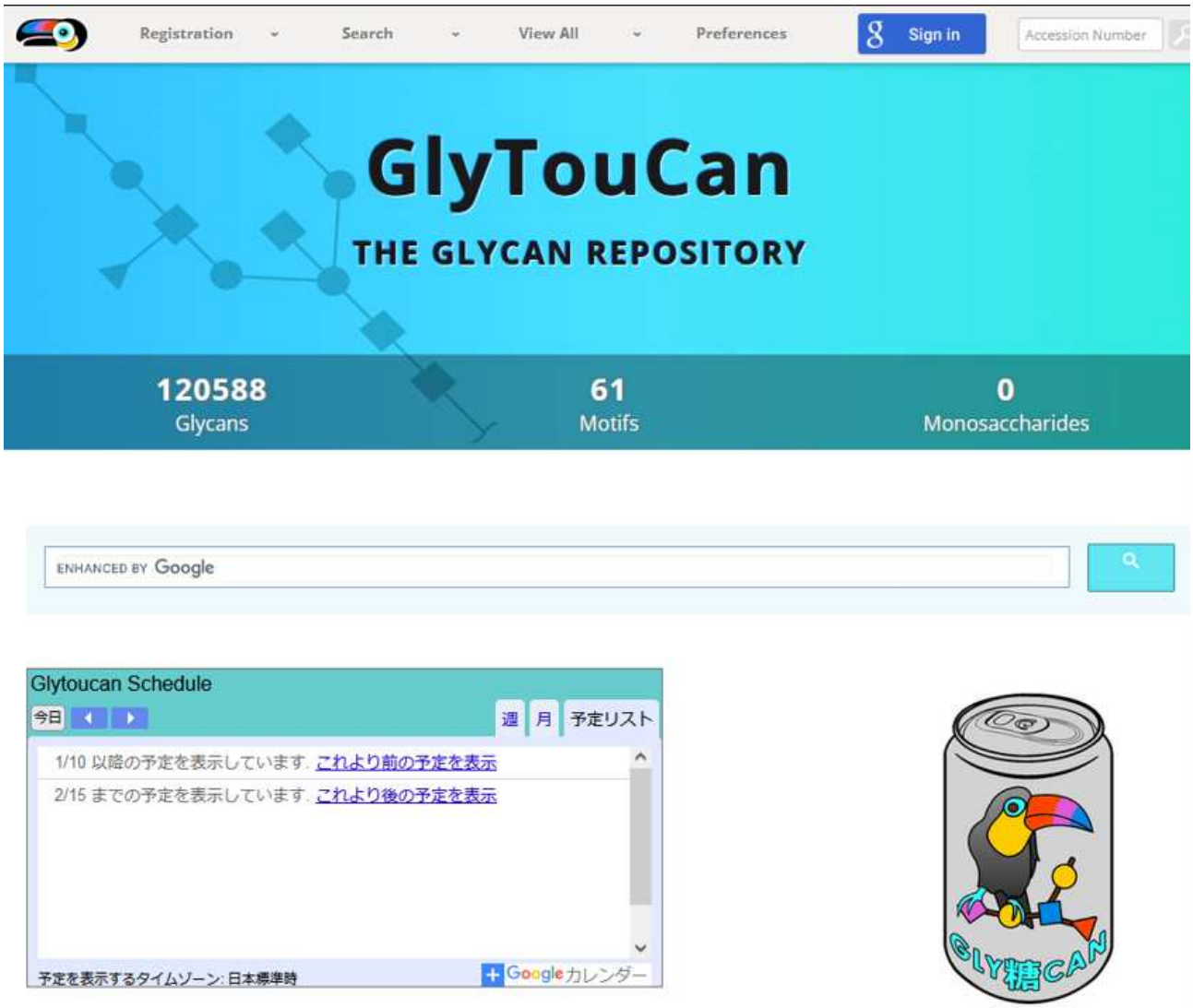


図 2.13 国際的な合意に基づいて開発された国際糖鎖構造リポジトリとしての役割を担う GlyTouCan の Web インターフェース

でのテキストデータの集積でしかないウェブではなく、ただのテキストではなく、XML 文章により記述され Resource Description Framework (RDF) [23] や Web Ontology Language (OWL) [24] など によって意味付けされたデータ構造にすることで人がテキストを解釈することなくデータの解釈を可能にさせる試みである。

2.3.1 RDF

RDF はデータリソースに対するメタデータを記述する枠組みである。RDF に従って生物学的なリソースを記述することで web 上のリソースを横断して検索することができるようになるため、

データベースはセマンティックウェブ化に取り組んでいる。RDF はデータリソースを主語、述語、目的語の” トリプル” の単位で記述することを要求する。RDF が扱うことができるデータは XML で定められた URI とリテラルであり、既存の Web 上のリソースを扱うには URL を用いることもでき、既存の Web リソースに対しメタデータを付加することも可能であるし、既存の Web リソースを分解して RDF による意味づけされたグラフ構造として記述することもできる。以下の図 2.14 は GlyTouCan が用いている RDF の一部である。主語の URI にアクセッションナンバーが含まれているため、人間は URI が糖鎖のことと理解できるが、マシンは理解できないため、マシンのためにそれが糖質（saccharide）だと指定している。

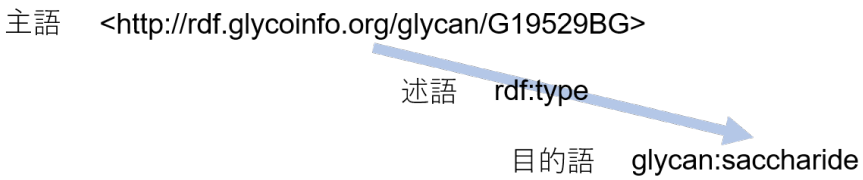


図 2.14 GlyTouCan で最も使われている RDF 表現のうちの一つ、主語の URI が糖質であることを示し、主語に数珠上につながるリソースにアクセスすることで、この糖質の情報を得ることができることをマシンに示唆している。

トリプルの目的語は主語になることもできるため、名前空間を共有する整理された RDF のデータは数珠繋ぎ状に連鎖的につながり、関連するデータを一度に問い合わせることが可能になる。トリプルはデータベースであるトリプルストアに保管され、問い合わせ言語である SPARQL [25] によって検索されることが一般的である。トリプルのグラフ構造をどのように設計するのかによって、SPARQL による問い合わせる効率が変わってくる [26] ことが知られており、トリプルをどのように形式化するのかは重要な問題である。

2.3.2 オントロジー

情報工学におけるオントロジーは Web 上の概念の形式化といってよい。オントロジーを明確にすることにより、RDF によって書かれたデータリソースがより明確になる。セマンティック・ウェブにおけるオントロジーの定義には RDF の言語拡張である OWL を用いる。OWL はトリプルを記述する際の URI の形式に対してそれがどういう意味の URI なのかを定義するための言語と

いってよい。糖鎖情報を RDF として記述するために必要な OWL の定義は GlycoRDF [27] として報告されている。図 2.14 で使われた”glycan:saccharide” という目的語は GlycoRDF の中の OWL ファイル [28] の一つにより定義されており、共通のルールとして認識されている。

第 3 章

GlyTouCan の継続的開発

GlyTouCan は WURCS 形式を用いて糖鎖構造を管理している。WURCS を扱う際には WURCS の仕様に従って Java によって開発された WURCSFrameWork[29] を用いている。WURCS は過去に 1 度、仕様レベルでのバージョンアップと日常的な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。この問題は WURCS による糖鎖構造をより強力に管理する理由となった。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。

さらに GlyTouCan は曖昧な糖鎖を指すことができる WURCS を巧みに用いることで図 3.1 のように糖鎖の曖昧さのレベルを定義し、糖鎖構造全体を整理することを可能にした。

3.1　背景

GlyTouCan のデータは、セマンティック Web に基づいており、全てのユーザーデータと糖鎖データが RDF トリプルストアに格納される。ユーザーはいつでも GlyTouCan のエンドポイント [30] に向けて RDF クエリ言語である SPARQL を用いて問い合わせを行うことができる。RDF クエリ言語を用いない一般的なユーザーは Web ユーザーインターフェースを用いてブラウザから

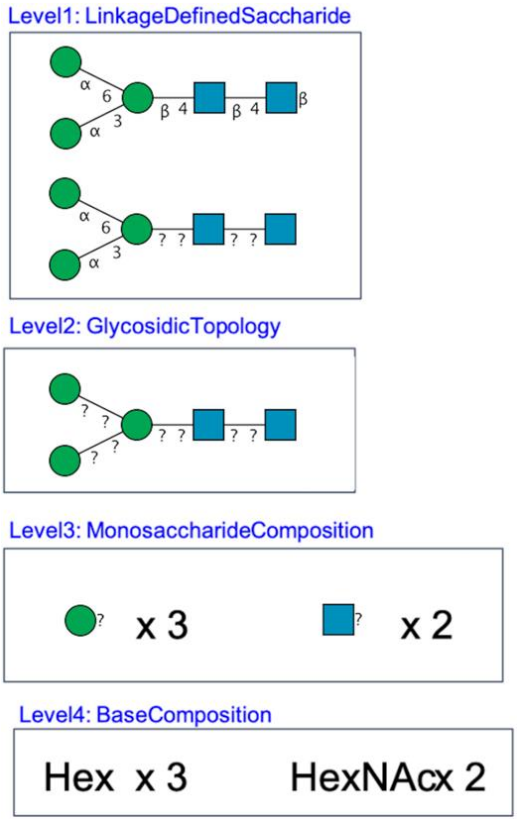


図 3.1 GlyTouCan に登録できる糖鎖構造の例、さまざまなレベルのあいまいさを定義できる

アクセスする。Web コンポーネントによって開発を行い、表示に用いられている。糖鎖構造を登録するにあたり、いままでは GlycoCT と WURCS による登録のみによって受け付け、即時にアクセッションナンバーを発行していたが、将来的な他の形式による登録やデータに対する検証を確かなものにする必要があった。また、入力された糖鎖構造データに関して一度しか検証を行うことができず、将来的に追加の検証が必要になった場合や、糖鎖表記がより進化し、データを修正したい場合などに柔軟に修正を行うことが困難だった。また、稀なケースだが、同時に同じ構造が登録された場合、同じ構造に複数の ID が発行されることがあった。これらの理由から GlyTouCan3.0 が開発されることになった。伴って既存のデータに対してもアーカイブの処理を行った。また、GlyCosmos Glycoscience Portal[31] が、糖鎖関連情報を統合するための Web ポータルとして正式にリリースされた。GlyCosmos には、リポジトリのセクションとデータリソースのセクションが含まれていて、GlyTouCan、GlycoPOST[32]、そして最近では UniCarb-DR[33] がリポジトリセクションのメンバーになった。これにより、GlyTouCan を簡素化し、検索機能の一部を GlyCosmos に移動することになった。そのため、質量やモチーフ、種による検索は GlyTouCan から削除され、

GlyCosmos の GlycanSearch で利用できるようなった。

3.2 方法

問題点を解決するため、新しい登録フローの設計をおこなった。以前のバージョンと比較して ID 発行の迅速さを犠牲にする代わりに慎重な登録手順を実行するような設計を行った。図 3.2 では開発を行う新しい登録システムの概略を示す。入力されたすべてのデータ、検証結果、生成された画像、および ID は、新しいユーザーの送信ページに一覧表示される。今後はバッチプログラムを修正することで追加の検証や、柔軟な修正が可能になる。

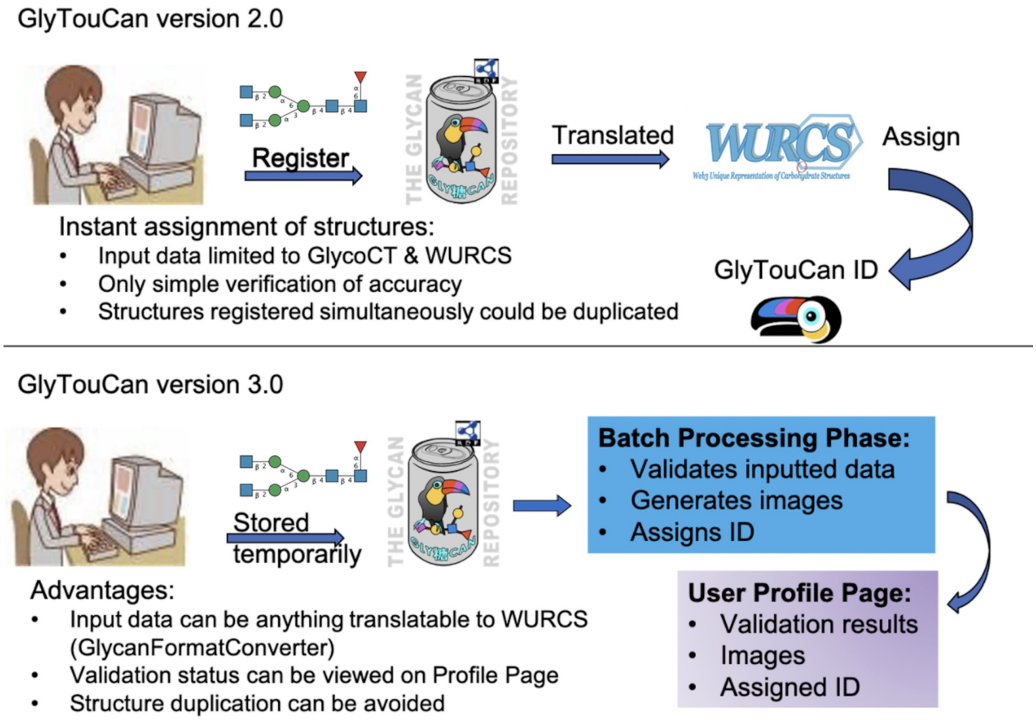


図 3.2 GlyTouCan の新しい登録手順、上部の GlyTouCan2.0 では、ID の登録、変換、および割り当てがすぐに実行されていたが、下部の GlyTouCan3.0 では、登録されたグリカンには最初にハッシュキーが割り当てられ、ユーザーの入力データを保存する。登録に関する処理は定期的なバッチプログラムで行われる。

また、誤ったエントリーだとわかったものについてはアーカイブの処理を行った。

3.3 結果

ここでは新しい登録プロセスが進行する様子を解説する。ユーザーは最初に Google アカウントを使用してログインすることができる。次に、以下の手順を実行する必要がある。

3.3.1 ユーザー登録

ユーザーは新しい構造を登録するために API キーを生成する必要がある。図 3.3 に示すように”user profile page”（`https://glytoucan.org/Users/profile`）では API キーを確認でき、生成されていない場合は、”Generate API Key” のリンクをクリックする事で新たに生成・更新できる。コントリビューター ID と API キーは、GlyTouCan への API アクセスに使用されるため、ユーザーはコンピュータープログラムから GlyTouCan を登録および検索できる。

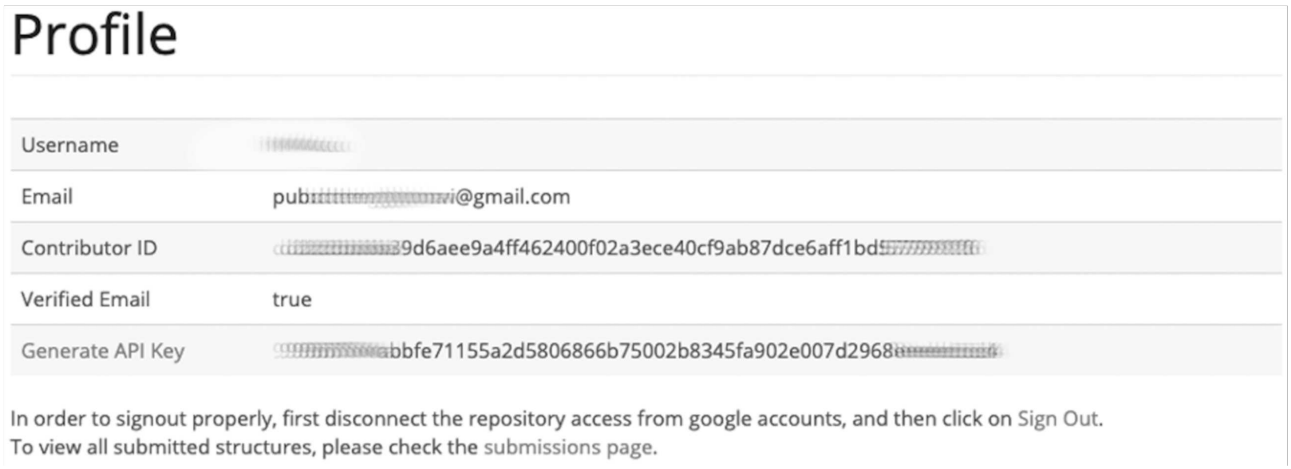


図 3.3 ”user profile page” では、登録済みのメールアドレス、コントリビューター ID、API キーなどの登録情報を確認できる。

3.3.2 糖鎖構造情報の登録

バージョン 3.0 では、GlyTouCan に登録されたすべてのデータに、サーバー上で最初にハッシュキーが割り当てられます。このハッシュキーには、データ、ユーザー情報、および登録の日時が紐付けられ、ハッシュキー自体も受付番号として機能する。

3.3.3 登録状況の確認

グラフィックツールまたは GlyTouCan でサポートされているテキスト形式で糖鎖構造情報を登録した後、ユーザーは図 3.4 に示すように ”submission page” で確認できる。ハッシュキーは即時に発行され、糖鎖構造の検証をするバッチプログラムが問題なく終了すればアクセッションナンバーが発行される。すでに登録されている場合には、割り当てられているアクセッション番号が表示される。

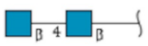
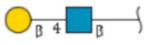
Date ▲	Submission Ref	Sequence	Accession Number
Wed, 09 Sep 2020 12:15:48 GMT	9fa0e3f3ec...	WURCS=2.0/1,2,1/[a2122h-1b_1-5_2*NCC/3=O]/1-1/a4-b1	G42666HT  G42666HT
Wed, 09 Sep 2020 11:58:54 GMT	9a181416e7...	WURCS=2.0/2,2,1/[a2122h-1b_1-5_2*NCC/3=O][a2112h-1b_1-5]/1-2/a4-b1	G00055MO  G00055MO
			G54173SV

図 3.4 ”submission page” で確認できる登録情報の例、データは送信日時に基づいて並べ替えられ、Submission Ref は、登録時に生成されるハッシュキーになる。Sequence には、ユーザーが送信した糖鎖構造データが表示される。バッチ処理でデータが有効であることが確認されると、アクセッション番号とグリカン画像が右端の列に表示される。

3.3.4 API による登録

GlyTouCan では、ブラウザを用いた糖鎖構造の登録の他に、プログラミング言語や、コマンドラインから登録できる API をつかうことができるようになった。API のリクエストの最もシンプルな方法は Listing 3.1 に示すコマンドを用いることで利用できる。

Listing 3.1 登録を行うコマンド

```
curl -X POST --header 'Content-Type: application/json' \
             --header 'Accept: application/json' \
             --user 'Contributer-ID:API-Key' \
```

```
4               -d '{ "sequence":"WURCS=2.0/1,1,0/[a2122h-1x_1-5]/1/" }' \
5               'https://api.glytoucan.org/glycan/register'
```

これに伴ってブラウザからの登録も API を経由するようになり今まで異なる登録方法が存在していた登録システムであったが、サーバーは全ての登録リクエストをこの API を通して受け付けることになった。情報学的に curl による API リクエストのコマンドは API のリファレンスリクエストと見なす慣習があり、このコマンドを元に、開発者は自身の糖鎖構造情報データを Java などのプログラミング言語から登録することができる。

3.3.5 パートナープログラム

パートナープログラムは、独自の糖鎖構造情報を管理する組織向けの特殊な機能である。図 3.5 に GlyTouCan パートナープログラムと、パートナー ID を GlyTouCan に登録するこのプロセスを示す。

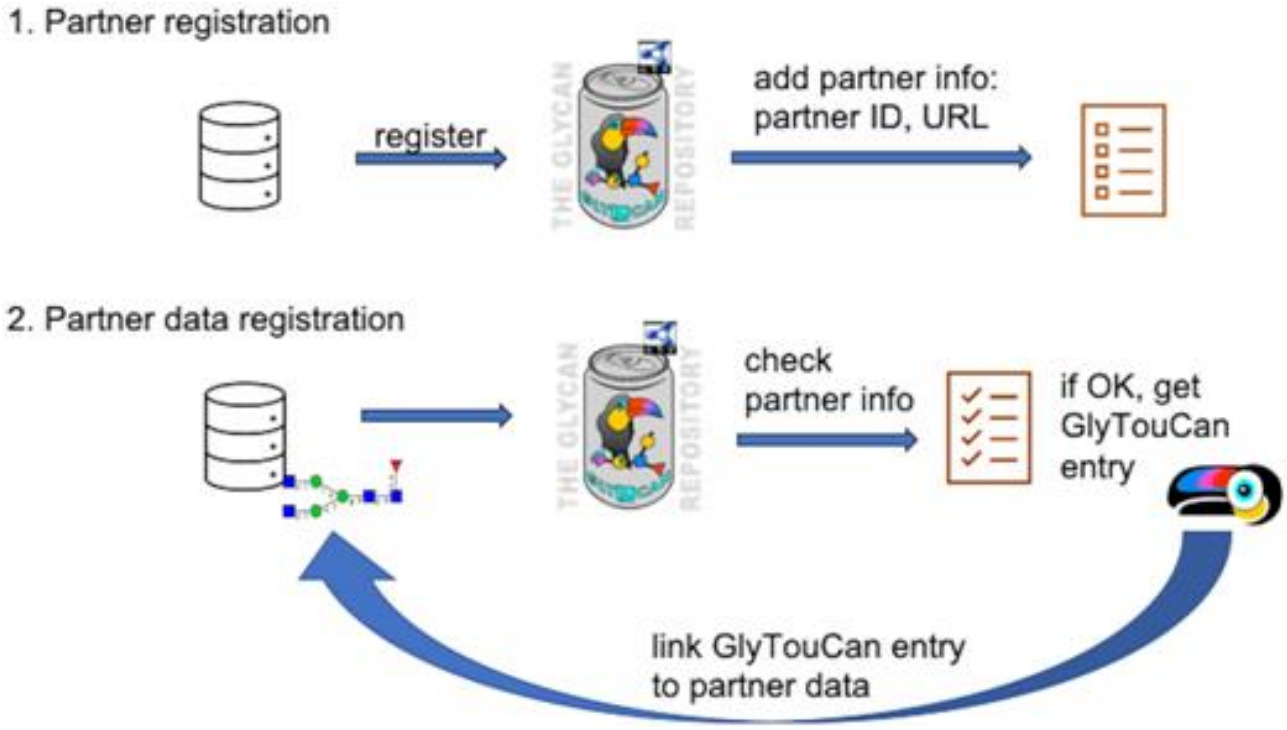


図 3.5 パートナープログラムの概略を示す。データベース管理者は、コントリビューター ID とデータベースの URL を提供することで、パートナーとして登録できる。パートナーは GlyTouCan ID に対応する糖鎖構造情報を登録できる。

主に、糖鎖の登録の自動化、およびパートナーが管理する Web サイトへのリンクを編集できる機能を備えている。このプログラムの主な目標は、糖鎖生物学コミュニティの間で、リンクされたデータネットワークを拡大することである。糖鎖構造のデータベース管理者が、GlyTouCan パートナーになった場合、特別な API をパートナーに提供し、パートナーはこのインターフェイスを使用して、GlyTouCan の特定のグリカンエントリへのリンクを追加、変更、および削除ができる。パートナーがアクセスすると、GlyTouCan はユーザーがパートナーメンバーであるかどうかを確認し、糖鎖構造を登録するときに、パートナーが管理するデータベースの ID を糖鎖構造情報と一緒に送信できる。図 3.6 に示すようにすでにパートナーによって GlyTouCan のエントリーに対して外部データベースの ID が紐づけられている。より多くのデータベース管理者が GlyTouCan に情報を追加できることで、ユーザーは関心のある糖鎖について詳細な情報を見つけることが期待できる。

3.3.6 アーカイブ処理

WURCSFrameWORK のアップデートによる WURCS データの更新と同じ WURCS に対して複数の ID が振られていた問題の解決のためにエントリーのアーカイブを行った。図 3.7 に示すようにアーカイブされたエントリーは上部に”This entry has been archived” と表示されアーカイブの理由に相当する処理のログを閲覧することができる。

3.4 本章の考察

この新しいシステムの利点は、ユーザーが GlycoCT および WURCS 形式に縛られることなく、糖鎖構造を登録できる点である。GlycanFormatConverter[34] が開発され、それをバッチプログラムに組み込むことで、サポートされている変換可能なフォーマットを登録できるようになった。現在、IUPAC 形式と KCF 形式が登録可能になっていて、今後さらに他の形式にも対応する予定である。バッチプログラムは、アクセッション番号を割り当てるために、入力された糖鎖構造情報を最新のライブラリで WURCS 形式に変換をおこなう。今回のアップデートでは既存の糖鎖構造情報のエラーや WURCS を扱うライブラリの更新によって見つかった適切な構造を示さないエン

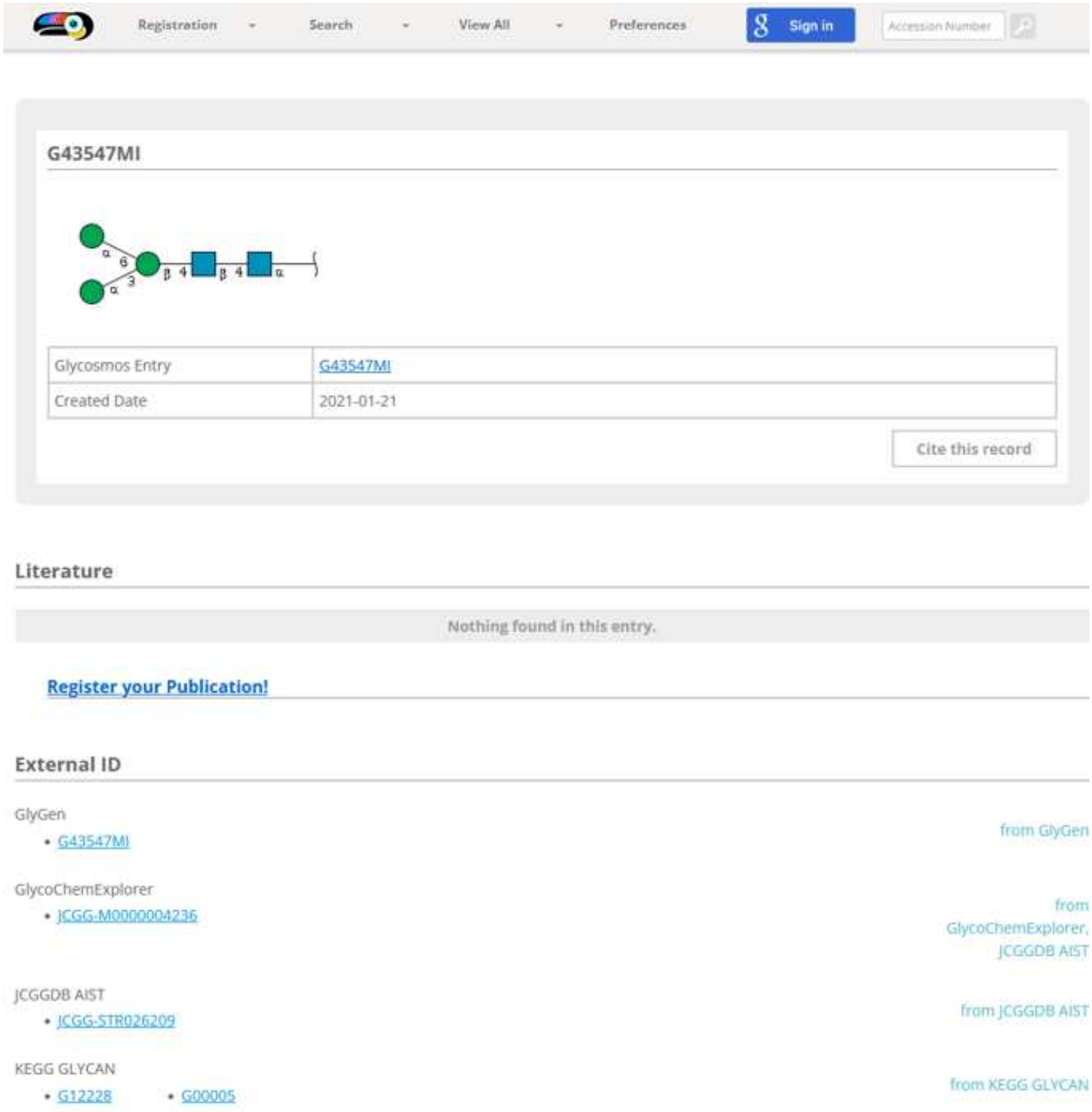


図 3.6 パートナーによって情報が追加されているエントリー、GlyTouCan のアクセッションナンバーに対し、External ID として、パートナーが管理するデータベースの ID が紐づけられている。

トリーを整理した。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と対応する有効なエントリーのリストを公開した。さらに、GlyCosmos には、GlyCosmos Glycans データリソースの下に、GlyTouCan で検証された糖鎖のリストが含まれるようになった。

また、パートナー機能を追加し、GlyTouCan 管理者がほのデータベースの結びつきを管理する

のではなく、データベース管理者が GlyTouCan との紐づけを管理することで、確かな情報を確実に管理できるようになった。

重要なことは、ユーザーの観点から、システムの観点から、アクセッション番号の割り当ての待機時間が増加しているにもかかわらず、新しい GlyTouCan システムはより安定して動作し、登録プロセス上のログをより詳細に保持でき、ユーザーは登録したデータを簡単に追跡できることである。また、運用上のバグや不具合を速やかに修正できる仕組みとなった。WURCS やそのライブラリの開発速度やその安定化にともなって、GlyTouCan でも柔軟なアップデートが必要になった。今後は関連するツールが開発されている速度に合わせて、新しいアップデートによって GlyTouCan を適切な状態で保守することができる。

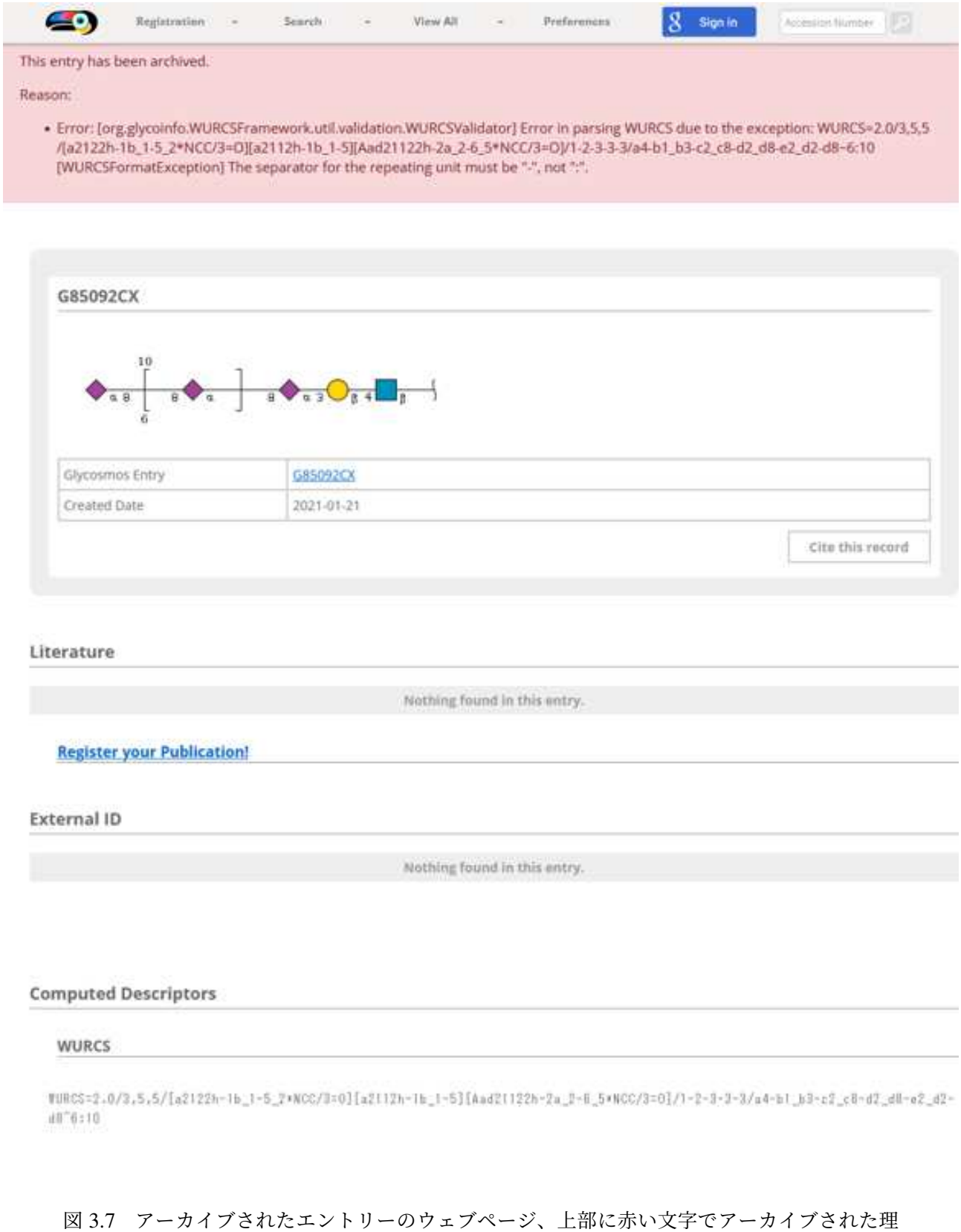


図 3.7　アーカイブされたエントリーのウェブページ、上部に赤い文字でアーカイブされた理由が明記される。この場合は WURCS の糖鎖構造配列に誤りが見つかったため。

第 4 章

発展的な糖鎖構造比較

検索には大きく分けて、一致検索と類似検索がある。一致検索やその派生である部分一致検索は単に文字列の一致あるいは、構造の一致を見つけることで達成できるが、類似検索は異なる。我々は効果的な類似検索を可能にするため単糖置換行列の作成を目指した。

4.1 方法

4.1.1 単糖データ

我々ははじめに SNFG で定義されている単糖のリストを参考に PubChem [5] より LINUCS 形式 [12] で単糖構造データを取得した。これらのデータは、SNFG によって定義された単糖構造のアノマー情報を指定していない。したがって、これらのデータにアノマー情報を追加して、各単糖のアノマー構造のアルファバージョンとベータバージョンをを生成し、完全な単糖情報の最終リストを作成した。次に、これらの LINUCS 文字列に基づいて、Glycosciences.de [35] から最終的に合計 118 個の単糖の原子レベルの 3 次元構造を取得した。Glycosciences.de から取得した PDB データはイレギュラーな情報が混じっていたため、処理できるように書き直すスクリプトを作成した。具体的には、取得した PDB ファイルの原子表記から原子を区別するための位置番号を削除し、簡略化した。さらに、原子「XX」が PDB ファイルに含まれている場合は、簡単にするために削除した。さらに、各構造の椅子の形状を調べて、単糖の立体配座異性体を示す 1C4 または 4C1 であるかどうかを確認した。単糖のイス型の種類を併記したリストは次の URL から参照できる。

https://gitlab.com/akihirof0005/TouCom/-/raw/master/which_ring/data

4.1.2 ソフトウェア

単糖間距離を計算するために、以下に説明するいくつかの機能を実装する次のプログラムと手順を開発した。開発したソフトウェアのソースコードは次の二つの URL で管理されている。

https://gitlab.com/akihirof0005/TouCom-sub

https://gitlab.com/akihirof0005/TouCom

- **Kabsch method**

  原子構造の比較のために、Kabsch 法 [36] による分子重ね合わせを行うライブラリを開発した。このライブラリを用いることによって同じ原子構造の 2 つの化合物間で最小の RMSD（二乗平均平方根偏差）を取得できるようになった。RMSD は式 4.1 に示すように比較する立体構造間の原子間距離を丸めて得られる数値であり、この値を二つの立体構造間距離として用いることができる。この式において、a_i と b_i はそれぞれ分子 A の原子と分子 B の原子を指し、n は入力構造の原子数を指す。

$$RMSD(A,B) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(a_i - b_i)^2} \tag{4.1}$$

- **階層的クラスタリング**

  クラスター間距離がグループ平均法によって決定される RMSD 距離を用いた階層的クラスタリングを実行する Java プログラムを開発した。

- **共通立体構造検索**

  また、異なる分子構造の立体構造である m_1 と m_2 の間で共通の立体構造を見つけ、RMSD を計算するソフトウェアを開発した。まず、水素原子を除いて、小さい方の 3D 構造の原子数を取得し、この数を n と呼ぶ。次に、より大きな 3D 構造から n の非水素原子をランダムに選択し、選択した原子のすべてのペア間 $n \times n$ で RMSD $V_{rmsd(m_1、m_2)}$ を計算する。$V_{rmsd(m_1,m_2)}$ は以下の式 4.2 で定義されている $F(n)$ と比較し、$F(n)$ より値が小さい場合に二つの立体構造間の最大共通構造が得られる。但し、$V_{rmsd(m_1,m_2)}$ が $F(n)$ より大きい

場合には $n = n - 1$ を設定し、再帰的に立体共通部分構造が見つかるまで繰り返す。式 4.2 は予備分析から得られた RMSD 値の分布に基づいて導き出された。

$$V_{rmsd(m_1,m_2)} < F(n) = 0.1 \times \log 10((n) \times (n-1) \times (n-2) \times 1.2 \div 6) + 1 \tag{4.2}$$

- **立体構造間距離の計算**

立体構造間距離を計算するために次の式 4.3 は RMSD $V_{rmsd(m_1,m_2)}$ と共通部分の最大原子数 n を用いて定義しました。m はデータセット全体で全ての n の最大数で決定される。定義した距離である spatialDistance は、一般的に重なり合う原子の数が多いほど距離が近くなるといえる定義した。

$$spatialDistance(m_1, m_2, n, m) = V_{rmsd(m_1,m_2)} + \sum_{i=n+1}^{m} F(i) \tag{4.3}$$

- **単糖特徴量を基にした距離の計算**

単糖の物理化学的性質を計算するために関数 4.4 を実装した。分子 m から官能基の数を数える関数である。これを用いて式 4.5 を定義した。

$$ct_m(functionalGroupName) \tag{4.4}$$

$$\begin{aligned} &functionalDistance(m_1, m_2) = \\ &\quad | \{count_{m_1}(OH) + count_{m_1}(COOH) + count_{m_1}(NH2) + count_{m_1}(NOC)\} - \\ &\quad\quad \{count_{m_2}(OH) + count_{m_2}(COOH) + count_{m_2}(NH2) + count_{m_2}(NOC)\} | + \\ &\quad | count_{m_1}(COOH) - count_{m_2}(COOH) | + \\ &\quad | count_{m_1}(NH2) - count_{m_2}(NH2) | + \\ &\quad | count_{m_1}(NOC) - count_{m_2}(NOC) | \end{aligned} \tag{4.5}$$

- **単糖間距離と類似度の計算**

spatialDistance と functionalDistance から式 4.6 の単糖間距離を定義した。また、距離から式 4.7 の類似度を定義した。

$$Distance(m_1, m_2) = spatialDistance + 2 \times functionalDistance(m_1, m_2) \tag{4.6}$$

$$similarity(m_1, m_2) = 1/(1 + Distance(m_1, m_2)) \tag{4.7}$$

4.1.3 ハードウェア

16G メモリと SSD ストレージ、Ryzen2600 の CPU を備えた一般消費者向けのパソコンで分析を行った。Java8 の実行環境を用いて速度を上げるために RAM ディスクに全てのファイルを設置し実行した。

4.1.4 手順

開発したソフトウェアを組み合わせて、以下の手順を構築した。この方法に従って、単糖の距離行列を計算する。

1. **各単糖に対して立体構造を予測する**

   単糖のデータセットに対して、Shape [37] を実行する。Shape は糖鎖向けの自動立体構造予測ソフトウェアで、mm3 力場をもちいて立体構造を評価し、遺伝的アルゴリズムで立体構造を検索する。Shape は単糖の立体構造データを基に素早く多数の取りうる立体構造を予測することができる。

2. **単糖の立体構造の数を絞り込む**

   各単糖に対しての多数の立体構造の中には非常によく似た構造が含まれているので、階層的クラスタリングを行い、立体構造数を少なく調整した。この手順は類似した立体構造をクラスターに分類し、代表を取り出すことで、類似した立体構造を取り除くために行う。図 4.1 に示すように、クラスターの数は、クラスターの直径とクラスター間の距離のバランスからクラスターの独立性を判断することによって動的に決定した。

3. **単糖置換行列の生成**

   階層的クラスタリングから選択したコンフォメーションを使用して、ソフトウェアのセクションで説明したソフトウェアを使用して、立体構造のすべてのペアの RMSD 値を取得した。この RMSD 距離行列は、式 4.3 を使用して距離値に変換され、次に各単糖ペアの平均

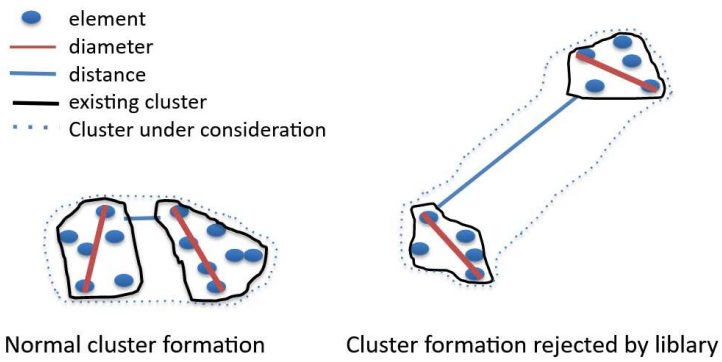


図 4.1 不自然に離れたクラスターが統合されないようにクラスターの独立性を判断した

距離値を取得することにより、$118 \times (118-1)/2 = 6903$ ペアの spatialDistance に変換した。また、式 4.5 を用いて置換基ベースの functionalDistance を計算し、式 4.4、式 4.5 を用いて、最終的に各単糖間の類似度を計算し、単糖置換行列としてまとめた。

4.1.5 単糖置換行列を用いた検索

既存の糖鎖構造解析ソフトに単糖置換行列を応用することで実際に単糖置換行列を用いることができる。いままでは単糖比較を一致または不一致で判断していた部分を類似度として計算できるので、検索などでいままでよりも良い結果が期待できる。糖鎖構造間のペアワイズアライメントを実装している KCaM に単糖置換行列を応用し、検索に用いた。

KCaM への実装

KCaM では 2 糖構造におけるスコア行列を用いているが、単糖間の類似性は真理値によって評価されるので、これを単糖置換行列による値に置き換えることで実現する。式 4.8 は KCaM がアライメントスコアを算出する計算式である。新しい KCAM のソースコードは次の URL で管理されている。`https://gitlab.com/akihirof0005/TouCom-sub/-/tree/master/kcam`

$$
\begin{aligned}
w(u,v) = & max[0, \\
& \alpha\delta[label(u), label(v)] \\
& - \beta(1 - \delta[ulabel(p(u),u), ulabel(p(v),v)]) \\
& - \beta(1 - \delta[dlabel(p(u),u), dlabel(p(v),v)])\ ]
\end{aligned}
\tag{4.8}
$$

$label$ は単糖を示し、$\delta[label(u), label(v)]$ は単糖間の比較スコアを示す。$\delta[ulabel(p(u),u),$ $ulabel(p(v),v)]$ と $\delta[dlabel(p(u),u), dlabel(p(v),v)])$ についてはグリコシド結合の比較結果を示している。KCaM は O-グリコシド結合を前提としており、前者は結合に使われる不斉炭素の位置の比較、後者は還元末端側の結合に使われる炭素位置の比較スコアを示す。従来の KCaM では $\delta[label(u), label(v)]$ が単純な一致不一致からなるスコアで構成されており、この実装を単糖置換行列から抽出した類似度を用いるように変更した。また従来の KCaM が想定している糖鎖のデータ構造では単糖名にアノメリック情報が含まれていないので、KCaM スコアの算定式を式 4.9 に修正し、KCaM のプログラムを変更した。これによって作成した単糖置換行列を用いた KCaM によるペアワイズアライメントを行えるようになった。

$$
\begin{aligned}
w(u,v) = & max[0, \\
& \alpha\delta[label(u), label(v)] \\
& - \beta(1 - \delta[dlabel(p(u),u), dlabel(p(v),v)])\ ]
\end{aligned}
\tag{4.9}
$$

4.2 結果

作成した単糖置換行列は次の URL で閲覧できる。`https://gitlab.com/akihirof0005/TouCom-sub/-/raw/master/kcam/similarity.txt`

4.2.1 データ

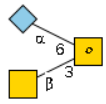
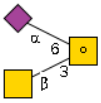
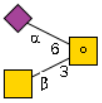
GlyCosmos よりヒトの糖鎖構造を IUPAC 形式で取得した。これを KCF 形式へ変換し、検索するデータに用いた。なお、もともと KEGG が管理していたデータにおいて Sia として表記されていた糖鎖構造のうち Neu5Ac と Neu5Gc に変換され登録されているが、Neu5Gc がヒトに含まれていない為、これを除外したものを用いた。用いたデータは次の URL から参照できる。

https://gitlab.com/akihirof0005/TouCom-sub/-/raw/master/kcam/all

4.2.2 G05768VS の検索について

表 4.1 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 4.1: 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を示す。KCaM 1.0 は従来の KCaM、KCaM 1.1 は改良した KCaM を示している。

QUERY			G05768VS		
					
KCaM 1.0			KCaM 1.1		
RANK	GlyTouCan ID	SCORE	RANK	GlyTouCan ID	SCORE
1	G75947FL 	88.3	1	G75947FL 	87
	G05235KE			G54953IZ	

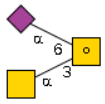
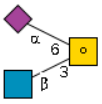
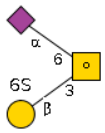
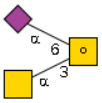
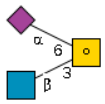
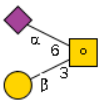
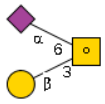
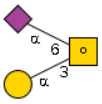
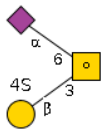
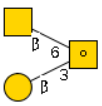
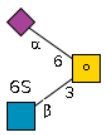
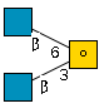
	
2 85	2 70.7
G29386CS	G05235KE
	
3 76.7	3 69.7
G54953IZ	G65285QO
	
3 76.7	4 69.3
G65285QO	G65790DQ
	
3 76.7	4 69.3
G65447BW	G45129XM
	
3 76.7	5 67
G93802CF	G61833XV
	
3 76.7	6 65.3

表 4.1 は表 4.1 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 $\alpha-$Neu5Gc は非常に高い類似度を示す $\alpha-$Neu5Ac に置換される。

表 4.2　表 4.1 示した結果の中で使用された単糖置換行列の値である。

Monosaccharide pair		Similarity
α-Neu5Gc	α-Neu5Ac	0.73056
β-GalNAc	β-GlcNAc	0.58792
β-GalNAc	α-GalNAc	0.55319
β-GalNAc	α-Gal	0.53561
β-GalNAc	β-Gal	0.53560

表 4.1 の二位の結果が特徴的であり、KCaM 1.0 では、$\beta-$GalNAc が $\alpha-$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 $\beta-$GalNAc が $\beta-$GlcNAc に置換された G54953IZ が二位に順位づけされた。これ

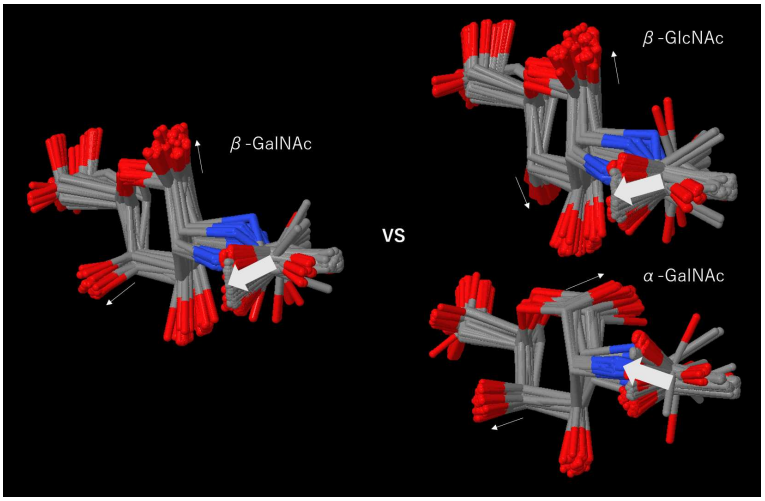


図 4.2　β-GlcNAc と α-GalNAc, β-GalNAc と β-GlcNAc　の 3 次元構造におけるコンフォメーションの傾向の違いを図示した図表 4.2 の数値を支持するようにみえる.

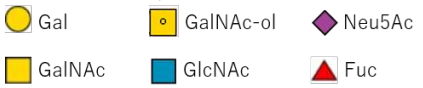
は表 4.2 に示すように $\beta-$GalNAc と $\beta-$GlcNAc の類似度が高いからである。表 4.2 で示した単糖対のうち、$\beta-$GalNAc と $\alpha-$GalNAc の違いは 1 位の炭素に結合する水酸基の向きが異なるが、$\beta-$GalNAc と $\beta-$GlcNAc の違いは 4 位の炭素に結合する水酸基の向きが異なるためであるが、その微妙な立体構造のスコアが算出されているのがわかる。図 4.2 に図示するように微妙に構造変化が起きており、これらの全体的な立体構造的特徴の変化が表 4.2 に反映していると考えられる。

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。

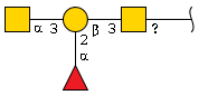
4.2.3 G71832QJ の検索について

G71832QJ は ABO 式血液型において A 型にみられるものとよく似た構造をしている糖鎖である。表 4.3 にはスコアが 75 以上の結果を KCaM 1.0、KCaM 1.1 それぞれで示す。

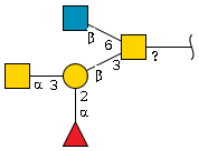
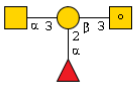
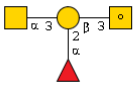
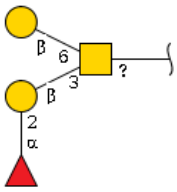
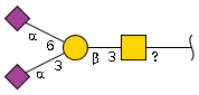
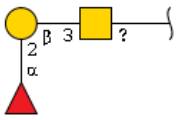
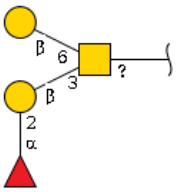
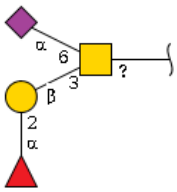
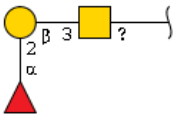
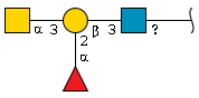
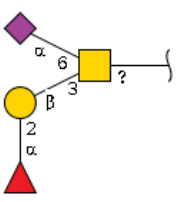
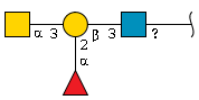
表 4.3: ヒトの糖鎖構造データを対象に G71832QJ をアライメントし、アライメントスコアで並べ替えた結果を示す。

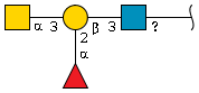
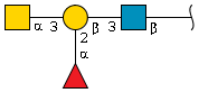
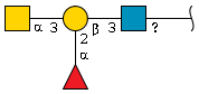
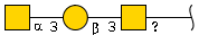
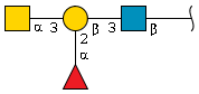
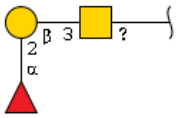
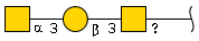
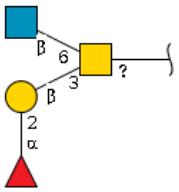
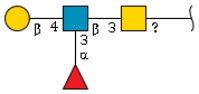
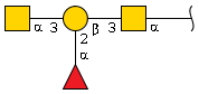
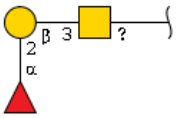


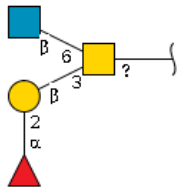
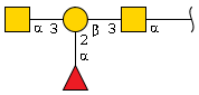
QUERY	G71832QJ



KCaM 1.0			KCaM 1.1		
RANK	GlyTouCan ID	SCORE	RANK	GlyTouCan ID	SCORE
1	G71832QJ	100	1	G71832QJ	100
2	G81466NG	81.3	2	G32588AA	80
	G32588AA			G02020YO	

3  80	3  75
G02020YO 4  75	G11876US 3  75
G03433NI 4  75	G46868TY 3  75
G11876US 4  75	G47391NN 3  75
G46868TY 4  75	G51828AY 3  75
G47391NN 4  75	G58507AZ 3  75

G51828AY 4  75	G66163TI 3  75
G58507AZ 4  75	G72914LC 3  75
G66163TI 4  75	G96057TO 3  75
G72914LC 4  75	G96463EY 3  75
G78108ZM 4  75	G99315PE 3  75
G96057TO 4  75	
G96463EY	

	4 75	
G99315PE 	4 75	

従来の KCaM ではトップの類似構造に G81466NG があげられる。この構造は Fuc が Neu5Ac に置換され、更に非還元末端側の GalNAc(a1-3)Gal のグリコシド結合部分が GalNAc(b1-4)Gal に置換されたものがアライメントされる。

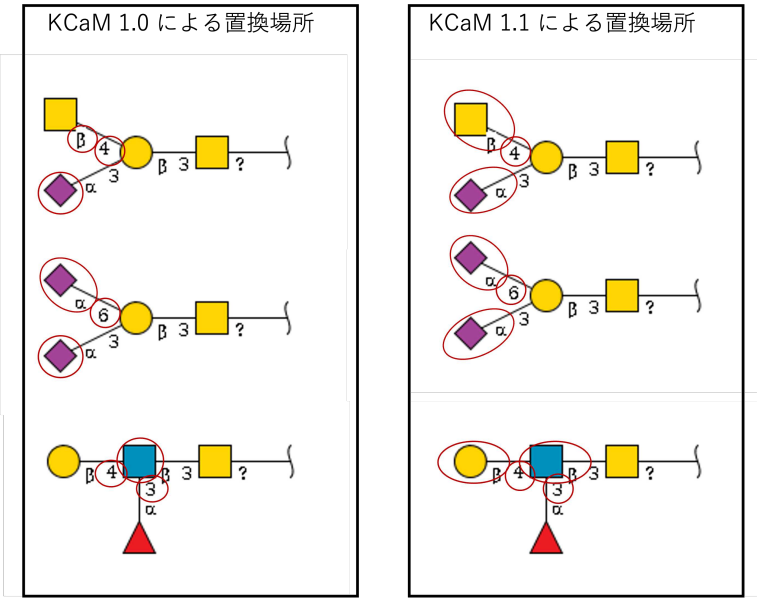


図 4.3　上から順に G81466NG、G03433NI、G78108ZM の構造がクエリである G71832QJ と置換している場所を赤で囲った図、KCaM 1.0 では少ない情報の置換で済んでいるが、KCaM 1.1 では大きな情報の置換が行われる。

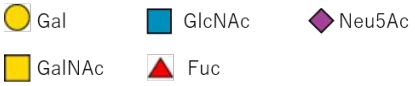
しかし、単糖置換行列を用いた KCaM 1.1 では $\alpha-$Fuc と $\alpha-$Neu5Ac の置換に加え、$\alpha-$GalNAc と $\beta-$GalNAc の置換、加えてグリコシド結合位置の置換の計 3 つの置換が入り、単糖の置換スコ

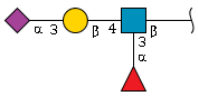
アは類似度を基に計算される。よってアライメントスコアが 80 以下の値となり、上位の結果では確認できない。代わりに β−GlcNAc が新たに結合した G32588AA が 81.3 のスコアになり 2 位にランキングされた。3 位にはスコアが 75 以下になる構造が並ぶが、KCaM 1.0 ではランクしていた G81466NG、G03433NI、G78108ZM がスコア 75 以下になって表では確認ができない。これらの事情を図 4.3 にまとめた。KCaM 1.1 になったことでアルゴリズムが変わり、アノマー情報のみの置換ができなくなり、大きな置換が必要となる事でスコアが下がっている。

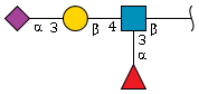
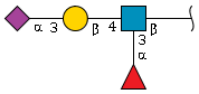
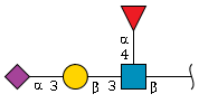
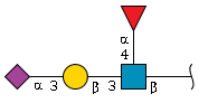
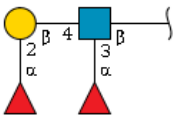
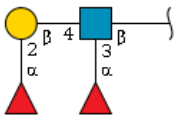
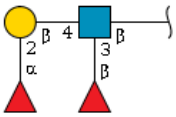
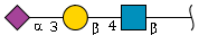
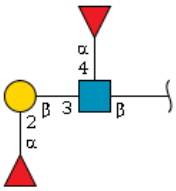
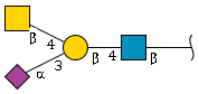
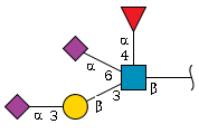
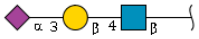
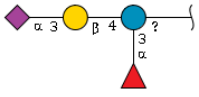
4.2.4 G00054MO の検索について

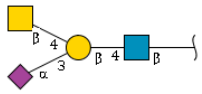
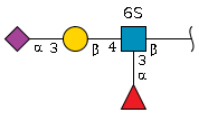
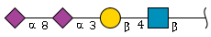
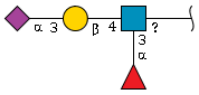
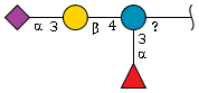
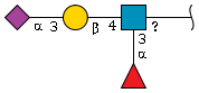
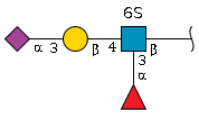
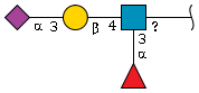
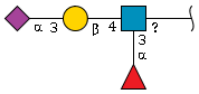
G00054MO は Sialyl Lewis x 構造として知られている構造である。表 4.4 にはスコアが 75 以上の結果を KCaM 1.0、KCaM 1.1 それぞれで示す。KCaM 1.1 の相違点は KCaM 1.0 の結果のうちから 4 位～6 位の結果がスコア 75 以下となって表にランキングされていない点である。それぞれ G58824S1、G00048MO、G93400AK がランク外になったのは G71832QJ の検索による結果と同様の事情である。

表 4.4: ヒトの糖鎖構造データを対象に G00054MO 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を示す。



QUERY			G00054MO		
					
KCaM 1.0			KCaM 1.1		
RANK	GlyTouCan ID	SCORE	RANK	GlyTouCan ID	SCORE
	G00054MO			G00054MO	

1		100	1		100
	G00053MO			G00053MO	
2		95	2		85
	G00052MO			G00052MO	
3		86.3	3		82.3
	G58824SI			G00065MO	
4		83.8	4		75
	G00048MO			G41052GC	
5		81.3	4		75
	G93400AK			G53957WR	
6		76	4		75
	G00065MO			G76568UL	
7		75	4		75

G41052GC 7  75	G80722US 4  75
G53957WR 7  75	G87205VF 4  75
G76568UL 7  75	G90387AM 4  75
G80722US 7  75	
G87205VF 7  75	
G90387AM 7  75	

4.2.5 まとめ

開発した単糖置換行列を用いて KCaM による応用例を示した。得られた新規の単糖置換行列は検索において次の効果をもたらした。つまり、G05768VS を検索した例では立体構造的により関連した順番に検索することを示し、G71832QJ と G00054MO を検索した例では立体的に関連性の低い糖鎖構造を類似構造の候補から外すことに成功した。

KCaM による検索結果はより構造的に似ているものが並ぶようになり、KCaM はより効率的に類似の糖鎖構造を検索することができ、今までより少ない数の糖鎖構造に絞り込むことができる。

4.3 本章の考察

結果からは、今までの KCaM が糖鎖構造を単なる木構造のデータ構造として扱っていたものを、単糖特異的なデータである単糖置換行列を用いて、KCaM が糖鎖構造データをより糖鎖らしく扱えることを示したといってよい。

現在は SNFG にリストされて単糖のみを扱っているが、対応する単糖を増やしていけば他生物種間において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を行うことができる。また、マルチプルアライメントを行う MCAW [38] 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。また、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。

また、KCaM に応用する以外にも、確率モデルを応用した解析ツールである、ProfilePSTMM [39] や OTMM[40] などのモデルについても応用が期待できる。これらのツールも単糖比較を一致するか不一致なのかをスコアリングするだけなので機能向上を期待できる。今後、単糖置換行列を組み込んだ糖鎖構造解析ツールは化学構造の類似性を考慮した糖鎖分析結果を得ることができる。先行研究であるスコア行列から発展して、いかなる糖鎖データベースも前提としない置換行列なので糖鎖解析において普遍的に用いることができる。

一方で、糖鎖を木構造として扱う以上は、糖鎖らしさを付加するには単糖置換行列の他にグリコシド結合を比較する際の重みづけデータが不足していることがあげられる。これを補完することでより糖鎖らしい類似性を定義することができ、検索機能の向上も見込むことができる。

第 5 章

全体の考察

5.1　まとめ

日々増え続ける糖鎖構造に対して、持続開発可能な GlyTouCan のシステムを作ることができた。GlyTouCan では、ユーザーが提出した糖鎖構造情報データをそのまま保存することになり、データの修正やシステムを拡張しやすくなった。また、可能な限り不具合を修正し、今後エントリーの整理が必要になった場合に直ちにアーカイブ出来るよう、アーカイブの仕組みを作成した。

糖鎖構造への理解を深めるため単糖置換行列の開発をおこない、糖鎖構造比較におけるその有用性を示すことができた。単糖置換行列は今後より発展していく必要があるが、ソフトウェアと比較する際の重みづけをファイルで分離することにより、柔軟にアップデートしていくことができる。

5.2　今後の課題

今後も WURCS や WURCSFrameWork のアップデートや増加する糖鎖構造に従って GlyTouCan もアップデートしていく必要がある。

単糖置換行列は、より多くの単糖や修飾に対応しより多くのツールに組み込まれる事が望ましい。すべての既知の糖鎖郡に対して、糖鎖間の類似度をトリプルストアに保存し公開することで、sparql による絞り込みをしながら類似情報付きの糖鎖を得ることができる。これは日々増加する

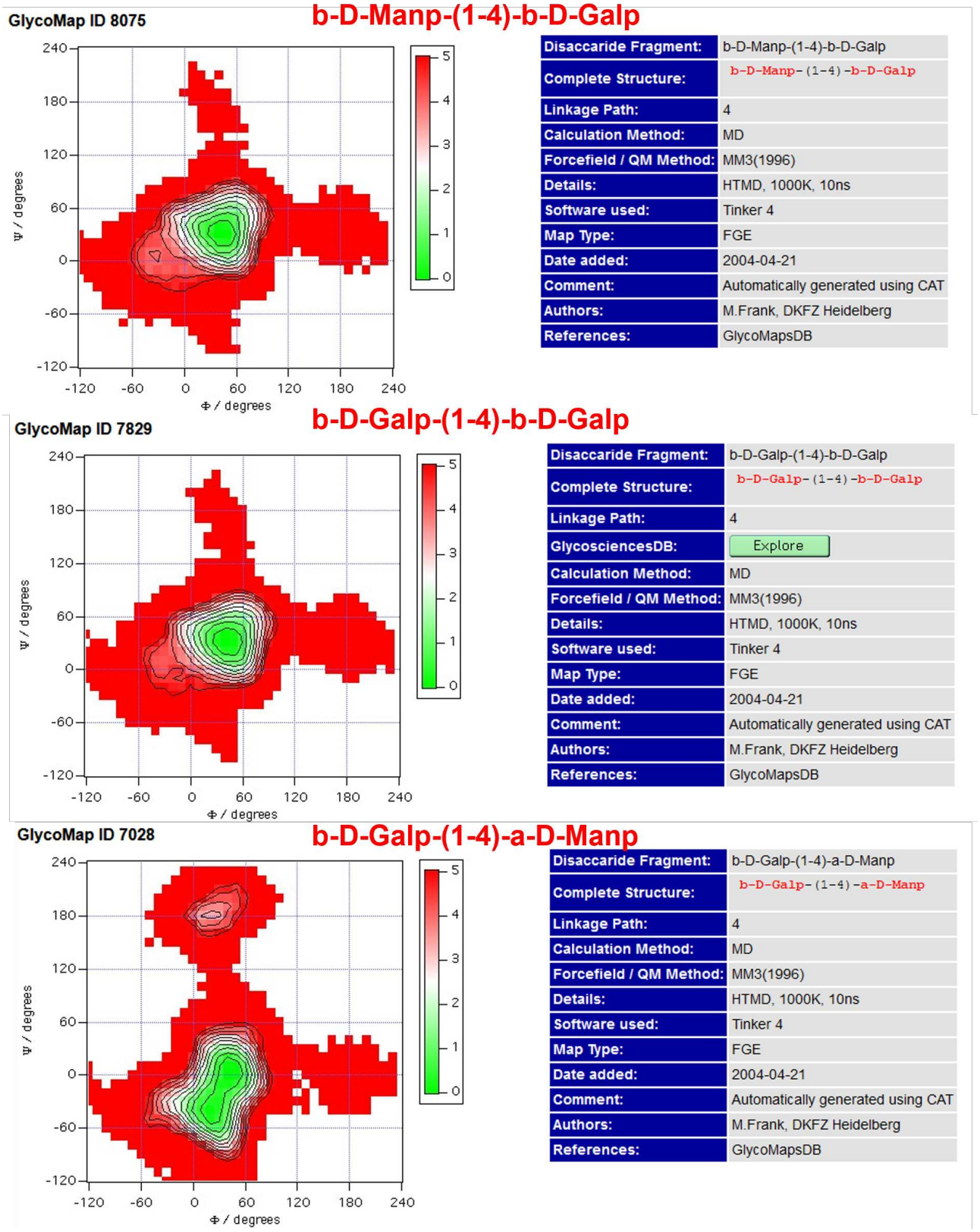


図 5.1　GlycoMapsDB より得た 3 つのラマチャンドランマップ、いずれもβ1-4 結合であるが、8075 のマップと 7829 のマップはよく似ているが、7028 のマップは他の二つとは異なり、より自由度の高い二面角を持っていることが分かる。

糖鎖構造に対して、リアルタイムな解析を可能にする。このために KCaM を始めとした各種糖鎖構造解析ツールが WURCS に対応することが理想である。また、グリコシド結合の比較は理想的にはすべてのラマチャンドランマップの比較した結果の行列を作成することが理想的には望ましい。なぜなら、今の結合様式を比較する方式は類似性を評価するには役不足であるからである。二単糖の単糖置換行列があれば解決するが、スコア行列のアプローチに近く、現実的ではないため二単糖をグリコシド結合の単位とするのは難しい。そのためより現実的なアプローチとして、グリコシド結合が形成する二面角をなす $C_{a+1} - C_a - O - C_b - C_{b+1}$ の構造を結合の最小単位とし、その立体構造のバリエーションでラマチャンドランマップを分類できるような方法論を構築し、グリコシド結合を立体的構造によって分類する。今の結合様式による理解ではなく、立体構造による理解に基づいてグリコシド結合の比較を行う。これらによって、糖鎖構造の配列による比較はより正確になる。

図 5.1 には GlycoMapsDB より得た 3 つのグリコシド結合のラマチャンドランマップを示す。これらはいずれも β1-4 結合である。これまでの比較ではこれらはすべて等しい結合と判断されるが、明らかに 7028 のマップは異なる傾向を持っていることが分かる。これらのマップの違いからグリコシド結合の距離や類似性を定義できる。実際のグリコシド結合のなす二面角にはいくつかの種類があり、すべての結合をこのように単純に論じることはできないが、以上のようなラマチャンドランマップを比較する方針でグリコシド結合の比較をより詳細に行うことができる。

5.3 展望

GlyTouCan の糖鎖構造が適切に管理され、セマンティックウェブとして公開され続ける事は糖鎖構造をとりまく解析や分析において今後ますます重要になる。今後、論文などで報告される新規の糖鎖配列に GlyTouCan で発行されるアクセッションナンバーが対応していくことで糖鎖配列が決定し登録された段階で、どのようなメタデータを持つ構造と相同性を持つのかがリアルタイムで理解できるシステムを目指すべきである。この度開発した単糖置換行列はいかなるデータベースからも独立しており、物理化学的性質に立脚した行列であるがゆえに GlyTouCan が管理するすべての糖鎖構造へ適応できる可能性を持っている。新たな単糖を増やし、単糖置換行列に追加す

ることは、構造が明確な単糖については容易であるが、修飾が重なる GAG のような構造のための単糖を追加することは、その立体構造の更なる検討が必要だ。また、糖鎖構造の比較においては単糖比較の重み付けのみならず、グリコシド結合に関しての比較に関しても見直し適切な重みをつける必要がある。これによってグリコシド結合間の類似性を数値化できるようになる。これによって糖鎖構造比較の精度を上げることができる。また、KCaM を代表した糖鎖構造解析ツールの解析結果を RDF 化することで GlyTouCan に登録されている糖鎖構造に対しての類似検索結果を素早く得ることができる。RDF はグラフ構造になっているのでそのグラフ構造を解析することで、類似性による糖鎖クラスターの発見にも期待できる。

また、本論文とは異なるアプローチとしてケモインフォマティクスの見地から糖鎖構造向けの比較のための構造記述子を定義することが考えられる。この場合は不斉炭素中心を考慮した構造記述子にする必要がある。不斉炭素中心を考慮した原子レベルの分子構造の記述子は存在するので、それらを糖鎖向けに整理し、ツールを開発する必要がある。

謝辞

辛抱強く指導にあたってくださった担当教員である主査の木下聖子先生、ありがとうございました。また、審査にあたってくださった池口雅道先生、西原祥子先生に感謝申し上げます。研究遂行にあたり、糖の構造情報について助言をくださった公益財団法人野口研究所の山田一昨先生に感謝申し上げます。

長い学生生活を支えてくださった学友の皆と両親に感謝申し上げます。

また、なによりも創価大学を見守ってくださった無名の創立者の皆さんと池田大作先生に感謝申し上げます。

参考文献

[1] Sriram Neelamegham, Kiyoko Aoki-Kinoshita, Evan Bolton, Martin Frank, Frederique Lisacek, Thomas Lütteke, Noel O' Boyle, Nicolle H Packer, Pamela Stanley, Philip Toukach, et al. Updates to the symbol nomenclature for glycans guidelines. *Glycobiology*, 29(9):620–624, 2019.

[2] Thomas Lütteke. Translation and Validation of Carbohydrate Residue Names with MonosaccharideDB Routines. In *A Practical Guide to Using Glycomics Databases*, pages 29–40. Springer, 2017.

[3] Arunima Singh, Matthew B Tessier, Kari Pederson, Xiaocong Wang, Andre P Venot, Geert-Jan Boons, James H Prestegard, and Robert J Woods. Extension and validation of the glycam force field parameters for modeling glycosaminoglycans. *Canadian journal of chemistry*, 94(11):927–935, 2016.

[4] DT Cremer and JA Pople. General definition of ring puckering coordinates. *Journal of the American Chemical Society*, 97(6):1354–1358, 1975.

[5] Sunghwan Kim, Paul A Thiessen, Evan E Bolton, Jie Chen, Gang Fu, Asta Gindulyte, Lianyi Han, Jane He, Siqian He, Benjamin A Shoemaker, et al. PubChem substance and compound databases. *Nucleic Acids Research*, 44(D1):D1202–D1213, 2016.

[6] Martin Frank, Thomas Lütteke, and C-W Von der Lieth. Glycomapsdb: a database of the accessible conformational space of glycosidic linkages. *Nucleic acids research*, 35(suppl_1):287–290, 2007.

[7] Kiyoko F Aoki, Atsuko Yamaguchi, Nobuhisa Ueda, Tatsuya Akutsu, Hiroshi Mamitsuka, Susumu Goto, and Minoru Kanehisa. KCaM (KEGG Carbohydrate Matcher): a software

tool for analyzing the structures of carbohydrate sugar chains. *Nucleic Acids Research*, 32(suppl_2):W267–W272, 2004.

[8] Kiyoko F Aoki, Hiroshi Mamitsuka, Tatsuya Akutsu, and Minoru Kanehisa. A score matrix to reveal the hidden links in glycans. *Bioinformatics*, 21(8):1457–1463, 2005.

[9] Steven Henikoff and Jorja G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[10] Kosuke Hashimoto, Susumu Goto, Shin Kawano, Kiyoko F Aoki-Kinoshita, Nobuhisa Ueda, Masami Hamajima, Toshisuke Kawasaki, and Minoru Kanehisa. KEGG as a glycome informatics resource. *Glycobiology*, 16(5):63R–70R, 2006.

[11] S Herget, R Ranzinger, K Maass, and C-WVD Lieth. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydrate research*, 343(12):2162–2171, 2008.

[12] Andreas Bohne-Lang, Elke Lang, Thomas Förster, and Claus-W von der Lieth. LINUCS: linear notation for unique description of carbohydrate sequences. *Carbohydrate Research*, 336(1):1–11, 2001.

[13] Kenichi Tanaka, Kiyoko F Aoki-Kinoshita, Masaaki Kotera, Hiromichi Sawaki, Shinichiro Tsuchiya, Noriaki Fujita, Toshihide Shikanai, Masaki Kato, Shin Kawano, Issaku Yamada, et al. WURCS: the Web3 unique representation of carbohydrate structures. *Journal of chemical information and modeling*, 54(6):1558–1566, 2014.

[14] Dennis A Benson, Mark Cavanaugh, Karen Clark, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. GenBank. *Nucleic acids research*, 41(D1):D36–D42, 2012.

[15] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, 1990.

[16] Mark Johnson, Irena Zaretskaya, Yan Raytselis, Yuri Merezhuk, Scott McGinnis, and Thomas L Madden. Ncbi blast: a better web interface. *Nucleic acids research*, 36(suppl_2):W5–W9, 2008.

[17] Protein data bank: the single global archive for 3d macromolecular structure data. *Nucleic acids research*, 47(D1):D520–D528, 2019.

[18] Akitsugu Suga, Masamichi Nagae, and Yoshiki Yamaguchi. Analysis of protein landscapes

around n-glycosylation sites from the pdb repository for understanding the structural basis of n-glycoprotein processing and maturation. *Glycobiology*, 28(10):774–785, 2018.

[19] Kiyoko F Aoki-Kinoshita, Hiromichi Sawaki, Hyun Joo An, Matthew Campbell, Qichen Cao, Richard Cummings, Daniel K Hsu, Masaki Kato, Toshisuke Kawasaki, Kay-Hooi Khoo, et al. The fifth acgg-db meeting report: towards an international glycan structure repository. *Glycobiology*, 23(12):1422–1424, 2013.

[20] Kiyoko Aoki-Kinoshita, Sanjay Agravat, Nobuyuki P Aoki, Sena Arpinar, Richard D Cummings, Akihiro Fujita, Noriaki Fujita, Gerald M Hart, Stuart M Haslam, Toshisuke Kawasaki, et al. GlyTouCan 1.0–The international glycan structure repository. *Nucleic Acids Research*, 44(D1):D1237–D1242, 2016.

[21] Michael Tiemeyer, Kazuhiro Aoki, James Paulson, Richard D Cummings, William S York, Niclas G Karlsson, Frederique Lisacek, Nicolle H Packer, Matthew P Campbell, Nobuyuki P Aoki, et al. Glytoucan: an accessible glycan structure repository. *Glycobiology*, 27(10):915–919, 2017.

[22] Tim Berners-Lee and James Hendler. Publishing on the semantic web. *Nature*, 410(6832):1023–1024, 2001.

[23] Eric J Miller. An introduction to the resource description framework. *Journal of library administration*, 34(3-4):245–255, 2001.

[24] Deborah L McGuinness, Frank Van Harmelen, et al. Owl web ontology language overview. *W3C recommendation*, 10(10):2004, 2004.

[25] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. Semantics of sparql, 2008.

[26] Jorge Pérez, Marcelo Arenas, and Claudio Gutierrez. Semantics and complexity of sparql. *ACM Transactions on Database Systems (TODS)*, 34(3):1–45, 2009.

[27] Rene Ranzinger, Kiyoko F Aoki-Kinoshita, Matthew P Campbell, Shin Kawano, Thomas Lütteke, Shujiro Okuda, Daisuke Shinmachi, Toshihide Shikanai, Hiromichi Sawaki, Philip Toukach, et al. Glycordf: an ontology to standardize glycomics data in rdf. *Bioinformatics*, 31(6):919–925, 2015.

[28] GlycoRDF/Glycan. https://github.com/glycoinfo/GlycoRDF/blob/master/ontology/glycan.owl.

[29] WURCSFramework. https://gitlab.com/glycoinfo/wurcsframework.

[30] GlyTouCan Endpoint. https://ts.glytoucan.org/sparql.

[31] Issaku Yamada, Masaaki Shiota, Daisuke Shinmachi, Tamiko Ono, Shinichiro Tsuchiya, Masae Hosoda, Akihiro Fujita, Nobuyuki P Aoki, Yu Watanabe, Noriaki Fujita, et al. The glycosmos portal: a unified and comprehensive web resource for the glycosciences. *Nature Methods*, 17(7):649–650, 2020.

[32] Yu Watanabe, Kiyoko F Aoki-Kinoshita, Yasushi Ishihama, and Shujiro Okuda. GlycoPOST realizes FAIR principles for glycomics mass spectrometry data. *Nucleic Acids Research*, 49(D1):D1523–D1528, 2021.

[33] MA Rojas-Macias, J Mariethoz, P Andersson, C Jin, V Venkatakrishnan, NP Aoki, D Shinmachi, C Ashwood, K Madunic, T Zhang, et al. Towards a standardized bioinformatics infrastructure for n-and o-glycomics. nat commun 10: 3275, 2019.

[34] Shinichiro Tsuchiya, Issaku Yamada, and Kiyoko F Aoki-Kinoshita. GlycanFormatConverter: a conversion tool for translating the complexities of glycans. *Bioinformatics*, 35(14):2434–2440, 2019.

[35] Kiyoko F Aoki-Kinoshita. *A Practical Guide to Using Glycomics Databases*. Springer, 2017.

[36] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976.

[37] Jimmy Rosen, Laurence Miguet, and Serge Pérez. Shape: automatic conformation prediction of carbohydrates using a genetic algorithm. *Journal of Cheminformatics*, 1(1):16, 2009.

[38] Masae Hosoda, Yukie Akune, and Kiyoko F Aoki-Kinoshita. Development and application of an algorithm to compute weighted multiple glycan alignments. *Bioinformatics*, 33(9):1317–1323, 2017.

[39] Kiyoko F Aoki-Kinoshita, Nobuhisa Ueda, Hiroshi Mamitsuka, and Minoru Kanehisa. Pro-

filePSTMM: capturing tree-structure motifs in carbohydrate sugar chains. *Bioinformatics*, 22(14):e25–e34, 2006.

[40] Kosuke Hashimoto, Kiyoko Flora Aoki-Kinoshita, Nobuhisa Ueda, Minoru Kanehisa, and Hiroshi Mamitsuka. A new efficient probabilistic model for mining labeled ordered trees applied to glycobiology. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2(1):1–30, 2008.

付録

簡便なアクセスのため、参照した URL のリンクを集めたページを用意した。

https://research-assets.gitlab.io/thesis-defense/sourcecode/

1.4 データ

1.4.1 用いた単糖のリスト

https://gitlab.com/akihirof0005/TouCom/-/raw/master/which_ring/data

1.4.2 用いた単糖データ

https://gitlab.com/akihirof0005/TouCom/-/tree/master/data/pdb

1.4.3 作成した単糖置換行列

https://gitlab.com/akihirof0005/TouCom-sub/-/raw/master/kcam/similarity.txt

1.4.4 検索に用いたヒト糖鎖構造データ

https://gitlab.com/akihirof0005/TouCom-sub/-/raw/master/kcam/all

1.5 単糖置換行列の作成に必要なソースコード

https://gitlab.com/akihirof0005/TouCom-sub

https://gitlab.com/akihirof0005/TouCom