Vol.2009-BIO-18 No.6
2009/9/17

糖鎖認識部位発見のための部分構造制約付きクラスタリング

寺井　はるな[†1]　瀬々 潤[†1]

生命活動に大きな影響を及ぼす糖鎖は，微妙な構造の違いがタンパク質との結合親和性に大きく影響するため，結合に関わる構造を知ることが重要である．本研究では377 種類の糖鎖構造について結合親和性を同時に計測できるグリカンアレイのデータを複数同時に解析し，複数の実験に共通して結合に関わる糖鎖部分構造の組み合わせを発見するため，糖鎖構造制約条件付きクラスタリングを開発した．本手法では結合部位である末端構造だけでなく，結合親和性を高める第二の認識部位も特定できるような解析を行い Galectin-1，Influenza B 型ウイルスが認識する既知の糖鎖構造が予測できることを確認した．

Constrained clustering for discovering high binding affinity glycan substructures

Haruna Terai [†1] and Jun Sese [†1]

Carbohydrate chains exerts a big influence on the vital activity. Because the slight difference of a structure greatly influences the binding affinity with the protein, it is important to know the structures related to the binding. In this reasearch, we analyzed plural data of glycan arrays which is able to measure the binding affinity of 377 kinds of carbohydrate structures at the same time. To discover the combination of the carbohydrate chain substructures related to the binding, we developed a clustering method constrained by the carbohydrate chain structures. In this method you can discover not only the primary binding site but the secondary sites which increase the binding affinity. We confirmed that our method could estimate the already-known carbohydrate structure which binds with Galectin-1 and Influenza B virus.

1. はじめに

DNA やタンパク質解析の進歩により遺伝子に関する情報が増えてきたが，ゲノムに書かれていないが生命活動に大きな影響を及ぼすものとして糖鎖が知られている．糖鎖は各種の単糖がグリコシド結合によってつながり合った一群の化合物を指す．DNA やタンパク質の素材である核酸塩基やアミノ酸は一列に並ぶしかないが，糖鎖は多数のヒドロキシ基が全て結合に活用できるので，枝分かれを有する複雑な構造を成し，構造異性体もあるため，構成する分子数が大きくなるに従って指数級数的な数の構造を取ることが可能となる．更に，糖鎖は微妙な構造の違いによりタンパク質やウィルスとの結合親和性に大きく影響するため，結合時に認識される構造を知ることが重要である．

図 1 は糖鎖の略図で，Gal, Neu5Ac, Glc などは各種の単糖，線は結合，線と単糖の間に書かれている a, b はそれぞれ α，β の構造異性体，数字は何番の炭素と結合しているかを表している．**図 2** は図 1 の糖鎖の構造を文字列で表した例であり，枝分かれ構造を示す際に括弧が使われる．この図 1，図 2 の右側は根で細胞表面に付着しており，左側が葉で通常タンパク質は葉に結合する．

本研究では，結合時に認識する糖鎖の部分構造を明確化できる予測方法を提案する．特に，近年グリカンアレイにより 300 を超える糖鎖構造の結合親和異性データを同時に採取する事が可能となっており，このデータを利用して糖鎖構造を部分構造に分解し，それらの組み合わせを考えることで，糖鎖構造上隣接しない部位であっても，結合に関連する部分構造を特定する．

本研究ではある特定条件下 (あるタンパク質を与えた場合) でタンパク質と高い結合親和性を示す糖鎖の部分構造を決定することが目的である．タンパク質は糖鎖の末端部分に結合することが知られているが，タンパク質が糖鎖に結合するときに認識する部位は結合する糖鎖末端だけでなく，結合親和性を高める部位が別にあると考えられる[1)]．このような部位を第二の認識部位と呼ばれ，本研究では結合部位である末端構造だけでなく結合親和性を高めるような第二の認識部位も特定できるような解析手法を開発する．

さらに本研究で利用したグリカンアレイのデータベースには一種類のタンパク質を複数の濃度で実験しているデータがあり，その一連の実験の流れを一度にまとめて解析した．本来

†1 お茶の水女子大学大学院　情報科学コース
Dept. of Computer Science, Ochanomizu University

Vol.2009-BIO-18 No.6
2009/9/17

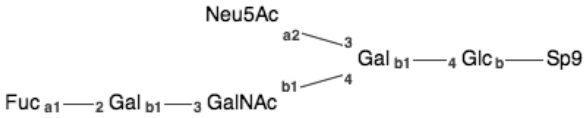


図 1 糖鎖の構造
Fig. 1 Glycan structure

Fuca1-2Galb1-3GalNAcb1-4(Neu5Aca2-3)Galb1-4Glcb-Sp9

図 2 糖鎖の構造（文字列の場合）
Fig. 2 Glycan structure (string)

SVM[2] など一般的な機械学習法ではクラスの値は一つだが，糖鎖構造をクラスタ分割の制約条件としたクラスタリングを行うことで，一連の実験に共通して結合親和性を高める部分構造を複数特定する．これにより一つの実験だけでは判断できなかった結果も複数の結果を用いることで，より精度の高い結果が得られる．

2. 関 連 研 究

化学構造と生物学的活性の間に成り立つ量的関係を線形計画法を用いて予測する研究[3] が行われている．津田らの研究では，回帰分類をおこなっているが，本研究で分類する対象は多次元ベクトルと対応しており，回帰分類は利用できない．また，本研究は木構造を有する糖鎖構造に絞り，分子量が大きい構造でも扱う点が異なる．

特定の環境下で反応した糖鎖構造を採取し，その糖鎖間で共通構造を見つける糖鎖構造のアラインメント[4]，あるいは，クラス分類問題として定式化しての共通構造の発見が行われている[5]．これらの研究で見つかった保存部位と，本研究で扱うグリカンアレイにおける結合親和度の関係を見つけることは，糖鎖認識部位を見つける一助となる．しかし，必ずしも保存されていなくても反応する構造がある場合，その構造を捕らえることができず，本手法に優位性がある．

3. 手　　法

本章では，グリカンアレイから得られるデータの構造を定義し，構造特異的な結合親和性を示す部分構造を抽出するための指標を定義する．

3.1 定　　義

本節では，糖鎖構造の定義及びグリカンアレイによって得られる結合親和度情報を定義する．以下，結合親和度を親和度と表記する．グリカンアレイには，糖鎖構造及びそれに対する実験に対する親和度が含まれている．

定義 1 (糖鎖構造とその部分構造群) 糖鎖 x はラベル有り根付き順序木 $T(x) = (V(x), E(x))$ として表される．ここで $V(x)$ 及び $E(x)$ は，それぞれ $T(x)$ の頂点集合及び辺集合を表す．頂点，辺ともにラベルがあり，頂点のラベルは単糖の種類，辺のラベルは構造異性体の種類及び結合している炭素番号を表す．$T(x)$ の根はタンパク質に結合している部位を示す．

木 S が木 T の部分木であるとき $S \subseteq T$ であると記す．糖鎖 x を表す木 $T(x)$ の全部分木の集合を $\mathcal{S}(x) = \{S \mid S \subseteq T(x)\}$ とする．

今回利用するグリカンアレイのデータにおいては，単糖の種類が 15 種，辺の種類が構造異性体 2 種と炭素番号の結合を組み合わせ合計 7 種存在している．図 1 の糖鎖から得られる全部分構造は 24 種ある．

グリカンアレイでは，各実験毎に全ての糖鎖がそれぞれ親和度を有する．

定義 2 X をグリカンアレイの全糖鎖集合，A を全実験集合とする．糖鎖 $x \in X$ に対し実験 $a \in A$ を行った親和度を $d(x, a)$ で表す．部分木 S に対し S を含む糖鎖集合 $\{x \mid S \subseteq T(x)\}$ を $X(S)$ と表す．

グリカンアレイにより，上記の情報が得られる．本研究ではどの部分構造が各実験の親和度に影響を与えているかを調べる．

3.2 制約条件付きクラスタ抽出

本節では，グリカンアレイ情報から結合親和性のある糖鎖構造を抽出するため，結合親和性の目安となる指標を定義する．糖鎖は構造特異的な結合が知られる一方，グリカンアレイは細胞外に構築された実験であるため，細胞内では見られない結合，あるいは，ノイズを含む可能性がある．これらの可能性に配慮し，適切な部分構造を抽出するため，3 つの条件を設定する．親和度が大きいこと，同一実験下における同一部分構造の親和度にバラツキが少ないこと，試薬の濃度を変化させた時，それに応じて親和度が変化することの 3 つである．それぞれの条件を設定する理由と，計測方法を以下に述べる．

まず，糖鎖の部分構造が実験において結合している場合に，その構造とグリカンアレイの親和度の間に成り立つ関係について考えてみよう．ある部分糖鎖構造 S が，実験 a において結合する場合，S を部分構造に持つ糖鎖集合 $X(S) = x \mid x \in X$ and $S \subseteq T(x)$ は $X(S)$

に含まれない糖鎖集合 $X - X(S)$ の親和度に比べ，大きい事が考えられる．グリカンアレイでは，実験 a が糖鎖 x に結合しない場合，0 に近い値を取っている (第 4 章参照) 事から，$X(S)$ 内の糖鎖の親和度が大きければ，結合していると考えられる．

この値は，次に示す着目する部分構造 S を有する糖鎖の親和度の平均を計算することにより評価できる．

定義 3 (親和度の平均) 着目する部分構造を S, S を持つ糖鎖集合を $X(S)$，全実験集合を A としたときの親和度の平均は，実験 $a \in A$ での親和度を $\mu(S,a)$ と定義すると，$\mu(S,a) = \frac{1}{|X(S)|}\sum_{x\in X(S)} d(x,a)$ であり，全実験集合に渡る平均親和度は，

$$\mu(S) = \frac{1}{|A|}\sum_{a\in A} d(x,a)$$

で表される．

次に，糖鎖の結合は非常に構造特異的であることが知られている．着目する部分構造の親和度が他の構造より大きかったとしても，その値にバラツキが大きい場合は，観測ノイズによる値の変動で偶然親和度が大きくなっている可能性が考えられる．このため，観測した親和度に，あまりバラツキが無い部分構造を採取したい．

値のバラツキ度合いは，分散により計測することが可能である．分散の小さな部分構造であれば，その値を取った偶然性が低い可能性が高くなる．複数の実験に対応させるため，次のように各実験における分散の平均に拡張した分散を定義する．

定義 4 着目する部分構造を S, S を持つ糖鎖集合を $X(S)$，全実験集合を A としたとき拡張した分散 $\sigma(S)$ を

$$\sigma(S) = \frac{1}{|A||X(S)|}\sum_{x\in X(S), a\in A}(d(x,a) - \mu(S,a))^2$$

最後に，糖鎖の結合は化学反応であり，結合する対象の濃度によって，飽和するまでは結合量が増加する事が予想される．よって，同一の試薬を複数の濃度で実験した場合に，濃度に応じて結合量が変化する部分構造であれば，より確実に結合していると考える事ができる．

結合の変化を捕らえるため，各実験における親和度の平均のばらつきを定義する．

定義 5 着目する部分構造を S, S を持つ糖鎖集合を $X(S)$，全実験集合を A としたとき変化量 $\delta(S)$ を，

$$\delta(S) = \frac{1}{|A|}\sum_{a\in A}(\mu(S,a) - \mu(S))^2$$

と定義する．

上記 3 つの条件を全て満たす構造を選ぶため，部分構造 S の良さを次の指標 $gindex(S)$ で計測する

定義 6 着目する部分構造を S, S を持つ糖鎖集合を $X(S)$，全実験集合を A としたとき，部分構造 S の結合親和度 $gindex(S)$ を

$$gindex(S) = \frac{\mu(S)\delta(S)}{\sigma(S)}$$

と定義する．この値を g-index と呼ぶ．

この指標において，全体平均 $\mu(S)$ 及び変化量 $\delta(S)$ が大きく，値の分散 $\sigma(S)$ が小さい構造が，より大きな値を有する事となる．

この値を利用する事で，全部分構造について指標を計算し，どの基質が特異的に結合しているかをランキングすることが可能となる．

3.2.1 部分構造の組み合わせへの拡張

上記の例では部分構造 S として木構造を考えた．しかし，グリカンアレイにおいてある糖鎖の部分構造 S を考えた場合，頂点ラベルと辺ラベルのバリエーションが多いため，S の大きさ（頂点数）が大きくなるに従って，S を部分構造に持つ糖鎖の数が急激に減少する．このため，基質特異性を示す部分構造を見つけられたとしても，その糖鎖の数が非常に少ないことが起こりうる．

本研究では，この問題点を部分構造の和集合を考える事で解決する．部分構造群 $\mathcal{S}$ に対し，糖鎖集合

$$X(\mathcal{S}) = \{x \in X \mid \exists S \in \mathcal{S} \text{ s.t. } S \subseteq T(x)\}$$

を定義する．この定義は，部分構造 S を持つ糖鎖群 $X(S)$ を $\mathcal{S}$ 内のいずれかの部分構造を含む糖鎖群に拡張したものである．定義 3, 4, 5 における各値の計算は容易に拡張する事が可能であり，g-index も同様に計算可能である．

複数の部分構造を利用して g-index を計算する際，$S, S' \subseteq T(x)$ に対して，$S \subseteq S'$ となる場合，$X(S) \supseteq X(S')$ が成立するので，$gindex(\{S\}) = gindex(\{S, S'\})$ が成り立つ．例えば，S = Neu5Acα2-8Neu5Acα2-，S' = Neu5Acα2-8Neu5Acα2-8Neu5Acβ- に対して $S \subseteq S'$ が成立するため，$gindex(\{S\}) = gindex(\{S, S'\})$ となる．このように包含関係にある場合，g-index を計算することなく，和集合を取ることで，計算の高速化を図った．

表 1 グリカンアレイデータ
Table 1 Glycan array data

糖鎖番号 Glycan No.	糖鎖名 Glycan name	結合親和度 Avg w/o Max & Min	...
1	Neu5Acα2-8Neu5Acα-Sp8	6883	...
2	Neu5Acα2-8Neu5Acβ-Sp17	20184	...
3	Neu5Acα2-8Neu5Acα2-8Neu5Acβ-Sp8	5720	...
4	Neu5Gcβ2-6Galβ1-4GlcNAc-Sp8	3978	...
⋮	⋮	⋮	
377	Neu5Acα2-3Galβ1-3(Neu5Acα2-6)GalNAc-Sp14	166	...

表 2 クラスの値に多次元ベクトルを持つ解析用データ
Table 2 Data with vector class values

サンプル: Glycan No.	属性: 部分構造 1 Neu5Acα2-	部分構造 2 -8Neu5Acα-	...	部分構造 x -8Neu5Acβ-	クラス: 結合親和度のベクトル (濃度 1, ..., 濃度 x)
1	巣	巣	...	-	(6883, ..., 3585)
2	巣	-	...	巣	(20184, ..., 7821)
3	巣	-	...	巣	(5720, ..., 8233)
4	-	-	...	-	(3978, ..., 2846)
⋮	⋮	⋮		⋮	⋮
377	巣	-	...	-	(166, ..., 2846)

4. 実行結果と考察

4.1 利用したデータ：グリカンアレイ

発現量解析で使用されるマイクロアレイのように，スライドガラス上に 300 個程度の糖鎖スポットを作成し，タンパク質やウィルスと反応させることで一度にスライド上の全糖鎖の結合親和性を測定することができるグリカンアレイの実験結果が Consortium for Functional Glycomics (CFG) の Functional Glycomics Gateway[6)7)] で公開されている．結合が強いほど値は高くなり，一部値がマイナスを表すものもある．
本研究では 6)7) にある Printed Array Ver. 3.0，3.1 のデータを使用した．各々320 種類，377 種類の糖鎖について実験が行われている．
データ例を**表 1** に示す．左から糖鎖番号，糖鎖名，結合親和度の平均を表している．

4.1.1 データの前処理

最も結合親和性が高くなるような糖鎖の部分構造を発見するために，各糖鎖構造を連結した部分木に分解し，各糖鎖がその部分構造を含むか否かを抽出する．結合親和性にはグリカンアレイデータの結合親和性の値をそのまま指定する．表 1 で示したようなグリカンアレイデータを複数枚一度に解析することを考えているので，複数サンプルでの結果を多次元ベクトルとして表す.（**表 2**）
また，異なる実験の結果を一つのベクトルで表すと，結合親和性の値が他の実験に比べて全体的に小さいときにその値は無視されてしまうので，各実験の値の偏りををなくすために，実験ごとの値を標準化したものをベクトルの値とする．実験はサンプルごとに行われているため，サンプルごとの値を標準化したものを値とする．

4.2 実 行 結 果

Galectin-1
糖結合タンパク質である Galectin-1 を 6) の Ver. 3.0 のグリカンアレイに濃度 5.62, 11.25, 22.5, 56.25, 112.5, 225μg/ml で与えた実験について解析した結果を示す.
上で示した指標を適用すると糖鎖構造が 3 つの場合が最も指標の値が大きくなることが分かり，その時の部分構造は，**表 3** のランキングの順番になる．表 3 の 1 位の構造は以下のいずれかを含んでいる糖鎖であることが分かった．-3GlcNAcβ1-3Galβ1-4Glcβ-，Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ1-3Galβ1-，Neu5Acα2-3Galβ1-4GlcNAcβ1-3Galβ1-4GlcNAcβ-．この結果から 3Galβ1 と 4GlcNAcβ1 が交互に続いている糖鎖が結合親和性が高くなる構造であることが分かり，糖鎖の末端部分だけでなく，糖鎖の真ん中の部位が結合時に認識されていることが予想できる．これらの部分構造の有無で分割したときの箱ひげ図を**図 3** で示す．図の青いグラフがこの部分構造を含む糖鎖，赤のグラフは部分構造を含まない糖鎖の結合親和度の分布を表す．このように本手法で提案した指標により，部分構造の有無で糖鎖を上手く分割できたことが示せた.

Influenza B variant#71
インフルエンザ B 型の変位型である Influenza B variant#71 を 6) の Ver. 3.1 のグリカンアレイに濃度 2000, 5000, 10000, 20000, 50000HAU/ml で与えた実験について解析した結果を示す.
指標を適用すると糖鎖構造が 3 つの場合が最も指標の値が大きくなることが分かり，その時の部分構造は，**表 4** のランキングの順番になる．1 位，2 位では 3 つの糖鎖構造を含み，3 位では 2 つのの糖鎖構造を含むものが予測できた．表 4 の 1 位は以下のいずれかを含んでいる

Vol.2009-BIO-18 No.6
2009/9/17

表 3 Galectin-1 の実行結果
Table 3 The results of Galectin-1

rank	substructure	g-index
1	*—3 GlcNAcb1—3 Galb1—4 Glcb—* Galb1—4 GlcNAcb1—3 Galb1—4 GlcNAcb1—3 Galb1—* Neu5Aca2—3 Galb1—4 GlcNAcb1—3 Galb1—4 GlcNAcb—*	372.7
2	*—3 GlcNAcb1—3 Galb1—4 Glcb—* *—4 GlcNAcb1—3 GalNAca—* Galb1—4 GlcNAcb1—3 Galb1—4 GlcNAcb1—3 Galb1—GlcNAc4—b1*	324.4
3	*—2 Galb1—3 GlcNAcb1—3 Galb1—4 Glcb—* *—4 GlcNAcb1—3 Galb1—4 GlcNAcb1—3 Galb1—4 GlcNAcb—*	292.4

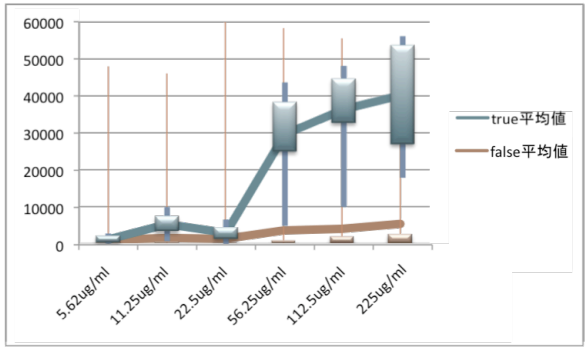


図 3 Galectin-1 の実行結果（箱ひげ図）
Fig. 3 The results of Galectin-1(box plot)

糖鎖である．Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-3(-6Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-，-6Galβ1-4GlcNAcβ1-2Manα1-3(Neu5Acα2-6Galβ1-4GlcNAcβ1-2Manα1-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-，-6GlcNAcβ1-．この結果からNeu5Acα2-6Galβ1-4GlcNAcβ1- を末端に持つ糖鎖が結合親和性が高くなる構造であることが分かり，それ以外にも-6GlcNAcβ1-の構造を認識していることが予想される．また 3 位の構造では 1 位と 2 位に比べて末端の Neu5Acα2-がなくなると g-index の値が小さくなっ

表 4 InfluenzaB の実行結果
Table 4 The results of InfluenzaB

rank	substructure	g-index
1	Neu5Aca2—6 Galb1—4 GlcNAcb1—2 Man a1—3 Manb1—4 GlcNAcb1—4 GlcNAcb—* *—6 Galb1—4 GlcNAcb1—2 Man a1—6 *—6 Galb1—4 GlcNAcb1—2 Man a1—3 Manb1—4 GlcNAcb1—4 GlcNAcb—* Neu5Aca2—6 Galb1—4 GlcNAcb1—2 Man a1—6 *—6 GlcNAcb1—*	293.4
2	Neu5Aca2—6 Galb1—4 GlcNAcb1—2 Man a1—3 Manb1—4 GlcNAcb1—4 GlcNAcb—* *—6 Galb1—4 GlcNAcb1—2 Man a1—6 *—6 Galb1—4 GlcNAcb1—2 Man a1—3 Manb1—4 GlcNAcb1—4 GlcNAcb—* Neu5Aca2—6 Galb1—4 GlcNAcb1—2 Man a1—6 *—6 Galb1—4 GlcNAcb1—3 Galb1—4 GalNAcb1—3 Galb1—*	290.5
3	Neu5Aca2—6 Galb1—4 GlcNAcb1—2 Man a1—3 Manb1—4 GlcNAcb1—4 GlcNAcb—* *—4 GlcNAcb1—2 Man a1—6 *—6 Galb1—4 GlcNAcb1—2 Man a1—3 Manb1—4 GlcNAcb1—4 GlcNAcb—* *—6 Galb1—4 GlcNAcb1—2 Man a1—6 *—6 Galb1—4 GlcNAcb1—3 Galb1—4 GalNAcb1—3 Galb1—*	120.2

ていることから，Neu5Acα2-の末端構造が特に重要であることが予想できる．

4.3 既知のデータとの比較

Galectin-1

糖結合タンパク質として知られている Galectin は現在 14 種類が同定されていて，一般的に β-ガラクトース含有糖鎖と結合することが知られている．特に Galectin-1 が認識する糖鎖部位としては N-アセチルラクトサミン（Galβ1-4GlcNAc）があり，本研究の結果と一致する．またこの繰り返しからなるポリ-N-アセチルラクトサミン｛(3Galβ1,4GlcNAcβ)n｝は多くの場合 N-アセチルラクトサミンよりも結合親和性が高くなることやポリラクトサミンの途中に枝分かれを有していないことがわかっており[8]，結果と一致した．さらに Galectin-1 では糖鎖の末端だけでなく，糖鎖の内側の構造を認識して結合することが知ら

Vol.2009-BIO-18 No.6
2009/9/17

れており，それによって結果にも現れているものと思われる．

Influenza B variant#71

一般的にインフルエンザ A 型や B 型はシアル酸末端（Neu5Acα-）を認識し結合することが知られている．特にインフルエンザ B 型はシアリル α2-6 ラクトテトラオシルセラミド（Neu5Acα2-6Galβ1-3GlcNAcβ1-4Galβ1-4Glcβ1-1'Cer）を強く認識することが知られている[9]．またこれ以外にもシアリル α2-3 ラクトネオテトラオシルセラミド，シアリル α2-3 パラグロボシドに対する反応も微弱にある．これらの構造に共通するのは Neu5Acα2-6(3)Galβ1-3(4)GlcNAcβ1-を含む構造であることである．これは本手法で予測した結果と一致した．

以上の結果からタンパク質やウィルスの既知の結合認識部位と一致し，本手法の有用性を示すことができた．

4.4 その他の実行結果

Phloem Protein

糖鎖に結合するレクチンである Phloem Protein2-A1（0.1μg/ml，1μg/ml，10μg/ml，50μg/ml, 100μg/ml）の解析結果を示す．

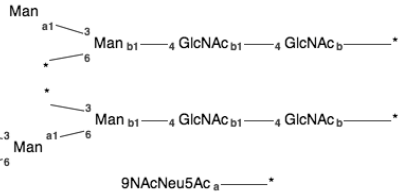
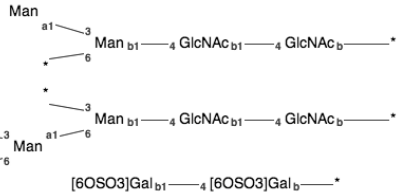
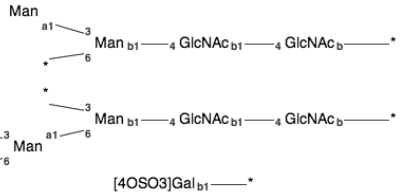
分割の指標を適用するとこの結果は，**表 5** で示した様になった．Phloem Protein においても糖鎖構造が三つの場合が最も指標の値が大きくなることが分かり，最も結合が強くなる部分構造は，9NAcNeu5Acα-，Manα1-3(-6)Manβ1-4GlcNAcβ1-4GlcNAcβ-，-3(-6)Manα1-6(-3)Manβ1-4GlcNAcβ1-4GlcNAcβ-，のいずれかを含んでいる糖鎖であることが分かった．この結果から Manα1-3(6)Manβ1-4GlcNAcβ1-4GlcNAcβ-の構造が末端にあっても，末端になくても結合に関わっていることが予想される．

またその構造とは別に 9NAcNeu5Acα-の末端構造も結合に関わっていることが解析の結果分かった．また 2 位，3 位では Manα1-3(6)Manβ1-4GlcNAcβ1-4GlcNAcβ-は共通して現れており，この構造以外に末端構造の違いにより結合の強さが変わることがわかった．このように糖結合部位がまだ知られていないものを予測することで，実験の時にある程度予想を付けることができるため実験が効率的に行えると考えた．

5. まとめと今後の課題

本研究ではタンパク質やウィルスが糖鎖と結合するときに糖鎖のどの部分構造を認識するかについての解析を行った．糖鎖の部分構造の有無に着目し，複数の実験において結合親和性が高くなるような糖鎖構造のを発見できる様な指標を開発し，その結果既知のタンパク質

表 5 Phloem Protein の実行結果
Table 5 The results of Phloem Protein

rank	substructure	g-index
1		27.0
2		16.4
3		16.1

である Galectin-1，Influenza B variant#71 について結果を確認することができた．またタンパク質が糖鎖と結合するときに認識する部分構造は複数ある場合があり，それは末端だけでなく，糖鎖の中間の部位にも起こりえることが考えられる．

現時点では組み合わせを取る数が 3 つまでと少なく，さらに 3 つの組み合わせを取ったとき計算時間が数時間かかるため，今後 3 つより多い組み合わせを取るために効率よく部分構造を探索できる方法を考え，計算時間を少なくするアルゴリズムを開発したい．また本手法で得られた結果を実際の実験でも確認し，本研究の有用性を検証していきたい．

Vol.2009-BIO-18 No.6
2009/9/17

参 考 文 献

1) J.M. Rini. *Lectin Structure*. Ann. Rev. Biophys. Biomolec. Struct., 24:551-557, 1995
2) V.N. Vapnik. *Statistical Learning Theory*. Wiley, New Tork, 1998.
3) H. Saigo, T. Kadowaki, and K. Tsuda. *A linear programming approach for molecular QSAR analysis*. In Proc. of the International Workshop on Mining and Learning with Graphs (MLG), 85–96, 2006.
4) K. Aoki-Kinoshita. *An introduction to bioinformatics for glycomics research*. PLoS Computational Biology, Vol. 4, No. 5, 2008.
5) Y. Yamanishi, F Bach, and J. Vert *Glycan classification with tree kernels*. Bioinformatics, Vol. 23 No. 10, 1211–1216, 2007.
6) Consortium for Functional Glycomics, Nature. Functional Glycomics Gateway http://www.functionalglycomics.org/
7) Raman R, Venkataraman M, Ramakrishnan S, Lang W, Raguram S, and Sasisekharan R. *Advancing glycomics: Implementation strategies at the Consortium for Functional Glycomics*. Glycobiology, 16(5), 82R:90R, 2006.
8) Cho, Moonjae; and Cummings, Richard D. *Galectin Structure*. Trends in Glyco-Science and Glycotechnology, Vol. 9 No. 45, 47–56, 1997.
9) Y. Suzuki. *Variation of Influenza Viruses and Their Recognition of the Receptor Sialo-Sugar Chains*. Journal of the Pharmaceutical Society of Japan, Vol. 113 No. 8, 556–578, 1993.