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

理学専攻 情報科学コース　寺井　はるな　(指導教員：瀬々　潤)

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．

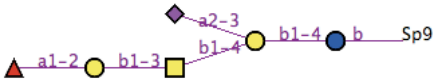


図 1: 糖鎖の構造

図 1 は糖鎖構造の略図で，ラベル有り根付き順序木として表される．右端は根で細胞表面に付着しており，左端は葉である．頂点，辺共にラベルがあり，頂点のラベルは単糖の種類 (丸や三角の記号)，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．

本研究では，結合時にタンパク質やウィルスに認識される糖鎖の部分構造を予測する方法を提案する．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度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，結合時に認識される部分構造の組み合わせを特定する．また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，結合特異性を高める第二の認識部位が存在すると考えられており [1]，本研究ではそのような第二の認識部位も考慮した解析を行った．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 Consortium for Functional Glycomics (CFG)[2][3] には 1 つのタンパク質を複数の濃度で実験したデータがあり，その一連の実験を一度に解析した．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と考えた．

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 [4] 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 [5]，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 [6]．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　研究手法

3.1　定義

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 1 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．

3.2　制約条件付きクラスタ抽出

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や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．

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 に近い値を取る事から，$X(S)$ 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，着目する部分構造 S を有する糖鎖の親和度の平均 $\mu(S)$ を計算することにより評価できる．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以下で定義する．

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

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．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 $\sigma(S)$ を定義する．

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

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変化量 $\delta(S)$ で定義する．

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

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．

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

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．

3.3 部分構造の組み合わせへの拡張

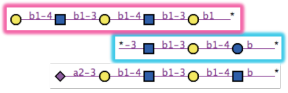
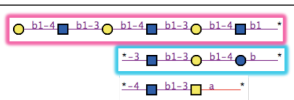
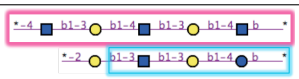
前述の通り，糖鎖には第二の認識部位があると考えられており，本研究ではその構造を抽出するために部分構造の和集合を考える事で解決する．部分構造群 $\mathcal{S}$ に対し，糖鎖集合 $X(\mathcal{S}) = \{x \in X \mid \exists S \in \mathcal{S} \text{ s.t. } S \subseteq T(x)\}$ を定義する．この定義は，部分構造 S を持つ糖鎖群 $X(S)$ を $\mathcal{S}$ 内のいずれかの部分構造を含む糖鎖群に拡張したものである．$\mu(S)$, $\sigma(S)$, $\delta(S)$ の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．

4 実行結果と既知データとの比較

Galectin-1

糖結合タンパク質である Galectin-1 を [3] 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, 225[μg/ml] で与えた実験について解析した結果を示す．上記の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 1 のランキングの順番になる．

表 1: Galectin-1 の実行結果

rank	substructure	g-index
1		372.7
2		324.4
3		292.4

Galectin-1 はポリ-N-アセチルラクトサミン $(3Galb1,4GlcNAcb)_n$ を認識することが知られており [7]，表 1 で赤く囲ったランキング上位の結果と一致した．なお青で囲った構造がランキング上位に入っており，第二の認識部位であると予測できる．

またこれらの部分構造の有無で糖鎖をクラスタに分割したときの箱ひげ図を図 2 に示す．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 型の変異型である Influenza B variant#71 を [3] の Ver. 3.1 のグリカンアレイに濃度 2000, 5000, 10000, 20000, 50000[HAU/ml] で与えた実験について解析した結果を表 2 に示す．

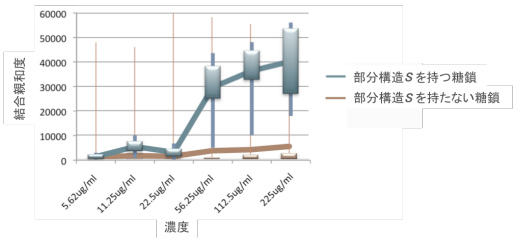
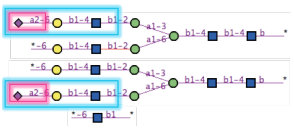
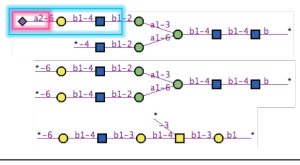


図 2: Galectin-1 の実行結果（箱ひげ図）

表 2: InfluenzaB の実行結果

rank	substructure	g-index
1		293.4
2		290.5
3		120.2

一般的にインフルエンザウィルスは赤で囲ったシアル酸末端 Neu5Aca-に結合する．特にインフルエンザ B 型では青で囲った Neu5Aca2-6(3)Galb1-3(4)GlcNAcb1-に対する反応があり [8]，結果と一致した．

5 まとめ

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どの糖鎖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糖鎖構造を発見できる指標を開発し，その結果 Galectin-1，Influenza B variant#71 について結果を確認した．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確認できた．

参考文献

[1] Rini, J.M., Lectin Structure, *Ann. Rev. Biophys. Biomolec. Struct.*, 24:551-557, 1995.

[2] 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.

[3] Consortium for Functional Glycomics (CFG), http://www.functionalglycomics.org/

[4] Saigo, H., Kadowaki, T., and Tsuda, K., A linear programming approach for molecular QSAR analysis, *In Proc. of the International Workshop on Mining and Learning with Graphs (MLG)*, 85–96, 2006.

[5] Aoki-Kinoshita, K., An introduction to bioinformatics for glycomics research, *PLoS Computational Biology*, Vol. 4, No. 5, 2008.

[6] Yamanishi, Y.,Bach, F., and Vert, J., Glycan classification with tree kernels, *Bioinformatics*, Vol. 23 No. 10, 1211–1216, 2007.

[7] Cho, Moonjae; and Cummings, Richard D., Galectin Structure, *Trends in GlycoScience and Glycotechnology*, Vol. 9 No. 45, 47–56, 1997.

[8] Suzuki, Y., 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.