クラスタリングによる糖鎖認識パターン解析

伊藤真和吏　(指導教員：瀬々潤)

1　はじめに

近年，核酸 (DNA)，アミノ酸に続く第三の鎖状生体分子として糖鎖が着目されている．この糖鎖は生体内のほとんどのタンパク質や脂質に結合し，免疫，細胞内輸送など生命機能に重要な役割を果たしている．例えば，インフルエンザウィルスは，細胞表面 にある特定の糖鎖を認識して感染する．

DNA やタンパク質の鎖は分岐構造が無く直線的に結合しているのに対し，糖鎖は糖が木構造になって結合している．図 1 に構造例を示す．図の四角や丸で表した頂点が糖を示しており，色や形は異なる種の糖である事を表している．また，糖と糖の間の結合が辺で表されている．また，辺に付けられたラベルは糖と糖が結合している位置と立体構造を示している．また，一般に糖鎖を書いた場合の右端がタンパク質に結合し，左端が細胞外の反応部位である．生体内で見られる糖は約 10 種類程度である．

このように生体内で重要な役割を示す糖鎖であるが実験サンプル（タンパク質，ウイルスの種類，以下サンプルという）によって認識するが糖鎖の構造が異なっている事が知られている．この機構解明に向け，近年開発されたグリカンアレイ [2] の利用が行われている．グリカンアレイは，スライドガラス上に多数の糖鎖をスポットし，その上からサンプルを流すことで糖鎖とサンプルを反応させ，各糖鎖と 1 種類のサンプルが結合する度合い (結合親和度) を一度に大量に計測することが可能な実験である．このグリカンアレイ実験結果は，Consortium for Functional Glycomics[1] (CFG) に蓄積されている．本研究では，CFG のデータを利用し，糖鎖とサンプルの間の反応を解析した．

2　手法

本研究の目的は，糖鎖間，サンプル間の類似度を計測することで，それらの反応の共通性を明らかにすることである．しかし，実験ノイズ及びデバイスの特性が未知であるため，糖鎖間，サンプル間の計測には考察が必要である．本研究では次の手段を用いてグリカンアレイに合う距離法の評価を行った．まず，様々な距離法とクラスタリング手法を用い，クラスタを生成する．次に各クラスタに対し，適切な機能を割り当てる（2.2 節）．機能の割り当てに関し，割り当ての妥当性と網羅性を検証するため，F 値の改良版を用いて評価する（2.4 節）．以上の内容に関して，以下サンプルの場合を例にして詳細を記すが，糖鎖構造間に対しても同様に適切な距離を導入する．

2.1　糖鎖アレイデータの形式

データベース中の全糖鎖集合を $G(|G| = n)$，全サンプル集合を $S(|S| = m)$ とする．また，糖鎖 $g \in G$ に対するサンプル $s \in S$ の結合親和度を a_{gs} と表す．更に，サンプル $s \in S$ に対し，$\mathbf{a}_s = (a_{1s}, ..., a_{ns})$ をサンプル s の親和度ベクトルと呼ぶ．本研究では，この親和度ベクトル間の距離を計測する適切な方法を求める事が目標となる．用いるデータは，糖鎖の数が 377 種類（$n = 377$），サンプル数が 291 種類（$m = 291$）のデータである．

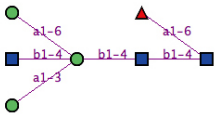

図 1: 糖鎖の木構造

2.2　距離法とクラスタリング

ベクトル間の距離を測る方法として 3 種類の異なる方法を調査した．サンプル s, t 間の距離について，それぞれの求め方は下記の通りである．

$$\text{Euclid 距離:} \quad d(s,t) = \sqrt{\sum_{g \in G} (a_{gs} - a_{gt})^2}$$

$$\text{Manhattan 距離:} \quad d(s,t) = \sum_{g \in G} |a_{gs} - a_{gt}|$$

$$\text{cosine 距離:} \quad d(s,t) = \frac{\mathbf{a}_s \cdot \mathbf{a}_t}{|\mathbf{a}_s||\mathbf{a}_t|}$$

クラスタリング手法に関して，ここではボトムアップ型の階層型クラスタリングを用いる．クラスタ間の距離を計測する手法として，以下の 3 種類を利用する．2 つのクラスタ C と C'(各クラスタはサンプルの集合) に対し，クラスタ C と C' 間の距離を $D(C, C')$ と表す．3 手法の距離の計測方法は下記の通りと成る．以上より同一データに対し合計 9 種類のクラスタが求まる．

$$\text{単連結法:} \quad D(C,C') = \min_{s \in C, t \in C'} d(s,t)$$

$$\text{完全連結法:} \quad D(C,C') = \max_{s \in C, t \in C'} d(s,t)$$

$$\text{Ward 法:} \quad D(C,C') = E(C \cup C') - E(C) - E(C')$$

$$\text{ただし } E(C) = \sum_{s \in C} \sum_{g \in G} (a_{gs} - c_g)^2, c_g = \frac{1}{|C|} \sum_{s \in C} a_{gs}$$

2.3　クラスタの評価

クラスタの評価としてサンプルによって濃度が異なるだけで同一のサンプルを用いている実験が多数あることに着目し，それらのサンプルが同一のクラスタに固まっているかを判定した．判定には以下の通り，二項検定を用いた．

着目するクラスタを C とする．$|C|$ はクラスタ内のサンプル数を表す．また，着目する同一サンプルから行われた実験グループを S とする．$|S|$ はその同一サンプルの実験数を表す．また $n = |C \cap S|$ とすると，クラスタ C に実験グループ S が有意に固まっているかは，以下の二項検定を用いて調べる事ができる．

$$p(C,S) = \sum_{i=n}^{|C|} {}_{|C|}\mathrm{C}_i \left(\frac{|S|}{291}\right)^i \left(\frac{291 - |S|}{291}\right)^{|C|-i}$$

あるクラスタ C に対し，最小の $p(C, S)$ を取るサンプルを S とすると，$p(C, S) \leq 0.01$ の場合，S は有意

[1] http://www.functionalglycomics.org

表 1: サンプルの F 値

	Ward	Complete	Single
Euclid	0.68	0.59	0.34
Manhattan	0.68	0.44	0.24
Cosine	0.71	0.59	0.34

表 2: 糖鎖の F 値

	Ward	Complete	Single
Euclid	0.52	0.42	0.17
Manhattan	0.53	0.43	0.18
Cosine	0.64	0.53	0.20

にそのクラスタに集まっていると考え，$p(C, S) > 0.01$ の場合は関連づけられるサンプルは無いとする．

2.4 F 値

良いクラスタ分割が出来た場合 (1) 多くのクラスタに対し実験グループが関連づけられる，(2) 多くの実験グループがクラスタに関連づけられる，の 2 つの関連づけができると考えられる．これらの 2 つの条件を満たすような指標として F 値の改良版を利用する．$|\mathcal{C}|$ をクラスタ数, $|\mathcal{S}|$ を実験の種類数とし，$|\mathcal{C}'|$ を実験が関連付いたクラスタ数，$|\mathcal{S}'|$ を関連付いた実験の種類の数とする．F 値を定義するために前述の (1), (2) に相当する 2 種類の指標を定義する．

$$\text{Enrichment} = \frac{|\mathcal{C}'|}{|\mathcal{C}|}, \text{Coverage} = \frac{|\mathcal{S}'|}{|\mathcal{S}|}$$

これらを用い，改良版の F 値を下記の通り定義する．

$$\text{F 値} \quad = \quad 2 \Big/ \left(\frac{1}{\text{Enrichment}} + \frac{1}{\text{Precision}} \right)$$

F 値の最大は 1 であるが，一般に Enrichment が上がると Coverage が下がり，逆も然りのため，値を大きくするには，両者のバランスが取れる必要がある．

3 実験と考察

3.1 実験

今回使用したグリカンアレイのデータは，CFG で配布されているものの内，グリカンアレイのバージョンが 3.1 のものを集め，合計 291 サンプル，377 個の糖鎖の結合親和度を含む．

このデータから結合親和度の分布を書くと，0 付近に数多く存在し，正規分布からは，かけ離れた分布である．更に，サンプル毎に見た場合，サンプルの反応する最大値に大きなバラツキがある．このサンプル間の感度の違いをならすため，本研究ではサンプル毎に平均 0，分散 1 にする標準化 を前処理として行った．

また，今回利用した全サンプルをサンプルの種類でグループでまとめるとグループ数は 51 であった．F 値の計算方法を鑑み，距離やクラスタ手法が異なっても公平に判断できるよう，この実験でのクラスタ数を一律 50 とした．糖鎖間距離の評価においても，各糖の有無で分けたグループ数が 50 であったため，クラスタ数を 50 とした．

3.2 結果と考察

本実験では，F 値を求める事で，どの手法が最も適切に糖鎖の距離を反映しているかを検証した．表 1 に

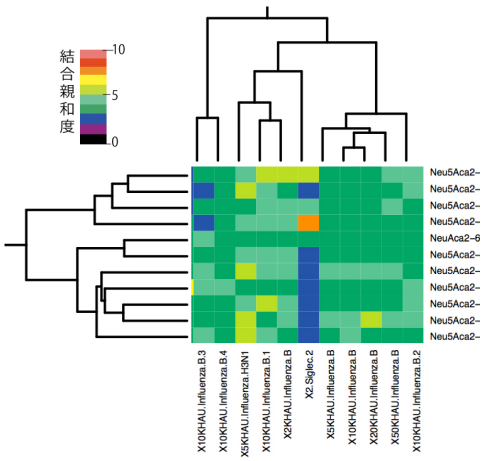

図 2: クラスタの拡大図

サンプル，表 2 に糖鎖の場合の結果を示している．

上記の表より，最も F 値が高いのは距離法は Cosine 距離かつクラスタリングの方法が Ward 法の場合であり，この手法が最もサンプル間距離を計る為に適した指標と考えられる．

図 2 はインフルエンザのクラスタとそれに強く反応する糖鎖のクラスタを表したものである。図の中心の明色部分は結合親和度が高いことを示し，左側と上側はデンドログラムである．右側は糖鎖を示し，下側はサンプルを示している．サンプルにおいて，たとえば x10KHAUInfluenza.B は，10KHAU[2]の濃度での Influenza.B の実験であることを示している．糖鎖の名称は，左側の文字列が糖鎖の末端構造である．図 2 からサンプルのクラスタは主に Influenza.B，糖鎖のクラスタは末端の構造が Neu5Ac（シアル酸 ･･･ 通常糖鎖の末端に存在し，細胞の認識などの役割を担っている）が主であることがわかる。このため，本クラスタリングの精度が高いことが分かる．更に，インフルエンザのクラスタの中にインフルエンザ以外にシアル酸を認識する事が知られている Siglec が含まれている．インフルエンザも，このシアル酸を認識する事が知られているため [3]，既存の知識との整合性がとれている．

4 まとめと今後の課題

本研究の結果，既知の生命科学の知識に沿うクラスタリング手法を得ることが出来た．生物的な知識を判別に全く用いずにこの結果が得られたことは非常に有用である．今度は，糖鎖のどの部分構造が特定のサンプルに反応しているのか調査するため，糖鎖の部分構造にも着目して糖鎖間の距離を測りたい．そして，糖鎖の構造と機能の解明についての手がかりを得たい．

参考文献

[1] KF. Aoki-Kinoshita. An Introduction to Bioinformatics for Glycomics Research. PLoS Computational Biology, 4(5), e1000075, 2008.

[2] O Blixt *et al.* Printed covalent glycan array for ligand profiling of diverse glycan bind ing proteins. Proc Natl Acad Sci, Vol. 101, No. 49, pp. 17033-17038, 2004.

[3] 鍔田武志．シアル酸の多様性と認識：感染と免疫の共進化．細胞工学．学研メディカル秀潤社，2007

[2] 赤血球凝集単位