論文の内容の要旨

論文題目　　タンパク質–糖鎖結合予測手法の開発

氏　　名　　葛　臻翼

１．背景

糖鎖結合タンパク質（carbohydrate-binding protein,CBP またレクチン）は、糖鎖に結合活性を示すタンパク質の総称で、細胞間の情報伝達や細胞種類の識別や細胞の免疫など多種多様な生体活動に関与している。タンパク質の糖鎖結合性の測定には、成分抽出/cDNA からのタンパク質発現によって得られたタンパク質に対し、アフィニティー/イオン交換クロマトグラフィーによる精製分画、凝集活性測定、コロニー形成阻害、阻害糖の実験などが行われている。これらは糖鎖結合タンパク質の同定や活性の定量・定性のためには必須の作業であるが、多くの時間と労力を必要とする。

また、糖鎖結合タンパク質には、多数の種類が存在する。分類のしかたにはさまざまな方法があり、糖鎖リガンドの情報を利用した分類がよく用いられているが、分子クローニングなどで明らかになったアミノ酸配列のホモロジーやモチーフの存在によって分類することができる。この方法により、糖を結合する際にカルシウムを必要とする C-型レクチン、糖鎖の中でガラクトース（galctose）を含む糖鎖構造（β-ガラクトシド）によく結合するガレクチンなどのタイプに分類されている。

本研究では、アミノ酸配列情報のみを用いて、糖鎖結合タンパク質を予測するとともに、糖鎖結合タンパク質のいくつかの主要なタイプを予測するシステムを開発し、その評価を行なった。本システムは、与えられたタンパク質が糖鎖と結合するかどうか、またそれらタンパク質の分類を、アミノ酸配列情報の

みから Support Vector Machine（SVM）を用いて学習・予測するというものであり、ゲノムワイドな解析にも適用できる。

２．材料と手法

本研究では、まず、糖鎖結合タンパク質を予測する手法を開発した（図 1）。研究対象としての糖鎖結合タンパク質としては、抗体以外の「糖鎖と構造特異的に相互作用し、抗体でなく、糖鎖を直接修飾しないタンパク質」を一括して扱うことにした。そこで、これらのタンパク質をデータベース UniProt Knowledgebase から抽出する際の検索条件の定式化を行った。さらに、糖鎖結合タンパク質の配列特徴を効果的に学習させるため、これらのアミノ酸配列に対し、BLAST によるクラスタリングを行い、配列冗長性を排除したデータセット（正例データセット）を作成した。一方、非糖鎖結合タンパク質のデータセット（負例データセット）としては、実際に発現が確認されているタンパク質の中から、糖鎖結合タンパク質の検索条件に合致しないものをランダムに収集し、上と同様にして冗長性を排除したものを用いた。さらに、多類分類のため、種類が明記されている糖鎖結合タンパク質のアミノ酸配列を収集した。糖鎖結合タンパク質の配列特徴を効果的に学習させるため、これらのアミノ酸配列に対し、クラスタリングを行い、配列冗長性を排除したデータセットを用いた。このデータセットの一部は、テストデータセットとして保留し、残り大部分をトレーニングデータセットとして SVM に投入した。

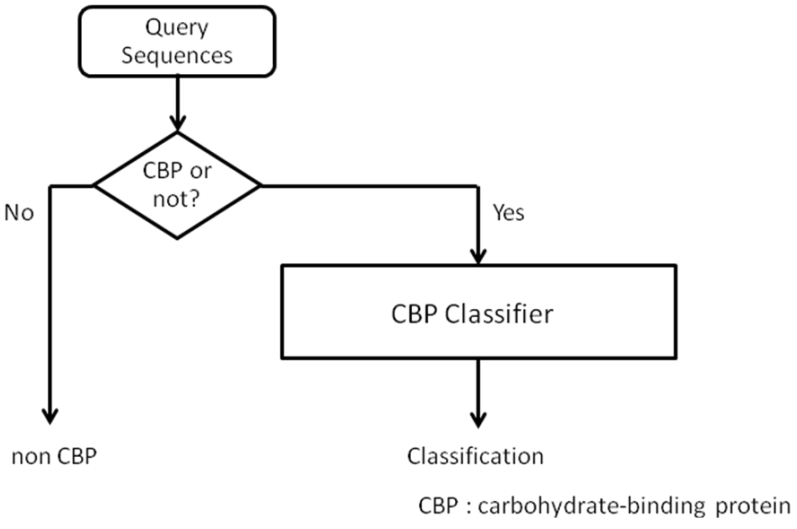

図 1　タンパク質–糖鎖結合予測手法の概要

学習については、アミノ酸配列から特徴ベクトルを作成し、SVM への入力とした。配列情報を特徴空間上に写像させるカーネル関数としては、アミノ酸の 3 つ組の出現パターンに基づく 3-spectrum kernel を用いた。5 分割交差確認

（5-fold cross validation）により SVM 最適なパラメータを求めて、モデルを構築して、テストデータセットの予測結果を評価した。

SVM は汎化性能が高く、未学習のデータの識別に優れる機械学習の方法である。二値分類器である SVM は、多値分類問題を解決するため、複数の SVM を組み合わせることで多値分類を実現する。本研究では、ある一つのクラスとそれを除く残りのすべてのクラスを分類する One-versus-Rest 法と、ある一つのクラスと別な一つのクラスの分類をすべてのクラスに対して適用する One-versus-One 法の 2 種類を用いた。

３．結果と考察

糖鎖結合タンパク質の予測では、AUC（Area Under the Curve）の値は 0.797 で、実用レベルの高い予測精度が達成できた。分類については、One-versus-Rest 法におけるトレーニング 5 分割交差確認精度は 93.93%（平均）で、テストデータセットを予測すると精度は 94.72%（平均）であった。一方、One-versus-One 法ではトレーニング 5 分割交差確認精度は 83.81%で、テストデータセットを予測すると精度は 85.98% であった。この結果、One-versus-Rest 法の方が高い精度で分類を行うことができたことがわかる。これは、One-versus-Rest 法では、SVM 予測精度に重要なパラメータ C（cost）とγ（gamma）を細かく調整できるが、One-versus-One 方法は多数のモデルを構築し、個別に最適なパラメータを求めていないためと考えられる。

なお、糖鎖結合タンパク質の各タイプの予測精度にも差があることをわかった。ドメインが配列全長に占める比が高いタイプは精度が高い傾向にあり、さらに、タイプ内の配列の類似性の傾向も予測精度と関与していることが示唆されている。

表 1　糖鎖結合タンパク質の分類性能（One-versus-Rest 法）

Type	Accuracy		Average coverage of the sequence by the domain (From Pfam)	Average identity of full alignment (From Pfam)
C type	94.02%	456/485	18.32 %	21 %
Galectin	98.35%	477/485	38.79 %	25 %
L type	98.35%	477/485	37.58 %	27 %
P type	N/A	N/A	25.28 %	45 %
R type	90.10%	437/485	21.22 %	19 %
R type like	92.78%	450/485	15.26 %	23 %