［ 別 紙 ２ ］

論 文 審 査 の 結 果 の 要 旨

申請者氏名　葛 臻翼

糖鎖結合タンパク質（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。

本論文では、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研究について述べたものである。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。本論文は、4 章より構成される。

本論文の第 1 章で、研究の背景および目的を述べた後、第 2 章では、本研究で用いた手法について述べている。まず、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とし、これらのタンパク質をデータベース UniProtKB 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を明記し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。

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、あるひとつのクラスとそれを除く残りのすべてのクラスに対する分類を適用可能なクラスに対して行う方法を実行した。

第 3 章では、本研究の結果と考察について述べている。まず、糖鎖結合タンパク質の予測は 0.8 以上の AUC 値を達成し、実用レベルで利用できることを示した。予測精度をさらに改善するため、負例データセットとの出現頻度の対数比がしきい値より低いもの（両者の差が小さいもの）を除き、正例データセットに特徴的に見られる 3 つ組だけをもとに学習・予測する手法の開発を試みた結果も述べており、精度の改善にはさらに検討が必要であることを述べている。

糖鎖結合タンパク質の種類（レクチンタイプ）の予測では、現在よく用いられている 2 つの多クラス予測手法（One-versus-One 法と One-versus-Rest 法）を試した。両方の手法で SVM の予測精度に重要なパラメータ C（cost）とγ（gamma）を細かく調整したが、One-versus-Rest 法が、AUC 値が 94.7%（平均）、One-versus-One 方法が、AUC 値が 86.0%（平均）という結果が得られた。One-versus-Rest 法の方が高い精度を得ることができた理由は、One-versus-One 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た。各タイプの予測精度に差があったが、これは使っているデータの配列相同性や保存度などの特徴に関係していると考えられることを述べている。

第 4 章では、本論文の内容をまとめており、開発したシステムの実用性、精度改善の余地、今後の課題について考察を行っている。以上、本論文の成果は、学術上応用上貢献するところが少なくない。よって、審査委員一同は、本論文が博士（農学）の学位論文として価値あるものと認めた。