［別紙２］

論文審査の結果の要旨

申請者氏名　　曹　巍

タンパク質の翻訳後脂質修飾は、タンパク質の構造や物理化学特性を変え、細胞機能にとって重要な役割をもっている。しかしながら、実験による翻訳後脂質修飾の解析は相応の時間とコストを要し、また、現在、タンパク質の配列データベースの登録数が増大している中で、翻訳後糖質修飾のアノテーションは十分になされていないのが現状である。本論文では、タンパク質翻訳後脂質修飾として、グリコシルホスファチジルイノシトール（GPI）アンカー型修飾とミリストイル修飾の２つを取り上げ、これらを、タンパク質のアミノ酸配列から予測する手法について述べている。本論文は、５章より構成されている。

第一章では、GPI アンカー型タンパク質とミリストイル化タンパク質の修飾に関する概要とこれまでの研究で得られている知見をまとめ、本研究の背景について記している。

第二章では、本論文で使用した手法について述べている。本論文では、翻訳後脂質修飾を特徴づける配列パターンを「学習」によって予測するアプローチをとることとし、機械学習の手法として近年よく用いられているサポートベクターマシン（SVM）を採用している。SVM は、２クラスの分類を行う機械学習手法であり、サポートベクトルと呼ばれるクラス境界近くに位置する学習点とのマージンを最大化するよう分離平面を構築するというものである。このマージン最大化という基準を用いることにより、高い汎化性能（未学習データに対する分類能）をもつところが、他の機械学習手法と比べてとくに優れている。本論文では、SVM の重要なパラメータである汎化パラメータ C、および Radial Basis Function（RBF）カーネル関数のパラメータ γ について最適化を行うとともに、学習の対象となるアミノ酸配列を簡易に表現することにより、予測精度の向上を目指している。パラメータ最適化の手段としては、Particle Swarm Optimization（PSO）法を利用している。また、性能の評価は、5-fold cross-validation (CV) テストを用いている。これは、データセットをランダムに５個のサンプル群に分け、４個のサンプル群で学習（訓練）を行った SVM を残りの１個のサンプル群に適用して予測を行うという手法である。

第三章では、GPI アンカー型タンパク質の予測について述べている。予測・学習に用いたデータセットは、positive データセットとしては、Swiss-Prot データベースを「GPI-ANCHOR」をキーワードとして検索して得られた 531 個のタンパク質、negative データセットとしては、Pascal Mäser らが、彼らの予測研究で使用したデータセットをそのまま使用している。予測に先立ち、C 末端側の配列に対して、残基

の疎水性を調べたところ、とくにC末端の20残基において他の部分と疎水性に大きな差があることが明らかとなり、疎水性が予測の鍵になることを発見した。また、SVMの入力（学習対象）として、最も効果的な学習が行える配列部分を調べた結果、C末端の60残基であるとの結論を得た。これらの結果をふまえ、本論文では、C末端の60残基のアミノ酸配列をKyte-Doolittleの疎水性指標の配列（実際には、周辺残基を含む9残基をウィンドウとしてその平均値を適用）に変換し、SVMの入力とすることで、96%の高い予測精度を達成している。

第四章では、ミリストイル化タンパク質の予測について述べている。予測・学習に用いたデータセットは、positiveデータセットとしては、Swiss-Protデータベースを「myristate」をキーワードとして検索して得られた449個のタンパク質、negativeデータセットとしては、GPIアンカー型タンパク質の予測で用いたPascal Mäserらが使用したデータセットをそのまま使用した。SVMに入力する配列部分は、モチーフパターンが得られているN末端の6残基のほか、それに続くNMT表面との相互作用部位、親水性リンカー領域を加えた合計17残基としている。これらのアミノ酸配列をKyte-Doolittleの疎水性指標の配列（実際には、周辺残基を含む3残基をウィンドウとしてその平均値を適用）に変換し、さらに、各残基位置におけるアミノ酸組成またはアミノ酸特性を数値化したAA-indexを加えて予測を行った。AA-indexについては、冗長性を除いた530種類の指標を網羅的に試した。その結果、アミノ酸組成を加えた手法では、97.2%の予測精度が得られ、AA-indexを加えた手法では、とくに予測精度の高い10個のAA-indexについて、98%～99%の予測精度が得られた。これら10個のAA-indexはすべて構造と安定性に関係するものであった。

第五章では、これらの結果をまとめ、今後の展望について述べている。

以上本論文は、機械学習SVMを用いてタンパク質の翻訳後修飾（GPIアンカー型タンパク質とミリストイル化タンパク質）を予測する、簡易で高速な手法を開発したものであり、その手法は、全ゲノムレベルで適用できるものと期待され、また、同様の手法を他の翻訳後修飾に適用できる拡張性を備えており、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。