機械学習（SVM）を用いたドメインリンカー予測		
黒田研究室	学籍番号：16251506	松沢佑紀

[背景・目的] タンパク質にはマルチドメインタンパク質と呼ばれる複数の構造的に独立した構造ドメインからなるタンパク質が存在する。マルチドメインタンパク質は巨大なタンパク質であり、一般的に発現や結晶化が困難であることが知られているため、タンパク質を個々の構造ドメインに分割することで、より容易に解析を行うことが有効とされている。そのため、ドメイン境界であるリンカー部位をアミノ酸配列から特定する方法が活発に研究されてきた。

現在、タンパク質のドメインリンカーの同定ツールは複数のものが知られており、当研究室においても、サポートベクターマシン（以下 SVM）を用いたドメインリンカー予測器 DROP や H-DROP が開発されてきた。しかし、これらの予測機の予測精度はあまり高くなく、マルチドメインタンパク質に限定しても精度は 30%を切っている。そこで本研究では、SVM を用いたドメインリンカー予測器の性能向上とその評価を目的とする。

[手法] SCOPe、CATH の 2 つのドメインデータベースを利用して、マルチドメインタンパク質のデータを取得した。これらのマルチドメインタンパク質を代表化、目視での確認を行いタンパク質のデータセットを作製した。次に、これらのデータセットから 33 種の特徴量を作成した。それらから最も分類の精度が良くなる 8 種の特徴量を選択し予測器の構築に使用した。予測器の構築はデータセット中の各残基についてドメイン、リンカーのどちらに含まれるかを学習して行った。そして、GridSearch を用いて SVM での学習パラメータの最適化を行った。

最後に、構築した予測器を用いてテストを行った。テストは 300 マルチドメインタンパク質を用いて評価した。評価方法はテスト用に用意したタンパク質のすべての残基に対して、予測器によりリンカーかドメインか予測し、最も確率の高い 1 残基をリンカーと判定した。この残基が実際のリンカー部位の±5 残基以内に含まれていれば予測を成功とした。そして、予測器の精度を示す Precision、感度を示す Sensitivity、それらを総合して評価するための F score の 3 つの値を算出した。同様の方法で既存のドメインリンカー予測器でも予測を行い性能の比較を行った。

[結果および考察] データセットの作成では 2428 個のマルチドメインタンパク質のデータセットを作成することができた。これは先行研究の DROP で利用したデータセットと比較して約 14 倍である。

SVM の特徴量は PSSM を 2 種、露出溶媒表面積、共進化を 3 種、タンパク質における残基の位置、Ca 炭素の平面角 Theata の 8 種を選出した。そして、最適化したパラメータを用いて予測器を構築した。構築した予測器を用いてテストを行ったところ、Precision が 0.41、Sensitivity が 0.24 という値が得られた。（下図）これは Precision については他の予測器と比較し最も高い値であり、誤検出が他の予測器よりも少ないといえる。この予測精度の向上の要因の 1 つとして、共進化情報の追加が考えられる。共進化はタンパク質内での相互作用予測に広く用いられており、DROP では利用していない情報であった。共進化の追加によって、より高精度なリンカー予測が達成できたと考えられる。

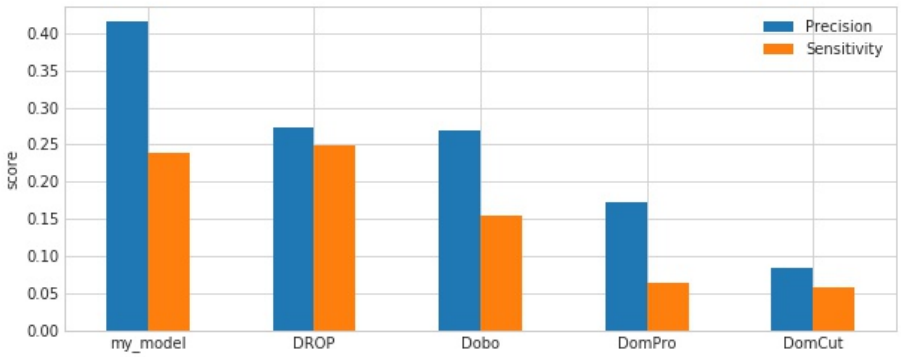

図 1. 構築した予測器の性能比較