SVM を用いたヘリカルリンカー領域の予測		
黒田研究室	学籍番号 : 09251029	鈴木 涼祐

【背景・目的】

発現精製が困難な構造未知タンパク質の新規ドメイン同定は、時間や労力の点から、実験的手法よりも計算的手法に利点がある。この手法はタンパク質のアミノ酸配列のみを用い、特に配列特徴が単純で検出が容易なドメイン境界を対象として発達してきた。しかし、ドメイン境界の多くがループ構造であり、ヘリックス構造をとるドメイン境界（ヘリカルリンカー）は、ループとは異なる特徴をもつため従来の方法で予測することは困難であった。そこで、本研究ではヘリカルリンカーに特化した学習データを構築し、機械学習法の一種である SVM (Support Vector Machine) を用いた予測機の開発を目的とした。

【方法】

1. ヘリカルリンカーデータセットの構築 : 独立して構造を取るドメインのデータセット IS-Dom からドメイン領域を取得し、ドメインを複数含む多ドメインタンパク質のデータセットを作成した。このデータセットから nrPDB を使って代表配列を選出した。次に代表選出したデータセットで、ドメイン境界にある残基と隣接したドメインとの間の相互作用数を計算した。相互作用は構造情報から、「水素結合」と「疎水性クラスタ」を算出した。相互作用数が閾値未満となるようにドメイン境界を前後に拡張し、最長となった領域をリンカー領域とした。そして、同定したリンカー領域の 70% 以上がヘリックス構造をとるものをヘリカルリンカーとした。
2. 特徴抽出 : 2840 個の特徴（二次構造、アミノ酸出現頻度、物理化学的特性など）から、Random Forest を用いてヘリカルリンカー予測に有効な特徴 41 個を抽出した。次に、この特徴群から Stepwise Selection で予測に最適な特徴組み合わせを決定した。最後に、この特徴を SVM の学習に利用し、予測機を構築した。
3. 予測機の評価 : 5-Fold Cross Validation Test にて予測機の評価を行い、予測効率を算出した。

【結果・考察】

構築したデータセットの、リンカー内とドメイン内のヘリックスのアミノ酸組成を比較した（図 1）。出現頻度の高いアミノ酸に注目すると、リンカー内では青色で示した親水性アミノ酸が多く、ドメイン内ヘリックスでは赤色で示した疎水性アミノ酸が多く見られることがわかった。これは、リンカー領域がドメイン内ヘリックスに比べて露出しているためと考えられる。

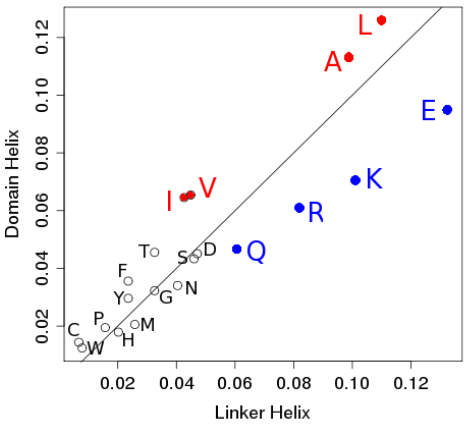

図 1. アミノ酸組成の比較

次に、予測機の性能を評価する。コントロールとして、タンパク質配列長の中心残基をリンカーと予測する手法 (Half Cut) 、ランダムで予測する手法 (Random Cut)、ドメイン境界予測機 (PPRODO)、ループリンカーに特化した予測機 (DROP) を用いた。SVM の出力値が最大の残基を中心に 1、11、21 残基の予測領域を設定し、これがデータセット内のリンカー領域と重複している場合を予測の正解とした。予測効率は、正解したリンカーの割合を意味する Sensitivity を用いた。表 1 に予測効率を示した。開発した予測機 (Stepwise 前、Stepwise 後) は、コントロールよりも高い予測精度を示した。また、Stepwise Selection により Sensitivity が最大で 12.9% 向上した。以上の結果は、ヘリカルリンカーに特化した予測機としての有効な性能を十分に示すものである。

表 1. 予測効率の比較

	Sensitivity		
	1 残基	11 残基	21 残基
Stepwise 前	0.357	0.457	0.571
Stepwise 後	0.486	0.586	0.629
Half Cut	0.257	0.300	0.357
Random Cut	0.120 ± 0.038	0.220±0.041	0.267±0.040
PPRODO	0.243	0.257	0.271
DROP	0.029	0.029	0.029