SVM 学習を用いたヘリカルリンカー予測機 H-DROP の高速化とその評価		
黒田研究室	学籍番号：10251010	尾園　早紀

【はじめに】

近年たんぱく質科学の分野においてたんぱく質の構造同定が進んでいるが、巨大たんぱく質はそのまま結晶化することが非常に困難である。そこでドメインという構造的・機能的な単位に分割して構造を解析するという手法が有効とされる。しかしながら実験的手法でドメイン境界領域(以下リンカーとする) を決定するためには多大な人員的・物的リソースが必要とされる。そのため、計算的にリンカーを予測する手法が開発されてきた。

本研究室の先行研究の成果として、機械学習法を用いたリンカー予測機 DROP と H-DROP があり、DROP はコイルリンカーの予測に、H-DROP はヘリカルリンカー(リンカー部分の 7 割以上がヘリックス構造をとっているもの、図 1) の予測に適している。既存の H-DROP は 1 配列の予測に 2~10 分かかり、インターネットツールとして速度改善が必要といえる。

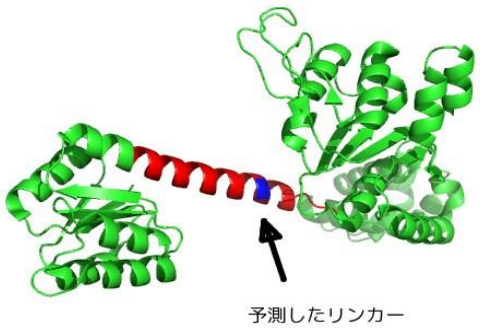
本研究では H-DROP の高速化を目的とし、改善に伴う予測精度への影響も調査した。

【方法】

H-DROP は SVM(サポートベクターマシン) という機械学習法を用いている。任意の配列を予測させ、残基単位でその残基がリンカーである可能性を評価する。各残基を様々な切り口でスコア化したものを特徴量(Feature) とし、これをベクターとして SVM の入力データとしている。SVM で教師あり学習をさせるため、学習データとして 255 配列から成るマルチドメインたんぱく質を選定しこれをデータセット(以下 DS-Helical とする) とした。この DS-Helical、もしくは DS-Helical の一部を SVM に学習させ任意の配列を評価させることでリンカーの予測が得られる。学習の際に用いる SVM パラメータ、Feature の組み合わせ(3000 種類用意し、その中から 30 個程度を選択して使う) を変えながら SVM 学習→配列の評価を繰り返すことで、予測精度が向上するよう最適な設定を追求した。最終的に決定された設定に合わせて H-DROP を構築し、現在 Web に公開されている H-DROP をアップデートした。

【結果と考察】

予測の際の各工程における所要時間を調査したところ、PSI-BLAST に最も時間がかかっていることがわかった。PSI-BLAST 以外の部分は合計で 2~4 秒、PSI-BLAST だけに数分かかっていた。PSI-BLAST は PSS と PSSM という、二次構造予測に基づいた特徴をスコア化した Feature を作成するために必要な工程である。そこで PSI-BLAST に変更を加えることで高速化出来ると考えた。PSI-BLAST で参照するデータベースをサイズダウンし、4 回検索を繰り返していた部分を 2 回に減らした。高速化の結果、1 配列の予測に既存の方法では 2~10 分かかっていた時間が 6~30 秒程度に短縮された。予測精度は(Sensitivity, Precision) = (33.3%, 36.4%) となり、先行研究で算出された (Sensitivity, Precision) = (35.2%, 38.8%) に多少劣るものの、速度向上と十分に引き換えられる結果であると考える(図 2)。ここで Sensitivity とはリンカーの全数に対する正解の割合、Precision は予測結果が出たもののうちの正解率である。



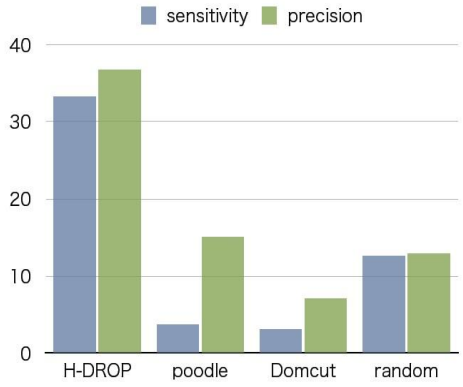


図 1. 1ny5_B というドメインのリンカーとされている部分(赤) と H-DROP で予測したリンカー(青)。

図 2. H-DROP と他の手法の場合の予測精度。poodle はタンパク質の disorder 領域予測プログラム、DomCut は教師用データからドメイン/リンカーにおける存在比を算出しその値により任意の配列の残基を評価する計算手法、random はタンパク質の両端から 40 残基ほど切り落としその内側でランダムに 1 残基をリンカーとして選びとった。