ヘリックスを含むドメイン境界配列の解析及びその予測		
黒田研究室	学籍番号： 09251506	辻 良太郎

[背景・目的]

生物学的に重要な機能を持つタンパク質は、いくつかのドメインにより構成される場合が多い。しかし、このような多ドメインタンパク質の中には、発現や精製が困難で解析が進んでいないものも存在する。このようなタンパク質の場合、単独で構造を取り、発現、精製が比較的容易な構造ドメインに分割して実験を行うという手法が取られる。この際に、ドメイン境界領域を計算的に予測する手法は、そもそも発現しないタンパク質の解析や解析の迅速化、実験コストの削減に貢献する。

先行研究では、ループ構造からなるドメイン境界領域予測手法が開発されており、既に高い効率が得られている。しかし、これらの予測手法では、既知のドメイン境界領域の約 30%を占めるヘリックスを含む境界領域の予測は困難である。そこで、本研究ではヘリックスを含むドメイン境界領域の予測に特化した予測機開発を目的とする。

[方法]

① 構造ドメインデータセットの作成： SCOP (Structural Classification of Proteins) によって定義された多ドメインタンパク質を取得し、そのドメイン間で「水素結合数 4 個未満、疎水性クラスター形成数 1 個未満、ジスルフィド結合数 1 個未満」ならば独立して構造を形成している構造ドメインであると定義して、それを選出した。

② リンカー領域の定義： 本研究では、リンカー領域を「前後のドメイン領域から独立しているドメイン境界領域」と定義した。①で選出された構造ドメインをもつタンパク質に対して、①と同様の定義で独立しているリンカー領域を決定した。さらに、取得したリンカー領域の中で、「70%以上の残基がヘリックスを形成するリンカー領域」をヘリカルリンカーと定義し、これを持つタンパク質をデータセットとして選出した。

③ 予測機の構築： 本研究では、統計的スコアを用いた手法と SVM による手法で予測機を開発した。統計的スコアは、各アミノ酸についてヘリカルリンカー領域と、ドメイン領域内ヘリックスの組成の差を用いて作成した。SVM ベクターのエンコーディングには、Binary Coding によるアミノ酸配列パターン情報と、上記の統計的スコアを用いた。

④ SVM 最適化： 学習に使用する Window 長や SVM パラメータについて最適化を行った。最適化の条件として AUC 値を用いて、5-Fold Cross Validation Test により検定した。また、予測法の Sensitivity と Precision を試験データのリンカー領域に前後 20 残基の誤差を認め、算出した。

[結果・考察]

本解析からヘリカルリンカーはドメイン内ヘリックスに比べて、Arg のような電荷を持ったアミノ酸を多く含むという傾向が明らかとなった (図 1) 。これはヘリカルリンカーが溶媒に露出した環境にあるためだと考えられる。さらに、これまで困難だったヘリカルリンカーの予測効率の向上が確認できた (表 1) 。また、SVM ベクターに統計的スコアを用いると有効であることが示唆された。

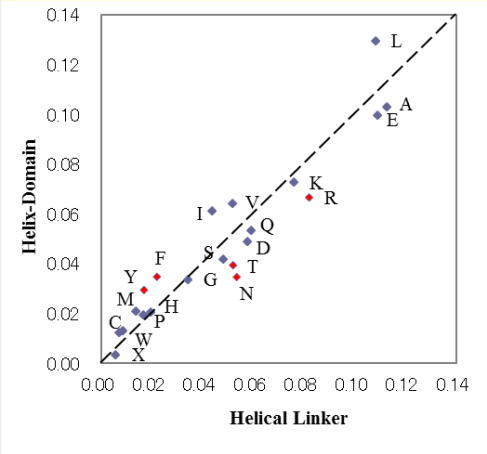

図 1. データセットのドメイン内ヘリックスとヘリカルリンカーのアミノ酸組成。赤点は χ^2 検定で有意差($P < 0.05$)が認められたアミノ酸を示している。

表 1. 各手法での Sensitivity と Precision および AUC 値

手法	Sensitivity	Precision	AUC
A	0.424	0.438	0.697
B	0.470	0.484	0.701
C	0.409	0.422	0.705
統計的スコア	0.409	0.422	0.621
DomCut	0.136	0.141	0.533
random	0.215	0.222	0.510

A、B、C は SVM による手法で、学習データが異なる。
A:Binary Coding、B:統計的スコア、C:A+B
DomCut は統計的手法でループ構造のドメイン境界領域を予測する方法である。
random は 1 タンパク質あたり 1 か所をランダムに予測領域として算出した。