高精度の disorder 領域予測法（POODLE）の開発


野口　保　　清水　佳奈
（タンパク質機能チーム 研究チーム長、　テクニカルスタッフ）

近年、立体構造を形成しないコード領域（disorder 領域：natively unfolded とも呼ばれています。図 1 参照）の中に、機能発現に関与する領域があることが実験的に明らかになり、このような領域は高等生物に特に多く見られる傾向があることがわかってきました。それらは、転写調節に関するタンパク質や DNA 結合タンパク質に多く存在することが示唆されています。また、disorder 領域は、X 線結晶解析や NMR によるタンパク質立体構造解析において、結晶化やスペクトルの帰属の妨げになるため、disorder 領域を予め取り除いて、解析を容易にすることも求められています。同様に、タンパク質立体構造予測においても、予測するべき配列を予め知ることにより、予測に費やす時間と労力を大幅に削減することができます。このような背景から、disorder 領域は注目を集めており、その予測手法の開発が近年活発に行われています。

我々は、disorder 領域の特徴がその長さによって異なることに着目し、タンパク質全体を対象にした disorder 予測と、長・短それぞれの disorder 領域予測に適した 3 種類の予測法（POODLE-W, L, S）を開発しました。ここでは、先行して開発を進め、既存の方法と比較して予測精度の向上が見られた POODLE-L（長い disorder 領域予測法）を紹介します。

既存の disorder 領域予測は、一般にアミノ酸配列やその配列の位置特異的なスコア行列 PSSM（Position Specific Scoring Matrix）と disorder 領域の関係を、統計的手法や機械学習などによって、処理（学習）して行っています。ここで紹介する POODLE-L は、アミノ酸残基特有の性質（10 項目）を、6 種類の物理化学的性質にグループ分けした指標を基に作成したスコア行列と、disorder 領域か否かの判別情報を Support Vector Machine(SVM) で学習することによって、disorder 領域予測を行います。この方法の特徴は、一般の予測法では、配列を window に区切り、その中心残基を予測するのに対して、window 全体を予測しながら window をずらす点と、その window サイズが 40 残基と広い点が挙げられます。さらに、window ごとに予測された結果は、配列上の各残基位置で集計され、その分布を基に 2 段目の SVM の学習を行い、各残基の disorder 予測を行います。6 種類の指標に関しては、どの組み合わせが最適か調べるために、全ての組み合わせで予測実験を行いました。その結果、6 種類全てを用いたモデルよりも上位の組み合わせが 9 モデルあり、その予測結果を統合することにより、予測精度を更に向上させることに成功しました。長い disorder 領域に関しては、公開されている既存の予測法の精度を超えることができました。

POODLE-L は、http://mbs.cbrc.jp/poodle（図 2 参照）で一般の方でも利用できるように、サービスを公開しています。POODLE-W, S についても、現在、公開を検討中です。

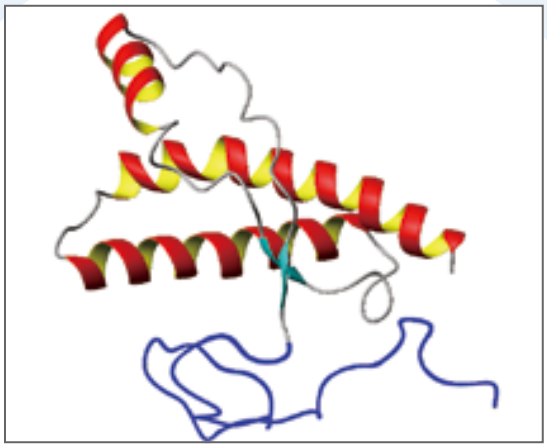

図 1：正常プリオンタンパク質 PrPC
N 末構造（青）が Disorder 領域のため、モデリングして表示

図 2：POODLE　(http://mbs.cbrc.jp/poodle)
正常プリオンタンパク質 PrPC の Disorder 予測結果。N 末に大きなピークが見られ、そこが Disorder 領域と予測されていることを示しています。