構造ドメインデータベースの作成及びSVM^{light}を用いた構造ドメイン領域の予測		
黒田研究室	02251002	荒井　隆雄

【背景】タンパク質には、複数のドメインからなるマルチドメインタンパク質が存在する。そのようなマルチドメインタンパク質において、ドメイン領域をアミノ酸配列から予測することができれば、NMR 構造解析や機能解析研究が容易に行うことができる。これまでのドメイン領域予測法は、Pfam、SMART、CDD、PROSITE といったドメインデータベースに登録されている配列との類似性を調べるものである。しかし、これらの方法での予測では登録されているタンパク質と類似な配列をもつタンパク質しか同定できない。すなわち、ドメインデータベースに登録されていない新規のドメイン領域を予測することはできない。そこで、新規の配列でもドメイン予測が可能なニューラルネットワークや SVM などの学習プログラムを用いて配列の特徴を学習させ、ドメイン領域を予測させていくというシステムの開発が求められている。しかしながら、ドメイン構造は多様で共通した配列の特徴を検出することは困難を極めるため、本研究ではドメイン境界の予測を試みる。

【方法及び目的】当研究室では、ドメインリンカー配列のデータベースを作成し、サポートヴェクターマシン（SVM^{light}）を用いて配列パターンを学習させ、新規の配列に対しドメイン領域を予測させることを最終目的とする。まず、データベースの作成に於いては、ドメインの独立性を判定するためドメイン間に水素結合や(接触)相互作用のない「構造ドメイン」とその構造ドメイン同士をつなぐ「ドメインリンカー」の二つを定義した。次に、SVM^{light}による学習に於いて、本研究で用いるコンピューターの性能の関係上、当研究室で現在までの研究により同定された 104 個のリンカー配列からなるデータベースを用いた。学習・予測効率では、ジャックナイフ法を用いて、SensitivityとSpecificityの二つの値で評価を行った。

【結果・考察】まず、データベースの作成では、700 個以上のドメインリンカー配列のサンプルが得られた。これは、従来の方法(目視)よりも 7 倍以上のサンプル数となる。

次に、SVM^{light}による学習効率の検証では、Window size及び学習させる際に設定する種々のパラメータを最適化した。Window sizeについては 19、SVM^{light}の各パラメータについてjでは 23、cでは 3、カーネル関数γでは 0.001 とすれば学習効率が最適化されることが示された。また、予測効率の検証では平均化のWindow及びリンカー定義値(cutoff value)をとることで、ランダムに予測した時よりもSensitivity、Specificityともに向上した(下図)。今後、Sensitivityに比べSpecificityが低いので、予測結果からリンカー候補の選び方をさらに検証する必要がある。