論　文　和　文　概　要

（２０００字程度）

報告番号	甲　第　　２９　　号	氏　名	安保　勲人

タンパク質は立体構造を形成し生体内で特異的に機能を発揮するため，タンパク質にとって立体構造形成は必要不可欠のイベントであると考えられてきた．しかし近年，この考えから逸脱したタンパク質である天然変性タンパク質の存在が明らかとなった．天然変性タンパク質は生理的条件下であっても，単独では立体構造を形成しない天然変性領域を保持し，さらに，天然変性領域中に相互作用パートナーと結合する数残基から数十残基の比較的短い領域を保持する．この機能部位を介した相互作用によって，天然変性タンパク質は転写調整やシグナル伝達などの生物学的プロセスにおいて重要な役割を果たしている．天然変性領域および機能部位はコンピュータを用いた予測プログラムによってアミノ酸配列より予測可能である．天然変性領域予測が実験を行う研究者が予測結果をある程度信頼し実験を進めることができるだけの実用精度に達する一方，機能部位の予測精度は充分ではない．実験的証拠が担保された機能部位データを提供するデータベースが存在するが，構造領域や天然変性領域のデータ数と比較して，機能部位データの数は圧倒的に少ない．このデータ数の不足が機能部位予測を難しくしている原因の１つとしてあげられる．

本研究では機能部位データを学習せずに，機能部位を予測するプログラム NeProc を開発した．機能部位は天然変性領域に存在する領域でありながら，構造領域的性質を持つことが知られている．つまり機能部位は天然変性領域中に存在する構造領域的性質を示す短い領域と見なすことができる．NeProc はこの機能部位の性質をターゲットとして予測を行う．そのため，データ数の不足している機能部位データの学習を必要とせず，構造領域と天然変性領域データのみを学習する．予測を実現するために，NeProc では比較的長い window を採用した Lmodel と，短い window を採用した Smodel の２つのモデルを組み合わせることで機能部位を予測する．Lmodel で天然変性領域を予測し，Smodel で予測された天然変性領域から，構造領域的性質を示す短い領域を予測する．Lmodel と Smodel は単層のニューラルネットワークとサポートベクターマシンのシンプルな機械学習法を複数回用いて予測を行う．特徴量には位置特異的スコア行列のみを採用し，クエリ配列の各アミノ酸を“機能部位”，“天然変性領域”，“構造領域”および“不明な領域”に分類する．

天然変性タンパク質データベース IDEAL が提供する実験的に検証された機能部位を用いて NeProc の予測精度を計測し，既存の機能部位予測プログラムの精度と比較した．その結果，NeProc は機能部位データを学習に用いていないにも関わらず，機能部位データを学習した既存の予測プログラムを上回る予測精度を達成した．特に，IDEAL に収録された機能部位の多くが分布する 10 残基から 50 残基の機能部位に対しての予測性精度が高かった．

IDEAL データでのテストでは，天然変性領域と注釈されたアミノ酸に対して，機能部位と予測をした場合，その予測は評価しなかった．これは，天然変性領域中に未知の機能部位が存在する可能性を，本研究では排除できないと判断したためである．そこで，ヒトタンパク質の天然変性領域中に予測法の性能評価で使用されるデータベースには収録されていない機能部位があるかを分析した．タンパク質データベース UniProt では実験的に検証されたタンパク質の機能領域の情報を提供している．そこで，複数の天然変性予測プログラムにより予測された天然変性領域中で，UniProt に記載されている機能領域を検索した．その結果，予測天然変性領域に多数の機能部位が存在していた．このことは，ヒトタンパク質の天然変性領域には現在でも，予測法の性能評価で使用されるデータベースには収録されていな

い天然変性領域中の機能部位が含まれていることを示唆している．また，抽出したデータを用いて機能部位予測のテストを行なったところ，NeProc は既存のプログラムを上回る予測精度を達成した．

NeProc は構造領域と天然変性領域データのみを学習することで，機能部位を予測することが可能である．そして本研究で実施した IDEAL データベースおよび未同定の機能部位データに対するテストにおいて，機能部位を学習した既存の予測プログラムを上回る予測精度を達成した．この結果は，機能部位予測を難しくしている原因の 1 つである“機能部位データの不足”を克服する可能性があることを示唆している．