

論文提出者	工学府博士後期課程 生命工学 専攻 平成20年度入学 学籍番号05831703 氏名 廣瀬 修一 印		
主指導教員 氏 名	黒田 裕	副指導教員 氏 名	朝倉 哲郎
論文題目	機械学習法を用いたタンパク質中の動的情報の予測 (Prediction of Disordered and Flexible Regions in Proteins Using Machine Learning Methods)		
論文要旨 (2000 字程度)			
<p>タンパク質は柔軟性を持つ分子であり、多くの場合、タンパク質は柔軟性を持つことによって、他のタンパク質や低分子との相互作用が可能となる。近年、生理条件化で三次元構造が欠如している領域の存在が多数報告されている。このような現象は一般的にディスオーダー領域 (disordered region) と呼ばれ、ラマチャンドラン角が常に変化し、一定の値をとっていない。一方、立体構造を保持している領域にも動きは存在する。これはフレキシブル領域 (flexible region) と呼ばれ、動きが引き起こされる要因から理論的に2種類の運動に分類できる。即ち、その領域自体が変形を起こしている内部運動と、隣接するアミノ酸残基の変形によって剛体として回転や並進を起こしている外部運動である。X線結晶解析やNMR等の実験手法は、タンパク質の動きに関する詳細な情報を与えるが、多くの時間と労力が必要であるため、現状では大規模な解析は不可能である。同様に、計算的手法の一種である分子シミュレーションは、動きを詳細に解析することが可能であるが、CPU負荷が高く、また構造情報が入力として必要であるため、その適用範囲が限られている。ゲノム解析の進展により膨大な数の配列情報を利用することができることから、配列情報からタンパク質の動的情報を得る計算機的手法の開発は重要であると考えられ、これによって、タンパク質の動きに関する大規模な解析が可能になることが期待される。</p> <p>博士論文では、アミノ酸配列情報から2種類の動的情報 (ディスオーダー領域および内部運動と外部運動) を予測する手法を構築し、配列特徴と動きとの関連に重要な知見を得た。さらに、開発した予測法の幾つかの応用事例について報告した。</p> <p>本論文は5章から構成されている。</p> <p>第一章「緒論」では、タンパク質の動きと機能との関連を説明し、タンパク質の動きが重要な性質であることを述べた。また、本研究で対象としたタンパク質の動きであるディスオーダー領域および内部運動と外部運動の定義を行った。さらに、タンパク質の動きに関連する計算機的手法の研究を概説し、現在までに開発されているデータベースや予測手法を紹介した。次に、予測手法において一般的に用いられる機械学習法について説明を行った。最後に、アミノ酸配列情報から動的情報の予測を行うことの意義について述べ、各章の意義と目的について述べた。</p>			

第二章「ディスオーダー領域予測法の開発」では、アミノ酸配列情報のみから長いディスオーダー領域を予測する手法を開発した。まず、10種類の単純なアミノ酸の物理化学的特徴を記述子として定義した。これらを利用して40残基のセグメントを数値表現し、二層のサポートベクターマシーン (SVM) からなる予測手法を構築した。次に、記述子の組み合わせを検討することにより、9種類において予測精度の向上が見られた。最後に、これらの予測器の出力を統合する手法を構築し、これを POODLE-L と名づけた。POODLE-L は、既存の手法よりも高い予測精度を示した。POODLE-L に組み込まれている10個の予測器に含まれる記述子の調査から、疎水性が長いディスオーダー領域を判別するのに重要な要因であることが示唆された。

第三章「フレキシブル領域予測手法の開発」では、アミノ酸配列情報のみから内部運動と外部運動の大きさを予測する手法を開発した。4種類のアミノ酸の物理化学的特長量 (アミノ酸の性質、運動の大きさ指標、予測された二次構造、予測された溶媒露出面積) を利用して隣接するアミノ酸ペア情報を記述子として定義し、アミノ酸配列を数値表現した。さらに、各残基が位置する二次構造に依存して運動の大きさが変化するという仮定に基づき、ランダムフォレストを用いて、複数の予測モデルを作成した。構築された手法は、二次構造中では動きが小さく、それ以外の領域では動きが大きいと仮定したナイーブモデルと比較して、高い予測精度を示した。予測モデルに含まれる変数の数と予測精度の関係から、二次構造と溶媒露出面積の情報が、内部運動および外部運動の予測と強く関連していることが示唆された。

第四章「予測プログラムの応用」では、本研究で開発された予測手法の3つの応用事例、(1) サーバ構築による予測手法の公開、(2) 我々の研究グループが開発した POODLE-L を含むディスオーダー領域予測プログラムである POODLE シリーズの統合によるディスオーダー領域の長さに依存しない予測システムである POODLE-I の開発、(3) ヒトゲノム情報を利用した GPCR ペプチドリガンド探索、について報告した。

第五章「結言」では、得られた成果を要約し、本研究の総括を行った。