氏　　名	朴（パク）　根（クン）　準（ジュン）
学位の種類	博　士（理　学）
学位記番号	理 博 第 2711 号
学位授与の日付	平成15年9月24日
学位授与の要件	学位規則第4条第1項該当
研究科・専攻	理学研究科生物科学専攻
学位論文題目	配列解析によるタンパク質の細胞内局在部位予測
論文調査委員	（主査）教授 金久　實　教授 宮田　隆　教授 藤吉好則

論　文　内　容　の　要　旨

タンパク質はリボソームで合成されたあと，細胞内の特定の部位に輸送されて，その機能を発揮する。アミノ酸配列からタンパク質の細胞内局在部位を予測する研究は，これまでにも数多く報告されている。本論文で申請者は，真核細胞の細胞内局在部位を12種類に分類し，それぞれ既知のアミノ酸配列データを集積し，サポートベクターマシン（SVM）と呼ばれる計算機科学の新しい方法に基づく予測法を開発した。

データセットの作成部分では，まず葉緑体，細胞質，細胞骨格，小胞体，細胞外（分泌），ゴルジ体，リソソーム，ミトコンドリア，核，ペルオキシソーム，細胞膜，液胞の12箇所の局在部位について，SWISS-PROTデータベースから，そのアノテーションを利用して既知のアミノ酸配列を抽出した。そこから配列類似性の高いものを除去して，全部で7589の配列エントリーを得た。このデータセットには，酵母，ヒト，マウス，ラット，線虫をはじめ，多数の生物種のデータが含まれている。

サポートベクターマシンとは，サンプル集合を2つのクラスに分類するパターン認識手法である。その際，従来の判別分析のように2つのクラスのデータ（分布関数）すべてを利用するのではなく，識別が難しいデータに着目して識別平面を求める。また，サポートベクターマシンはカーネル法と呼ばれる方法の1つで，線形分離可能でないサンプルデータでも高次元空間（特徴空間）へ写像して，線形サポートベクターマシンを適用することができる。写像そのものを計算する必要はなく，写像された2つのデータ間の距離（類似度）に相当する内積が計算できればよく，これを与えるのがカーネルである。カーネルにはいくつか標準的なものがよく利用されており，ここでは線形カーネル，多項式カーネル，RBFカーネルのどれを使うかの選択をまず行った。これまでの研究で，アミノ酸組成がタンパク質の局在部位を識別するのに有効であることが知られているので，葉緑体とそれ以外といった識別を12箇所すべての局在部位について行い，結果としてRBFカーネルを使うこと，またデータセットのサイズが大きい部位と小さい部位でパラメータを変化させるやり方（RBF mixture）が最も高い予測率を与えることが分かった。

次にタンパク質を表現する特徴量として，アミノ酸組成だけでなく，隣接したアミノ酸ペアと1，2，3個のギャップをはさんだアミノ酸ペアを考慮した。これら5種類（アミノ酸組成と4種類のアミノ酸ペアの組成）の特徴量それぞれに対し，RBF mixtureを用いたサポートベクターマシンを作成し，さらに与えられたアミノ酸配列から局在部位を予測する際に，5つのサポートベクターマシンの予測結果を総合的に判断する voting scheme を考案した。正答のタンパク質の割合は78%，正答の局在部位の割合は58%という予測率が得られ，これまでの研究の中でも最も高い予測率であった。とくに従来の方法ではデータサイズの大きな部位に最適化する傾向があり，一見全体の予測率はよく見えるが，データサイズの小さな部位の予測率が極端に悪くなっていたが，ここではその点が大きく改良されている。なお，開発したプログラムはPLOCと名付けた予測システムとして，ゲノムネットでサービスを行っている。

論文審査の結果の要旨

ゲノムにコードされたタンパク質の機能を推定するためにホモロジー検索が広く利用されているが，データベース中に類似配列が存在しないか，類似配列はあってもその機能が不明の場合も多く，ホモロジー検索には限界がある。その際，タンパク質の細胞内局在部位に関する手掛かりが得られれば，ある程度の機能推定が可能となる。アミノ酸配列から局在部位を予測することは，バイオインフォマティクスの課題の１つとして，これまでに多くの研究がなされてきた。当初は局在化シグナルといったアミノ酸配列の特徴パターンを利用して，例えば Nakai & Kanehisa による PSORT システムなどが作られた。その後，タンパク質のアミノ酸組成が局在部位の識別に有効であることが示され，生物的な知見は別にして，予測率という実用面の観点から，アミノ酸配列を組成のような特徴量のベクトルに変換するアプローチが主流となっている。また予測法としても，古典的な判別分析から始まり，ニューラルネットワーク，隠れマルコフモデル，サポートベクターマシンといった計算機科学の最先端の方法が用いられている。

本論文では，１つのアミノ酸配列をアミノ酸組成（20次元のベクトル），隣接したアミノ酸ペアおよび１，２，３個のギャップを含むアミノ酸ペアの組成（それぞれ400次元のベクトル）で表現し，サポートベクターマシンの方法を用いて，局在部位予測を行ったものである。本論文独自のアイディアはギャップを含むアミノ酸のペア（特定アミノ酸の周期性を反映すると考えられる）を考慮したことだけであるが，総合的に様々な解析を系統的に行っている点は評価に値する。すなわち，多く（12種類）の局在部位に対するデータセットを作成したこと，サポートベクターマシンのカーネルとパラメータを系統的に調べたこと，５種類の特徴ベクトルから得られた結果を統合する voting scheme を考察したことなどである。また，以前に核移行シグナル配列パターンを隠れマルコフモデルで学習することで核局在化予測を試み，その反省からアミノ酸またはアミノ酸ペア組成ベクトルをサポートベクターマシンで学習する方向に転換しており，局在部位予測の分野全体に対する幅広い知識と経験が生かされた論文となっている。

本論文の方法は，全体の予測精度（total accuracy）では従来の最もよい方法と同レベルであった。一方，12箇所の場所ごとの予測精度をデータサイズの重みをつけずに単純平均をとった値（location accuracy）では，従来のものよりも大幅に改善されていた。つまりデータサイズの小さな部位でも比較的精度がよくなっており，これはデータサイズの違いを RBF カーネルのパラメータに反映させたこと，および voting scheme を用いたことの結果であると考えられる。開発した予測法は PLOC システムとして実用化しており，12箇所の局在部位は，実際の予測システムでは10箇所をもつ動物細胞，11箇所をもつ植物細胞，10箇所をもつ酵母細胞と区別して行うようになっている。

以上申請者が行ったタンパク質局在部位予測法の開発は，基礎的な研究成果としても，ゲノム解析の実用的なツールとしても高く評価できる。よって，本論文は博士（理学）の学位論文として価値あるものと認める。なお，論文内容とそれに関連した試問の結果，合格と認めた。