

(続紙 1)

京都大学	博士（情報学）	氏名	John Ellsworth Brown III
論文題目	Kernel Methods in Biochemical Informatics and Applications to DNA Repair Research		
（論文内容の要旨） 本論文はタンパク質配列データからDNA修復タンパク質の認識と分類のため予測方式、および、化合物の構造データからの活性の予測方式について述べられており、6章から構成されている。 第1章では本論文で対象とする問題の生物学的化学的背景について述べている。 第2章では本論文で必要となる生物学および化学に関する基礎的事項や、本論文で利用する統計的パターン認識手法、カーネル法、および、サポートベクターマシンなどのカーネル法に基づく分類予測法の概要について説明している。 第3章ではタンパク質配列データからのDNA修復タンパク質の認識とその分類のための予測法を示し、さらに、その解析法による実際のゲノム配列の解析結果について議論している。まず、配列データからの修復タンパク質の予測のために類似配列検索に基づく手法や既存の配列データに対するカーネル関数について説明するとともに、いくつかの新たなカーネル関数を提案している。特にタンパク質の二次構造の予測結果や類似配列の検索結果とスペクトラムカーネルを組み合わせたカーネル関数を提案している。次に、それらのカーネル関数の有効性を示すために、サポートベクターマシンとの組み合わせにより実際のタンパク質配列データを用いた計算機実験を行っている。その結果として、利用可能な類似配列が少ない場合などにおいては提案手法がより高い予測精度を持つことを示している。さらに、提案した予測法および類似配列検索による予測法を31種類の生物種のゲノム配列解析に適用し、その結果としてゲノムサイズの増加に従いDNA修復タンパク質が増加傾向にあるなどの知見を得るとともに、未知のDNA修復タンパク質候補配列を見出している。また、提案した予測法を実装したWeb Serverについても紹介している。 第4章では化合物の構造データから活性を予測するためのカーネル関数を提案している。化合物の構造データは通常無向グラフとして表現されるが、既存のグラフカーネル関数では立体異性体を原理的に区別できなかった。そこで、立体異性体を区別可能なカーネル関数を既存の木パターンカーネルを拡張することにより提案している。具体的には既存のカーネルで同一視されていた部分構造を適切に区別することにより立体異性体を区別できるようにしている。提案したカーネル関数の有効性を評価するためにサポートベクターマシンおよびサポートベクター回帰と組み合わせ、ステロイドやビタミンD誘導体などの化合物に関する実際の活性値データを用いた計算機実験を行っている。その結果として、立体異性体が多く含まれるデータに関して、特にそのようなデータが多く含まれるビタミンD誘導体の転写調節能や分化誘導能といった活性値の予測に関しては、提案手法が既存手法と比較し高い予測精度を持つことが示されている。 第5章では化合物の物理化学的な特性を考慮したカーネル関数を提案している。まず、原子の部分電荷に関する情報をグラフのマッチングと組み合わせたカーネル関数を提案し、次に、そのカーネル関数に化合物のサイズや剛性などの影響を取り入れたカーネル関数の検討を行い、さらに、提案したカーネルの正定値性などについて議論している。そして、ここで提案したカーネル関数に対してもサポートベクター回帰と組み合わせ、ステロイドやビタミンD誘導体などの化合物の活性値データを用いた計算機実験を行っている。その結果として、ステロイドに関するデータについては第4章で示したカーネル関数も含め既存のカーネル関数より高い予測精度が得られることを示している。 第6章は結論であり、本研究のまとめと今後の課題について述べている。			

(論文審査の結果の要旨)

本論文は、機械学習手法の一つであるカーネル法に基づくタンパク質配列データからのDNA修復タンパク質の認識分類のための予測方式と化合物の構造データからの活性の予測方式、および、それらを用いた情報解析結果について述べたもので、得られた成果は以下のとおりである。

(1) 細胞内でDNA修復タンパク質が果たす役割が大きい、詳細な機能が不明なDNA修復タンパク質も数多く、また、多くの生物種においてはDNA修復タンパク質がどの程度あるのかもよくわかっていない。そこで、本論文ではタンパク質配列データからDNA修復タンパク質であるかないかを予測し、さらに、修復タンパク質であると予測された場合にはより詳細な機能クラスへの分類を行うためのカーネル関数を開発した。そのために、配列データに対する既存のカーネルであるスペクトラムカーネルに、配列の二次構造に関する予測結果や類似配列の検索結果を組み合わせたカーネル関数を設計した。そして、既知のDNA修復タンパク質配列データを用いた計算機実験により、既存のカーネルや類似配列に基づく検索手法との比較を行い、提案したカーネル関数の有用性を示した。さらに、既存手法や提案したカーネル関数を31種類の生物種のゲノムデータに対して適用し、DNA修復タンパク質の分布を調べ、その結果として、ゲノムサイズの増加とともにDNA修復タンパク質も増加する傾向があるなどの知見を得た。

(2) 化合物の構造データからの活性予測は薬剤設計などに応用を持つ重要な問題であり、これまでも数多くの研究が行われてきた。近年ではカーネル法に基づく手法も数多く提案されるようになり、その有効性が認識されつつある。しかしながら、カーネル法に基づく既存研究では化合物をグラフ構造としてのみ表現しており、化合物において重要な立体異性に関する情報が考慮されていなかった。本論文では既存の木パターンカーネルを拡張することにより、立体異性体を区別できる新たなカーネル関数を開発した。そして、活性値が既知のステロイドに関するデータやビタミンD誘導体に関するデータを用いて計算機実験を行い、立体異性体が多く含まれるデータに対する提案手法の有効性を示した。

(3) 化合物の活性の予測においては原子や部分構造の物理化学的性質を考慮することも重要である。本論文では原子の部分電荷とグラフのマッチングを組み合わせた新たなカーネル関数を開発し、さらに、そのカーネルに化合物のサイズや剛性に関する情報も取り入れたカーネル関数も提案した。開発したカーネル関数についても活性値が既知の化合物データを用いた計算機実験を行い、特にステロイドデータの活性値の予測において提案手法が有効であることを示した。

以上、本論文ではバイオインフォマティクスおよびケモインフォマティクスにおける重要な研究テーマである、配列データからのタンパク質機能予測および化学構造データからの活性予測の両者に関して、新規な手法を提案するとともに、実際のタンパク質データおよび化合物データを用いた計算機実験により、それらの有効性を示している。提案手法のいずれもが新規性、有用性、拡張性の高いものであり、当該分野の発展のために十分な寄与をしている。よって、本論文は博士(情報学)の学位論文として価値あるものと認める。

また、平成22年3月1日論文内容とそれに関連した口頭試問を行った結果合格と認めた。

