（ 続紙 １ ）

京都大学	博士（情報学）	氏名	WANG Feiqi（王　菲琪）
論文題目	Design of Computational Models for Analyzing Graph-Structured Biological Data （グラフ構造を持つ生物情報データに対する計算モデルのデザイン）		

（論文内容の要旨）

本論文は、グラフ構造を持つ生物情報データ、具体的にはRNA二次構造データとキナーゼ阻害剤の化学構造データに対する計算モデルおよび計算手法の設計と評価について述べられており、５章から構成されている。

第１章では、研究の背景と動機、分野横断型研究の重要性、成果の概要、論文の構成について述べている。

第２章では、まず、本論文で利用するデータ、具体的には、文字列、根つき木、平面グラフ、および、それらを表現するためのデータ構造について説明するとともに、木構造間、および、グラフ構造間の編集距離という概念について説明している。次に、本論文の後半で利用するニューラルネットワークに関して、グラフ畳み込み、プーリングを中心に説明している。

第３章では、RNA二次構造のうち、擬似ノットと呼ばれる複雑な部分構造に対して比較を行うための計算モデルと手法を提案し、実際のRNA二次構造データを用いた計算機実験により有効性を評価している。提案モデルでは、まず、外平面グラフとして表現された擬似ノットつきRNA二次構造に対してPeelingという既存手法を用いて中心を計算し、それをもとに根つき木に変換する。そして、２個のRNA二次構造から変換された２個の根つき木に対し既存の編集距離計算手法を適用することにより、擬似ノット構造の比較を行う。提案モデルの評価については、実際のRNA分子の擬似ノット構造を集めたデータベースのデータを用いて、RNA配列データの編集距離を用いた場合との比較を行っている。具体的には、データのペアすべてについて距離を計算し、それをもとに階層型クラスタリングを行い、その結果をGene Ontology（遺伝子の機能分類）と照合することにより定性的な比較を行っている。さらに類似の手順により、配列の編集距離のみを用いた場合、提案モデルによる距離のみを用いた場合、配列の編集距離と提案モデルによる距離を併用した場合の比較も行っている。

第４章では、キナーゼタンパク質の阻害剤となる化学構造（グラフ構造）からタンパク質の結合部位（原子）を予測するための新規モデルを提案している。近年、グラフの同型性判定のためのヒューリスティックな計算手法であるWeisfeiler-Lehman(WL)アルゴリズムに基づくグラフ畳み込みニューラルネットワークが盛んに研究・応用されているが、その概念に基づき、WL Boxという名の計算モデルを新たに提案している。このモデルの特長は、辺の情報に基づき頂点に割り当てられた特徴量の畳み込みを行う従来手法とは異なり、switch weightという重みを用いて、より柔軟な畳み込みを行う点にある。さらに、このWL Boxによる特徴量の畳み込みとグラフの行列表現に対する従来の二次元畳み込みを個別に行い、それにPyramid Spatial Poolingという既存の統合モデルを適用することにより学習・予測を行う階層型ニューラルネットワークであるPISPKI(Prediction of Interaction Sites of Protein Kinase Inhibitors)モデルを構築している。そして、実際の阻害剤の化学構造データを用いて、従来の畳み込みニューラルネットワーク、および、サポートベクターマシンという既存の機械学習手法との計算機実験による比較を行っている。さらに、結合部位をシャッフルしたデータを用いた計算機実験、および、モデルの構成要素の一部を取り除いた手法を比較するAblation Studyにより、提案モデルが適切に学習を行えているかを評価している。

第５章は結論であり、本研究をまとめるとともに、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、擬似ノットつきRNA二次構造の比較のためのグラフ変換と木構造編集距離に基づく計算モデル、および、キナーゼタンパク質阻害剤の化学構造（グラフ構造）データからの結合部位（原子）予測のためのニューラルネットワーク・モデルについて述べたものであり、得られた成果は以下のとおりである。

（１）RNA（リボ核酸）分子は塩基が並んだ文字列構造をしているが、生体内では一部の塩基対が結合することにより安定した構造をとる。結合塩基対の集合はグラフ構造として表現され、RNA二次構造とよばれる。RNA二次構造はその機能と密接に関連するため多くの構造比較計算モデルが提案されてきたが、擬似ノットという複雑な部分二次構造については十分に研究されていなかった。そこで、グラフの木構造への変換と木構造に対する編集距離という従来手法を巧妙に組み合わせることにより新規な計算モデルを開発した。具体的には、Peelingという操作に基づき外平面グラフの中心を一意に計算する手法を二次構造に適用することにより木構造に変換し、その木構造に対して木編集距離という（非）類似性の尺度モデルを適用することにより、擬似ノットつき二次構造間の類似度を評価する計算モデルを開発した。実際のRNA二次構造データを用いてクラスタリングを通じた計算機実験を行った結果、文字列データ間の編集距離を用いた場合として比較し、より適切な結果が得られることを示した。さらに、文字列間編集距離のみの場合、提案モデル、文字列間編集距離と提案モデルの組み合わせについても比較実験を行い、組み合わせた場合が最も適切なクラスタリング結果が得られたことを示した。

（２）キナーゼタンパク質は酵素として働くタンパク質であり、細胞の分化やアポトーシスを始めとする様々な機構に関わる重要なタンパク質である。そして、その阻害剤となる化合物に対して、タンパク質との結合部位（結合原子）を予測することは阻害剤の機能解析において重要であるが、有効な計算手法が開発されていなかった。そこで、化学構造（グラフ構造）が与えられた際に結合部位を予測するために、グラフニューラルネットワーク(GNN)に基づく新規な計算モデルを開発した。具体的には、従来のGNNで用いられていたWeisfeiler-Lehmanアルゴリズムに基づきつつも、複数の重みを切り替えて利用することによりグラフの特徴量を効果的に計算するWL Boxというモジュールを新規に提案し、それを従来の２次元畳み込みネットワークとPyramid Spatial Pooling層を用いて統合したPISPKIモデルを設計した。そして実際の阻害剤データを用いて、従来の畳み込みニューラルネットワーク、および、サポートベクターマシンという既存の機械学習手法との計算機実験による比較を行い、提案モデルがより高い予測精度を持つことを示した。さらに、結合部位シャッフルによる計算機実験により適切に学習が行えていること、および、Ablation StudyによりWL Boxが有効に機能していることを示した。

以上、本論文ではグラフ構造データに対する計算モデルという情報学における重要な研究課題に取り組み、擬似ノットつきRNA二次構造比較、および、キナーゼタンパク質の阻害剤となる化学グラフ構造における結合部位予測のための新規モデル・手法を提案し、それぞれの手法を実際の生物学データなど用いた計算機実験により評価した。提案手法のいずれもが新規性、有用性が高く、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、令和４年２月１７日、論文内容とそれに関連した事項につ

いて試問を行った結果、合格と認めた。なお、本論文のインターネットでの全文公表についても支障がないことを確認した。

要旨公開可能日：　　年　　月　　日以降