（ 続紙 １ ）

京都大学	博士（ 情報学 ）	氏名	森 智弥
論文題目	Methods for Analyzing Tree-Structured Data and their Applications to Computational Biology （木構造データの解析手法とその計算生物学への応用）		

（論文内容の要旨）

本論文は、データ工学において重要な根付き無順序木データの比較問題と類似部分構造検索問題に対する効率的な計算手法、および、そのバイオインフォマティクスへの応用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、結果の概要、および、論文の構成について述べている。特に、木構造データが情報学の様々な分野で利用されていること、木構造データ比較の重要性、生物学データへの応用などについて述べている。

第２章では、木構造データ、特に根付き無順序木の比較手法について詳細なサーベイを行っている。まず、木構造データの比較に広く用いられている編集距離の定義について説明し、順序木と無順序木の場合に距離が大きく異なる場合があること、および、順序木の編集距離は多項式時間で計算可能であるが無順序木の編集距離計算はNP困難であることを述べている。次にその困難性に対処するために開発されてきた様々な既存手法について説明している。さらに、制約がついた場合の編集距離、近似アルゴリズム、木構造のアラインメント、および、tree inclusionについて説明している。

第３章では、無順序木の編集距離を計算するための最大重みクリークと動的計画法を組み合わせた計算手法を提案している。無順序木の編集距離計算のための最大重みクリークに基づく既存手法について説明した後、本論文で提案する動的計画法との組み合わせ、および、計算時間削減のための一連のヒューリスティクス操作について説明している。次に、提案手法の有効性を示すために、ベンチマークとして広く利用されているWEBデータであるCSLOGSデータ、および、計算生物学関連データベースKEGGに格納されている糖鎖構造データを用いた計算機実験により提案手法と既存の最大重みクリークに基づく手法を比較し、データサイズが小さい場合を除いて、提案手法の方が圧倒的に高速であることを示している。

第４章では、木構造データベースの検索などのために提案されたtree inclusionを実用的なものにするために、無順序木に対するtree inclusionの拡張を提案し、その計算アルゴリズムを示している。具体的には、tree inclusionでは頂点の挿入操作のみによりパターン木がテキスト木に変換可能かどうか判定するのみであったため、置換操作を許し、かつ、テキスト木の部分構造とパターン木の間にスコアを導入するという拡張について説明している。そして、拡張したtree inclusionのスコアを計算するための動的計画法に基づくアルゴリズムを示し、その計算量を解析している。さらに、テキスト木内にある各頂点の子の順序を誘導部分木のサイズによりソートすることで領域計算量を減らす手法について説明している。また、より柔軟なマッチングを可能にするため、少数の削除操作も許した場合のアルゴリズムも示し、その計算量を解析している。そして、提案手法の有効性を示すために行った様々な計算機実験とその結果について説明している。具体的には、シミュレーション・データを用いた計算機実験によりパターン木の最大出次数を定数で抑えた場合の計算時間がパターン木およびテキスト木のサイズにほぼ線形に比例すること、WEBのログデータを用いた計算機実験により実データに対しても高速に動作すること、DBLPなどの文献データベース・データを用いた計算機実験により既存の順序木編集距離計算ツールを用いた検索より優れた精度で部分類似構造検索を行えること、KEGGデータベース・データを用いた計算機実験により糖鎖構造検索も高速に実行可能であることなどを示している。

第５章は結論であり、本研究をまとめるとともに、今後の研究の方向性や課題について述べている。

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し、合わせて、３，０００字を標準とすること。
論文内容の要旨を英語で記入する場合は、４００～１，１００wordsで作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること。

（続紙 ２ ）

（論文審査の結果の要旨）

本論文は、データ工学およびバイオインフォマティクスにおいて広く利用される根付き無順序木データの構造類似度の計算手法と類似部分構造の検索手法、および、その応用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）無順序木の編集距離計算はNP困難であり、既存手法として編集距離計算問題を最大重みクリーク問題に変換し、それに既存の最大重みクリーク・アルゴリズムを適用するという手法が知られていた。本論文では大幅な高速化を図るために、この手法と動的計画法を組み合わせ、さらにいくつものヒューリスティクスを組み込んだ手法を提案し、ヒューリスティクスにより解の最適性が損なわれないことを証明した。そして、WEBログデータ、糖鎖構造データを用いた計算機実験により、データサイズがある程度以上大きくなると、提案手法が最大重みクリークに基づく既存手法より、はるかに高速に動作することを示した。特に糖鎖構造データについては、既存手法では計算時間がかかりすぎて全データペアに対する編集距離計算を行うことができなかったが、提案手法ではそれを可能にするという結果を得た。

（２）既存研究において木構造データベースの検索のためにtree inclusionという概念が提案され、その計算が無順序木に対してはNP困難であるが、パターン木の最大出次数がD以下である場合には、O(2^{2D}mn)時間で可能であることが示されていた（m,nはパターン木、テキスト木の頂点数を表す）。しかしながら、tree inclusion は頂点の挿入操作のみに基づく判定問題であるため、柔軟な検索を行うことができなかった。そこで、頂点の置換操作も許し、かつ、スコアも導入するという拡張を提案し、それに対するO(2^{2D}mn)時間アルゴリズムを開発した。そして、シミュレーションデータ、WEBログデータ、糖鎖構造データを用いた計算機実験により、木の最大出次数が７程度以下の場合であれば、高速にスコアを計算できることを示すとともに、文献データを用いた計算機実験により文献検索における有用性を示した。

（３）上記（２）で提案したtree inclusionの拡張に対するアルゴリムの改良および拡張を提案した。一つは、スコアを計算するだけで良い場合には、データの前処理を工夫することにより領域計算量をO(2^{D}mn)からO(n+2^{D}m log n)に削減できることを示し、もう一つは、上記アルゴリズムを少数の頂点の削除を許した場合に拡張し、最大出次数および削除頂点の個数の両者が定数で抑えられている場合にはO(mn)時間で計算できることを示した。さらに、実データを用いた計算機実験により、これらのアルゴリズムの妥当性を示した。

以上、本論文ではデータ工学およびバイオインフォマティクスにおける重要な研究課題である無順序木データの類似度計算および類似部分構造の検索に関して、効率的な計算手法、および、新たな定式化を提案するとともに、シミュレーションデータおよび実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２７年８月２７日に実施した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。

注）論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　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。」）を付け加えること。

Webでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。
要旨公開可能日：　　年　　月　　日以降