

## 審査の結果の要旨

氏名 モハammed サヒリ

本論文は、生物のDNA配列の断片データ（DNAシーケンスデータ）から、ゲノム配列を再構築するゲノムアセンブリ問題の解決を試みている。現在の計測技術では一度に計測できるDNA断片長が全ゲノムに比して極端に短いことから、複雑な構造を有する生物種のゲノム配列の決定には様々な困難が存在する。そのためゲノムアセンブリ問題は、データの前処理を含む多種の部分問題からなり、最終的に精度を向上させるためには、それぞれの問題に対して深い洞察を基に適切に対処し、システム全体を構築する必要がある。本論文では、シーケンスデータの前処理と、アセンブリを行う後処理とを明確に分け、それぞれの問題に対してアプローチを考えることで計測技術の差異に過度に依存しない、汎用的かつ精度の高いアセンブリシステムの開発を試みている。計測データ前処理に対しては、シーケンスデータに含まれる計測エラーおよびクロニングベクターとよばれる他種の生物由来の配列を、それぞれ検出および除去する新規アルゴリズム群を考案し、既存手法を上回る精度を示している。後処理であるアセンブリにおいては、シーケンスデータを基に構成されるグラフ構造から、配列情報の曖昧性を表現する部分構造の解消に主に着目し、解消段階において考慮すべき新規部分構造群を提案し、それぞれの解消アルゴリズムを考案している。またグラフ構造からDNA配列を推定する段階においても、従前使われてきたグラフ理論的に依拠した手法ではなく、より計測データの信頼性に依拠した新規アルゴリズムを考案し、システム全体として既存手法と比べて高い精度でゲノムアセンブリを行うことに成功している。

本論文は七章からなり、第一章では、ゲノムアセンブリ問題の背景および動機を明らかにし、上記問題に対する提案手法群の貢献の概要を述べている。第二章では、ゲノムアセンブリ問題を解くための元データである、DNAシーケンスデータを産生する技術のサーベイを行っている。また既存のゲノムアセンブリ手法が、Overlap graphに依拠するものと、de Bruijn graphに依拠するものとの二種に大別されることを述べ、それぞれの特徴を説明し、本研究でde Bruijn graphに基づく戦略を採用した動機を述べている。第三章では、提案手法に基づき実装したシステムの構成を述べている。システム全体は、シーケンスの生データの前処理を行うものと、処理後のシーケンスデータを入力としてアセンブルを行うアセンブラからなることを示し、それぞれの役割と開発の指針を説明している。第四章では、シーケンス生データの前処理を行うために提案

した手法群 (DNA ScissorおよびQamar) のアルゴリズムの説明および性能評価を行っている。DNA Scissorはシークエンスデータに混入するクローニングベクター由来の配列の除去を行う手法であり、Qamarはシークエンスデータに含まれる配列読み取りエラーの訂正アルゴリズムである。提案手法を様々なシミュレーションデータおよび実データセットに対して適用し、各提案手法がそれぞれの既存手法よりも精度が高いことを明らかにしている。第五章では、提案した二種のアセンブラー (Arapan-SおよびArapan-M) のアルゴリズムを説明し、それぞれのアセンブラーを種々の実データに対して適用し既存手法と性能を比較し提案手法の有効性を明らかにしている。Arapan-Sは小規模ゲノム用のアセンブラーであり、de Bruijn graph中のバブルと呼ばれる部分グラフ構造の解消に高速なアルゴリズムを提案している。またアセンブリアルゴリズムとして、従前使われてきたグラフ中の最短経路を探索する方法ではなく、グラフ中のノードに付随するk-mer配列の頻度または長さの情報を利用する新規手法も提案し、双方の性能を比較し結果を議論している。一方、Arapan-Mは、Arapan-Sを基に開発された長い繰り返し配列を含む中規模ゲノム用のアセンブラーであり、申請者はde Bruijn graphの注意深い観察から、グラフ構造のクリーニングの段階において解消されるべきものとして四種の新規部分グラフ構造を見だし、それぞれの解消アルゴリズムを提案している。更に各部分グラフ構造のアセンブル精度への寄与を実データに対する実験により示し結果を議論している。第六章では、本研究を進めるに当たり開発した種々のツール群のアルゴリズムを説明し、性能を評価している。第七章では、研究の概括と今後の展望を述べている。

本論文は、ゲノムアセンブリにおいて直面する多種多様な問題に対して現実的かつ効果的な解決策を与えるものであり、今後の研究への指針を示すという点で大きな貢献をなすものであり評価に値する。

よって本論文は博士 (情報理工学) の学位請求論文として合格と認められる。