

琉球大学学術リポジトリ

DNA配列アセンブリの信頼性向上に関する研究

メタデータ	言語: 出版者: 琉球大学 公開日: 2017-10-30 キーワード (Ja): キーワード (En): 作成者: 大城, 絢子, Ohshiro, Ayako メールアドレス: 所属:
URL	http://hdl.handle.net/20.500.12000/37365

平成 29年 8月 8日

琉球大学大学院
理工学研究科長 殿

論文審査委員

主査 氏名 岡崎 威生

副査 氏名 遠藤 聡志

副査 氏名 名嘉村 盛和



学位（博士）論文審査及び最終試験の終了報告書

学位（博士）の申請に対し、学位論文の審査及び最終試験を終了したので、下記のとおり報告します。

記

申請者	専攻名 総合知能工学 氏名 大城 絢子	
指導教員名	岡崎 威生	
成績評価	学位論文 <input checked="" type="checkbox"/> 合格 <input type="checkbox"/> 不合格	最終試験 <input checked="" type="checkbox"/> 合格 <input type="checkbox"/> 不合格
論文題目	DNA配列アセンブリの信頼性向上に関する研究 A Study on accuracy improvement for DNA sequence assembly	
審査要旨（2000字以内） ギガシークエンサーの性能の飛躍的な向上と低コスト化により、膨大なDNA配列データの高速な獲得が可能になったことで、読み取り配列であるリードの結合過程であるDNAアセンブリの研究が進んでいる。生成されたリードデータはDNA配列解析に用いられるが、配列解析には正確な配列データが前提であるため、アセンブリ精度は非常に重要である。高速並列処理により1日あたりに獲得できる配列数は大幅に増えたものの、シークエンサーによる読み取りエラー部位への対応、正しいアセンブリの実行は現在でも大きな課題となっている。代表的な読み取りエラー除去法としてk-merがあるが、適用するk値によりアセンブリの結果が大きく異なるため、最適かつ頑健なk値の決定が困難という課題がある。		

(次頁へ続く)

審査要旨

また特に与える引数や用いる手法などの、実行条件へのアセンブリ結果の依存は、未知配列の復元であるde novo assemblyにおいて解決すべき大きな課題となっている。

本研究では、ギガシークエンサーによるDNA配列のアセンブリアルゴリズムの信頼性改善を目的として、機械学習アルゴリズムを用いることで以下の提案を行った。

1. 重み優先探索と機械学習アルゴリズムによるアセンブリ手法
2. 複数のk-merと手法を統合しアセンブリの依存を改善し頑健性を強化した、ダブル・アセンブリ手法
3. 複合決定木による配列結合ルール獲得のための、ダブル・アセンブリ手法

第1の提案手法においては、リード間の重複長を重みとした重み優先探索法と、重複長やcontig長の合成関数を説明変数、結合の正誤を目的変数とした学習データのガウシアンSVMへの適用により生成し判別ルールを獲得した。検証実験の結果、誤結合配列の除去が可能になったこと、リード本数の少ない場合は5 base程度の結合下限値を設定し判別ルールとの組み合わせ・リード本数の多い場合は高結合下限値設定により、重複部位の精度を維持しながら誤結合配列の除去が可能であるという成果を得た。

第2の提案手法においては、複数のk-merと手法によるcontigを統合するダブル・アセンブリを提案した。さらにcontig上におけるk-merのcoverage valueの分布を波形として表現し、周波数解析による特徴量を用いて結合を決定する手法を提案した。k-merの分布特徴量のフーリエ変換成分や勾配、結合対象であるcontigの波形の類似度など18の変数を説明変数とし、結合の正誤を目的変数とした学習データを用いて、C4.5への適用により判別ルールを獲得した手法を提案した。検証実験の結果、従来のアセンブリで生成が困難であった長い正結合配列の生成し、被覆率を改善できた。

第3の提案手法においては、結合の信頼性を示す結合正誤に加え、従来のアセンブリにて結合精度の指標としてk-merのcoverage valueの最小値、重複長を目的変数として適用した。さらに、正誤ルールを同時に適用することで、両ルールの対応が困難なcontigの取得または除去を試みた。検証実験の結果、従来の決定木による判別ルールでは対応できなかったcontigへの判別が可能になった。

本研究にて提案した複合決定木による判別ルールを用いたDNAダブル・アセンブリにより、適用するk-merや手法に影響されずかつ被覆率が改善されたアセンブリ、さらに従来の機械学習アルゴリズムでは対応できなかった判別ルールの生成により、より幅広くcontig結合の正誤の判別が可能になった。関連した特徴量の適用により目的変数を拡張するという複合決定木は、アセンブリに限らず機械学習アルゴリズム全般の分野において応用が可能である。

したがって、本研究成果は工学的に有用であり、提出された学位論文は博士の学位論文に相当するものと判断し学位論文の審査を合格とする。また、論文発表会における発表ならびに質疑応答において、申請者は専門分野および関連分野の十分な知識ならびに十分な研究能力を有していることが確認できたので最終試験を合格とする。