様式８の１の１　別紙１

論文の内容の要旨

専攻名　システム創成工学専攻
氏　名　遠藤友基

大規模ゲノムの解読は，生命システムの解明だけでなく，医療科学，薬学，農学などの多様な応用が考えられるため，生命に関する多くの分野において重要な要素となっている．ゲノムを解読するためには，DNAの全塩基配列を解読することが必要となる．一方，次世代シーケンサと呼ばれる技術の進歩により，DNAからリードと呼ばれる短い塩基配列の断片を短時間で大量に読み取ることが可能となってきた．一般的な塩基配列の決定は，この大量のリードを計算機で読み込み，適切に繋ぎ合わせることでコンティグと呼ばれる長い塩基配列を決定することで実現される.この技術はアセンブリアルゴリズムと呼ばれ，そのうち未知である塩基配列を決定するものはde novoアセンブリアルゴリズムと呼ばれる.de novoアセンブリアルゴリズムを実現するプログラム・ソフトウェアはde novoアセンブラと呼ばれ，様々な手法が提案されている．de novoアセンブラのうち，比較的少ない消費メモリ量で高精度のコンティグが得られるため，VelvetやSOAPdenovoが現在では特に広く普及している．しかし，大規模なゲノムのde novoアセンブリでは，極めて膨大な数のリードを扱う必要があるため，VelvetやSOAPdenovoを用いても消費メモリ量が非常に膨大となってしまいメモリ不足となりやすい．本論文ではこうした背景を踏まえ，de novoアセンブリアルゴリズムにおける課題について検討し，次世代シーケンサから得られた大量のリードを用いて，大規模なゲノムに対してもde novoアセンブリが可能となるように，消費メモリ量の少ないde novoアセンブリアルゴリズムを提案する．

以下に本論文の構成と各章の内容について述べる．

第１章では，本研究に至る背景と研究の目的を述べる．まず，ゲノムの解読と塩基配列の決定について，塩基配列の決定に必要となるde novoアセンブリアルゴリズムの概要と問題点について述べ，本研究の目的と意義を明らかにする．

第２章では，DNAの構造や塩基配列とゲノムとの関係について概説する．ゲノムとは膨大な量の遺伝情報であり，その生物一個体を形成するために必要な全ての情報が含まれている．そのゲノムの解読には，DNAを構成する塩基配列の決定が必要であることを述べる．さらに，ゲノムの持つ役割とゲノム解読の有用性について説明する．

第３章では，ゲノム解読のための重要な最初のステップとなる，塩基配列の決定方法の手順について述べる．まず，リードを生成するためのシーケンシングについて述べ，次にリードをつなぎ合わせコンティグを生成するde novoアセンブリについて述べる．さらに，リードの一般的な表現方法(フォーマット)について概説する．

第４章では，de Bruijnグラフを用いたde novoアセンブリアルゴリズムについて述べる．Velvetをはじめとする多くのde novoアセンブラでは，de Bruijnグラフと呼ばれるグラフを用いて

実現されており，提案手法においてもVelvetをベースにde novoアセンブリを実現している．したがって，提案手法のベースとなる，de Bruijnグラフを用いたde novoアセンブリアルゴリズムの原理と課題について説明し，提案手法の方針について述べる．

第５章では，消費メモリ量を大幅に削減したde novoアセンブリアルゴリズムである提案手法について述べる．リードからde Bruijnグラフを構築し，de Bruijnグラフからコンティグを得るまでの提案手法の各手順について説明する．提案手法は，de Bruijnグラフを用いたde novoアセンブラであるVelvetをベースに実現しているが，最低限の情報のみをメモリ上に保持することで大幅な消費メモリ量の削減を行っている．

第６章では，提案手法の有効性を確かめるため実験を行う．実際の次世代シーケンサより得られたE. coli K-12 strain MG1655及びヒトの14番染色体のリードを用いてアセンブリを行う．比較手法として，VelvetとSOAPdenovo2(SOAPdenovoの後継)と比較し，その結果について考察する．実験の結果，本手法はE. coliに対しては他手法の約20%，ヒト14番染色体に対しては他手法の約60%の消費メモリ量でde novoアセンブリが可能という結果が得られた．

第７章では，本研究の成果を総括し，得られた知見をまとめる．また，今回の実験結果を踏まえた今後の課題・展望について述べる．