論文の内容の要旨

論文題目　大規模生物配列情報における特異領域と反復領域の探索に関する研究
氏　　名　山田智之

ゲノム配列解析が多様化・大規模化するにつれて、遺伝子配列のゲノム上での位置を特定するといった単純なアライメントの課題だけではなく、遺伝子の発現量や多様性を解析するという課題が生まれてきた。それに伴って、ゲノム配列や、遺伝子配列においてその配列を特徴づける配列を発見する課題が重要性を増してきたといえる。具体的には高頻度に出現する配列領域を発見し取り除くという課題や、特異性の高い領域を発見するといった課題である。これらは配列全体を概観する際に重要である他、SAGE等のタグ配列の解析や、遺伝子配列に対するプライマーやsiRNA等の配列設計を行う際に有用である。

本論文では大規模な生物配列情報から特異性の高い領域と繰り返し領域を効率的に発見するアルゴリズムを提案し、その応用として基礎的な生物実験を設計する方法について述べている。いくつかのミスマッチが許されたとしても依然として他の配列にアライメントされることがないような極めて特異性の高い配列を発見するためには、類似の配列を見落としなく列挙する方法が必要である。本論文ではオーバーラップシードを用いた高速な配列アライメントアルゴリズムを提案しているが、このアルゴリズムはその要請を満たしている。オーバーラップシードを用いることによって短い配列がクエリーである場合には検索に必要な計算を少なくすることができ、高速な検索が可能になっている。一方BLAST、SSAHA、BLATやPatternHunterのアルゴリズムは高い確率で類似の配列を発見する方法であるため、一定の割合で類似配列を見落としてしまい、ユニーク領域を発見するアルゴリズムには利用することができない。

オーバーラップシードを用いるアルゴリズムによって、すべての類似配列を高速に列挙することができるようになったが、クエリー配列の長さが長くなると、それでも特異性の正確な指標を計算は困難であるため、特異性の指標の下界を高速に求める方法を検討した。特異性の指標の下界を用いると、ユニーク領域の全てを発見することはできないが、発見された領域はユニーク領域であることを保証することができる。この計算を行うために接尾辞配列等いくつかのデータ構造を導入しているが、このデータ構造はリピート領域を発見にも利用することができる。

最後に、これらのアルゴリズムの生物実験設計への応用例を紹介している。siRNA配列、マイクロアレイ配列、マルチプレックスPCRプライマーを用いる実験において、精度の高い実験を効率よく行うためには対象配列に対しての特異性を保証することは重要であり、本論文のアルゴリズムを活用して設計システムを構築している。

これまでにも、大規模なゲノム配列が続々と解読されているが、今後はsolexaや454などの新たなシークエンシング技術によって更に解読される配列数は莫大なものになると考えられる。この新たな局面を迎えて、ゲノム科学はゲノム配列の解読からゲノム配列の活用へとその領域を拡大してきている。また、遺伝子配列の決定方法や発現量の測定方法も大きく変わろうとしている。本論文で紹介する手法はそのような今後の新しい実験手法にも応用可能なものである。