氏　　名　粕川雄也（かすかわたけや）

博士の専攻分野の名称　博　士（情報科学）

学位記番号　第　19677　号

学位授与年月日　平成17年3月25日

学位授与の要件　学位規則第4条第2項該当

学位論文名　**遺伝子機能解析のための大規模配列データ解析手法に関する研究**

論文審査委員　（主査）
教　授　松田　秀雄

（副査）
教　授　清水　浩　　教　授　柏原　敏伸　　教　授　赤澤　堅造

## 論文内容の要旨

近年の大規模配列決定プロジェクトにより、様々な生物種について、大量の遺伝子の配列が入手可能となり、現在は同定された遺伝子の機能を調べる研究（遺伝子機能解析）が進められている。その中で、特にmRNA（メッセンジャーRNA）を逆転写させて得られたcDNA（相補DNA）を用いた遺伝子機能解析では、各々の遺伝子を代表するcDNAはどれか、各々のcDNAからどのようなタンパク質が生成されるか、各々の遺伝子についてどのような機能が知られているのか、といった情報が必要不可欠である。

本論文では、遺伝子機能解析のために必要とされるこれらの情報を得るために行われる以下の3種類の配列データ解析について、特に大規模に実施する際に起こる問題点とその解決法の提案を行う。

遺伝子クラスタリングは重複して得られた遺伝子配列やcDNAの塩基配列を同じ遺伝子由来のものごとにグループ化する解析であり、従来、手動による方法や、配列間の類似度をもとにまとめるなどの方法が取られていた。しかし、前者には大規模なデータへの適用が困難であり、後者は精度が良くないといった問題があった。そこで、精度良く、大規模なデータセットにも適用できることを目指した遺伝子クラスタリング法を提案し、実際のデータを適用した結果の評価により本手法の有効性を示す。

タンパク質コード領域の予測はcDNAの塩基配列の中で翻訳されるタンパク質のアミノ酸の並びを決める領域（タンパク質コード領域）を同定する解析である。この解析はcDNAの塩基配列が全長でない場合、塩基に誤りがある場合、タンパク質コード領域がそもそも存在しないmRNA由来である場合などがあり、予測が難しいという問題がある。そこで、現在提案されているコード領域予測法の評価を行うことで、予測法の考察を行う。

遺伝子機能アノテーションは遺伝子について過去の研究ですでに知られている機能についての情報を収集し、遺伝子の機能情報として付与することを目的とした解析である。従来は小規模で、手動により行われることが多かったが、大規模なcDNAを使った機能解析のためには、コンピュータにより自動化された機能アノテーション法が必要となっている。そこで、機能アノテーションの自動化法について提案し、手動による結果との比較から本提案手法の有効性を示す。

また、これらの解析を大規模に実施するためには、効率的なデータ交換（公開データの収集および解析したデータの配布）が必要である。そこで、外部で公開されている複数の生物データベースから、より効率的にデータを収集することを目的とした、分散データベース上での実体化ビューの効率的な管理法を提案し、実際のデータベースから抽出したパラメータを用いて評価する。また、XML（extensible markup language）を用いて定義された、機能アノ

テーション結果のデータ交換のためのデータ形式を提案し、有効性を示す。

最後に結論として本研究で得られた成果を要約した後、今後の課題について述べる。

## 論文審査の結果の要旨

本論文では、遺伝子の機能を解明するために必要な、「遺伝子クラスタリング」、「タンパク質コード領域予測」、「遺伝子機能アノテーション」の３種類の配列データ解析について、大規模な計算機解析を行うための提案を行っている。従来これらの解析は手動で行われることが多かった。しかし、近年、遺伝子配列データが大量に決定可能となり、そのため、これらの配列データ解析を、計算機上で自動的に処理でき、かつ手動による結果と同程度の精度の結果を得ることのできるような方法が必要とされていた。そこで、遺伝子クラスタリングに対しては、数十万個規模の大量の配列データセットに対して適用可能な自動による手法を提案している。提案手法では、手動による結果に近づけるため、外部のクラスタリング結果を統合する方法が中心となっている。また、実際のデータセットに提案手法を適用して有用性を示し、さらに、他のクラスタリング手法の結果と提案手法による計算結果とを比較して、ある遺伝子の逆鎖側に重なって存在している遺伝子が別のクラスタとして正しく分けられているか等のいくつかの観点について、結果の改善が見られたことも示している。タンパク質コード領域予測に対しては、既存の６種類の予測法と手動による結果とを比較して、いくつかの手法を組み合わせることで結果の精度を向上できることを示している。遺伝子機能アノテーションに対しては、手動による結果とよく一致することを目指した自動化手法を提案し、実際のデータに適用して７割から８割程度一致することを示している。

さらに本論文では、配列データ解析を効率よく進めるために、大量のデータを外部と交換するためのデータ交換手法に関する２つの提案を行っている。１つめに、大量に生成された機能情報を効率よく収集できるようにすることを目的とした、分散データベース環境における効率的な実体化ビューの更新反映法を提案している。そして、この提案手法の正当性を示し、分子生物学データベースから抽出したパラメータを用いて有効性を示している。２つめに、従来は存在していなかったcDNAの機能アノテーション情報を交換するための標準データ形式を提案している。このデータ形式は機能アノテーションのデータベースで実際に利用されている。

以上により、本論文の成果は計算機による遺伝子機能解析に関する研究の発展に貢献するものである。よって博士（情報科学）の学位論文として価値あるものとして認める。