審 査 の 結 果 の 要 旨

氏　　名　小野寺　拓

計算生物学において、次世代シーケンサーや分子動力学シミュレーションによって質・量の面でこれまでのゲノムデータをはるかに上回るデータが利用可能となってきており、そのような多様な「ポスト」ゲノムデータの質・量に応じて高次解析を行うためのアルゴリズムの開発が重要となっている。本論文はこうした計算生物学の研究進展に応じた高性能アルゴリズムの開発を行っている。データの規模に対応する効率性を達成するため、新たにアラインメントフリーなアルゴリズムという枠組みを構築し、メタゲノムおよび真核生物ゲノムなどの質・量ともに複雑なゲノムから有益な情報を抽出するための新たなゲノム解析モデルも提案している。

ここで、アラインメントフリーな解析法として、長い計算時間を必要とする一般のアラインメントアルゴリズムに基づかないデータ解析法で、配列・構造のアノテーションを与える方法が提案されている。機能予測におけるアラインメントフリー的手法として部分文字列などの頻度に基づく文字列カーネルとサポートベクターマシン(SVM) を組合せた手法があるが、そこではスペクトラム・カーネル法では比較的高い予測精度に加え、接尾辞木を用いた高速な計算が可能であり、文字列アルゴリズムの分野で発展してきたものより高度なデータ構造も援用して拡張アルゴリズムの開発が行われている。

本論文の研究貢献部分は３章から６章であり、各章は次のような内容となっている。まず３章においては、次章で提案するカーネル関数を特徴空間の次元によらない計算量で求めることを目標に、接尾辞配列の一般化であるb 接尾辞配列というデータ構造が与えられている。このデータ構造は接尾辞配列の一般化であり、特定の位置にワイルドカードを含むパターンの検索にも応用可能である。このようなパターンマッチングは実際にspaced seed を用いた配列相同性検索において必要とされている。

４章ではSVMのスペクトラム・カーネルと同様に部分文字列の頻度に基づくギャップ付スペクトラム・カーネルという文字列カーネルが提案されている。部分文字列の中にギャップ（パターンマッチにおけるワイルドカード）を導入する点が拡張となっている。類似のアイディアとしては、ワイルドカード・カーネルという既存手法があり、これは特定のサイズと重みのギャップパターンを全て同時に考慮するものであるが、複数のギャップパターンを考慮した場合に起こり得る事態については解明されないままだった。この課題の解決に３章の結果を応用して取り組んでおり、ギャップ付スペクトラム・カーネルを求める効率的なアルゴリズムが提案されている。また、SVM を用いた予測に必要な計算をサポートベクターのサイズによらない計算量で行う手法も提案されており、これらの独自な方法によりSVM のための効率的なアルゴリズムを得た上で、その有効性が計算機実験を通して示されている。

５章ではタンパク質立体構造に対するアノテーションに関する新提案がなされている。タンパク質の構造は配列よりも直接的に機能に関係しており、既知であれば機能推定のための重要な手がかりになる。一方立体構造の類似度の指標として標準的に用いられている構造アラインメントとよばれる手法は配列アラインメントよりもさらに多く

の計算量を必要とする。そこで５章では、４章において模索した部分文字列の頻度に基づくアラインメントフリーな文字列カーネルのテクニックを立体構造に応用し、構造に対するアラインメントフリーなカーネルを与えている。また、二次元接尾辞木とよばれる既存のデータ構造を用いた効率的なカーネルの計算法と、サポートベクターのサイズによらない予測法も提案している。タンパク質superfamily の予測実験では、提案手法が類似の既存手法のうち最も精度のよいものと同程度の予測精度を、300 倍以上高速に達成できることを示している。

６章においてはゲノムアセンブリが研究対象となっている。既存手法では、リードのオーバーラップを表現するグラフを作り、そのグラフのパスをたどることで元の配列、またはその中に含まれる長い部分文字列を復元するというアプローチが主流である。グラフの構築に関しては多くの先行研究があるが、グラフから元の配列を復元する部分に関してはまだ確立された方法は存在しない。シーケンスエラー、リピート配列、多倍体ゲノムなどによりグラフの中に導入される部分構造をいかに処理するかが課題となっており、既存手法ではグラフに含まれる単純なモチーフを用いてこのような部分構造を検出していた。一方、そうした単純なモチーフでは検出できない複雑な部分構造が出現する場合があることがわかっていた。最終的にはこれらの部分構造をどう処理するかが問題になるが、それ以前にこれらの部分構造はより単純なモチーフと異なり検出自体が非自明である。そこで申請者は複雑な部分構造のグラフ理論的な特徴づけとしてスーパーバブルと呼ぶ部分グラフのクラスを定義し、その性質を明らかにした。また、与えられたグラフに含まれる全てのスーパーバブルを検出するための効率的な手法を提案している。提案手法はグラフの頂点数に対して最大で二乗の計算量を必要とするが、実際には非常に効率的に動作していることを確率的なモデルのもと証明に成功している。

全体として、本論文では網羅的な生物学的配列・構造研究にむけたアラインメントフリーアルゴリズムを開発することを目指し、多様なポストゲノムデータの高次解析を可能にする成果をあげている。

なお、本論文の一部は共同研究によって得られたものであるが、申請者が主体的に研究して得られた成果であることを確認している。

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。