審 査 の 結 果 の 要 旨

氏　　名　　坂内　英夫

本論文は，生物配列からのパターン発見手法に関して研究をしたものである．配列に潜む生命情報を解明する手掛りを得るために，共通の性質や機能を持つ配列集合から効率的に意味のあるパターンを発見する手法は，莫大な量の文字列データを生産している分子生物学の分野においては非常に応用性が高く，必要とされる技術である．

まず，アミノ酸配列のＮ末端側に存在する３種の局在化シグナルの特徴抽出の問題に取り組み，パターン等の配列に関するルールを生物学的な知見に基いて設計し，徹底したパラメータ空間の探索を行なった．結果として，比較的簡素なルールであるにもかかわらず，ニューラルネット法によって得られる最高性能に迫る予測性能を得ることに成功した．得られたルールを基に，局在化シグナル予測システム iPSORT を構築し，サービスとして公開をしている．

局在化シグナルの問題に取り組んだ際に，従来の配列集合に共通するパターン発見の技術は有効ではあったが，それらに加え，アミノ酸の様々な物理化学的・生化学的性質を数値として表わした情報（アミノ酸指標）を解析に用いたことが成功に寄与した．近年においてはマイクロアレイを始めとする観測技術の発達により，配列情報と深く関係した様々な定量データが大量に生産されつつある．そのような背景から，従来は文字列情報単独で行っていたパターン発見の手法に，数値・カテゴリ属性を加えることで，より有効に確かなパターンを抽出することのできる新しいパターン発見の手法について研究を進めた．そこで，入力として文字列集合が与えられ，各文字列に対応付けられた数値属性が存在する時に，パターンが文字列中に出現する・しないという事象と数値データとの相関が高くなるようなパターンを探す，相関パターン発見問題を定義した．このように文字列属性と数値属性の両方の情報をアルゴリズムの中で同時に扱うことで，それぞれを単独に用いるだけでは見逃してしまう関係性を発見できることが期待できる．

本論文では，数種類のパターンクラスに関して，相関パターン発見問題を最適にかつ効率良く解くためのアルゴリズムを示した．まず，部分文字列パターンに関しては接尾辞木(suffix tree）を用いた線形時間アルゴリズムを示し，更に接尾辞配列（suffix array）を用いた効率の良い実装を与えた．次に，二つの部分文字列パターンを論理演算で組合せたパターンに関して，最適な組合せを $O(N^2)$ で求めることのできるアルゴリズムを示した．更にそのアルゴリズムを一般に k 個のパターンを組合せた場合に一般化し，最適解が $O(N^k)$

時間で求められることを示した．より複雑なパターンクラスに関しては，この問題は多くの場合に NP-困難であることが示されるが，様々なパターンクラスに関して有効な分枝限定法を用いたアルゴリズムを与えた．

最後に，それぞれのアルゴリズムを実際の生物配列データに対して適用した．具体的には，1）遺伝子の発現量と相関するパターンを遺伝子のコード領域の上流配列から探す，転写因子結合部位に関するパターン発見，2）mRNA の分解速度と相関するパターンを 3' UTR 配列から探す，mRNA の寿命エレメントに関するパターン発見，3）イントロンの長さと相関するパターンをイントロンの acceptor site 配列から探す，長大イントロンの特徴に関するパターン発見を行なった．いずれの例においても，考案したアルゴリズムは素朴なアルゴリズムと比べて速度の面で効率が良く，素朴なアルゴリズムでは現実的な時間で計算が困難な場合でも問題を解くことができた．発見されたパターンに関しても生物学的に妥当な解釈ができるものであり，問題設定及びアルゴリズムの有効性・有用性を確認する事ができた．

このように，本論文は情報理工学のコンピュータ科学分野において、特にバイオインフォマティクスにおける顕著な研究成果をあげたものである．なお，本論文の研究内容は共同研究者との研究遂行により得られたものであるが，申請者が主体となって行った研究による成果であると認めた．

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．