論文の内容の要旨

論文題目　Efficient Algorithms and their Applications for Optimal Pattern Discovery from Biosequences

生物配列からの効率的な最適パターン発見アルゴリズムとその応用

氏名　　坂内 英夫

生物配列からのパターン発見手法に関して研究をする．近年様々なゲノムプロジェクト等により，DNA，RNA，タンパク質などの生物配列データが大量に蓄積されて来ている．しかし，何種類もの生物種の完全ゲノムが決定されて来てはいるものの，未だにこれらの配列にどのような情報がどのように埋め込まれているかに関してわかっていない事は多い．配列に潜む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，意味のありそうな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．

本研究ではまず，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に関する特徴抽出を行なう．徹底したパラメータ空間の探索の結果，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、比較的簡単なルールの組であげることに成功する．その際に従来の配列集合に共通するパターンを探索する技術は有効であったが，それらに加え，アミノ酸指標データベース (AAindex) に含まれる，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を用いた事が成功に繋がった．

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，他にも配列情報と深く関係した様々な定量的なデータが大量に生産されつつあり，このような新しい情

報をどのように用いて知識を抽出すれば良いのかが重要な問題となって来ている. そこで, 本研究では従来文字列情報単独で行っていたパターン発見手法に数値・カテゴリ属性を加え, より有効に確かなパターンを抽出するための新しい手法の開発に取り組む. 文字列データと対応付けられた数値データが存在する時に, 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式化する. 文字列属性とそれに深く関係した数値・カテゴリー属性の両方の情報をアルゴリズムの中で同時に扱う事で, 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事が期待される.

本研究では特に, 適当な条件を満たす評価関数の基で, 最適なパターンを発見するアルゴリズムについて研究し, 幾つかのパターンクラスに関して相関パターン発見問題を効率的に解くためのアルゴリズムを与える. まず, 部分文字列パターンに対して接尾辞木 (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) を用いた効率的な実装を与える. 次に, 複数の部分文字列パターンを論理的に組合せたパターンに関するアルゴリズムを示す. また, より複雑なパターンクラスを考えた場合, この問題は一般的には NP-hard であることが示されるが, 多くの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る.

開発したアルゴリズは実際の生物配列データに対してを適用し, 手法の有効性を確かめる. 具体的には転写因子結合部位に関するパターン発見, mRNA の寿命エレメントに関するパターン発見, 長大イントロン配列に関するパターン発見, を行なう. いずれの例においても, 開発したアルゴリズムは素朴なアルゴリズムに比べて速度の面で効率的である事が示された. また, 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った.