5N-08

配列データベースから正規表現のモチーフを抽出する方法

井上 雄貴† 福本 翔平† 森 康真† 北上 始†
広島市立大学情報科学研究科†

1. はじめに

配列データベースから頻出なパターンを抽出する手法は，DNA やアミノ酸などの生物配列データのモチーフ抽出など多くの問題解決に有用であるといわれている．この配列データベースからモチーフを抽出する方法は，数多く知られている．本稿では，Lawrence らによって提案された有名な GS（ギブスサンプリング）法に着目する．この手法は，初期解（ある部分配列集合）から始めて部分配列データ集合のプロファイル（位置依存スコア行列）に最も近い類似部分配列の位置を配列データごとに確率探索している．そのため，全域的な最適解に収束することが保障されておらず，局所的な最適解になってしまう可能性がある．

そこで，配列データベースを予め多重整列化をおこない，スライディングウインドウ法を用いて相対エントロピーが最大になる領域を見つけることで，高い出現頻度をもつミスマッチクラスタを抽出する手法[1]が提案されている．

本稿では，GS 法と多重整列に基づく手法で得た 2 つのミスマッチクラスタから段階的一般化法[2]を用いて最小汎化集合を抽出する．抽出された最小汎化集合から規則性を把握し，それらの精度等を評価，比較する．

2.ミスマッチクラスタの抽出

2.1 従来手法

従来手法として知られている GS 法の主な目的は図 1 のように，*M* 種類のアルファベットと配列数 *N* 本によって定義される配列データベース DB からユーザが定めた長さ *K* の部分配列集合（*K*-部分配列集合）を取り出し，できるだけ類似した部分配列集合となるように計算していくことである．*K* の部分配列集合とは，*N* 本の配列からなる配列データベースの各配列から取り出される長さ *K* の部分文字列集合のことである．

2.2 多重整列に基づく手法

まず，多重整列化により，ギャップと呼ばれる記号(-)を配列データの各文字を類似した部分

Method for extracting motifs with regular expression from sequence databases.
†Yuuki Inoue, Shouhei Fukumoto, Yasuma Mori, Hajime Kitakami, Graduate School of Information Sciences, Hiroshima City University

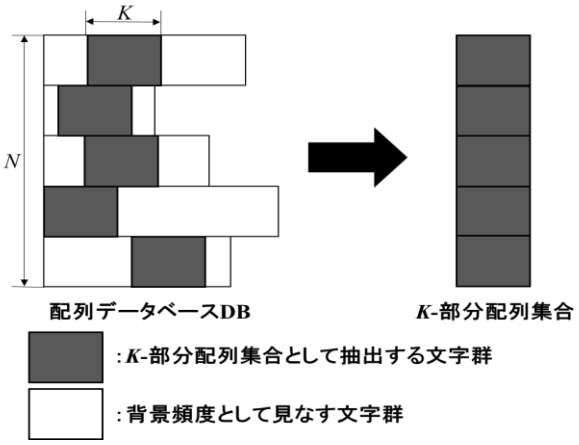

図 1：GS 法

で特定できるように挿入し，配列データベースの長さを統一することで多重整列を得る．

次に，この得られた多重整列から，相対エントロピーが最大となる *K*-部分配列集合の領域を選択し，その選択された結果を初期値として，新しいプロファイル計算法を用いて，GS 法を適用する．

3.段階的一般化法

段階的一般化法は，ミスマッチクラスタに含まれる要素を少しずつ組み合わせることにより最小汎化パターンを探索するボトムアップアプローチである．このとき小さなサイズの集合要素から探索をおこない，探索過程において，列挙された各集合要素に対する最汎パターンの計算，不要な部分列挙木の枝刈り，冗長パターンの除去などを実施することにより，高速にミスマッチクラスタの最小汎化集合を抽出する．しかし，ミスマッチクラスタを構成する部分文字列が増えると計算時間が膨大になるという問題もある．

4.処理手順

配列データベースから最小汎化集合を抽出するまでの処理手順は以下の通りである．また，それらの流れを図 2 に示す．

(1)配列データベース DB に対して，GS 法および多重整列に基づく手法を用い，各々に対するミスマッチクラスタ MIS を得る．

(2)MIS に対して多重整列化をおこない，ギャップ記号(-)を挿入することで MIS の各配列の長さを等しくする．本稿では多重整列化のために ClustalX と呼ばれる系統解析用のプログラムを使

用した.
(3)多重整列化された MIS を構成する各配列の同じ列に，アミノ酸残基の物理化学的性質を表す行列において，同じ性質を持たないとされる文字が存在していた場合，その列を全てワイルドカード記号(*)に置き換える.
(4)ワイルドカードを含んだ MIS に対して$変換をおこなう．$変換とは，MIS を構成する各配列を列ごとに比較し，列についてまとまりのある部分列を$(*n*)と表記するように変換することである.
(5)$変換によって各配列が$(*n*)の記号列に変換された MIS に対して，段階的一般化法を用いて汎化処理を実施し，最小汎化集合を抽出する.

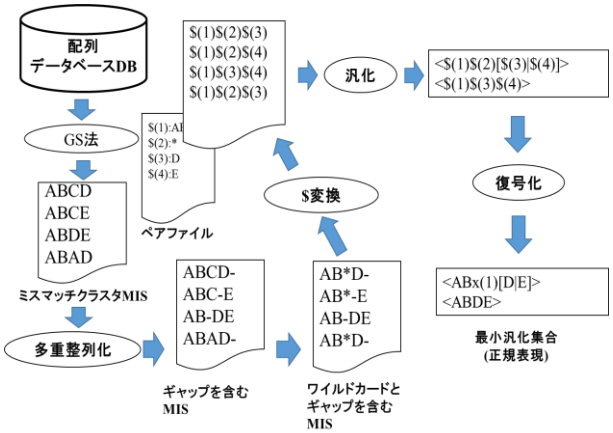

図 2：処理手順

5.評価実験

ミスマッチクラスタの抽出精度の違いが，最小汎化集合の汎化配列パターンの抽出精度に与える影響を比較する.

5.1 実験方法

評価をするために，従来手法と多重整列化を予め行う 2 種類の方法で得られた MIS を用いて最小汎化集合を抽出する．そして抽出された最小汎化集合の汎化配列パターンの支持数の多い順にランキングを行う．上位 3 位までにランキングされた汎化配列パターンについて評価を行うために，以下で定義される精度の式(1)を用いる.

$$\text{精度}(\%) = \frac{B}{B+C} \times 100 \qquad (1)$$

モチーフ配列パターンと一致している汎化配列パターンの文字数を B，それ以外のノイズ文字を C とする．この値が大きいほどモチーフ配列パターンを取り出せているとする.

使用したデータセットの一部を表 1 に示す.

表 1：使用したデータセット

データセット	長さ	データ件数
Kringle	14	95
PTS_EIIA	17	51

また，従来手法と多重整列に基づく手法で求めた MIS の精度を表 2 に示す.

表 2：提案手法と従来手法の精度の比較 (MIS)

データセット	従来手法(%)	多重整列(%)
Kringle	64.44	77.22
PTS_EIIA	49.18	75.58

5.2 実験結果

表 3 に評価実験の結果得られたデータセットごとの従来手法と多重配列に基づく手法の精度を示す.

表 3：提案手法と従来手法の精度の比較 (最小汎化集合)

データセット	従来手法(%)	多重整列(%)
Kringle	61.63	88.1
PTS_EIIA	17.65	100

5.3 考察

表 3 の結果を見ると，従来手法から得られた汎化配列パターンより，多重整列化に基づく手法によって得られた汎化配列パターンの方が，精度がよくなっていることがわかる．最小汎化集合を計算し，支持数でランキングをしたことにより，モチーフ配列パターンを含まない配列を除外し，モチーフ配列パターンを多く含む配列を上位にランキングできたことで精度を上げることができたと考えられる．また，従来手法の精度が下がっているが，これは多くの文字がワイルドカードに置換されたため，モチーフを見つけることができなくなったからだと考えられる.

6.まとめ

従来手法と多重整列に基づく手法の 2 種類の方法で得られたミスマッチクラスタから最小汎化集合を計算し，モチーフ配列パターンの抽出精度を求めた．結果として，多重整列に基づく手法を用いたデータセットでは精度が上がり，従来手法の精度は下がってしまった.

今後の課題として，今回はワイルドカード文字に置換する条件としてアミノ酸残基の物理化学的性質を表す行列を用いたが，BLOSUM 行列などの別の条件を用いて実験を行ってみることが挙げられる.

参考文献

[1] 福本 翔平，北上 始，森 康真: 多重整列に基づくモチーフの統計的抽出法，FIT2014, 3 pages (2014).
[2] 田村 慶一，木村 浩明，荒木 康太郎，北上 始: 段階的一般化法によるミスマッチクラスタからの最小汎化パターン抽出，電子情報通信学会論文誌, J93-D(3), pp.189-202 (2010).