2ZB-2

ギブスサンプリングを用いた可変長配列パターンの抽出

加藤 智之† 森 康真‡ 荒木 康太郎† 黒木 進‡ 北上 始‡
†広島市立大学大学院情報科学研究科 ‡広島市立大学情報科学部

1．はじめに

分子配列データベース中の頻出パターンに含まれるモチーフの発見支援のために，パターン成長アプローチを用いた手法が数多く提案されている．しかしながら，明らかに不要であるパターンが数多く抽出されるという問題がある．

本論文では，可変長パターン抽出法[2]にLawrence[1]らのギブスサンプリングの手法を新たに取り入れ，不要パターンを削除する方法について提案する．

2．用語の定義

2.1 ストリングとワイルドカード

配列データベース *DB* の各配列は，アルファベット文字で構成される．アルファベット文字と記号*で表されるワイルドカード文字(以下，ワイルドカードと呼ぶ)で構成される有限の文字列をストリングと呼ぶ．ワイルドカードは任意の 1 文字を表す．*k*-ストリングとは，*k* 個の文字で構成されるストリングを意味する．

2.2 パターン

パターンとは，複数の配列に共通に含まれている，*k*-ストリングの集合に対する表現形式である．*k* 個のアルファベット文字で構成されるパターン$<pat^k>$は以下の式(1)のように表現される．ただし，α_iは 1 文字のアルファベット文字とする．

$$<pat^k>=<\alpha_1\text{-}x(i_1,j_1)\text{-}\alpha_2\text{-}x(i_2,j_2)\text{-}\ldots\text{-}x(i_{k-1},j_{k-1})\text{-}\alpha_k>:cnt \quad (1)$$

式(1)中の *cnt* は支持数を表しており，$<pat^k>$が存在する，異なる配列番号の数を表している．また，ユーザが与えた最小支持数以上の支持数をもつパターンを頻出パターンと呼ぶ．

式(1)中の $x(i,j)$は，ワイルドカード領域と呼ばれ，各文字間に *i* 個から *j* 個のワイルドカードが含まれていることを表している．$i<j$ のとき，その領域を可変長ワイルドカード領域と呼ぶ．また，$\varepsilon = j - i$ を誤差と呼ぶ．ワイルドカード領域の範囲は，ユーザにより与えられた最大ワイルドカード数 wc_{max} 及び最大誤差数 ε_{max} により制限され，それぞれ $i \leqq wc_{max}$，$\varepsilon = j - i \leqq \varepsilon_{max}$ という関係が成り立つ．

3．可変長パターン抽出法

パターン成長アプローチを用いたスコープデータベース *SDB*[2]の方法は，頻出パターン中に含まれる可変長ワイルドカード領域の極小化，冗長性の除去を可能としており，優れた抽出能力をもっている．以下に，*SDB* による頻出パターン抽出法の処理手順を示す．

(1)入力パラメータの最小支持数，ワイルドカード数∈[0,wc_{max}]，誤差数∈[0,ε_{max}]を与える．
(2)*DB* をスキャンし，1-頻出パターン$<pat^1>$を全て求め，これらを F_1 とする．
(3)各$<pat^k>\in F_k$ 対して，新たに頻出パターンが抽出されなくなるまで以下の処理を繰り返す．
- F_k に対してスコープデータベースを構築する．
- 構築されたスコープデータベースから極小かつ，非冗長な(*k*+1)-パターンを生成する．
- 支持数を計算し，頻出な(*k*+1)-パターンを抽出する．
- (*k*+1)-頻出パターンに含まれる全ての可変長ワイルドカード領域を極小化し，F_{k+1} に追加する．
- $k = k + 1$

(4)全ての頻出パターンを出力する．

表 1 の配列データベースに対して上記処理手順を適用すると，図 1 の列挙木が得られる．

表 1:配列データベース

sid	配列データ
1	FKYAKWLCDN
2	SFVKTAEHNQC
3	ALR
4	MSKPL
5	FSKFLMAWEH

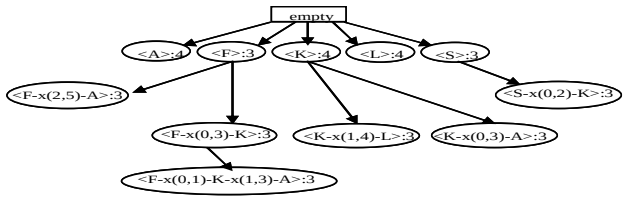

図 1:SDB により抽出される頻出パターン

一般に，パターン成長アプローチでは，頻出パターンを精密に抽出することができるが，多

Flexible Sequence Pattern Extraction using Gibbs sampler
†Tomoyuki KATO, ‡Yasuma MORI, †Kotaro ARAKI, ‡Susumu KUROKI, ‡Hajime KITAKAMI
†Graduate School of Information Science, Hiroshima City University
‡Faculty of Information Science, Hiroshima City University

量の頻出パターンが抽出されてしまう．その中の多くは明らかに不要なパターンである．

4．ギブスサンプリング

ギブスサンプリング[1]は，配列データベースの各配列から，指定した長さ k のできるだけ互いに類似した部分文字列(k-部分文字列)を抽出するアルゴリズムである．各配列の k-部分文字列の出現確率を計算し，できるだけ出現確率が大きい k-部分文字列を選択していくことで，互いに類似した k-部分文字列の集合を求めることができる．

5．提案手法

以下に我々が提案する処理手順を示す．

(1)ギブスサンプリングを用い，配列データベースに含まれる各配列データから類似する k-ストリング集合を切り出す．

(2)切り出された k-ストリング集合内の各要素は，お互いに類似しているが，同一ではないので，スコープデータベース *SDB* の方法を用いて，頻出な可変長配列パターンを抽出する．

以上により，提案手法は，配列データベースにスコープデータベースの方法を適用するよりも，着目する箇所が限定されるので，重要な頻出パターンだけを抽出することが期待される．さらに，k-ストリング集合は，互いにできるだけ類似したストリングの集合であるので，よりチーフの形式に近い頻出パターンの抽出が期待できる．

6．性能評価

従来手法と提案手法を比較するために，Zinc Finger モチーフを含むデータセットで実験を行った．Zinc Finger の形式は，<*C*-*x*(2,4)-*C*-*x*(3)-[*LIVMFYWC*]-*x*(8)-*H*-*x*(3,5)-*H*>である．配列データベースから上記モチーフを抽出するために，入力パラメータの最大ワイルドカード数を 8，最大誤差数を 2 として実験した．その結果を表 1,2 に示す．なお，表 2 中の k の値はギブスサンプリングで抽出する文字列の長さ，各表のモチーフ数は抽出された頻出パターン集合中に含まれるモチーフの数を表している．

Zinc Finger モチーフは，ワイルドカード領域を考慮すると最大 25 文字なので，k の最小値を 25 とした．k が 25 のときは，支持率 90%ではモチーフを発見できなかったが，支持率 80%以下では，提案手法と同等のモチーフを発見できた．この時，抽出される頻出パターン数は最大で約 2%まで減少している．さらに k を 100 とすると，支持率 90%でも従来手法と同等のモチーフを見つけることができた．このとき，抽出される頻出パターン数は最大で約 16%まで減少している．これより，ギブスサンプリングで抽出する k の値は，モチーフ長より長い場合のほうが良い性能を発揮することがわかる．

以上のことから，従来手法にギブスサンプリングを取り入れることで，モチーフの数を減少することなく，抽出される頻出パターン数を減少させることができるため，優れた抽出能力をもっているといえる．

表 1:従来手法の結果

比較項目/最小支持数	90%	80%	70%
頻出パターン数(件)	1042	9673	293218
モチーフ数(件)	9	9	9
計算時間(秒)	115.47	744.41	11572.22

表 2:提案手法の結果

比較項目/最小支持数		90%	80%	70%
k = 25	頻出パターン数(件)	44	242	935
	モチーフ数(件)	0	9	9
	計算時間(秒)	0.17	0.87	2.39
k = 100	頻出パターン数(件)	168	2143	47566
	モチーフ数(件)	9	9	9
	計算時間(秒)	5.29	55.65	985.15

7．まとめ

本論文では，パターン成長アプローチにギブスサンプリングを適用することで，不要パターンの除去を行った．Zinc Finger モチーフを含むデータセットで実験し，提案手法が有効であることを示した．今後の研究課題としては，あいまい文字を考慮した頻出パターンの抽出法の研究などがあげられる．

謝　辞

本研究の一部は，日本学術振興会・科学研究費補助金（基盤研究（C）（一般），課題番号：17500097），広島市立大学・特定研究費（一般研究費（コード番号:31006））の支援により行われた．

参考文献

[1] Lawrence,C.E., Altschul,S.F., Boguski,M.S., Liu,J.S., Neuwald,A.N. and Wotton,J.: Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment, Science,263,208-214, 1993.

[2] 加藤 智之，北上 始，森 康真，田村 慶一，黒木 進：極小かつ非冗長な可変長ワイルドカード領域を持つ頻出配列パターンの抽出，電子情報通信学会論文誌 D，データ工学特集号, Vol.J90-D, No.2, 2007 年 2 月出版予定