可変長配列パターン抽出法におけるギブスサンプリングを用いた不要パターンの除去方式

Elimination of Background Patterns using Gibbs Sampling on Flexible Sequence Pattern Extraction

加藤 智之♥ 森 康真♦ 荒木 康太郎♥
黒木 進♦ 北上 始♦

Tomoyuki KATO Yasuma MORI Kotaro ARAKI
Susumu KUROKI Hajime KITAKAMI

著者らは，パターン成長アプローチにより抽出される多くの不要パターンを除去するために，パターン成長アプローチとギブスサンプリングを用いた頻出配列パターン抽出法を提案する．ギブスサンプリングを用いて，各配列から長さ k の部分文字列集合を抽出し，抽出された集合に対してパターン成長アプローチを適用する．これにより，配列データベースの参照箇所が限定されるため，不要なパターンを除去することができる．本研究の有効性を示すために，Leucine Zipper モチーフ及び，Zinc Finger モチーフを含むデータセットを用いて実験を行ったところ，抽出される頻出パターン数を大幅に減らすことに成功した．

We propose a method for extracting frequent sequential patterns which used pattern-growth approach and Gibbs sampling, in order to extract candidates of a motif and remove unnecessary patterns from amino acid sequence databases. A Gibbs sampling is used and a set of *k*-subsequence is extracted from sequences, and pattern-growth approach is applied to the extracted set of *k*-subsequence. Therefore, the extracted frequent sequential patterns are made strict and unnecessary patterns can be removed. We evaluated using datasets that includes the Leucine Zipper and Zinc Finger motifs. As a result, we succeeded in reducing the large number of the frequent sequential patterns extracted.

1. はじめに

配列データベースから，頻出パターンを抽出することは，アミノ酸などの生物配列データのモチーフ抽出などの多くの問題解決に有効であるといわれている．モチーフとは，PROSITE[1]などで見られる生物学的に重要な機能をもつ特徴的なパターンである．アミノ酸配列や，テキスト情報などを含むデータベースに対する配列データマイニングでは，固定長や可変長のワイルドカード領域をもつ頻出配列パターンを抽出する方法の研究[2]が進められてきた．しかしながら，配列データマイニングのアプローチでは，数学的な規則性を漏れなく精密に抽出できるが，明らかに不要と思われる可変長配列パターンが大量に抽出されるという問題がある．

本論文では，頻出パターンが大量に抽出されるという問題点を解決することに着目している．そのために,従来の手法にLawrence らのギブスサンプリング[3][4]の手法を新たに採り入れ，配列データベースを正の部分文字列集合と負の部分文字列集合に分割することにより不要パターンを削除する方法について提案する．

2. 用語と問題の定義

配列データベース $DB = \{t_1, t_2, \cdots, t_n\}$において，各配列は s_{sid} と表現される(n は配列数，sid は配列番号)．各 s_{sid} 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(以降，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．ただし，ストリングの両端は，アルファベット文字に限定する．ワイルドカードは任意の 1 文字を表す記号である．ストリングの長さ k はストリングを構成するアルファベット文字数で決まる．例えば，<*FLMA*>は 4-ストリングであり，ワイルドカードを含むストリング<*F*K*A*>は 3-ストリングである．

2.1 パターン

k-パターンとは，複数の配列データに共通に含まれている k-ストリング(k 個のアルファベット文字をもつ)からなる特定の集合に対する表現形式である．例えば，2-ストリングの集合{<*F**K*>,<*FK*>,<*F**K*>}を説明する 2-パターン<pat^2>は，<F-$x(0,2)$-K>と表現される．a_i をアルファベット$\sum$の要素(1 文字)とすると，k 個のアルファベット文字をもつ k-パターン<pat^k>は以下のように表現される．

$$<pat^k> = <a_1\text{-}x(i_1,j_1)\text{-}a_2\text{-}x(i_2,j_2)\text{-}\ldots\text{-}x(i_{k-1},j_{k-1})\text{-}a_k>:cnt \qquad (1)$$

cnt は支持数を表し，<pat^k>が出現する配列の数(異なる配列番号 sid の数)を表している．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．

$x(i,j)$は，ワイルドカード領域と呼ばれ，文字 a_i と a_{i+1} の間にワイルドカードが i 個から j 個含まれていることを表している．$x(i,j)$の領域において，$i<j$ のとき，その領域を可変長ワイルドカード領域と呼ぶ．$i=j$ のとき，それを固定長ワイルドカード領域と呼び，この領域を $x(i)$で簡略表現する．また，$\varepsilon = j - i$ を可変長ワイルドカード領域の誤差と呼ぶ．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．ある k-パターンにおいて，全てのワイルドカード領域が固定長である場合，そのパターンを固定長パターンと呼ぶ．一つでも可変長のワイルドカード領域を含むパターンを可変長パターンと呼ぶ．

2.2 正パターンと負パターン

配列データベースに対してギブスサンプリングを適用すると，各配列から指定した長さ k の部分文字列が抽出される．これら k-部分文字列の集合を正の部分文字列集合と呼ぶ．ま

♥ 学生会員 広島市立大学大学院情報科学研究科
{kato,kotaro}@db.its.hiroshima-cu.ac.jp
♦ 正会員 広島市立大学大学院情報科学研究科
{mori, kuroki, kitakami}@its.hiroshima-cu.ac.jp

た，ギブスサンプリングで抽出されなかった部分，つまり，正の部分文字列集合以外の集合を負の部分文字列集合と呼ぶ．これらの間の関係を図 1 に示す．

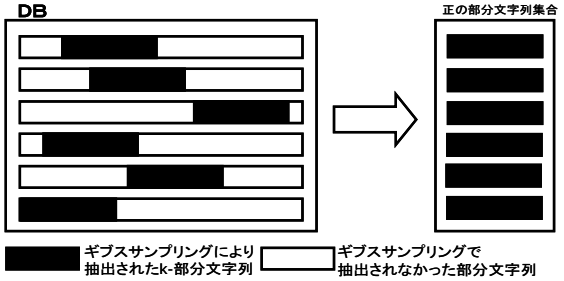

図 1:正の部分文字列集合と負の部分文字列集合
Fig.1 Both of the positive and negative sets

正の部分文字列集合及び負の部分文字列集合からそれぞれ抽出されるパターンを，正及び負のパターンと呼ぶ．

3. 従来手法

射影データベースを用いたパターン成長アプローチにおける可変長パターン抽出法は，スコープデータベース[2]という手法が提案されている．スコープデータベースは，従来の射影データベースに含まれるスキャン開始位置の情報に加えて，ユーザにより定められた参照範囲の情報と，それまでに求まった，可変長の k-頻出パターンに対する全ての k-ストリングの位置情報から構成される．これにより，極小かつ，非冗長な可変長ワイルドカード領域を求めることができる．

以下で，スコープデータベースによる頻出パターンの抽出手順について述べる．

(1) 入力パラメータとして，最小支持数を与える．さらに，ワイルドカード数を$[0, wc_{max}]$の範囲内から，誤差数を$[0, \varepsilon_{max}]$の範囲内からそれぞれ選択し，与える．
(2) 配列データベース DB を 1 回スキャンし，1-頻出パターン$\langle pat^1 \rangle$を全て求め，これらを F_1 とする．
(3) $F_k = \phi$ になるまで，各$\langle pat^k \rangle \in F_k$対して，以下の処理を繰り返す．
- $\langle pat^k \rangle$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) $F_k = F_1 \cup F_2 \cup \cdots \cup F_k$を出力する．

上記の処理手順を表 1 の配列データベースに適用すると，図 2 の列挙木が得られる．

表 1:配列データベース
Table 1 An example of a sequence database

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

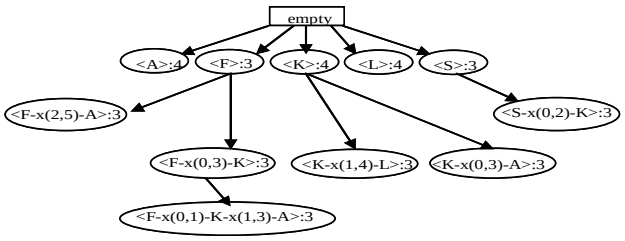

図 2：スコープデータベースにより抽出される頻出パターン
Fig.2 Patterns extracted by scope database

パターン成長アプローチによる可変長頻出パターン抽出法では，精密な表現の頻出パターンを抽出することができるが，一方で，大量の頻出パターンが抽出されるという問題がある．特に可変長のワイルドカード領域をもつモチーフの探索においては，膨大な量の頻出パターンが抽出され，明らかに不必要なパターンが数多く見られる．

4. 不要パターンの除去方式

パターン成長アプローチにより数多く抽出される不要なパターンを除去するために，統計学手法の一つであるギブスサンプリングを用いて，不要なパターンの除去を行う方法について提案する．

4.1 ギブスサンプリング

ギブスサンプリングは，配列データベース DB の各配列から，指定した長さ k をもち，互いにできるだけ類似している部分文字列集合(k-部分文字列集合)を求めることができる．ギブスサンプリングのアルゴリズム[3][4]を図 3 に，ギブスサンプリングの抽出処理の例を図 4 に示す．なお，図 4 中の各番号は図 3 のアルゴリズムの各ステップ番号に対応する．

① t 本の配列をもつ DB の各配列に対して，k-部分文字列の開始点 st_i をランダムに選び，それらを配列順に並べた集合を $S = \{s_1, s_2, \cdots, s_t\}$とする．
② DB からランダムに 1 つの配列 Z を選択する．
③ 配列 Z 以外の残りの $t-1$ 本の配列から取り出される k-部分文字列集合から各文字の各位置 i での出現確率 A_i を計算する．
④ DB の k-部分文字列集合に含まれない部分から各文字 a の背景的出現確率 Q_a を計算する．
⑤ Z 上の各位置 i を k-部分文字列を開始点とし，それぞれについて，k-部分文字列の評価値 P_i を出現確率と背景的出現確率を用いて計算する．
⑥ $\{P_1, P_2, \ldots, P_n\}$の各評価値の中から，ランダムに P_j を選択し，P_j に対応する k-部分文字列の配列上の新しい開始点 st'_j を選ぶ．ただし，P_j はできるだけ値が大きいものが選ばれるものとする．
⑦ 収束するまで 2～6 を繰り返す．

図 3:ギブスサンプリングのアルゴリズム
Fig.3 The Gibbs sampling algorithm

ギブスサンプリングにより切り出された k-部分文字列の集合の各要素は，お互いに類似しているが，どのような規則性があるのか明確ではないという問題がある．

4.2 提案方法

従来手法及び，ギブスサンプリングのそれぞれが抱えている問題点を相互に補完するために，以下の処理手順を提案する．

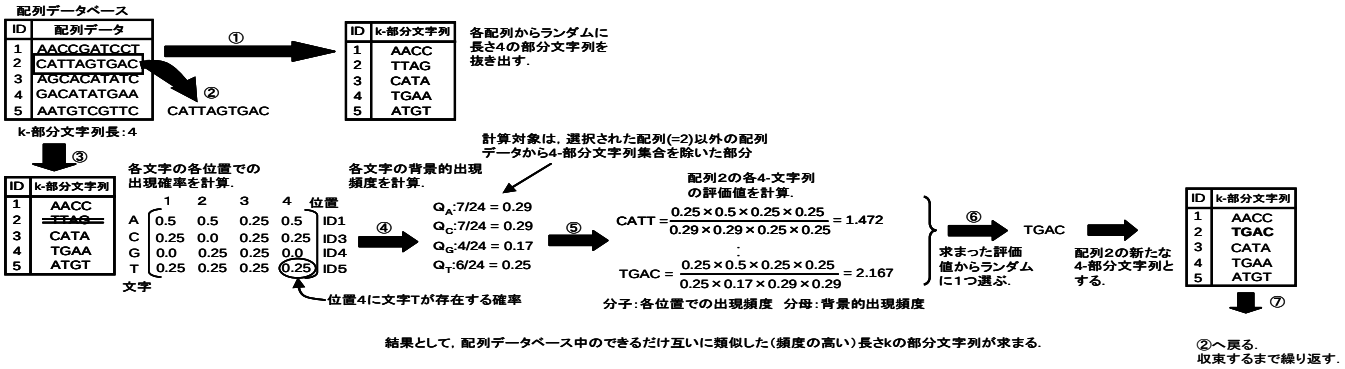

図 4:ギブスサンプリングによるパターン抽出処理の例

Fig.4 An example of the pattern extraction processing by Gibbs sampling

(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，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，配列データベースの全領域にスコープデータベースの方法を適用するよりも，着目領域が限定されるので，不要な頻出パターンを排除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，よりモチーフの形式に近い頻出パターンの抽出が期待できる.

5. 評価

ここでは，従来手法と提案手法よる頻出パターン抽出の計算結果を比較する．性能評価のために使用したデータセットは，Leucine Zipper モチーフ及び Zinc Finger モチーフを含む配列データベースである．なお，ここでは，ギブスサンプリングにより抽出された正の部分文字列集合に対して，従来手法を適用した結果について述べる.

5.1 Leucine Zipper データセット

ここでは，PROSITE から Leucine Zipper モチーフを含むデータセットを選択するために,登録番号として PS00036 を用いた．このデータセットには，125 件の配列データが含まれており，47250 文字で構成されている．また，Leucine Zipper モチーフの形式は,〈[*KR*]-*x*(1,3)-[*RKSAQ*]-*N*-*x*(2)-[*SAQ*](2)-*x*-[*RKTAENQ*]-*x*-*R*-*x*-[*RK*]〉である．この配列データベースから上記のモチーフを抽出するために，入力パラメータを，最大ワイルドカード数,最大誤差数をともに 2 とした.表 2 は，従来手法による頻出パターン抽出結果を，表 3 は提案手法による頻出パターン抽出結果を表している．ただし，表 3 中の *k* は，ギブスサンプリングで抽出する正の部分文字列集合の文字列の長さ，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．また，最小支持率は，抽出されるワイルドカード領域を固定長ワイルドカード領域に限定した際にモチーフが抽出される直前付近の支持率から設定した．なお，Leucine Zipper モチーフは，ワイルドカード領域を含めると最大で 16 文字であるため，*k* の値を 16 の倍数に設定し，実験を行った.

k の値を 16 としたところ，抽出された頻出パターン中にモチーフを見つけることはできなかった.次に *k* を 32 としたところ，抽出される頻出パターン数が約 1/3 に減少しているにもかかわらず，各支持率において，提案手法と同じ数のモチーフを得ることができた．さらに，*k* の値を増加させても同様の結果が得られた.

表 2:Leucine Zipper の計算処理(従来手法)

Table2 Calculation result of a dataset that includes the Leucine Zipper motif (existing method)

比較項目/最小支持数	37%	36%	30%	25%
頻出パターン数(件)	52129	58029	205808	711663
モチーフ数(件)	6	6	11	11
計算時間(秒)	57.46	63.99	341.69	2141.13

表 3:Leucine Zipper の計算処理(提案手法)

Table3 Calculation result of a dataset that includes the Leucine Zipper motif (proposed method)

比較項目/最小支持数		37%	36%	30%	25%
k = 16	頻出パターン数(件)	1539	1726	3572	5274
	モチーフ数(件)	0	0	0	0
	計算時間(秒)	0.43	0.45	0.72	1.03
k = 32	頻出パターン数(件)	14034	16280	93638	243616
	モチーフ数(件)	6	6	11	11
	計算時間(秒)	16.16	19.31	169.81	466.55
k = 64	頻出パターン数(件)	18878	22037	129736	566330
	モチーフ数(件)	6	6	11	11
	計算時間(秒)	25.07	28.93	262.15	1628.88
k = 128	頻出パターン数(件)	21611	24924	137137	579804
	モチーフ数(件)	6	6	11	11
	計算時間(秒)	26.10	30.53	252.12	1666.38

5.2 Zinc Finger データセット

ここでは，PROSITE から Zinc Finger モチーフを含むデータセットを選択するために,登録番号として PS00028 を用いた．このデータセットには 744 件の配列データが含まれており，426011 文字で構成されている．また，Zinc Finger モチーフの形式は,〈*C*-*x*(2,4)-*C*-*x*(3)-[*LIVMFYWC*]-*x*(8)-*H*-*x*(3,5)-*H*〉である．この配列データベースから上記のモチーフを抽出するために，入力パラメータを，最大ワイルドカード数を 8，最大誤差数を 2 とした．表 4 に従来手法による頻出パターン抽出結果を，表 5 に提案手法による頻出パターン抽出結果を示す．最小支持率は，Leucine Zipper モチーフと同様に，抽出されるワイルドカード領域を固定長ワイルドカード領域に限定した際にモチーフが抽出される直前付近の支持率から設定した．なお，Zinc Finger モチーフは，ワイルドカード領域を考慮すると最大で 25 文字であるため，*k* を 25 及び 50

として実験を行った.

kの値が 25 及び 50 の場合,支持率が 90%のときは抽出された頻出パターン中にモチーフを見つけ出すことはできなかったが，支持率 80%以下では提案手法と同じ数のモチーフを見つけることができた．さらに，k を 100 としたところ，支持率 90%でも従来手法と同じ数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約 1/6 に減少している.

表 4:Zinc Finger の計算処理(従来手法)
Table4 Calculation result of a dataset that includes the Zinc Finger motif (existing method)

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 5:Zinc Finger の計算処理(提案手法)
Table5 Calculation result of a dataset that includes the Zinc Finger motif (proposed method)

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 = 50	頻出パターン数(件)	65	463	3633
	モチーフ数(件)	0	9	9
	計算時間(秒)	0.92	5.11	31.62
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
k = 150	頻出パターン数(件)	233	2559	77727
	モチーフ数(件)	9	9	9
	計算時間(秒)	10.20	93.17	1893.55

5.3 考察

前述の 2 つのデータセットの実験結果について考察する．両データセットともに，ギブスサンプリングを適用しない場合,頻出パターンの探索範囲はデータセット全体であるため，抽出される頻出パターン数が多くなっている．一方，ギブスサンプリングを適用した場合，データセットの着目範囲が限定される．例えば，Leucine Zipper モチーフを含むデータセットで,kの値が 16 のとき,着目範囲の文字数は 16×125 で，2000 文字となり，ギブスサンプリングを適用しない場合に対して，探索範囲が約 4%に減少しているため，抽出される頻出パターン数が減少している．それにも関わらず，ギブスサンプリングを適用しない場合と同数のモチーフを発見できている．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.

6. まとめ

本論文では，ギブスサンプリングを用いてパターン成長アプローチによって抽出される不要なパターンの除去を行った．Leucine Zipper モチーフ及び Zinc Finger モチーフを含むデータセットを PROSITE から取り出し実験を行った．ギブスサンプリングで抽出された長さ k の部分文字列集合に対して，スコープデータベースを適用した結果，頻出パターン中に含まれるモチーフ数を減少することなく，抽出される頻出パターン数を,Leucine Zipper データセットで 1/3 に,Zinc Finger データセットで 1/6 まで減少することに成功した.

今後の課題として，負パターン集合用いて頻出パターンを削減する方法の検討が残されている．そのためには，配列データベース中のパターンの存在位置の解析が重要である.

[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）の支援により行われた.

[文献]

[1] PROSITE：http://kr.expasy.org/prosite

[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

[3] 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.

[4] Liu, J.S., Neuwald, A.N. and Lawrence,C.E.:Bayesian models for multiple local sequence alignment and Gibbs sampling strategies, JASA, 90, 1156-1170, 1995.

加藤 智之 Tomoyuki KATO
広島市立大学大学院 情報科学研究科博士前期課程在学中．2006 広島市立大学情報科学部卒業．日本データベース学会，電子情報通信学会 各学生会員.

森 康真 Yasuma MORI
広島市立大学大学院 情報科学研究科助教．1994 北陸先端科学技術大学院大学情報科学研究科博士前期課程修了．データベースシステムの研究・開発に従事．情報処理学会，人工知能学会，日本データベース学会，ACM 各会員.

荒木 康太郎 Kotaro ARAKI
広島市立大学大学院 情報科学研究科博士前期課程在学中．2006 広島市立大学情報科学部卒業．日本データベース学会，情報処理学会 各学生会員.

黒木 進 Susumu KUROKI
広島市立大学大学院 情報科学研究科准教授．1990 東京大学大学院工学系研究科修士課程修了．博士(工学)．空間データベースの研究に従事.日本データベース学会,情報処理学会,電子情報通信学会，ACM, IEEE 各会員.

北上 始 Hajime KITAKAMI
広島市立大学大学院 情報科学研究科教授．1976 東北大学大学院工学研究科博士前期課程修了．博士(工学)．データベースおよび人工知能などの研究開発に従事．電子情報通信学会データ工学ワークショップ DEWS2007 in Hiroshima 組織委員，日本データベース学会 BI 研究グループ運営委員，人工知能学会評議員，情報処理学会（TOM）論文誌編集委員，IEEE および ACM 各会員.