73th SIGKBS原稿拡大版 (第73回人工知能学会知識ベースシステム研究会, 平成18年3月, 大阪大学)

深さ優先探索に基づく変数制限つき極大モチーフの高速マイニング

A Depth-First Algorithm for Efficient Discovery of Maximal Motif with Maximum Gap Length Constraint

有村 博紀 [1*]　宇野 毅明 [2]
Hiroki Arimura[1] and Takeaki Uno[2]

[1] 北海道大学大学院情報科学研究科
[1] Graduate School of Information Science and Technology, Hokkaido University
[2] 国立情報学研究所
[2] National Institute of Informatics

Abstract: In this paper, we study the problem of discovering all maximal motifs in a sequence, which is considered in bioinformatcs area since 2000. In the previous paper, we developed an efficient maximal motif discovery algorithm based on depth-first search over the hypothesis space for the class of repeated motifs with wildcards. In this paper, we extend this algorithm for the class of motifs with wildcards with a constraint on the maximum number of consecutive occurrences of wildcards. The resulting algorithm is a fast and lightweight enumeration algorithm, which has polynomial delay and polynomial space complexities.

1　導入

パターン発見は，与えられた族の**組合せパターン**から，入力データ中に出現し指定された制約を満たすような全てのパターンを見つけることである．パターン発見は，計算生物学や，時系列解析，テキストマイニングなどの分野において，中心的な役割を果たしており，1990年代後半以来，集合や，系列，木，グラフ等のさまざまな組合せパターンの族に対して，効率よいパターン発見アルゴリズムの研究が行われている．

文字ワイルドカードをもつ**モチーフ**(motif) とは，定数文字と**文字ワイルドカード**（文字変数）$\circ$ からなる $x = \mathtt{AB} \circ \mathtt{B} \circ \mathtt{BC}$ のような文字列である．**最大変数長制限付きモチーフ**(motif with maximum gap length) とは，非負整数 $K \geq 0$ に対して，高々K 個より真に多い連続した変数出現をもたないモチーフである．この最大変数長制限は，計算生物学等における応用で有用な制約である．

本稿では，文字ワイルドカードをもつモチーフの族に対する**極大モチーフ列挙問題**(maximal motif enumeration problem) を考察する．モチーフ x が，入力文字列 s 上の**極大モチーフ**であるとは，x が s 上に指定回数 θ 回以上出現し，しかも s 上で同じ出現リストを持つようなより長いモチーフに含まれないことをいう．極大モチーフ列挙問題は，重複をさけてすべての極大モチーフを計算する問題である．これは，**頻出閉集合発見** (closed itemset discovery) の文字列版にあたる問題である．

この極大モチーフ発見問題は，他の多くのデータマイニング問題と同様に，解であるモチーフの総数が一般に入力長の指数個になる**列挙問題**である．そこで列挙アルゴリズムの計算量として，解一個あたりの計算時間と記憶領域量を考えるのが普通である．ワイルドカードをもつモチーフの族に対して，極大モチーフ列挙問題が効率よい列挙アルゴリズム，すなわち出力多項式時間アルゴリズムを持つかどうかは，2000 年の [6] らによる問題の導入以来の未解決問題であった．これに対して，2005 年に Arimura と Uno [2] は，深さ優先探索に基づく多項式遅延多項式領域列挙アルゴリズム MaxMotif を与えて，この問題が出力多項式時間計算可能であることを示した．これは，解一個あたりの計算時間と使用領域の両方が，入力のみの多項式でおさえらえる軽量高速アルゴリズムである．

*連絡先：北海道大学大学院情報科学研究科
〒 060-0814 札幌市北区北 14 条西 9 丁目
E-mail: arim@ist.hokudai.ac.jp

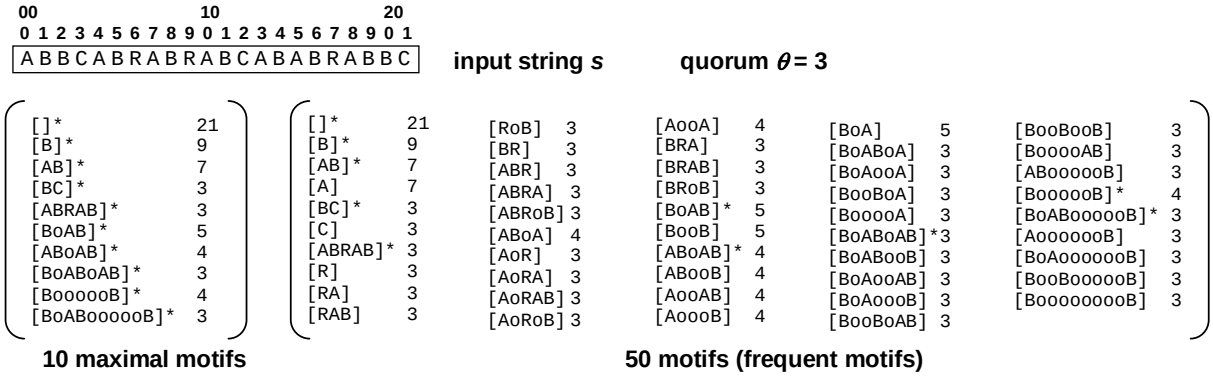

図 1: Examples of maximal motifs (left) and motifs (right) for an input string s and a quorum θ, where * indicates a maximal motif and the number associated to each motif indicates its frequency. These 10 maximal motifs are representatives containing the whole information on the occurrences of all motifs in s.

1.1 本稿の目的

本発表の目的は次の二点である．第一に，MaxMotif アルゴリズムの基本的アイディアと実装の詳細について述べることである．MaxMotif アルゴリズムは，発見対象となる極大モチーフからなる仮説空間上の深さ優先探索に基づいており，閉集合発見のための LCM アルゴリズムで開発された PPC 拡張 (Prefix-Preserving Closure Extension) を用いて，非極大な頻出モチーフを列挙することなく，すべての極大モチーフを入力の多項式時間（多項式遅延）と多項式領域で高速に発見する．

第二に，最大変数長制限付きモチーフの族 $\mathcal{M}_k$ ($k \geq 0$) における極大モチーフ発見問題の効率よい解法について議論する．最大変数長制限つき極大モチーフは，制限無しの極大モチーフと異なる性質をもっており，アルゴリズム MaxMotif で出力時に最大変数長に関する検査を行うだけの素朴な方法では正しい結果が得られない．そのため，MaxMotif アルゴリズムを拡張し，最大変数長制限付きモチーフの族における極大モチーフ発見問題に対する効率よいアルゴリズム BoundedGapMaxMotif を与える．次は，本稿の主結果である．

定理 1 (主結果) 入力文字列を $s \in \Sigma^*$ とおく．最大変数長 $k \geq 0$ のワイルドカードつきモチーフの族 $\mathcal{M}_k$ に対して，アルゴリズム BoundedGapMaxMotif は，モチーフ一つあたり $O(mn^2)$ 時間と $O(n)$ 領域で，s 上のすべての最大変数長制限 k つき極大モチーフを重複なしに発見する．ここに，$n = |s|$, $m = |\mathcal{L}(x)|$ は列挙される極大モチーフ x の出現リストの長さである．

最大変数長制限付きモチーフの族における極大モチーフ発見問題に対する多項式遅延多項式領域列挙アルゴリズムが存在する．

極大モチーフの例: 図 3 の上方に，入力文字列 s と最小頻度パラメータ $\theta = 3$ に対するモチーフと極大モチーフの例を示す．例えば，R∘B と ABRAB は，s 上で同じ出現リスト $\{6, 9, 17\}$ をもつ頻出モチーフであるが，前者は極大であるのに対して，後者はそうでない．長さ 20 文字の入力に対して，頻出モチーフ（図下右）は全部で 50 個あるのに対して，極大モチーフ（図下左）は 10 個である．

関連研究: 文字ワイルドカードをもつモチーフの族に対する極大モチーフ発見問題は，2000 年に Parida らによって導入された [6]．Parida ら [7] は，始めに文献 [6] のアルゴリズムを用いて入力文字列 s から, 非冗長モチーフという極く少数のモチーフを取り出し，次に非冗長モチーフの集合から全ての極大モチーフを生成するという二段階アルゴリズムを提案し，この問題の出力多項式時間計算可能性を主張した．この結果の正当性は，文献 [6] の非冗長モチーフの総数が入力長 n と最小頻度 θ の多項式個（線形）であるという主張に基づいていた．しかし，2003 年に Pisanti ら [10] が非冗長モチーフの総数は最悪時に θ の指数となることを示し，[6] の証明に誤りがあることを指摘した．そのため，文字ワイルドカードをもつモチーフの族に対する極大モチーフ発見問題はふたたび未解決問題となった．これに対して，2005 年に Arimura と Uno [2] は，深さ優先探索に基づく多項式遅延多項式領域列挙アルゴリズム MaxMotif を与え，この問題が出力多項式時間計算可能であることを示し，この問題を肯定的に解決した．本研究は，この結果を最大変数長制限をもつモチーフの族に拡張するものである．

2 準備

2.1 モチーフと出現リスト

自然数全体の集合を $\mathsf{N} = \{0, 1, 2, \ldots\}$ と書き，集合 A の要素数を $|A|$ で表わす．**定数文字**のアルファベットを Σ とおき，$\{\circ\} \notin \Sigma$ を**変数**または**ワイルドカード**という．

長さ n の**入力文字列**は定数文字列 $s = a[1]\dots a[n-1] \in \Sigma^*$ である．**パターン**は，定数文字と変数からなる文字列で，両端が定数記号で始まるもの，すなわち，$\{\varepsilon\} \cup \Sigma \cup (\Sigma \cdot (\Sigma \cup \{\circ\})^* \cdot \Sigma)$ の元をいう．**パターン全体の族**を $\mathcal{P}$ で表わす．パターン x の**最大変数長**とは，x 中の連続した変数 $\circ$ の個数の最大値をいう．正整数 $K \geq 0$ に対して，$\mathcal{P}_k$ で**最大変数長が高々K のパターンの族**を表わす．

$\Sigma \cup \{\circ\}$ 上に，任意の文字 $a \in \Sigma$ に対して $a \preceq a$ かつ $a \preceq \circ, \circ \preceq \circ$ が成立するように，半順序 $\preceq$ を導入する．次に二項関係 $\preceq$ を，パターンと文字列上に拡張する．**文字列 x が位置 $0 \leq p \leq n-1$ で文字列 y に照合する**とは，任意の $i = 0, \dots, m-1$ に対して，$x[i] \preceq y[p+i]$ が成立することをいう．このとき，$x \preceq y$ と書き，x が y に**含まれる**，または x が y を**包摂する**という．添字 p を x の y 中の出現位置という．関係 $\preceq$ は $\mathcal{P}$ 上の半順序であることは容易にわかる．

パターン x の**出現リスト**とは，集合 $\mathcal{L}(x) = \{\, 0 \leq p \leq |s|-1 \mid p \text{ は } x \text{ の } y \text{ 中の出現位置} \,\}$ である．出現リスト $\mathcal{L} \subseteq \{0, \dots, |s|-1\}$ と整数 $d \in \mathsf{N}$ に対して，$\mathcal{L}$ の**幅 d の並行移動**を $\mathcal{L} + d = \{\, \ell + d \mid \ell \in \mathcal{L} \,\}$ と定義する．二つのパターン x, y が s に関して**等価**である（$x \equiv_s y$ と書く）とは，ある $d \in \mathsf{N}$ に対して，$\mathcal{L}(x) = \mathcal{L}(y) + d$ が成立することを言う．明らかに，等価関係は $\mathcal{P}$ 上の同値関係である．

2.2 極大モチーフ発見問題

パターン x と非負整数 $0 \leq \theta \leq |s| - 1$ に対して，$|\mathcal{L}(x)| \geq \theta$ ならば，x は s 上の θ-**モチーフ**（または単に**モチーフ**）という．

定義 1 (Parida *et al* [6], Pisanti *et al.* [10]) モチーフ x が s 上の**極大モチーフ**(maximal motif) であるとは，x を含む任意のモチーフ $y \in \mathcal{P}$ $(x \preceq y)$ に対して，$\mathcal{L}(y) = \mathcal{L}(x) + d$ となる $d \in \mathsf{N}$ が存在しないことを言う．

定義 2 パターンの任意の族を $\mathcal{C} \subseteq \mathcal{P}$ とおく．モチーフ $x \in \mathcal{C}$ が $\mathcal{C}$ **に関する** s **上の極大モチーフ**(maximal motif on s w.r.t. $\mathcal{C}$) であるとは，x を含む任意のモチーフ $y \in \mathcal{C}$ $(x \preceq y)$ に対して，$\mathcal{L}(y) = \mathcal{L}(x) + d$ となる $d \in \mathsf{N}$ が存在しないことを言う．

最大変数長 k つき極大モチーフとは，族 $\mathcal{P}_k$ に関する極大モチーフである．$\mathcal{M}$ と $\mathcal{M}_k$ で，それぞれ，極大モチーフの族と最大変数長 k つき極大モチーフの族を表わす．[1]

[1]等価な言い方では，極大モチーフ x は $\equiv_s$ に関する同値類 $\{\, y \in \mathcal{P} \mid x \equiv_s y \,\}$ の半順序 $\preceq$ に関する極大元である．極大モチーフは同値類の一意な最大元であることが保障されるが，最大変数長 k つき極大モチーフはそうではない．

モチーフの族 $\mathcal{C} \subseteq \mathcal{P}$ に対する**極大モチーフ列挙問題**(maximal motif enumeration problem) は，長さ n の入力文字列 $s = s[0] \cdots s[n-1] \in \Sigma^*$ と最小頻度 $0 \leq \theta \leq n$ を受け取り，$\mathcal{C}$ のすべての極大モチーフを重複なく見つける問題である．

3 マージ演算と極大モチーフの特徴づけ

極大モチーフ発見において鍵となる演算が，パターンのマージ (merge) $\oplus$ である [1, 3, 10, 9]. 帰納学習における最小汎化 [12] と密接な関係をもつ．まず有限長の文字列を $x \in (\Sigma \cup \{\circ\})^*$ を，関数 $\lfloor x \rfloor : dom(s) \to \Sigma$ と同一視する．ここに，集合 $dom(s) = \{0, \dots, |s|-1\}$ を x の領域といい，任意の $i \in dom(s)$ に対して $\lfloor x \rfloor(i) = x[i]$, 任意の $i \notin dom(s)$ に対して $\lfloor x \rfloor(i) = \circ$ と定義する．これは文字列の両側に無限個の $\circ$ を追加した**無限文字列**を考えることに等しい．逆に，無限文字列 $x : dom(s) \to \Sigma$ に対して，$\lceil x \rceil$ で，x の**両端から無限個の連続した $\circ$ を除去して得られる有限文字列**を表わす．もし x が定数文字を含まなければ，$\lceil x \rceil = \varepsilon$ である．

次に任意の整数 $d \in \mathsf{N}$ に対して，x **を正方向に d だけずらして得られる文字列**を $x + d$ と書く．ここに，任意の $i \in \mathsf{N}$ に対して，$\lfloor x + d \rfloor(i) = \lfloor x \rfloor(i - d)$ が成立する．したがって，$x + d$（または $x - d$）は，それぞれ x の原点を負方向に（正方向に）幅 d だけ移動して得られる文字列である．例えば，文字列 $s = \mathtt{EABCDABED}$ に対して，右への 2 文字移動して $(\lfloor s \rfloor + 2) = \cdots\circ\,\mathtt{EA}\downarrow\mathtt{BCDABED}\circ\cdots$ であり，$\lceil (\lfloor s \rfloor + 2) \rceil = \mathtt{EABCDABED} = \mathtt{s}$ である．ここに添え字 0 の位置を $\downarrow$ で示す．

今，文字のマージを，文字 $a, b \in \Sigma \cup \{\circ\}$ に対して，$a \oplus a = a$ とし，$a \neq b$ ならば $a \oplus b = \circ$ と定義する．このとき，**文字列 x, y のマージ** $x \oplus y$ を，任意の $i \in \mathsf{N}$ に対して $(x \oplus y)[i] = x[i] \oplus y[i]$ を満たす無限長文字列として定義する．マージ演算の例をあげる．

$$\begin{array}{rl} x_1 = & \mathtt{AB}\downarrow\mathtt{BCABRABRA} \\ x_2 = & \mathtt{CA}\downarrow\mathtt{BRABCABCA} \\ \hline x_1 \oplus x_2 = & \circ\circ\downarrow\mathtt{B}\circ\mathtt{AB}\circ\mathtt{AB}\circ\circ \end{array}$$

定義 3 出現リスト $\mathcal{L} \subseteq \{0, \dots, n-1\}$ のマージを $\bigoplus \mathcal{L} = \lceil \bigoplus_{d \in \mathcal{L}} (s - d) \rceil \in \mathcal{P}$ と定義する．

モチーフ $x \in \mathcal{P}$ に対して，$Clo(x) = \bigoplus \mathcal{L}(x) \in \mathcal{P}$ を，s における x の**閉包**(closure) という．

補題 2 任意の x, y と任意の出現リスト $\mathcal{L}, \mathcal{L}'$ に対して，次の性質が成立する．

1. $\mathcal{L}(xy) = \mathcal{L}(x) \cap (\mathcal{L}(y) - |x|)$
2. $\mathcal{L} \supseteq \mathcal{L}'$ ならば $\bigoplus \mathcal{L} \preceq \bigoplus \mathcal{L}'$.
3. 任意の整数 $d \in \mathsf{N}$ に対して，$\bigoplus \mathcal{L} = \bigoplus (\mathcal{L} + d)$.
4. $x \preceq Clo(x)$

Algorithm MaxMotif(θ: 最小頻度, s: 入力文字列)
1 Expand($\bot, \theta, s$);

Proc. Expand(x: モチーフ, $\theta \geq 0$, s : 入力文字列)
2 **if** $|\mathcal{L}(x)| < \theta$ **then return**;
3 **output** x;
4 **for** $k := core\_i(x) + 1$ **to** $|s|$ **do**
5 **foreach** $c \in \Sigma$ **do begin**
6 $y = Clo(x\langle k \leftarrow c\rangle)$;
7 **if** $x[0..k-1] = y[0..k-1]$ **then**
8 **call** Expand(y, s, θ);
9 **end for**

図 2: A polynomial space polynomial delay enumeration algorithm for $\mathcal{M}$.

5. $x \preceq y$ ならば $Clo(x) \preceq Clo(y)$
6. 閉包 $Clo(x)$ は，x を含む極大モチーフで，$\preceq$ に関して最小のものである．

補題 3 任意のモチーフ x に対して (i)–(iii) は等価:

(i) x は極大モチーフ.
(ii) ある出現リスト $\mathcal{L} \subseteq \{0, \ldots, n-1\}$ に対して，$|\mathcal{L}| \geq \theta$ かつ $x = \bigoplus \mathcal{L}$.
(iii) $x = Clo(x)$.

極大モチーフに対しては，モチーフの包摂関係と出現リストの包含関係が一致する：すなわち，極大モチーフ $x, y \in \mathcal{M}$ に対して，$x \preceq y \iff \exists d \in \mathsf{N}\,[\,\mathcal{L}(x) \supseteq \mathcal{L}(y) + d\,]$ が成立する．以上の議論から，族 $\mathcal{P}$ に関して出現リストの等価性に関する同値類 $[x]_{\equiv_s}$ は，代表元として一意な極大モチーフをもつこともわかる．

4 アルゴリズム MaxMotif

4.1 アルゴリズムの概要

図 2 に，極大モチーフ発見アルゴリズムを解く多項式遅延多項式領域アルゴリズム MaxMotif の概略を示す．このアルゴリズムは最小の極大モチーフ $\bot$ からスタートして，極大モチーフ全体 $\mathcal{M}$ を含む全域木上を，バックトラックを用いて深さ優先探索し，入力文字列 s 中のすべての極大モチーフを重複なく発見する．以下では，アルゴリズムの実現のための詳細を述べる．

4.2 極大モチーフの全域探索木

本節では，極大モチーフ全体 $\mathcal{M}$ の全域木 $\mathcal{M}$ を構築する．初めに，$\bot = Clo(\varepsilon)$ を**根**と定義する．これは，$\preceq$ に関する $\mathcal{M}$ の一意な最小元である．

次に，任意の根でない極大モチーフ $y \in \mathcal{M}$（**子**という）に対して，その一意な**親** $x = \mathcal{P}(y)$ を，y の末尾から順に文字列を削っていき，その出現リストが初めて変化した時点で，その閉包をとって得られるパターンを割り当てる．

より正確には次のように定義する：まず子モチーフ y の**核添え字**(core index) を，$core\_i(y) = \min\{\, 0 \leq \ell \leq |y| - 1 \mid \mathcal{L}(x) = \mathcal{L}(y) \,\}$ と定義する．ここで，y のコア添字を $\ell = core\_i(y) \geq 0$ とする．このとき，$\mathcal{M}$ の探索木における y の**親**(parent) をパターン $\mathcal{P}(y) \stackrel{\text{def}}{=} Clo(\lceil y[0..\ell-1] \rceil)$ と定義する．このとき，$\mathcal{P}(y)$ は常に一意に定まり，極大モチーフとなる．

$\mathcal{M}$ を頂点集合とし，$\mathcal{P}$ を有向辺集合，$\bot$ を根とする有向グラフを $\mathcal{T} = (\mathcal{M}, \mathcal{P}, \bot)$ と書くと，次の定理が成立する．

定理 4 $\mathcal{T} = (\mathcal{M}, \mathcal{P}, \bot)$ は M の全域木である．

これによって，意図どおり $\bot$ を根とし，極大モチーフだけを結ぶ全域木 $\mathcal{T}$ を構成することができた．

4.3 PPC 拡張

先の全域木 $\mathcal{T}$ では，子から親への辺で木が表現されているため，これを直接に深さ優先探索に利用することはできない．そこで，この辺を逆向きに変えて，親から子への移動を可能にする必要がある．

モチーフ $x \in \mathcal{P}$ に対する**代入**とは，対 $\xi = \langle i \leftarrow c\rangle \in \{0, \ldots, |x|-1\} \times \Sigma$ である．ただし，$x[i] = \circ$ は変数と仮定する．代入 ξ の x への**適用** $x\xi$ とは，パターン x で位置 i に文字 c を代入して得られるパターン，すなわち，$x\xi[i] = c$ かつ，任意の $j \neq i$ に対して $x\xi[j] = x[j]$ となるパターン $x\xi$ である．例えば，$x = \mathtt{BA} \circ \mathtt{B}$ の場合に，$x\langle -1 \leftarrow \mathtt{C}\rangle = \mathtt{CBA} \circ \mathtt{B}$ であり，$x\langle 2 \leftarrow \mathtt{C}\rangle = \mathtt{BACB}$, $x\langle 6 \leftarrow \mathtt{C}\rangle = \mathtt{BA} \circ \mathtt{B} \circ \circ \mathtt{C}$ である．

定義 4 (ppc-extension) 空でない任意の極大モチーフ x $(y \neq \bot)$ に対して，モチーフ y が x の**接頭辞保存閉包拡張**(PPC 拡張，prefix-preserving closure expansion) であるとは，次の (i)–(iii) が成立することである．

(i) ある代入 $\xi = \langle k \leftarrow c\rangle$ に対して, $y = Clo(x\langle k \leftarrow c\rangle)$ が成立する．
(ii) 添え字 k が $k > core\_i(x)$ を満たす．
(iii) $x[0..k-1] = y[0..k-1]$ が成立する．すなわち長さ $i-1$ の接頭辞は変化しない．

微妙な問題として，もし $|x| < k$ の場合は，$x[0..k-1]$ は x の右側に必要なだけ変数 $\circ$ を詰めて得られる文字列，つまり $\lfloor x \rfloor[0..k-1]$ だと定義する．このとき，代入された位置 k が新しい子モチーフの核添え字になることが証明できる．次が成立する．

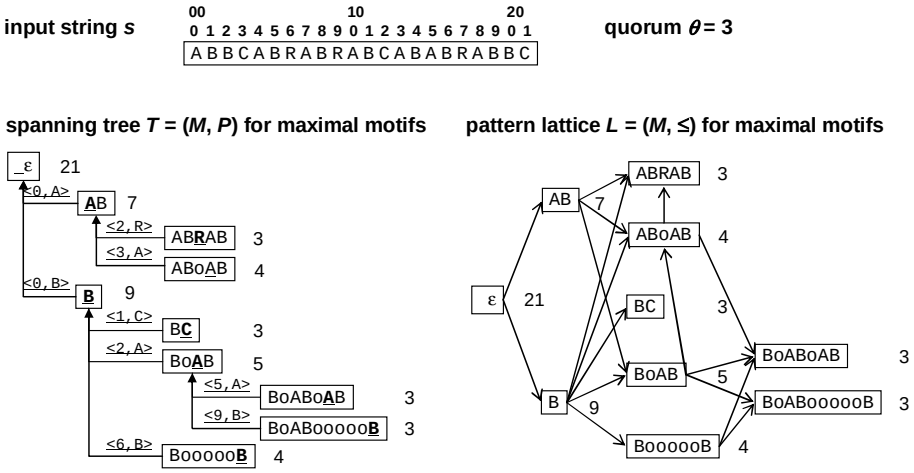

図 3: The spanning tree $\mathcal{T} = (\mathcal{M}, \mathcal{P})$ (left) and the pattern lattice $\mathcal{L} = (\mathcal{M}, \preceq)$ (right) for maximal motifs on quorum $\theta = 3$ and the input string s (top).

補題 5 *PPC* 拡張において，親 x と代入 ξ から，対応する子 y と核添え字 $core\_i(y)$ が $O(mn)$ 時間で計算可能である．

補題 6 (PPC 拡張の正当性) 空でない任意の極大モチーフ x *(*$y \neq \bot$*)* に対して，*(i)* $x = \mathcal{P}(y)$ であることと，*(ii)* ある代入 $\xi = \langle k \leftarrow c \rangle$ に対して，$y = Clo(x\xi)$ が x の *PPC* 拡張であることは，同値である．さらに，*(ii)* において，異なる代入 ξ に対しては異なる y が対応する．

以上より，図 2 のアルゴリズムの正当性が示された．計算量の解析とあわせて次の定理が示される．

定理 7 (極大モチーフ列挙，有村・宇野 [2]) $K \geq 0$ を正整数とし，$s \in \Sigma^*$ を入力文字列とする．最大変数長 K のワイルドカードつきモチーフの族に対して，アルゴリズム MaxMotif は，モチーフ一つあたり $O(mn^2)$ 時間と $O(n)$ 領域で，s 上のすべての極大モチーフを重複なしに発見する．ここに，$n = |s|$, $m = |\mathcal{L}(x)|$ は列挙される極大モチーフ x の出現リストの長さである．

5 最大変数長制限つき極大モチーフの発見

本節では，最大変数長制限つきモチーフ族 $\mathcal{P}_k$ $(k \geq 0)$ に関する，極大モチーフの多項式遅延多項式領域列挙アルゴリズムを与える．任意の正整数を k とおく．

観察: $\mathcal{M}$ の場合と異なり，$\mathcal{M}_k$ の最大変数長制限つき極大モチーフは，$\equiv_s$ に関する同値類の一意な最大元であることが保障されない．例として最大変数長 $k = 2$ の場合を考える．$\mathcal{P}_2$ の三つのモチーフ $x = \mathtt{a} \circ \mathtt{a}$ と，$y = \mathtt{b} \circ \mathtt{b}$，$z = \mathtt{c} \circ \mathtt{c}$ が，入力文字列中 s 中にちょうど 3 文字ずつ間隔をあけて出現すると仮定する

$$x_1 = \texttt{ADADDDBDBDDDCDC}$$
$$x_2 = \texttt{AEAEEEBEBEEECEC}$$
$$x_3 = \texttt{AEAEEEBEBEEECDC}$$

また x, y, z のぞれぞれは，これらの位置にしか出現しないとする．このとき，$\mathcal{L}(x) = \mathcal{L}(y) - 6 = \mathcal{L}(z) - 12$ であり，族 $\mathcal{P}$ に関して x と等価な極大モチーフは $w \stackrel{\text{def}}{=} Clo(x) = Clo(y) = Clo(z) = a \circ a \circ \circ \circ b \circ b \circ \circ \circ c \circ c$ であるが，これは P_2 のパターンでない．族 P_2 を考えたとき，x, y, z の三つが，ずらしに関して出現位置 $\mathcal{L}(x)$ と同じ出現リストをもつ $\mathcal{P}_2$ の極大モチーフである．すなわち，最大変数長制限つきモチーフは一般に複数の等価な極大元をもつ．

5.1 最大変数長制限つき閉包

以上の観察より，最大変数長制限の場合は，制限された閉包演算が必要であることがわかる．非負整数を $k \geq 0$ とおく．s 上の出現リスト $\mathcal{L} \subseteq \{0, \ldots, n-1\}$ に対して，その最大変数長 k のマージを $\bigoplus^{(k)} \mathcal{L} = M[a..b]$ と定義する．ここに，$M \stackrel{\text{def}}{=} \bigoplus_{d \in \mathcal{L}}(s - d)$ であり，$a \leq 0$ と $b \geq 0$ は，それぞれ $M[a..b] \in \mathcal{P}_k$ をみたす最小と最大の添え字と定める．ここで，M は無限文字列であり，マージ後のずらしおよび両端の $\circ$ の除去を行っていないことに注意されたい．

定義 5 (最大変数長制限つき閉包) 非負整数を $k \geq 0$ とおく．最大変数長制限つきモチーフ $x \in \mathcal{P}_k$ に対して，その**最大変数長制限つき閉包**を $Clo(x, k) = \oplus^{(k)}\mathcal{L}(x)$ と定義する．

補題 8 任意の $k \geq 0$ と，任意の最大変数長つきパターン $x, y \in \mathcal{P}_k$，任意の出現リスト $\mathcal{L}, \mathcal{L}'$ に対して，次の性質が成立する．

1. $\mathcal{L} \supseteq \mathcal{L}'$ ならば $\bigoplus^{(k)} \mathcal{L} \preceq \bigoplus^{(k)} \mathcal{L}'$.
2. $x \preceq Clo(x, k)$
3. $x \preceq y$ ならば $Clo(x, k) \preceq Clo(y, k)$
4. 閉包 $Clo(x)$ は，x を含む極大モチーフで，$\preceq$ に関して最小のものである．

変数長制限つきマージ演算 $\bigoplus^{(k)}$ は，ずらしに対して不変でない．すなわち先の観察でみたように，一般のモチーフの場合には，任意の整数 $d \in \mathsf{N}$ に対して $\bigoplus^{(k)} \mathcal{L} = \bigoplus^{(k)}(\mathcal{L}+d)$ となるという性質が成立したが，変数長制限つきモチーフに対しては，これは成立しない．

補題 9 任意のモチーフ x に対して (i)–(iii) は等価:

(i) x は $\mathcal{P}_k$ の極大モチーフ．
(ii) ある出現リスト $\mathcal{L} \subseteq \{0,\ldots,n-1\}$ に対して，$|\mathcal{L}| \geq \theta$ かつ $x = \bigoplus^{(k)} \mathcal{L}$．
(iii) $x = Clo(x,k)$．

証明: (ii)⇒ (i): ある出現リスト $\mathcal{L}$ に対して $x = \bigoplus^{(k)} \mathcal{L}$ と仮定する．$\mathcal{L} = \mathcal{L}(x) = \mathcal{L}(y) + d$ かつ $x \prec y$ となるモチーフ $y \in \mathcal{P}_k$ が存在すると仮定する．このとき，y 中の x の出現位置は d であり，y は x に代入 $i \leftarrow c$ で定数文字を一つ追加して得られると仮定しても一般を失わない．したがって，$y \in \mathcal{P}_k$ より，$i \in \{-k-1,\ldots,|x|+k\}$ であり，各 $p \in \mathcal{L}(x)$ に対して，s の位置 $p+i \in \{p-k-1,\ldots,p+|x|+k\}$ に文字 c が出現する．よって，$y = \bigoplus^{(k)} \mathcal{L}$ であって仮定に反するので，このような y は存在しない．(i)⇒ (iii): 性質から，$x \preceq Clo(x,k)$ であり，ある d に対して $\mathcal{L}(x) = \mathcal{L}(Clo(x,k)) + d$ が成立する．ここで，仮定より x は $\mathcal{P}_k$ の極大モチーフであるから，$x \prec Clo(x,k)$ ではありえない．よって示された．(iii)⇒ (ii): 明らか． ■

5.2 最大変数長制限つき極大モチーフ発見アルゴリズム

まず $\mathcal{M}_k$ に対して，全域探索木 $\mathcal{T}$ を拡張する．非負整数を $k \geq 0$ とおく．

定義 6 任意の根でない極大モチーフ $y \in \mathcal{M}$（**子**という）に対して，その一意な**親** $x = \mathcal{P}(y)$ を，パターン $\mathcal{P}^{(k)}(y) \stackrel{\text{def}}{=} Clo(\lceil y[0..\ell-1] \rceil, k)$ と定義する．ここに，核添え字の定義は前節のものと同一であり，$\ell = core\_i(y) \geq 0$ は y のコア添字である．

補題 9 より，$\mathcal{P}(y)$ は常に一意に定まり，変数長制限 k つきの極大モチーフ，つまり，$\mathcal{M}_k$ の元であることが保障される．次は M_k に対する探索木が存在することをいう．

定理 10 $\mathcal{T}^{(k)} = (\mathcal{M}_k, \mathcal{P}^{(k)}, \bot)$ は M_k の全域木である．

次に，$\mathcal{T}^{(k)}$ 上の深さ優先探索の方法を与える．$\mathcal{M}$ の探索木の場合と同様に，$\mathcal{T}^{(k)}$ の辺を逆向きに変えて，親から子への移動を可能にする．

定義 7 (最大変数長制限つき PPC 拡張) 空でない任意の極大モチーフ x $(y \neq \bot)$ に対して，モチーフ y が x の**接頭辞保存閉包拡張**(PPC 拡張，prefix-preserving closure expansion) であるとは，次の (i)–(iii) が成立することである．

(i) ある代入 $\xi = \langle k \leftarrow c \rangle$ に対して，$y = Clo(x\langle k \leftarrow c \rangle)$ が成立する．
(ii) 添え字 k が条件 $core\_i(x) < k$ かつ条件 $k \leq |x| + k + 1$ を満たす．
(iii) $x[0..k-1] = y[0..k-1]$ が成立する．すなわち長さ $i-1$ の接頭辞は変化しない．

上の定義と，M に対する PPC 拡張（定義 4）との違いは，(ii) 節で条件 $k \leq |x| + k + 1$ が追加されていることである．これにより，拡張で得られる子モチーフが $\mathcal{P}_k$ に属することが保障される．

補題 11 *PPC* 拡張において，親 x と代入 ξ から，対応する子 y と核添え字 $core\_i(y)$ が $O(mn)$ 時間で計算可能である．

補題 12 (最大変数長制限つき PPC 拡張の正当性) 空でない任意の極大モチーフ x *(*$y \neq \bot$*)* に対して，*(i)* $x = \mathcal{P}^{(k)}(y)$ であることと，*(ii)* 最大変数長 k 制限をみたすある代入 $\xi = \langle \ell \leftarrow c \rangle$ に対して，$y = Clo(x\xi, k)$ が x の最大変数長 k 制限つきの *PPC* 拡張であることは，同値である．さらに，*(ii)* において，異なる代入 ξ に対しては異なる y が対応する．

証明: (i)⇒ (ii): 明らかに $x = \mathcal{P}^{(k)}(y)$ ならば $x = \mathcal{P}(y)$ である．$\ell = core\_i(y)$ と $c = y[\ell]$ に対して $x = Clo(y[0..\ell-1], k)$ である．子 y は高々 k 個の連続した変数 $\circ$ しかふくまないので，補題 12 の証明と同様にして，$y = Clo(x\langle \ell \leftarrow c \rangle, k)$ が成立することがわかる．(ii)⇒ (i): x の最大変数長 k 制限つきの PPC 拡張 $y = Clo(x\xi, k)$ において，$y \in \mathcal{P}_k$ なので y は高々 k 個の連続した変数 $\circ$ しかふくまない．さらに，$core\_i(x) < core\_i(y) = \ell$ である．また制限つき PPC 拡張の定義から $y[0..\ell-1] = x[0..\ell-1]$ であるので，$\mathcal{P}^{(k)}(y) = Clo(y[0..\ell-1], k) = Clo(x[0..\ell-1], k)$ が成立する．補題 12 の証明と同様にして，$m = core\_i(x) < core\_i(y) = \ell$ であり，$m < \alpha < \ell$ となる核添え字 α はない．これと，x が極大であることから，$Clo(x[0..\ell-1], k) = Clo(x) = x$ が導かれる．よって，$\mathcal{P}^{(k)}(y) = x$ が成立する． ■

定義 8 (BoundedGapMaxMotif アルゴリズム) 図 2 のアルゴリズム MaxMotif において，PPC 拡張を上で導入した最大変数長制限つき PPC 拡張で置き換えて得られるアルゴリズムを BoundedGapMaxMotif アルゴリズムと呼ぶ．

以上により，本稿の主結果である定理 1 が示される．

定理 1 の証明: 上記の修正によっても，BoundedGapMaxMotif アルゴリズムの計算量は MaxMotif アルゴリズムと同じであることが容易にわかる．したがって，補題 12 の最大変数長制限つき PPC 拡張の正当性から定理が示される．

6 おわりに

本稿では，文字ワイルドカードをもつモチーフの族に対する**極大モチーフ列挙問題**(maximal motif enumeration problem) を考察し，先に著者等が与えた MaxMotif アルゴリズムを拡張し，最大変数長制限付きモチーフの族における極大モチーフ発見問題に対する効率よいアルゴリズム BoundedGapMaxMotif を与えた．これにより，最大変数長制限付きモチーフの族における極大モチーフ発見問題が入力サイズの多項式遅延多項式領域で列挙可能であることが示された．

今後の研究課題として，さまざまな制約のもとでの極大モチーフ発見問題に対して，MaxMotif アルゴリズムに基づいて，効率よい列挙アルゴリズムを与えるための一般的な構成法の確立があげられる．

また，集合や，木，グラフなどのさまざまな組合せパターン族に対する MaxMotif アルゴリズムの拡張も今後の課題である．この方向に関しては，著者等は文献 [3] において，アイテム集合の階層的な一般化であり，同時に順序木と無順序木の自明でない部分族であるような**属性木**(attribute trees) と呼ばれるラベルつき木の族に対して，多項式遅延多項式領域アルゴリズムを与えている．今後はより複雑なグラフ族への拡張を行いたい．

最後に，生物情報学等の応用では，ノイズを含むデータや不適正な仮説空間の問題で，ここで用いたような厳密な閉包計算では有用で明確なモチーフが得られないことがある．ここで考察したような極大モチーフ計算と，確率的モデルをもち多系列解析 [5] との融合も，今後の課題である．

参考文献

[1] A. Apostolico and L. Parida, Compression and the wheel of fortune, In *Proc. the 2003 Data Compression Conference (DCC'03)*, IEEE, 2003.

[2] H. Arimura, T. Uno, A Polynomial Space and Polynomial Delay Algorithm for Enumeration of Maximal Motifs in a Sequence, Proc. ISAAC'05, LNCS 3827, Springer, Dec. 2005.

[3] H. Arimura, T. Uno, An Output-Polynomial Time Algorithm for Mining Frequent Closed Attribute Trees, In *Proc. Inductive Logic Programming*, LNAI 3625, Springer, 1–19, August 2005. (to appear)

[4] H. Arimura, T. Shinohara, S. Otsuki, Finding minimal generalizations for unions of pattern languages and its application to inductive inference from positive data, In *STACS'94*, LNCS 775, Springer-Verlag, 649–660, 1994.

[5] R. Durbin, S. R. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, 1998.

[6] L. Parida, I. Rigoutsos, A. Floratos, D. Platt, and Y. Gao, Pattern discovery on character sets and real-valued data: linear bound on irredundant motifs and effcient polynomial time algorithm, In *Proc. SODA'00*, 297–308, 2000.

[7] L. Parida, I. Rigoutsos, D. E. Platt, An Output-Sensitive Flexible Pattern Discovery Algorithm In *Proc. the 12th International Conference on Combinatorial Pattern Matching (CPM'01)*, LNCS 2089, 131–142, 2001.

[8] N. Pasquier, Y. Bastide, R. Taouil, L. Lakhal, Discovering Frequent Closed Itemsets for Association Rules, In *Proc. ICDT'99*, 398–416, 1999.

[9] J. Pelfrêne, S. Abdeddaïm, and J. Alexandre, Extending Approximate Patterns, In *Proc. CPM'03*, LNCS 2676, 328–347, 2003.

[10] N. Pisanti, M. Crochemore, R. Grossi, and M.-F. Sagot, A basis of tiling motifs for generating repeated patterns and its complexity for higher quorum, In *Proc. MFCS'03*, LNCS 2747, 622–631, 2003.

[11] N. Pisanti, M. Crochemore, R. Grossi, and M.-F. Sagot, A comparative study of bases for motif inference, In *String Algorithmics*, C. Iliopoulos and T. Lecroq editors, KCL publications, 2004.

[12] G. Plotkin, A note on inductive generalization, In B. Meltzer and D. Mitchie, editors, Machine Intelligence, volume 5, 153–163, Edinburgh University Press, 1970.

[13] T. Uno, T. Asai, Y. Uchida, H. Arimura, An efficient algorithm for enumerating closed patterns in transaction databases, In *Proc. DS'04*, LNAI 3245, Springer-Verlag, 16-30, 2004.

[14] X. Yan, J. Han, CloseGraph: Mining Closed Frequent Graph Patterns In *Proc. SIGKDD'03*, 2003.