Vol.2014-BIO-40 No.1
2014/12/18

データ分割を適用したGibbs Sampling Algorithmの拡張によるモチーフ抽出

青木 史林[1,a)]　岡崎 威生[†1]　名嘉村 盛和[†1]

概要：モチーフ抽出問題に対して, 計算コストを削減できる Gibbs Sampling Algorithm(GS) を拡張し, 分割されたデータを利用して初期状態におけるモチーフの影響を抑え, 大域的モチーフを安定的に抽出できる分割型 GS について提案する. さらに, 分割型 GS を用いて抽出された候補モチーフ群に対して, 適合性指標としての尤度を導出する.

Motif extraction with Gibbs Sampling Algorithm using data dividing

Abstract: In order to extract global motifs, we proposed divided Gibbs Sampling Algorithm(GS) prevent influence of initial motifs in subset by data dividing. Futhermore, this paper estimates likelihood of motif appearence frequency as compability criterion of extracted motif candidates by divided GS.

1. はじめに

モチーフとは複数のアミノ酸または塩基によって構成される特徴的なパターンを指し, 複数のアミノ酸における相互作用や DNA への結合など, 生物的に非常に重要な機能を果たす. 未知なモチーフの発見においては, 1 ゲノム配列からのモチーフ抽出は不可能であるため, 同一のモチーフ配列が含まれると推測される複数のゲノム配列の最頻出パターンを抽出することによって推定される. しかし, ゲノム配列における正確な最頻出パターンの抽出は, 入力データサイズに対して膨大な計算量や計算空間を必要とするため困難とされている. この問題に対し, 複数ゲノム配列における最頻出パターンを抽出するための評価値最適化に関する研究が行われている.

Gibbs Sampling Algorithm(GS)[1][2] は, 入力された複数ゲノム配列 (データセット) とモチーフ長に対する類似性の高い部分配列集合を抽出することによってモチーフを推定する手法であり, 1 部分配列集合に関する計算処理をすることで MAFFT[3] のようなアライメント処理に対して計算空間を削減できる利点を持つ. また, GS は入力データセットの各配列における部分配列のみを扱うことで, 一般的なローカルアライメント法とは異なり, 各配列におけるそれぞれのモチーフの位置関係に影響されないモチーフ抽出が実行できる. しかし, 局所的に分布するモチーフ (局所解モチーフ) が多く含まれる場合もあり, GS は初期状態の部分配列集合に分布するモチーフに影響されやすいため, 頻出パターン (大域的モチーフ) を安定的に抽出できない. さらに, GS には部分配列集合内の類似性に対する評価指標 F 値が用いられるが, F 値は部分配列集合における相同性に影響され, 最頻出パターンを持つ部分配列集合が必ずしも最大の F 値を獲得するとは限らない.

これらの問題に対し, 本研究では GS におけるモチーフ抽出精度を改善するため GS にデータ分割を適用し, 初期状態におけるモチーフの影響を抑え, 大域的モチーフの安定的な抽出ができる分割型 GS を提案した. また, 大域的なモチーフが複数存在し, それらのモチーフ長が異なる場合は GS に対して正確なモチーフ長を与えることができない. そこで, 各モチーフを完全に抽出するために, 一般的なモチーフ長に対して十分に大きな部分配列長を与えて分割型 GS を繰り返し実行し, 獲得された部分配列集合からモチーフ配列のみを再抽出し, モチーフ候補群を出力する手法を提案した. さらに, モチーフ候補群に対する適合性指標として, モチーフ出現頻度の尤度関数を導出した.

[1] 琉球大学 理工学研究科 情報工学専攻
Information Engineering Cource, Graduate School of Engineering and Science, University of the Ryukyus
[†1] 現在, 琉球大学 工学部 情報工学科
Presently with Faculty of Engineering, University of the Ryukyus
[a)] k138561@ie.u-ryukyu.ac.jp

2. GS によるモチーフ抽出

GS は, アミノ酸配列や塩基配列で構成されるゲノム配列データセット $S = \{s_1, s_2, \cdots, s_{n-1}, s_n\}$ の各 $s_i = \{a_1^{s_i} a_2^{s_i} \cdots a_{l_{s_i}-1}^{s_i} a_{l_{s_i}}^{s_i}\}$ におけるモチーフ推定のための最頻出パターン抽出を目的とする. GS は, 入力された S とモチーフ長 k によって構成される部分配列集合 $U = \{u_1, u_2, \cdots, u_{n-1}, u_n\}, u_i = \{a_{\eta_{u_i}}^{s_i} a_{\eta_{u_i}+1}^{s_i} \cdots a_{\eta_{u_i}+k-2}^{s_i} a_{\eta_{u_i}+k-1}^{s_i}\}$ を作成し, 部分配列 $u_r (1 \leq r \leq n)$ を確率的な更新を反復的に実行することで頻出パターンを抽出する. 従来 GS の u_r の更新時には遷移指標 E_x^U(式 1) が適用され, 出力される部分配列集合における類似性が F 値 (式 2) として計算される.

$$E_x^U = \prod_{j=0}^{k-1} \frac{C^U_{j+1,a^{s_r}_{x+j}}}{nP^U_{a^{s_r}_{x+j}}} \quad (1)$$

$$F = \sum_{i=1}^{k} \sum_{j=1}^{l_\omega} C^U_{i,a_j^\omega} log \frac{Q^U_{i,a_j^\omega}}{P^U_{a_j^\omega}}, Q^U_{i,a_j^\omega} = \frac{C^U_{i,a_j^\omega} + b^S_{a_j^\omega}}{n - 1 + B} \quad (2)$$

$C^U_{j+1,a^{s_r}_{x+j}}$ は U の $j+1$ 番目の位置におけるアミノ酸配列要素 $a^{s_r}_{x+j}$ の出現数であり, $P^U_{a^{s_r}_{x+j}}$ は背景頻度として適用され, $\overline{U \cup s_r}$ におけるアミノ酸要素 $a^{s_r}_{x+j}$ の出現頻度によって定義される. また, 式 2 において, S を構成する各アミノ酸要素の集合を $\omega = \{a_1^\omega, a_2^\omega, \cdots\}$ とすると, $b^S_{a_j^\omega}$ は C^U_{j+1,a_j^ω} が 0 となるときに, Q^U_{i,a_j^ω} が 0 となることを防ぐために用いられ, 各 $b^S_{a_j^\omega}$ は $f^S_{a_j^\omega} \times B$ によって決定される. ここで,$f^S_{a_j^\omega}$ は, 文字 a_j^ω の S に対する相対出現頻度により決定される. また, B は $\sqrt{n}$ としてよいことがわかっている [4]. 各遷移指標 E_x^U は s_r に対して $l_{s_r} - k + 1$ パターン作成され, 指標値に比例した確率で新たな u_r が決定される.

アルゴリズム 1 にて, 従来 GS におけるモチーフ抽出手続きを示した.

アルゴリズム 1 Gibbs Sampling Algorithm

入力: ゲノム配列データセット $S = \{s_1, s_2, \cdots, s_{n-1}, s_n\}$, モチーフ長 k

出力: 部分配列集合 $U = \{u_1, u_2, \cdots, u_{n-1}, u_n\}$

1: 各 $s_i (1 \leq i \leq n)$ に対して, $u_i \subset s_i$ をみたすモチーフ長 k の部分配列集合 U をランダムに作成する.
2: U に対する更新対象配列として, 配列番号 $r (1 \leq r \leq n)$ をランダムに決定する.
3: $U = U \cap \overline{u_r}$ とし, U におけるアミノ酸要素 a_j^ω の背景頻度 $P^U_{a_j^\omega}$ と, 各位置 $q (1 \leq q \leq k)$ における出現数 C^U_{q,a_j^ω} を求める.
4: s_r に対し, $l_{s_r} - k + 1$ パターンの各部分配列に対する遷移指標群 $E^U = \{E_1^U, E_2^U, \cdots, E^U_{l_{s_r}-k}, E^U_{l_{s_r}-k+1}\}$ を作成する.
5: E^U に対し, 遷移指標値に比例した確率で E_x^U を選択する.
6: $u_r = \{a_x^{s_r} a_{x+1}^{s_r} \cdots a_{x+k-2}^{s_r}, a_{x+k-1}^{s_r}\}$ として部分配列を更新し, $U = U \cup u_r$ とする.
7: U の収束が得られるまで, 2 ～ 6 を繰り返す.

一般的にモチーフは各 s_i に対して非常に配列長が小さいため, 初期状態の U に分布するモチーフの出現数は非常に少ないことが予想される. しかし, 従来 GS は遷移指標 (式 1) の性質上, 初期状態における U に含まれるモチーフに強く影響され, 必ずしも頻出性の高いモチーフを抽出できない問題がある. また, 自然界におけるモチーフには多様なパターンがあり, 入力データセットの配列数によっては $C^{U_i}_{j+1,a^{s_r}_{x+j}}$ が 0 を示すことにより, モチーフが抽出されない場合がある. さらに, F 値はモチーフの相同性やモチーフ長 k に影響されるため, 最頻出パターンを含む部分配列集合が必ずしも最大の F 値を示すとは限らない.

3. 分割型 GS による大域的モチーフの安定的抽出

従来 GS は, 初期状態 U に分布するモチーフに強く影響されるためモチーフ抽出精度が安定せず, さらにモチーフパターンによっては U における一部のアミノ酸要素の出現数が 0 に陥った場合, U に対して類似性の高い部分配列の遷移指標値が 0 を示す問題があった.

本研究では, 従来 GS におけるモチーフ抽出精度の改善を目指し, まず初期状態の部分配列集合に分布するモチーフの影響を抑制するため, GS へのデータ分割を適用した. 初期状態の部分配列集合に含まれるモチーフの分布数は非常に少ないことが予想されるため, データ分割を適用すると, 初期状態に分布するモチーフに対して影響を受けない分割データが複数作成され, 大域的モチーフが導出される可能性が向上すると考えられる. また, 大域的モチーフは分割データの多くに分布するため, 各分割データに対して生成される分割部分配列集合に出現する可能性が最も高く, かつ高い指標値を獲得できると推測される. そこで, 初期状態の部分配列集合に分布するモチーフの影響を最小限に抑えるため, 2 種類の分割部分配列集合を適用した部分配列更新を実行し, 従来 GS よりも安定的に大域的モチーフを抽出できる分割型 GS を提案した.

分割型 GS は, 入力データセット S に対して $W (2 \leq W \leq \frac{n}{2})$ 種類の分割データセット群 S' および分割部分配列集合群 U' を作成し, $1 \leq i \leq W$, $1 \leq j \leq W, i \neq j$ のとき, S', U' はそれぞれ式 3, 4 のように定義される.

$$S' = \{S'_1, S'_2, \cdots S'_{W-1}, S'_W\}, S'_i \subset S, S'_i \cap S'_j = \emptyset \quad (3)$$

$$U' = \{U'_1, U'_2, \cdots U'_{W-1}, U'_W\}, U'_i \subset U, U'_i \subset S'_j \quad (4)$$

次に, 従来 GS が U に対して類似性の高い部分配列の遷移指標値を 0 とみなす問題に対して, 出現数に対する擬似度数を追加することで遷移指標値が 0 を示すことを防ぐようにした. ただし, 擬似度数の大きさによっては配列長の大きな局所解モチーフへの影響を受けやすくなり, 大域的モチーフの抽出が困難となるため, 背景頻度に対して十分に低い数値を与えなければならない. そこで, 背景頻度を重みとした擬似度数 $R^{U'_i}_{a^{s_r}_{x+j}} = f^{S'}_{a^{s_r}_{x+j}} \times P^{U'_i}_{a^{s_r}_{x+j}}$ を適用した新たな

遷移指標 $E_x^{U'_i}$(式 5) を分割型 GS へ適用した. なお, $f^{S'}_{a^{s_r}_{x+j}}$ は, アミノ酸要素 $a^{s_r}_{x+j}$ の U'_i に対する相対出現頻度によって定義される.

$$E_x^{U'_i} = \prod_{j=0}^{k-1} \frac{C^{U'_i}_{j+1,a^{s_r}_{x+j}} + R^{U'_i}_{a^{s_r}_{x+j}}}{nP^{U'_i}_{a^{s_r}_{x+j}}} \tag{5}$$

次に, 分割型 GS において部分配列 u_r を更新するときは, s_r を含む $U'_\alpha(1 < \alpha < W+1)$ と, ランダムに選択された分割部分配列集合 $U'_\beta(1 < \beta < W+1, \alpha \neq \beta)$ の 2 種類の各遷移指標 (式 5) を応用した U'_α, U'_β 間の統合遷移指標 E'_x(式 6) を適用した.

$$E'_x = \frac{E_x^{U'_\alpha} + E_x^{U'_\beta}}{1 + 2(l_{s_r} - k + 1) - p(E_x^{U'_\alpha}) - p(E_x^{U'_\beta})} \tag{6}$$

$p(E_x^{U'_\alpha}), p(E_x^{U'_\beta})$ は, 各遷移指標群 U'_α, U'_β における $E_x^{U'_\alpha}, E_x^{U'_\beta}$ のそれぞれの p 値を意味し, 統合遷移指標に対する重みとして適用され, 遷移指標値における各 p 値が高いほどモチーフ配列としての適合性が高く, 統合遷移指標がより高い指標値として示される. アルゴリズム 2 にて, 分割型 GS のモチーフ抽出手続きを示した.

アルゴリズム 2 分割型 GS

入力: ゲノム配列データセット $S = \{s_1, s_2, \cdots, s_{n-1}, s_n\}$, モチーフ長 k

出力: 部分配列集合 $U = \{u_1, u_2, \cdots, u_{n-1}, u_n\}$

1: 各 $s_i(1 \leq i \leq n)$ に対して, $u_i \subset s_i$ をみたすモチーフ長 k の部分配列集合 U をランダムに作成する.
2: 分割数 W に対して, W 種類に構成される分割データセット群 S' および分割部分配列集合群 U' を作成する.
3: U に対する更新対象の配列番号 $r(1 \leq r \leq n)$ をランダムに決定する.
4: s_r, u_r を含む分割領域をそれぞれ S'_α, U'_α とし, それ以外の分割領域 $S'_\beta, U'_\beta(1 \leq \beta \leq W, \beta \neq \alpha)$ をランダムに選択する.
5: $U'_\alpha = U'_\alpha \cap \overline{u_r}$ とし, $S'_\alpha \cap \overline{U'_\alpha \cup s_r}, S'_\beta \cap \overline{U'_\beta}$ における各アミノ酸要素 a^ω_j の背景頻度 $P^{U'_\alpha}_{a^\omega_j}, P^{U'_\beta}_{a^\omega_j}$ と, 各位置 $q(1 \leq q \leq k)$ における出現数 $C^{U'_\alpha}_{q,a^\omega_j}, C^{U'_\beta}_{q,a^\omega_j}$ を求める.
6: s_r に対し, $l_{s_r} - k + 1$ パターンの各部分配列に対する遷移指標群 $E^{U'_\alpha}, E^{U'_\beta}$ を作成する.
7: $E^{U'_\alpha}, E^{U'_\beta}$ に対する統合遷移指標値群 $E' = \{E'_1, E'_2, \cdots, E'_{l_{s_r}-k}, E'_{l_{s_r}-k+1}\}$ の各指標値に比例した確率で E'_x を選択する.
8: $u_r = \{a^{s_r}_x a^{s_r}_{x+1} \cdots a^{s_r}_{x+k-2}, a^{s_r}_{x+k-1}\}$ として部分配列を更新し, $U'_\alpha = U'_\alpha \cup u_r$ とする.
9: U の収束が得られるまで, 3 ～ 8 を繰り返す.

分割型 GS は従来 GS に対してより大域的なモチーフの抽出が期待されるが, 実行の際に分割数 W が小さく設定されると分割データの配列数の増加によって局所解モチーフの影響を受けやすく, 大域的モチーフの抽出が困難となる. また, W が大きすぎる場合は分割データの配列数の減少に伴い, 自然界における多様なパターンを持つモチーフに対して分割部分配列集合が高い遷移指標値を獲得できず, モチーフをほとんど抽出できない場合がある. そのため, 分割型 GS は局所解モチーフの影響を考慮しつつ, 多様なモチーフのパターンに対して十分に高い指標値が獲得されるように配列数をある程度大きく設定しなければならない. 以上により, 分割型 GS は配列数の大きなデータセットに対して大域的なモチーフがより安定的に抽出されると期待される.

4. 分割型 GS によるモチーフ候補群の抽出

分割型 GS によって入力データセットにおける大域的モチーフの安定的な抽出が期待されるが, 大域的モチーフが複数存在する場合, 1 種類のみの部分配列集合の出力では推定結果が不十分とされるため, 正確なモチーフ推定には複数のモチーフ候補 (モチーフ候補群) の抽出が必要とされる.

また, 入力データセットに含まれる各モチーフを $M = \{M_1, M_2, \cdots\}, M_i = \{m^i_1, m^i_2, \cdots\}$ とし, 各 M_i のモチーフ長を K_i とすると, 入力データセット S において分布する各モチーフの配列長や出現頻度はそれぞれ異なる場合が多く, 入力データセットの特定の配列ではモチーフ配列間において重複が観測されることがある. これらのモチーフの分布を考慮すると, 従来 GS や分割型 GS を適用したモチーフ候補群の抽出を実行する際には, 正確なモチーフ長 k を与えることができない. そこで, 本研究では分割型 GS に対して十分に大きな部分配列長 k を入力し実行することにより, 完全に大域的モチーフを含んだ部分配列集合を抽出した. 十分に大きなモチーフ長を検討する際には, モチーフデータベースとして用いられる PROSITE[5] に登録されるモチーフ長の多くが $5 \leq k \leq 40$ に分布することから, $k = 40$ を分割型 GS に与えることとした. しかし, 分割型 GS に適用された遷移指標 (式 5) は類似性の低い領域に影響されやすく, 部分配列集合に対して類似性の高い部分配列の遷移指標値が小さくなる場合がある. そのため, 類似性の低い領域に影響されにくく, かつ類似性の高い領域に比例して高い指標値を求められる遷移指標の適用が必要とされる. そこで, 従来 GS の遷移指標を改善する際に適用した擬似度数を変更し, 背景頻度に近い擬似度数を与えると, モチーフを含む部分配列内の類似性の低い領域に影響されにくい. したがって, モチーフ候補群抽出における擬似度数には, 背景頻度 $P^{U'_i}_{a^\omega_j}$ に十分近い値を示すと予想される $f^{S'_i}_{a^\omega_j}$ を適用した.

モチーフ候補群の抽出においては, 同一のモチーフが繰り返し抽出されない効率的なモチーフ探索が必要とされる. ただし, 自然界におけるゲノム配列にはモチーフ配列同士が重複する場合があり, このようなパターンを考慮したモチーフ抽出が必要とされる.

十分に大きいモチーフ長によって抽出された各部分配列 u_r には, u_r に含まれるモチーフ配列 $m^i_{x_i}, i \geq 0, x_i \geq 0$ に関して以下のように判別される.

(1) モチーフ配列との完全一致の有無 ($u_r = m^i_{x_i}$, または $u_r \neq m^i_{x_i}$)
(2) 各モチーフ配列の完全獲得の有無 ($u_r \subset m^i_{x_i}$, または $u_r \not\subset m^i_{x_i}$)
(3) 各モチーフ配列に対する部分一致の有無 ($u_r \cap m^i_{x_i} = 0$, または $u_r \cap m^i_{x_i} > 0$)
(4) 各モチーフ配列における重複の有無 ($m^i_{x_i} \cap m^{i+1}_{x_{i+1}} > 0$, または $m^i_{x_i} \cap m^{i+1}_{x_{i+1}} = 0$)
(5) モチーフでない配列の有無 ($u_r \cap \overline{m^i_{x_i}} \cap \overline{m^{i+1}_{x_{i+1}}} \cdots > 0$, または $u_r \cap \overline{m^i_{x_i}} \cap \overline{m^{i+1}_{x_{i+1}}} \cdots = 0$)

各部分配列のモチーフ間の関係を考慮したとき, 以前に獲得された部分配列が必ずしも不要な配列になるとは限らないため, 同一の部分配列であってもある程度高い遷移指標を導出でき, かつ複数のモチーフ候補を効率的に抽出できることが望ましい. ただし, モチーフ配列間における重複が発生する現象はごくわずかであり, データセットにおけるモチーフの分布状態に対し, 複数のモチーフが部分配列に含まれる可能性は低い. そのため, 以前に抽出された部分配列集合との重複サイズに比例したペナルティを遷移指標に与えると, 初期状態における遷移指標は比較的低い数値を示すためペナルティの影響を受けやすく, 新たなモチーフ候補を優先的に導出できる. また, 部分配列集合における新たなモチーフ候補配列が多く獲得されたとき, 以前抽出された部分配列の遷移指標はペナルティの影響を受けにくく, かつ高い遷移指標を示すことにより適切なモチーフ候補として再抽出できると期待される.

モチーフ候補群抽出における分割型 GS の各部分配列 u_r に与えるペナルティ $\epsilon_{r,x}$ は, 抽出済みのモチーフ候補数 $t(t \geq 0)$ における各 u^t_r を用いて式 7 のように決定され, 以前抽出された部分配列との重複長に比例して増加する. $\epsilon_{r,x}$ を適用した分割部分配列集合における遷移指標は式 8 として定義される.

$$\epsilon_{r,x} = \sum_{i=1}^{t} (\{a^{s_r}_x a^{s_r}_{x+1} \cdots a^{s_r}_{x+k-2} a^{s_r}_{x+k-1}\} \cap u^i_r)! \quad (7)$$

$$E^{U'_i}_x = \frac{1}{1+\epsilon_{r,x}} \prod_{j=0}^{k-1} \frac{C^{U'_i}_{j+1,a^{s_r}_{x+j}} + f^{S'_i}_{a^{s_r}_{x+j}}}{nP^{U'_i}_{a^{s_r}_{x+j}}} \quad (8)$$

提案法にて出力されるモチーフ候補の部分配列長はモチーフに対して十分に大きいことを仮定しているため, 正確なモチーフのみを抽出しなければならない. 正確にモチーフが抽出されたモチーフ候補に対して F 値を適用し, F 値を構成するモチーフ候補の各位置の評価値 $\{f_1, f_2, \cdots, f_{k-1}, f_k\}$ を観測すると, 類似性の高いモチーフ領域には当然高い値が分布する. さらに, 部分列長 $k'(1 \leq k' \leq k)$ を用いて, k' のとりうる値域にて獲得される F 値 $F_{k'} = \{\sum_{i=1}^{k'} f_i, \sum_{i=2}^{k'+1} f_i, \cdots, \sum_{i=k-k'-1}^{k-1} f_i, \sum_{i=k-k'}^{k} f_i\}$ の分布を比較したとき, 部分列長 k' がモチーフ長の真値と一致したとき, 最も鋭いピークが発見される. そこで, 各部分列長における F 値の推移から尖度 $\theta_{k'}$(式 9) が最も高くなる k' をモチーフ長の真値として推定し, その部分列長にて最も高い F 値をモチーフとして再抽出した.

最後に, 各 U_i におけるモチーフ出現頻度の尤度関数 $L(U_i)$(式 12) を適用し, $L(U_i)$ にしたがって, U_i のモチーフ適合性を評価した.

$$\theta_{k'} = \frac{\sum_{i=1}^{k-k'-1} (\sum_{j=0}^{k'-1} f_{i+j} - \mu)^4}{k'\sigma^4} \quad (9)$$

$$\sigma^2 = \frac{1}{n-1} \sum_{i=1}^{k-k'-1} (\sum_{j=0}^{k'-1} f_{i+j} - \mu)^2 \quad (10)$$

$$\mu = \frac{1}{n} \sum_{i=1}^{k-k'+1} \sum_{j=0}^{k'-1} f_{i+j} \quad (11)$$

$$L(U_i) = \frac{1}{n} \sum_{j=1}^{n} \frac{p(E^{U_i}_{u_j})}{l_{s_j} - k + 1} \quad (12)$$

式 12 は, 各部分配列のモチーフとしての適合性の高さに比例して高い指標値が算出されることから, p 値を適用すると各 s_j におけるモチーフである確率として示されるため, 全部分配列に対して適用することで U_i におけるモチーフ出現頻度の尤度として導出される.

モチーフ候補群抽出における終了条件としては, モチーフ候補 $L(U_i)$ の変動を利用し, U_i, U_{i+1} 間にて $L(U_i) > L(U_{i+1})$ を示し, かつその後抽出された U_{i+2} において閾値 $\varphi = 0.1$ よりも尤度の変動が確認されなかった場合に実行を完了することとした. アルゴリズム 3 に, モチーフ候補群抽出における手続きを示した.

アルゴリズム 3 分割型 GS を適用したモチーフ候補群抽出法

入力: ゲノム配列データセット $S = \{s_1, s_2, \cdots, s_{n-1}, s_n\}$
出力: モチーフ候補群 $Z = \{U_1, U_2, \cdots, U_{t-2}, U_{t-1}\}$, 尤度 $L = \{L(U_1), L(U_2), \cdots, L(U_{t-2}), L(U_{t-1})\}$

1: 抽出済モチーフ候補数 t の初期値として, $t = 0$ とする.
2: ペナルティを考慮した遷移指標 (式 8) を適用した分割型 GS により U_{t+1} を抽出する.
3: S と部分配列長 $k = 40$ を入力した分割型 GS によって抽出された部分配列集合 U_{t+1} に対するモチーフ出現頻度の尤度 $L(U_{t+1})$ を求める.
4: U_{t+1} に対する F 値 $\{f_1, f_2, \cdots, f_{k-1}, f_k\}$ を求め, U_j に含まれるモチーフ長の真値 k' の推定を行う. また, k' に対し, 最も高い F 値を示す U_j の部分的な領域を新たな U_j として更新する.
5: U_j に対し, モチーフ出現頻度の尤度 $L(U_j)$ を求める.
6: モチーフ候補群 $Z = \{U_1, U_2, \cdots, U_{t'-2}, U_{t'-1}\}, t' \geq 3$ に対し, $L(U_{t'-3}) > L(U_{t'-2})$ かつ $|L(U_{t'-2}) - L(U_{t-1})| < \varphi$ をみたすまで, モチーフ候補群 Z に U_{t+1} を追加し, $t = t + 1$ として 2 ∼ 4 を繰り返す.

5. 複数の頻出パターン抽出における性能評価比較検証実験

ゲノム配列データセットに含まれる各モチーフと, モチーフ候補群抽出法によって出力される各モチーフ候補の関連性を検証し, MAFFT と法に対するモチーフ抽出精度を比較する実験を行った.

本研究において提案したモチーフ候補群抽出法は複数のモチーフ候補の出力を前提としているため, 1 種類のみのモチーフを含むゲノム配列データセットと, 複数のモチーフを含むゲノム配列データセットを適用することでモチーフ候補の各モチーフに対する抽出精度や, モチーフ出現頻度の尤度などの精度の変化を観測できると考えられる. なお, 従来 GS に対してはモチーフ候補群抽出法と同様に部分配列長 $k = 40$ を入力し, 抽出された部分配列集合に対して性能評価を行った. また, MAFFT はアライメントされたゲノム配列データセットの全配列を出力するため, $k \leq 40$ における出力からモチーフが最も多く含まれる部分配列集合 U を抽出し, 性能評価対象とした.

モチーフ抽出精度の性能評価尺度には nucleotide-level sensitivity(nSn), site-level sensitivity(sSn), nucleotide-level positive predictive value($nPPV$) を適用した [6]. 各評価尺度は, True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN) をそれぞれ nucleotide-level, site-level として適用したものが用いられ, $nSn, sSn, nPPV$ は式 13, 14, 15 によって定義される.

$$nSn = \frac{nTP}{nTP + nFN} \tag{13}$$

$$sSn = \frac{sTP}{sTP + sFN} \tag{14}$$

$$nPPV = \frac{nTP}{nTP + nFP} \tag{15}$$

本実験では比較検証実験に人工データを適用し, 1 種類のみのモチーフが含まれるゲノム配列データセット S_1, 2 種類のモチーフが含まれるゲノム配列データセット S_2 を作成した. ゲノム配列データセットの配列数 n, 各配列長 $l_{s_i}(1 \leq i \leq n)$ は $n = 500, 1000 \leq l_{s_i} \leq 2000$ として定義される. 人工データセットに設定されるモチーフは表 1 に示され, アミノ酸要素によって無作為に生成された人工データであり, いずれも一意のパターンを持つ. また, モチーフはアミノ酸要素によって無作為に生成された各ゲノム配列のランダムな位置に挿入し, 一配列に同一のモチーフ配列が複数含まれないように作成された.

表 1 人工データセットに適用した各モチーフの出現頻度および共通パターン

	出現頻度	配列長	PROSITE パターン形式
M_1	0.9	24	P-C-G-N-H-R-C-G-K-P-M-K-G-H-D-K-L-F-H-G-V-I-T-C
M_2	0.7	27	M-Y-M-T-G-C-G-I-P-L-V-Y-Q-Y-M-D-E-W-E-Q-K-Y-V-Y-Y-I-Y

表 2 S_1 に対する従来 GS のモチーフ抽出精度およびモチーフ候補群抽出法によって抽出された各モチーフ候補のモチーフ抽出精度, モチーフ長推定値 k', モチーフ出現頻度の尤度 $L(U_i)$

Algorithm	MAFFT	従来 GS	モチーフ候補群抽出法		
出力	$U(k = 24)$	U	U_1	U_2	U_3
nSn	0.01	0.99	0.99	0.00	0.00
sSn	0.02	0.99	0.99	0.00	0.00
$nPPV$	0.01	0.53	0.63	0.00	0.00
k'			34	3	12
$L(U_i)$			0.99	0.60	0.60

表 3 S_2 に対する従来 GS のモチーフ抽出精度およびモチーフ候補群抽出法によって抽出された各モチーフ候補のモチーフ抽出精度, モチーフ長推定値 k', モチーフ出現頻度の尤度 $L(U_i)$

Algorithm		MAFFT	従来 GS	モチーフ候補群抽出法		
出力		$U(k = 27)$	U	U_1	U_2	U_3
M_1	nSn	0.00	0.84	0.64	0.00	0.00
	sSn	0.01	0.96	0.96	0.00	0.00
	$nPPV$	0.00	0.45	0.87	0.00	0.00
M_2	nSn	0.01	0.01	0.08	0.36	0.81
	sSn	0.02	0.02	0.13	0.99	0.99
	$nPPV$	0.00	0.00	0.09	0.69	0.69
k'				16	10	22
$L(U_i)$				0.99	0.95	0.98

表 2, 3 に MAFFT, 法, モチーフ候補群抽出法によって抽出された各出力の検証実験結果を示した. MAFFT は S_1, S_2 に対してほとんどモチーフをアライメントできなかったが, よりモチーフ出現頻度の高いモチーフを含むゲノム配列データセットに対してはモチーフ抽出精度の向上が期待される. モチーフ候補群抽出法ではモチーフ長 k' の推定法の適用により, 従来 GS に対する $nPPV$ の向上に成功したが, M_1 のモチーフ長の真値に対しては推定値 k' が十分とはいえない. また, S_2 に対してはモチーフ長の推定値 k' が必要以上に非常に小さくなり, 不要な配列とともにモチーフ配列が除去される結果を示した. これらの現象を防ぐためには, F 値による尖度のより正確な観測のため, モチーフ候補に複数のモチーフが混在しないようにし, かつ不要な部分配列を除去したモチーフ候補を導出する必要があると考えられる.

また, S_1 に対してモチーフ候補群抽出法を実行した際, U_2, U_3 にはモチーフが含まれていない状態にも関わらず, 部分配列の p 値の高さによってモチーフ出現頻度の尤度が 0.6 を示した. これらの尤度をよりモチーフの出現頻度に近似させるには, 算出された各部分配列の遷移指標値など, 順位尺度以外の数値を考慮に含める必要がある.

6. まとめ

本研究では従来 GS に対して大域的なモチーフの安定的な抽出を実行する分割型 GS を提案した. さらに, 複数の頻出パターンの抽出のため, 分割型 GS を応用したモチーフ候補群抽出法を用い, それぞれのモチーフ候補に対してモチーフ長 k を推定することによるモチーフ出現頻度の尤度を導出し, それらの適性を確認した. 実データを適用した際のモチーフ候補群抽出法のモチーフ抽出精度については当日報告する.

参考文献

[1] Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F. and Wootton, J. C.: Detecting Subtle Sequence Signals A Gibbs Sampling Strategy for Multiple Alignment, *Science*, Vol. 262, No. 5131, pp. 208–214 (1993).

[2] Liu, J. S., Neuwald, A. F. and Lawrence, C. E.: Baysian Models for Multiple Local Sequence Alignment and Gibbs Sampling Strategies, *American Statistical Association*, Vol. 90, No. 432, pp. 1156–1170 (1995).

[3] Katoh, K. and Standley, D. M.: MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability, *Molecular Biology and Evolution*, Vol. 30, No. 4, pp. 772–780 (2013).

[4] 河野修久, 田村慶一, 森康真, 北上始：ギブスサンプリングとアラインメント処理に基づく類似部分配列の抽出方式, 研究報告数理モデル化と問題解決, Vol. 2009-MPS-76, No. 46, pp. 1–8 (2009).

[5] Bairoch, A.: PROSITE, Swiss Institute of Bioinformatics, Centre Medical Universitaire and Structural Biology and Bioinformatics Department, University of Geneva, 1 rue Michel Servet, CH-1211 Geneva 4 (online), available from ⟨http://prosite.expasy.org/⟩ (accessed 2014-11-14).

[6] Tompa, M., Li, N., Bailey, T. L., Church, G. M., Moor, B. D., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Regnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. and Zhu, Z.: Assessing computational tools for the discovery of transcription factor binding sites, *Nat. Biotechnol.*, Vol. 23, No. 1, pp. 137–144 (2005).