Vol.2019-BIO-57 No.7
2019/3/9

SELEX 法を用いた核酸アプタマー推定のための高速クラスタリング手法とその性能評価

小野 貴義[1,a)]　加藤　信太郎[2,1]　伊藤　康一[1]　皆川 宏貴[2]　堀井 克紀[2]　白鳥 行大[2]　和賀 巌[2]
青木 孝文[1]

概要：標的分子と特異的に結合する核酸分子（アプタマー）は，Systematic Evolution of Ligands by EXponential enrichment (SELEX) 法により選択される．SELEX 法とは，ランダムに初期化された核酸ライブラリーから標的分子による選択と Polymerase Chain Reaction (PCR) による増幅を繰り返す実験手法である．SELEX 法から得られた核酸ライブラリーを次世代ゲノムシーケンサーで読み取り，大量のアプタマー候補の塩基配列を得る．候補配列すべてを実験的に評価することは現実的ではないため，至適配列を効率的に選択する必要がある．本稿では，大量の候補配列をクラスター化し，その代表配列をアプタマーの至適配列とするための高速なクラスタリング手法を提案する．性能評価実験を通して，速度と精度における提案手法の有効性を示す．

キーワード：アプタマー，SELEX 法，次世代ゲノムシーケンサー，クラスタリング

A Fast Clustering Method for Aptamer Estimation Using SELEX Sequence Data and Its Performance Evaluation

1. はじめに

アプタマー [1] とは，その塩基配列・三次構造の両方により，標的分子と特異的に結合する一本鎖の核酸分子（RNA または DNA）である．タンパク質，ペプチド，低分子化合物，金属イオンなどの標的分子に対し，強い結合力と特異な選択性を持つため，分子バイオセンサー [2] や診断薬，治療薬 [3], [4] に応用されている．

アプタマーの実験的な選別方法として，Systematic Evolution of Ligands by EXponential enrichment (SELEX) 法 [5] が用いられている．SELEX 法の概要を **図 1** に示す．SELEX 法は，(i) 初期の核酸ライブラリーの作成，(ii) 標的分子による核酸分子の選択，(iii) 標的分子と結合しない核酸分子の洗浄，(iv) 標的分子と結合する核酸分子の溶出，(v) ポリメラーゼ連鎖反応 (Polymerase Chain Reaction: PCR)[*1]による溶出した核酸分子の増幅の 5 ステップで構成される．核酸分子は，PCR のためのプライマー領域が両端に配置され，その間をランダムな塩基配列で構成される．初期ライブラリーには，約 10^{14}〜10^{15} 個の核酸分子が格納される [7]．(ii)〜(v) を 1 ラウンドとして，これを

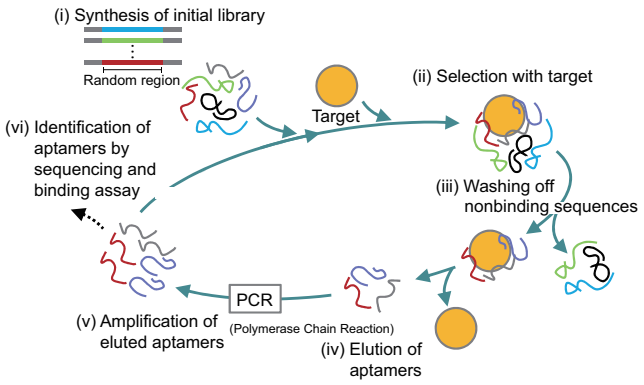

図 1　SELEX 法の手順
Fig. 1　SELEX procedure.

[1] 東北大学 大学院情報科学研究科　〒 980-8579 仙台市青葉区荒巻字青葉 6-6-05
[2] NEC ソリューションイノベータ株式会社　〒 136-8627 東京都江東区新木場 1-18-7
[a)] ono@aoki.ecei.tohoku.ac.jp

[*1] 鋳型 DNA，DNA ポリメラーゼとその補助因子，プライマー（鋳型 DNA の両端の配列を含む 2 種類の短い一本鎖 DNA）を含む溶液の温度を周期的に変化させることで，DNA を指数関数的に増幅させる [6]．

© 2019 Information Processing Society of Japan

繰り返すことで最終的に，標的分子と結合する核酸分子（アプタマー）が試験管内に濃縮される．各ラウンドにおける核酸ライブラリーを次世代ゲノムシーケンサー (Next Generation Sequencer: NGS) で読み取ることで，大量のアプタマー候補の塩基配列を得ることができる．

核酸分子が十分に濃縮された核酸ライブラリーを，NGS で読み取り得られた候補配列から，アプタマーの配列を推定することができる．しかし，観測されたすべての候補配列が標的分子と結合しているわけではない．標的分子による選択以外に，PCR バイアス [8], [9] や標的分子以外への結合により，濃縮された核酸分子が存在する．そのため，候補配列が標的分子に対する強い結合力と特異な選択性を有しているかを実験的に評価する必要がある（図 1 (iv)）．一方で，NGS で読み取った大量の候補配列すべてを実験的に評価することは現実的ではない．そのため，類似する配列をクラスター化し，その代表配列をアプタマーの至適配列とする必要がある．

SELEX 法から得られた大量の配列データをクラスタリングする既存手法に，FASTAptamer [10] や AptaCluster [11]，APTANI [12] がある．FASTAptamer，AptaCluster，APTANI は，それぞれ Levenshtein Distance (LD)，局所性鋭敏型ハッシュ (Locality Sensitive Hashing: LSH) と k-mer 類似度，マルチプルアラインメント [13] に基づいてクラスタリングを行う．通常，アプタマーは，結合に関与する共通の部分配列（モチーフ）により，標的分子と結合しているため，モチーフに基づいてクラスタリングを行うべきである．しかし，これらのクラスタリング手法は，配列全体の類似度を考慮しているが，類似したモチーフに基づいて配列をクラスタリングするようには設計されていない．

そこで，本稿では，SELEX 法において NGS で読み取った大量の候補配列から効率的かつ正確にアプタマーの至適配列を選択するため，配列の類似度ではなく，モチーフを考慮した新たなクラスタリング手法を提案する．精度評価実験を通して，速度と精度における提案手法の有効性を示す．

2. 提案手法

配列データが与えられたとき，長さ l_{min} から l_{max} の部分配列を探索しながらモチーフを推定し，モチーフをもとに効率的にクラスタリングを行う方法について説明する．2.1 では，スコアの定義で必要となる，特定の部分配列を含む配列が出現する確率の計算方法を説明する．2.2 では，モチーフを推定する際の指標となる統計量 Z スコアを定義する．2.3 では，Z スコアをもとに，長さ l_{min} から l_{max} の部分配列を探索しながらモチーフを推定し，モチーフをもとにクラスタリングを行う手法について説明する．

2.1 特定の部分配列を含む配列の出現確率

長さ L の配列 s 内に長さ l の部分配列 m が出現する確率 $P_a(m, L)$ を考える．全長配列 s の i 番目の文字を $s[i]$ と表し，s の i_1 番目から i_2 番目までの部分文字列を $s[i_1..i_2]$ と表す．m に対しても同様とする．Ω を文字の集合する．塩基を扱う場合は，$\Omega = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T(U)}\}$ となる．s の i 番目の文字が $c\ (\in \Omega)$ であるという事象は互いに独立であるとする．各文字の出現確率を $p_0(c)\ (c \in \Omega)$ とすると，長さ l の配列が m である確率 $Q(m)$ は，以下の式で表せる．

$$Q(m) = \prod_{i=1}^{l} p_0(m[i]) \tag{1}$$

また，m 内で自己重複している文字列の集合を $\mathcal{T}$ とする．m が “ATATA” であるとき，$\mathcal{T} = \{\mathrm{A}, \mathrm{ATA}\}$ となる．m 内で自己重複している文字列 $t \in \mathcal{T}$ は，$m[1], m[1..2], \ldots, m[1..l-1]$ のいずれかの文字列である．$|t|$ を t の長さとするとき，長さ $|t|$ の配列が t である確率 $q(t)$ は，以下の式で表すことができる．

$$q(t) = \prod_{i=1}^{|t|} p_0(t[i]) \tag{2}$$

2.1.1 部分配列 m に自己重複がないとき

長さ l の部分配列 m に自己重複する文字列がないとき，長さ L の配列 s に部分配列 m が出現する確率 $P_a(m, L)$ は，以下の漸化式で表される．

$$P_a(m, L) = \begin{cases} 0 & (L < l) \\ P_a(m, L-1) + Q(m)[1 - P_a(m, L-l)] & (l \leq L) \end{cases} \tag{3}$$

$P_a(m, L)$ は，長さ $L-1$ の配列内に部分配列 m が出現する確率 $P_a(m, L-1)$ と，$s[L-l+1..L]$ が m かつ $s[1..L-l]$ に m が出現しない確率 $Q(m)[1 - P_a(m, L-l)]$ との和により求まる．**図 2 (a) (1)** に示すように，長さ $L-1$ の配列の長さを L に伸長すると，$s[L-l+1..L]$ が m である事象が新たに生じる．$s[L-l+1..L]$ に m が存在する事象の確率を計算すると，$s[1..L-l]$ における文字は任意であるため，$\prod_{i=1}^{L-l} \Sigma_{c \in \Omega} p_0(c) \cdot \prod_{j=1}^{l} p_0(m[j]) = 1 \cdot Q(m) = Q(m)$ となる．しかし，この事象の確率のうち，$s[1..L-l]$ に m が存在する事象（図 2 (a) (2)）の確率は $P_a(m, L-1)$ で計算しているため，$s[L-l+1..L]$ が m かつ $s[1..L-l]$ に m が存在する事象の確率 $Q(m)P_a(m, L-l)$ を引く必要がある．したがって，部分配列 m に自己重複する文字列がないとき，配列 s に m が出現する確率は，式 (3) で表される．

2.1.2 部分配列 m に自己重複があるとき

部分配列 m に自己重複する文字列があるとき，配列 s に

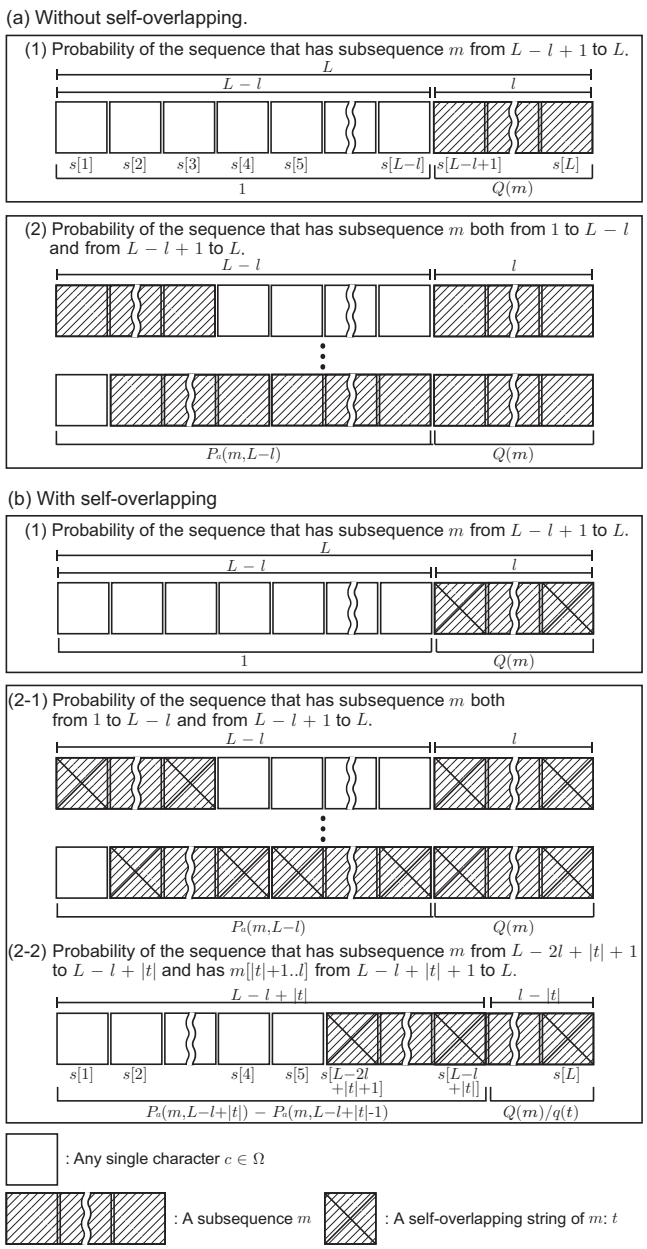

図 2 漸化式（式 (3)，(4)）の説明

Fig. 2 Illustration of recurrence formulas Eq. (3) and Eq. (4).

m が出現する確率 $P_a(m, L)$ は，以下の漸化式で表される．

$$P_a(m,L) = \begin{cases} 0 & (L < l) \\ P_a(m,L-1) + Q(m)[1 - P_a(m,L-l)] \\ -\sum_{t \in \mathcal{T}} \frac{Q(m)}{q(t)}[P_a(m,L-l+|t|) \\ -P_a(m,L-l+|t|-1)] & (l \le L) \end{cases} \tag{4}$$

式 (3) と同様に，長さ $L-1$ の配列内に部分配列 m が出現する確率 $P_a(m, L-1)$ と，$s[L-l+1..L]$ が m かつ $s[1..L-l]$ に m が出現しない確率 $Q(m)[1-P_a(m, L-l)]$ を項に含む．加えて，図 2 (b) (2-2) に示すように，$s[L-2l+|t|+1..L-l+|t|]$ が m かつ，$s[L-l+|t|+1..L]$ が $m[|t|+1..l]$ である事象の確率は，$P_a(m, L-1)$ で計算しているため，この事象の確率を引く必要がある．この事象の確率は，$s[L-l+|t|+1..L]$ が $m[|t|+1..l]$ である確率のうち，長さ $s[1..L-l+|t|]$ に部分配列 m が出現する確率から，$s[1..L-l+|t|-1]$ に部分配列 m が出現する確率を引いた $Q(m)/q(t)[P_a(m, L-l+|t|) - P_a(m, L-l+|t|-1)]$ で表すことができる．また，部分配列 m には，自己重複する文字列が複数存在する場合を考慮すると，$P_a(m, L)$ は式 (4) で表される．

2.2 スコアの定義

配列データ内で有意に現れる部分配列が，モチーフであると推定するための統計量を定義する．SELEX 法では，核酸分子に対して塩基の挿入や削除が起こるために，NGS で読み取った各候補配列の長さが異なることがある．このような，配列データ内に異なる長さの配列が含まれることを考慮して，配列データ $S = \{s_1, s_2, \cdots, s_{|S|}\}$ における部分配列 m を含む配列の出現確率 $P_d(m, S)$ を，以下の式により求める．

$$P_d(m,S) = \frac{1}{|S|}\sum_{i}^{|S|} P_a(m, |s_i|) \tag{5}$$

ここで，$|s_i|$ は配列 s_i の長さを表す．統計量 Z スコア $Z(m, S)$ を $P_d(m, S)$ を用いて，以下の式により定義する．

$$Z(m,S) = \frac{\frac{F_m}{|S|} - P_d(m,S)}{\sqrt{\frac{P_d(m,S)[1-P_d(m,S)]}{|S|}}} \tag{6}$$

ここで，F_m は部分配列 m を含む配列の観測数を表す．

2.3 高速なクラスタリング手法

2.3.1 モチーフの推定

配列データ内に現れる部分配列を探索し，Z スコアをもとにモチーフを推定する．モチーフの推定を行う前に，各文字の出現確率を推定する．配列データ $S = \{s_1, s_2, \cdots, s_{|S|}\}$ が与えられたとき，配列 s_i の長さを $|s_i|$，s_i に含まれる各文字の数を n_i^c $(c \in \Omega)$ として，各文字の出現確率を以下の式により推定する．

$$\hat{p}_0(c) = \sum_{i=1}^{|S|} \frac{n_i^c}{|s_i|} \tag{7}$$

各文字の出現確率の推定値 $\hat{p}_0(c)$ を Z スコアの算出に用いる．これにより，配列データ全体における文字の比率の偏りを考慮することができる．

モチーフの推定は，以下の (i) ∼ (v) の処理により行う．長さ l_{min} から l_{max} にわたり，部分配列を 1 文字ずつ伸長して，伸長前後の部分配列の Z スコアを比較することで，探索する部分配列の数を削減しながらモチーフを推定する．

(i) 長さ l_{min} のすべての部分配列に対して，それぞれ Z

© 2019 Information Processing Society of Japan

スコアを計算する．Z スコアが 0 より大きい部分配列をモチーフとする．

(ii) $l \leftarrow l_{min}$ とする．

(iii) 推定したモチーフに Ω の任意の 1 文字を付加して伸長し，伸長した部分配列の Z スコアを計算する．Z スコアが，伸長前の Z スコアよりも大きい部分配列をモチーフとする．

(iv) $l+1 > l_{max}$ のとき，モチーフの推定を終了する．

(v) $l \leftarrow l+1$ として (iii) に戻る．

Algorithm 1 Procedure for motifs estimation

Require: Sequence data: $S = \{s_1, s_2, \cdots, s_{|S|}\}$;
A character set: Ω;
Subsequences: $M = [m_1, m_2, \cdots, m_{|\Omega|^{l_{min}}}]$;
The minimum/maximum length of subsequences: l_{min}, l_{max};
Ensure: The estimated motifs: M';
for $i = 1$ to $|\Omega|^{l_{min}}$ **do**
$z \leftarrow Z(m_i, S)$
if $z > 0$ **then**
Add m_i to the end of the array M^*
Add z to the end of the array Z
end if
end for
$M \leftarrow M^*$
Add elements of M^* to M'
Remove all elements of M^*
$l \leftarrow l_{min}$
while $l+1 \leq l_{max}$ **do**
for $c \in \Omega$ **do**
for $i = 1$ to $|M|$ **do**
$m \leftarrow$ a string concatenated c with the i-th element of M
$z \leftarrow Z(m, S)$
if $z_i < z$ **then**
// z_i is the i-th element of Z
Add m to the end of the array M^*
Add z to the end of the array Z^*
end if
end for
end for
$M \leftarrow M^*$, Add elements of M^* to M'
$Z \leftarrow Z^*$
Remove all elements of M^* and Z^*
$l \leftarrow l+1$
end while

2.3.2 クラスタリング

推定したモチーフをもとに配列データのクラスタリングを行う．長さが異なるモチーフの Z スコアを比較するために，モチーフ m' の Z スコアを以下の式より正規化する．

$$Z^*(m', S) = \frac{Z(m', S) - \hat{\mu}_{|m'|}}{\hat{\sigma}_{|m'|}} \tag{8}$$

ここで，$|m'|$ を m' の長さ，$\hat{\mu}_{|m'|}$ と $\hat{\sigma}_{|m'|}$ を，長さ $|m'|$ のモチーフの Z スコアの平均と標準偏差とする．モチーフを $Z^*(m', S)$ により降順に整列し，モチーフをもとに配列をクラスタリングする．モチーフに基づいたクラスタリングの詳細を以下に示す．

(i) $i \leftarrow 1$ とする．

(ii) i 番目のモチーフをクラスターシードとして選択する．配列データから i 番目のモチーフを有する配列を抽出し，i 番目のクラスターに含める．ここで，i 番目のモチーフを含む配列が存在しないとき，クラスタリングを終了する．配列データから抽出した配列を取り除く．

(iii) $i \leftarrow i+1$ として (2) に戻る．

Algorithm 2 Procedure for clustering

Require: Unique sequences of sequence data S sorted by frequency in descending order: S';
The motifs sorted by $Z^*(m', S)$ in descending order: M'';
Ensure: The clusters $C = \{C_1, C_2, \ldots\}$;
$i \leftarrow 1$
while $|S'| > 0$ and M'' has the i-th element **do**
for $j = 1$ to $|S'|$ **do**
if m''_i in s'_j **then**
// m''_i is the i-th element of M''
// s'_j is the j-th element of S'
Add s'_j to the cluster C_i
else
Add s'_j to the array S''
end if
end for
$i \leftarrow i+1$
$S' \leftarrow S''$, Remove all elements of S''
end while

3. 実験・結果

ヒト ES 細胞 H1 株を標的として，SELEX 法を 5 ラウンドまで行なって得られた配列データ [14] を用いて，既存手法と提案手法の速度と精度を評価する．本実験で用いた [14] の 5 ラウンド目のデータは，実験的に結合することが確認された配列と結合しないことが確認された配列を含む 15,327,604 個のデータからなり，そのうち一意な配列は 4,381,160 個存在する．実験で用いる計算機は，OS が Ubuntu 16.04 (Xenial Xerus) 64 bit で，CPU が Intel(R) Xenon(R) CPU E5-1650v4@3.60Ghz，メモリが 64 GB である．実験を行う際，FASTAptamer-Cluster における LD の閾値は，FASTAptamer のユーザガイドに従い $d = 7$ とする．APTANI は，配列の出現頻度による標準のフィルタリング機能を用いず，配列の長さをデータに適した値に設定する．AptaCluster は，既定の設定を用いる．また，提案手法においては，$l_{min} = 5$，$l_{max} = 10$ とする．

3.1 処理時間

出現頻度 f $(\geq 1, 10, 100)$ でフィルタリングした配列データを用いて，既存手法と提案手法の処理速度を比較する．出現頻度とは，配列データ内に同じ配列が重複して存在す

© 2019 Information Processing Society of Japan

表 1　異なる大きさのデータに対する各手法の処理時間
Table 1　Processing time of each method for different data size.

Method	All sequences	Sequences (≥ 10)	Sequences (≥ 100)
FASTAptamer	DNF[1]	15h51m6s	17m57s
AptaCluster	7m53s	1m30s	46s
APTANI	DNF[2]	32m52s	1m47s
Ours	4h50m8s	9m11s	35s

[1] FASTAptamer did not finish the process in 7 days with all sequences.
[2] APTANI exited with an error message after the prediction of secondary structure which took 25 hours.

る数を表す．出現頻度 $f \geq 10$ の核酸配列は 8,799,219 個であり，そのうち一意な配列は 156,587 個である．出現頻度 $f \geq 100$ の核酸配列は 4,947,522 個であり，そのうち一意な配列は 6,193 個である．既存手法と提案手法の処理時間を **表 1** に示す．$f \geq 100$ でフィルタリングしたデータに対しては，提案手法の処理時間が 35 秒で最も短かった．$f \geq 10$ でフィルタリングしたデータに対して，処理時間は AptaCluster が 1 分 30 秒と最も短く，提案手法が 9 分 11 秒と 2 番目に短かった．全データに対して処理を行ったとき，FASTAptamer-Cluster は，7 日以上計算しても終了せず，APTANI はエラーを生じた．提案手法の処理時間は 4 時間 50 分 8 秒であった．提案手法は，既存手法の中で 2 番目に高速に処理することができ，大量の配列データに対しても処理を行うことができた．

3.2　精度

出現頻度 $f \geq 10$ でフィルタリングした配列データに対して，既存手法と提案手法によりクラスタリングを行い，標的分子と結合する配列と結合しない配列のクラスタリング結果から，クラスタリングの精度を比較する．クラスタリング結果を **表 2** に示す．表 2 の項目は，標的分子と結合する・しないことが確認されている配列と，データ内における出現頻度の順位，出現頻度，標的分子と結合するか否か，各手法におけるクラスターの順位が示されている．ここで，クラスター順位 (ranking) は，各クラスターに対して上位のものから順に割り当てた番号である．また，AptaCluster と APTANI における丸括弧内の “Freq.”・“Div.” は，クラスターに含まれる配列の出現頻度・配列の多様性（一意な配列の数）によってクラスター順位を決定したことを表す．クラスタリング結果において，上位のクラスターに含まれる配列がアプタマーの至適配列として選択される．FASTAptamer-Cluster は結合する配列を 6 番目以降，AptaCluster (Freq.) は 7 番目以降，AptaCluster (Div.) は 5 番目以降，APTANI (Freq.) は 7 番目以降，APTANI (Div.) は 870 番目以降の複数のクラスターに分類しているが，提案手法は，1 番目と 5 番目のクラスターに分類している．FASTAptamer-Cluster や，AptaCluster，APTANI はデータ内で出現頻度の最も高い配列を 2 番目以上の上位のクラスターに分類しているが，これは標的分子に結合しない配列である．提案手法は，この配列を 26 番目のクラスターに分類し，結合する配列を含むクラスターよりも下位のクラスターに分類している．したがって，提案手法は，既存手法よりも正確に標的分子と結合する・結合しない配列を分類できているといえる．

また，この結果をもとに計算した両手法の受信者操作特性 (Receiver Operating Characteristic: ROC) 曲線と曲線下面積 (Area Under the Curve: AUC) を **図 3** に示す．ここで，カットオフ値以下の順位・より大きい順位をもつクラスターに配列が分類されていることを，それぞれ陽性 (positive) ・陰性 (negative) とする．また，結合配列が陽性・陰性と判断されるとき，それぞれ真陽性 (True Positive: TP) ・偽陽性 (False Positive: FP) とし，結合しない配列が陽性・陰性と判断されるとき，それぞれ偽陰性 (False Negative: FN) ・真陰性 (True Negative: TN) とする．このとき，真陽性率 (True Positive Rate: TPR) と偽陽性率 (False Positive Rate: FPR) はそれぞれ以下の式で求められる．

$$TPR = \frac{TP}{TP + FN},\ FPR = \frac{FP}{FP + TN} \tag{9}$$

ROC 曲線は，縦軸に真陽性率 (TPR)，横軸に偽陽性率 (FPR) をとり，カットオフ値を変動させることで得られる曲線であり，AUC が 1 に近いほど，精度が高い．図 3 より，提案手法の AUC は，1 であり，既存手法よりも高精度であった．以上より，提案手法を用いることで高速かつ正確にアプタマーの至適配列を推定することができる．

4.　まとめ・今後の課題

本稿では，SELEX 法から得られる大量の核酸配列データに対する高速なクラスタリング手法を提案した．長さの短いモチーフから探索し，推定されたモチーフを伸長しながら探索を行うことで可変長のモチーフを高速に探索し，推定されたモチーフをもつ配列同士でクラスターを形成する手法を提案した．精度評価実験を通して，速度と精度における提案手法の有効性を示した．今後は，他のラウンドのデータや，標的分子の異なる SELEX 法から得られた配列データを用いて，性能評価実験を行う予定である．

参考文献

[1] Ellington, A. D. and Szostak, J. W.: In vitro selection of RNA molecules that bind specific ligands, *Nature*, Vol. 346, pp. 818–822 (1990).
[2] Song, S., Wang, L., Li, J., Fan, C. and Zhao, J.: Aptamer-based biosensors, *TrAC Trends in Analytical Chemistry*, Vol. 27, No. 2, pp. 108–117 (online), DOI: https://doi.org/10.1016/j.trac.2007.12.004 (2008).
[3] Shukla, D., Namperumalsamy, P., Goldbaum, M. and Cunningham, E. T.: Pegaptanib sodium for ocular vascular disease, *Indian journal of ophthalmology*, Vol. 55, No. 6, pp. 427–430 (online), DOI: 10.4103/0301-4738.36476 (2007).

© 2019 Information Processing Society of Japan

表 2　結合する・しない配列のクラスター順位

Table 2　Cluster ranking for binding/non-binding sequences.

Sequence information				Cluster ranking					
Sequence	Ranking	Frequency	Binding	FASTAptamer	AptaCluster (Freq.)	AptaCluster (Div.)	APTANI (Freq.)	APTANI (Div.)	Ours
aggagggg**GACTT**aggactgggtttaggg	6	92237	Yes	6	7	5	7	870	5
agggTATGGACTTCgacgtctcggctgaa	24	20057	Yes	15	17	15	15	699	1
cgcacaggaaggTATGGACTTCgacgttt	63	8750	Yes	24	64	65	58	290	1
ggTATGGACTTCgacgtcttctgacctaa	82	6753	Yes	15	81	72	68	2188	1
gaaaTATGGACTTCgatacgccggctgag	255	1483	Yes	60	229	112740	102	626[2]	1
agtatctatcc**GACTT**ggatttacgttcg	8459	84	Yes	546	9921	28056	1993	626[2]	5
tatcc**GACTT**ggatggctgagcaaggcta	100914	15	Yes	731	94490	125262	2038	626[2]	5
aggagggg**GACTT**aggactgggtttatga[1]	281478	4	Yes	NA	NA	NA	NA	NA	NA
gcaggtgtggtttgctgagg**TGGGCC**ctg	1	583447	No	1	1	2	1	125	26
tttggtttgctg**TATGG**tgggctctgtta	8	70095	No	7	8	10	8	916[2]	16
gtgagggtg**AGGAC**aggttagcgtggtgg	10	51669	No	9	11	9	16	916[2]	54
ggtgaggcg**GACGT**atcttttagcaaatc	12	45038	No	10	12	13	13	520	41
tcgcttgaacggggaactactcca**GACGT**	23	20380	No	14	21	23	45	2270	41
gTGGGCgcacttagacggggtgatcgtaa	375	831	No	75	335	76783	387	1739	37
ACTTAtttgtcttaagtggcgggtcaatg	398	771	No	78	238	556	460	2188	47
gggtcc**CTTCG**gggtgacgatggtatcta	520	504	No	107	466	120874	1758	2253	11
ggt**GTGGG**agggtcgtattgtgtcctgt	3847	126	No	388	4568	59849	92	1	66
cttatttgtgtttagtggcgggc**GTTTG**t	29324	41	No	50	539	110	44	323	92
ctatttg**TTCTA**gtggcggtcatctaagg	44000	31	No	50	9134	4859	2043	2253	88

Underlined capital letters represent estimated motifs.

[1] This sequence was filtered due to the frequency is less than the cutoff.

[2] These cluster rankings are just tied, those sequences are not grouped in the same cluster.

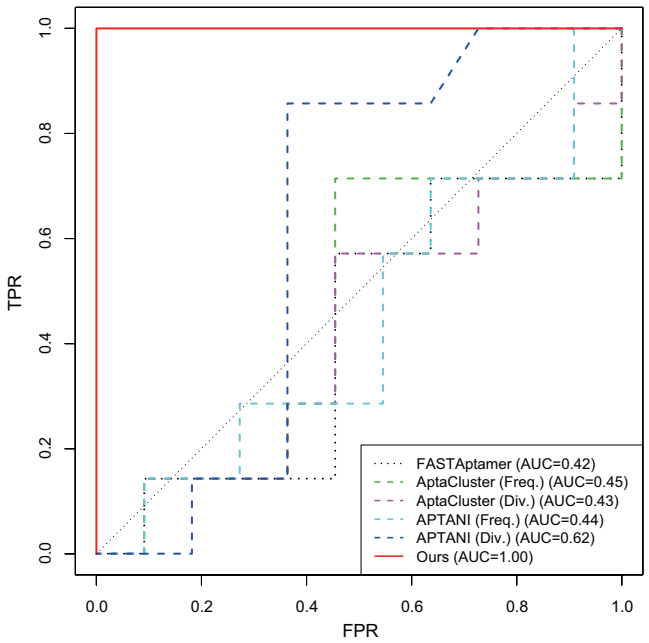

図 3　各クラスタリング手法の ROC 曲線

Fig. 3　ROC curves of each clustering method. "Freq." and "Div." in the parenthesis after AptaCluster and APTANI mean the cluster ranking with frequency and diversity (the number of unique sequences) in the cluster for each method.

[4] Bunka, D. H., Platonova, O. and Stockley, P. G.: Development of aptamer therapeutics, *Current Opinion in Pharmacology*, Vol. 10, No. 5, pp. 557–562 (online), DOI: https://doi.org/10.1016/j.coph.2010.06.009 (2010).

[5] Tuerk, C. and Gold, L.: Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase, *Science*, Vol. 249, No. 4968, pp. 505–510 (online), DOI: 10.1126/science.2200121 (1990).

[6] 田中隆明，村松正實：基礎分子生物学（第 4 版），東京化学同人 (2016).

[7] Bowser, M. T.: SELEX: Just another separation?, *Analyst*, Vol. 130, No. 2, pp. 128–130 (2005).

[8] Acinas, S. G., Sarma-Rupavtarm, R., Klepac-Ceraj, V. and Polz, M. F.: PCR-Induced Sequence Artifacts and Bias: Insights from Comparison of Two 16S rRNA Clone Libraries Constructed from the Same Sample, *Appl. Environ. Microbiol.*, Vol. 71, No. 12, pp. 8966–8969 (2005).

[9] Shao, K., Ding, W., Wang, F., Li, H., Ma, D. and Wang, H.: Emulsion PCR: A High Efficient Way of PCR Amplification of Random DNA Libraries in Aptamer Selection, *PloS one*, Vol. 6, No. 9, p. e24910 (online), DOI: 10.1371/journal.pone.0024910 (2011).

[10] Alam, K. K., Chang, J. L. and Burke, D. H.: FASTAptamer: A bioinformatic Toolkit for High-throughput Sequence Analysis of Combinatorial Selections, *Molecular therapy. Nucleic acids*, Vol. 4, No. 3, p. e230 (online), DOI: 10.1038/mtna.2015.4 (2015).

[11] Hoinka, J., Berezhnoy, A., Sauna, Z. E., Gilboa, E. and Przytycka, T. M.: AptaCluster - A Method to Cluster HT-SELEX Aptamer Pools and Lessons from its Application, *Research in Computational Molecular Biology*, Vol. 8394, pp. 115–128 (2014).

[12] Caroli, J., Taccioli, C., Fuente, A. D. L., Serafini, P. and Bicciato, S.: APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data, *Bioinformatics*, Vol. 32, pp. 161–164 (2016).

[13] Blackshields, G., Sievers, F., Shi, W., Wilm, A. and Higgins, D. G.: Sequence embedding for fast construction of guide trees for multiple sequence alignment, *Algorithms for Molecular Biology*, Vol. 5, No. 1, p. 21 (online), DOI: 10.1186/1748-7188-5-21 (2010).

[14] Jiang, P., Meyer, S., Hou, Z., Propson, N. E., Soh, H. T., Thomson, J. A. and Stewart, R.: MPBind: A meta-motif-based statistical framework and pipeline to predict binding potential of SELEX-derived aptamers, *Bioinformatics*, Vol. 30, No. 18, pp. 2665–2667 (2014).

© 2019 Information Processing Society of Japan