

HT-SELEX法を用いた核酸アプタマー推定のための クラスタリング手法の高速化

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 巖²
青木 孝文¹

概要：アプタマーは、高親和性・特異性により標的分子と結合する核酸分子のことであり、Systematic Evolution of Ligands by EXponential enrichment (SELEX) 法と次世代シーケンシング (Next-Generation Sequencing: NGS) を組み合わせた High-Throughput SELEX (HT-SELEX) 法を用いて核酸ライブラリの中から選択される。HT-SELEX 法から得られる大量の配列データには、標的に対する親和性が低い配列も多く含まれている。それらの中からアプタマー候補を選択するために、Fast Motif-Based Clustering method (FMBC) を開発した。FMBC においてモチーフ推定が全体の処理時間の大半を占めるが、各モチーフのスコア算出が互いに独立であるため、並列化が可能である。本研究では、並列処理により FMBC を高速化する。公開データを用いた評価実験により、従来よりも最大で約 50 倍に高速化したことを示す。

キーワード：アプタマー，HT-SELEX 法，次世代シーケンシング，クラスタリング，並列処理，高速化

Accelerating a clustering method for aptamer estimation using HT-SELEX

1. はじめに

アプタマー [1] とは、高親和性・特異性により、標的分子と結合する短い一本鎖の核酸分子 (RNA または DNA) のことである。アプタマーは、タンパク質 [2] や低分子 [3]、イオン [4]、細胞 [5] などの様々な分子に対して、強い結合力と特異な選択性をもつため、診断薬や治療薬 [6]、[7]、バイオセンサー [8]、[9] 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 [10] が用いられてきた。近年では、SELEX 法に次世代シーケンシング (Next-Generation Sequencing: NGS) を組み合わせた High-Throughput SELEX (HT-SELEX) 法が広く用いられている [11]。HT-SELEX 法の手順を 図 1 に示す。HT-SELEX 法は、(i) 初期の核酸ライブラリーの作成、(ii)

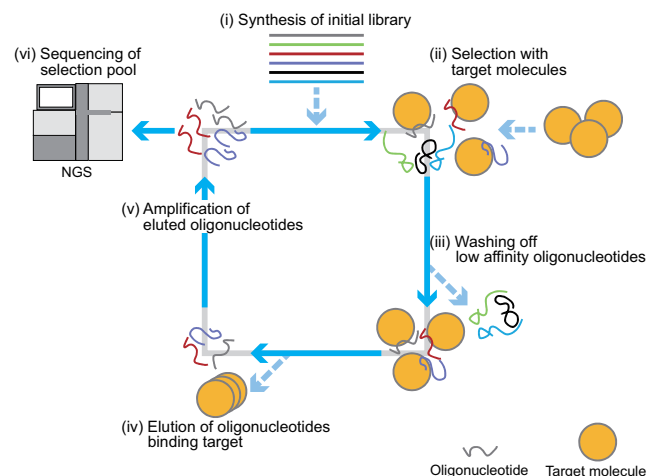


図 1 HT-SELEX 法の手順

Fig. 1 HT-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

標的分子による核酸分子の選択、(iii) 標的分子に対する親和性が低い核酸分子の洗浄、(iv) 標的分子と結合している核酸分子の溶出、(v) ポリメラーゼ連鎖反応 (Polymerase Chain Reaction: PCR) による溶出した核酸分子の増幅、

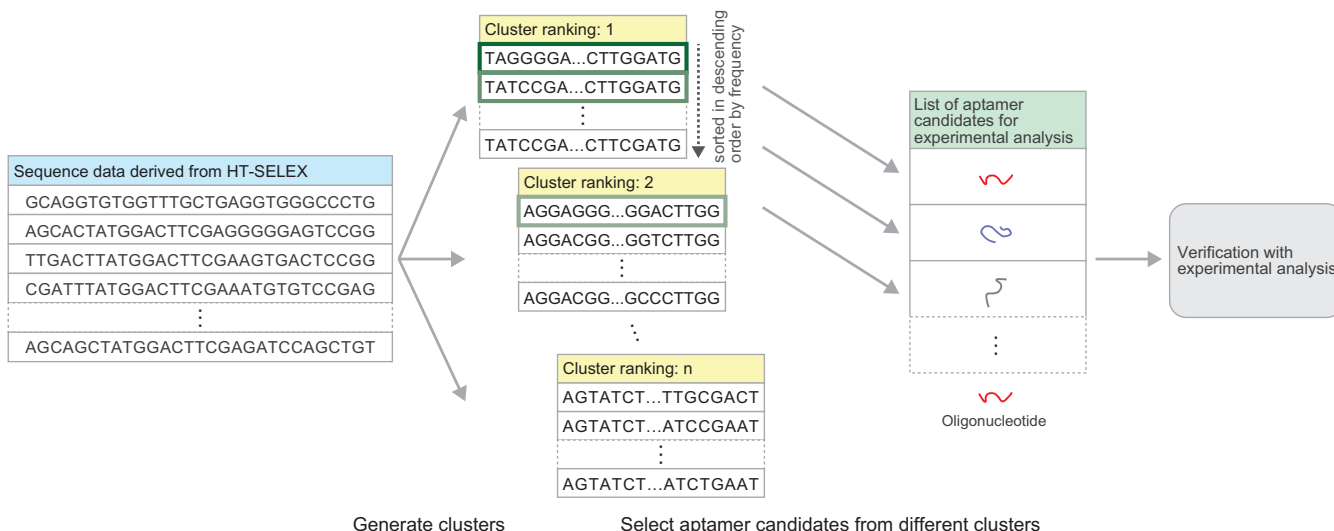


図 2 クラスタリング結果から実験解析までの手順
Fig. 2 Procedure from clustering results to experimental analysis.

(vi) NGS による増幅させた核酸分子のシーケンシングの 6 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 $10^{14} \sim 10^{15}$ 個の核酸分子が格納される [12]。(ii)~(v) を 1 ラウンドとして、これを繰り返すことで最終的に、標的分子に対して親和性の強い核酸分子が試験管内に濃縮される。

HT-SELEX 法からは、各ラウンドにおける大量の核酸分子の塩基配列を得ることができ、これらの大量の配列データからアプタマーの配列を推定することができる。しかし、HT-SELEX 法から得られる大量の配列データには、標的に対する親和性が低い配列も多く含まれ、これらの配列は、PCR パイアス [13], [14] や標的分子以外への結合により、濃縮されたと考えられる。また、配列データ内で出現頻度の高いすべてのアプタマーが標的分子と結合するわけではないため、配列データ内の核酸分子が標的分子に対して高い親和性を有しているかを実験的に評価する必要がある [15]。一方で、HT-SELEX 法から得られる配列データには、 $10^6 \sim 10^7$ 個程度の大量の配列が含まれており、それら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、配列データから、標的分子に対して親和性が低い配列をふるい分け、アプタマーの至適配列を選択し、標的分子との結合親和性を評価するための候補とする必要がある。

図 2 に、HT-SELEX 法により得られた配列データから結合評価を行う候補配列を選択する手順を示す。配列データをクラスタリングし、生成されたクラスターに順位をつけ、上位のクラスターの代表配列を至適配列とし、結合評価の候補配列とする。

HT-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 [16] や AptaCluster

[17], APTANI [18] 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 [19] 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には設計されていない。そこで、HT-SELEX 法から得られた大量の配列データから、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開発した [20]。この手法を Fast Motif-Based Clustering method (FMBC) と呼ぶ。FMBC は、文献 [21] の公開データを用いた性能評価実験を通して、AptaCluster に次いで 2 番目に高速であり、4 つの従来手法に比べ高精度に配列データをクラスタリングできることを示した。FMBC は、モチーフ推定とクラスタリングの 2 つの処理から構成され、モチーフの推定が全体の処理時間の大半を占めている。モチーフ推定において各モチーフのスコアを算出する処理は、互いに独立であることから並列化が可能である。そこで、本稿では、並列処理によって FMBC を高速化する。公開データを用いた評価実験を通して、並列処理の有効性を示す。

2. FMBC のアルゴリズム

FMBC は、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、推定したモチーフをもとに効率的にクラスタリングを行う手法である。2.1 では、FMBC におけるモチーフ推定アルゴ

リズムについて説明する．2.2 では，推定モチーフに基づいたクラスタリングのアルゴリズムについて説明する．

2.1 モチーフ推定

配列データ内に現れる部分配列を探索し，統計量 Z スコアをもとにモチーフを推定する． Z スコアは，ある観測値が平均値からどれくらい離れているかを示す値であり，観測値と平均値の差を標準偏差で割ることにより求まる．FMBC は Z スコアを，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際の指標として用いる． Z スコアを計算するための確率分布には，配列データにおける特定の部分配列を含む配列の出現確率を用いる．2.1.1 では，配列データにおける特定の部分配列を含む配列の出現確率を定義する．2.1.2 では，モチーフを推定する際の指標となる Z スコアを定義する．2.1.3 では， Z スコアをもとにモチーフを推定するアルゴリズムについて説明する．

2.1.1 配列データにおける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． m の i 番目の文字を $m[i]$ と表す． Ω を文字の集合とする．塩基を扱う場合は $\Omega = \{A, C, G, T(U)\}$ となる． s の i 番目の文字が c ($c \in \Omega$) 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 $p(c)$ ($c \in \Omega$) 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．

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

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， $\mathcal{T} = \{A, ATA\}$ となる． t の長さを $|t|$ とし， t の i 番目の文字を $t[i]$ と表すとき，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．

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

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 $P_a(m, L)$ は，以下の漸化式で表される．

$$P_a(m, L) = \begin{cases} 0 & (L < l) \\ P_a(m, L-1) + Q(m)[1 - P_a(m, L-l)] - \sum_{t \in \mathcal{T}} \frac{Q(m)}{q(t)} [P_a(m, L-l+|t|) - P_a(m, L-l+|t|-1)] & (l \leq L) \end{cases} \quad (3)$$

HT-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 により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 $S = \{s_1, s_2, \dots, s_{|S|}\}$ 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．

$$P_d(m, S) = \frac{1}{|S|} \sum_i^{|S|} P_a(m, |s_i|) \quad (4)$$

ここで， $|s_i|$ は配列 s_i の長さを表す．

2.1.2 Z スコアの定義

2.1.1 で定義した $P_a(m, L)$ は，長さ L の配列が部分配列 m を含むか含まないかのベルヌーイ試行の確率を表しており，確率分布 $P_d(m, S)$ は，データ数 $|S|$ を試行回数としたときの二項分布を表している．したがって， Z スコア $Z(m, S)$ は $P_d(m, S)$ と部分配列 m を含む配列の観測数 F_m を用いて，以下の式により表される．

$$Z(m, S) = \frac{F_m - |S|P_d(m, S)}{\sqrt{|S|P_d(m, S)[1 - P_d(m, S)]}} \quad (5)$$

2.1.3 モチーフの推定手順

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 $S = \{s_1, s_2, \dots, s_{|S|}\}$ 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，各文字の出現確率を以下の式により推定する．

$$\hat{p}(c) = \sum_{i=1}^{|S|} \frac{n_i^c}{|s_i|} \quad (6)$$

各文字の出現確率の推定値 $\hat{p}(c)$ 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 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．

図 3 に示すように，FMBC のモチーフ推定では，ある部分配列をモチーフと推定すると，その部分配列に任意の 1 塩基を付加し伸長した部分配列をモチーフと推定するか判定する（分枝操作）．伸長した部分配列の Z スコアが伸長

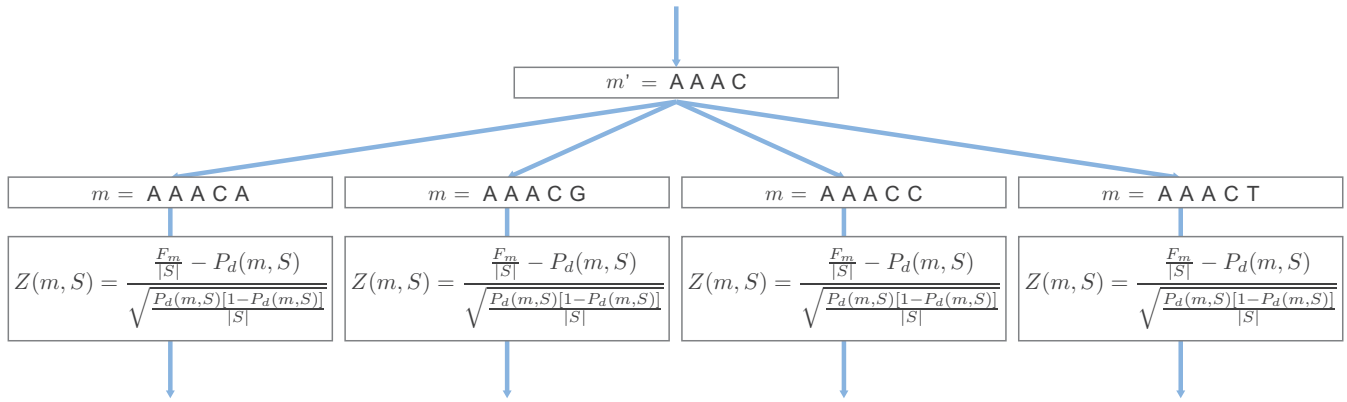


図 4 並列化した Z スコア計算の説明

Fig. 4 Illustration of parallelized Z score calculation.



図 3 分枝限定法を用いて探索する部分配列を減らす説明

Fig. 3 Illustration of reducing subsequences to be search using branch and bound.

前よりも小さい場合、その部分配列を伸長して得られる部分配列の探索を行わず（刈り込み）、伸長前より大きい場合はその部分配列をモチーフと推定する（限定操作）。分枝限定法の考え方を取り入れることにより、探索する部分配列を減らすことでアルゴリズムを高速化している。

2.2 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。

$$Z^*(m', S) = \frac{Z(m', S) - \hat{\mu}_{|m'|}}{\hat{\sigma}_{|m'|}} \quad (7)$$

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。

3. FMBC の実装と並列化

FMBC は、(i) 配列の出現頻度の計算、(ii) モチーフの推定、(iii) 推定モチーフをもとにしたクラスタリングの 3 つの処理を実行する。(i) 配列の出現頻度の計算では、FASTA 形式、または FASTQ 形式の配列データを入力として、配列データに含まれる一意な配列の出現頻度を計算する。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る数を表す。配列データを一意な配列とその出現頻度に要約することでデータを処理する際のメモリ使用量を削減する。(ii) モチーフの推定では、2.2 で説明したアルゴリズムを用いて、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する。(iii) 推定モチーフをもとにしたクラスタリングでは、2.2 で説明したアルゴリズムを用いて、配列データ内の一意な配列をクラスタリングする。

FMBC において、(i) (iii) は、逐次処理であるが、図 4 に示すように、(ii) における長さ l の部分配列の Z スコアを計算する処理は互いに独立であり、並列化が可能である。Z スコア算出では、配列データ内の一意な配列それぞれに対して、ある部分配列が含まれているかどうかを判定する必要があり、一意な配列の数が多ければ処理に時間がかかる。また、探索する部分配列が多いほど、全配列データを参照する回数が増えるため処理時間が長くなる。分枝限定法の考え方により探索する部分配列を減らすことに加えて、全配列データを参照し各部分配列を有する配列数を計算する処理を並列化することによってさらなる高速化が実現できると考えられる。

4. 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 [21] を用いて、速度

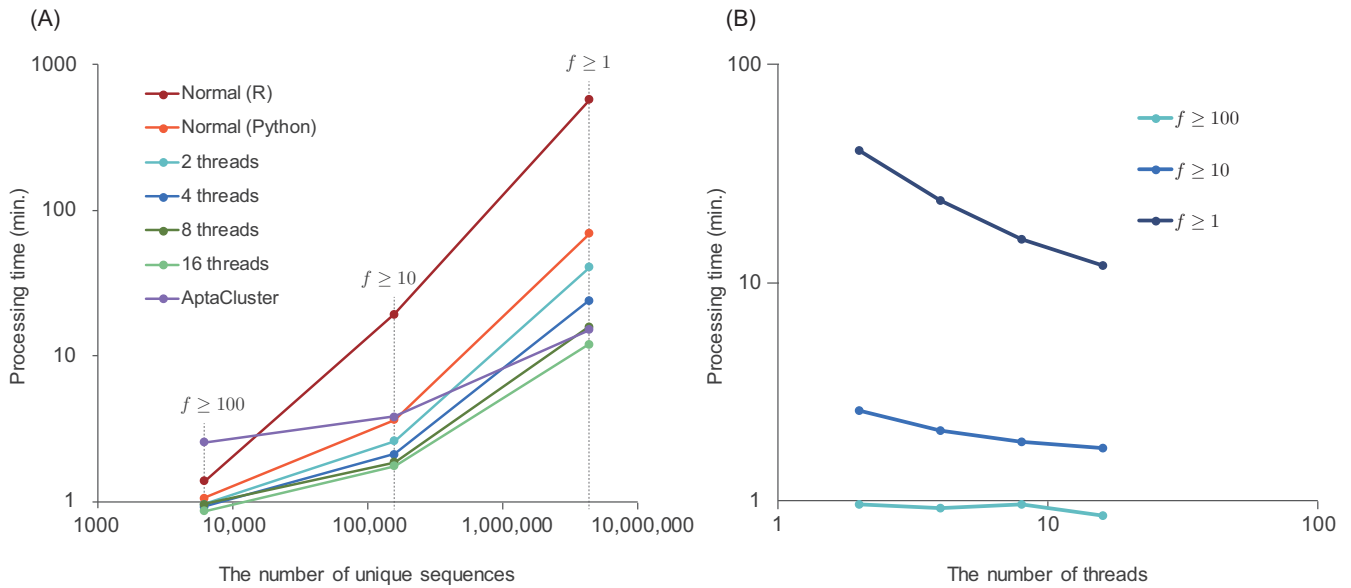


図 5 (A) 異なる大きさのデータに対する各プログラムの処理時間と (B) 異なるスレッド数に対する処理時間

Fig. 5 Processing time of each program for different data size (A) and using different numbers of threads (B).

評価を行う。本実験で用いた [21] の 5 ラウンド目のデータにおける総リード数は 15,327,604 であり、そのうち一意な配列の数は 4,381,160 である。実験で用いる計算機は、OS が CentOS 6.10 で、CPU が Intel(R) Xenon(R) CPU E5-2680@2.70Ghz(8 コア)2 基、メモリが 132 GB である。従来のプログラムと FMBC に並列処理を実装したプログラムを用いて実験を行う。従来のプログラムは R によって実装されているが、本実験では新たに Python によって実装する。並列化には、Python の標準ライブラリーに含まれる multiprocessing を用いる。各プログラムの処理時間は、プログラムの開始から終了までの時間である総経過時間とする。また、[20] において処理時間が従来のプログラムよりも高速であった AptaCluster [17] も比較対象として用いる。AptaCluster は、配列データを扱いやすいデータに変換する事前処理 (-parse) と変換したデータを用いたクラスタリング (-cluster) を実行することで配列データのクラスタリングを行うため、処理時間は事前処理とクラスタリングの総経過時間の和とする。各手法の設定において、AptaCluster は既定の設定を用い、FMBC においては、 $l_{min} = 5, l_{max} = 10$ とする。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 $f \geq 10$ 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。FMBC に並列処理を実装したプログラムにおいて、処理を割り当てるスレッド数を 2, 4, 8, 16 としたときのそれぞれの処理時間を測定する。

従来のプログラムと今回実装したプログラム、AptaCluster の処理時間を図 5 (A) に示す。全データに対して処理を行ったとき、従来のプログラム (normal (R)) の処理時間は 574.02 分で、Python により並列化せず実装したプログラム (normal (Python)) の処理時間は 69.55 分、AptaCluster の処理時間は 15.08 分であった。また、並列処理を実装したプログラムの各スレッド数での処理時間 (2, 4, 8, 16 threads) は、スレッド数を 16 としたとき 12.08 分で最も高速であった。これは、従来のプログラムよりも最大で約 50 倍高速であり、Python により並列化せず実装したプログラムよりも約 6 倍高速であり、AptaCluster よりも高速であった。 $f \geq 10$ でフィルタリングしたデータに対しては、並列化を実装したプログラムのスレッド数を 16 としたとき 1.75 分で最も高速であり、AptaCluster の処理時間 3.81 分よりも高速であった。 $f \geq 100$ でフィルタリングしたデータに対して、FMBC を実装したプログラムの処理時間はいずれも 1 分前後であり、AptaCluster の処理時間である 2.55 分よりも高速であった。FMBC に並列処理を取り入れることで従来よりも高速化することができた。

処理時間とスレッド数の関係を図 5 (B) に示す。処理する配列数が多いほど、割り当てるスレッド数により処理時間が大きく変化した。配列数が 6,193 個 ($f \geq 100$) と少ない場合は、スレッド数を増やしても処理時間は短縮されなかった。配列数が 156,587 個 ($f \geq 10$) の場合にも、スレッド数を多くするほど処理時間を短縮することができたが、全体の配列に対して処理を行った場合に比べて変化は小さかった。これは、処理する配列数が少ない場合に、一意な

配列の出現頻度・配列長を計算する前処理や、推定したモチーフによる配列のクラスタリングに要する時間の方が、並列化したモチーフのスコア算出に要する時間よりも長くなったためと考えられる。また、データの入出力に時間がかかるため、並列化により演算速度が向上しても全体の処理速度の改善には至らなかったと考えられる。HT-SELEX法から得られるデータのような、膨大なデータを処理する際には、FMBCの並列化は処理速度を大幅に高速化でき、より効率的なアプタマー開発に貢献できると考えられる。

5. まとめ・今後の課題

本稿では、HT-SELEX法から得られる大量の核酸配列データに対するクラスタリング手法であるFMBCに並列処理を実装することでさらなる高速化を実現した。FMBCのモチーフ推定において、同じ長さの部分配列のZスコアを計算する処理は独立であるため並列化を行った。公開データを用いた評価実験を通して、FMBCに並列処理を実装することで処理時間が改善されることを示し、並列処理の有効性を示した。今後は、標的分子の異なるHT-SELEX法から得られた配列データを用いてFMBCの性能評価を行うことや、HT-SELEX実験から得られたデータから、FMBCを用いて至適配列を選択し、結合評価を行うことでその有効性を示す予定である。

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] Bock, L. C., Griffin, L. C., Latham, J. A., Vermaas, E. H. and Toole, J. J.: Selection of single-stranded DNA molecules that bind and inhibit human thrombin, *Nature*, Vol. 355, No. 6360, p. 564 (1992).
- [3] Zimmermann, G. R., Wick, C. L., Shields, T. P., Jenison, R. D. and Pardi, A.: Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer, *Rna*, Vol. 6, No. 5, pp. 659–667 (2000).
- [4] Qu, H., Csordas, A. T., Wang, J., Oh, S. S., Eisenstein, M. S. and Soh, H. T.: Rapid and label-free strategy to isolate aptamers for metal ions, *ACS Nano*, Vol. 10, No. 8, pp. 7558–7565 (2016).
- [5] Marton, S., Cleto, F., Krieger, M. A. and Cardoso, J.: Isolation of an aptamer that binds specifically to E. coli, *PLoS ONE*, Vol. 11, No. 4, p. e0153637 (2016).
- [6] 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).
- [7] 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).
- [8] Song, S., Wang, L., Li, J., Fan, C. and Zhao, J.: Aptamer-based biosensors, *Trends in Analytical Chemistry*, Vol. 27, No. 2, pp. 108–117 (online), DOI: <https://doi.org/10.1016/j.trac.2007.12.004> (2008).

- [9] Kaneko, N., Horii, K., Akitomi, J., Kato, S., Shiratori, I. and Waga, I.: An Aptamer-Based Biosensor for Direct, Label-Free Detection of Melamine in Raw Milk, *Sensors*, Vol. 18, No. 10, p. 3227 (2018).
- [10] 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).
- [11] Zhuo, Z., Yu, Y., Wang, M., Li, J., Zhang, Z., Liu, J., Wu, X., Lu, A., Zhang, G. and Zhang, B.: Recent Advances in SELEX Technology and Aptamer Applications in Biomedicine, *International Journal of Molecular Sciences*, Vol. 18, No. 10, p. 2142 (online), DOI: 10.3390/ijms1810214 (2017).
- [12] Bowser, M. T.: SELEX: Just another separation?, *Analytst*, Vol. 130, No. 2, pp. 128–130 (2005).
- [13] 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, *Applied and Environmental Microbiology*, Vol. 71, No. 12, pp. 8966–8969 (2005).
- [14] 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).
- [15] Tolle, F., Wilke, J., Wengel, J. and Mayer, G.: By-product formation in repetitive PCR amplification of DNA libraries during SELEX, *PLoS ONE*, Vol. 9, p. e114693 (online), DOI: 10.1371/journal.pone.0114693 (2014).
- [16] 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).
- [17] 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).
- [18] Caroli, J., Taccioli, C., Fuente, A. D. L., Serafini, P. and Biciato, S.: APTANI: a computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data, *Bioinformatics*, Vol. 32, pp. 161–164 (2016).
- [19] 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).
- [20] 小野貴義, 加藤信太郎, 伊藤康一, 皆川宏貴, 堀井克紀, 白鳥行大, 和賀巖, 青木孝文: 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, 情報処理学会研究報告, Vol. 2019-BIO-57, No. 7, pp. 1–6 (2019).
- [21] 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).