博 士 論 文

HT-SELEX データを用いた アプタマーの設計に関する研究

情報基礎科学専攻
加藤　信太郎

目次

第 1 章

緒言

近年，バイオテクノロジーの発展に伴い，ゲノム，トランスクリプトーム，プロテオームに代表される生命に関するビッグデータの取得が可能となった．このビッグデータを効率よく解析し，生命現象の理解のみならず，医療や創薬などの産業応用に必要な情報を取得するため，バイオインフォマティクスと呼ばれる学問分野の技術が重要となっている．バイオインフォマティクスとは，生物学（バイオロジー）と情報学（インフォマティクス）を組み合わせた学問分野であり，そこでは主に，生物に由来するデータを精度よく高速に解析する手法を開発すること，また，その手法を利用して有用な情報を探し出すことが行われている．対象となるデータとして，たとえばゲノム解析であれば DNA の配列情報，トランスクリプトーム解析であれば遺伝子の発現情報があげられる．一本鎖 DNA や RNA から構成される核酸アプタマー（以下，アプタマーと呼ぶ）と呼ばれる人工核酸も，バイオインフォマティクスの分野で研究されている．アプタマーは塩基配列[*1]が設計できるという特徴を有するが，その設計方法の効率化が課題となっている．そこで本論文では，バイオインフォマティクスの新たな手法を提案し，効率的なアプタマーの設計方法を確立することを目的とする．

アプタマーは分子認識能[*2]を有し，立体構造を形成して標的分子と特異的に結合する核酸分子である．現在，市場で広く利用されている分子認識能を有する物質として抗体があるが，抗体は実験動物や細胞培養により生産を行うため，コストや品質などに課題を抱えている．アプタマーは核酸分子であるため，配列が一度決まれば，化学合成により安価か

[*1] 塩基とは，アデニン (A)，シトシン (C)，グアニン (G)，チミン (T)（RNA であればウラシル (U)）であり，塩基配列とは塩基の並びである．

[*2] 分子が他の分子を見分ける能力のこと．分子認識能を有する生体物質には，アプタマー以外に，抗体や酵素などがある．

つ高い品質で大量に合成することができる．そのため，抗体がもつ上記のような課題がなく，新たな分子認識能を有する物質として注目されている．アプタマーが結合する標的分子の種類は多様であり，タンパク質 [1,2]，細胞 [3]，ウィルス [4]，病原菌 [5]，毒物 [6]，低分子 [7] などと結合することが報告されている．この特性により，アプタマーはバイオセンサー [8,9]，医薬品 [10]，診断薬 [11,12] などに利用されている．

アプタマーは，SELEX (Systematic evolution of ligands by exponential enrichment) [13,14] という実験手法を用いて，さまざまな核酸分子から選別される．2010 年以降，SELEX に次世代シーケンシング (Next-generation sequencing: NGS) を利用し, SELEX から大量の配列情報を取得する方法が利用されている (High-throughput SELEX: HT-SELEX) [15,16]．本論文ではこの大量の配列情報を，HT-SELEX データと呼ぶ．図 1.1 に，HT-SELEX の概要を示す．はじめに，ランダム核酸ライブラリ（ランダムな塩基配列を含む分子数が 10^{14} から 10^{16} 個の核酸分子）を試験管内に準備する（図中の 1）．次に，標的分子と核酸分子を結合させ，結合しない核酸分子を洗浄により除去する．結合した核酸分子を標的分子から溶出し，ポリメラーゼ連鎖反応[*3](Polymerase chain reaction: PCR) により増幅する（図中の 2 から 5）．これを一つのラウンドとして繰り返し実施することで，徐々にアプタマーを試験管内に濃縮させる．通常 8 から 12 ラウンド終えた後，NGS により配列を読み取り，HT-SELEX データを得る（図中の 6）．アプタマーの設計は，この HT-SELEX データを利用して行う．

本論文において，アプタマーは以下に示す二つの工程で設計される．

1. HT-SELEX データのクラスタリングの結果に基づく，アプタマーの候補配列の選択.
2. アプタマーの小型化（アプタマーの配列を可能な限り短くすること）.

ここで，一つ目の工程にあげたクラスタリングの必要性と，精度の高いクラスタリングの条件に関して説明する．アプタマーの配列の決定には，HT-SELEX データから選択したアプタマーの候補配列を化学合成し，標的分子との結合実験による評価を行う必要がある．ただし，結合実験により評価できる配列の数が数十本であるのに対し，HT-SELEX データには 1,000 万本以上の配列が含まれているため，効率よくアプタマーの候補配列を選択する必要がある．HT-SELEX データの中には，標的分子との結合様式が異なる多様なアプタマーの配列や，標的分子とは結合しない核酸分子の配列が混在している．そのた

[*3] 特定の DNA 断片を選択的に増やす方法．ウィルスや細菌の検査にも利用される．

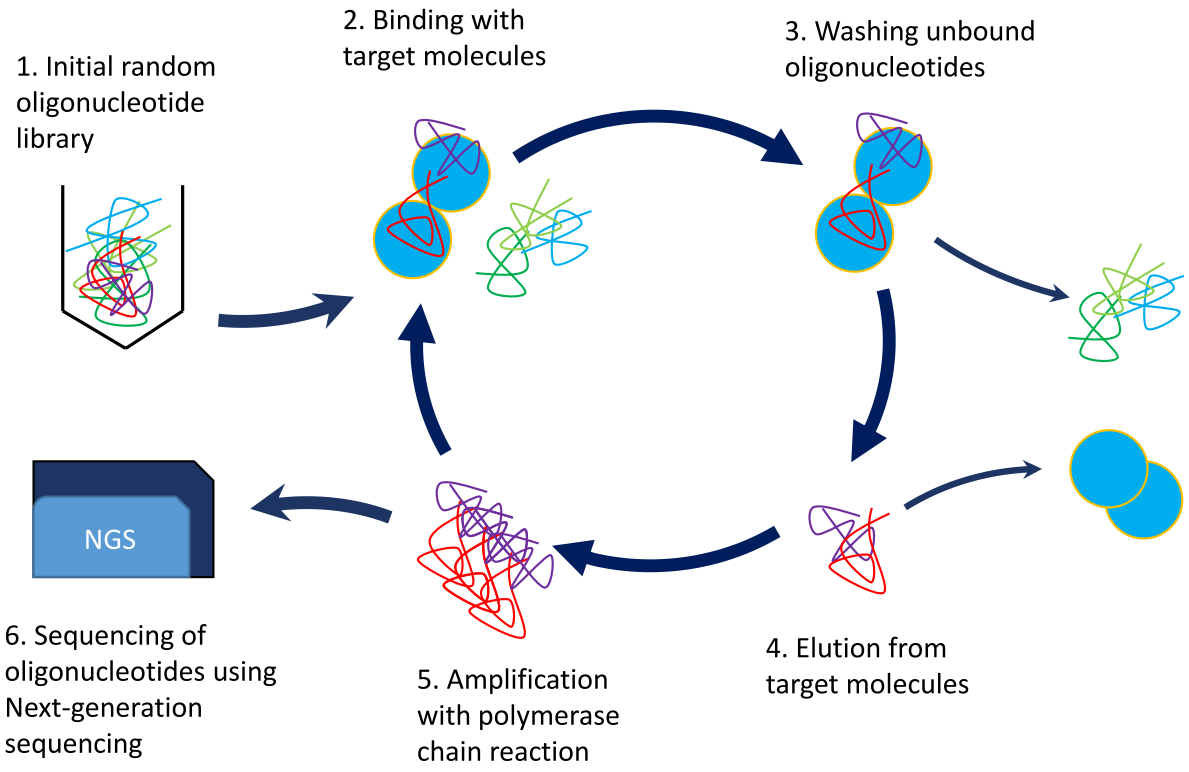

図 1.1　HT-SELEX の概要

め，これらをクラスタリングにより精度良く分けること，得られたクラスタに順位をつけてどのクラスタからアプタマーの候補配列を選択すべきかわかることが，効率よくアプタマーの候補配列を選択することにつながる．HT-SELEX データをクラスタリングにより精度良く分けるためには，アプタマーの標的分子との結合領域を考慮する必要がある．これは，結合領域の配列と立体構造が共通していれば，全長の配列が等しくないアプタマーでも同様の結合様式をとるからである．そのため，HT-SELEX データからアプタマーの結合領域を探し出し，同じ結合領域を有する配列同士でクラスタを作成することが，精度の高いクラスタリングとなる．

HT-SELEX データの従来のクラスタリング手法として，FASTAptamer [17]，AptaCluster [18]，APTANI [19]，AptaTRACE [20] があげられる．また，ゲノム解析のために開発された UCLUST [21] と UNOISE3 [22] も，HT-SELEX データのクラスタリングに利用可能であることが報告されている [23]．FASTAptamer，AptaCluster，APTANI，UCLUST，UNOISE3 は，全長配列の類似性を考慮したクラスタリング手法であり，アプタマーの結合領域を考慮しない．そのため，アプタマーの結合様式が異なる配列でクラス

タを作成してしまう恐れがある．また，これらの手法は配列の頻度に依存したクラスタの順位付けを行うため，頻度は高いが標的分子と結合しない配列（たとえば，PCR で増幅されやすい核酸分子の配列など）で，順位の高いクラスタを作成してしまう恐れもある．AptaTRACE は，結合領域を考慮したクラスタリング手法であるが，推定可能な結合領域の長さに制限があるため，長い結合領域を推定することができない．また，AptaTRACE は複数ラウンドの HT-SELEX データを用いるため，より長い計算時間とシーケンシングのための追加の費用が必要となってしまう．以上を考慮し，本論文では新たなクラスタリング手法 Fast string-based clustering (FSBC) [24–32] を提案する．

FSBC は，高速に任意の長さの結合領域を推定し，推定した結合領域を利用して配列を分けるクラスタリング手法である．費用と計算時間の節約のため，FSBC は単一ラウンドの HT-SELEX データを利用する．FSBC の流れを以下に示す．はじめに，特定の文字列（特定の塩基の並び）を含む配列の出現確率と観測される配列の比率を用いて，文字列のスコアを Z スコアとして定義する．Z スコアが任意の値より大きい文字列を，アプタマーと標的分子との結合領域とする．次に，高速に任意の長さの結合領域を推定する．結合領域の長さが長くなると，塩基を並べて作られる文字列の個数が指数関数的に増大するため，探索領域の削減を行いながら結合領域の推定を行う．最後に，同じ結合領域を含む配列同士でクラスタを作成する．FSBC の有用性を検証するため，実際の HT-SELEX データを用いて評価を行う．評価には，ヒトの胚性幹細胞 (human embryonic stem cell: hESC) [33] と，ヒトインターロイキン 10 レセプター α サブユニット (human interleukin-10 receptor alpha subunit: IL10RA) を標的とした HT-SELEX データ [34,35] を用いる．FSBC と従来手法の計算速度とクラスタリングの精度を比較することで，FSBC の有用性を示す．

候補配列を結合評価しアプタマーの配列を決定した後，二つ目の工程であるアプタマーの小型化を行う．HT-SELEX で得られるアプタマーの長さはおよそ 80 塩基から 100 塩基であり，このうち一部の領域が標的分子と結合している．アプタマーを実用化するために，結合に関与しない領域を削除し，アプタマーの小型化を行う．アプタマーを小型化する理由として，化学合成のコストと合成エラーの削減，標的分子との親和性の向上 [36,37] などがあげられる．アプタマーの小型化に関する研究はあまり行われておらず，現状，経験を有する専門家がアプタマーの二次構造予測[*4]の結果を利用し，勘と経験によりアプタ

[*4] 二次構造とは，核酸分子が形成する立体構造を二次元で表現したもの．二次構造予測とは，二次構造を塩基配列より予測すること．

マーの小型化配列を推定している．そのため，アプタマーの小型化配列の推定に明確な指標がなく，結果に再現性がない．また，アプタマーには構造多様性（多様な立体構造を取り得ること）があり，複数の二次構造を取り得るが，人ではそれらを網羅的に考慮することが難しい．さらに，通常は結合領域が不明瞭なままでアプタマーの小型化配列が推定されているため，結合領域を削除した小型化配列を推定してしまう恐れもある．以上の問題を踏まえて，本論文では新たなアプタマーの小型化配列を推定する手法を提案する [38].

アプタマーは立体構造を形成して標的分子と結合するため，小型化されたアプタマーの結合領域が，もとの長さのアプタマーの結合領域の構造を保持している必要がある．また，精度良く小型化配列を推定するためには，アプタマーの構造多様性を考慮する必要がある．本論文で提案するアプタマーの小型化配列を推定する手法では，これらの条件を最適化問題として定式化する．最適化問題の定式化により明確な指標が定まり，人に依存しない再現性のあるアプタマーの小型化配列の推定が可能となる．また，FSBC により推定したアプタマーの標的分子との結合領域を利用するため，小型化配列の推定において，結合領域を除去してしまう過ちを回避することができる．提案する手法の有用性を検証するため，実際のアプタマーを用いて評価を行う．提案する手法の検証には，血管内皮細胞増殖因子 (Vascular endothelial growth factor: VEGF) と，C 反応性蛋白 (C-reactive protein: CRP) に結合するアプタマーを用いる．

以上述べたように，本論文では効率的なアプタマーの設計方法を確立することを目的とし，HT-SELEX データを高精度に分けるクラスタリング手法とアプタマーの小型化配列を推定する手法を提案する．さらに，提案する二つの手法を，実際の HT-SELEX データとアプタマーを用いて，その有用性を示す．

本論文は，以下の 5 章より構成される．

第 1 章（緒言）は，本論文の背景と目的，および概要を述べたものである．

第 2 章（HT-SELEX データを用いたアプタマーの設計に関する基礎的考察）では，HT-SELEX データを用いたアプタマーの設計に関する基礎的考察に関して述べる．まず，アプタマーと HT-SELEX に関する詳細を述べる．次に，HT-SELEX データのクラスタリングについて述べる．最後に，アプタマーの小型化に関して述べる．

第 3 章（HT-SELEX データのクラスタリング）では，提案するクラスタリング手法である FSBC に関して述べる．はじめに，FSBC の概要に関して述べる．次に，特定の文字列を含む配列の出現確率に関して述べる．続いて，出現確率を用いて文字列のスコアを定義し，スコアの高い文字列を高速に探索する方法を述べる．さらに，選ばれた文字列を利用してクラスタを作成する方法を述べる．また，FSBC を高速化するための並列処理と，

FSBC により得られたクラスタの分布を比較する方法について述べる．最後に，hESC と IL10RA を標的とした HT-SELEX データを用いて FSBC を評価し，その有用性を示す．

第 4 章（アプタマーの小型化）では，提案するアプタマーの小型化配列を推定する手法に関して述べる．まず，アプタマーの小型化配列を推定するための最適化問題を定式化する．次に，最適化問題を解くためのアルゴリズムについて述べる．最後に，VEGF と CRP に結合するアプタマーを用いて，提案する手法の精度を評価し，その有用性を示す．

第 5 章（結言）では，本論文の結言を述べる．

第 2 章

HT-SELEX データを用いたアプタマーの設計に関する基礎的考察

2.1 まえがき

SELEX が開発された 1990 年は，サンガー法によるシーケンシングを利用し，数十の配列情報からアプタマーの候補配列を選択していた．2005 年に新たなシーケンシングが登場し，大量の配列情報の取得が可能となり，サンガー法と比較して次世代シーケンシング (NGS) と呼ばれるようになった．2010 年以降，SELEX に NGS を利用し，大量の配列情報からアプタマーを選択する方法 (HT-SELEX) が実施されるようになった [15, 16]．現在は HT-SELEX のおかげで，多様なアプタマーの候補配列を選択することや，SELEX における早期ラウンドでアプタマーの候補配列を選択することが可能である．

第 1 章で述べたとおり，本論文ではアプタマーの設計を，クラスタリングに基づいたアプタマーの候補配列の選択と，アプタマーの小型化の二つの工程により行う．図 2.1 に，HT-SELEX データを用いたアプタマーの設計と評価の流れを示す．まず，HT-SELEX データをクラスタリングにより分ける（図の 1）．アプタマーが含まれる可能性の高いクラスタよりアプタマーの候補配列を選び，結合実験により評価する．十分な親和性を示すアプタマーの候補配列をアプタマーの配列として決定する（図の 2）．アプタマーの配列を，結合能を損なうことなく，できる限り短くした配列を推定する（図の 3）．推定した配列を結合実験により評価する．結合実験により標的分子と十分な親和性を示した配列を，小型化アプタマーの配列として決定する（図の 4）．以上の工程がアプタマーの設計であり，設計された小型化アプタマーはバイオセンサーや医薬品などの目的に合わせて利用される．

本章では，HT-SELEX データを用いたアプタマーの設計に関する基礎的考察として，

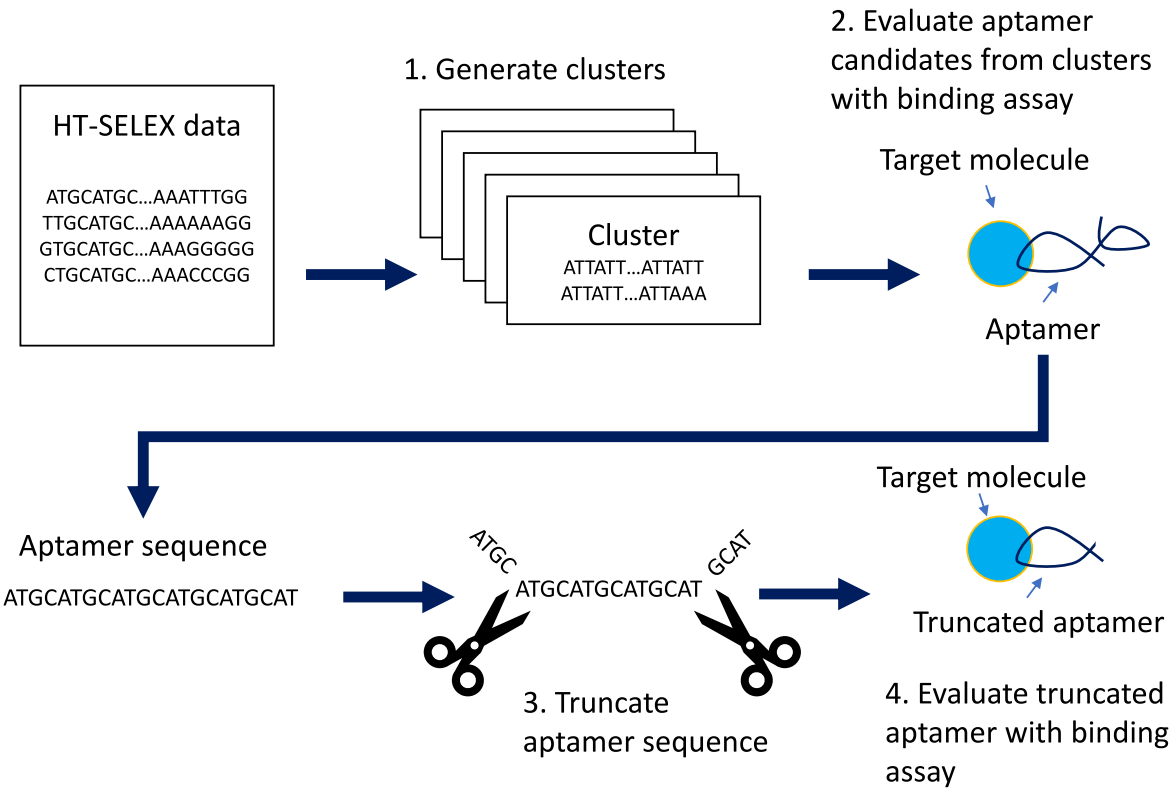

図 2.1　アプタマーの設計と評価の流れ

まず，第 2.2 節で，研究対象であるアプタマーの特性に関して述べる．次に，第 2.3 節で，HT-SELEX の概要に関して述べる．続いて，第 2.4 節で，HT-SELEX データのクラスタリングに関して述べ，第 2.5 節で，アプタマーの小型化について述べる．最後に，第 2.6 節で本章の結びを述べる．

2.2　アプタマーの特性

アプタマーの有用性

アプタマーは一本鎖 DNA もしくは RNA から構成される核酸分子であり[*1]，標的分子に特異的に結合する分子認識能を持つ．現在，市場で幅広く利用されている分子認識能を持つ物質として抗体がある．アプタマーは，抗体と比較して優れた点があるため，新た

[*1] ペプチドから構成されるペプチドアプタマーもあるが，本論文では核酸から構成される核酸アプタマーを対象とする．

表 2.1　アプタマーと抗体との比較

特徴	アプタマー	抗体
特異性	高い	高い
親和性	高い	高い
製造コスト	安価	高価
品質	均一	製造ロット差あり
安定性	常温保存可能	凍結保存が望ましい
標的の種類	幅広い標的分子	抗原抗体反応のあるものに限定
化学修飾の導入	容易	困難
構造スイッチによるセンサー化	可能	不可能

な分子認識能を持つ物質として注目されている．表 2.1 にアプタマーと抗体との比較を示す．

以下に，アプタマーの抗体に対する優位性に関して詳細を説明する．抗体は，抗体を生成する動物（ポリクローナル抗体[*2]の場合）や，ハイブリドーマ細胞[*3]（モノクローナル抗体[*4]の場合）を利用して生産するため，費用が掛かり，また，製造ロットの間でばらつきが出やすい．一方アプタマーは，一度塩基配列が決定すれば，合成装置を利用し安価に均一な品質で大量に合成が可能である．抗体はタンパク質で構成されているため，熱変性により分子認識能が失われる．そのため，抗体は熱変性を防ぐために凍結保存が望ましいが，アプタマーは凍結乾燥させれば常温でも長期保存が可能である．アプタマーは常温で長期保存が可能なため輸送も容易であり，冷蔵庫が使用できない場所での利用にも優れている．

アプタマーが注目されている大きな理由は，幅広い標的分子に対するアプタマーが取得できることである．抗体は生物の防御機能として確立された免疫機構により作成されるため，標的分子が抗原として認識されないと抗体の生成は行われない．たとえば，非常に小さな分子などは抗原として認識されないため，抗体を作成することができない．一方アプタマーにはこのような制限はなく，低分子やさらに小さいイオンを認識するアプタマーの

[*2] 抗原の複数のエピトープを認識する抗体．エピトープとは抗原が抗体と結合する部分．

[*3] ハイブリドーマ法により作成され得た細胞．ハイブリドーマ法は，抗体を生成する B 細胞にがん細胞を融合し，寿命なく抗体を生産させる方法．

[*4] 抗原の一つのエピトープを認識する抗体．多くのモノクローナル抗体が医薬品に利用されている．

報告もされている [39–41]．さらに，アプタマーはテオフィリン[*5]とカフェインのような構造が類似した低分子の違いも認識でき，低分子への特異度にも優れている [7]．結晶構造解析により，アプタマーが低分子を包み込むような形で結合していることが明らかになっており [42]，これは，アプタマーと高分子であるタンパク質との結合様式 [43] とは異なる．このように，アプタマーは標的分子の種類に合わせた結合様式をとり，さまざまな種類の標的分子に結合することができる．また，非常に毒性の強い毒物などは，抗体を生成する生物が抗体を十分に生成できないまま命を落としてしまうため，抗体を得ることが困難であるが，SELEX は試験管内でアプタマーを濃縮させる方法なので，そのような制限はない．以上のように，アプタマーは抗体より多種多様な標的分子を識別できるため，幅広い利用が可能と考えられる．

もう一つアプタマーが注目されている大きな理由が，人工的な介入の余地が大きく，標的分子との高い親和性による結合が実現できることである．抗体は，動物の免疫機構により生成されるため，人工的に作成したアミノ酸[*6]を抗体の部品として用いることは難しい．一方，アプタマーは核酸分子であるため，核酸の合成装置による化学合成で作成する．そのため，合成する際に天然型の塩基ではなく，修飾塩基と呼ばれる人工的に改良した塩基を核酸分子の部品として利用することができる．過去にさまざまな修飾塩基が開発されており，標的分子との強い親和性を持つものが報告されている [1, 44–46]．また，Ds と Px という修飾塩基による新たな人工塩基対も作成され，アプタマーに利用されている [47]．修飾塩基を導入することで，天然型の 5 塩基だけでは作り出せない特徴を作ることができ，結合様式のバリエーションを増やすことが可能となる．さらに，修飾塩基ではないが，Spiegelmer という核酸の鏡像異性体[*7]を用いたアプタマーも開発されている [48]．Spiegelmer は核酸ではあるが鏡像異性体のため，生体内にあるヌクレアーゼ[*8]などの影響を受けない．そのため，医薬品にとって重要である半減期の延長が可能である．Spiegelmer はドイツの NOXXON 社が創薬開発に利用している．以上のように，アプタマーにはさまざまな人工的な改良の余地があるため，さらなる特異性と親和性の向上の可能性がある．

抗体にはないアプタマーの特徴として，可逆的な構造変化を起させることができると

[*5] 喘息などに利用される医薬品.

[*6] タンパク質を構成する物質．ヒトのタンパク質は 20 種類のアミノ酸の組み合わせにより構成されている．

[*7] 立体配置が互いに鏡像の関係となっている立体異性体.

[*8] 核酸を分解する酵素.

いう点があげられる．この特徴を生かして，アプタマーが標的分子と結合したときにだけ構造を変化させて，標的分子を検出するセンサーの開発が可能である．抗体は構造が変化しないため，このような構造スイッチによるセンサーの開発ができない．この構造スイッチによるセンサーは，煩雑な手順を必要としないため，簡易検査の作成に適している．アプタマーを用いた構造スイッチによるセンサーの作成には，モレキュラービーコン法，DNAzyme のペルオキシダーゼ活性による検出法，酸化還元反応による電気化学的検出法などが利用されている．図 2.2 に，アプタマーを用いた構造スイッチによるセンサーの例を示す．モレキュラービーコン法では，アプタマーとブロック配列（アプタマーの一部の領域と相補鎖を形成する配列）を連結したセンサーの 5’ 末端と 3’ 末端に蛍光基 (Fluorophore) と消光基 (Quencher) を取り付ける．標的分子と結合する前は，アプタマーはブロック配列と塩基対を形成しているため，蛍光基は消光基により消光されている．標的分子とアプタマーが結合すると，取り付けた蛍光基と消光基との距離が開き，蛍光検出が可能となる [49,50]（図中の A）．DNAzyme を用いたセンサーは，共に核酸分子であるアプタマーと DNAzyme を連結させたものである．通常はブロック配列と塩基対を形成して非活性の状態にしてある DNAzyme が，アプタマーと標的分子が結合して構造変化を起こし，酵素活性を持つようになる（図中の B）．アプタマーと DNAzyme を用いた検出方法として，メラミン，パツリン[*9]，アデノシン三リン酸 (ATP)[*10]を検出するセンサーが報告されている [51–53]．酸化還元反応による電気化学的検出法では，電極上に酸化還元指示薬（メチレンブルーなど）を付加したアプタマーを固定する．標的分子とアプタマーが結合すると，アプタマーが構造を変化し，酸化還元指示薬と電極との位置が変わるため，電子の流れる量が変化し検出が可能となる（図中の C）．電気化学的検出法による例として，水中の水銀や鉛を検出するセンサーが報告されている [54]．

アプタマーの利用例

アプタマーの商用利用の例として，加齢黄斑変性症の治療薬 Macugen[*11]がある [10]．Macugen は VEGF の機能を阻害し，血管新生および血管透過性を抑制する．現在，承認されているアプタマー医薬品は Macugen だけであるが，Macugen 以外のアプタマー医薬品の臨床試験も進んでおり [11]，日本国内ではリボミック社が RNA アプタマーの医薬

[*9] カビ毒の一種．

[*10] アデノシンに 3 つのリン酸基がついた物質．アデノシンはアデニンという塩基とリボースという糖からなる．ATP は，生体内でエネルギーを貯蔵したり利用したりする際に媒体となる．

[*11] 一般名 Pegaptanib．2004 年に米国で，2008 年に日本で承認．

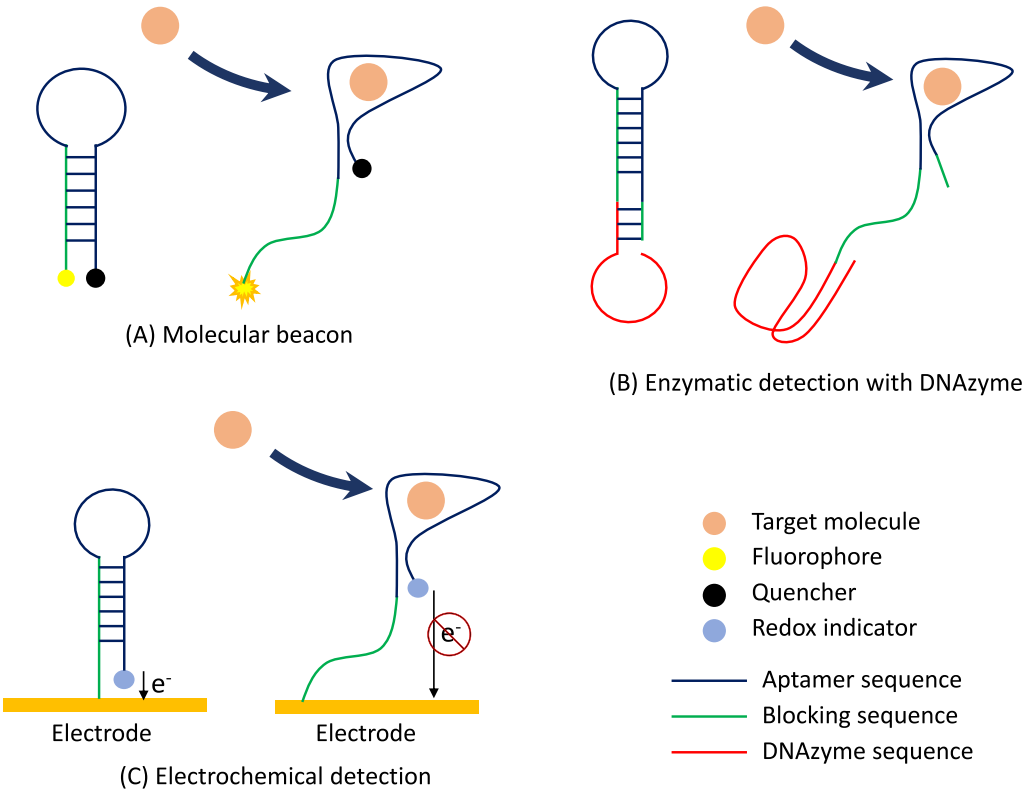

図 2.2　アプタマーを用いた構造スイッチによるセンサーの例

品を開発している [55].

他の商用利用の例として，SomaLogic 社の SomaScan があげられる．SomaLogic 社は，修飾塩基を導入したアプタマーである Slow off-rate modified aptamer (SOMAmer) を用いて，およそ 5,000 種類のタンパク質を一度に定量する SomaScan を作成した [56]. 5,000 種類の測定には，アプタマーが核酸分子であることを利用し，DNA マイクロアレイの技術が利用されている．SomaScan の利用により，疾患と関連するタンパク質バイオマーカーが発見されている（肺がん [57]，中皮腫 [58]，心血管疾患 [59]，腎機能 [6] など）. また，SomaScan はバイオマーカーの探索だけでなく，バイオマーカーを利用した疾患予測モデルの作成にも利用されている．バイオマーカーを組み合わせた心血管疾患の再発予測モデルは，従来の予測モデルである Framingham secondary risk score [60] より，心血管疾患の再発を高精度に予測できることが報告されている [61]．また，この心血管疾患

の再発予測モデルが，投薬に対するモニタリングに利用できる可能性があることも報告されている [62]．さらに，SomaScan を利用することにより，さまざまな健康状態を同時に予測可能であることも報告されている（心血管疾患のリスク，経口ブドウ糖負荷試験の結果，肝臓脂肪，腎機能，除脂肪体重，中性脂肪，体脂肪率，最大酸素摂取量など）[63]．

アプタマーを用いたバイオセンサーの例として，メラミンを測定するバイオセンサーがあげられる [51]．2008 年，中国でメラミンを意図的に牛乳やペットフードに混入させる事件が多発し大きな社会問題となった [64]．これは，メラミンに多く含まれている窒素を利用し，牛乳中のタンパク質の量を多く見せかけるために乳製品メーカーが行ったことである．牛乳からメラミンを検出するためには，実験室で遠心分離器による前処理を行ってから検査するのが一般的であるが，このメラミンを測定するアプタマーを用いたバイオセンサーは，前処理をすることなくメラミンを測定できることが報告されている．このバイオセンサーは，オンサイト（現場）で，迅速な，低コストで，簡便な測定を実現したものである．

以上，アプタマーは抗体と比較して有利な点があり，新たな分子認識能を持つ物質として期待されている．また，前述した SomaScan やメラミンのバイオセンサーは，アプタマーが核酸分子である特徴を利用したものであり，抗体で実現することは難しい．そのため，アプタマーは単に抗体の代替品というわけではなく，さらに広い実用化が可能な物質であるといえる．次節にて，アプタマーを実験的に選別する方法である SELEX に，NGS を適用して大量のデータを取得する方法である HT-SELEX の概要に関して述べる．

2.3　HT-SELEX の概要

HT-SELEX とは，実験的にアプタマーを選別する方法である SELEX に NGS を適用して大量の情報を取得する方法である．まず，NGS に関して説明する．2005 年に，454 Life Sciences 社よりそれまでのサンガー法とは異なるパイロシーケンシング法と呼ばれる新たなシーケンシングが発表された．パイロシーケンシング法により，それまでのサンガー法と比較して大量の配列情報を読み取ることが可能となった．このパイロシーケンシング法が最初の NGS である．その後も，シーケンシング技術は改良されており，現在は，Illumina 社の Sequencing by synthesis (SBS) 法を利用した MiSeq で 1,000 万本以上の配列を一度の稼働で読むことができる．

サンガー法は，DNA 合成に用いる dNTP[*12]に，蛍光色素で標識した ddNTP[*13]を加えて DNA の伸長を行う．dNTP が取り込まれると伸長が続くが，ddNTP が取り込まれると，DNA の伸長反応が止まる．この特徴を利用して，ゲルで満たされた毛細管（キャピラリー）を用いた電気泳動を行い，短いものから順番に蛍光色素を特定し塩基を決定する．パイロシーケンシング法は，はじめに，ビーズ上に固定した DNA をエマルジョン PCR[*14]により増幅させる．次に，増幅された DNA を鋳型として塩基を伸長し，塩基ごとに異なる蛍光反応を検出して塩基を決定する．SBS 法は，はじめに，ガラス基板上にブリッジの形で固定された DNA を PCR により増幅する．次に，ブリッジの片側を切断して一本鎖にし，増幅した DNA を鋳型として蛍光標識した dNTP を取り込ませる．この際に，取り込まれた塩基の蛍光の違いにより塩基を決定する．パイロシーケンシング法と SBS 法では，塩基配列の決定を超並列で行うことができるため，サンガー法と比較して大量の配列を読むことができる．本来ゲノム解析のために開発されたシーケンシングであるが，核酸分子を扱う SELEX にも適用が可能である．

NGS の装置は発売当初は高額であったが，その後小型化され，また価格も徐々に安価となり，一般の実験室でも NGS の装置を保有することができるようになった．そのため，近年では SELEX に NGS を適用する HT-SELEX が幅広く利用されている．以下に，HT-SELEX により大量の配列情報（HT-SELEX データ）を得る流れを示す．はじめに，10^{14} から 10^{16} 個の核酸分子を含むランダム核酸ライブラリを試験管内に準備する．核酸分子は，塩基，リン酸，糖（DNA であればデオキシリボース，RNA であればリボース）から構成されており，糖鎖の両端のうち，5 ’ にリン酸が結合している側を 5 ’ 末端，その逆を 3 ’ 末端と呼ぶ．ランダム領域（ランダムな塩基配列からなる領域）は PCR のため 5’ 末端と 3’ 末端にプライマー領域をもつ．通常，SELEX に用いられる核酸分子は，プライマー領域はそれぞれ 25 塩基ほど，ランダム領域は 30 から 50 塩基ほどで作成される．そのため，SELEX で用いられる核酸分子の長さは，ランダム領域とプライマー領域を含めて 80 塩基から 100 塩基ほどである．図 2.3 に SELEX に用いる核酸分子の一次構造を示す．図の核酸分子の左側が 5 ’ 末端であり，右側が 3 ’ 末端である．5 ’ 側と 3 ’ 側の破線部分は 5 ’ 側のプライマー領域と 3 ’ 側のプライマー領域である．プライマー領域に挟まれた実線の部分がランダム領域である．続いて，準備したランダム核酸ライブラリに，以下の工程を繰り返し行い，試験管内に標的分子と結合するアプタマーを濃縮させる．

[*12] デオキシヌクレオチド; dATP, dGTP, dCTP, dTTP.
[*13] ジデオキシヌクレオチド; ddATP, ddGTP, ddCTP, ddTTP.
[*14] 油中で PCR を行う方法.

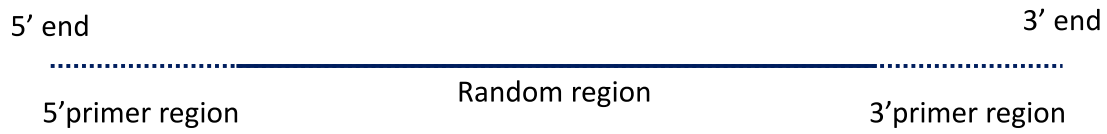

図 2.3　SELEX に用いられる核酸分子の一次構造

1. 標的分子と核酸分子を結合させる．
2. 標的分子と結合しない核酸分子は除去する．
3. 標的分子と核酸分子の複合体から核酸分子を溶出する．
4. 溶出された核酸分子を PCR により増幅する．

上記工程を十分な回数のラウンド繰り返すことにより（通常 8 から 12 ラウンドほど），試験管内には標的分子と結合するアプタマーで満たされる．最後に，試験管内に含まれる核酸分子の配列を NGS で読み取る．通常は，SELEX の最終ラウンドから HT-SELEX データを取得するが，すべてのラウンドの核酸分子を分注により保存しておき，NGS によりすべてラウンドの HT-SELEX データを取得することも実施されている．すべてのラウンドの HT-SELEX データの取得は，ラウンド間の比較ができるため，アプタマーの濃縮の過程などの理解にも有用である．

2.4　HT-SELEX データのクラスタリング

SELEX のラウンドを繰り返すと，理論上はアプタマーのみが濃縮され，HT-SELEX データには，アプタマーの配列のみが高い頻度で観測されるはずである．しかし実際には，高い頻度で観測される配列が標的分子と高い親和性を示さないことがしばしば起こる．これは，PCR により増幅されやすい核酸分子，標的分子と電荷により弱く結合している核酸分子[*15]，担体[*16]や試験管に付着した核酸分子なども濃縮されているためである．これらの濃縮を抑えるために，洗浄の強化による弱い結合をする核酸分子の除去や，Negative SELEX [65] による担体と結合する核酸分子の除去といった工夫がなされている．ただしそれでもなお，HT-SELEX データには標的分子と結合しない核酸分子の配列が高い頻度で観測されてしまう．また，SELEX で濃縮されるアプタマーは一種類ではなく，異なる結合様式のアプタマーが混在している．その中には，特異性や親和性の異なる

[*15] 核酸分子は負の電荷を帯びているため，正の電荷をもつものに対して非特異的に弱く結合する．
[*16] 標的分子を固定するための土台となる物質．

ものや，標的分子の異なるエピトープを認識するアプタマーも混在している可能性がある．これらの異なる結合様式のアプタマーは，利用目的により使い分けることができる．たとえば，異なるエピトープを認識する二つのアプタマーは，片方のアプタマーで標的分子を固定化し，もう片方のアプタマーで検出するようなサンドイッチ法に利用可能である．このように，標的分子と結合しない核酸分子の配列，異なる結合様式のアプタマーの配列が HT-SELEX データに混在しているため，それらをクラスタリングにより分けることが率よくアプタマーの候補配列を選択することにつながる．図 2.4 に，クラスタリングにより HT-SELEX データを分割し，アプタマーの候補配列を評価するまでの流れを示す．図中の左部が HT-SELEX データを示す．得られた HT-SELEX データを分割し，クラスタごとに順位付けを行う（図中の 1）．上位のクラスタの代表配列をアプタマーの候補配列とする．通常，上位のクラスタの中で頻度の高い配列を 1~3 本選択する（図中の 2）．アプタマーの候補配列を化学合成し，結合評価を行い，標的分子と高い親和性を示す配列をアプタマーとする（図中の 3）．赤で表示されたアプタマーと紫で表示されたアプタマーは異なる結合様式で標的分子に結合しており，緑は標的分子と結合しない核酸分子を表す．

従来の HT-SELEX データのクラスタリング手法として，FASTAptamer [17], AptaCluster [18], APTANI [19], AptaTRACE [20] があげられる．また，ゲノム解析のために開発されたクラスタリング手法 UCLUST [21] と UNOISE3 [22] も, HT-SELEX データのクラスタリングに利用可能である．表 2.2 に従来手法の名称，使用言語，クラスタリングの概要を示す．以下に，順を追って従来のクラスタリング手法について記載する．

FASTAptamer

FASTAptamer は全長配列の類似性をレーベンシュタイン距離 (Levenshtein distance: LD) により計算し，その距離が近い配列でクラスタを作成するクラスタリング手法である．LD とは，二つの文字列があったとき，一つの文字列からもう一方の文字列へ書き換えるために必要な最小の修正回数である．たとえば，文字列 1 が ATGC で文字列 2 が AAAA であったとき，文字列 1 が文字列 2 へと書き換えられるためには，AAGC, AAAC, AAAA と 3 回の修正が必要となる．このとき LD は 3 となる．HT-SELEX データの頻度の高い配列をクラスタの基本配列とし，基本配列と基本配列からの LD が閾値より短い配列で同一クラスタを作成する．頻度の高い配列から順番にクラスタの基本配列とするため，クラスタの順位が配列の頻度に依存する．基本配列に対してクラスタに含まれていない配列すべての LD を計算するため，計算に時間がかかる．そのため，計算時間を

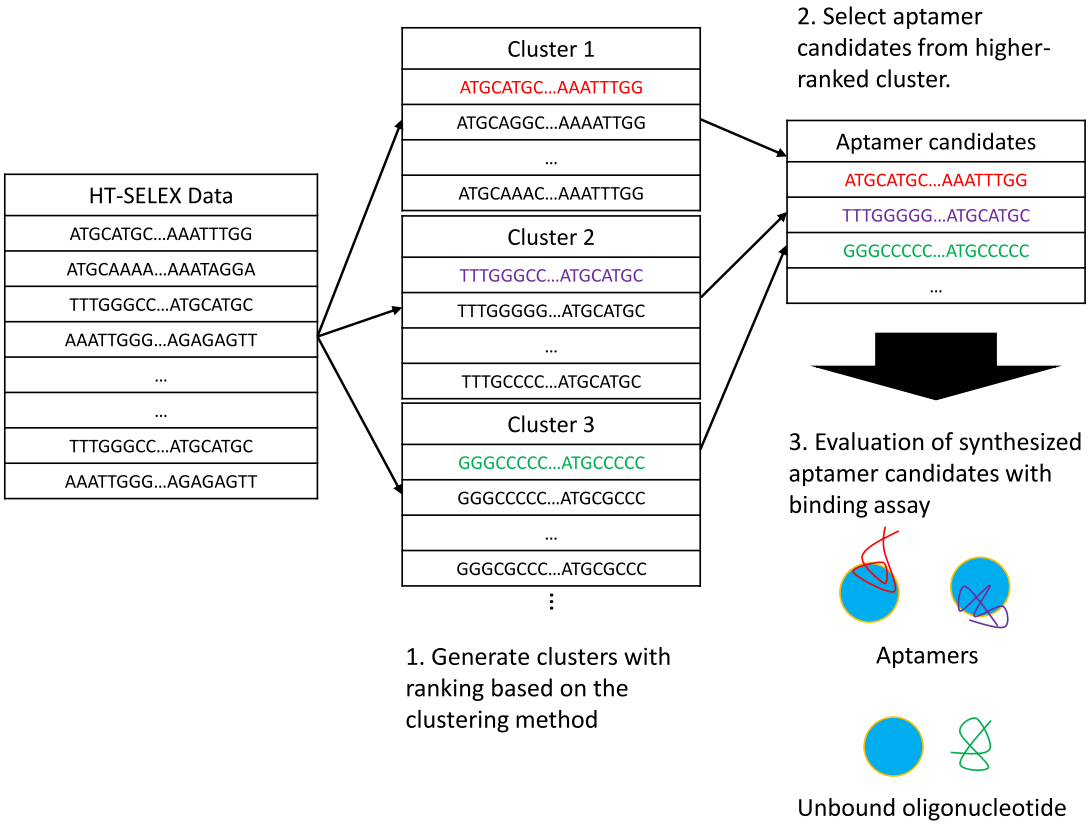

図 2.4　HT-SELEX データにクラスタリングを用いたアプタマー候補配列の選択と評価

短縮するためには，HT-SELEX データを頻度によりフィルタリングするなどの処理が求められる．また，HT-SELEX データに対して，適切な LD の閾値が明らかではないため，閾値を変更して解析結果を評価する試行錯誤の手間がかかる問題点もある．

AptaCluster

AptaCluster は二つの工程から構成されている．最初の工程が，局所性鋭敏型ハッシュ (Locality sensitive hashing: LSH) で，次の工程が k-mer counting[*17] [66] である．LSH では，配列の位置を指定して，その位置の塩基が共通するものを同じバケットに分類する．最後に，k-mer counting により類似するバケットを結合してクラスタを作成する．AptaCluster では，クラスタの順位付けにクラスタの総配列数とクラスタの配列の多様性

[*17] k-mer とは，長さ k の部分配列を意味し，k-mer counting とは長さ k の部分配列の個数である．

表 2.2　クラスタリング手法の比較

クラスタリング手法	使用言語	概要
FASTAptamer	Perl	全長配列の類似性をレーベンシュタイン距離 (Levenshtein distance: LD) で計算し，その距離が近い配列でクラスタを作成する.
AptaCluster	C++	最初に局所性鋭敏型ハッシュ (Local sensitive hashing: LSH) を用いて配列を大まかなグループに分類し，次に k-mer counting の近いグループでクラスタを作成する.
APTANI(mBed)	Python(C++)	任意の数の基準配列を探索し，基準配列からの距離が近い配列でクラスタを作成する.
UCLUST	C++	ペアワイズアラインメントの結果をもとに配列の類似性を計算し，類似性の高い配列でクラスタを作成する.
UNOISE3	C++	配列の頻度で補正した LD で配列間の類似性を計算し，その類似性が高い配列でクラスタを作成する.
AptaTRACE	C++, Java	最初に，ラウンド間で大きく頻度が増加する配列構造モチーフを探索し，次に，配列構造モチーフを含む配列でクラスタを作成する.

（非重複配列の数）の二つの指標を用いる．LSH では，配列の任意の位置で塩基が等しいかどうか調べるため，すべての配列の長さが等しくなくてはならない．SELEX の工程では，欠損や挿入により長さが変わる核酸分子も存在しており，AptaCluster ではそれら長さが変化した核酸分子を取り扱うことができない問題がある.

APTANI

APTANI は Aptamotif [67] を HT-SELEX データの解析ができるように改良したものである．APTANI は Vienna RNA package [68] を用いて配列の二次構造を予測し，配列構造モチーフ (Sequence-structure motif) を推定する方法である．配列構造モチーフとは，塩基の並びとその塩基が形成する二次構造の要素 (Secondary structural element)

を組み合わせた結合に関連する文字列である．APTANI は推定した配列構造モチーフを利用してクラスタを作成せず，代わりとして追加された機能である mBed [69] を用いてクラスタを作成する．mBed は Clustal Omega [70] に実装されているクラスタリング手法である．Clustal Omega は配列のマルチプルアライメント[*18]を行うソフトウェアである．大量配列のマルチプルアライメントは非常に計算時間を要するため，Clustal Omega はマルチプルアライメントを行う前に，mBed を用いて類似した配列でクラスタを作成する前処理を行う．mBed は二つの工程からなり，一つ目はクラスタ基準配列の探索であり，二つ目はクラスタ基準配列をもとにしたクラスタリングである．なお，クラスタ基準配列の探索には，usePivotObjects heuristic と usePivotGroups heuristic という二つのオプションがある．以下に，mBed によるクラスタ作成の手順を記す．$d(\cdot,\cdot)$ は二つの配列の LD を表す．

usePivotObjects heuristic の場合のクラスタ基準配列の選択.

1. 配列情報 X から任意の数のクラスタ基準配列 R を選択する．
2. クラスタ基準配列 R の各々の要素 s に対し 3 から 5 を繰り返す．
3. $d(l,s)$ を最大化する配列 l を X から選択する．
4. $d(m,l)$ を最大化する配列 m を X から選択する．
5. m を新たなクラスタ基準配列とする．

usePivotGroups heuristic の場合のクラスタ基準配列の選択.

1. 配列情報 X から任意の数のクラスタ基準配列 R を選択する．
2. クラスタ基準配列 R の各々の要素 s に対し 3 から 6 を繰り返す．
3. $d(l,s)$ を最大化する配列 l を X から選択する．
4. $d(m,s)+d(m,l)$ を最大化する配列 m を X から選択する．
5. $d(n,s)+d(n,l)+d(n,m)+\cdots$etc. を最大化する配列 n を X から選択する．
6. 同じ配列が選択された場合，もしくはクラスタの最大数に達した場合に，すべての選択された配列をクラスタの基準配列とする．

得られたクラスタ基準配列 $R_i, i=1,\cdots,t$ に対して，配列 $s\in X$ の embedded vector を定義する: $F(s)=[d(s,R_1),d(s,R_2),\cdots d(s,R_t)]$．二つの配列 $x,y\in X$ の embedded vector である $F(x),F(y)$ のユークリッド距離を計算し，距離が近いものを同一クラスタ

[*18] 三種類以上の配列を並べて比較する方法で，配列の類似した領域を特定するために用いられる．

とする．基準配列 R の探索のために，LD の計算を繰り返し行うため，計算に時間がかかる問題点がある．FASTAptamer と同様に，HT-SELEX データに対し頻度によるフィルタリングを行い，HT-SELEX データのサイズを小さくすることが求められる．

UCLUST

UCLUST は，ペアワイズアライメント[*19]により配列の類似性を求め，クラスタを作成する手法である．配列の類似性は，“identity” として次式で定義されている:

$$\text{identity} = \frac{\text{一致する塩基の数}}{\text{二つの配列の短いほうの配列の長さ}}. \tag{2.1}$$

閾値である identity threshold T よりも大きい配列同士で同じクラスタを作成する．二つの配列のすべての組み合わせに対して，ペアワイズアラインメントを行うため計算に時間がかかる．また，HT-SELEX に対するペアワイズアラインメントの適切なパラメータが分からないため，試行錯誤によるパラメータの探索も求められる．

UNOISE3

UNOISE3 は配列情報から PCR による読み取りエラーの配列を取り除くために作成された手法である．頻度の高い配列をクラスタの基準配列として，基準配列と類似した配列でクラスタを作成する．類似度の計算には，配列の頻度で補正した LD を用いる．頻度の高い配列 C をクラスタの基準配列とし，別の配列 M とする．このとき頻度の割合を “abundance skew” として定義する:

$$skew(M, C) = a_M / a_C. \tag{2.2}$$

a_M と a_C を，それぞれ配列 C と M の頻度とする．abundance skew が LD を考慮した閾値

$$\beta(d) = 1/2^{\alpha d+1} \tag{2.3}$$

より小さい場合，M はクラスタ基準配列 C と同じクラスタに分類される．d は C と M の LD を表し，α は LD に対する重みのパラメータである．FASTAptamer と同様に，基準配列に対する LD をすべて計算するため，クラスタリングに長い時間を要する．

[*19] 二つの配列を並べて比較し，類似する領域を特定する方法.

AptaTRACE

AptaTRACE は，複数のラウンドの HT-SELEX データを比較し，ラウンド間で大きく頻度が増加する配列構造モチーフを特定し，配列構造モチーフをもとにクラスタを作成する手法である．以下に AptaTRACE の手順を示す．

1. 配列情報を配列の頻度によりフィルタリングする．
2. SFOLD [71] を用いて二次構造を予測し，二次構造の要素であるヘアピンループ，バルジ，インターナルループ，マルチループ，ダングリングエンド/エクステリアーループ，ベースペア/ステムの位置情報を得る（二次構造の要素に関する詳細は後述する）．
3. 各々のラウンドで k-mer および k-context の分布を計算する．k-mer は長さ k の部分配列を示し，k-context は長さ k の二次構造の要素を示す．
4. k-mer および k-context のラウンド間の比較情報を得る．
5. 複数のラウンド間の比較により k-context の分布の変化を推定し, k-mer の context shifting score を求める．context shift score は k-mer の構造の推定に用いられる．
6. 頻度の低い配列を用いて k-mer の context shifting score の帰無仮説の経験分布を作成し，その分布と比較してスコアの高い k-mer を選択する．
7. 類似する k-mer の中で二次構造の要素が一致しているものを集約して配列モチーフの作成を行う．
8. 配列モチーフを含む配列でクラスタを作成する．

精度の高いクラスタリングの条件

SELEX より得られるアプタマーは，配列の全長ではなく，一部の領域が標的分子と結合している．この領域を，アプタマーと標的分子との結合領域と呼ぶ．結合領域の配列と立体構造が同じアプタマーは，たとえ他の領域が異なっていても同様の結合様式をとる．結合様式が同じアプタマーは同じクラスタに分けられるべきであるため，クラスタリングには結合領域を考慮するべきである．また，PCR のバイアスや，電荷の影響により濃縮している核酸分子は，仮に試験管内に濃縮していたとしても特定の結合領域を含まない．そのため，アプタマーの結合領域を推定し，それを利用したクラスタリングは，標的分子と結合しない核酸分子を除外するためにも有効である．

従来のクラスタリング手法の問題点

FASTAptamer, AptaCluster, APTANI, UCLUST, UNOISE3 は，全長配列の類似性を考慮したクラスタリング手法であり，結合領域を考慮していない．そのため，アプタマーの結合様式の異なる配列を同一クラスタに分けてしまう恐れがある．これは，異なる結合様式のアプタマーを取得するためには問題となる．また，配列の頻度に依存したクラスタの順位付けを行うため，標的分子と結合しない核酸分子の配列であっても，高い頻度であれば順位の高いクラスタに分けられてしまう．これは，アプタマーの候補配列の中に標的分子と結合しない核酸分子の配列を選んでしまう問題につながる．

AptaTRACE は従来手法のなかでも，結合領域を考慮したクラスタリング手法である．AptaTRACE は，結合領域を配列構造モチーフとして推定するが，推定する長さが 5 から 8 塩基という制限がある．そのため，長い結合領域によりアプタマーが標的分子と結合している場合には正しく結合領域を推定することができないという問題がある．また，複数のラウンドの HT-SELEX データを用いるため，シークエンシングの追加のコストが掛かること，データが増えることにより計算時間が余計に掛かることが問題となる．AptaTRACE は二次構造予測の結果を用いて配列構造モチーフを推定するが，二次構造予測に用意されている熱力学的パラメータには天然型の塩基のパラメータしかない [72, 73]．そのため，修飾塩基を用いたときは構造予測の精度が悪くなる可能性があり，正しい配列構造モチーフを推定することができなくなる恐れがある．このことは，クラスタリング精度の悪化につながる．

このように，過去に HT-SELEX データのクラスタリングに関して研究されてきてはいるが，それぞれの手法が問題を抱えている．精度の悪いクラスタリングの結果により，誤ったアプタマーの候補配列を選択してしまった場合，結合実験の失敗となり，追加の実験を行う必要がある．そのため，従来のクラスタリング手法の問題を解決するための新たなクラスタリング手法が求められている．

2.5　アプタマーの小型化

第 2.3 節で述べたように，SELEX により取得されたアプタマーは 5’ 末端と 3’ 末端のプライマー領域とランダム領域からなり，通常その長さは 80 塩基から 100 塩基である．アプタマーを実用化するためには，結合領域を保持しつつ，アプタマーの小型化（アプタマーの配列をできるだけ短くすること）を行う．図 2.5 に，アプタマーの小型化の概要を

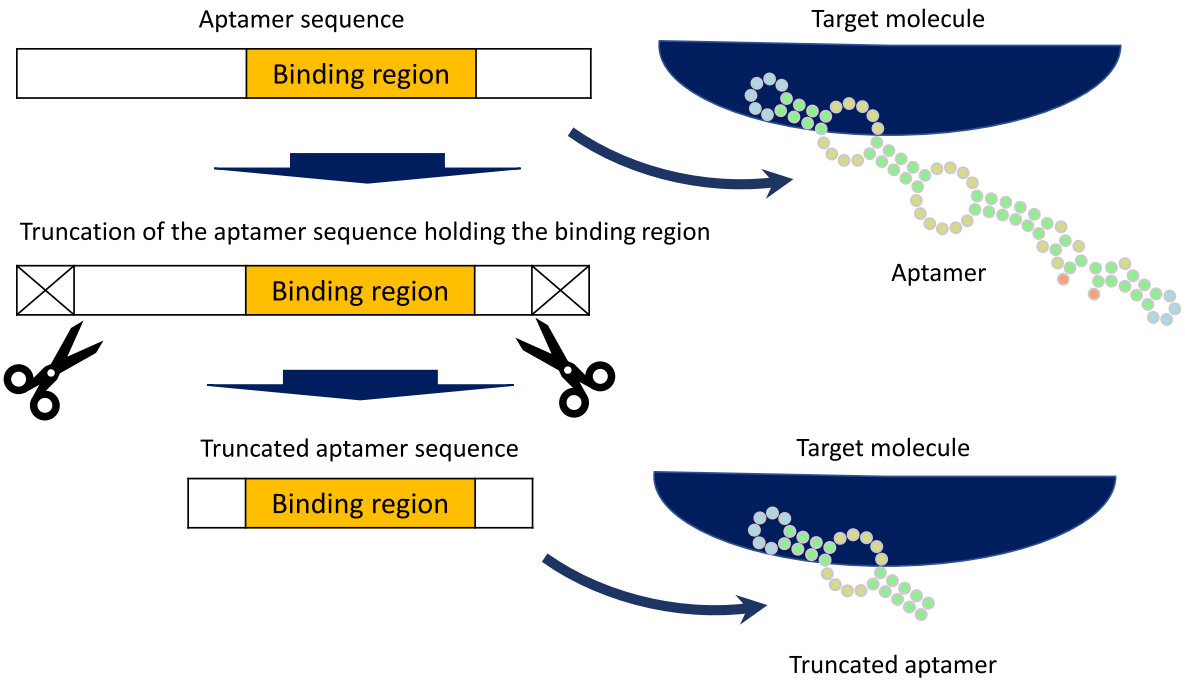

図 2.5　アプタマーの小型化

示す．図の上段は，アプタマーの全長配列と結合領域の位置関係と，アプタマーが立体構造を形成し，標的分子と結合している様子を示す．図の中段は，アプタマーの結合に不要な領域の削除を示す．図の下段は，小型化アプタマーと，小型化アプタマーが立体構造を形成して標的分子と結合している様子を示す．図に示されているように，小型化アプタマーの結合領域の立体構造は，もとのアプタマーの結合領域の立体構造を保持している必要がある．

小型化アプタマーの例

小型化アプタマーの例として，トロンビン DNA アプタマーがあるが，これは 15 塩基と非常に短い（トロンビンの RNA アプタマーは 25 塩基である）[74, 75]．さらに例をあげると，Macugen は 28 塩基の長さであり [10], interleukin-6 に結合する SOMAmer は 32 塩基の長さである [43]．Protein data bank (PDB) [76] には，アプタマーと標的分子との複合体の立体構造が登録されている．表 2.3 に，PDB に登録されているアプタマー

の長さを示す．各列は，PDB の ID，PubMed の ID[*20]，文献のタイトル，アプタマーの長さを示す．登録されているほとんどのアプタマーの長さは 40 塩基以下である．

表 2.3　PDB に登録されているアプタマーの塩基の長さ

PDB ID	PubMed ID	Title	Length
1HUT	8102368	The structure of alpha-thrombin inhibited by A 15-mer single-stranded DNA aptamer	15
6EO7	29684204	X-ray structure of the complex between human alpha-thrombin and modified 15-mer DNA aptamer containing 5-(3-(acetamide-N-yl)-1-propen-1-yl)-2'-deoxyuridine residue	15
5EF6	28473536	Structure of HOXB13 complex with methylated DNA	18
2FY1	17318228	A dual mode of RNA recognition by the RBMY protein	21
5UC6	28993621	Structural insights into IL-1 alpha recognition by a naphthyl-modified aptamer that mimics IL-1RI Domain III	23
3AGV	20675355	Crystal structure of a human IgG-aptamer complex	24
4HQU	23139410	Crystal structure of human PDGF-BB in complex with a modified nucleotide aptamer (SOMAmer SL5)	24
5DO4	27566147	Thrombin-RNA aptamer complex	25
6GN7	30357392	X-ray structure of the complex between human alpha thrombin and NU172, a duplex/quadruplex 26-mer DNA aptamer, in the presence of sodium ions.	26
3DD2	18971322	Crystal structure of an RNA aptamer bound to human thrombin	26
4I7Y	24311581	Crystal structure of human alpha thrombin in complex with a 27-mer aptamer bound to exosite II	27
4ZBN	26027732	Non-helical DNA triplex forms a unique aptamer scaffold for high affinity recognition of nerve growth factor	28

[*20] 文献検索サイト PubMed が，各文献に割り当てた ID 番号．PubMed は，医学，生物学などの文献の検索エンジンである．

5CMX	26673709	X-ray structure of the complex between human alpha thrombin and a duplex/quadruplex 31-mer DNA aptamer	31
4NI7	24415767	Crystal structure of human interleukin 6 in complex with a modified nucleotide aptamer (SOMAmer SL1025)	32
5HRT	27043297	Crystal structure of mouse autotaxin in complex with a DNA aptamer	34
5VOE	29863725	DesGla-XaS195A bound to aptamer 11F7t	36
6U82	32020675	Crystal structure of the double homeodomain of DUX4 in complex with a DNA aptamer containing bulge and loop	38
4R8I	25901662	High resolution structure of a mirror-Image RNA oligonucleotide aptamer in complex with the chemokine CCL2	40
3HXO	19913482	Crystal structure of von Willebrand factor (VWF) A1 domain in complex with DNA aptamer ARC1172, an inhibitor of VWF-platelet binding	42
3EGZ	18940672	Crystal structure of an in vitro evolved tetracycline aptamer and artificial riboswitch	65
4YB1	25818298	20A mutant c-di-GMP Vc2 riboswitch bound with 3',3'-cGAMP	91
3MUT	20690679	Crystal structure of the G20A/C92U mutant c-di-GMP riboswith bound to c-di-GMP	92
3IRW	19898477	Structure of a c-di-GMP riboswitch from V. cholerae	92

アプタマーの小型化の目的

アプタマーの小型化が必要な理由として，小型化アプタマーがさらに強い結合を示す可能性があること [36, 37]，アプタマーの構造多様性を抑えること，核酸の化学合成のコストと合成エラーを減らすことなどがあげられる．アプタマーは立体構造を形成して標的分子と結合するために，正しい立体構造を形成しないと標的分子と結合することができない．長いアプタマーでは多様な構造をとり得るため，たとえ配列が同じだとしても，その中のある割合が目的の構造をとらず標的分子と結合しない．これは，医薬品であれば薬効の低下，診断薬やバイオセンサーであれば精度の低下につながる可能性がある．そのため，アプタマーの小型化を行い，取り得る構造の多様性を抑制して，目的の構造を形成する割合を上げることが重要である．図 2.6 にアプタマーの塩基配列の長さと構造多様性と

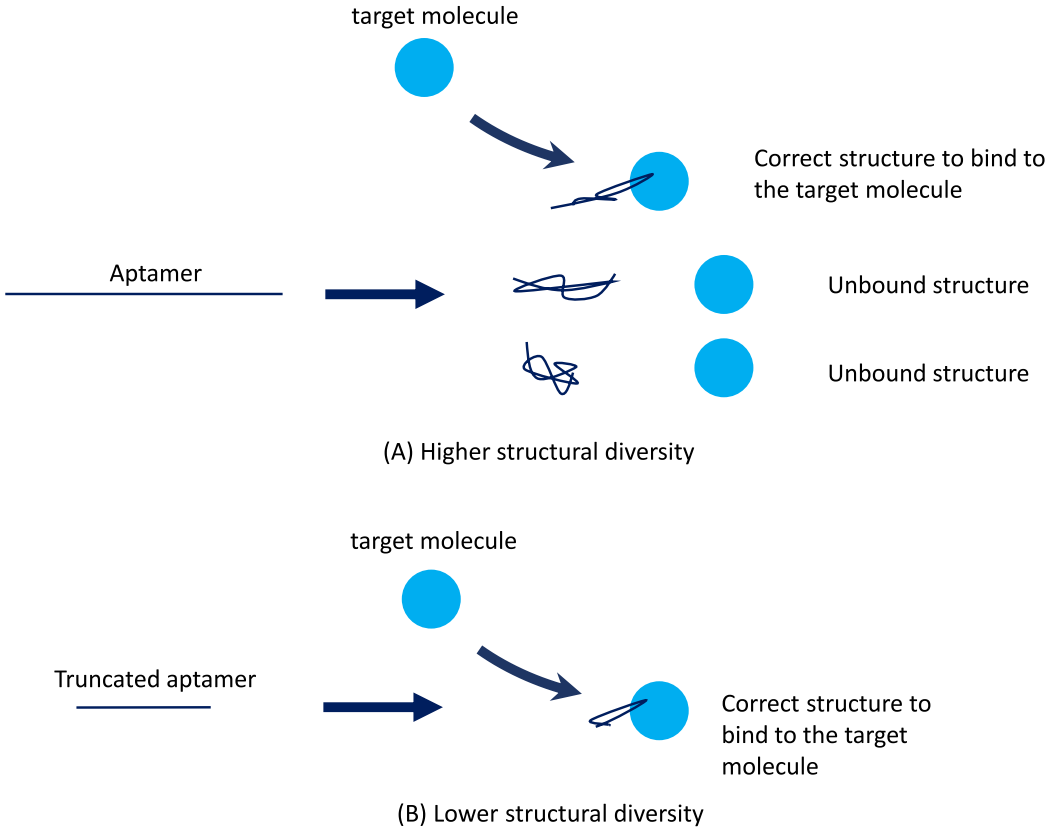

図 2.6　アプタマーの塩基配列の長さと構造多様性との関係

の関係を示す．長いアプタマーの構造多様性は高く，結合する構造をとる割合が低い（図中の A）．短いアプタマーの構造多様性は低く，結合する構造をとる割合が高い（図中の B）．他にもアプタマーを小型化するメリットがあり，たとえば，第 2.2 節で述べたように，小型化アプタマーは他の酵素活性を持つ核酸分子である DNAzyme [77] と連結させ，標的分子と結合したときにだけ DNAzyme 活性を持つセンサーにすることもできる [51]．また，小型化アプタマーはタンパク質との分子量の比率が大きくなるため，蛍光偏光法への利用にも適している [78]．蛍光偏光法とは，蛍光標識したアプタマーが溶液中で回転運動する速度が，タンパク質のような大きな標的分子と結合すると遅くなることを利用した検出方法である．

アプタマーの構造予測の種類

アプタマーは立体構造を形成して標的分子と結合するので，配列から構造を特定し，それを利用して小型化配列の推定を行うことが重要である．核酸の構造予測としては二次構造予測と三次構造予測があり，二次構造とは三次構造を二次元に射影したものである．核酸分子の三次構造予測には，フラグメントアセンブリという二次構造をフラグメント化したものを，三次構造のレファレンスと比較して組み合わせる方法が報告されている．たとえば，RNAComposer [79] は，最初に二次構造を予測し，二次構造のフラグメントと RNA FRAbase [80] に登録されている三次構造のフラグメントを比較して，全体の三次構造を推定する．MC-Sym [81] も RNAComposer と同様に，二次構造のフラグメントと準備されたフラグメントの三次構造を組み合わせて，全体の三次構造を予測する．ただ，通常アプタマーの小型化には三次構造ではなく二次構造を予測したものを利用する．これは，高次構造の予測になるほど予測精度が低くなること，また，アプタマーの小型化には二次構造の予測でも十分であることが理由である．

アプタマーの二次構造予測とその表現

核酸分子がとり得る二次構造には，ステムループ構造 [82]，シュードノット構造 [83]，G-quadruplex 構造 [74,77] がある．これらの構造を形成するために，ワトソンクリック型[*21]の塩基対だけでなく，グアニンとウラシルが形成するゆらぎ型塩基対 [84] や，三重鎖構造 [85] など特殊な塩基対も存在する．図 2.7 に，ステムループ構造（図中の A），シュードノット構造（図中の B），G-quadruplex 構造（図中の C）を示す．ステムループ構造とシュードノット構造は基本的に塩基対からなる構造であるが，G-quadruplex 構造は特殊な二次構造であり，4 つのグアニンが一つの平面を作り，その平面が重なるような構造である．トロンビンアプタマーは G-quadruplex 構造を形成してトロンビンと結合する [86].

核酸の二次構造予測には，熱力学的エネルギーの安定性を考慮して計算される手法 [68,87,88] と，機械学習による手法がある [89,90]．二次構造と三次構造を含めて，ほとんどの核酸分子の構造予測は RNA の予測に限定されている．RNA の構造予測が DNA の構造予測より先行している理由は，non-coding RNA [91] のようなタンパク質に翻訳されないが生体内で機能を持つ RNA（transfer RNA も non-coding RNA の一つである）

[*21] グアニンとシトシン，アデニンとチミンの塩基対.

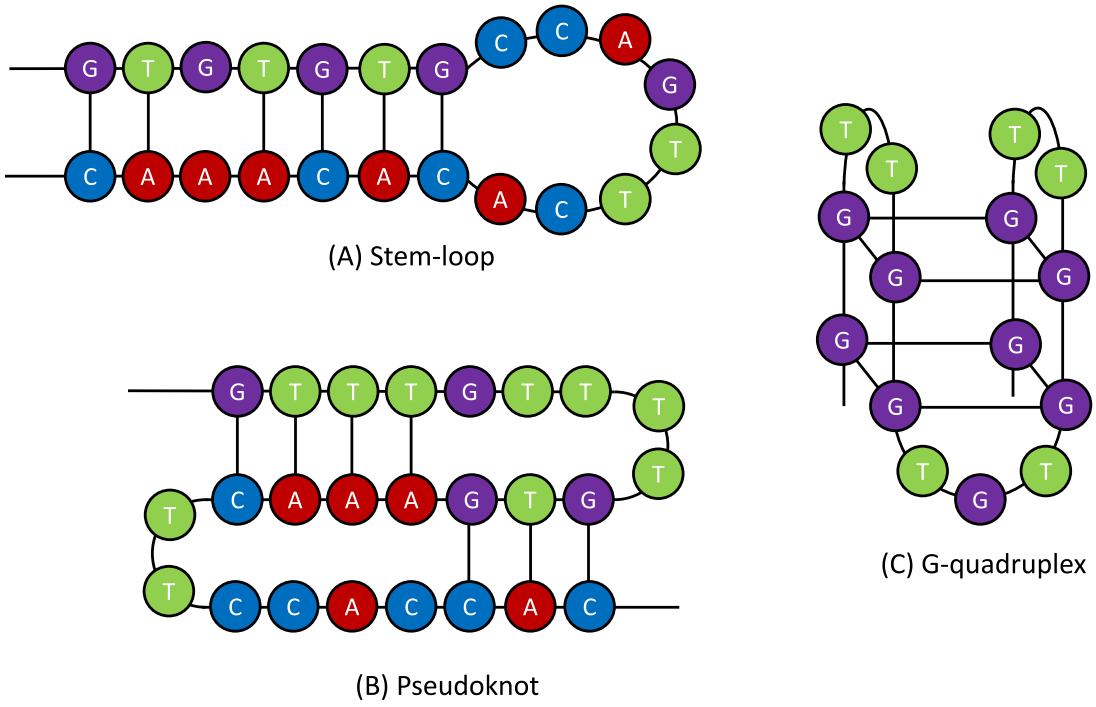

図 2.7　アプタマーの二次構造の例

の構造予測の研究が先行しているためである．ただ，熱力学的エネルギーの安定性を計算する手法では，DNA と RNA の熱力学的パラメータが用意されており，DNA と RNA の二次構造を推定することができる [72, 73]．予測された二次構造は Dot-bracket という Fasta フォーマットを拡張した書式で示される．Fasta フォーマットは，配列情報を “>” で記載し，その下の行に配列を記載する．仮に，配列情報として配列名が “Sample sequence”，配列情報が GGGGGAAAAACCCCC であれば以下のように記載される．

```
> Sample sequence
GGGGGAAAAACCCCC
```

Dot-bracket フォーマットでは，塩基が塩基対を形成する場合は，“()” で表現する．塩基対を組まない場合は “.” で表現される．たとえば, 塩基配列を GGGGGAAAAACCCCC としたとき，G と C が塩基対を組み “(((((.....)))))” と表現される．また，シュードノット構造の場合は “[]” でしめされ，G-quadruplex 構造は “+++” で示される．以下に，RNAfold [68] で計算された予測結果を示す．

```
> Sample sequence, RNA, 37 degree
GGGGGAAAAACCCCC
(((((.....))))) ( -8.60)
```

カッコ内の数値は，自由エネルギーの値を表し，低い値であるほど安定した構造である（単位は kcal/mol である）．上記は RNA の配列で温度が 37C° の場合の自由エネルギーであるが，以下に DNA の場合で温度が 25C° の場合を記す．

```
> Sample sequence, DNA, 25 degree
GGGGGAAAAACCCCC
(((((.....))))) ( -6.39)
```

このように同じ二次構造であっても，RNA と DNA の違いや温度の違いによって核酸分子は構造安定性が異なる．続いて，シュードノット構造と，G-quadruplex 構造が予測された場合の Dot-bracket 形式による表現の例を以下に示す．シュードノット構造の予測には，ipknot [92] を用いた．

```
> Sample Pseudoknot
AAUUAAUUAAUUAAUUAAUUAAUUAAUUAAUUAAUUAA
.(([[[[(((([[[[((....]]))))]]]]))))]].
```

G-quadruplex 構造の例を以下に示す．G-quadruplex 構造の予測には，RNAfold を用いた．以下の配列は酵素活性を持つ DNAzyme であり [77]，他のアプタマーと連結したセンサーに応用されている [51].

```
> G-quadruplex
GGGUGGGAGGGUCGGG
+++.+++.+++..+++ (-35.85)
```

アプタマーの二次構造の可視化

次に，Dot-bracket フォーマットを可視化する方法に関して記載する．以下に RNAfold により予測されたスフィンゴシルホスホリルコリン (Sphingosylphosphorylcholine : SPC)) [93] に結合する RNA アプタマーの Dot-bracket フォーマットの結果を示す．

```
> SPC aptamer
```

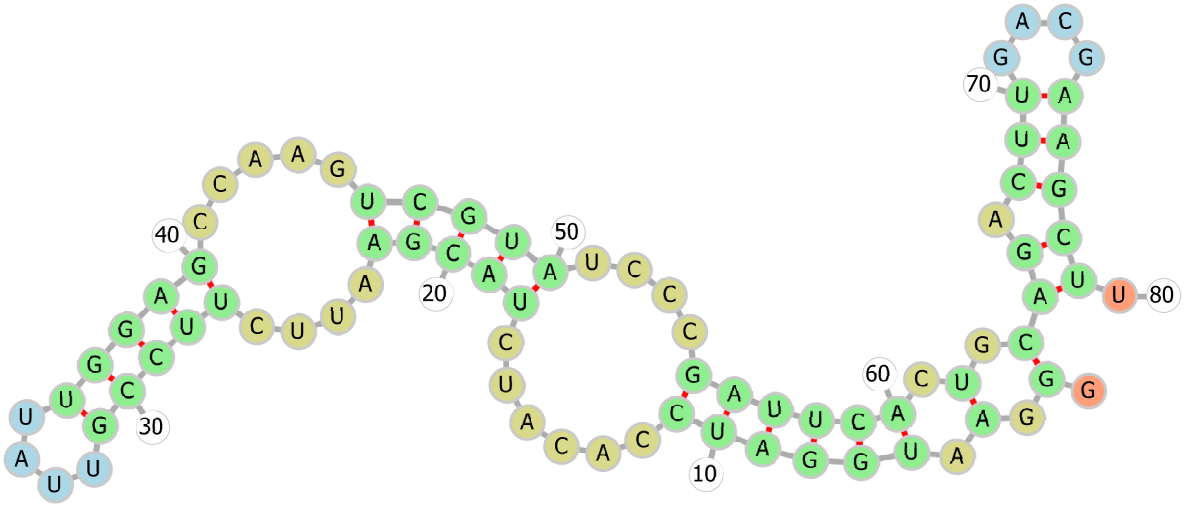

図 2.8　二次構造予測の例

```
GGGAAUGGAUCCACAUCUACGAAUUCUUCCGUUAUUGGAGCCAAGUCGUAUCCCGAUUCACUGCAGACUUGACGAAGCUU
.((((.((((.(........).)))))))).....(((((((((((((.....)))).))))))).....)))) (-19.42)
```

この Dot-bracket の書式を可視化すると 図 2.8 で表現される．二次構造に付随する番号は，5’ 末端からの塩基の数を表している．そのため，この SPC アプタマーは 80 塩基であることがわかる．二次構造の可視化には Forna [94] を用いた．

二次構造には特徴的な部分構造があり，それらを二次構造の要素と呼ぶ．二次構造の要素の種類には，ヘアピンループ (H)，バルジ (B)，インターナルループ (I)，マルチループ (M)，ダングリングエンド/エクステリアーループ (E)，ベースペア/ステム (S) がある．塩基の並び方だけでなく，どの二次構造の要素を形成しているのかもアプタマーと標的分子との結合に影響を与える．図 2.9 に，二次構造の要素を図示する．

アプタマーの構造多様性

アプタマーは一つの配列から最安定構造，準安定構造などの異なる構造を形成する．異なる構造は RNAsubopt [68] で求めることができる．以下に，RNAsubopt で推定された SPC の最安定構造と準安定構造を示す．塩基配列の右にある数値は最安定構造における推定された自由エネルギーを示し，二次構造の右に記載されている数値は各二次構造に対する自由エネルギーを示す．この中で，最安定構造は自由エネルギーが -19.42 である構造であり，それ以外の構造は準安定構造である．

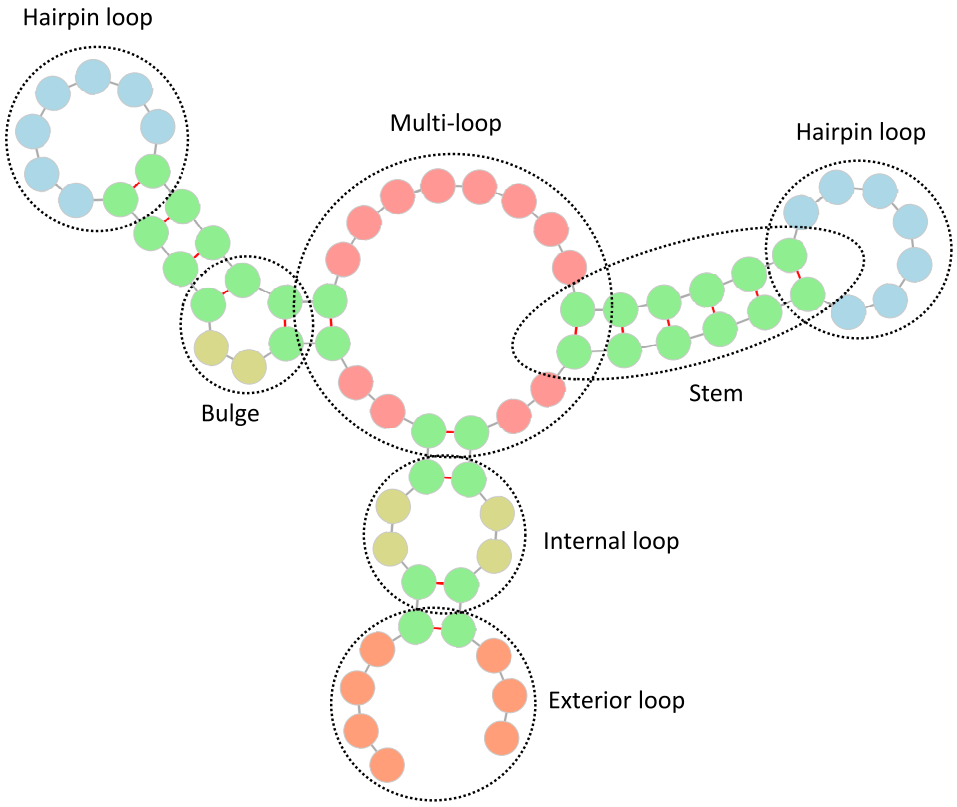

図 2.9　二次構造の要素

```
> SPC aptamer
GGGAAUGGAUCCACAUCUACGAAUUCUUCCGUUAUUGGAGCCAAGUCGUAUCCCGAUUCACUGCAGACUUGACGAAGCUU -19.42
.((((.((((.(........).)))))))).......((((((((((.................))))))).....)))) -18.95
.((((.((((.(........).)))))))).......((((((((((((.....)))).))))))).....)))) -19.42
.(.(.((((((.....((((....(((.....)))).....))))))....))))))).)((.(((....))))). -18.51
.(.(.((((((.....((((....((((....))))).....))))))....))))))).)((.(((....))))). -18.91
(((((((.....))).((((((....((((....))))).....)))))))).........((.(((....))))). -18.79
```

以下に小型化された SPC アプタマーの最安定構造と準安定構造を示す．小型化アプタマーにはステム構造を強化するため，5’ 末端に GG と 3’ 末端に CC が加えられている．

```
> SPC truncated aptamer
GGCGAAUUCUUCCGUUAUUGGAGCCAAGUCGCC -13.39
```

```
(((((.....((((....)))).....))))) -12.50
(((((....(((.....)))).....))))) -12.99
(((((....((((....)))).....))))) -13.39
(((((.((.(((((....)))))..)).))))) -12.42
```

ここで重要なことは，小型化前の SPC アプタマーの最安定構造と小型化後の構造が異なり，小型化前の SPC アプタマーの準安定構造が小型化後の構造と一致することである．つまり，小型化前の SPC アプタマーの最安定構造が，標的分子である SPC と結合せず，準安定構造が SPC と結合していると考えられる．以下に，小型化前の最安定構造と小型化後の最安定構造の比較を示す．

```
                    GGCGAAUUCUUCCGUUAUUGGAGCCAAGUCGCC
GGGAAUGGAUCCACAUCUACGAAUUCUUCCGUUAUUGGAGCCAAGUCGUAUCCCGAUUCACUGCAGACUUGACGAAGCUU
.((((.((((.(........).)))))))).......(((((((((((((.....)))).))))))).....))))
                    (((((....(((((....)))))).....)))))
```

予測された構造の中で最も構造が安定なものが標的分子と結合しているわけではなく，実際どの構造でアプタマーが標的分子と結合しているのかは不明である．どのような構造で標的分子と結合しているのか調べるためには，X 線結晶構造解析のような実験を新たに行う必要があるが，時間とコストを要するため通常は行われない．標的分子と結合しているアプタマーの真の二次構造がわからないため，二次構造予測から小型化アプタマーを推定する際には，さまざまな二次構造を考慮する必要がある．

現状のアプタマーの小型化の課題

アプタマーの小型化配列の推定には二次構造予測が必要であることを述べたが，従来は経験を有する専門家が，二次構造予測の結果を見ながら小型化配列を推定してきた．これでは明確な基準が存在せず，結果の再現性がないため，予測の方法が正しいのかどうか検証できない．また，通常はアプタマーの標的分子との結合領域が分からないまま小型化配列が推定されているため，結合領域を削除した小型化配列を推定してしまう可能性がある．さらに，アプタマーには構造多様性があるため，複数の二次構造を考慮して小型化配列を推定する必要があるが，人では複数の構造を網羅的に考慮することが難しい．以上述べてきたように，アプタマーの小型化に関するバイオインフォマティクスの研究はあまりなされておらず，小型化配列の予測は専門家の手腕に依存している．そのため，バイオインフォマティクスによる再現性のあるアプタマーの小型化配列を推定する手法が求められ

ている．このときアプタマーの小型化配列を推定する手法には，アプタマーの構造多様性，結合領域の二次構造が考慮される必要がある．

2.6　むすび

本章では，HT-SELEX データを用いたアプタマーの設計に関する基礎的考察に関して記した．最初に，第 2.2 節で，アプタマーの特性を述べた．その中で，抗体と比較したアプタマーの優位性を述べ，アプタマーの実用例に関しても述べた．次に，第 2.3 節で，HT-SELEX に関する概要を述べた．続いて，第 2.4 節で，HT-SELEX データのクラスタリングに関して述べた．その中で，従来のクラスタリング手法の特徴に関して詳細を述べた．最後に，第 2.5 節で，アプタマーの二次構造と小型化に関して述べた．

第 3 章

HT-SELEX データのクラスタリング

3.1 まえがき

本章では，HT-SELEX データを高速かつ高精度に分けるクラスタリング手法 FSBC [24, 27–29, 32] に関して述べる．FSBC は主に 2 つの工程から構成される．アプタマーの標的分子との結合領域を高速に推定する工程と，推定した結合領域を用いてクラスタを作成する工程である．FSBC の概要を図 3.1 に示す．図の上段は，HT-SELEX データからアプタマーの結合領域を推定する工程を示しており，図の下段は，推定した結合領域を含む配列でクラスタを作成する工程を表す．

SELEX において利用される核酸分子の長さは，プライマー領域を含めて 80 塩基から 100 塩基ほどの長さである．アプタマーは全長の配列で標的分子と結合しているわけではなく，一部の領域が結合している．HT-SELEX データの中でアプタマーの標的分子との結合領域は，特定の文字列として高い頻度で観測される．これを，本論文では Over-represented string (ORS) と呼ぶ．そのため，HT-SELEX データの中で ORS を探すことが，アプタマーの標的分子との結合領域を推定することになる．共通する ORS を含むアプタマーは，同様の結合様式を示すことが期待されるため，共通する ORS を含む配列でクラスタを作成することが，精度の高いクラスタリングとなる．FSBC は，ORS を高速に探索し，ORS をもとに配列を分けるクラスタリング手法である．

本章では，まず第 3.2 節で，FSBC の概要を説明する．第 3.3 節で，特定の文字列を含む配列の出現確率に関して述べる．第 3.4 節で，特定の文字列を含む配列の出現確率を用い，新たな文字列のスコアとして Z スコアを定義する．第 3.5 節で，ORS を高速に探索する方法を述べる．第 3.6 節では，第 3.5 節で選択された ORS を用いてクラスタを作成する方法を述べる．第 3.7 節では，並列処理による FSBC のさらなる高速化に関し

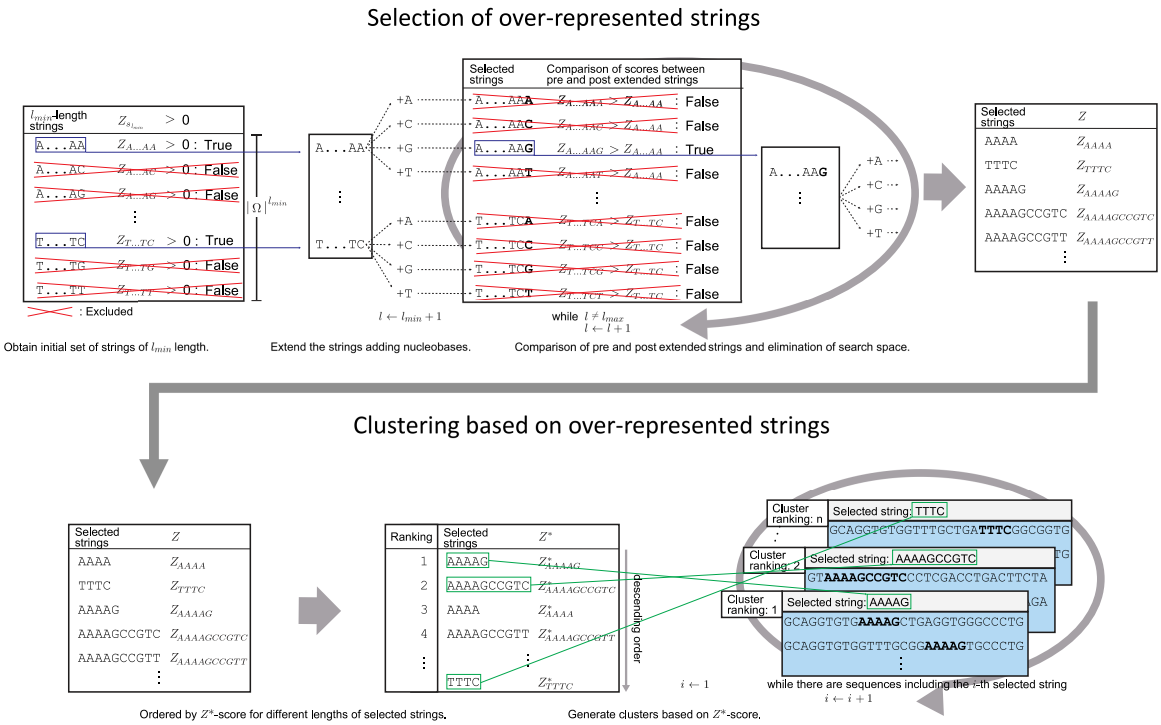

図 3.1　FSBC の概要

て述べる．第 3.8 節では，配列の多様性を定義し，FSBC で得られたクラスタがどのように分布しているのか評価する方法を述べる．第 3.9 節では，hESC を標的分子とした HT-SELEX データにより, 提案手法の有用性を検証した結果を述べる．第 3.10 節では，第 3.9 節と異なる標的分子である，IL10RA を標的分子とした HT-SELEX データを用いて，FSBC の有用性を検証した結果を述べる．最後に第 3.11 節で本章のむすびを述べる．

3.2　FSBC の概要

FSBC では，文字列のスコアを Z スコアと定義し，この Z スコアが任意の値より高い文字列を ORS とする．SELEX の過程において，試験管内の塩基のバランスが均一である保障がないため，Z スコアの定義には HT-SELEX データから観測される塩基の比率を

考慮する．塩基のバランスが均一であるというのは，観測されるアデニン (A)，シトシン (C)，グアニン (G)，チミン (T)/ウラシル (U) が各々およそ 25% 観測されることである．たとえば，グアニンが多く含まれる HT-SELEX データにおいて AAAAA と GGGGG が同じ数だけ観測された場合，提案する Z スコアの定義では，AAAAA の方が高い Z スコアとなる．塩基のバランスを考慮している理由は，FSBC ではシーケンシングによるコスト削減のため最終ラウンドの HT-SELEX データのみを利用することを想定しており，最終ラウンドにおける塩基の不均一性に対応するためである．

探索する ORS が長くなると，指数関数的に文字列の組み合わせが増大する．そのため，すべての文字列を列挙して，その中から ORS を選ぶことは非常に長い計算時間を要してしまう．この問題を解決するため，FSBC では短い文字列と長い文字列の Z スコアを比較しながら，探索する必要のない文字列を探索領域から削除する．この探索領域の削減により，長い ORS を探索することが可能となる．この探索領域削減の方法は，短い文字列を親ノードとし長い文字列を子ノードとすると，ちょうど分枝限定法における分枝の刈り込みと同義となる．

分枝限定法により，長い ORS ほど Z スコアが高くなるため，異なる長さの ORS を Z スコアにより直接比較することができない．そこで，ORS の長さごとに Z スコアを規格化し（Z^* スコア），異なる長さの ORS の順位付けを行う．たとえば，探索された ORS に ATGCA と ATGCATGCA があったとき，$Z(\mathrm{ATGCA}) < Z(\mathrm{ATGCATGCA})$ であっても，$Z^*(\mathrm{ATGCA}) > Z^*(\mathrm{ATGCATGCA})$ であれば，ATGCA の方が結合領域として尤もらしいことになる．順位付けされた上位の ORS から順番に，その ORS を含む配列で同一クラスタを作成する．

以上が FSBC の概要である．FSBC は，ORS を探索することによって結合領域を推定し，その推定された結合領域でクラスタを作成する．そのため，PCR のバイアスで濃縮されてしまう配列のような，頻度は高いが結合領域を持たない配列の影響に対して頑健であり，精度の高いクラスタリングが期待できる．また，ORS を探索する際に探索領域の削減を行うため，長い ORS の探索も高速に行うことができる．

3.3　特定の文字列を含む配列の出現確率

本節では，特定の文字列を含む配列の出現確率に関して述べる．塩基の集合を $\Omega = \{\mathrm{A}, \mathrm{C}, \mathrm{G}, \mathrm{T}\ (\mathrm{U})\}$ とする．アルファベットの A，C，G，T (U) はそれぞれアデニン，シトシン，グアニン，チミン（ウラシル）を示す．DNA アプタマーの場合はチミンとなり，

RNA アプタマーの場合はウラシルである．各々の塩基の出現確率を $p_j, j \in \Omega$ とする．塩基の集合の要素からなる文字列を s としてその長さを $|s|$ とする．また，s に含まれる各々の文字の数を $n(s,j), j \in \Omega$ とする．このとき，s を含む長さ L の配列が出現する確率 $P(s,L)$ は以下の漸化式となる:

$$\begin{aligned} P(s,L) &= P(s,L-1) + Q(s) - Q(s)P(s,L-|s|) \\ &\quad - Q(s)\sum_{t\in\mathcal{T}} \frac{1}{q(t)} \left\{P(s,L-|s|+|t|) - P(s,L-|s|+|t|-1)\right\}, \\ &\qquad Q(s) = \prod_{j\in\Omega} p_j^{n(s,j)}, \quad q(t) = \prod_{j\in\Omega} p_j^{n(t,j)}, \quad L \geq |s|. \end{aligned} \tag{3.1}$$

ただし，$\mathcal{T}$ は s の自己重複領域を表し，$n(t,j), t \in \mathcal{T}, j \in \Omega$ は自己重複領域に含まれる塩基の数である．たとえば，文字列 s が ATATA の場合，自己重複領域は $\mathcal{T} = \{\text{A, ATA}\}$ となる．また，自己重複領域以外の部分は，自己重複領域 A に対して，TATA となり，ATA に対して TA となる．$L < |s|$ の場合は，$P(s,L) = 0$ となる．これは，文字列 s が長さ L の配列より長いため，確率が 0 となるのは自明である．式 3.1 は，$Q(s)$ および $q(t)$ の計算において，塩基の出現確率を適用している．そのため，塩基のバランスが均一でないプールにおいても，塩基の出現確率を考慮した文字列を含む配列の確率が計算できる．塩基の出現確率は，互いに独立に同一分布に従うとする (independent and identically distributed: i.i.d.)．これは，生体内における DNA 配列は，3 つの塩基の並びがコドンとなりアミノ酸への翻訳のための情報となっているが，アプタマーに関してはこのような隣り合う塩基同士の法則性がないためである．式 3.1 を $Q(s)$ でまとめると次式となる:

$$\begin{aligned} &P(s,L) = P(s,L-1) \\ &+ Q(s)\left[1 - P(s,L-|s|) - \sum_{t\in\mathcal{T}} \frac{1}{q(t)} \left\{P(s,L-|s|+|t|) - P(s,L-|s|+|t|-1)\right\}\right], \\ &\qquad Q(s) = \prod_{j\in\Omega} p_j^{n(s,j)}, \quad q(t) = \prod_{j\in\Omega} p_j^{n(t,j)}, \quad L \geq |s|. \end{aligned} \tag{3.2}$$

また，文字列 s が自己重複領域を持たない場合は，式 3.1 は簡略化でき式 3.3 となる:

$$\begin{aligned} P(s,L) &= P(s,L-1) + Q(s)\{1 - P(s,L-|s|)\}, \\ &\qquad Q(s) = \prod_{j\in\Omega} p_j^{n(s,j)}, \quad L \geq |s|. \end{aligned} \tag{3.3}$$

次に，式 3.1 に関して詳細を記載する．最初の項 $P(s, L-1)$ は s が 1 から $L-|s|$ の間に出現する確率である．第 2 項である $Q(s)$ は $L-|s|+1$ から L に s が出現する確率である．$Q(s)$ は 1 から $L-|s|$ に s が含まれる確率も含んでいるため，$P(s, L-1)$ の確率と重複する確率がある．そこで，第 3 項の $Q(s)P(s, L-|s|)$ で s が 1 から $L-|s|$ に出現し，かつ $L-|s|+1$ から L にも出現する同時確率を除去する．文字列 s が自己重複領域を持たなければ，第 3 項で長さ L の配列が文字列 s を含む確率を計算できる．文字列 s が自己重複領域を持つ場合は，自己重複領域以外の部分が $L-2|s|+1+|t|$ から $L-|s|$ の間に含まれる確率を $Q(s)$ が含んでしまうため，この重複する確率をさらに除去する必要がある．そこで，文字列 s が $L-|s|+1$ から L に出現し，かつ自己重複領域以外の部分が $L-2|s|+1+|t|$ から $L-|s|$ に出現する同時確率を削除する．これは，第 4 項の $Q(s)\sum_{t\in\mathcal{T}}\frac{1}{q(t)}\{P(s, L-|s|+|t|)-P(s, L-|s|+|t|-1)\}$ で表される．

第 4 項の一部である，自己重複以外の部分が $L-2|s|+1+|t|$ から $L-|s|$ に出現する確率に関して補足する．文字列 s が $L-2|s|+1+|t|$ から $L-|s|+|t|$ にだけ存在する確率は，$Q(s)-Q(s)P(s, L-2|s|+|t|)$ であり，式 3.3 を変形して，次式で表される:

$$
\begin{aligned}
&Q(s)-Q(s)P(s, L-2|s|+|t|)= \\
&\qquad P(s, L-|s|+|t|)-P(s, L-|s|+|t|-1) \\
&\qquad Q(s)=\prod_{j\in\Omega}p_j^{n(s,j)},\quad q(t)=\prod_{j\in\Omega}p_j^{n(t,j)},\quad t\in\mathcal{T},\quad L\geq|s|. \quad (3.4)
\end{aligned}
$$

長さ $L-1$ 以下の s を含む確率の計算はすでに自己重複領域を考慮して計算済みのため，式 3.4 で自己重複領域を考える必要はない．式 3.4 から文字列 s の自己重複領域以外の部分が $L-2|s|+1+|t|$ から $L-|s|$ にのみ出現する確率に変換するためには，上式を $q(t)$ で割ればよいので次式となる:

$$
\frac{1}{q(t)}\left\{P(s, L-|s|+|t|)-P(s, L-|s|+|t|-1)\right\}. \quad (3.5)
$$

図 3.2 に，式 3.1 の例（文字列 s が ATA の場合）を示す．配列の長さが $L=3$ の場合，ATA が取りうる場合は一種類のみであるため，塩基の出現確率の掛け算 $P(\mathrm{ATA}, 3)=p_Ap_Tp_A=Q(\mathrm{ATA})$ となる．配列の長さが $L=4$ の場合，ATAN および NATA の 2 つの場合が考えられる．N は任意の塩基を示す．お互いの事象は背反であるため，$P(\mathrm{ATA}, 3)$ に $Q(\mathrm{ATA})$ を加えた $P(\mathrm{ATA}, 4)=P(\mathrm{ATA}, 3)+Q(\mathrm{ATA})$ が確率となる．配列の長さが $L=5$ の場合，$P(\mathrm{ATA}, 4)$ に $Q(\mathrm{ATA})$ を加えると，ATA の A の部分が自己重複しているため, ATATA を含む確率を二度計算してしまう．そこで, ATA が位置 3 から 5 に出現し，

L	1	2	3	4	5	6	7	Probability
3	A	T	A					$P(\mathrm{ATA},3)=p_{\mathrm{A}}p_{\mathrm{T}}p_{\mathrm{A}}=Q(\mathrm{ATA})$
4	A N	T A	A T	N A				$P(\mathrm{ATA},4)=P(\mathrm{ATA},3)$ $+Q(\mathrm{ATA})$
5	A N N A	T A N T	A T A A	N A T T	N N A A			$P(\mathrm{ATA},5)=P(\mathrm{ATA},4)$ $+Q(\mathrm{ATA})$ $-Q(\mathrm{ATA})P(\mathrm{ATA},3)/q(\mathrm{A})$
6	A N N N A	T A N N T A	A T A N A T	N A T A A A	N N A T T T	N N N A A A		$P(\mathrm{ATA},6)=P(\mathrm{ATA},5)$ $+Q(\mathrm{ATA})$ $-Q(\mathrm{ATA})P(\mathrm{ATA},3)$ $-Q(\mathrm{ATA})\{P(\mathrm{ATA},4)-P(\mathrm{ATA},3)\}/q(\mathrm{A})$
7	A N N N N A N	T A N N N T A	A T A N N A T A	N A T A N N A T	N N A T A A A A	N N N A T T T T	N N N N A A A A	$P(\mathrm{ATA},7)=P(\mathrm{ATA},6)$ $+Q(\mathrm{ATA})$ $-Q(\mathrm{ATA})P(\mathrm{ATA},4)$ $-Q(\mathrm{ATA})\{P(\mathrm{ATA},5)-P(\mathrm{ATA},4)\}/q(\mathrm{A})$

$p_j, j \in \{\mathrm{A},\mathrm{T},\mathrm{G},\mathrm{C}\}$ is a probability of nucleobase.

図 3.2　式 3.1 の例（文字列を ATA とした場合）

AT が位置 1 から 2 に出現する同時確率 (Joint probability) $Q(\mathrm{ATA})P(\mathrm{ATA},3)/q(\mathrm{A})$ を引く必要がある．配列の長さが $L=6$ の場合，$P(\mathrm{ATA},5)$ に $Q(\mathrm{ATA})$ を加えると，ATAATA を含む確率を二度計算してしまう．そこで，ATA が位置 1 から 3 と位置 4 から 6 に出現する同時確率 $Q(\mathrm{ATA})P(\mathrm{ATA},3)$ を引く必要がある．さらに，ATA の自己重複領域 A があるため，$L=5$ のときと同様に，位置 2 から 3 に AT があり，位置 4 から 6 に ATA がある同時確率 $Q(\mathrm{ATA})\{P(\mathrm{ATA},4)–P(\mathrm{ATA}),3)\}/q(\mathrm{A})$ を引く必要がある．同じように，配列の長さが $L=7$ の場合も，$P(\mathrm{ATA},6)$ に $Q(\mathrm{ATA})$ を加えて，同時確率 $Q(\mathrm{ATA})P(\mathrm{ATA},4)$ と $Q(\mathrm{ATA})\{P(\mathrm{ATA},5)–P(\mathrm{ATA}),5)\}/q(\mathrm{A})$ を引く必要がある．以下同様に，配列の長さが $L=8$ 以降も，任意の長さの配列が特定の文字列を含む確率を正確に計算することができる．

図 3.3 に式 3.1 の各項の意味を示す．1 行目 $P(s,L-1)$ は位置 1 から $L-1$ までに s を

<table>
<tr><td></td><td>1</td><td>2</td><td>…</td><td>…</td><td>$L-\lvert s\rvert$</td><td>$L-\lvert s\rvert+1$</td><td>…</td><td>$L-1$</td><td>L</td></tr>
<tr><td>1</td><td colspan="8">$P(s,L-1)$</td><td></td></tr>
<tr><td>2</td><td colspan="5">Any sequences</td><td colspan="4">$Q(s)$</td></tr>
<tr><td>3</td><td colspan="5">$P(s,L-\lvert s\rvert)$</td><td colspan="4">$Q(s)$</td></tr>
<tr><td>4</td><td colspan="3">String s is not included</td><td colspan="2">$\frac{1}{q(t)}\{P(s,L-\lvert s\rvert+\lvert t\rvert)-P(s,L-\lvert s\rvert+\lvert t\rvert-1)\}$</td><td colspan="4">$Q(s)$</td></tr>
</table>

図 3.3　式 3.1 の各項の説明

含む配列が出現する確率である．2 行目は $Q(s)$ は，位置 $L-|s|+1$ から L に s を含む配列が出現する確率である．確率 $P(s,L)$ を求めるために，まず 1 行目 $P(s,L-1)$ に 2 行目 $Q(s)$ を加える．ただし，$Q(s)$ は，位置 1 から $L-|s|$ はいかなる配列でも良い．そのため，位置 1 から $L-|s|$ に文字列 s がある配列の確率を二度計算することになる．そこで，3 行目と 4 行目で二度計算した確率を除去する必要がある．3 行目は，位置 1 から $L-|s|$ までに s があり，かつ位置 $L-|s|+1$ から L に s がある同時確率 $P(s,L-|s|)Q(s)$ である．4 行目は，自己重複のある文字列 s の，自己重複領域以外の部分が位置 $L-2|s|-|t|$ から $L-|s|$ に出現し，1 から $L-2|s|-|t|-1$ に s が出現せず，かつ s が $L-|s|+1$ から L に出現する同時確率である．配列の位置が 1 から $L-2|s|-|t|-1$ に s が出現しない条件が必要なのは，すでに 3 行目の $P(s,L-|s|)Q(s)$ で確率が除去されているためである．

式 3.1 は非常に汎用的な式であり，Ω の定義を変更することにより拡張が容易である．たとえば, Ω をアミノ酸の集合とすれば, 特定のアミノ酸の並びを含む配列の確率へと応用可能である．ほかにも，$\Omega=\{\mathrm{AAA, AAG, AAT, AAC,}\cdots\mathrm{, TTA, TTC, TTG, TTT}\}$ のような集合を考えれば，ゲノム上のコドンの探索にも応用できる．さらに，二次構造の要素を $\Omega_{\mathrm{structure}}=\{\mathrm{H, B, S, M, E, I, G}\}$ とし，$\Omega_{\mathrm{nucleobase}}=\{\mathrm{A, C, G, T(U)}\}$ としたとき，その直積である $\Omega_{\mathrm{nucleobase}}\times\Omega_{\mathrm{structure}}$ を用いると，二次構造を加味した文字列を含む配列の存在確率にも応用可能である．

3.4　Z スコアの定義

本節では，前節で求めた特定の文字列を含む配列の出現確率を用いて，文字列のスコアである Z スコアの定義を行う．式 3.1 は，文字列 s が長さ L の配列に出現する確率を示すものであるが，実際の配列には塩基の挿入や欠損による影響があり均一の長さではない．そのため s が出現する確率 $P(L,s)$ を，長さが不均一である配列データに対応させる必要がある．観測された配列の数を N とし i 番目の配列の長さを L_i とする．このとき，観測された HT-SELEX データにおいて s を含む配列が出現する確率は次式で表される:

$$P(s) = \frac{1}{N}\sum_{i=1}^{N} P(s, L_i). \tag{3.6}$$

ただし，文字列 s を含む配列の出現確率は他の配列に依存せず，互いに独立に同一分布に従うとする (i.i.d.).

式 3.6 より，HT-SELEX データに対して文字列 s を含む配列が出現する確率が導かれた．次に，実際に観測された文字列 s を含む配列数と s が出現する期待値を比較して文字列 s のスコアを定義する．配列が文字列 s を含むもしくは含まないという二値の結果を N 個の配列に対して行うことは，N 回のベルヌーイ試行と同義である．そのため，文字列 s が含まれる配列の数は $B(N, P(s))$ の二項分布に従う．図 3.4 に，母集団から N 本の配列をランダムサンプリングし，特定の文字列を含むもしくは含まないという試行（N 回のベルヌーイ試行）の概要を示す．配列の数 N が十分に大きい値のとき，文字列 s を含む配列の頻度 $F(s)$ と s を含む配列が出現する期待値 $NP(s)$ の差を，二項分布の標準偏差である $\sqrt{NP(s)(1-P(s))}$ で割ると，その値 $Z(s)$（文字列 s の Z スコア）は漸近的に標準正規分布に従う:

$$Z(s) = \frac{F(s) - NP(s)}{\sqrt{NP(s)(1-P(s))}} = \frac{\frac{F(s)}{N} - P(s)}{\sqrt{\frac{P(s)(1-P(s))}{N}}} \sim N(0,1). \tag{3.7}$$

この $Z(s)$ を文字列 s のスコアとし（Z スコア），Z スコアが高いほど期待値と比較して多く観測される ORS ということになる．一方，たとえ頻度の高い配列であっても，Z スコアの高い文字列を含まなければ，結合領域を持たないことになるため，標的分子と結合しない核酸分子である可能性が高いことになる．

式 3.7 の補足として，二項分布の N を増加させた場合に二項分布が漸近的に正規分布に従うことを示す例を図 3.5 に示す．N が 10 の場合には，二項分布は歪んだ形状をして

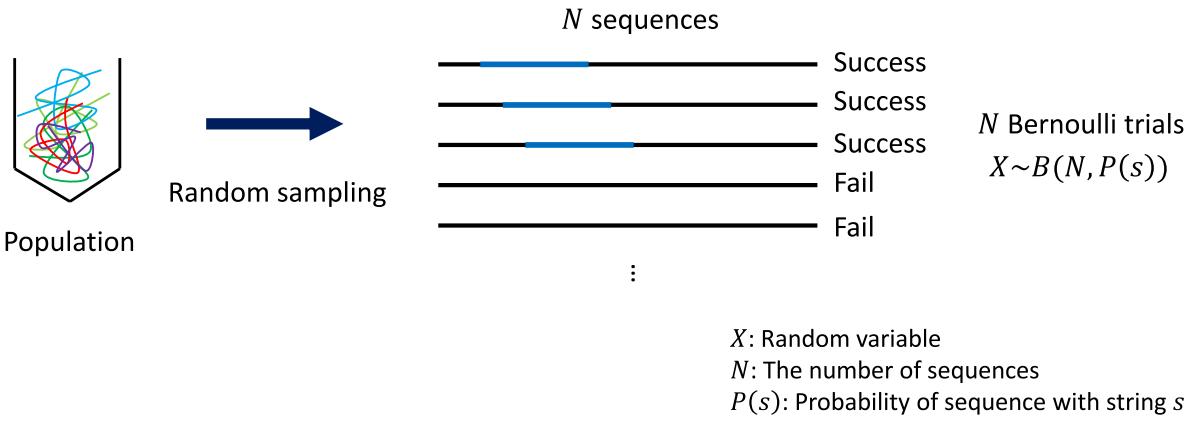

図 3.4　特定の文字列を含む配列が出現する分布

いるが，N を増やして 160 にすると，二項分布が正規分布の形状に類似することが図から読み取れる．

3.5　アプタマーの結合領域の高速な探索

前節で，文字列 s の Z スコアに関して定義をした．本節では，任意の長さの ORS を高速に探索するためのアルゴリズムに関して述べる．文字列 s の Z スコアを計算するためには，各々の塩基の存在確率が必要である．そのため，式 3.8 で塩基の存在確率を推定する:

$$\hat{p}_j = \frac{n_j}{\sum_{i=1}^{N} L_i}, \quad j \in \Omega. \tag{3.8}$$

前述したとおり，SELEX の工程のため各々の塩基のバランスは均一である保証がない．そこで，式 3.8 の塩基の存在確率の推定値を用いることにより，試験管内の塩基の偏りを考慮した Z スコアの計算ができる．

探索したい ORS の最小の長さを l_{min} として，最大の長さを l_{max} とする．すべての文字列を列挙するために必要な組み合わせの数は，$\sum_{l=l_{min}}^{l_{max}} |\Omega|^l$ である．組み合わせの数は指数関数的に増加するため，l_{max} が大きい場合，すべての文字列を列挙しその Z スコアを計算することは長い計算時間を要する．FSBC は短い ORS から順番に長い ORS を探索する．このとき，文字列の Z スコアと文字列に Ω の要素を加えた文字列の Z スコアを比較し，不要な探索領域を削減しながら長い ORS の探索を行う．たとえば，文字列が AAA であるとき，Ω の要素を伸長した文字列は AAAA，AAAT，AAAG，AAAC であ

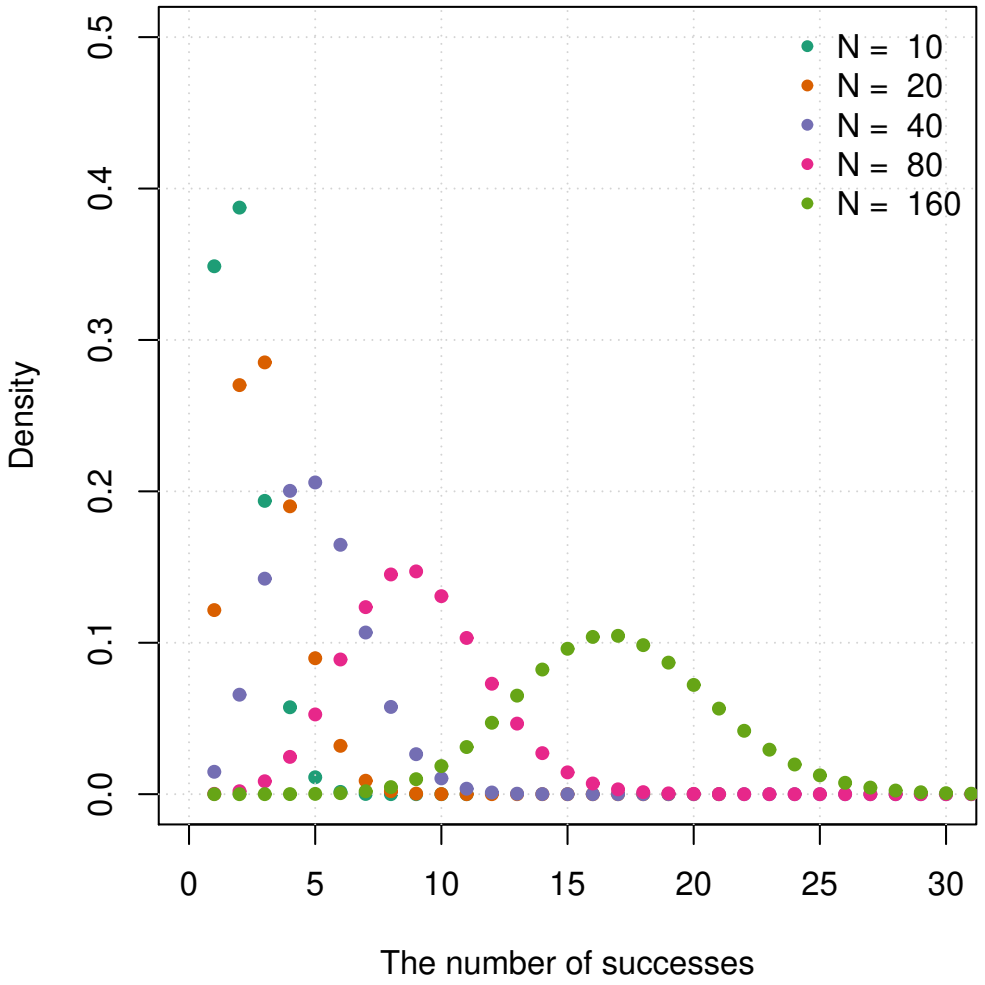

図 3.5　試行回数を増やしたときの二項分布の変化

る．このとき，AAA の Z スコアより低い Z スコアの文字列は探索領域から削除する．仮に，$Z(\mathrm{AAA}) > Z(\mathrm{AAAA})$ であったとき，AAAA は探索領域より削除され，AAAA に塩基を加えた AAAAA，AAAAT，AAAAG，AAAAC も探索されない．これは，AAAA に塩基を伸長した段階で，Z スコアが下がるため，それ以上塩基を伸長して ORS を探索することが効果的でないためである．以下に l_{min} から l_{max} の長さの ORS を探索する手

順を示す．

1. l_{min} の長さの文字列をすべて列挙し Z スコアを計算する．
2. Z スコアが 0 より大きい文字列を選択する．
3. $l \leftarrow l_{min} + 1$ とする．
4. 文字列に Ω の要素を加えて Z スコアを計算する．
5. 伸長した文字列の Z スコアが伸長する前の文字列の Z スコアより高い場合，その伸長した文字列を選択する．
6. $l \leftarrow l + 1$.
7. もし l が l_{max} より大きければ終了．そうでなければ 4 へ戻る．

この探索領域の削減方法は，伸長する前の文字列を親ノードとして伸長した後の文字列を子ノードとしたときの分枝限定法である．図 3.6 に，分枝限定法による ORS の探索方法を示す．まず，l_{min} の長さの文字列をすべて列挙し，その Z スコアを計算する．図中では $l_{min} = 3$ とする．Z スコアが 0 より小さい文字列は削除する．これは，Z スコアが 0 より小さい文字列は，期待値よりも低く出現しているためである．図中では，AAT と AAG が削除される．次に選択された文字列，AAA と AAG を一塩基伸長する．AAA を伸長した，AAAA, AAAT, AAAG, AAAC の Z スコアが AAA の Z スコアより高いものを選択し，そうでないものを除去する．図中では，AAAT と AAAG が削除される．同様に，AAAA から一塩基伸長して再び Z スコアを比較し選択する文字列と削除する文字列を決める．同様の作業を l_{max} まで行う．ここで，削除されなかった文字列が FSBC により選択された ORS である．

選択された ORS の集合を $\mathcal{S}$ とし，その数を $|\mathcal{S}|$ とする．このとき，すべての文字列を組み合わせた場合との関係は，$|\mathcal{S}| \ll \sum_{l=l_{min}}^{l^{max}} |\Omega|^l$ であり，分枝限定法による大幅な計算時間の削減ができる．また，l_{max} が 20 を超えるような長い ORS の探索でも現実的な時間で計算を終えることができる．アルゴリズム 1 に，ORS を探索するための疑似コードを記載する．

アルゴリズム 1 では，l_{min} の文字列の集合から $Z(s) > 0$ の文字列を選択しているが，$Z(0) > \zeta_0$ と拡張することができる．帰無仮説のもと l_{min} の文字列の集合から得られる $Z(s)$ は漸近的に標準正規分布に従うため，ζ_0 として片側 0.05% の $Z(s)$ を選択するということに変更することもできる．また，塩基を伸長する際の評価を $Z(s') > Z(i)$ としているが，この比較条件を拡張子 $Z(s') > \zeta Z(i)$ として選択条件を変更することも可能である．このように条件を拡張することにより，限定する分枝の数を増やしてさらに計算速度

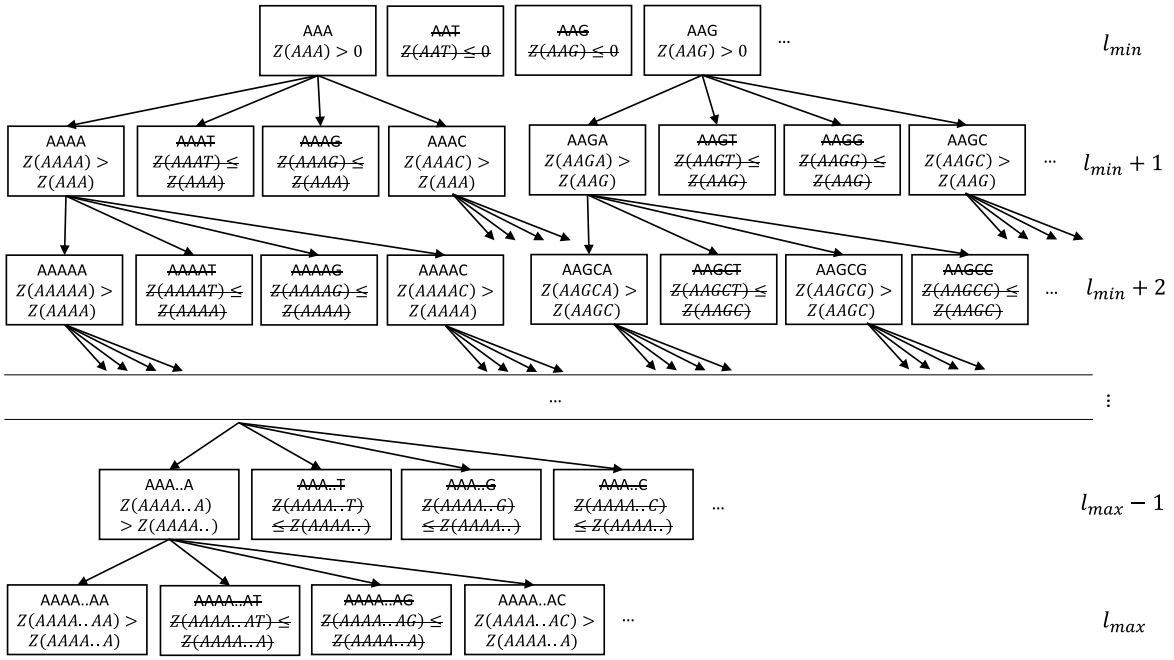

図 3.6　分枝限定法による ORS の探索

を上げることも可能である．

3.6　結合領域を考慮したクラスタリング

前節で，異なる長さ（l_{min} から l_{max} まで）の ORS を探索する方法を述べた．ORS の探索はアプタマーの標的分子との結合領域の推定と考えることができる．本節では，推定したアプタマーの結合領域を含む配列でクラスタを作成する手順を記載する．標的分子の種類，アプタマーと標的分子との結合様式，アプタマーが結合する標的分子のエピトープの違い等により，アプタマーの標的分子との結合領域の長さが異なる．そのため，異なる長さの ORS を比較し，どの長さの ORS が結合領域なのか判断する必要がある．ORS の探索において，高い Z スコアを選びながら長い ORS を探索するため，l_{max} の ORS が最

Algorithm 1: Selection of over-represented strings

Data: $\mathcal{D}$: Sequence data; l_{min}, l_{max}: The minimum and maximum length of strings for searching; $\Omega = \{A, T, G, C\}$: A set of nucleobases; $\Omega^* = \{\epsilon, A, T, G, C, AA, AT, AG, AC, \cdots\}$: Ω^* is a Kleene closure of Ω, and ϵ is an empty string;

Result: $\mathcal{S}$: A set of over-represented strings;

$\mathcal{S} \leftarrow \{s \mid s \in \Omega^* \wedge Z(s) > 0 \wedge |s| = l_{min}\}$;

for $l = l_{min}$ **to** $l_{max} - 1$ **do**

  $\mathcal{S}_l \leftarrow \{s \mid s \in \mathcal{S} \wedge |s| = l\}$;

  if $S_l = \emptyset$ **then**

    return $\mathcal{S}$

  end

  for $i \in \mathcal{S}_l$ **do**

    for $j \in \Omega$ **do**

      $s' \leftarrow \text{concatenate}(i, j)$;

      if $Z(s') > Z(i)$ **then**

        $\mathcal{S} \leftarrow \mathcal{S} \cup \{s'\}$;

      end

    end

  end

end

return $\mathcal{S}$;

も Z スコアが高くなる．そのため，Z スコアにより異なる長さの ORS を比較することはできない．そこで，選択された ORS の Z スコアを経験分布により各々の長さで正規化し，異なる長さの ORS がお互い比較できるように補正する．正規化した Z スコアを Z^* スコアとする．式 3.9 に，文字列 s の Z^* スコアを示す:

$$Z^*(s) = \frac{Z(s) - \hat{\mu}_{|s|}}{\hat{\sigma}_{|s|}}. \tag{3.9}$$

このとき，$\hat{\mu}_{|s|}$ および $\hat{\sigma}_{|s|}$ はそれぞれ s の長さの Z スコアの平均と標準偏差の推定値とし，探索された ORS から計算する．

Algorithm 2: Clustering

Data: $\mathcal{D}$: Sequence data; $\mathcal{S}$: A set of sorted over-represented strings by Z^*-score;
Result: $\mathcal{C}$: Clusters (a family of sets of sequences);
$\mathcal{C} \leftarrow \{\emptyset\}$;
while $\mathcal{D} \neq \emptyset$ **do**
  for $j = 1$ **to** $|\mathcal{S}|$ **do**
    $s_j \leftarrow$ the j-th string from $\mathcal{S}$;
    $\mathcal{C}_{s_j} \leftarrow \{d \mid d \in \mathcal{D} \wedge d$ includes $s_j\}$;
    if $\mathcal{C}_{s_j} \neq \emptyset$ **then**
      $\mathcal{C} \leftarrow \{\mathcal{C}, \mathcal{C}_{s_j}\}$;
      $\mathcal{D} \leftarrow \mathcal{D} - \mathcal{C}_{s_j}$;
    end
  end
end
return $\mathcal{C}$;

選択された ORS の集合 $\mathcal{S}$ を Z^* のスコアの高い順より並べ替えて，以下の手順により配列をクラスタに分ける．

1. HT-SELEX データを $\mathcal{D}$ とする．
2. $j \leftarrow 1$，$i \leftarrow 1$ とする．
3. j 番目の ORS を含む配列を $\mathcal{D}$ より選択する．
4. もし，j 番目の ORS を含む配列がなければ，$j \leftarrow j + 1$ として 2 へ．
5. i 番目のクラスタに選択した配列を格納し，$\mathcal{D}$ から選択した配列を取り除く．
6. $\mathcal{D}$ に配列がなければクラスタリングを終了する．
7. $i \leftarrow i + 1$, $j \leftarrow j + 1$ として 2 へ移動する．

ORS を用いたクラスタリングの疑似コードをアルゴリズム 2 に示す．

Z^* スコアにより，l_{min} から l_{max} の長さの ORS を比較し，アプタマーの標的分子との結合領域とその長さを推定することができる．この推定された結合領域は，アプタマーの

小型化の際にも利用される．

3.7　並列処理によるクラスタリングの高速化

本節では，高速化のため ORS の探索を並列化することを述べる．図 3.6 が示すように，伸長前の文字列の Z スコアと伸長後の文字列の Z スコアを比較しながら高いスコアの文字列の探索を行う．そのため，計算の依存関係は伸長前の文字列（親のノード）と伸長後の文字列（子のノード）にしかなく，子のノード同士では計算の依存関係がない．たとえば，文字列 AAA を伸長して文字列 AAAA，AAAT，AAAG，AAAC の Z スコアの比較を行うとき， AAAA，AAAT，AAAG，AAAC は AAA にのみ依存しており，お互いに独立した計算となる．また，異なるノードに関しても計算に依存関係がない．たとえば，AAA のノードと GGG のノードから塩基を伸長することは，お互いに独立した計算である．図 3.6 と同様に，アルゴリズム 1 でも計算の依存関係が示されている．まず，$\mathcal{S}_l$ のループにおいて，要素同士の依存関係はないことが示されている．つまり，$\mathcal{S}_l$ の各々の要素に対する計算は独立しており，いかなる順番で解析しても結果に影響はない．$\mathcal{S}_l$ のループと同様に Ω に関するループにおいても，要素同士の依存関係がない．そのため，$\mathcal{S}_l$ のループと Ω のループはそれぞれ並列処理が可能である．また，$\mathcal{S}_l$ のループと Ω のループ同士においても依存関係はない．つまり，一つ目のループを Ω のループとし，二つ目のループを $\mathcal{S}_l$ としても計算結果に影響がない．そこで，$\mathcal{S}_l$ と Ω のループに関して並列処理を実装した Parallel implementation of FSBC (pFSBC) の開発を行った [25, 26, 30]．また，pFSBC の実装では R ではなく Python [95] を用いた．これは，一般的に Python の方が R より高速処理に長けているためである．

3.8　クラスタの分布の比較

本節では，配列の多様性を定義し，FSBC により得られたクラスタの分布を比較する方法を述べる．HT-SELEX データのばらつきに対して，FSBC により作成されたクラスタがどのように分布しているのか確認することは，類似したクラスタからアプタマーの候補配列を選ぶことの回避につながる．これは，多様なアプタマーの候補配列を選択するうえで重要なことである．過去の研究において，HT-SELEX データの中でクラスタがどのように分布しているのか比較を行う方法は報告されていない．そこで，クラスタの分布を比較する手法を新たに提案する [25, 31]．クラスタの分布を比較するために，配列の特徴量

を ORS により定め（特徴ベクトルによる配列の表現），特徴量を用いて配列の多様性を定義する．

HT-SELEX データを，配列数 N と選択された ORS の数 $|\mathcal{S}|$ の次元の行列で表現する．この行列を，配列の多様性を表す行列として定義する．また，定義された行列に対して，次元圧縮法である Uniform manifold approximation and projection (UMAP) [96] を用いて，配列の多様性を二次元に射影する．次元圧縮の方法としては，主成分分析，t-Distributed Stochastic Neighbor Embedding (t-SNE) [97] などもあるが，主成分分析が非線形データへの対応が難しいこと，t-SNE は計算時間を要するため HT-SELEX のような大量データに向かないことを踏まえ，UMAP を次元圧縮の方法として利用する．UMAP は，リーマン多様体上にデータが一様分布している，距離は局所的に一定，多様体は局所的に接続されているという仮定のもと，次元を圧縮する方法である．近年では，遺伝情報の解析にも利用されている [98].

$\mathcal{D}'$ を冗長な配列を除いた HT-SELEX データとし，$\mathcal{D}'$ の配列の数を $|\mathcal{D}'$ とする．$\mathcal{S}$ を選択された ORS の集合とし，$|\mathcal{S}|$ を ORS の数とする．また，i 番目の配列を $d_i, i = 1, \cdots, |\mathcal{D}'$ とし，$b_i \in R^{|\mathcal{S}|}, i = 1, \cdots, |\mathcal{D}'$ を配列 d_i の特徴ベクトルとする．このとき，もし i 番目の配列 d_i が j 番目の文字列を含む場合，$b_{ij} = 1$ とし，それ以外は $b_{ij} = 0$ とする．もしくは，i 番目の配列 d_i が j 番目の文字列を含む場合は，$b_{ij} = Z^*_j$ として，それ以外は $b_{ij} = 0$ とする．前者は，ORS を含むかどうかを特徴としているが，後者はさらに ORS の重みを Z^* で考慮している．求めたベクトル b_i を組み合わせて $B \in R^{|D' \times |S|}$ を作成し，HT-SELEX データ D' の多様性を表す行列とする．0 もしくは 1 の二値から作成された多様性を表す行列と Z^* スコアより作成された多様性を表す行列の例を，表 3.1 と表 3.2 に示す．B に対して UMAP を適用して二次元の空間へと射影し ($B \to B_2 \in R^{|D'2}$)，配列の多様性を可視化する．

3.9　hESC を標的分子とした HT-SELEX データによる性能評価

hESC を標的分子とした SELEX に NGS を適用して得られた HT-SELEX データ [33] を，FSBC の性能評価に用いた．hESC は分化多能性[*1]を保有し，ほぼ無限に増殖が可能な細胞である．hESC は，その特徴から再生医療において大きく期待されている．た

[*1] 生体を構成する多様な種類の細胞に分化する能力.

表 3.1　二値で配列の多様性を表す行列を作成した例

	GTGCGA	GGTGC	GGTGCGATTG	GTGGTGCGAT	GTGCG
Seq1	1	1	1	1	1
Seq2	1	1	0	0	1
Seq3	0	0	0	0	0
Seq4	0	1	0	0	0
Seq5	0	0	0	0	0
Seq6	0	0	0	0	0

表 3.2　Z^* スコアで配列の多様性を表す行列を作成した例

	GTGCGA	GGTGC	GGTGCGATTG	GTGGTGCGAT	GTGCG
seq1	2.845069	2.623834	2.607035	2.596472	2.586775
seq2	2.845069	2.623834	0.000000	0.000000	2.586775
seq3	0.000000	0.000000	0.000000	0.000000	0.000000
seq4	0.000000	2.623834	0.000000	0.000000	0.000000
seq5	0.000000	0.000000	0.000000	0.000000	0.000000
seq6	0.000000	0.000000	0.000000	0.000000	0.000000

だし，hESC は受精卵を用いるために倫理的な問題がある．hESC に対するアプタマーは，細胞の標識や細胞の精製に用いることができる．hESC の SELEX は 5 ラウンド実施されており，HT-SELEX のデータも 5 ラウンドまでのデータが公開されている．19 の配列がフローサイトメトリー[*2]により結合評価されており，その中で 8 配列が結合を示し，11 配列が結合を示していない．実験における解析実行環境は，CPU が Intel(R) Xeon(R) CPU E5-1650v4@3.60GHz，メモリが 64GB，オペレーティングシステムは Ubuntu 16.04 (Xenial Xerus) である．

[*2] 蛍光などで標識した細胞をふるい分ける手法.

3.9.1　実験方法

計算速度

計算速度の評価には，hESC を標的分子とした HT-SELEX の第 5 ラウンドの HT-SELEX データを用いた（ただし，AptaTRACE は複数のラウンドの HT-SELEX データが必要なため，第 4 ラウンドと第 5 ラウンドの HT-SELEX データを用いた）．HT-SELEX データのサイズを配列の頻度によるフィルタリングで変更し，クラスタリングにかかる計算速度を計測した．配列の頻度のフィルタリングには，フィルタリングなし（全配列），頻度が 10 以上の配列，頻度が 100 以上の配列とした．このとき，配列の数は，全配列の場合は 15,327,604，頻度が 10 以上の配列の場合は 8,799,219，頻度が 100 以上の配列の場合は 4,947,522 であった．また，重複配列を除いた配列数は，全配列の場合は 4,381,160，頻度が 10 以上の配列の場合は 156,587，頻度が 100 以上の配列の場合は 6,193 であった．

FSBC の比較対象として，従来の HT-SELEX データのクラスタリング手法である FASTAptamer，AptaCluster，APTANI，AptaTRACE を用いた．FASTAptamer はユーザーガイドに従い同一クラスタに含める LD の閾値を 7 とし，最大のクラスタ数を 100 と設定した．AptaCluster は既定のオプションで実装した．APTANI は規定のオプションでは頻度によるフィルタリングを自動的に行うため，このオプションを適用しなかった．APTANI のそれ以外のオプションは規定とした．AptaTRACE に関しては，background sequence のオプションを 1,000 とした．これは，ほかの値を background sequence オプションに指定した場合と比較して，精度の高い結果が得られたためである．FSBC は $l_{min} = 5$，$l_{max} = 10$ として計算を行った．FSBC の実装には R [99] と Bioconductor package [100] を用いた．FASTAptamer，AptaCluster，APTANI，AptaTRACE は配布されている実行形式のプログラムを利用した．

クラスタリング精度

クラスタリングによりアプタマーが順位の高いクラスタに分類され，hESC に結合しない核酸分子が下位のクラスタに分類されることが望ましい．まず，結合評価された 19 配列が属するクラスタの順位を求めた．このとき，19 配列と頻度との関係を調べ，FSBC が頻度の高い hESC に結合しない核酸分子に対して頑健であることを調べた．次に，結合評価された 19 の配列の結合の有無と属するクラスタの順位により，受信者操作

特性 (Receiver operating characteristic: ROC) 曲線を図示し，ROC に対する曲線下面積 (Area under the curve: AUC) により評価を行った．

クラスタリングの精度評価には，hESC を標的分子とした HT-SELEX の第 5 ラウンドの頻度が 10 以上の HT-SELEX データを用いた（ただし，AptaTRACE は複数のラウンドの HT-SELEX データが必要なため，計算速度の評価と同様に第 4 ラウンドと第 5 ラウンドの HT-SELEX データを用いた）．頻度が 10 以上の配列を用いた理由は，FASTAptamer と APTANI が HT-SELEX データの全配列で計算が終了しなかったためである．

計算速度の評価と同様に，FSBC の比較対象として FASTAptamer，AptaCluster，APTANI，AptaTRACE を用いた．すべての手法において計算速度の評価と同じパラメータを利用した（FASTAptamer の LD の閾値を変化させても精度に大きな変化は見られなかったため，計算速度の評価と同じパラメータを利用している）．ただし，例外として FASTAptamer の最大クラスタ数のパラメータの指定は除外した．FSBC に関しては，オプションによる精度の違いを調べるため，$l_{min} = 3, 4, 5$，$l_{max} = 10$ とした．AptaCluster は，クラスタ内の配列の総頻度によるクラスタの順位と，クラスタ内の非重複配列の数によるクラスタの順位を計算するため，精度評価には二つのクラスタの順位を適用した．APTANI にはクラスタの順位を出力する機能がないため，AptaCluster に倣い，クラスタ内の配列の総頻度とクラスタ内の非重複配列の数をもとにクラスタの順位として評価を行った．

FSBC は Z スコアを比較した分枝限定法により，長い ORS の探索を実施している．分枝限定法のため，$l_{min} + 1$ 以上の長さの探索した ORS に最大の Z スコアの ORS が含まれる保証がない．そこで，l_{min} から l_{max} までの文字列を全列挙して Z スコアを計算し，最も高い Z スコアの ORS が，探索した ORS に含まれるかどうか評価を行った．また，上位 10 の Z スコアの ORS が，探索した ORS にいくつ含まれるか評価した．さらに，選択された ORS の Z スコアの分布が ORS の長さに依存することを図示し，Z スコアでは異なる長さの ORS を直接比較できないことを示した．一方，ORS の長さで補正した Z^* スコアでは，異なる長さの ORS が比較できることを示した．

FSBC を第 5 ラウンドの HT-SELEX データの他に第 3 ラウンド，第 4 ラウンドの HT-SELEX データ（全配列）にも適用した．これは，アプタマーの濃縮が十分でない早期ラウンドの SELEX において，アプタマーと hESC と結合しない核酸分子の配列を FSBC により正しく分割できるのか評価するためである．評価方法は，5 ラウンドのときと同様に，まず，結合評価された配列が属するクラスタの順位を求める．次に，クラスタ

の順位と配列の頻度との関連を調べ，FSBC が頻度の高い hESC に結合しない核酸分子の配列に対して頑健であることを調べる．最後に，クラスタの順位と hESC への結合の有無により AUC を計算する．第 1 ラウンドと第 2 ラウンドの HT-SELEX データは評価対象から外した．これは，結合評価を行った 19 配列が第 1 ラウンドと第 2 ラウンドの HT-SELEX データから観測されなかったためである．

3.9.2　結果

計算速度

表 3.3 に FASTAptamer, AptaCluster, APTANI, AptaTRACE, FSBC の計算時間を示す．最初の列にクラスタリング手法の名称を記載し, 2 列目以降に全配列の HT-SELEX データ，頻度が 10 以上の HT-SELEX データ，頻度が 100 以上の HT-SELEX データの計算時間（実時間と CPU 時間）を記す．

頻度が 10 以上と 100 以上の HT-SELEX データでは，すべてのクラスタリング手法で解析が終了した．しかし, 全配列の HT-SELEX データでは, FASTAptamer と APTANI が解析を終了できなかった．FASTAptamer は全配列の HT-SELEX データの解析において，7 日以上計算を続けたが解析を終了しなかった．APTANI は全配列の HT-SELEX データの解析において，配列の二次構造を推定後にエラーを出力し解析を終了した（二次構造の推定には 25 時間を要した）．そのため，本来の HT-SELEX のデータを解析可能なクラスタリング手法は，AptaCluster，AptaTRACE，FSBC に限られる．

すべてのデータサイズにおいて AptaCluster が最も速く，提案したクラスタリング手法である FSBC は 2 番目の速さであった．AptaCluster は全配列の HT-SELEX データの計算において, 実時間が 3 分 45 秒と他のクラスタリング手法と比較して, 非常に速い計算時間を示した．FSBC は全配列の HT-SELEX データの計算において, 実時間が 4 時間 40 分 51 秒で解析を終了し，AptaCluster と比較して長い計算時間を要しているが，現実的に利用できる計算時間と考えられる．一方，AptaTRACE は全配列の HT-SELEX データの解析において, 71 時間 38 分 35 秒を要したため（およそ 3 日），全配列の HT-SELEX データ解析はあまり現実的ではない．AptaTRACE の実時間と CPU 時間が大きく異なるのは，AptaTRACE に実装されている並列処理の効果である．

表 3.3　各クラスタリング手法の計算時間

	All sequences		Sequences ($\geq$ 10)		Sequences ($\geq$ 100)	
Method	Real time	CPU time	Real time	CPU time	Real time	CPU time
FASTAptamer	DNF[1]	DNF[1]	5 h 16 m 4 s	5 h 16 m 3 s	10 m 40 s	10 m 40 s
AptaCluster	3 m 45 s	4 m 9 s	33 s	26 s	28 s	17 s
APTANI	DNF[2]	DNF[2]	32 m 52	34 m 59 s	1 m 47 s	1 m 20 s
AptaTRACE	71 h 38 m 35 s	246 h 15 m 12 s	1 h 1 m 17 s	2 h 2 m 50 s	3 m 52 s	5 m 44 s
FSBC	4 h 40 m 51 s	4 h 40 m34 s	9 m 25 s	9 m 17 s	51 s	46 s

Abbreviation: DNF, did not finish
DNF[1]: FASTAptamer did not finish the calculation in one week.
DNF[2]: APTANI stopped the calculation with an error message after the secondary structure estimation which took 25 hours.

クラスタリング精度

FASTAptamer，AptaCluster，APTANI，AptaTRACE，FSBC による，結合評価された 19 配列のクラスタリングの結果を表 3.4 に示す．1 列目は結合評価された配列の ID，2 列目が配列の頻度による順位，3 列目が配列の頻度，4 列目が結合評価結果，5 列目以降は各クラスタリング手法によって作成されたクラスタの順位を示す．たとえば，seq1 は FASTAptamer において 6 番目のクラスタに属し，AptaTRACE では 1 番目のクラスタに属している．AptaCluster (Frequency)，AptaCluster (Diversity)，APTANI (Frequency)，APTANI (Diversity) は，それぞれ AptaCluster と APTANI におけるクラスタ内の配列の総頻度によるクラスタの順位と，クラスタ内の非重複配列の数によるクラスタの順位である．seq1 から seq8 までが hESC と結合するアプタマーであり，seq9 から seq19 までが hESC と結合しなかった核酸分子の配列である．アプタマーの配列と hESC に結合しない核酸分子の配列は，それぞれ配列の頻度により降順に並べられている．アプタマーの中で最も頻度の低い seq8 は（頻度が 4，割合は $4/8,799,219 = 4.54 \times 10^{-7}$），クラスタリング精度評価に用いた HT-SELEX データ（頻度が 10 以上）に含まれていない．そのため，結果に NA と記載する．ただし，seq8 は hESC と結合するアプタマーである．FASTAptamer，APTANI は全配列の HT-SELEX データでの解析ができないため，seq8 のような頻度が低いアプタマーを取りこぼしてしまう．APTANI (Diversity) において，seq5，seq6，seq7 と seq10，seq11 が同一順位のクラスタに属しているが，クラスタの順位がたまたま同順位であるだけで実際には異なるクラスタに属している．AptaTRACE の解析では，seq6，seq13，seq14，seq16，seq18，seq19 がいかなるクラスタにも属さなかっ

たため, NA と記載する．FSBC に関しては, 最も良い結果を示した $l_{min} = 5$, $l_{max} = 10$ の結果のみを記載する．FASTAptamer, AptaCluster (Frequency/Diversity), APTANI (Frequency/Diversity), AptaTRACE, FSBC が作成したクラスタの数はそれぞれ 2,380, 136,350，2,348，13，155 であった.

FSBC と AptaTRACE では，1 番目のクラスタにアプタマーが属している．しかし，FASTAptamer，AptaCluster (Frequency/Diversity)，APTANI (Frequency/Diversity) では 1 番目のクラスタ，もしくは 2 番目のクラスタに hESC と結合しない核酸分子の配列が属している．FASTAptamer，AptaCluster (Frequency/Diversity)，APTANI (Frequency) の 1 番目，もしくは 2 番目のクラスタには，最も頻度の高い配列である seq9 が含まれており，クラスタの順位が配列の頻度に強く影響を受けていることがわかる．HT-SELEX において 1 番頻度の高い配列は，しばしば標的分子と結合を示さないことがある．これは，PCR のバイアスによって増えやすい核酸分子，担体に非特異的に結合する核酸分子などの影響である．FASTAptamer，AptaCluster，APTANI は，クラスタの順位が頻度に強く依存するため，頻度の高い標的分子と結合しない核酸分子の配列を高い順位のクラスタから除外することができない.

FSBC では，アプタマーが属するすべてのクラスタの順位が，標的分子と結合しない配列が属するクラスタの順位より上位である．AptaTRACE も FSBC と同様に，1 番目のクラスタにアプタマーを含んでいるが，アプタマーである seq6 はいかなるクラスタにも属さず，またアプタマーである seq7 は hESC に結合しない核酸分子である seq9 と seq10 より下位のクラスタに属している．さらに，seq17 の hESC と結合しない核酸分子の配列が，アプタマーである seq6 と同じクラスタに属しており，同一クラスタ内にアプタマーと標的分子と結合しない配列が混在してしまっている．これは，AptaTRACE が正しい結合領域を推定していないためと考えられる．AptaTRACE は複数のラウンドの HT-SELEX データを利用するのに対し，FSBC は単一のラウンドの HT-SELEX データしか利用しないが，AptaTRACE よりも精度の高いクラスタリング結果を示した．AptaTRACE が，ノイズ配列（頻度は高いが標的分子と結合しない配列）の影響を受けて正しく結合領域を推定していないことに対し，FSBC はノイズ配列に対して頑健であったことが考えられる．以上より，FSBC はアプタマーの候補配列を効率よく選択するために有用であるといえる.

表 3.4　各手法によるクラスタリングの結果

Sequence information				Cluster ranking						
ID	Ranking	Frequency	Binding	FASTAptamer	AptaCluster (Frequency)	AptaCluster (Diversity)	APTANI (Frequency)	APTANI (Diversity)	AptaTRACE	FSBC ($l_{min} = 5$)
seq1	6	92237	Yes	6	7	5	7	870	1	5
seq2	24	20057	Yes	15	17	15	15	699	1	1
seq3	63	8750	Yes	24	64	65	58	290	1	1
seq4	82	6753	Yes	15	81	72	68	2188	1	1
seq5	255	1483	Yes	60	229	112740	102	626**	1	1
seq6	8459	84	Yes	546	9921	28056	1993	626**	NA***	5
seq7	100914	15	Yes	731	94490	125262	2038	626**	5	5
seq8*	281478	4	Yes	NA	NA	NA	NA	NA	NA	NA
seq9	1	583447	No	1	1	2	1	125	4	26
seq10	8	70095	No	7	8	10	8	916**	4	16
seq11	10	51669	No	9	11	9	16	916**	7	54
seq12	12	45038	No	10	12	13	13	520	11	41
seq13	23	20380	No	14	21	23	45	2270	NA***	41
seq14	375	831	No	75	335	76783	387	1739	NA***	37
seq15	398	771	No	78	238	556	460	2188	8	47
seq16	520	504	No	107	466	120874	1758	2253	NA***	11
seq17	3847	126	No	388	4568	59849	92	1	5	66
seq18	29324	41	No	50	539	110	44	323	NA***	92
seq19	44000	31	No	50	9134	4859	2043	2253	NA***	88

*: seq8 was excluded because the frequency was lower than the cutoff.

**: The cluster ranks were tied; however, the sequences are grouped into different clusters.

***: The sequences are not clustered. Because these sequences do not include any sequence-structure motif.

図 3.7 に，配列が属するクラスタの順位と配列の頻度との関連を示す．赤と青い点は，表 3.4 の配列を示し，赤は hESC に結合した配列（アプタマー），青は結合しなかった核酸分子の配列を示す．灰色の点は結合評価されていない配列を示す．表 3.4 と同様に，FASTAptamer，AptaCluster (Frequency)，AptaCluster (Diversity)，APTANI (Frequency) では，配列の頻度とクラスタの順位が強い関連を示す．一方，AptaTRACE，FSBC では配列の頻度とクラスタの順位との関連が FASTAptamer，AptaCluster，APTANI と比較して小さい．AptaTRACE と FSBC では，アプタマーが 1 番目のクラスタに属しており，FASTAptamer，AptaCluster，APTANI より良い精度であることがわかる．すべてのクラスタリング手法の中で，FSBC が最も頻度の高い標的分子と結合しない配列を下位のクラスタに分けていることがわかり，FSBC が頻度は高いが標的分子と結合しない配列に対して最も頑健であるといえる．

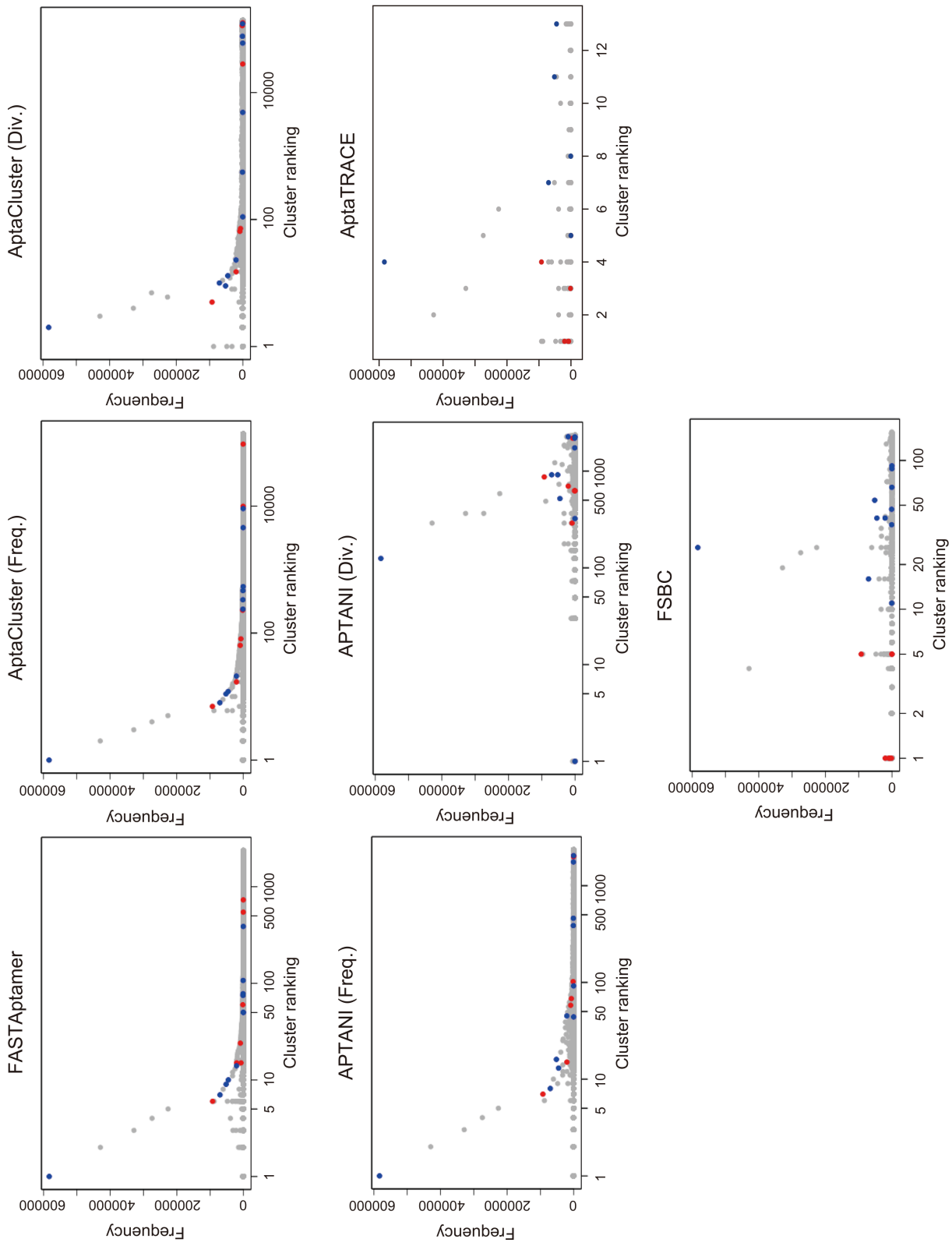

図 3.7　配列が属するクラスタの順位と頻度との関係

図 3.8 に，配列が所属するクラスタの順位と，結合評価の結果をもとにした ROC 曲線を示す．ROC 曲線の図の右下には，各手法とオプションによる AUC を記載する．FSBC は，$l_{min} = \{4, 5\}$，$l_{max} = 10$ において，アプタマーと hESC と結合しない核酸分子の配列を完全に分けるクラスタを作成した (AUC = 1.00)．また，FSBC では，$l_{min} = 3$，$l_{max} = 10$ においても，0.96 という高い AUC を示した．AptaTRACE も同様に，0.86 と高い AUC を示したが，すべてのオプションにおいて FSBC が最も良い AUC を示した．AptaTRACE でクラスタに属さなかった配列があるが，これらは結合評価を行う候補配列から除外されることになる．そのため，どこにも属さなかった配列は最も下位のクラスタより低いクラスタに属しているとして AUC の計算を行った．FASTAptamer, AptaCluster，APTANI からは，高い AUC が得られなかった．これは，前述したように頻度の高い hESC と結合しない核酸分子の影響を大きく受けているためである．以上より，全長配列の類似度からクラスタを作成する手法 (FASTAptamer, AptaCluster, APTANI) より，AptaTRCE や FSBC のようにアプタマーの結合領域を考慮したクラスタリングの方が良い結果が得られた．さらに，FSBC は AptaTRACE より高い精度を示しており，従来の手法すべてと比較して最も精度の高いクラスタリングの結果を示した．このことは，アプタマーの候補配列を効率よく選択するうえで，FSBC が有用であることを示す．

FSBC の異なるオプションにおいて，結合評価した配列が属するクラスタの順位と頻度との関係を図 3.9 に示す．濃い青い点が結合を示した配列であり，薄い青が結合しなかった配列である．点に付随する番号は，頻度による順位を示す．FSBC では，探索する文字列の最小の長さを短くすると（$l_{min} = 3$ の場合），4 番目のクラスタにアプタマーと標的分子と結合しない核酸分子の配列を含んでしまう．ただ，$l_{min} = 3$ のときに，最も頻度が高く標的分子と結合しない核酸分子の配列が属するクラスタの順位が，最も低い順位となっている．そのため，$l_{min} = 3$ のときは，頻度の高い標的分子と結合しない核酸分子の配列に対して，最も頑健であるといえる．$l_{min} = 4$ と $l_{min} = 5$ の場合，すべてのアプタマーが標的分子と結合しない核酸分子の配列より上位のクラスタに分けられた．そのため，hESC の HT-SELEX データにおいて，アプタマーと hESC に結合しない核酸分子の配列を分ける十分な感度を得るには，l_{min} が少なくとも 4 以上である必要がある．

FSBC ($l_{min} = 5$，$l_{max} = 10$) はクラスタリングにおいて，1,003 の ORS を選択して 155 のクラスタを作成した．選択した ORS の数は，すべての文字列の組み合わせと比較すると非常に少ない．$l_{min} = 5$，$l_{max} = 10$ の文字列の組み合わせの数は，$\sum_{l=5}^{10} 4^l = 1,397,760$ であり，実際に選択された文字列の組み合わせとの比率は，

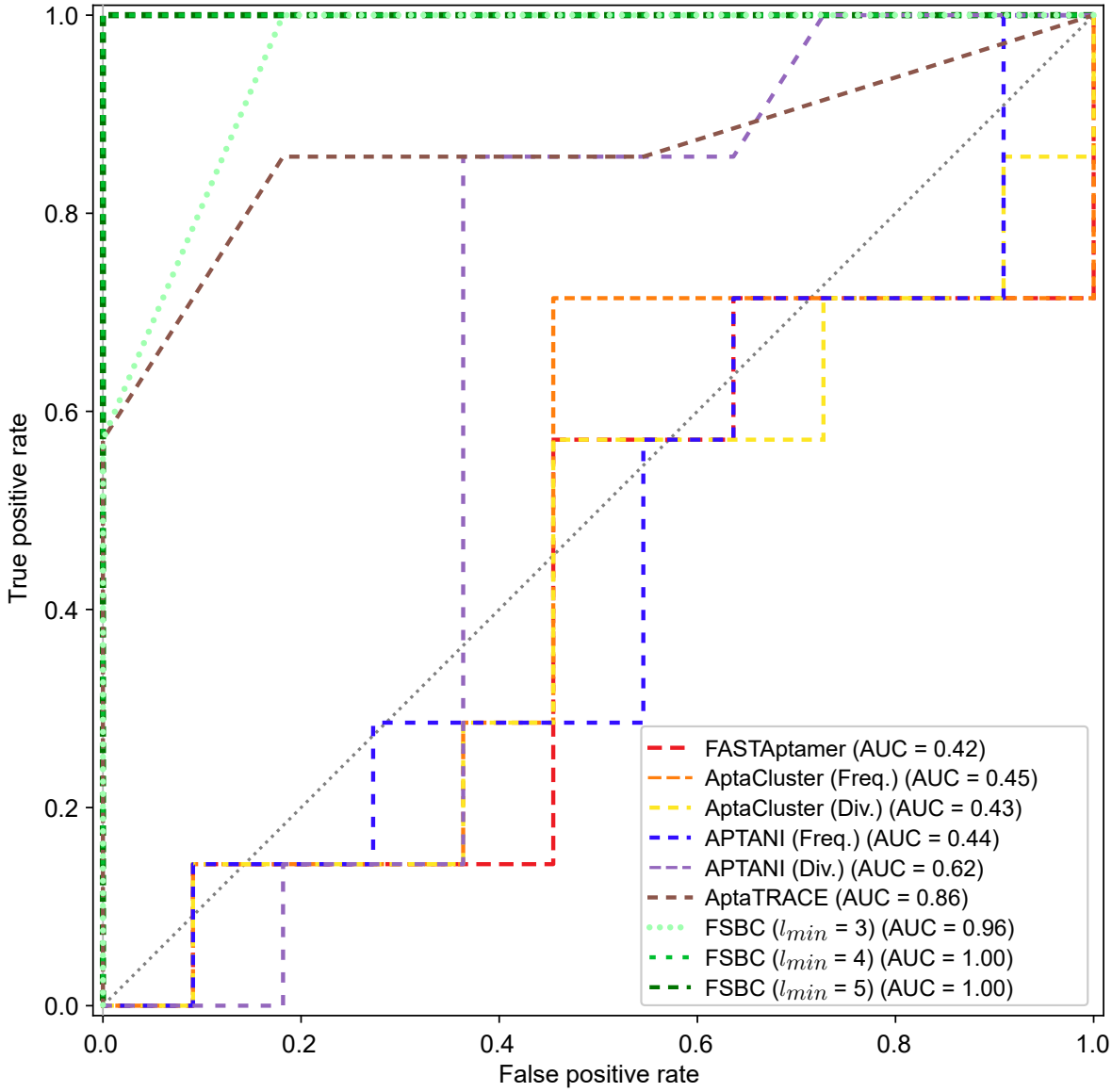

図 3.8　ROC 曲線

$1,003/1,397,760 = 0.0007175767$ である．この比率は，FSBC が探索領域の削減により，高速に ORS を探索した効果である．しかしながら，FSBC の ORS の探索には分枝限定法を用いているため，すべての組み合わせの文字列の中で最も高い Z スコアの ORS が選ばれている保証がない．そこで，長さ 5 から 10 までの文字列をすべて列挙し，選択された ORS の中に，最も Z スコアの高い ORS が含まれているのか確認した．同時に，

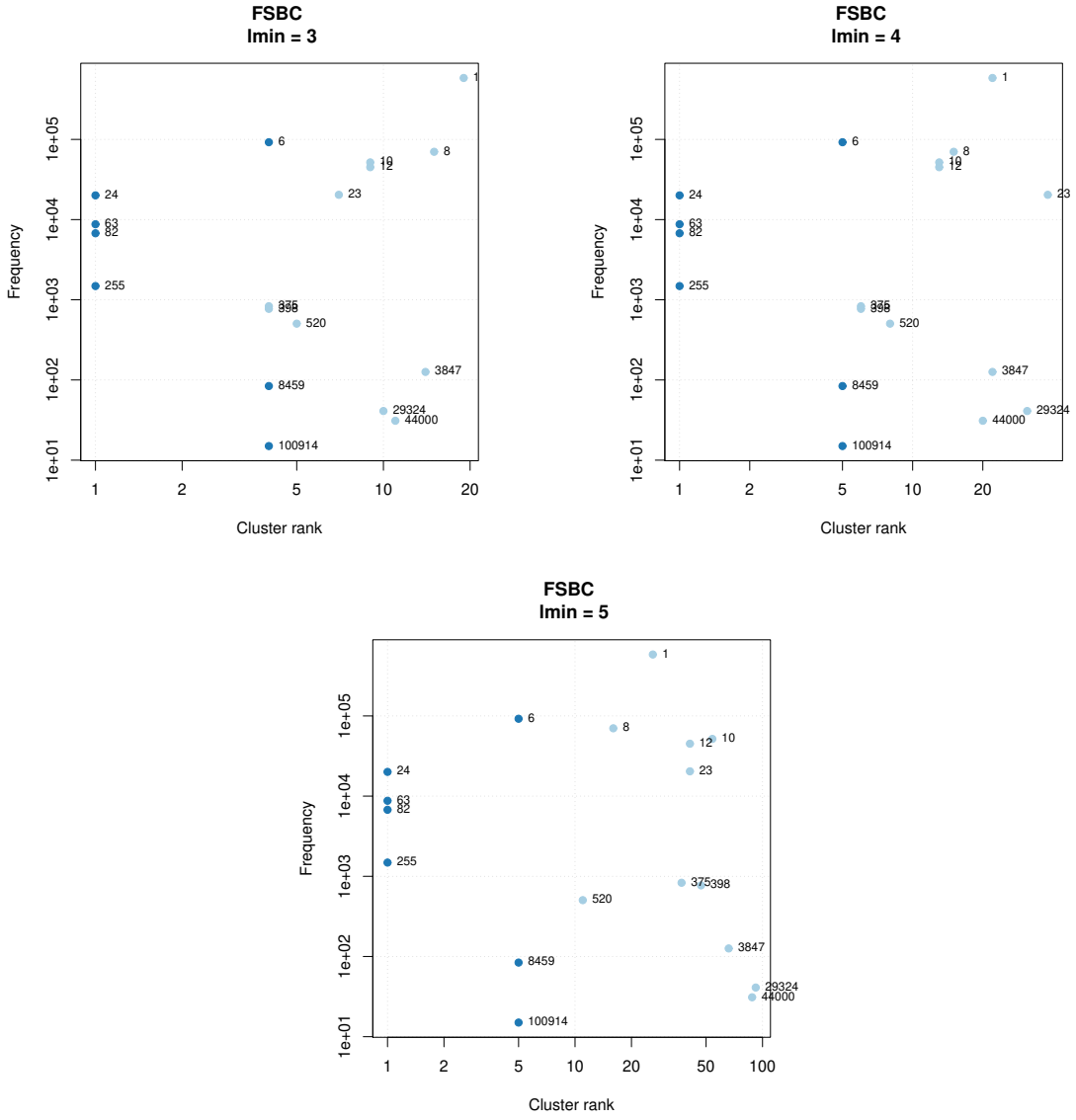

図 3.9　FSBC の各オプションにおける結合評価した配列が属するクラスタの順位と頻度との関係

上位 10 の Z スコアの ORS が，選択された ORS にいくつ含まれているのかも確認した．表 3.5 に全文字列と選択された文字列との関係を示す．1 列目が文字列の長さ，2 列目が最も Z スコアの高い ORS が，選択した ORS に含まれているかどうかを表し，3 列目が上位 10 の Z スコアの ORS が，選択された ORS に含まれる数である．FSBC では，最も Z スコアの高い ORS は，すべての長さの ORS でも選択されている．また，上位 10 の

表 3.5　探索した ORS と全列挙による ORS との比較

ORS の長さ	最も Z スコアの高い ORS が選択できたか	上位 10 の Z スコアの ORS が選ばれた数
5	Yes	10
6	Yes	7
7	Yes	7
8	Yes	6
9	Yes	6
10	Yes	6

Z スコアの ORS も半分以上は選択されている．そのため，分枝限定法による長い ORS の探索は効果的であることがいえる．

選択された 1,003 の ORS の Z スコア，Z^* スコアと ORS の長さに対する分布を図 3.10 に示す．左の図が，選択された ORS の Z スコアの分布を示し，右の図が選択された ORS の Z^* スコアの分布を示す．高い Z スコアの ORS を探索するというアルゴリズムの特性上，選択された ORS が長くなると Z スコアの分布も段階的に高くなる．そのため，長さの異なる ORS を比較する際には長さごとの補正が必要になる．Z スコアを補正し Z^* スコアとすることにより，異なる長さの ORS の比較ができる．アプタマーの結合領域の長さは，標的分子の種類，標的分子との結合様式やエピトープに依存し，どのような長さの結合領域となるか不明である．Z^* スコアにより，異なる長さの ORS が比較できるため，異なる長さの ORS の中から結合領域を推定するのに Z^* は有用である．

FSBC を第 5 ラウンドの HT-SELEX データだけでなく，第 3 ラウンドと第 4 ラウンドにも適用した結果を表 3.6 に示す．1 列目と 2 列目は，それぞれ配列の ID と結合結果を示す．3 列目から 5 列目は第 3 ラウンドの配列の頻度と頻度の順位とクラスタの順位であり，6 列目から 8 列目は第 4 ラウンドの配列の頻度と頻度の順位とクラスタの順位である．第 3 ラウンドと第 4 ラウンドにおいて観測されない配列は NA と記載した．表 3.4 の第 5 ラウンドの結果と同様に，FSBC は第 3 ラウンド，第 4 ラウンドでも順位の高いクラスタにアプタマーが属している．そのため，第 3 ラウンドと第 4 ラウンドにおける AUC はそれぞれ 0.89 と 1.00 であった．このように，アプタマーの濃縮が十分でない早期ラウンドにおいても，FSBC は精度の高いクラスタリング結果を示した．FASTAptamer と APTANI は，頻度のフィルタリングを行わないと HT-SELEX データを計算できないた

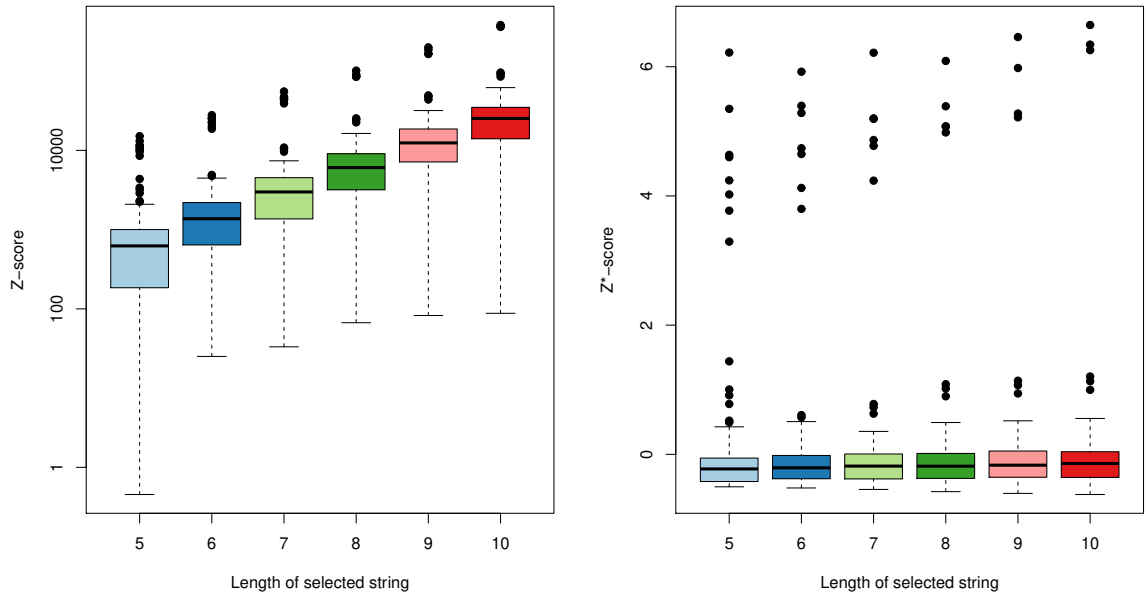

図 3.10　選択された ORS の長さに対する Z スコアと Z^* スコアの分布

め，そもそも第 3 ラウンドのようなアプタマーがあまり濃縮されていないラウンドで，アプタマーの探索を行うことができない．しかし，FSBC では第 3 ラウンドにおいて，頻度が 10 以下のアプタマーは上位のクラスタに属し，頻度が 89 の標的分子と結合しない核酸分子の配列は下位のクラスタに属している．

結合評価した配列が所属するクラスタの順位と頻度との関係を図 3.11 に示す．茶色がアプタマーであり，緑が標的分子と結合しない核酸分子の配列である．第 3 ラウンドにおいて，非常に低い頻度のアプタマーが上位のクラスタに属していることが図に示されている．特に，頻度が 10 のアプタマーは 1 番目のクラスタに分けられている．一方で，最も頻度の高い標的分子と結合しない核酸分子の配列は，順位の低いクラスタに分けられている．よって，たとえアプタマーの濃縮が十分でない，早期ラウンドにおいても FSBC の精度が高いことがわかる．FSBC の利用により，早期ラウンドで精度の高いアプタマーの候補配列が選択できれば，SELEX のラウンドの短縮によるコストと時間の削減にもつながる．

表 3.6　第 3 ラウンドと第 4 ラウンドの FSBC のクラスタリング結果

Sequence information		The third round			The fourth round		
ID	Binding	Ranking	Frequency	Cluster ranking	Ranking	Frequency	Cluster ranking
seq1	Yes	38	10	3	15	6837	2
seq2	Yes	82	5	5	53	1639	1
seq3	Yes	457	3	5	112	926	1
seq4	Yes	10134	2	5	130	747	1
seq5	Yes	474449	1	34	308	306	1
seq6	Yes	474449	1	5	14892	39	8
seq7	Yes	NA	NA	NA	109981	13	8
seq8	Yes	NA	NA	NA	NA	NA	NA
seq9	No	2	89	25	1	66644	32
seq10	No	2	89	25	2	59338	12
seq11	No	47	8	11	27	4259	18
seq12	No	59	6	45	36	3103	66
seq13	No	10134	2	16	90	1121	39
seq14	No	474449	1	36	239	405	40
seq15	No	NA	NA	NA	1463	89	55
seq16	No	NA	NA	NA	10173	45	40
seq17	No	NA	NA	NA	23905	32	10
seq18	No	NA	NA	NA	54930	21	40
seq19	No	NA	NA	NA	120767	12	39

3.10　IL10RA を標的分子とした HT-SELEX データによる性能評価

FSBC の並列処理を実装する pFSBC を，IL10RA を標的分子とした HT-SELEX データ [34,35] に適用し，並列処理による計算速度とクラスタリングの精度の評価を行った．IL10RA は細胞膜タンパク質であり IL10 のレセプターである．IL10 は抗炎症サイトカインであり，自己免疫疾患などの炎症をともなう疾患の創薬ターゲットである．IL10RA の HT-SELEX データは NCBI BioProject に登録されており，プロジェクトの登録番号は PRJNA315881 である．このプロジェクトの登録番号には二つの HT-SELEX のデータが

登録されており, それぞれ HT-SELEX データの登録番号は SRR3279660 と SRR3279661 である．SRR3279660 には IL10RA の HT-SELEX の第 1 ラウンドから第 4 ラウンドの HT-SELEX データが格納され，SRR3279661 には第 5 ラウンドの HT-SELEX データが格納されている．

Allnutt らは IL10RA の HT-SELEX データを用いて，複数のクラスタリング手法の評価を実施した [23]．本節では，すでに実施された Allnutt らの評価方法に倣い，pFSBC の評価実験を行った．Allnutt らの文献では，既存のクラスタリング手法の比較結果がすでに報告されているが，本論文では再度実行した．その理由は，同じ環境下における速度比較を行うため，また報告されている配列の数がもとのデータから再現できなかったためである．クラスタリングの精度の評価を平等に行うためにも，同じ配列で再実験を行った．実験における実行環境は，CPU が Intel(R) Xeon(R) CPU E5-1650 v4@3.60GHz, メモリが 64GB, オペレーティングシステムは Ubuntu 16.04 (Xenial Xerus) である．

3.10.1　実験方法

計算速度

pFSBC の評価には，IL10RA を標的分子とした HT-SELEX データの第 5 ラウンドの HT-SELEX データを用いた．配列数は 3,946,124 であり，重複配列を除く配列数は 607,198 であった．第 3.9 節の実験において，AptaCluster が最も高速であったため，今回の実験においては AptaCluster を比較対象とした．AptaCluster の実装には規定のオプションを利用した．pFSBC（非並列処理）のオプションには，$l_{min} = \{4, 5, 6\}$ と $l_{max} = \{10\}$ を用いた．pFSBC（32 並列処理）では，$l_{min} = \{4, 5, 6\}$ と $l_{max} = \{10\}$ のオプションの組み合わせと，$l_{min} = 5$, $l_{max} = 25$ のオプションに関しても計算速度の評価をした．これは，長い ORS の探索に要する計算時間を評価するためである．25 の長さの文字列を網羅的に探索する場合は, 4^{25} の組み合わせがあり, その数は 1.1258999×10^{15} である．仮に，一つの文字列に関して一秒で計算するとしたら 35702051.6235 年かかる．もし，$l_{min} = 5$ から $l_{max} = 25$ の文字列の組み合わせを探索する場合，総数にすると $\sum_{i=5}^{25} 4^i$ の探索領域から文字列を探すことになる．

クラスタリングの精度

pFSBC のクラスタリングの精度の評価には，計算速度の評価と同様に，IL10RA を標的分子とした HT-SELEX の第 5 ラウンドの HT-SELEX データを用いた．IL10RA の

HT-SELEX データの中で評価された配列の数は 33 であり，その中で結合を示す配列（アプタマー）は 19 で，標的分子と結合を示さない配列は 14 である．結合の可否判断には解離定数を用い，Allnutt らと同様の条件である解離定数 (K_d) が 100 nM[*3]未満のものを標的分子と結合するアプタマーとして定義した．解離定数とはアプタマーと標的分子との親和性を表すものであり，低いほど強く結合していることになる．解離定数はアプタマーの濃度と標的分子の濃度を掛けたものを，複合体（アプタマーと標的分子が結合したもの）の濃度で割った値で計算されるため，複合体の濃度が濃ければ濃いほど解離定数は 0 に近づくことになる．

Allnutt らはクラスタリング手法を評価するために 3 つの指標を定義した．それらの指標は，クラスタの順位と結合力の順位とのスピアマンの順位相関 r_s，クラスタの順位と解離定数 (K_d) のピアソンの相関 r，そして Top 10 correct である．Top 10 correct とは，アプタマーが上位 10 のクラスタに含まれる数を示す．ただし，Top 10 correct はクラスタリングの感度のみを指標としているため，上位のクラスタの擬陽性を無視した指標となっている．たとえば，全配列を一つのクラスタにしてしまえば，第 1 のクラスタにすべてのアプタマーが含まれてしまい，Top 10 correct が最も良い結果となってしまうという問題点がある．そこで，本節で行う実験ではさらに，Top 10 incorrect と Positive predictive value (PPV) を別途設けて，クラスタリングの評価を行った．Top 10 incorrect とは，上位 10 のクラスタに結合しない配列が含まれる数であり，PPV は上位 10 のクラスタに含まれた配列の中で結合した配列の割合を示す．式で記載すると，

$$\mathrm{PPV} = \frac{\text{Top 10 correct}}{\text{Top 10 correct} + \text{Top 10 incorrect}} \tag{3.10}$$

となる．精度評価においては，PPV を優先順位の高い評価指標として考える．

pFSBC の比較対象として，Allnutt の文献で用いられている FASTAptamer, UCLUST, UNOISE, AptaCluster を適用した．AptaTRACE は複数ラウンドの HT-SELEX データを用いるため，本実験では除外した．FASTAptamer のオプションでは，同一クラスタに含める LD の閾値を 7 とし，頻度のフィルタリングオプションとして 10, 100 を用いた．UCLUST では，類似性の閾値を 97% および 90% に設定した．UNOISE3, AptaCluster においては既定の設定を用いた．pFSBC では，$l_{min} = \{4, 5, 6\}$, $l_{max} = \{10\}$ と $l_{min} = 5$, $l_{max} = 25$ のオプションを用いた．

[*3] モル濃度を表す単位．1 nM は 1 リットル中に分子が 10^{-9} モル含まれる濃度を示す．

クラスタの分布の比較

pFSBC($l_{min} = 5$, $l_{max} = 10$) で探索された ORS を用いて，非冗長な HT-SELEX データを多様性を表す行列に変換した．多様性を表す行列は，ORS を含むか含まないかという二値の情報と，ORS の Z^* スコアを用いて作成した．同じ特徴ベクトルを持つ配列は除去し ($B \to B'$)，多様性を表す行列 B' のランクを最大とした (rank($B' = |\mathcal{S}|$)．多様性を表す行列に UMPA を用いて次元の圧縮を行った．UMAP の実装には uwot ライブラリ [101] を利用し，パラメータは既定とした．UMAP による次元圧縮後のデータを可視化し，FSBC が作成したクラスタの分布と結合評価を行った配列の分布を，HT-SELEX データ全体の分布と比較する．分布の比較により，結合評価された配列が大域的に分布するのか，それとも局所的に分布するのか評価する．また，結合評価された配列の中でアプタマーと標的分子と結合しない核酸分子の位置関係に関しても評価する．

3.10.2　結果

計算速度

表 3.7 に，AptaCluster，FSBC，pFSBC の計算時間を示す．1 列目に手法の名称を記載し，2 列目に手法に対するオプションを記載し，3 列目にクラスタリングに要する時間を示す．R で記載されたオリジナルの FSBC と，並列処理を行わない pFSBC(1 core) を比較すると，コンピュータ言語を変更するだけで 7 倍以上の高速化ができた．さらに，pFSBC(1 core) と pFSBC(32 core) を比較すると，$l_{min} = 6$, $l_{max} = 10$ のオプションにおいておよそ 7 倍の計算速度の向上があった．このように，FSBC に並列処理が効果的であることが，実際の HT-SELEX データで示された．また，$l_{min} = 5$, $l_{max} = 25$ として，25 塩基の長さの ORS を探索した場合でも，1 分 32 秒と非常に高速な時間で ORS の探索が完了している．この，$l_{min} = 5$, $l_{max} = 25$ の計算時間は，$l_{min} = 6$, $l_{max} = 10$ よりも短い．FSBC のアルゴリズムでは，最初に l_{min} の文字列に関してはすべて列挙して評価を行う．そのため，アルゴリズムの特性上 l_{min} の値に大きく計算時間が依存する．このことが，$l_{min} = 5$, $l_{max} = 25$ より $l_{min} = 6$, $l_{max} = 10$ の方が，計算時間が長い理由である．逆に，非常に l_{max} が長いとしても l_{min} を適切な長さに抑えれば計算時間はそれほどかからない．AptaCluster と pFSBC(32 core) を比較したところ，すべてのオプションにおいて pFSBC の計算速度が速かった．

表 3.7　AptaCluster，FSBC，pFSBC のクラスタリングに要する計算時間

Method	Option	Processing time
AptaCluster	Default options	4 min 18 sec
Original FSBC (R)	$l_{min} = 4$ $l_{max} = 10$	30 min 9 sec
	$l_{min} = 5$ $l_{max} = 10$	54 min 27sec
	$l_{min} = 6$ $l_{max} = 10$	159 min 2sec
pFSBC (1 core, Python)	$l_{min} = 4$ $l_{max} = 10$	4 min 8 sec
	$l_{min} = 5$ $l_{max} = 10$	8 min 1 sec
	$l_{min} = 6$ $l_{max} = 10$	20 min 52 sec
pFSBC (32 cores, Python)	$l_{min} = 4$ $l_{max} = 10$	38 sec
	$l_{min} = 5$ $l_{max} = 10$	1 min 3 sec
	$l_{min} = 6$ $l_{max} = 10$	2 min 55 sec
	$l_{min} = 5$ $l_{max} = 25$	1 min 32 sec

クラスタリングの精度

表 3.8 に，結合評価された配列に対する各手法によるクラスタリングの結果を示す．各列は，配列の ID，解離定数 (K_d)，解離定数による順位と各クラスタリング手法によって作成されたクラスタの順位を示す．解離定数の単位は nano molar (nM) である．表の下部に記載されている 6 行は，解離定数の順位とクラスタの順位とのスピアマンの順位相関係数 (r_s)，解離定数とクラスタの順位のピアソンの相関係数 (r)，Top 10 correct，Top 10 incorrect，PPV，そして解析に要した時間を示す．

スピアマンの順位相関係数，ピアソンの相関係数，Top 10 correct，Top 10 incorrect，PPV において，最も良い結果を示した手法は，pFSBC ($l_{min} = 6, l_{max} = 10$)，UCLUST

(97%)，pFSBC ($l_{min} = 4$, $l_{max} = 10$)，pFSBC (($l_{min} = 5$, $l_{max} = 10$), ($l_{min} = 6$, $l_{max} = 10$), ($l_{min} = 5$, $l_{max} = 25$))，pFSBC (($l_{min} = 5$, $l_{max} = 10$), ($l_{min} = 5$, $l_{max} = 25$)) であった．つまり，ピアソンの相関係数以外では，pFSBC はほかの従来手法と比較して，最も精度の高い結果を示した．今回の実験で最も着目している指標である PPV においては，pFSBC が最も良い値 (PPV = 0.83) であり，そのときのオプションは ($l_{min} = 5$, $l_{max} = 10$) と ($l_{min} = 5$, $l_{max} = 25$) であった．また，pFSBC のほかのオプションである ($l_{min} = 4$, $l_{max} = 10$) と ($l_{min} = 6$, $l_{max} = 10$) においても，PPV が 0.67 と他の手法と同程度の結果を示した．pFSBC はオプション ($l_{min} = 5$, $l_{max} = 10$) と ($l_{min} = 5$, $l_{max} = 25$) で，ピアソンの相関係数以外すべて同じ結果であった．そのため, pFSBC のクラスタリングの精度は l_{min} の影響が強いと考えらえる．計算時間に関しては, pFSBC ($l_{min} = 3$, $l_{max} = 10$) が最も速い計算速度を示した．pFSBC ($l_{min} = 6$, $l_{max} = 10$) の計算時間が FASTAptamer(f≥100) より長いが，これは FASTAptamer が頻度によるフィルタリングで HT-SELEX データのサイズを小さくしている影響である．そのため，HT-SELEX データの全配列を用いた解析では，pFSBC が最も高速である．以上より，pFSBC が従来手法と比較して高精度であり，また高速に解析可能であることが示された．Allnutt の報告では，複数ラウンドを用いた解析も実施しており UNOISE3 の PPV が 0.875 と最も良い結果である．pFSBC (($l_{min} = 5$, $l_{max} = 10$), ($l_{min} = 5$, l_{max}25)) は，PPV が 0.83 と，複数ラウンドを用いた UNOISE3 の PPV とほぼ同等の結果であった．このように，pFSBC は単一ラウンドの HT-SELEX データで，他の手法で複数ラウンドを用いた結果とほぼ同等の結果を示した．

表 3.8 で最も注目すべき点は，pFSBC のみが上位 10 のクラスタに最も結合力の高い配列 L462 を含むことである．さらに，pFSBC (($l_{min} = 4$, $l_{max} = 10$), ($l_{min} = 5$, $l_{max} = 10$)) では，最も結合力の高い配列が 1 番目のクラスタに属している．既存の手法では，上位 10 に最も強く結合する配列を含んでいないため，この配列を結合評価するための候補配列から取りこぼしてしまう．以上より，従来の手法と比較して，pFSBC は強く結合するアプタマーを効率よく選択するために有用である．

図 3.12 に，各手法のクラスタの順位と配列の頻度との関連を示す．FASTAptamer, UCLUST に関しては精度の高かったオプションのみを記載する．横軸はクラスタの順位を示し，縦軸に配列の頻度を示す．濃い青色の点は結合した配列を示し，薄い青色は結合しなかった配列を示す．縦の破線は 10 番目のクラスタを示す．点に付随する番号は解離定数による順序を示す．IL10RA の HT-SELEX データで最も結合の強い配列は L462 であり ($K_d = 2$)，頻度による順位が 12 番目である．一方，頻度の最も高い配列 H0 の

解離定数は $K_d = 25$ である．図に示されるように，pFSBC 以外のクラスタリング手法は，クラスタの順位と頻度との相関が非常に高い．そのため，頻度の高い配列を選択し，頻度はそれほど高くないが強く結合するアプタマーを取りこぼしてしまう恐れがある．一方，pFSBC は頻度の影響を大きく受けないため，オプション ($l_{min} = 5, l_{max} = 10$) のとき，第 1 のクラスタが最も強く結合するアプタマーである L462 を含んでいる．以上より，pFSBC は頻度による影響に対してほかの手法より頑健であり，配列の頻度が低くても強く結合するアプタマーの探索に有効である．ただし，標的分子と結合しないが頻度の高い配列 H1 も同時に第 1 のクラスタに属してしまっているため，さらなるクラスタリングの精度の向上が求められる．従来手法でも，H1 はすべて第 2 のクラスタに属しているため，現存するクラスタリング手法では H1 を上位のクラスタから排除することは困難であるといえる．

図 3.13 に，クラスタの順位と解離定数との関係を示す．横軸にクラスタの順位を示し，縦軸に解離定数 K_d を示す．図 3.12 と同様に，濃い青色の点は結合した配列を示し，薄い青色は結合しなかった配列を示す．縦の破線は 10 番目のクラスタを示す．点に付随する番号は配列の頻度を表す．いずれの手法もクラスタの順位と解離定数に対して強い相関がみられなかった．

クラスタの分布の比較

図 3.14 に，HT-SELEX データを二値の行列に変換し，多次元圧縮法である UMAP により二次元へと射影した図を示す．大きな点は結合評価された配列を示し，濃い青は標的分子に強く結合したアプタマー ($K_d < 100$ nM) を示し，薄い青は標的分子と結合しなかった核酸分子の配列を示す．結合評価された配列に付随するラベルは結合評価した配列の ID を示す．小さい点は結合評価がされなかった配列を示しており，色は FSBC により作成されたクラスタの順位を表している．結合評価されなかった配列の色は，赤くなるほど FSBC のクラスタの順位が高く，紫になるほどクラスタの順位が低いことを示す．この図は，FSBC によって求められたクラスタの分布と結合評価した配列の分布を示す．HT-SELEX のデータは大きく 4 つのグループに分布しており（右，上，右下，左下），FSBC から計算された上位のクラスタが，右のグループに多く属していることがわかる．配列 H0 が右のグループに近い結合評価された配列である．右にあるグループ以外は，高い順位から低い順位のクラスタが混在している．特に，左下のグループには幅広い順位のクラスタが混在していることがわかる．結合評価を行った配列は，HT-SELEX データの配列の多様性の中で，広い範囲にわたり選択されていることがわかる．つまり，結合評価

した配列は局所的に分布しておらず，多様なアプタマーの候補が選択されていると考えられる．配列 L462，H3，H7 は，結合評価されていない配列が作るクラスタから離れており，他の配列と比較して個性的な配列であると考えられる．

図 3.15 に，図 3.14 から結合評価されてない配列を除いた図を示す．結合評価された配列のいくつかは点が重なっており，類似した配列であることがわかる．また，異なる色の点の重なりもあり，アプタマーと標的分子と結合しない配列の分類があまりうまく行われていないことがわかる．アプタマー（濃い青）と標的分子と結合しない配列（薄い青）が全体にまんべんなく配置されており，特に法則性は見当たらない．よって，HT-SELEX データを二値の行列に変換し UMAP で次元圧縮された結果は，FSBC により得られたクラスタの分布の比較，結合評価した配列とクラスタとの分布の比較を行うことはできるが，射影した空間でアプタマーと標的分子と結合しない配列を見分けるのは難しい．

図 3.16 に HT-SELEX データを Z^* スコアを用いて行列に変換し，多次元圧縮法である UMAP により二次元へと射影した図を示す．図 3.14 と同様に，大きい点は結合評価をした配列，小さい点は結合評価をしていない配列を表す．図 3.14 と同様に，結合評価された配列の色は結合の有無を示し，濃い青はアプタマー ($K_d < 100$ nM) を示し，薄い青は標的分子に結合しなかった配列を示す．結合評価された配列に付随するラベルは，結合評価した配列の ID を示す．小さい点は，結合評価がされなかった配列を示しており，色は FSBC により作成されたクラスタの順位を表している．赤くなるほど順位の高いクラスタを示し，紫になるほど低いクラスタの順位を示す．色のグラデーションが右側では赤いのに対して，左側では紫である．つまり，横軸が FSBC で推定したクラスタの順位と強い相関を示す．図 3.14 と比較すると，結合評価されていない配列が極端に密集していないことがあげられる．これは，二値の多様性を表す行列と比較して Z^* を用いた多様性を表す行列に，連続的な情報も加わったため，極端に配列が密集しなかったためと考えられる．図 3.14 と同様に，結合評価された配列は全体に分布しており，HT-SELEX データの中で局所的に評価されていないことがわかる．

図 3.16 から結合評価をしていない配列を取り除いた図を，図 3.17 に示す．図 3.15 と異なり，アプタマー（濃い青）と標的分子と結合しない配列（薄い青）が離れて配置されている．たとえば，グラフの上部には比較的アプタマー（濃い青）が配置されており，中央部分には標的分子と結合しない配列しかない．上部に配置されているアプタマーの中には，H33，H4 のような FSBC では下位のクラスタに分類されているものも含まれている．そのため，Z^* により多様性を表現した行列を，UMAP で二次元に射影した結果は，新たなクラスタリング手法としても活用できる可能性がある．

3.11　むすび

本章では，まず第 3.2 節で，FSBC の概要を説明した．第 3.3 節で，特定の文字列を含む配列の出現確率に関して述べた．第 3.4 節で結合領域を推定するために新たに文字列スコアを Z スコアとして定義した．第 3.5 節に，提案した Z スコアを比較しながら，分枝限定法により探索領域を削減し，高速に長い ORS を探索するための方法を述べた．第 3.6 節に，探索した ORS をもとに，HT-SELEX データをクラスタに分ける方法を示した．このとき，異なる長さの ORS を比較するために，長さで規格化した Z スコアである Z^* スコアを定義した．長さの異なる ORS の Z^* スコアを比較することで，結合領域の長さの推定が可能となった．第 3.7 に，ORS の探索を並列化することにより FSBC を高速化することを述べた．第 3.8 節に，配列の多様性を定義し，FSBC により得られたクラスタの分布を比較する方法を述べた．第 3.9 節で，hESC を標的分子とした HT-SELEX データを用いて FSBC の性能を評価した．FSBC は従来のクラスタリング手法と比較して，最も高いクラスタリング精度を示した．また，SELEX の第 3 ラウンドのような早期ラウンドでも，高いクラスタリングの精度を示した．第 3.10 節で，IL10RA を標的分子とした HT-SELEX データを用いて pFSBC の性能評価を行った．pFSBC は，hESC の HT-SELEX のときと同様に，従来手法と比較して最も良いクラスタリングの精度を示した．また，pFSBC は従来手法と比較して最も高速であった．HT-SELEX データの中で，FSBC が作成したクラスタと，結合評価した配列の分布を比較し，結合評価した配列が局所的に分布しておらず，多様な配列が評価されていることを確認した．クラスタリングの精度と速度の結果より，本章で提案した手法 FSBC (pFSBC) が HT-SELEX データに対するクラスタリング手法として有用であることを示した．

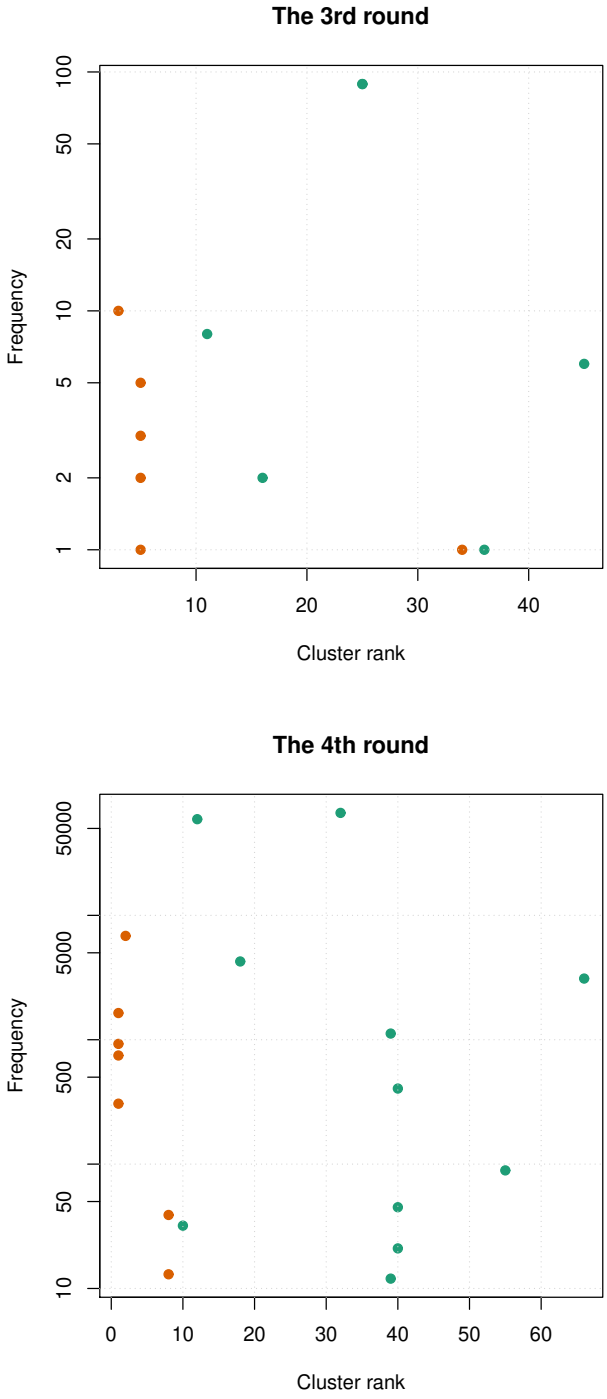

図 3.11　第 3 ラウンドと第 4 ラウンドにおける結合評価した配列が属するクラスタの順位と頻度との関係

表 3.8　各手法によるクラスタリングの結果

			FASTAptamer		**UCLUST**		**UNOISE3**	**AptaCluster**	**pFSBC**			
ID	**Kd**	**rank**	**f≥100**	**f≥10**	**90%**	**97%**			$l_{min}=4$ $l_{max}=10$	$l_{min}=5$ $l_{max}=10$	$l_{min}=6$ $l_{max}=10$	$l_{min}=5$ $l_{max}=25$
L462	2	1	12	12	12	16	12	13	1	1	3	3
L464	4	2	22	22	26	30	21	25	4	12	40	14
L455	4	3	20	20	24	29	20	23	20	29	137	31
L454	8	4	10	11	11	25	11	11	14	75	430	77
H33	10	5	39	39	51	70	38	48	37	75	84	77
L463	12	6	25	25	30	35	25	28	27	54	179	56
H4	18	7	5	5	5	8	5	5	15	144	156	146
H12	20	8	14	14	14	18	14	15	6	7	12	9
H22	20	8	26	27	31	42	26	29	6	7	12	9
H30	25	9	37	37	47	59	35	42	32	78	412	80
H0	25	9	1	1	1	1	1	1	1	1	1	1
L465	25	9	27	28	32	39	27	30	16	81	179	83
L418	35	10	13	13	13	17	13	14	9	148	486	150
L413	40	11	17	17	20	23	17	20	21	30	40	32
H6	50	12	7	7	7	10	7	7	8	100	140	102
H3	60	13	4	4	4	7	4	4	2	2	183	4
H2	65	14	3	3	3	4	3	3	7	11	92	13
H8	80	15	8	9	9	11	9	9	4	11	107	13
L420	80	15	21	21	33	34	24	24	13	78	93	80
H40	120	16	44	44	59	93	44	56	7	11	107	13
H1	120	16	2	2	15	2	2	2	1	1	3	3
L412	120	16	32	33	40	51	31	37	14	33	57	35
H14	123	17	16	16	19	21	16	19	15	81	121	83
H16	375	18	19	19	23	27	19	22	15	48	179	50
H7	375	18	9	8	8	12	8	8	4	11	17	13
H9	375	18	11	10	10	14	10	10	11	43	179	45
H20	375	18	24	23	27	31	22	26	35	81	179	83
L409	500	19	33	32	38	120	37	36	39	81	179	83
H26	500	19	36	36	43	57	34	40	28	81	179	83
L417	500	19	23	24	28	33	23	27	10	160	453	166
H5	500	19	6	6	6	9	6	6	2	74	179	76
H15	500	19	18	18	22	48	18	21	17	20	27	22
H24	500	19	29	29	34	44	28	32	11	52	171	54
r_s			0.11	0.08	0.11	0.14	0.12	0.08	0.09	0.22	<u>0.23</u>	0.22
r			0.12	0.14	0.12	<u>0.26</u>	0.19	0.13	0.18	0.20	0.13	0.12
Top 10 correct			7	6	6	5	6	6	<u>10</u>	5	2	5
Top 10 incorrect			3	4	3	2	4	4	5	<u>1</u>	<u>1</u>	<u>1</u>
PPV			0.70	0.60	0.67	0.71	0.60	0.60	0.67	<u>0.83</u>	0.67	<u>0.83</u>
Processing time (s)			131	3,772	1,674	2,230	539	258	<u>38</u>	63	175	152

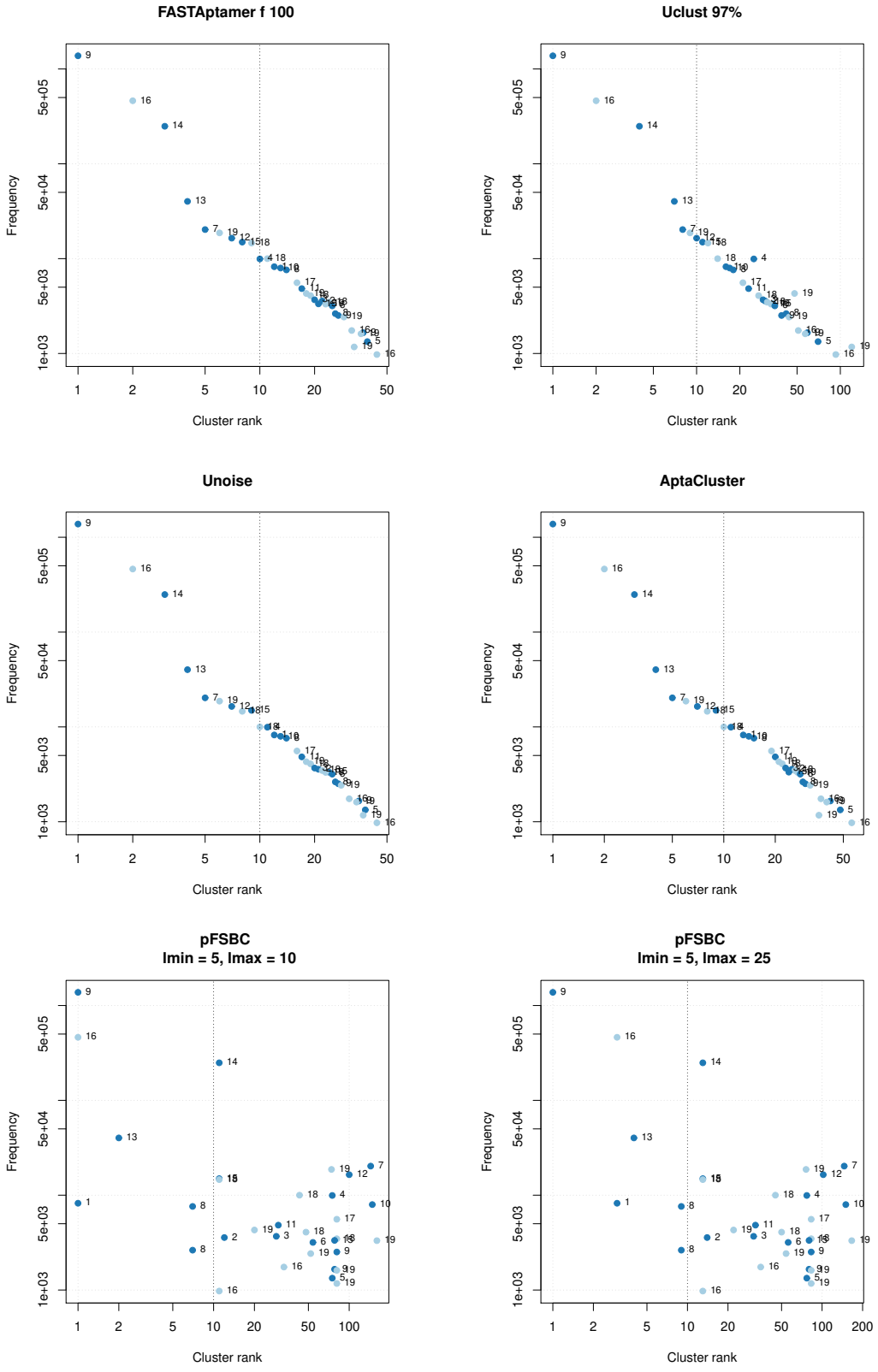

図 3.12　結合評価した配列が属するクラスタの順位と頻度との関係

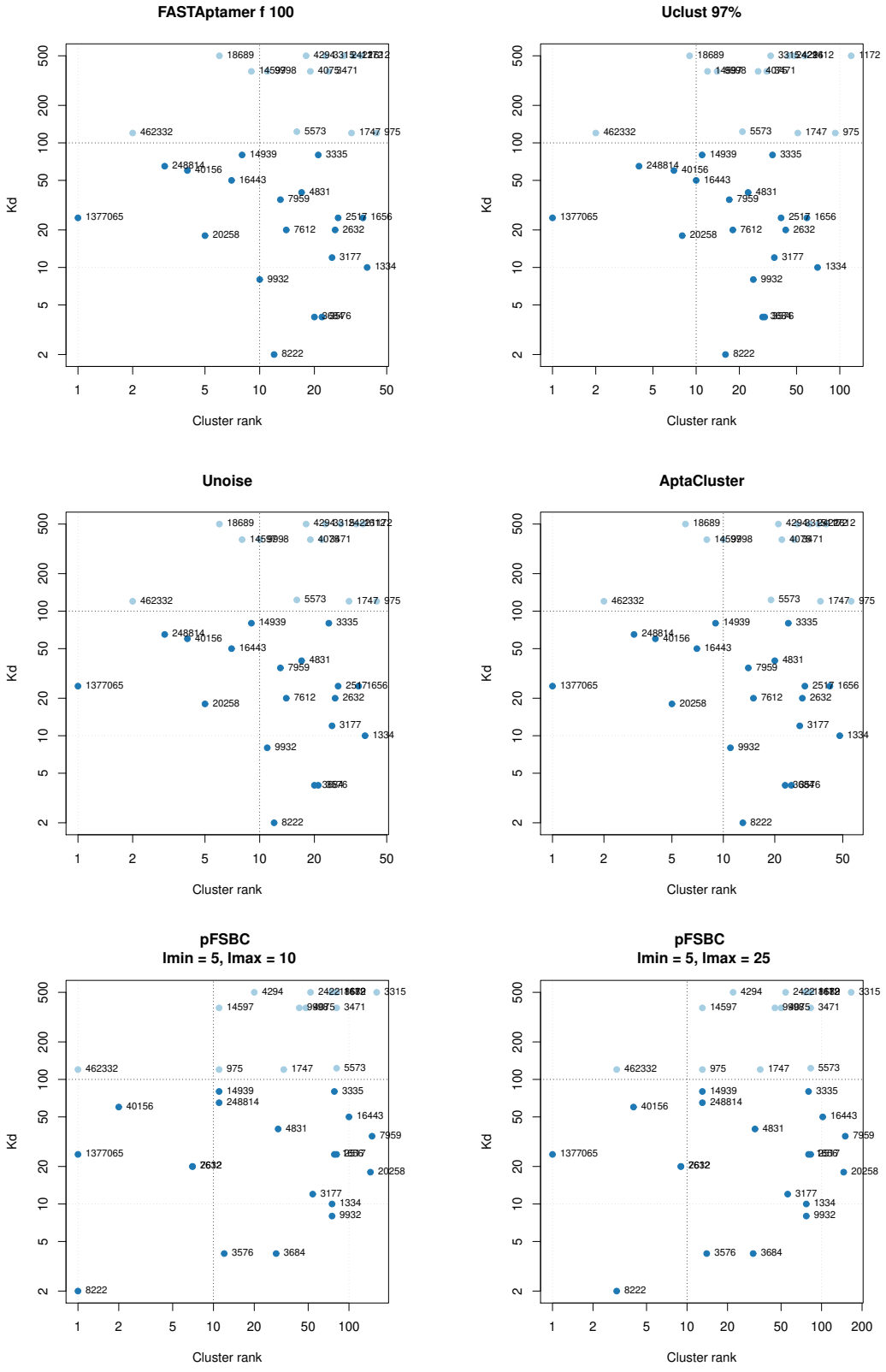

図 3.13　結合評価した配列が属するクラスタの順位と解離定数との関係

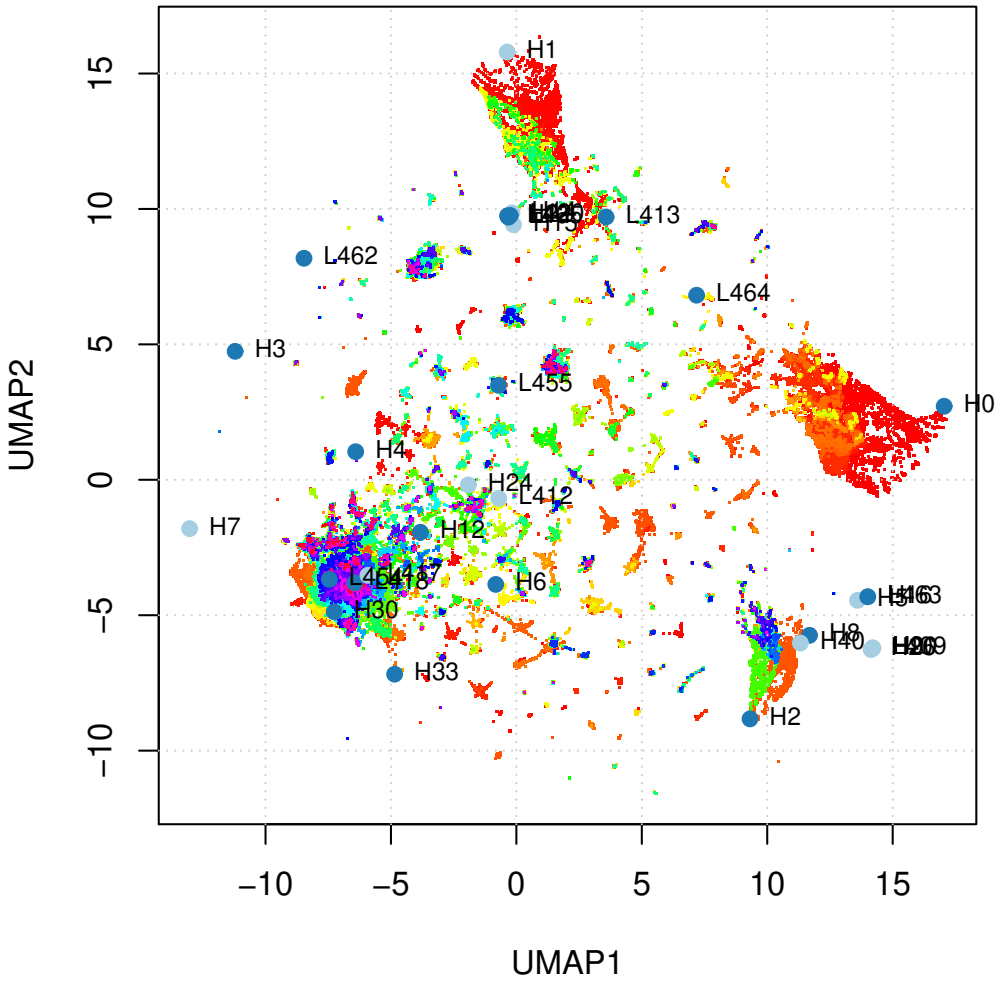

図 3.14　配列の多様性とクラスタとの比較（二値）

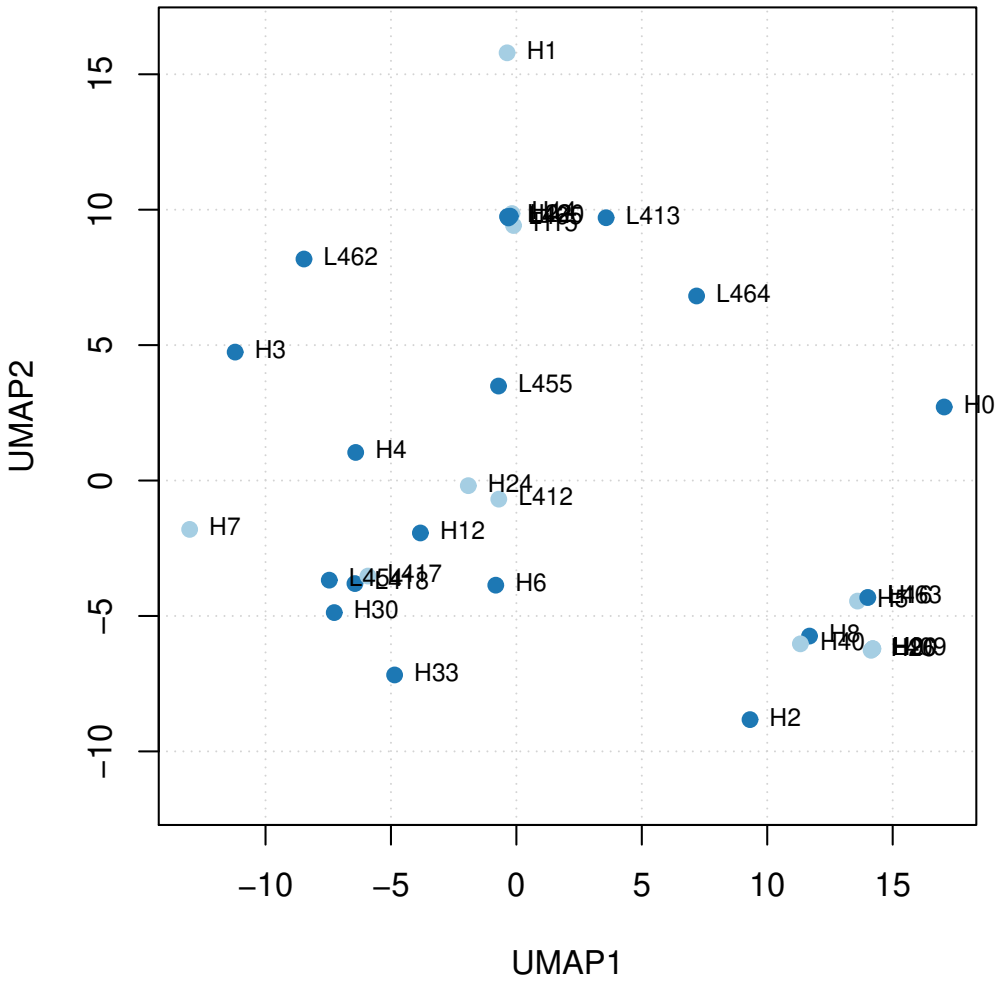

図 3.15　結合評価を行った配列の関係（二値）

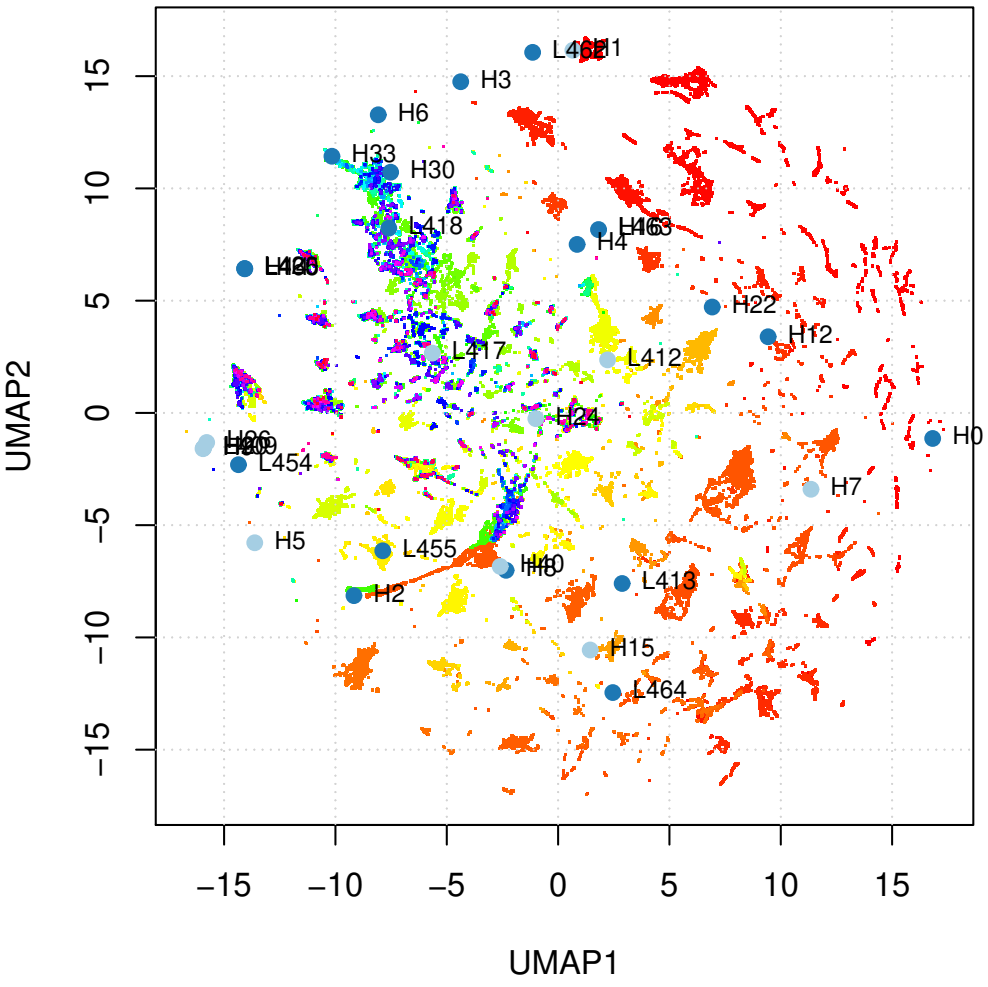

図 3.16　配列の多様性とクラスタとの比較（Z^* スコア）

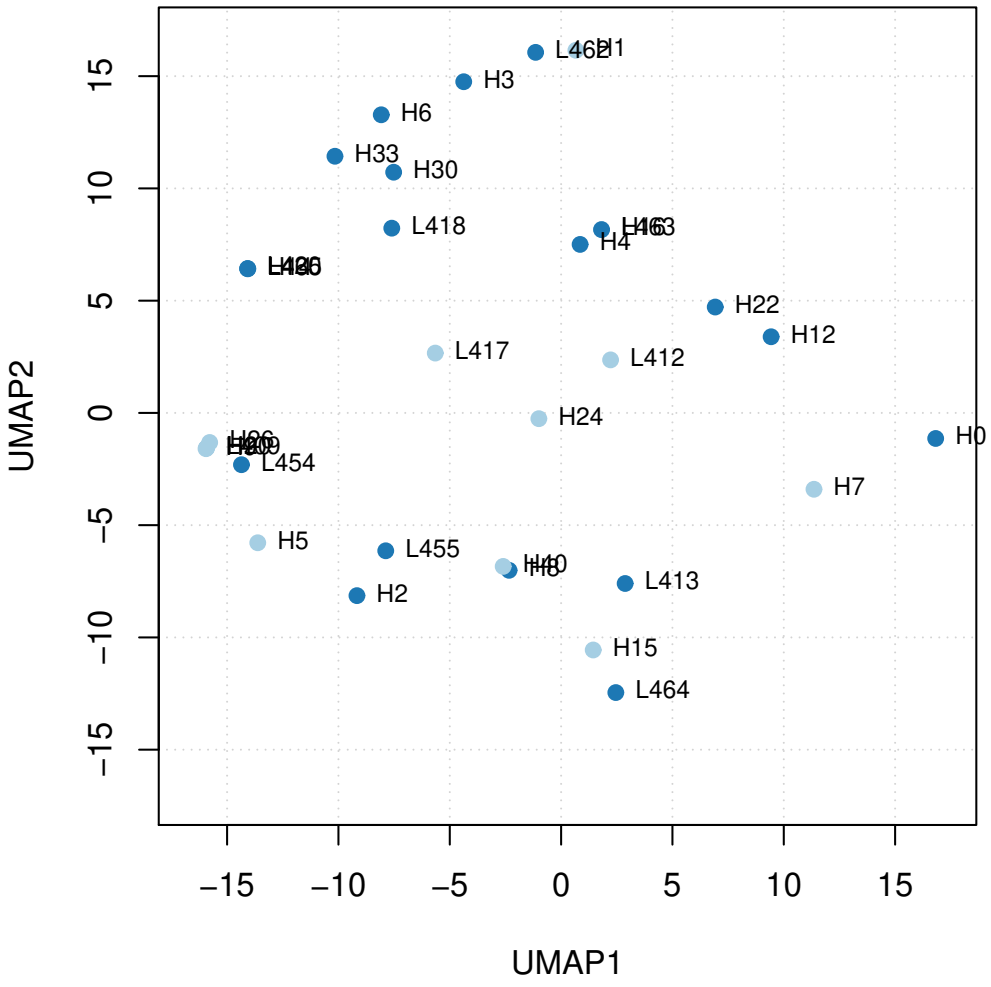

図 3.17　結合評価を行った配列の比較（Z^* スコア）

第 4 章

アプタマーの小型化

4.1 まえがき

アプタマーをバイオセンサーや医薬品などに利用するため，アプタマーの結合に不要な領域をできるだけ削除して，アプタマーの小型化を行う．アプタマーを小型化する理由として，化学合成のコストと合成エラーの削減，標的分子との結合能の向上，構造多様性の抑制などがあげられる．アプタマーは立体構造を形成して標的分子と結合するため，結合領域の構造が保持されたまま小型化される必要がある．現状，アプタマーの小型化配列を推定するバイオインフォマティクスの手法の報告はなく，専門家がアプタマーの二次構造予測の結果を見ながら，経験によりアプタマーの小型化配列を推定している．そのため，アプタマーの小型化配列の推定に明確な基準がなく，結果の再現性がないという問題がある．また，通常アプタマーの結合領域が不明瞭のまま，アプタマーの小型化配列が推定されており，結合領域を削除した小型化配列を推定してしまう恐れもある．さらに，人ではアプタマーの構造多様性による二次構造を網羅的に考慮することが困難である．以上の問題を踏まえて，アプタマーの構造多様性と結合領域の構造の保持を考慮したアプタマーの小型化配列を推定する手法を提案する [38]．また，提案する手法を実際のアプタマーに適用して，その有用性を示す．

本章では，第 4.2 節で，アプタマーを小型化配列を推定するための最適化問題を定式化する．第 4.3 節で，最適化問題の解を求めるアルゴリズムに関して述べる．第 4.4 節で，提案したアプタマーの小型化配列を推定する手法を，VEGF に結合するアプタマーと，CRP に結合するアプタマーに適用し，有用性を示す．最後に，第 4.5 節で，本章の結びを述べる．

4.2　最適化問題の定式化

SELEX より取得したアプタマーの全長配列を S とする．5' 末端と 3' 末端より，m 塩基と n 塩基短くした小型化配列を $S(m,n)$ とする．S および $S(m,n)$ の塩基の長さをそれぞれ $|S|$ と $|S(m,n)|$ とする．結合領域を B とし，その長さを $|B|$ とする．考慮する核酸分子の状態は二次構造のみとし，S の二次構造の集合を Ω_S とし，$S(m,n)$ の二次構造の集合を $\Omega_{S(m,n)}$ とする．そこで，配列 S の二次構造 $U \in \Omega_S$ が存在する確率 $\Pr(U), U \in \Omega_S$ を，ボルツマン分布と二次構造 U の自由エネルギー $E_U, U \in \Omega_S$ を用いて次式で定義する:

$$\Pr(U) = \frac{\exp(-E_U/kT)}{\sum_{j\in\Omega_S}\exp(-E_j/kT)}, \quad U \in \Omega_S. \tag{4.1}$$

ここで，k はボルツマン定数，T は絶対温度を示す．この確率の計算は，McCaskill による RNA の二次構造の存在確率の計算と同じである [102]．同様に，小型化配列 $S(m,n)$ に対する二次構造 $U(m,n) \in \Omega_{S(m,n)}$ が存在する確率を次式で定義する:

$$\Pr(U(m,n)) = \frac{\exp(-E_{U(m,n)}/kT)}{\sum_{j\in\Omega_{S(m,n)}}\exp(-E_j/kT)}, \quad U(m,n) \in \Omega_{S(m,n)}. \tag{4.2}$$

全長配列の二次構造の集合 Ω_S の中から，結合領域 B の部分の二次構造を抜き出し，重複を除いたものを結合領域 B の二次構造の集合 Ω_B とする．このとき，結合領域 B が全長配列 S の中でとり得る二次構造 $u \in \Omega_B$ の存在確率は，u を含む全長配列の二次構造 $U \in \Omega_S$ の存在確率の和で計算できる:

$$\Pr(u,S) = \bigcup_{j\in\mathcal{J}} \Pr(j), \quad \mathcal{J} = \{U \mid U \text{ includes } u \wedge U \in \Omega_S\}, \quad u \in \Omega_B. \tag{4.3}$$

小型化配列 $S(m,n)$ の中で，結合領域 B がとる二次構造 $u \in \Omega_B$ の存在確率は，u を含む小型化配列の二次構造 $U(m,n) \in \Omega_{S(m,n)}$ の存在確率の和で計算できる:

$$\Pr(u,S(m,n)) = \bigcup_{j\in\mathcal{J}} \Pr(j),$$
$$\mathcal{J} = \{U(m,n) \mid U(m,n) \text{ includes } u \wedge U(m,n) \in \Omega_{S(m,n)}\}, \quad u \in \Omega_B. \tag{4.4}$$

次に，小型化する前と小型化した後の結合領域 B の二次構造の変化を考える．全長配列に含まれる結合領域 B の二次構造 $u \in \Omega_B$ の存在確率と，小型化したときの存在確率

の比 (Probability ratio) は次式で計算される:

$$R(u, m, n, S) = \frac{\Pr(u, S(m, n))}{\Pr(u, S)}, \quad u \in \Omega_B. \tag{4.5}$$

アプタマーの配列を除去しすぎて，結合領域 B が結合時の二次構造を維持できないとき，小型化配列は標的分子と結合することができない．そこで，少なくとも存在確率の比 $R(u, m, n, S)$ が 0 より大きい必要がある．これを一般化し，存在確率の比 $R(u, m, n, S)$ が任意の閾値 θ より大きいとすると，その条件は次式となる:

$$R(u, m, n, S) = \frac{\Pr(u, S(m, n))}{\Pr(u, S)} > \theta, \quad u \in \Omega_B. \tag{4.6}$$

アプタマーの二次構造予測に用いる熱力学的パラメータは，天然の塩基を対象としたものしかなく，修飾塩基を対象としたものがない [72]．そのため，修飾塩基を用いたアプタマーの自由エネルギーや，自由エネルギーから導かれる存在確率の精度の保証がない．熱力学的パラメータは，異なるアプタマーの長さにおいて共通して利用されるため，修飾塩基に対するバイアス（偏り）は同じ向きになると考えられる．そのため，絶対的な数値である確率を利用するのではなく，存在確率の比を利用することにより，バイアスの影響を抑える．

図 4.1 に存在確率の比の計算例を図示する．図の 1 では，アプタマーの二次構造を求め，各二次構造に対する存在確率を計算する．表の各列は，二次構造の ID，存在確率，結合領域の二次構造を色で表わす．図中では，7 つの二次構造が推定され，結合領域では 3 つの二次構造が推定されている（黄色，黄緑色，水色）．図の 2 では，アプタマーの結合領域の各二次構造の存在確率を計算する．図中では，結合領域の構造は 3 種類あり，図の 1 で得られた存在確率から計算される．たとえば，結合領域が黄色い二次構造であれば，$0.1 + 0.1 + 0.2 = 0.4$ となる．図の 3 では，図の 1 と同様に小型化アプタマーの二次構造を求め，各二次構造の存在確率を計算する．図中では，5 つの二次構造が推定されたことになる．図の 4 では，図の 2 と同様に，アプタマーの結合領域の各二次構造の存在確率を計算する．図の 5 では，図の 2 と図の 4 で計算した結合領域の各二次構造の存在確率から，存在確率の比を計算する．図では，結合領域が水色の構造の場合，もとの長さの存在確率が 0.2 であり，小型化した際の存在確率が 0.8 である．そのため，存在確率の比は $0.8/0.2 = 4$ となる．

式 4.6 の条件のもと，アプタマーの結合領域の各二次構造 $u \in \Omega_B$ に対し，最も短くア

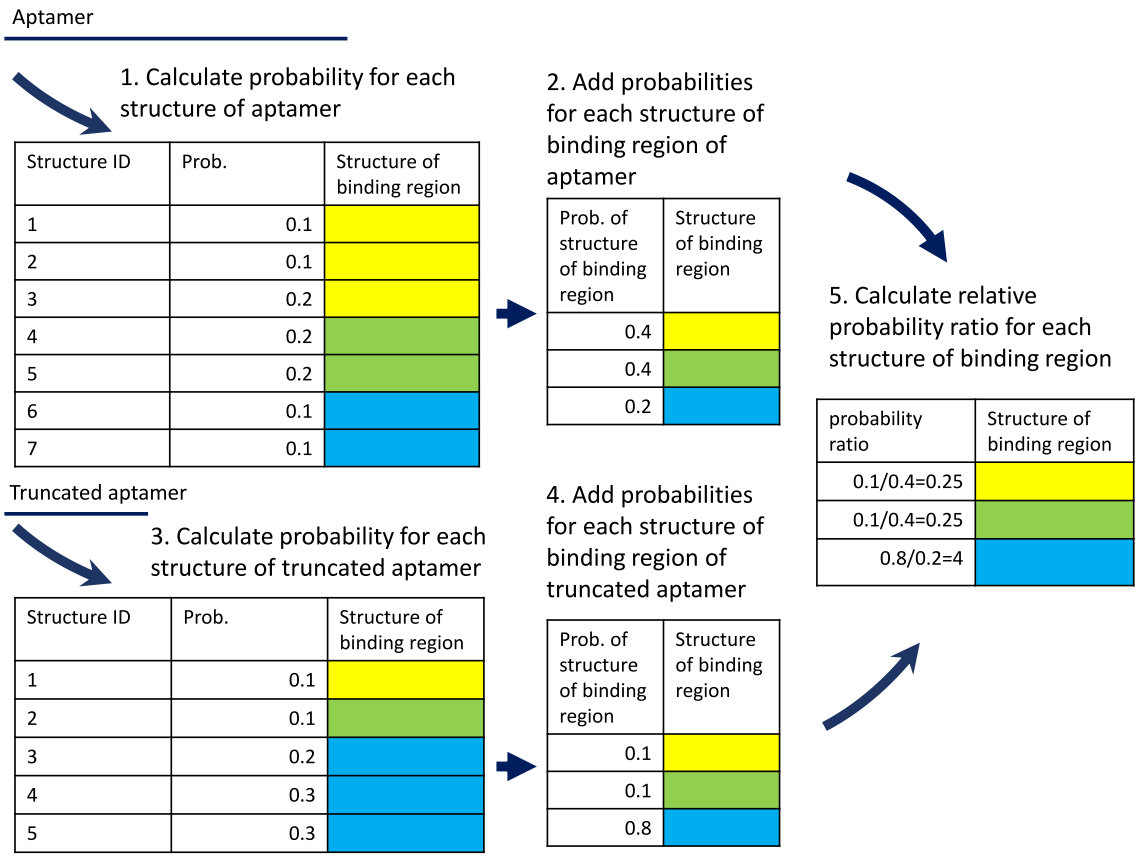

図 4.1　存在確率の比の計算例

プタマーを小型化することは，以下の最適化問題で記述できる:

$$\begin{aligned} \underset{m,n}{\arg\min} \quad & |S(m,n)| \\ \text{subject to} \quad & R(u,m,n,S) \geq \theta, \\ & R(u,m,n,S) = \frac{\Pr(u, S(m,n))}{\Pr(u,S)}. \end{aligned} \tag{4.7}$$

式 4.7 より，各二次構造 $u \in \Omega_B$ に対するアプタマーの小型化配列が推定できる．さらに，候補配列の数を限定するため，各二次構造 $u \in \Omega_B$ の中で最も短い配列を選択する．これは，$R(u,m,n,S), u \in \Omega_B$ の最大値が θ 以上であることを満たせばよいので次式で

表現される:

$$\begin{aligned} \underset{m,n}{\arg\min} \quad & |S(m,n)| \\ \text{subject to} \quad & \max_{u \in \Omega_B} R(u,m,n,S) \geq \theta, \\ & R(u,m,n,S) = \frac{\Pr(u, S(m,n))}{\Pr(u,S)}. \end{aligned} \tag{4.8}$$

式 4.8 では，結合領域 B の二次構造 $u \in \Omega_B$ から，最も小型化できる u を選択して，アプタマーを小型化する．

4.3　アプタマーの小型化配列を推定するアルゴリズム

前節で定義した式 4.8 を満たす m, n を求めるアルゴリズムを以下に示す．

1. S の二次構造 Ω_S を求める．
2. 二次構造 Ω_S の存在確率 $\Pr(U,S), U \in \Omega_S$ を求める．
3. S の結合領域 B の二次構造 Ω_B を求める．
4. 結合領域 B の二次構造 Ω_B の存在確率 $\Pr(u,S), u \in \Omega_B$ を求める．
5. $m \leftarrow 1$, $n \leftarrow 1$ とする．
6. S の 5' 末端から m 塩基削除，3' 末端から n 塩基削除し $S(m,n)$ を得る．
7. $S(m,n)$ の二次構造 $\Omega_{S(m,n)}$ を求める．
8. 二次構造 $\Omega_{S(m,n)}$ の存在確率 $\Pr(U, S(m,n)) \in \Omega_{S(m,n)}$ を求める．
9. $S(m,n)$ の結合領域 B の二次構造 Ω_B の存在確率 $\Pr(u, S(m,n)) \in \Omega_B$ を求める．
10. 存在確率の比 $R(u,m,n,S) \leftarrow \frac{\Pr(u,S(m,n))}{\Pr(u,S)}, u \in \Omega_B$ を計算する．
11. もし，5' 末端が結合領域でなければ $m \leftarrow m+1$ として，6 へ戻る．
12. もし，3' 末端が結合領域でなければ $n \leftarrow n+1$ として，6 へ戻る．
13. 存在確率の比 $R(u,m,n,S)$ が閾値 θ 以上の中で最も $S(m,n)$ が短い配列を，アプタマーの小型化配列として選択する．

最適化問題を満たす m と n を求める疑似コードを，アルゴリズム 3 に記載する．

Algorithm 3: Estimation of the shortest aptamer sequence

Data: S: The aptamer sequence; B: The estimated binding region; θ: A cutoff of probability ratio;

Result: $S[m..n]$: The estimated shortest aptamer;

$\Omega_S \leftarrow$ optimal and sub-optimal secondary structures of S;

$\Pr(U, S) \leftarrow$ probability of secondary structure $U \in \Omega_S$ of S;

$\Omega_B \leftarrow$ secondary structures of the binding region B;

$\Pr(u, S) \leftarrow$ probability of secondary structure $u \in \Omega_B$ of S;

for $u \in \Omega_B$ **do**

  for $m = 1$ **to** *the left position of* B **do**

    for $n = |S|$ **to** *the right position of* B **do**

      $\Omega_{S[m..n]} \leftarrow$ optimal and sub-optimal secondary structures of $S[m..n]$;

      $\Pr(U_{S[m..n]}, S[m..n])) \leftarrow$ probability of secondary structure $U_{S[m..n]} \in \Omega_{S[m..n]}$;

      $\Pr(u, S[m..n]) \leftarrow$ probability of secondary structure u of $S[m..n]$;

      $R(u, m, n, S) \leftarrow \frac{\Pr(u, S[m..n])}{\Pr(u, S)}$;

    end

  end

end

return The shortest sequence $S[m..n]$ with $R(u, m, n, S) > \theta$;

4.4 VEGF と CRP アプタマーを用いた性能評価

4.4.1 実験方法

VEGF に結合するアプタマーと，CRP に結合するアプタマーを用いて，提案したアプタマーの小型化配列を推定する手法の有用性を示す．VEGF は血管新生を促すタンパク質であり，心血管疾患 [103]，糖尿病 [104]，がん [105] などの疾患との関連が報告されている．また，加齢黄斑変性症のアプタマー医薬品である Macugen の標的分子も VEGF である [10]．CRP は炎症と関連のあるバイオマーカーであり，

たとえば関節リウマチの検査項目として利用されている．CRP は感染症との関連も示されており，COVID-19 の予後との関連も報告されている [106]．VEGF アプタマーの配列は，GGATTGCCCGATTCACCGTAAACAGTTCTGGTTCCTTAGTTTAAAGTCACGTCTTAGTTTAAGGCATTCTGGAGCGGCATAAC であり長さは 83 塩基である．HT-SELEX データから推定された VEGF アプタマーの標的分子との推定結合領域は，AGTTTAAAGTCACGTCTTAG であり，20 塩基の長さである．VEGF アプタマーは天然型の塩基を利用した DNA アプタマーである．CRP アプタマーの配列は，GGTTACGCCGCACATCAGTTTAGTGATAAAAGCCCGGTTACAGATGATCAGGGGCATTCGACAGGCTGGACATATC であり長さは 76 塩基である．HT-SELEX データから推定された CRP アプタマーの標的分子との推定結合領域は，CGGTTACAGATGATCA であり，長さは 16 塩基である．CRP アプタマーはチミンを修飾塩基 [45] で置き換えたアプタマーである．アプタマーと VEGF および CRP との結合領域の推定には，各 HT-SELEX データと FSBC ($l_{\min} = 5$，$l_{\max} = 25$) を用いた．VEGF アプタマーと CRP アプタマーが含む ORS の中で，最も Z^* スコアの高い ORS を推定結合領域とした．VEGF と CRP の推定された結合領域の長さは，それぞれ 20 と 16 であるが，AptaTRACE でこの長さの結合領域を推定することはできない．

VEGF アプタマーと CRP アプタマーの小型化配列を，提案手法により推定した．アプタマーとアプタマーの小型化配列の二次構造予測には，RNAsubopt を用いた [68]．VEGF アプタマーと CRP アプタマーは DNA アプタマーであるため，二次構造予測には DNA の熱力学的パラメータを適用した [72, 107]．アプタマーの小型化配列を推定する手法は Python で実装した．これは，RNAsubopt を含む Vienna RNA library [68] に Python のインターフェースが用意されているためである．アプタマーの小型化において，推定結合領域の二次構造の存在確率の比に対する閾値は $\theta = 1$ と設定した．これは，アプタマーを小型化したとき，推定結合領域の構造の存在確率が，もと長さのアプタマーの推定結合領域の構造の存在確率より高いことが条件であることを意味する．また，RNAsubopt では，最安定構造と準安定構造のみを求め，それ以外の構造の存在確率は無視できるとした．

提案手法により推定したアプタマーの小型化配列を化学合成し，標的分子との結合評価を行った．CRP の小型化アプタマーは修飾塩基を導入しているため，合成の都合上によりプライマー領域が必要となる．そのため，推定された小型化配列の最も 5’ 側にある修飾塩基であるチミンに，天然型の塩基より成るプライマー領域を 15 塩基分加えた配列を評価配列とした．つまり，もとの推定された小型化配列より 10 塩基長い配列を評価配列

としている．また，推定した小型化アプタマーと，さらに短い結合しないことが予測される配列も結合評価し，推定した小型化アプタマーが，真に最も短いアプタマーであるか検証した．標的分子との結合評価には，表面プラズモン共鳴 (Surface plasmon resonance: SPR) 分析[*1]を用いた．

4.4.2　結果

VEGF と CRP の HT-SELEX データに FSBC を適用して，VEGF と CRP アプタマーの結合領域を推定した．図 4.2 に，ORS の Z^* スコアと ORS を含む配列の総頻度との関係（上図）と，配列が属するクラスタの順位と頻度との関連（下図）を示す．上図において，濃い青の点は VEGF アプタマーに含まれる ORS であり，薄い青の点はそれ以外の ORS である．下図において，濃い青の点は VEGF アプタマーを示し，薄い青の点はそれ以外の配列を示す．同じ内容の図を CRP アプタマーに関しても作図し，図 4.3 に示す．図 4.2 より，VEGF アプタマーの ORS より高い Z^* スコアの ORS が，他の配列で数多く存在することがわかる．VEGF アプタマーは最も頻度が高いアプタマーであるが，最も順位の高いクラスタには属していない．そこで，VEGF アプタマーより順位の高いクラスタに属する他の配列も，VEGF に結合することが考えられる．図 4.3 より，CRP アプタマーの ORS は他の ORS と比較して高い Z^* スコアを示した．CRP アプタマーは，1 番目のクラスタに属している．

VEGF と CRP の HT-SELEX データを FSBC により解析したことで，VEGF と CRP アプタマーの推定結合領域の結果を得た．表 4.1 に，アプタマーの全長配列と推定結合領域との位置関係を示す．表におけるハイフンは，塩基が存在しないことを示す．表における各列は，配列の名前，配列の長さ，配列を示す．

[*1] センサーチップ上の金や銀などの金属に光を照射すると，金属の電子の振動による表面プラズモン波が発生し反射光の一部分が吸光される．アプタマーが標的分子と結合すると，分子量が増加し表面プラズモン波による吸光される光の角度が変化する．この吸光される光の角度の変化を検出し，アプタマーと標的の結合を評価する方法が SPR 分析である．

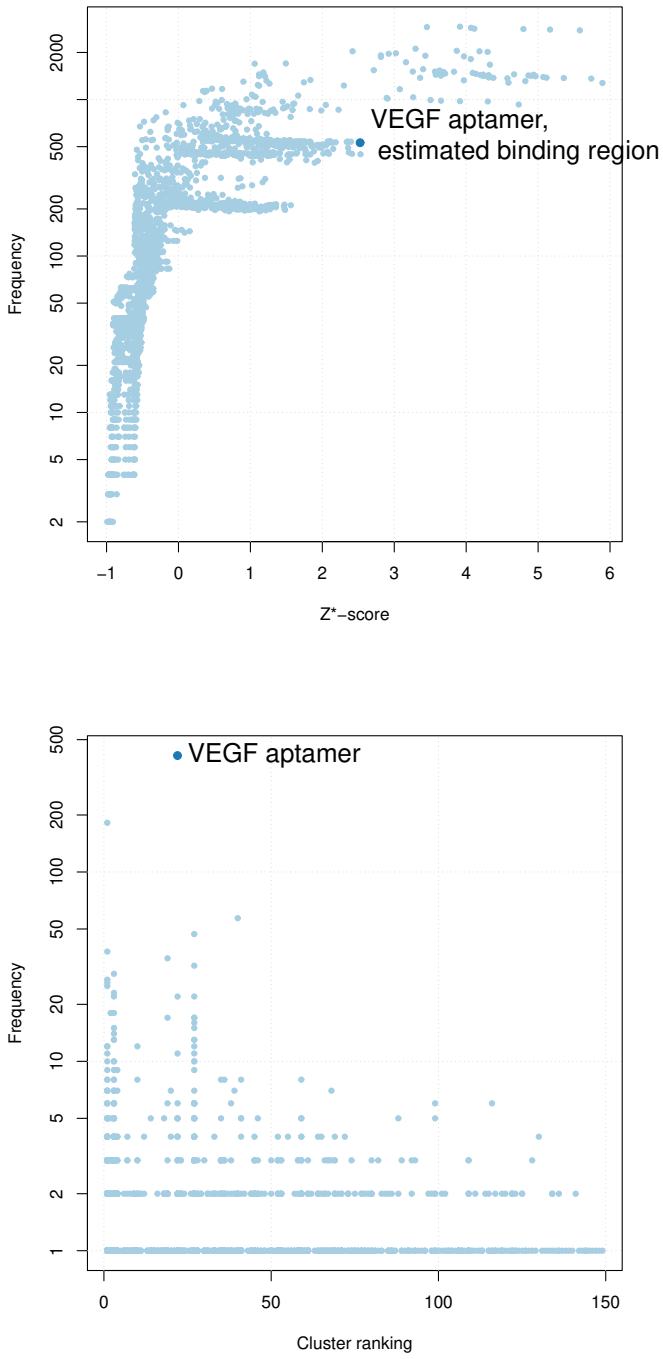

図 4.2　VEGF を標的分子とした HT-SELEX データの解析結果; 上図: ORS の Z^* スコアと ORS を含む配列の総頻度との関係, 下図: 配列が属するクラスタの順位と頻度の関係

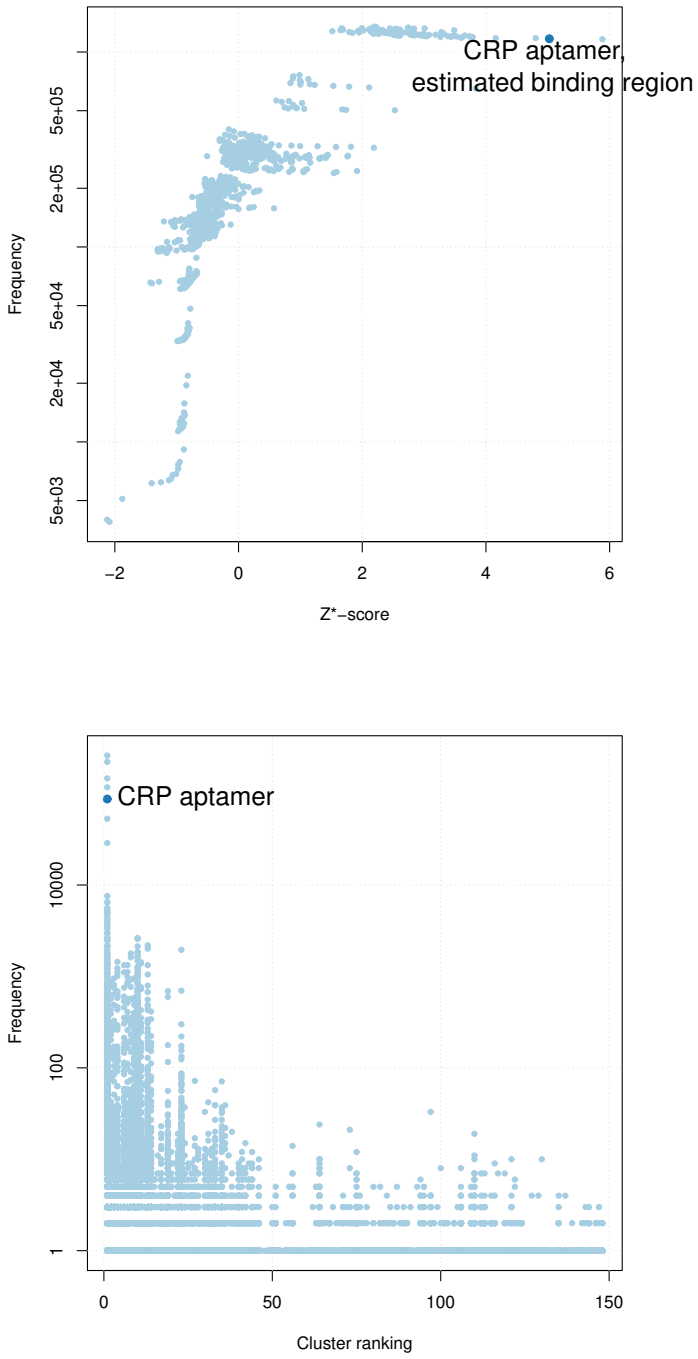

図 4.3　CRP を標的分子とした HT-SELEX データの解析結果; 上図: ORS の Z^* スコアと ORS を含む配列の総頻度との関係, 下図: 配列が属するクラスタの順位と頻度の関係

表 4.1　VEGF と CRP アプタマーとその推定結合領域

Target Name	Length	Sequence
VEGF	83	GGATTGCCCGATTCACCGTAAACAGTTCTGGTTCCTTAGTTTAAAGTCACGTCTTAGTTTAAGGCATTCTGGAGCGGCATAAC
VEGF EBR	20	----------AGTTTAAAGTCACGTCTTAG----------
CRP	76	GGTTACGCCGCACATCAGTTTAGTGATAAAAGCCCGGTTACAGATGATCAGGGGCATTCGACAGGCTGGACATATC
CRP EBR	16	----------CGGTTACAGATGATCA----------

Abbreviation: EBR, estimated binding region.

VEGF アプタマーと CRP アプタマーの全長配列と結合領域の二次構造の多様性を，表 4.2 と表 4.3 に示す．1 列目に，アプタマーの全長配列における最安定構造と準安定構造を示す（最も自由エネルギーが低いものが最安定構造である）．2 列目に，DNA パラメータより得られた自由エネルギー（単位は kcal/mol）を示す．3 列目に，存在確率を示す．4 列目に，推定結合領域の二次構造を示す．VEGF の二次構造は，上の 8 つの二次構造と下の 4 つの二次構造と大きく二つの二次構造のグループがあることがわかる．上の 8 つの二次構造の方が安定しているため，存在確率は下の 4 つの二次構造より高い．CRP の二次構造も中央部分に長いヘアピンループ構造を持つものと，そうでないものの二つのグループに分かれる．中央部分にヘアピンループ構造を持つ二次構造の方が安定しており存在確率も高い．また，推定された二次構造の数が VEGF と比較して CRP の方が多く，CRP のほうが構造多様性が高いことがわかる．

表 4.2　VEGF アプタマーの二次構造とその存在確率

GGATTGCCCGATTCACCGTAAACAGTTCTGGTTCCTTAGTTTAAAGTCACGTCTTAGTTTAAGGCATTCTGGAGCGGCATAAC	Free energy	Probability	AGTTTAAAGTCACGTCTTAG
....((.(((.((((((.(((((...........(((......)))..........))))).)).....)))).)))))....	-7.39	0.052	......)))..........)
....((.(((.((((((.(((((.......((..(((......)))..))......))))).)).....)))).)))))....	-8.06	0.160	......)))..))......)
....(((.((.((((((.(((((...........(((......)))..........))))).)).....)))).)))))....	-7.39	0.052	......)))..........)
....(((.((.((((((.(((((.......((..(((......)))..))......))))).)).....)))).)))))....	-8.06	0.160	......)))..))......)
....((((.(.((((((.(((((...........(((......)))..........))))).)).....)))).)))))....	-7.39	0.052	......)))..........)
....((((.(.((((((.(((((.......((..(((......)))..))......))))).)).....)))).)))))....	-8.06	0.160	......)))..))......)
.......(((.((((((.(((((...........(((......)))..........))))).)).....)))).)))......	-7.14	0.034	......)))..........)
.......(((.((((((.(((((.......((..(((......)))..))......))))).)).....)))).)))......	-7.81	0.105	......)))..))......)
((.....))......(((....(((...((.(........(((((.(........).)))))).))..)))...)))......	-7.16	0.035	...(((((.(........).
((.....))......((((...(((...((...((((.....(((.(........).)))))))))..)))..))))......	-7.30	0.044	.....(((.(........).
((.....))......((((...(((...((...(((((.....(((......)))....)))))))..)))..))))......	-7.40	0.053	(.....(((......)))..
((.....))......((((...(((...((.(........(((((.(........).)))))).))..)))..))))......	-7.73	0.092	...(((((.(........).

表 4.3　CRP アプタマーの二次構造のその存在確率

GGTTACGCCGCACATCAGTTTAGTGATAAAAGCCCGGTTACAGATGATCAGGGGCATTCGACAGGCTGGACATATC	Free energy	Probability	CGGTTACAGATGATCA
((.....)).....((((......)))).....((((.................)))).....((...))((....))	-6.62	0.018	(...............
((.....)).(((.........))).....((((.................)))).....((...))((....))	-7.04	0.037	(...............
((.....)).....((((......)))).....((((((......((((........))))))...)))).((....))	-6.72	0.022	((......((((....
((.....)).(((.........))).....((((((......((((........))))))...)))).((....))	-7.14	0.044	((......((((....
((.....)).(((.........))).....((((((......((((.(....).))))))...)))).((....))	-6.96	0.033	((......((((.(..
((.....)).(((.........))).....((((((......(((.........))))))...)))).((....))	-7.01	0.036	(((......(((....
((.....)).(((.........))).....((((((......(((.(....).))))))...)))).((....))	-6.83	0.026	(((......(((.(..
((.....)).(((.........)))...((.((((.................)))).)).........((....))	-6.72	0.022	(...............
((.....)).....((((......)))).....((((.................)))).............((....))	-6.96	0.033	(...............
((.....)).(((.........))).....((((.................)))).............((....))	-7.38	0.066	(...............
((.....)).(((.........))).....((((.................)))).........(.....).....	-6.96	0.033	(...............
((.....)).(((.........))).....((((.................))))..((.........)).....	-6.76	0.023	(...............
((.....)).(((.........)))...((.((((.................)))).))...((...)).....	-6.61	0.018	(...............
((.....)).....((((......)))).....((((.................)))).....((...)).....	-6.85	0.027	(...............
((.....)).(((.........))).....((((.................)))).....((...)).....	-7.27	0.055	(...............
((.....)).....((((......)))).....((((((......((((........))))))...)))).....	-6.95	0.032	((......((((....
((.....)).(((.........))).....((((((......((((........))))))...)))).....	-7.37	0.065	((......((((....
((.....)).....((((......)))).....((((((......((((.(....).))))))...)))).....	-6.77	0.024	((......((((.(..
((.....)).(((.........))).....((((((......((((.(....).))))))...)))).....	-7.19	0.048	((......((((.(..
((.....)).....((((......)))).....(((((((......(((.........))))))...)))).....	-6.82	0.026	(((......(((....
((.....)).(((.........))).....(((((((......(((.........))))))...)))).....	-7.24	0.052	(((......(((....
((.....)).....((((......)))).....(((((((......(((.(....).))))))...)))).....	-6.64	0.019	(((......(((.(..
((.....)).(((.........))).....(((((((......(((.(....).))))))...)))).....	-7.06	0.039	(((......(((.(..
((.....)).(((.........)))...((.((((.................)))).)).....	-6.95	0.032	(...............
((.....)).....((((......)))).....((((.................)))).....	-7.19	0.048	(...............
((.....)).(((.........))).....((((.................)))).....	-7.61	0.098	(...............
(((...))).(((.........))).....((((.................)))).....	-6.74	0.023	(...............

表 4.4　VEGF の推定結合領域がとり得る二次構造の存在確率

AGTTTAAAGTCACGTCTTAG	Probability
......)))..........)	0.189
...(((((.(........).	0.127
......)))..))......)	0.587
(.....(((......))))..	0.053
.....(((.(........).	0.044

表 4.5　CRP の推定結合領域がとり得る二次構造の存在確率

CGGTTACAGATGATCA	Probability
(((......(((.(..	0.084
((......((((....	0.164
(...............	0.534
(((......(((....	0.114
((......((((.(..	0.105

表 4.2 と表 4.3 から，VEGF と CRP の推定結合領域がとり得る二次構造の存在確率を計算すると，表 4.4 と表 4.5 となる．1 列目に推定結合領域がとり得る二次構造を示し，2 列目にその構造の存在確率を示す．VEGF の推定結合領域の二次構造には一貫性が無く，大きく分けて三つ存在する．推定結合領域より 5' 側の配列とステム構造を保有するもの（1 行目と 3 行目）と，推定結合領域より 3' 側の配列とステム構造を形成するもの（2 行目と 5 行目），そして推定結合領域にヘアピンループ構造をもつもの（4 行目）である．CRP の推定結合領域の二次構造は大きく分けて二つあり，推定結合領域の中央部分に，3' 側の配列とステム構造を形成するもの（1, 2, 4, 5 行目）とそうでないもの（3 行目）に分かれる．

図 4.4 と図 4.5 に，VEGF アプタマーと CRP アプタマーを小型化した配列の長さと推定結合領域の二次構造の存在確率の比の関係を示す．横軸が配列の長さを示し，縦軸が存在確率の比を示す．点の色は結合領域がとり得る二次構造を表し，凡例を図の左下に示す．水平の点線は存在確率の比が 1 の場合を表しており，今回の解析では，閾値を $\theta = 1$ としているため，この線より上にある点の中で最も短いものが，推定されたアプタマーの小型化配列となる．

VEGF において，推定結合領域の二次構造が薄い青と濃い青の場合は，短いアプタマーの小型化配列が推定されていないが，推定結合領域が薄緑と濃い緑の場合では，短いアプタマーの小型化配列が推定されている．特に，薄緑の二次構造で最も短い VEGF アプタマーの小型化配列が推定されている．さらに，薄緑と濃い緑の二次構造の場合では，存在確率の比も 5 倍以上であり，VEGF アプタマーの構造多様性が抑えられていることがわかる．CRP アプタマーにおいては，推定結合領域の二次構造が薄い青の場合に，最も短いアプタマーの小型化配列が推定された．ただし，VEGF アプタマーの時ほど構造多様性は抑えられてはいない．

表 4.6 と表 4.7 に，推定結合領域の各二次構造において推定された VEGF アプタマーと CRP アプタマーの小型化配列を示す．1 列目には推定された小型化配列，2 列目は小型化配列の長さ，3 列目と 4 列目は 5’ 末端と 3’ 末端から削除した塩基の数，5 列目に推定結合領域の二次構造を示す．VEGF の小型化配列の長さは，推定結合領域の二次構造に大きく依存する．そのため，短いもので 23 塩基であるが長いものは 51 塩基である．1 段目と 2 段目は類似した二次構造であるため，ほぼ同じ長さのアプタマーの小型化配列が推定されている．同様に 4 段目と 5 段目の二次構造も類似しており，完全に同じアプタマーの小型化配列が推定された．CRP は VEGF ほど推定結合領域の二次構造の影響はなく，短い配列で 22 塩基，長い配列で 37 塩基である．中段の 3 配列の二次構造は非常に類似しており，推定されたアプタマーの小型化配列も完全に等しい．

本来は，アプタマーの結合領域の真の二次構造は分からないので，時間と費用に余裕があれば，各々の構造から候補配列をすべて選択することが望ましい．その場合は，VEGF から 4 配列，CRP から 3 配列選択することになる．ただし，本論文では推定結合領域の各二次構造から選ばれた配列の中で，最も短い配列を小型化アプタマーの候補とする．つまり，VEGF アプタマーから一つ，CRP アプタマーから一つ候補配列を選択する．代わりに，結合が予測されない小型化配列の結合実験も行い，推定されたアプタマーの小型化配列が真の小型化配列であるか評価する．ただし，前節でも述べたように，CRP アプタマーに関しては，10 塩基分 5’ 側に長い候補配列を実験で評価する．プライマー領域を含んだ CRP アプタマーの小型化配列を，表 4.8 に示す．1 行目に，推定された CRP の小型化アプタマーを記載し，2 行目に，プライマー領域を含んだ CRP の小型化アプタマーを示し，3 行目に，修飾塩基を利用している位置を示す．1 行目の，ハイフンの部分には塩基がないことを示し，3 行目のハイフンは天然型の塩基を示し，アスタリスクは修飾塩基を示す．

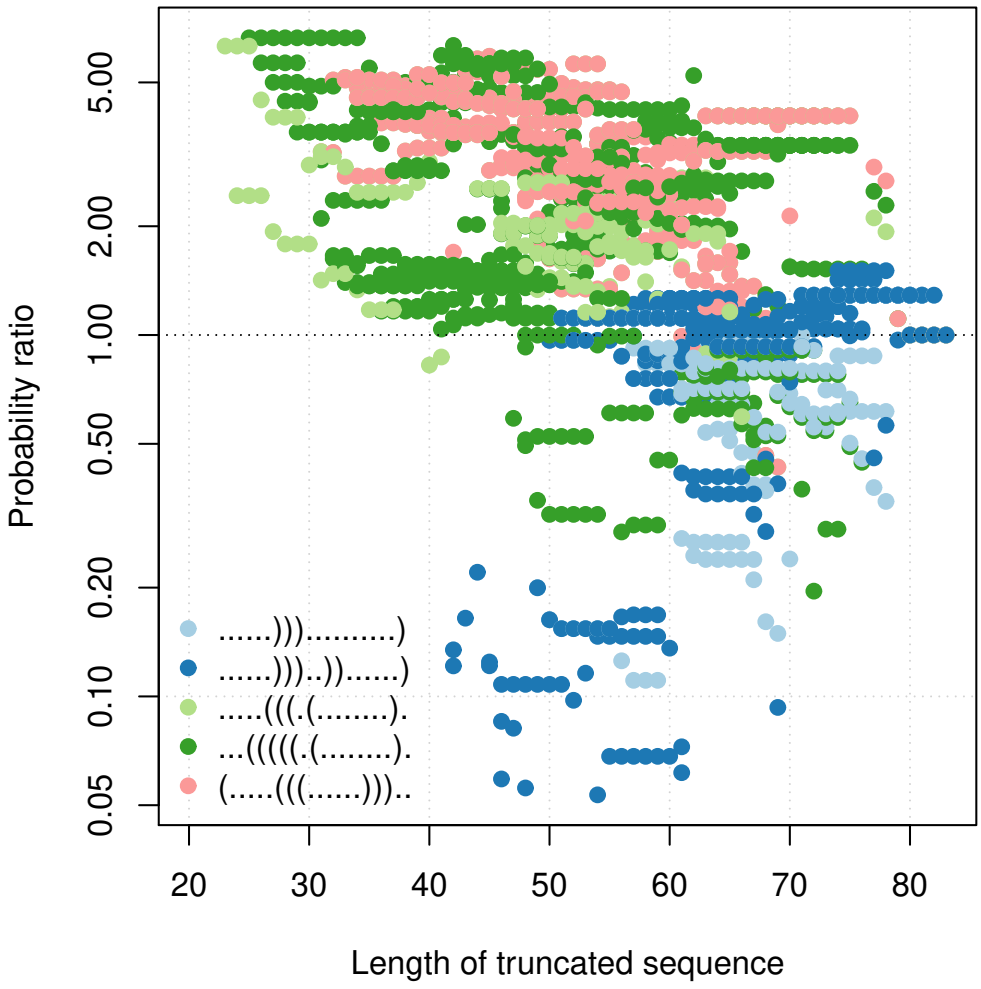

図 4.4　小型化配列の長さに対する推定結合領域の二次構造の存在確率の比 (VEGF)

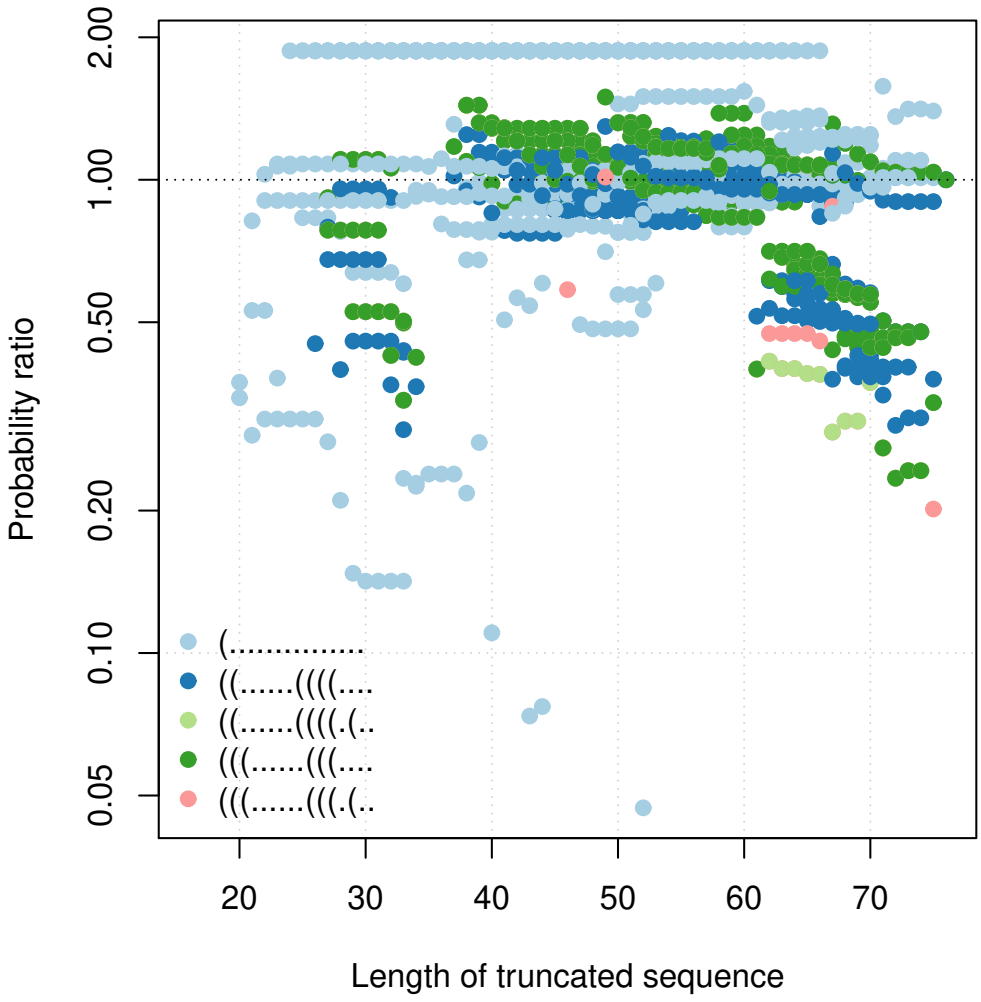

図 4.5　小型化配列の長さに対する推定結合領域の二次構造の存在確率の比 (CRP)

表 4.6　推定結合領域の各二次構造に対して推定されたアプタマーの小型化配列 (VEGF)

Truncated sequence	Length	m	n	PR	AGTTTAAAGTCACGTCTTAG
AGTTTAAAGTCACGTCTTAGTTT	23	37	23	6.302	.....(((.(........).
AGTTTAAAGTCACGTCTTAGTTTAA	25	37	21	6.643	...(((((.(........).
CCTTAGTTTAAAGTCACGTCTTAGTTTAAGG	31	33	19	3.218	(.....(((......))))..
ACCGTAAACAGTTCTGGTTCCTTAGTTTAAAGTCACGTCTTAGTTTAAGGC	51	14	18	1.114	......)))..........)
ACCGTAAACAGTTCTGGTTCCTTAGTTTAAAGTCACGTCTTAGTTTAAGGC	51	14	18	1.114	......)))..))......)

Abbreviation: PR, Probability ratio.

表 4.7　推定結合領域の各二次構造に対して推定されたアプタマーの小型化配列 (CRP)

Truncated sequence	Length	m	n	PR	CGGTTACAGATGATCA
CCCGGTTACAGATGATCAGGGG	22	32	22	1.027	(...............
CCGGTTACAGATGATCAGGGGCATTCGA	28	33	15	1.104	((......((((.(..
CCGGTTACAGATGATCAGGGGCATTCGA	28	33	15	1.104	(((......(((....
CCGGTTACAGATGATCAGGGGCATTCGA	28	33	15	1.104	(((......(((.(..
AGCCCGGTTACAGATGATCAGGGGCATTCGACAGGCT	37	30	9	1.019	((......((((....

Abbreviation: PR, probability ratio.

表 4.8　プライマー領域を含んだ CRP アプタマーの小型化配列

	Sequence
Estaimted CRP truncated aptamer	----------CCCGGTTACAGATGATCAGGGG
Estimated CRP truncated aptamer with 5'-primer region	GTGATAAAAGCCCGGTTACAGATGATCAGGGG
Position of modified base	---------------**-----*--*------

図 4.6，図 4.7 に，VEGF アプタマーと CRP アプタマーの推定結合領域の，各二次構造に対する，存在確率の比と除去された塩基の数との関係をヒートマップで示す．ヒートマップの上部に，対応する推定結合領域の二次構造を示す．縦軸の数値は 5' 末端から削除した塩基の数を示しており，横軸の数値は 3' 末端から削除した塩基の数を示す．よって，ヒートマップの右下に行くほど短い配列で，左上に行くほどもとの配列に近い．右下まで両端の塩基を削除すると，推定結合領域だけが残ることになる．ヒートマップの色は存在確率の比を表しており，色が濃いほど高い存在確率の比となる．逆に，色が白い場合は推定結合領域がもとの二次構造を維持していないことを示す．色は，0 から 0.5 が白，0.5 から 1，1 から 2，2 から 4，4 から 8 と区別している．左上の格子は，5' 側と 3' 側の塩基を一塩基も削っていないため，もとのアプタマーの配列を示す．そのため，必然的に存在確率の比は 1 となる．今回の解析において，閾値を $\theta = 1$ と設定したため，下から三番目の濃さの青色であれば，その小型化配列は標的分子と結合すると予測する．ヒートマップ上にある文字 T (True) および F (False) は，結合実験の結果を表し，T なら標的分子との結合を，F なら非結合を示す．

VEGF では，推定結合領域の二次構造に依存してヒートマップの色のバランスが異なる．特に，上段のヒートマップと比較して，中段と下段は濃い色が右下まで行き届いている．興味深いことに中段の図はお互いの色の濃淡が逆転している．つまり，中段左の図の中央付近は濃い青で満たされているが，中段右の図の中央付近は色が薄い．逆に，中段右の 3' 末端を 21 塩基除去したところは濃い青が 5' 側の塩基の除去数に依存せず広がっているが，中段左のその部分はちょうど白くなっている．実際には，中段左の二次構造と中段右の二次構造は非常に類似しているため，アプタマーの推定結合領域はこの二つの二次構造の平衡状態にあると考えられる．ただ，今回の手法では細かな平衡状態までは加味しておらず，異なる二次構造として扱っているため，どちらかのヒートマップでは，あたかも存在確率の比が下がっているように見える．そのため，類似した二次構造に関しては存在確率の和を適用して，平衡状態を考慮できるように拡張することもできる．CRP では，左上図のみで，存在確率の比の高い領域が広い範囲で見られるが，ほかの 4 つの図では 3' 末端を 10 塩基削ったところで，推定結合領域が二次構造を維持できていない．左上図以外の二次構造は非常に類似しており，4 つの二次構造が平衡状態にあると考えられる．以上のように，各二次構造に対するヒートマップを図示することは，二次構造同士の関連性や平衡状態を考察するのに有用である．

次に，図 4.6 と図 4.7 を一つのヒートマップにマージし，結合評価を行った配列と比較する．マージには，存在確率の比が最も高い二次構造の値を用いる．これは，式 4.8 の条

件式である，存在確率の比の最大値が θ より大きいという条件と同じである．

推定された VEGF アプタマーの小型化配列は，5' 末端から 37 塩基削除し，3' 末端から 23 塩基削除したもの $S_{\mathrm{VEGF}}(37, 23)$ である．つまり，ヒートマップの右下の濃い青の箇所が推定された最も短いアプタマーの小型化配列となる．このとき，右下の候補配列 $S_{\mathrm{VEGF}}(37, 23)$ は非常に濃い青色を示しており，構造多様性がもとの配列よりも抑えられていることがわかる．VEGF のアプタマーの推定結合領域の二次構造の最安定構造は，このアプタマーの小型化配列の二次構造と異なる．ゆえに，最安定構造だけでなく，準安定構造を考慮して，アプタマーの小型化配列を推定することが重要であるといえる．ヒートマップの色と文字 T と F を比較すると，ちょうど $S_{\mathrm{VEGF}}(37, 23)$ が VEGF との結合を示し，$S_{\mathrm{VEGF}}(37, 24)$ が結合を示さない．また，図には記載していないが結合領域部分を削除した $S_{\mathrm{VEGF}}(38, 23)$ も結合を示さなかった．以上より，真の最短の小型化 VEGF アプタマーは，推定された小型化配列と完全に一致した．参考ではあるが，VEGF のアプタマーは提案した手法が実施される前にアプタマーの小型化が実施されており，その配列は $S_{\mathrm{VEGF}}(33, 19)$ であり，長さは $|S_{\mathrm{VEGF}}(33, 19)| = 31$ 塩基であった．よって，今回の手法の適用により，既に作成されているアプタマーの小型化配列をさらに 8 塩基短くすることができた．

推定された CRP アプタマーの小型化配列は，5' 側から 32 塩基削除し 3' 側から 22 塩基削ったものである．前述したように CRP アプタマーは，推定した小型化配列より 10 塩基長い配列で結合評価を行っている．そのため，ヒートマップに記載されている結合可否の情報である文字 T と F は，5' 側から 22 塩基削除したところ $S_{\mathrm{CRP}}(22, 22)$ となっている．推定した CRP アプタマーの小型化配列 $S_{\mathrm{CRP}}(22, 22)$ は CRP との結合を示さなかったが，推定したアプタマーの小型化配列の 3' 側に一塩基加えた配列 $S_{\mathrm{CRP}}(22, 21)$ は結合を示した．つまり，真の最短の小型化 CRP アプタマーからたった一塩基だけ異なる小型化配列の推定に成功したことになる．CRP アプタマーの小型化配列も，VEGF と同様に提案手法を実現する以前に作成されており，そのアプタマーの小型化配列は，$S_{\mathrm{CRP}}(8, 21)$ であった．文字 T が $S_{\mathrm{CRP}}(8, 21)$ にあるのは，既にこのアプタマーの小型化配列が作製されていたためである．提案手法により，既に作成された CRP アプタマーの小型化配列 $S_{\mathrm{CRP}}(8, 21)$ をさらに 14 塩基削減することができた．また，CRP アプタマーの小型化配列の推定結果は，提案手法が修飾塩基を用いたアプタマーにも利用できることを示した．今回，閾値 θ を 1 と設定したが，仮に $\theta = 1.5$ としていれば，CPR アプタマーに関しても真のアプタマーの小型化に成功していた．

表 4.9 と表 4.10 に，VEGF アプタマーと CRP アプタマーの小型化配列と結合結果を

示す．各列は，アプタマーの小型化配列，結合の可否，小型化配列の長さ，5’ 末端から削除した塩基の数，3’ 末端から削除した塩基の数，推定結合領域の小型化前と後の存在確率の比を表す．小型化配列のハイフンは塩基がないことを示す，これは異なる配列の位置を合わせるためである．表 4.9 の NA は，推定結合領域まで削除しているため存在確率の比が計算されていないことを示す．VEGF に結合する配列の存在確率の比は，すべて 1.0 を超えており，結合しない配列の存在確率の比は 0 と予測されている．つまり，最も短い VEGF アプタマーの小型化配列を提案手法は推定したことになる．また，CRP アプタマーに関しては，結合すると推定した小型化配列 $S_{\mathrm{CRP}}(22, 22)$ が結合を示さなかった．しかし，一塩基 5’ 側に長くした小型化配列 $S_{\mathrm{CRP}}(22, 21)$ は CRP に結合した．よって，提案手法は最も短い CRP アプタマーの小型化配列を一塩基違いで推定できたことになる．

VEGF と CRP の小型化配列の SPR による結合実験の結果の一部を，図 4.10 と図 4.11 に示す．横軸は時間を示し，縦軸は標的分子のアプタマーとの結合量を示す．異なる色の線は標的分子の濃度を表す．時間 0 で，標的分子とアプタマーを結合させ，その後ゆっくりと洗浄を行う．アプタマーが標的分子と結合すると，結合量が増え，線が右肩上がりとなる．図 4.10 では，VEGF アプタマーの小型化配列である $S_{\mathrm{VEGF}}(33, 19)$ と $S_{\mathrm{VEGF}}(37, 23)$ が結合している．結合しない小型化配列では，結合量が変わらないため，線が水平のままとなる．VEGF と結合するアプタマーの小型化配列では，洗浄のときの解離がみられず，非常に結合力の強い小型化アプタマーであることがわかる．図 4.11 では，CRP アプタマーの小型化配列の中で $S_{\mathrm{CRP}}(8, 21)$ と $S_{\mathrm{CRP}}(22, 21)$ が結合している．

4.5　むすび

本章では，推定結合領域をもとにアプタマーの小型化配列を推定する新たな手法を提案した．まず，第 4.2 節で，アプタマーの小型化配列を推定するための最適化問題を定式化した．次に，第 4.3 節で，最適化問題を解くためのアルゴリズムを説明した．最後に，第 4.4 節で，提案手法を VEGF アプタマーと CRP アプタマーに適用して，評価を行った．推定された VEGF アプタマーの小型化配列は，真に最も短いアプタマーの小型化配列と一致した．また，修飾塩基を導入した CRP アプタマーに対しても，真に最も短いアプタマーと一塩基異なる小型化配列を推定し，精度が高いことを示した．アプタマーの小型化配列を推定する手法の精度が高かったもう一つの理由として，FSBC によるアプタマーの結合領域の推定の精度が高かったことも考えられる．以上のように，結合領域の構造を考

表 4.9　VEGF アプタマーの小型化配列の結合結果

Truncated sequence	Bind	Length	m	n	PR
--------TAAAGTCACGTCTTA--------	No	15	41	27	NA*
-------TTAAAGTCACGTCTTAG-------	No	17	40	26	NA*
------TTTAAAGTCACGTCTTAGT------	No	19	39	25	NA*
-----GTTTAAAGTCACGTCTTAGTT-----	No	21	38	24	NA*
-----GTTTAAAGTCACGTCTTAGTTT----	No	22	38	23	NA*
----AGTTTAAAGTCACGTCTTAGTT-----	No	22	37	24	0.000
----AGTTTAAAGTCACGTCTTAGTTT----	Yes	23	37	23	6.302
---TAGTTTAAAGTCACGTCTTAGTTTA---	Yes	25	36	22	2.430
---TAGTTTAAAGTCACGTCTTAGTTTAA--	Yes	26	36	21	6.643
--TTAGTTTAAAGTCACGTCTTAGTTTA---	Yes	26	35	22	2.430
--TTAGTTTAAAGTCACGTCTTAGTTTAA--	Yes	27	35	21	6.643
-CTTAGTTTAAAGTCACGTCTTAGTTTAA--	Yes	28	34	21	6.643
--TTAGTTTAAAGTCACGTCTTAGTTTAAG-	Yes	28	35	20	5.665
-CTTAGTTTAAAGTCACGTCTTAGTTTAAG-	Yes	29	34	20	5.665
CCTTAGTTTAAAGTCACGTCTTAGTTTAAG-	Yes	30	33	20	4.882
-CTTAGTTTAAAGTCACGTCTTAGTTTAAGG	Yes	30	34	19	4.391
CCTTAGTTTAAAGTCACGTCTTAGTTTAAGG	Yes	31	33	19	3.218

Abbreviation: PR, probability ratio.
*: PRs were not caluculated due to the elimination of estimated binding region.

慮しながら小型化配列を推定する提案手法は，実際のアプタマーを用いた検証において，高い精度でアプタマーの小型化配列を推定できることを示した．また，高精度に CRP アプタマーの小型化配列を推定できたことは，提案手法が修飾塩基を導入したアプタマーにも有用であることを示した．

表 4.10　CRP アプタマーの小型化配列の結合結果

Truncated sequence	Bind	Length	m	n	PR
----------GTGATAAAAGCCCGGTTACAGATGATCAG	No	29	22	25	0.000
----------GTGATAAAAGCCCGGTTACAGATGATCAGG	No	30	22	24	0.142
----------GTGATAAAAGCCCGGTTACAGATGATCAGGG	No	31	22	23	0.904
----------GTGATAAAAGCCCGGTTACAGATGATCAGGGG	No	32	22	22	1.080
----------GTGATAAAAGCCCGGTTACAGATGATCAGGGGC	Yes	33	22	21	1.873
CGCACATCAGTTTAGTGATAAAAGCCCGGTTACAGATGATCAGGGGC	Yes	47	8	21	1.873

Abbreviation: PR, probability ratio.

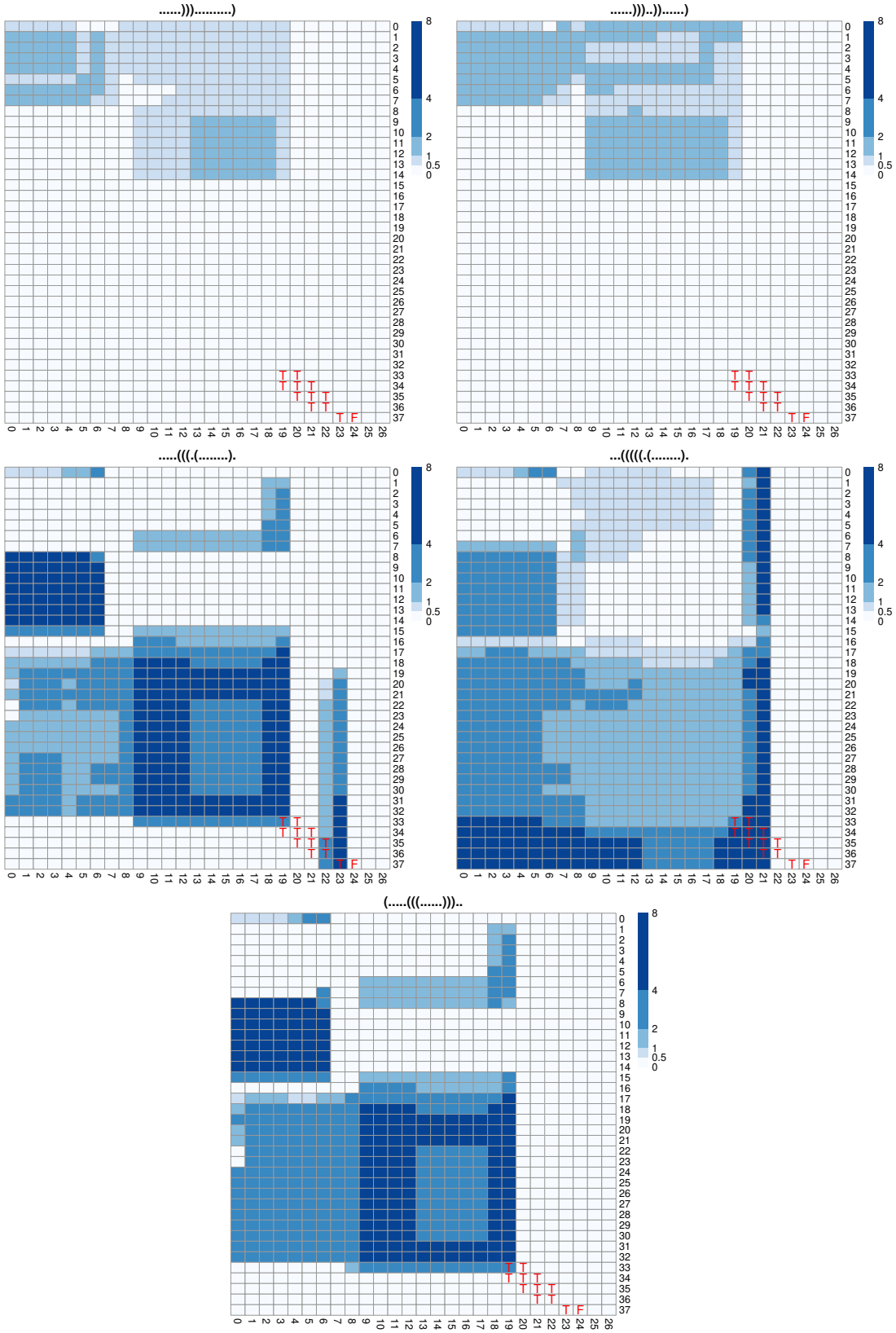

図 4.6　推定結合領域の各二次構造における存在確率の比 (VEGF)

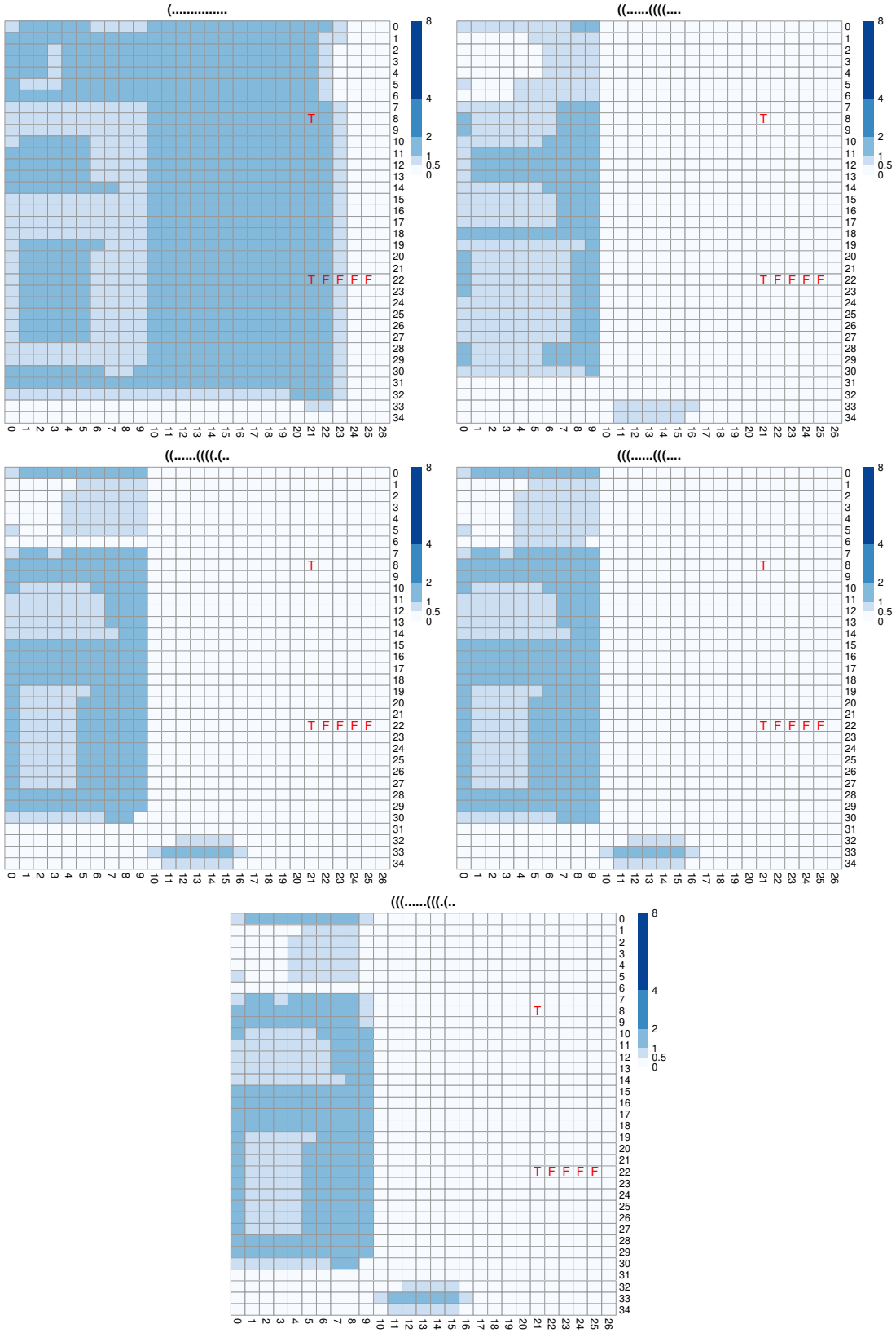

図 4.7　推定結合領域の各二次構造における存在確率の比 (CRP)

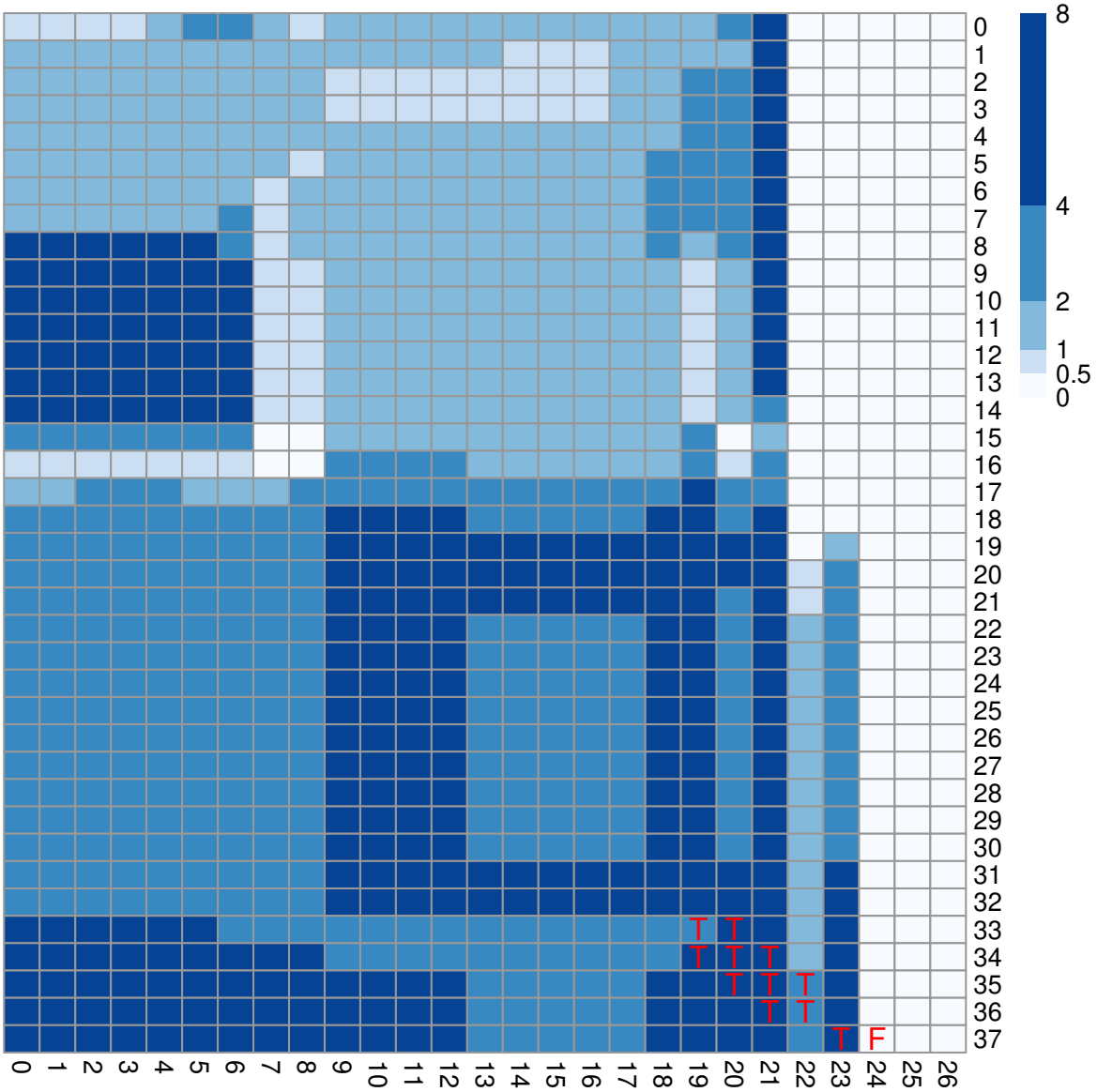

図 4.8　推定結合領域の二次構造の存在確率の比 (VEGF)

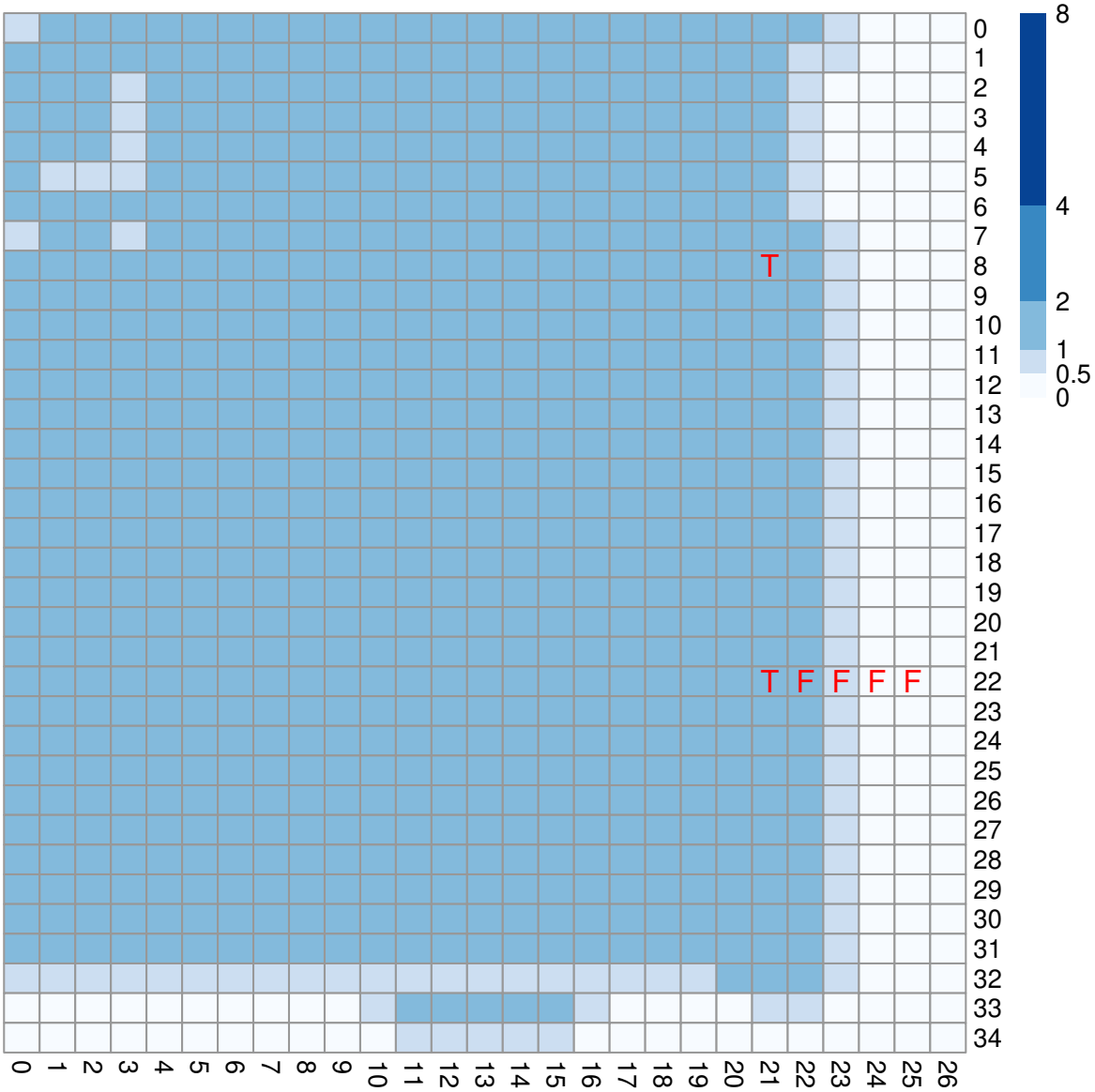

図 4.9　推定結合領域の二次構造の存在確率の比 (CRP)

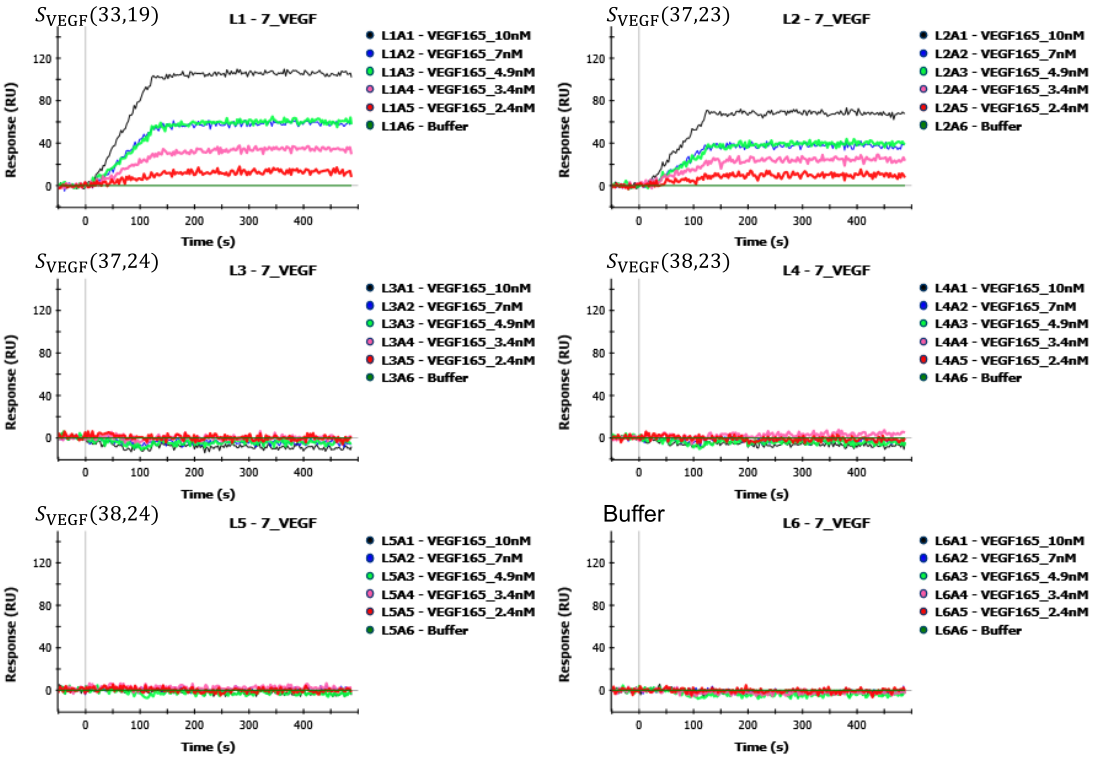

図 4.10　小型化配列の結合評価 (VEGF)

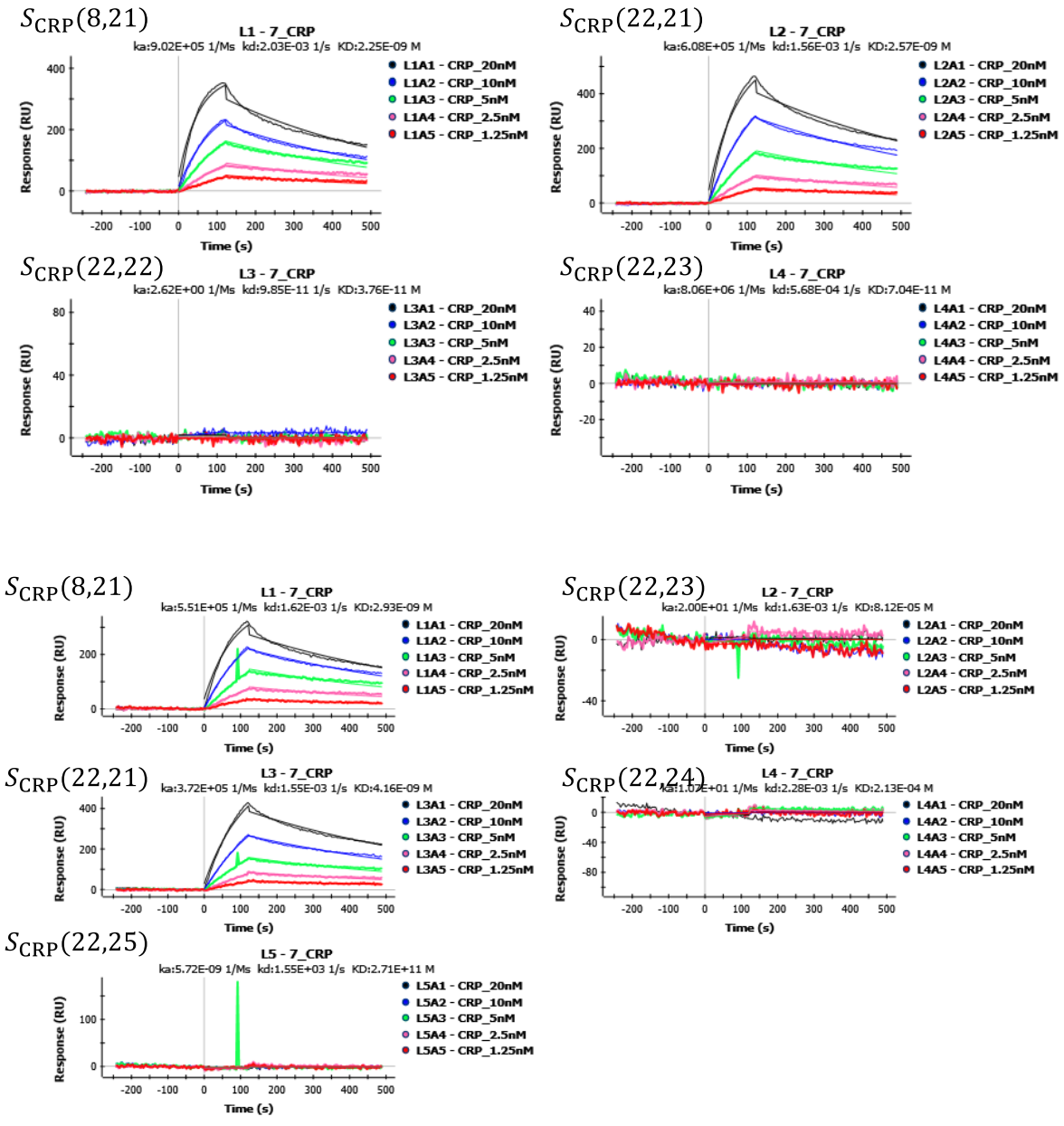

図 4.11　小型化配列の結合評価 (CRP)

第 5 章

結言

本論文の目的は，HT-SELEX データを用いたアプタマーの設計方法を確立することである．その目的のため，二つの新規手法を提案した．一つ目は，HT-SELEX データを高精度かつ高速に分けるクラスタリング手法である．二つ目は，アプタマーの小型化配列を高精度に予測する手法である．提案したクラスタリング手法である FSBC は，従来の手法と比較して最も高精度にクラスタを作成した．また，並列処理を行う pFSBC は，従来手法と比較して最も高速であった．提案したアプタマーの小型化配列を推定する手法は，高精度にアプタマーの小型化配列を推定した．以上より，提案する二つの手法の有用性が示された．本論文の成果により，HT-SELEX データを用いた効率的なアプタマーの設計方法が確立されたといえる．このことは，アプタマーの実用化においても意義があるといえる．

以下に，本論文における各章の要約を記述する．第 1 章では，本論文の背景と目的，および概要を述べた．第 2 章では，HT-SELEX データを用いたアプタマーの設計に関する基礎的考察に関して述べた．まず，アプタマーと HT-SELEX に関して述べた．次に，HT-SELEX データのクラスタリングに関して述べた．最後に，アプタマーの小型化に関して述べた．第 3 章では，HT-SELEX データに対する新たなクラスタリング手法 FSBC を提案した．はじめに，FSBC の概要について述べた．次に，特定の文字列を含む配列の出現確率について述べ，この出現確率を用いて，文字列のスコアである Z スコアを定義した．そして，探索領域を削減しながら高速に結合領域を推定する方法を述べ，推定した結合領域を利用してクラスタを作成する方法を述べた．また，FSBC を高速化するための並列処理と，FSBC から得られるクラスタの分布を比較する方法に関して述べた．最後に，FSBC を hESC と IL10RA を標的とした二つの HT-SELEX データを用いて評価し

た．評価の結果，FSBC が従来手法と比較して最も精度よくクラスタを作成し，並列化した提案手法 pFSBC は従来の手法と比較して最も高速であった．第 4 章では，アプタマーの小型化配列を推定する手法を提案した．まず，アプタマーの小型化配列を推定するための最適化問題を定式化した．次に，最適化問題を解くためのアルゴリズムを述べた．最後に，提案した手法を VEGF アプタマーと CRP アプタマーに適用し，手法の評価を行った．評価の結果より，提案した手法が高精度にアプタマーの小型化配列を推定することが示された．

今後の展望として，本論文で提案する手法により，アプタマーとアプタマーの小型化配列の取得が効率化され，より多くのアプタマーが実用化されると考えられる．そして，アプタマーを用いたバイオセンサー，医薬品，診断薬などが数多く普及し，アプタマーがより身近な存在となると考えられる．また，FSBC とアプタマーの小型化配列を推定する手法の結果を蓄積し，それらを統合的に評価することで，アプタマーの配列の法則性を見出すことができる可能性がある．法則性が見出されれば，将来的に，SELEX の実験を伴わないアプタマーの設計が可能になると考えられる．

参考文献

[1] H. Minagawa, K. Onodera, H. Fujita, T. Sakamoto, J. Akitomi, N. Kaneko, I. Shiratori, M. Kuwahara, K. Horii, and I. Waga, "Selection, characterization and application of artificial DNA aptamer containing appended bases with subnanomolar affinity for a salivary biomarker," Scientific Reports, vol.7, p.42716, March 2017.

[2] L.C. Bock, L.C. Griffin, J.A. Latham, E.H. Vermaas, and J.J. Toole, "Selection of single-stranded DNA molecules that bind and inhibit human thrombin," Nature, vol.355, no.6360, pp.564–566, Feb. 1992.

[3] S. Catuogno and C.L. Esposito, "Aptamer cell-based selection: Overview and advances," Biomedicines, vol.5, no.3, p.49, Sept. 2017.

[4] I. Shiratori, J. Akitomi, D.A. Boltz, K. Horii, M. Furuichi, and I. Waga, "Selection of DNA aptamers that bind to influenza A viruses with high affinity and broad subtype specificity," Biochemical and Biophysical Research Communications, vol.443, no.1, pp.37–41, Jan. 2014.

[5] S. Marton, F. Cleto, M.A. Krieger, and J. Cardoso, "Isolation of an aptamer that binds specifically to E. coli," PLOS ONE, vol.11, no.4, p.e0153637, April 2016.

[6] I. Cunha, R. Biltes, M.G.F. Sales, and V. Vasconcelos, "Aptamer-based biosensors to detect aquatic phycotoxins and cyanotoxins," Sensors, vol.18, no.7, p.2367, July 2018.

[7] G.R. Zimmermann, C.L. Wick, T.P. Shields, R.D. Jenison, and A. Pardi, "Molecular interactions and metal binding in the theophylline-binding core of an RNA aptamer.," RNA, vol.6, no.5, pp.659–667, Jan. 2000.

[8] E.M. McConnell, J. Nguyen, and Y. Li, "Aptamer-based biosensors for environ-

mental monitoring," Frontiers in Chemistry, vol.8, p.434, May 2020.

[9] G.K. Mishra, V. Sharma, and R.K. Mishra, "Electrochemical aptasensors for food and environmental safeguarding: A review," Biosensors, vol.8, no.2, p.28, June 2018.

[10] E.W.M. Ng, D.T. Shima, P. Calias, E.T. Cunningham, D.R. Guyer, and A.P. Adamis, "Pegaptanib, a targeted anti-VEGF aptamer for ocular vascular disease," Nature Reviews. Drug Discovery, vol.5, no.2, pp.123–132, Feb. 2006.

[11] P. Kumar Kulabhusan, B. Hussain, and M. Yüce, "Current perspectives on aptamers as diagnostic tools and therapeutic agents," Pharmaceutics, vol.12, no.7, p.646, July 2020.

[12] D. Ruiz Ciancio, M.R. Vargas, W.H. Thiel, M.A. Bruno, P.H. Giangrande, and M.B. Mestre, "Aptamers as diagnostic tools in cancer," Pharmaceuticals, vol.11, no.3, p.86, Sept. 2018.

[13] C. Tuerk and L. Gold, "Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase," Science, vol.249, no.4968, pp.505–510, Aug. 1990.

[14] A.D. Ellington and J.W. Szostak, "In vitro selection of RNA molecules that bind specific ligands," Nature, vol.346, no.6287, pp.818–822, Aug. 1990.

[15] M. Cho, Y. Xiao, J. Nie, R. Stewart, A.T. Csordas, S.S. Oh, J.A. Thomson, and H.T. Soh, "Quantitative selection of DNA aptamers through microfluidic selection and high-throughput sequencing," Proceedings of the National Academy of Sciences, vol.107, no.35, pp.15373–15378, Aug. 2010.

[16] B. Zimmermann, T. Gesell, D. Chen, C. Lorenz, and R. Schroeder, "Monitoring genomic sequences during SELEX using high-throughput sequencing: Neutral SELEX," PLOS ONE, vol.5, no.2, p.e9169, Feb. 2010.

[17] K.K. Alam, J.L. Chang, and D.H. Burke, "FASTAptamer: A bioinformatic toolkit for high-throughput sequence analysis of combinatorial selections," Molecular Therapy - Nucleic Acids, vol.4, p.e230, Jan. 2015.

[18] J. Hoinka, A. Berezhnoy, Z.E. Sauna, E. Gilboa, and T.M. Przytycka, "AptaCluster - A method to cluster HT-SELEX aptamer pools and lessons from its application," International Conference on Research in Computational Molecular Biology, vol.8394, pp.115–128, 2014.

[19] J. Caroli, C. Taccioli, A. De La Fuente, P. Serafini, and S. Bicciato, “APTANI: A computational tool to select aptamers through sequence-structure motif analysis of HT-SELEX data,” Bioinformatics, vol.32, no.2, pp.161–164, Jan. 2016.

[20] P. Dao, J. Hoinka, M. Takahashi, J. Zhou, M. Ho, Y. Wang, F. Costa, J.J. Rossi, R. Backofen, J. Burnett, and T.M. Przytycka, “AptaTRACE elucidates RNA sequence-structure motifs from selection trends in HT-SELEX experiments,” Cell Systems, vol.3, no.1, pp.62–70, July 2016.

[21] R.C. Edgar, “Search and clustering orders of magnitude faster than BLAST,” Bioinformatics, vol.26, no.19, pp.2460–2461, Oct. 2010.

[22] R.C. Edgar, “UNOISE2: Improved error-correction for Illumina 16S and ITS amplicon sequencing,” bioRxiv, p.081257, Oct. 2016.

[23] T.R. Allnutt, T.P. Quinn, M.F. Richardson, and T.M. Crowley, “Shortlisting aptamer candidates from HT-SELEX data,” Aptamers, vol.2, pp.36–44, May 2018.

[24] S. Kato, T. Ono, H. Minagawa, K. Horii, I. Shiratori, I. Waga, K. Ito, and T. Aoki, “FSBC: Fast string-based clustering for HT-SELEX data,” BMC Bioinformatics, vol.21, no.1, p.263, June 2020.

[25] S. Kato, T. Ono, M. Ito, K. Ito, H. Minagawa, K. Horii, I. Shiratori, I. Waga, and T. Aoki, “Parallel implementation of string-based clustering for HT-SELEX data,” EAI Endorsed Transactions on Bioengineering and Bioinformatics, vol.bebi 20, no.e4, pp.1–11, Oct. 2020.

[26] T. Ono, S. Kato, H. Minagawa, K. Horii, I. Shiratori, I. Waga, K. Ito, and T. Aoki, “Parallel implementation of motif-based clustering for HT-SELEX dataset,” Proc. of The 19th Annual IEEE International Conference on Bioinformatics and Bioengineering, pp.50–55, Oct. 2019.

[27] 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，和賀巌，青木孝文，“大量核酸配列データのための高速なクラスタリング手法の検討，” 平成 30 年度電気関係学会東北支部連合大会講演論文集，p.144，Sep. 2018.

[28] 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，和賀巌，青木孝文，“大量核酸配列データのための高速なクラスタリング手法とその性能評価，” 映像情報メディア学会技術報告，第 42 巻，pp.75–78，Aug. 2018.

[29] 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，和賀巌，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，” 研究報告バイオ情報学 (BIO)，第 2019-BIO-57 巻，pp.1–6，Mar. 2019.

[30] 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，和賀巌，青木孝文，“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，” 研究報告バイオ情報学 (BIO)，第 2019-BIO-58 巻，pp.1–6，Jun. 2019.

[31] 伊藤雅基，加藤信太郎，伊藤康一，皆川宏貴，堀井克紀，白鳥行大，和賀巌，青木孝文，“核酸アプタマー推定のためのクラスタリング手法とベクトル量子化を用いた解析，” 映像情報メディア学会技術報告，第 44 巻，pp.17–20，Sep. 2020.

[32] 加藤信太郎, 秋冨穣, 酒井伸也, 増田洋美, 辻祥太郎, 大津敬, 和賀巌, “Fast aptamer detection for large quantities of sequence data,” 第 32 回日本分子生物学会年会, Dec. 2009. ポスター発表.

[33] P. Jiang, S. Meyer, Z. Hou, N.E. Propson, H.T. Soh, J.A. Thomson, and R. Stewart, “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, Sept. 2014.

[34] J. Hoinka, A. Berezhnoy, P. Dao, Z.E. Sauna, E. Gilboa, and T.M. Przytycka, “Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery,” Nucleic Acids Research, vol.43, no.12, pp.5699–5707, July 2015.

[35] A. Levay, R. Brenneman, J. Hoinka, D. Sant, M. Cardone, G. Trinchieri, T.M. Przytycka, and A. Berezhnoy, “Identifying high-affinity aptamer ligands with defined cross-reactivity using high-throughput guided systematic evolution of ligands by exponential enrichment,” Nucleic Acids Research, vol.43, no.12, pp.e82–e82, July 2015.

[36] M. Jia, J. Sha, Z. Li, W. Wang, and H. Zhang, “High affinity truncated aptamers for ultra-sensitive colorimetric detection of bisphenol A with label-free aptasensor,” Food Chemistry, vol.317, p.126459, July 2020.

[37] H.A. Alhadrami, R. Chinnappan, S. Eissa, A.A. Rahamn, and M. Zourob, “High affinity truncated DNA aptamers for the development of fluorescence based progesterone biosensors,” Analytical Biochemistry, vol.525, pp.78–84, May 2017.

[38] 伊藤雅基，加藤信太郎，伊藤康一，皆川宏貴，堀井克紀，白鳥行大，和賀巌，青木

孝文，“二次構造予測を用いたアプタマーの小型化に関する検討，” 2020 年度電気関係学会東北支部連合大会，no.T01，Aug. 2020.

[39] D.E. Huizenga and J.W. Szostak, “A DNA aptamer that binds adenosine and ATP,” Biochemistry, vol.34, no.2, pp.656–665, Jan. 1995.

[40] A. Liang, L. Zhou, H. Qin, Y. Zhang, H. Ouyang, and Z. Jiang, “A highly sensitive aptamer-nanogold catalytic resonance scattering spectral assay for melamine,” Journal of Fluorescence, vol.21, no.5, pp.1907–1912, Sept. 2011.

[41] S. Dalirirad and A.J. Steckl, “Aptamer-based lateral flow assay for point of care cortisol detection in sweat,” Sensors and Actuators B: Chemical, vol.283, pp.79–86, March 2019.

[42] C.H. Lin and D.J. Patei, “Structural basis of DNA folding and recognition in an AMP-DNA aptamer complex: Distinct architectures but common recognition motifs for DNA and RNA aptamers complexed to AMP,” Chemistry & Biology, vol.4, no.11, pp.817–832, Nov. 1997.

[43] A.D. Gelinas, D.R. Davies, T.E. Edwards, J.C. Rohloff, J.D. Carter, C. Zhang, S. Gupta, Y. Ishikawa, M. Hirota, Y. Nakaishi, T.C. Jarvis, and N. Janjic, “Crystal structure of interleukin-6 in complex with a modified nucleic acid ligand,” Journal of Biological Chemistry, vol.289, no.12, pp.8720–8734, March 2014.

[44] J.C. Rohloff, A.D. Gelinas, T.C. Jarvis, U.A. Ochsner, D.J. Schneider, L. Gold, and N. Janjic, “Nucleic acid ligands with protein-like side chains: Modified aptamers and their use as diagnostic and therapeutic agents,” Molecular Therapy - Nucleic Acids, vol.3, p.e201, Jan. 2014.

[45] Y. Imaizumi, Y. Kasahara, H. Fujita, S. Kitadume, H. Ozaki, T. Endoh, M. Kuwahara, and N. Sugimoto, “Efficacy of base-modification on target binding of small molecule DNA aptamers,” Journal of the American Chemical Society, vol.135, no.25, pp.9412–9419, June 2013.

[46] H. Minagawa, Y. Kataoka, H. Fujita, M. Kuwahara, K. Horii, I. Shiratori, and I. Waga, “Modified DNA aptamers for C-reactive protein and lactate dehydrogenase-5 with sub-nanomolar affinities,” International Journal of Molecular Sciences, vol.21, no.8, p.2683, Jan. 2020.

[47] M. Kimoto, R. Yamashige, K.-i. Matsunaga, S. Yokoyama, and I. Hirao, “Gen-

eration of high-affinity DNA aptamers using an expanded genetic alphabet," Nature Biotechnology, vol.31, no.5, pp.453–457, May 2013.

[48] S. Klußmann, A. Nolte, R. Bald, V.A. Erdmann, and J.P. Fürste, "Mirror-image RNA that binds D-adenosine," Nature Biotechnology, vol.14, no.9, pp.1112–1115, Sept. 1996.

[49] R.E. Armstrong and G.F. Strouse, "Rationally manipulating aptamer binding affinities in a stem-loop molecular beacon," Bioconjugate Chemistry, vol.25, no.10, pp.1769–1776, Oct. 2014.

[50] T.A. Feagin, N. Maganzini, and H.T. Soh, "Strategies for creating structure-switching aptamers," ACS Sensors, vol.3, no.9, pp.1611–1615, Sept. 2018.

[51] N. Kaneko, K. Horii, J. Akitomi, S. Kato, I. Shiratori, and I. Waga, "An aptamer-based biosensor for direct, label-free detection of melamine in raw milk," Sensors, vol.18, no.10, p.3227, Oct. 2018.

[52] Y. Tomita, Y. Morita, H. Suga, and D. Fujiwara, "DNA module platform for developing colorimetric aptamer sensors," BioTechniques, vol.60, no.6, pp.285–292, June 2016.

[53] C. Teller, S. Shimron, and I. Willner, "Aptamer-DNAzyme hairpins for amplified biosensing," Analytical Chemistry, vol.81, no.21, pp.9114–9119, Nov. 2009.

[54] H. Abu-Ali, A. Nabok, and T.J. Smith, "Development of novel and highly specific ssDNA-aptamer-based electrochemical biosensor for rapid detection of mercury (II) and lead (II) ions in water," Chemosensors, vol.7, no.2, p.27, June 2019.

[55] L. Jin, Y. Nonaka, S. Miyakawa, M. Fujiwara, and Y. Nakamura, "Dual therapeutic action of a neutralizing anti-FGF2 aptamer in bone disease and bone cancer pain," Molecular Therapy: The Journal of the American Society of Gene Therapy, vol.24, no.11, pp.1974–1986, Nov. 2016.

[56] L. Gold, D. Ayers, J. Bertino, C. Bock, A. Bock, E.N. Brody, J. Carter, A.B. Dalby, B.E. Eaton, T. Fitzwater, D. Flather, A. Forbes, T. Foreman, C. Fowler, B. Gawande, M. Goss, M. Gunn, S. Gupta, D. Halladay, J. Heil, J. Heilig, B. Hicke, G. Husar, N. Janjic, T. Jarvis, S. Jennings, E. Katilius, T.R. Keeney, N. Kim, T.H. Koch, S. Kraemer, L. Kroiss, N. Le, D. Levine, W. Lindsey, B. Lollo, W. Mayfield, M. Mehan, R. Mehler, S.K. Nelson, M. Nelson, D. Nieuwlandt,

M. Nikrad, U. Ochsner, R.M. Ostroff, M. Otis, T. Parker, S. Pietrasiewicz, D.I. Resnicow, J. Rohloff, G. Sanders, S. Sattin, D. Schneider, B. Singer, M. Stanton, A. Sterkel, A. Stewart, S. Stratford, J.D. Vaught, M. Vrkljan, J.J. Walker, M. Watrobka, S. Waugh, A. Weiss, S.K. Wilcox, A. Wolfson, S.K. Wolk, C. Zhang, and D. Zichi, "Aptamer-based multiplexed proteomic technology for biomarker discovery," PLOS ONE, vol.5, no.12, p.e15004, Dec. 2010.

[57] M.R. Mehan, S.A. Williams, J.M. Siegfried, W.L. Bigbee, J.L. Weissfeld, D.O. Wilson, H.I. Pass, W.N. Rom, T. Muley, M. Meister, W. Franklin, Y.E. Miller, E.N. Brody, and R.M. Ostroff, "Validation of a blood protein signature for non-small cell lung cancer," Clinical Proteomics, vol.11, no.1, p.32, Aug. 2014.

[58] R.M. Ostroff, M.R. Mehan, A. Stewart, D. Ayers, E.N. Brody, S.A. Williams, S. Levin, B. Black, M. Harbut, M. Carbone, C. Goparaju, and H.I. Pass, "Early detection of malignant pleural mesothelioma in asbestos-exposed individuals with a noninvasive proteomics-based surveillance tool," PLOS ONE, vol.7, no.10, p.e46091, Oct. 2012.

[59] K.A. Olson, A.L. Beatty, B. Heidecker, M.C. Regan, E.N. Brody, T. Foreman, S. Kato, R.E. Mehler, B.S. Singer, K. Hveem, H. Dalen, D.G. Sterling, R.M. Lawn, N.B. Schiller, S.A. Williams, M.A. Whooley, and P. Ganz, "Association of growth differentiation factor 11/8, putative anti-ageing factor, with cardiovascular outcomes and overall mortality in humans: Analysis of the Heart and Soul and HUNT3 cohorts," European Heart Journal, vol.36, no.48, pp.3426–3434, Dec. 2015.

[60] R.B. D'Agostino, M.W. Russell, D.M. Huse, R.C. Ellison, H. Silbershatz, P.W.F. Wilson, and S.C. Hartz, "Primary and subsequent coronary risk appraisal: New results from the Framingham study," American Heart Journal, vol.139, no.2, Part 1, pp.272–281, Feb. 2000.

[61] P. Ganz, B. Heidecker, K. Hveem, C. Jonasson, S. Kato, M.R. Segal, D.G. Sterling, and S.A. Williams, "Development and validation of a protein-based risk score for cardiovascular outcomes among patients with stable coronary heart disease," JAMA, vol.315, no.23, pp.2532–2541, June 2016.

[62] Williams Stephen A., Murthy Ashwin C., DeLisle Robert K., Hyde Craig, Malarstig Anders, Ostroff Rachel, Weiss Sophie J., Segal Mark R., and Ganz

Peter, “Improving assessment of drug safety through proteomics,” Circulation, vol.137, no.10, pp.999–1010, March 2018.

[63] S.A. Williams, M. Kivimaki, C. Langenberg, A.D. Hingorani, J.P. Casas, C. Bouchard, C. Jonasson, M.A. Sarzynski, M.J. Shipley, L. Alexander, J. Ash, T. Bauer, J. Chadwick, G. Datta, R.K. DeLisle, Y. Hagar, M. Hinterberg, R. Ostroff, S. Weiss, P. Ganz, and N.J. Wareham, “Plasma protein patterns as comprehensive indicators of health,” Nature Medicine, vol.25, no.12, pp.1851–1857, Dec. 2019.

[64] N. Guan, Q. Fan, J. Ding, Y. Zhao, J. Lu, Y. Ai, G. Xu, S. Zhu, C. Yao, L. Jiang, J. Miao, H. Zhang, D. Zhao, X. Liu, and Y. Yao, “Melamine-contaminated powdered formula and urolithiasis in young children,” New England Journal of Medicine, vol.360, no.11, pp.1067–1074, March 2009.

[65] A.D. Ellington and J.W. Szostak, “Selection in vitro of single-stranded DNA molecules that fold into specific ligand-binding structures,” Nature, vol.355, no.6363, pp.850–852, Feb. 1992.

[66] K. Yang and L. Zhang, “Performance comparison between k-tuple distance and four model-based distances in phylogenetic tree reconstruction,” Nucleic Acids Research, vol.36, no.5, pp.e33–e33, March 2008.

[67] J. Hoinka, E. Zotenko, A. Friedman, Z.E. Sauna, and T.M. Przytycka, “Identification of sequence–structure RNA binding motifs for SELEX-derived aptamers,” Bioinformatics, vol.28, no.12, pp.i215–i223, June 2012.

[68] R. Lorenz, S.H. Bernhart, C. Höner zu Siederdissen, H. Tafer, C. Flamm, P.F. Stadler, and I.L. Hofacker, “ViennaRNA Package 2.0,” Algorithms for Molecular Biology, vol.6, no.1, p.26, Nov. 2011.

[69] G. Blackshields, F. Sievers, W. Shi, A. Wilm, and D.G. Higgins, “Sequence embedding for fast construction of guide trees for multiple sequence alignment,” Algorithms for Molecular Biology, vol.5, no.1, p.21, May 2010.

[70] F. Sievers, A. Wilm, D. Dineen, T.J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, J.D. Thompson, and D.G. Higgins, “Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega,” Molecular Systems Biology, vol.7, no.1, p.539, Jan. 2011.

[71] Y. Ding, C.Y. Chan, and C.E. Lawrence, “Sfold web server for statistical folding

and rational design of nucleic acids," Nucleic Acids Research, vol.32, no.suppl_2, pp.W135–W141, July 2004.

[72] D.H. Turner and D.H. Mathews, "NNDB: The nearest neighbor parameter database for predicting stability of nucleic acid secondary structure," Nucleic Acids Research, vol.38, no.suppl_1, pp.D280–D282, Jan. 2010.

[73] M. Andronescu, A. Condon, H.H. Hoos, D.H. Mathews, and K.P. Murphy, "Efficient parameter estimation for RNA secondary structure prediction," Bioinformatics, vol.23, no.13, pp.i19–i28, July 2007.

[74] K. Padmanabhan, K.P. Padmanabhan, J.D. Ferrara, J.E. Sadler, and A. Tulinsky, "The structure of alpha-thrombin inhibited by a 15-mer single-stranded DNA aptamer.," Journal of Biological Chemistry, vol.268, no.24, pp.17651–17654, Aug. 1993.

[75] S.B. Long, M.B. Long, R.R. White, and B.A. Sullenger, "Crystal structure of an RNA aptamer bound to thrombin," RNA, vol.14, no.12, pp.2504–2512, Jan. 2008.

[76] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The protein data bank," Nucleic Acids Research, vol.28, no.1, pp.235–242, Jan. 2000.

[77] N. Kaneko, K. Horii, S. Kato, J. Akitomi, and I. Waga, "High-throughput quantitative screening of peroxidase-mimicking DNAzymes on a microarray by using electrochemical detection," Analytical Chemistry, vol.85, no.11, pp.5430–5435, June 2013.

[78] H. Minagawa, A. Shimizu, Y. Kataoka, M. Kuwahara, S. Kato, K. Horii, I. Shiratori, and I. Waga, "Fluorescence polarization-based rapid detection system for salivary biomarkers using modified DNA aptamers containing base-appended bases," Analytical Chemistry, vol.92, no.2, pp.1780–1787, Jan. 2020.

[79] M. Popenda, M. Szachniuk, M. Antczak, K.J. Purzycka, P. Lukasiak, N. Bartol, J. Blazewicz, and R.W. Adamiak, "Automated 3D structure composition for large RNAs," Nucleic Acids Research, vol.40, no.14, pp.e112–e112, Aug. 2012.

[80] M. Popenda, M. Szachniuk, M. Blazewicz, S. Wasik, E.K. Burke, J. Blazewicz, and R.W. Adamiak, "RNA FRABASE 2.0: An advanced web-accessible database with the capacity to search the three-dimensional fragments within

RNA structures," BMC Bioinformatics, vol.11, no.1, p.231, May 2010.

[81] M. Parisien and F. Major, "The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data," Nature, vol.452, no.7183, pp.51–55, March 2008.

[82] P. Svoboda and A.D. Cara, "Hairpin RNA: A secondary structure of primary importance," Cellular and Molecular Life Sciences CMLS, vol.63, no.7, pp.901–908, April 2006.

[83] D.R. Davies, A.D. Gelinas, C. Zhang, J.C. Rohloff, J.D. Carter, D. O'Connell, S.M. Waugh, S.K. Wolk, W.S. Mayfield, A.B. Burgin, T.E. Edwards, L.J. Stewart, L. Gold, N. Janjic, and T.C. Jarvis, "Unique motifs and hydrophobic interactions shape the binding of modified DNA ligands to protein targets," Proceedings of the National Academy of Sciences, vol.109, no.49, pp.19971–19976, Dec. 2012.

[84] G. Varani and W.H. McClain, "The G·U wobble base pair," EMBO reports, vol.1, no.1, pp.18–23, July 2000.

[85] S. Rhee, Z.-j. Han, K. Liu, H.T. Miles, and D.R. Davies, "Structure of a triple helical DNA with a triplex-duplex junction," Biochemistry, vol.38, no.51, pp.16810–16815, Dec. 1999.

[86] A. Pica, I. Russo Krauss, V. Parente, H. Tateishi-Karimata, S. Nagatoishi, K. Tsumoto, N. Sugimoto, and F. Sica, "Through-bond effects in the ternary complexes of thrombin sandwiched by two DNA aptamers," Nucleic Acids Research, vol.45, no.1, pp.461–469, Jan. 2017.

[87] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," Nucleic Acids Research, vol.31, no.13, pp.3406–3415, July 2003.

[88] J. Akitomi, S. Kato, Y. Yoshida, K. Horii, M. Furuichi, and I. Waga, "ValFold: Program for the aptamer truncation process," Bioinformation, vol.7, no.1, pp.38–40, Aug. 2011.

[89] C.B. Do, D.A. Woods, and S. Batzoglou, "CONTRAfold: RNA secondary structure prediction without physics-based models," Bioinformatics, vol.22, no.14, pp.e90–e98, July 2006.

[90] J. Singh, J. Hanson, K. Paliwal, and Y. Zhou, "RNA secondary structure prediction using an ensemble of two-dimensional deep neural networks and transfer learning," Nature Communications, vol.10, no.1, p.5407, Nov. 2019.

[91] J.S. Mattick and I.V. Makunin, "Non-coding RNA," Human Molecular Genetics, vol.15, no.suppl_1, pp.R17–R29, April 2006.

[92] K. Sato, Y. Kato, M. Hamada, T. Akutsu, and K. Asai, "IPknot: Fast and accurate prediction of RNA secondary structures with pseudoknots using integer programming," Bioinformatics, vol.27, no.13, pp.i85–i93, July 2011.

[93] K. Horii, K. Omi, Y. Yoshida, Y. Imai, N. Sakai, A. Oka, H. Masuda, M. Furuichi, T. Tanimoto, and I. Waga, "Development of a sphingosylphosphorylcholine detection system using RNA aptamers," Molecules, vol.15, no.8, pp.5742–5755, Aug. 2010.

[94] P. Kerpedjiev, S. Hammer, and I.L. Hofacker, "Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams," Bioinformatics, vol.31, no.20, pp.3377–3379, Oct. 2015.

[95] G. Van Rossum and F.L. Drake, Python 3 Reference Manual, CreateSpace, Scotts Valley, CA, 2009.

[96] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," arXiv, p.1802.03426, Sept. 2020.

[97] L. Laurens van derMaaten and G. Hinton, "Visualizing data using t-SNE," Journal of Machine Learning Research, vol.9, no.86, pp.2579–2605, 2008.

[98] K. Ishigaki, M. Akiyama, M. Kanai, A. Takahashi, E. Kawakami, H. Sugishita, S. Sakaue, N. Matoba, S.-K. Low, Y. Okada, C. Terao, T. Amariuta, S. Gazal, Y. Kochi, M. Horikoshi, K. Suzuki, K. Ito, S. Koyama, K. Ozaki, S. Niida, Y. Sakata, Y. Sakata, T. Kohno, K. Shiraishi, Y. Momozawa, M. Hirata, K. Matsuda, M. Ikeda, N. Iwata, S. Ikegawa, I. Kou, T. Tanaka, H. Nakagawa, A. Suzuki, T. Hirota, M. Tamari, K. Chayama, D. Miki, M. Mori, S. Nagayama, Y. Daigo, Y. Miki, T. Katagiri, O. Ogawa, W. Obara, H. Ito, T. Yoshida, I. Imoto, T. Takahashi, C. Tanikawa, T. Suzuki, N. Sinozaki, S. Minami, H. Yamaguchi, S. Asai, Y. Takahashi, K. Yamaji, K. Takahashi, T. Fujioka, R. Takata, H. Yanai, A. Masumoto, Y. Koretsune, H. Kutsumi, M. Higashiyama, S. Murayama, N. Minegishi, K. Suzuki, K. Tanno, A. Shimizu, T. Yamaji, M. Iwasaki, N. Sawada, H. Uemura, K. Tanaka, M. Naito, M. Sasaki, K. Wakai, S. Tsugane, M. Yamamoto, K. Yamamoto, Y. Murakami, Y. Nakamura, S. Raychaudhuri, J. Inazawa, T. Yamauchi, T. Kadowaki, M. Kubo, and Y. Kamatani, "Large-scale

genome-wide association study in a Japanese population identifies novel susceptibility loci across different diseases," Nature Genetics, vol.52, no.7, pp.669–679, July 2020.

[99] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria, 2020.

[100] R.C. Gentleman, V.J. Carey, D.M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A.J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J.Y. Yang, and J. Zhang, "Bioconductor: Open software development for computational biology and bioinformatics," Genome Biology, vol.5, no.10, p.R80, Sept. 2004.

[101] J. Melville, "uwot: The uniform manifold approximation and projection (UMAP) method for dimensionality reduction," 2020. R package version 0.1.10.

[102] J.S. McCaskill, "The equilibrium partition function and base pair binding probabilities for RNA secondary structure," Biopolymers, vol.29, no.6-7, pp.1105–1119, 1990 May-Jun.

[103] M. Thulliez, D. Angoulvant, M.L. Le Lez, A.-P. Jonville-Bera, P.-J. Pisella, F. Gueyffier, and T. Bejan-Angoulvant, "Cardiovascular events and bleeding risk associated with intravitreal antivascular endothelial growth factor monoclonal antibodies: Systematic review and meta-analysis," JAMA Ophthalmology, vol.132, no.11, pp.1317–1326, Nov. 2014.

[104] N. Gupta, S. Mansoor, A. Sharma, A. Sapkal, J. Sheth, P. Falatoonzadeh, B. Kuppermann, and M. Kenney, "Diabetic retinopathy and VEGF," The Open Ophthalmology Journal, vol.7, pp.4–10, Feb. 2013.

[105] H.L. Goel and A.M. Mercurio, "VEGF targets the tumour cell," Nature Reviews Cancer, vol.13, no.12, pp.871–882, Dec. 2013.

[106] W.J. Guan, Z.Y. Ni, Y. Hu, W.H. Liang, C.Q. Ou, J.X. He, L. Liu, H. Shan, C.L. Lei, D.S.C. Hui, B. Du, L.J. Li, G. Zeng, K.Y. Yuen, R.C. Chen, C.L. Tang, T. Wang, P.Y. Chen, J. Xiang, S.Y. Li, J.L. Wang, Z.J. Liang, Y.X. Peng, L. Wei, Y. Liu, Y.H. Hu, P. Peng, J.M. Wang, J.Y. Liu, Z. Chen, G. Li, Z.J. Zheng, S.Q. Qiu, J. Luo, C.J. Ye, S.Y. Zhu, N.S. Zhong, and China Medical Treatment Expert Group for COVID-19, "Clinical characteristics of

coronavirus disease 2019 in China," New England Journal of Medicine, vol.382, no.18, pp.1708–1720, April 2020.

[107] D.H. Mathews, M.D. Disney, J.L. Childs, S.J. Schroeder, M. Zuker, and D.H. Turner, "Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure," Proceedings of the National Academy of Sciences, vol.101, no.19, pp.7287–7292, May 2004.

謝辞

本論文の執筆に当たり，多くの方々にご支援いただきました．

指導教員である青木孝文教授は，私の社会人編入学を受け入れてくださり，本論文のための研究を実施する機会を与えてくださいました．また，研究に関しては終始温かい激励とご指導，ご鞭撻を賜りました．ここに深く感謝の意を表します．

副査を引き受けてくださった，中尾光之教授ならびに木下賢吾教授より，審査にて貴重なご助言をいただきました．今後は，先生方にいただいたご助言を心に留めつつ研究を続けていきたいと思います．ここに深く感謝の意を表します．

伊藤康一准教授には，本論文の執筆に関して，細部に至るまで懇切丁寧に指導していただきました．ここに深く感謝の意を表します．

筆者と同じ研究室の卒業生である小野貴義氏，博士前期課程の伊藤雅基氏とは，ともに同じ研究を行い，プログラムの実装に関して多大な協力をしていただきました．また，有意義な議論により本論文を大きく発展させることができました．同研究室の博士後期課程の河合洋弥氏には，論文の校閲に協力していただきました．心より感謝いたします．

NEC ソリューションイノベータ株式会社プロフェッショナルフェロー和賀巌氏，同主席プロフェッショナル坂崎純次氏には，私が学術研究の道に進むことを許していただきました．心よりお礼申し上げます．また，NEC ソリューションイノベータ株式会社シニアプロフェッショナル堀井克紀氏，同シニアプロフェッショナル白鳥行大氏，同主任皆川宏貴氏には，本研究における評価実験の結果の取得など多大な支援をしていただきました．心よりお礼申し上げます．

最後に，これまで私をあたたかく応援してくれた亡き父，母，姉，いつもそばで支えてくれた妻，いつも私にエネルギーをくれる二人の息子たちに心から感謝します．

2021 年 1 月 16 日

加藤　信太郎