RNA デザイン

早稲田大学 理工学術院[*1]、産業技術総合研究所 生体システムビッグデータ解析オープンイノベーションラボラトリ[*2]、日本医科大学 大学院医学研究科[*3]
浜田道昭[*1~3]

RNAdesign

In this review, the author, who specializes in bioinformatics, provides a comprehensive explanation of RNA design, from fundamentals to cutting-edge technologies. The review introduces RNA design methodologies, including RNA inverse folding and mRNA design, based on the relationships between RNA sequence, structure, and function. Furthermore, it details new deep learning-based methods, RaptGen and RfamGen, demonstrating efficient design approaches for RNA aptamer sequences and functional RNA family sequences. The review also discusses future perspectives, including integration with high-throughput experimental technologies, large language models, and applications of quantum computing.

本稿では、バイオインフォマティクス(生命情報科学)を専門とする筆者が、RNA デザインの基礎から最新技術までを解説する。RNA デザインの手法として、RNA の配列・構造と機能の関係性を基盤に、RNA 逆フォールディングや mRNA デザインなどの手法を紹介する。さらに、深層学習を活用した新手法である RaptGen と RfamGen について詳述し、RNA アプタマー配列や機能性 RNA ファミリー配列の効率的な設計方法を示す。さらに、ハイスループット実験技術との融合や、大規模言語モデル、量子コンピューティングの応用など、将来の展望も議論する。

Michiaki Hamada[*1~3]

Keyword: RNA design, RNA structure, RNA inverse folding, large language model

1. 序論

RNA(リボ核酸, Ribonucleic acid)は、生命の起源から現代の複雑な生命システムに至るまで、生物学的プロセスにおいて中心的な役割を果たしてきた分子である[1)]。かつては単なる遺伝情報の仲介者(メッセンジャー RNA, mRNA)と考えられていた RNA だが、今日ではその多様な機能が明らかになり、遺伝子発現調節、触媒活性、構造的役割など、細胞内で多岐にわたる役割を担っていることがわかってきた[2)]。この認識の変化に伴い、RNA を人工的に設計し、操作する技術、すなわち RNA デザイン(RNA design)の重要性が急速に高まっている。

RNA デザインは、特定の機能や配列・構造をもつ RNA 分子を人工的に設計する過程を指す。この分野は、合成生物学、医薬品開発、ナノテクノロジーなど、幅広い応用可能性を秘めている。例えば、mRNA ワクチン、RNAi治療薬など、RNA デザインの成果はすでに実用化され、社会に大きなインパクトを与えている[3)]。一方で、RNA デザインには多くの課題が存在する。RNA の配列、構造、機能の関係は複雑であり、望みの機能をもつ RNA 配列を設計することは容易ではない。また、RNA の構造は環境に応じて動的に変化するため、設計した RNA が実際の生物学的環境下で期待どおりに機能するかどうかを予測することも困難である。

これらの課題に対処するため、RNA デザインの分野ではさまざまなアプローチが開発されてきた。

[*1] Faculty of Science and Engineering, Waseda University
[*2] Computational Bio Big-Data Open Innovation Laboratory(CBBD-OIL), National Institute of Advanced Industrial Science and Technology(AIST)
[*3] Graduate School of Medicine, Nippon Medical University

RNAの機能と構造には密接な関連性があることが知られているため、初期のアプローチとしては、RNA逆フォールディング(RNA inverse folding)のアルゴリズムの開発が主流であった。これは、目的の二次構造を形成するRNA配列を見つけ出すアプローチである。さらに最近では、深層学習に代表される人工知能(AI)技術と、その学習データを取得するための大規模な実験技術の発展に伴い、「データ駆動型」のアプローチがRNAデザイン研究において多く用いられるようになってきている。

本稿では、これらのRNAデザインの基礎から最新の技術動向までを概観する。まず、2節で、RNAの構造と機能の基本的な関係性について説明し、続いて3節で、RNAデザインの主要な手法を紹介する。さらに、4節では、RNAデザインの将来展望について議論する。本稿の執筆者は、バイオインフォマティクス(生命情報科学)を専門とする研究者であり、そのため本稿も情報科学を中心としたRNAデザインに関する内容がメインとなっていることをご注意いただきたい。RNAデザインは、基礎生物学の理解を深めるだけでなく、医療やバイオテクノロジーに革新をもたらす可能性を秘めた、重要かつ興味深い研究分野である。本稿が、RNAデザインに関心をもつ研究者や学生にとって有用な情報源となることを切に願う。

2. RNAの構造と機能

RNAの化学構造を理解することは、その多様な機能と挙動を把握し、デザインにつなげる上で重要である。RNAは、リボース糖、リン酸基、そして4種類の主要な塩基から構成されている。これらの要素が組み合わさり、RNAの複雑かつ動的な構造を形成している。

RNAの基本的な構成要素から見ていく。まず、リボース糖は5つの炭素原子からなる環状構造で、2'位に特徴的な水酸基(-OH)をもっている。この2'-OHがRNAの化学的反応性と構造的柔軟性に大きく寄与しており、DNAとの主な違いとなっている。リン酸基は、このリボース糖の3'位と5'位を連結する役割を果たしている。負電荷をもつリン酸基は、RNAの親水性と溶解性に寄与するとともに、骨格の主要な構成要素として分子に安定性を提供している。

RNAを構成する4つの主要な塩基は、プリン塩基であるアデニン(A)とグアニン(G)、そしてピリミジン塩基であるシトシン(C)とウラシル(U)である。これらの塩基は、水素結合を介して特異的な塩基対(base-pair)を形成し、RNAの二次構造や三次構造の形成に重要な役割を果たしている。また、特にtRNAやrRNAでは、これらの主要塩基に加えて、メチル化やシュードウリジン化などの修飾塩基(modified base)が存在し、RNAの構造安定性や機能に影響を与えている。

RNAの一次構造は、ヌクレオチド(リボース糖+リン酸基+塩基)が直鎖状に連なった配列(塩基の文字列)として表現される(**図1A**)。この配列は5'末端から3'末端への方向性をもち、リン酸ジエステル結合によって連結されている。通常、RNA配列は5'から3'方向に記述される(例:5'-AUGC-3')。

RNAの二次構造は、一次構造の塩基間で形成される水素結合によって生じる高次構造である(**図1B**)。最も一般的な塩基対はWatson-Crick塩基対で、A-UペアとG-Cペアがこれに該当する。G-Cペアは3本の水素結合を形成するため、A-Uペアよりも安定である。また、1本の水素結合を形成するG-Uのようなウォブル塩基対も頻繁に観察され、RNAの機能や構造の多様性に寄与している。

RNAの三次構造は、これらの二次構造要素がさらに折りたたまれ、相互作用することで形成される複雑な立体構造である(**図1C**)。三次構造の安定化には、塩基スタッキング、金属イオン(特にMg^{2+})との相互作用、水素結合ネットワークなどが重要な役割を果たしている。特徴的な三次構造モチーフとしては、シュードノット、キッシングヘアピン、コアキシャルスタッキング、A-minorモチーフなどがある。これらの構造は、RNAの触媒活性や分子認識など、特定の機能の発現に重要である。

さらに、複数のRNA分子間やRNAとタンパク質間の相互作用によって形成される四次構造(複合体立体構造)も存在する。リボソームやスプライソソーム、RNAi誘導サイレンシング複合体(RISC)な

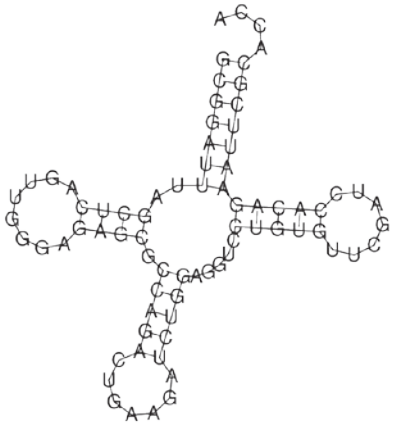
A）1 次構造（配列）

GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUGUUCGAUCCACAGAAUUCGCACCA

B1）2 次構造（カッコ表記）

(((((((..((((........))))).((((((.......))))))).....((((((.......))))))))))))....

B2）2 次構造（グラフ表記）



C）3 次構造（立体構造）



図1　RNA の配列と構造（酵母の $tRNA^{Phe}$ の例）

（A）一次構造は RNA の塩基配列を5’ から3’ の方向に文字列として表現したもの（配列）である。（B）二次構造（secondary structure）は、塩基対（一般には G-C、A-U、G-U 塩基対を考える場合が多い）の集合として表現できる抽象化された高次構造である。（C）三次構造は原子配置まで含めて立体構造の詳細を表現したものである。

どがその例である。機能性の分子は単体で機能を実現することは稀であり、これらの複合体は、生命活動に不可欠なさまざまなプロセスを担っている。加えて、RNA は環境の変化、例えば温度変化、イオン濃度の変化、リガンドとの結合、タンパク質との相互作用などに応じて、その構造を変化させることができる。この構造の柔軟性が、RNA の多様な機能を可能にしている。

以上のような RNA 構造の重要性を背景に、RNA 構造を予測・解析するための情報科学的手法が数多く開発されている。例えば、RNA の二次構造予測（RNA フォールディング）はバイオインフォマティクスの初期の頃から長く研究されている歴史のある研究テーマであり、多くの優れたアルゴリズム・ソフトウェアが存在している[4]。近年では、RNA 立体構造予測、RNA と他の分子の複合体構造予測などの研究も進んできているが、タンパク質の立体構造・複合体構造予測に比べると予測の精度などに未だ改善の余地を残している[5]。一方で、RNA の構造を解析するための実験的手法も進歩している。化学プロービング法（SHAPE や DMS プロービングなど）、X 線結晶構造解析、NMR 分光法、クライオ電子顕微鏡などの技術が、RNA の構造解明に大きく貢献している。これらの技術は RNA デザイン研究においても重要な役割を果たすことが期待されている。

3. RNA デザインの手法

本節では、RNA デザインの手法について述べる。長年の研究実績を有する RNA 逆フォールディング、および、COVID-19 のワクチン開発で特に注目されている mRNA デザインに引き続き、筆者らが近年提案した RNA アプタマー配列デザインの手法 RaptGen、および、機能性 RNA デザインの手法 RfamGen についての紹介を行う。

3-1. RNA逆フォールディング

RNA逆フォールディング(RNA inverse folding)は、研究の歴史が長く、かつRNAデザインにおいて重要な役割を果たす計算手法である。この手法は、目的とする構造(一般には二次構造を考えることが多い)に折れ畳むRNA配列を見つけ出すことを目的としている。通常のRNAフォールディング(RNA folding)問題が、与えられたRNA配列から構造を予測するのに対し、逆フォールディングはこの逆の問題、つまり与えられた構造から配列を設計する問題に取り組む(**図2**)。前述のとおりRNAの構造と機能には密接な関連性があるため、目的の構造をもつRNA配列を設計することは、RNAデザインにおいて有用なアプローチの一つとなる。

RNA逆フォールディングの代表的な手法を**表1**にまとめた。この表が示すとおり、多くの手法はRNAの二次構造をターゲットにした逆フォールディングアルゴリズムであるが、一部はシュードノット構造や三次構造をターゲットにした設計が可能である。また、最適化手法としては、遺伝的アルゴリズムなどの古典的なものから、強化学習やニューラルネットワークなど、幅広く用いられている。RNA逆フォールディングは歴史が長いため多くの手法が提案されている。さらに包括なレビューに関しては、例えば、文献6)を参照していただきたい。

このようなRNA逆フォールディング手法は、リボザイムやリボスイッチなどの機能性RNAの設計

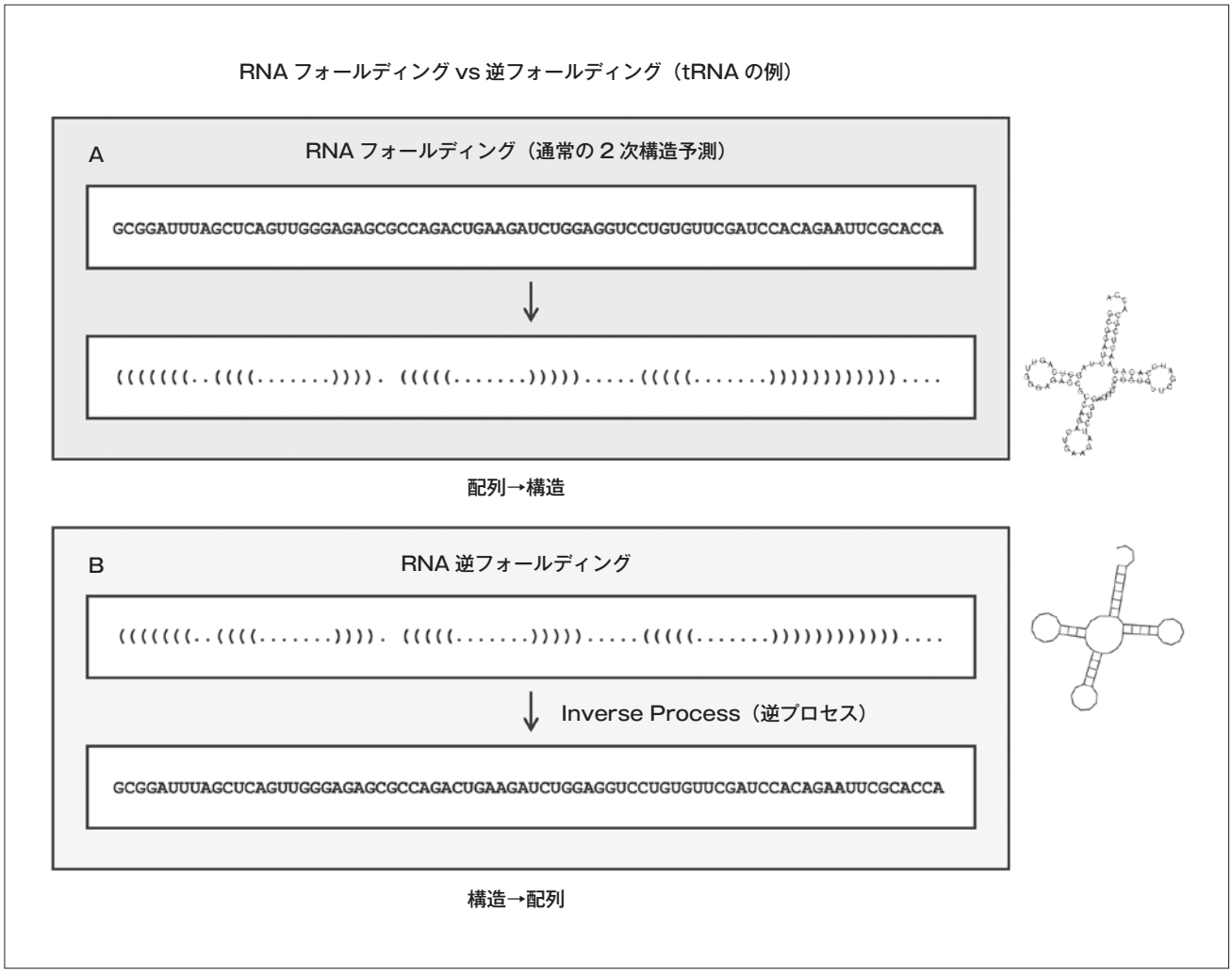


図2　RNA逆フォールディングの概要
通常のRNAフォールディングは、与えられたRNA配列からその二次構造の予測を行う。これに対して、逆フォールディングは与えられた二次構造からその構造に折れ畳まるRNA配列の予測を行う。

表1　RNA逆フォールディング手法

ソフトウェア名	発表年	PK	3D	特徴	文献
RNAInverse	1994			最も初期のRNA逆フォールディングアルゴリズムの1つ。Vienna RNA packageに含まれている	28
INFO-RNA	2006			動的計画法による初期配列生成と、エネルギーに基づく近傍選択を用いた、確率的局所探索を組み合わせた高速アルゴリズム	29
MODERNA	2011			多目的遺伝的アルゴリズムを用いて、安定性と構造類似性の2つの目的関数を同時に最適化	30
antaRNA	2015			アントコロニー最適化を用いる。二次構造制約に加えてGC含量制約や配列制約を満たすRNA配列を設計可能	31
MCTS-RNA	2017	✓		モンテカルロ木探索を用いる。GC含量を制御しながら効率的に目的の二次構造を持つRNA配列を設計	32
SentRNA	2018			EteRNAゲームのプレイヤーが提出した解答データを用いてニューラルネットワークを訓練し、人間の設計戦略を学習	33
LEARNA	2019			強化学習を用いる手法であり、新しいRNA設計タスクごとにゼロから学習を行うLEARNAと多数のRNA設計学習を事前に学習したMeta-LEARNAと組み合わせたMeta-LEARNA-Adaptがある	34
SAMFEO	2023			アンサンブル目的関数の最適化とマルチフロンティア探索を組み合わせ、構造情報を活用した効率的なRNA設計を実現し、多数のMFE解を生成可能	35
RiboDiffusion	2024	✓	✓	拡散モデルを用いてRNA 3次構造からRNA配列を生成する逆折りたたみ手法で、グラフニューラルネットワークベースの構造モジュールとTransformerベースの配列モジュールを組み合わせた方法	36
ES＋eM2dRNAs	2024			進化戦略を用いてターゲットRNA構造を分解して最適化	37
gRNAde	2024	✓	✓	幾何学的深層学習を用いて3D RNAフォールディングを行う手法	38

PK：シュードノット構造を考慮可能であるか。3D：立体構造を考慮可能であるか。

に応用されており、一定の成功を収めている。例えば、ハンマーヘッドリボザイムの活性を向上させた変異体の設計や、新規のリボスイッチの設計などが報告されている[7,8]。しかし、配列－構造－機能の関係は複雑であり、構造情報のみに基づいた設計では限界がある。そのため、最新のRNA設計手法では、逆フォールディングを基盤としつつ、進化的情報、実験データ、機械学習による予測など、多様なアプローチを統合することが重要となる。今後のRNA逆フォールディングの発展方向としては、以下のような点が挙げられる。

1. 二次構造だけでなく、三次構造や機能的側面を考慮した逆フォールディング技術の開発。
2. 大規模データの活用：実験的に検証されたRNA配列-構造-機能データを学習に取り入れ、より現実的な設計を可能にする。
3. 動的な構造変化の考慮[9]：リボスイッチなどの環境応答性RNAの設計のため、複数の構造状態間の遷移を考慮した逆フォールディングアルゴリズムの開発。
4. 計算効率の向上：より大規模で複雑なRNAの設計を可能にするため、並列計算や量子アニーリングなどの新しい計算技術の活用。

RNA逆フォールディングは、RNAデザインの基礎的なアプローチとして重要な役割を果たしており、今後も新しい技術との融合や改良が進むことで、より高度で実用的なRNA設計ツールとして発展していくことが期待される。

3-2. mRNAデザイン

mRNAデザインは、COVID-19パンデミックにおけるmRNAワクチンの成功により、その重要性と潜在的可能性が広く認識されるようになった。mRNAは、5'キャップ、5'非翻訳領域(UTR)、コーディング配列(CDS)、3'UTR、ポリA鎖から構成される(**図3**)。これらの各要素は、mRNAの安定性(分解のされにくさ)や翻訳効率(翻訳のされやす

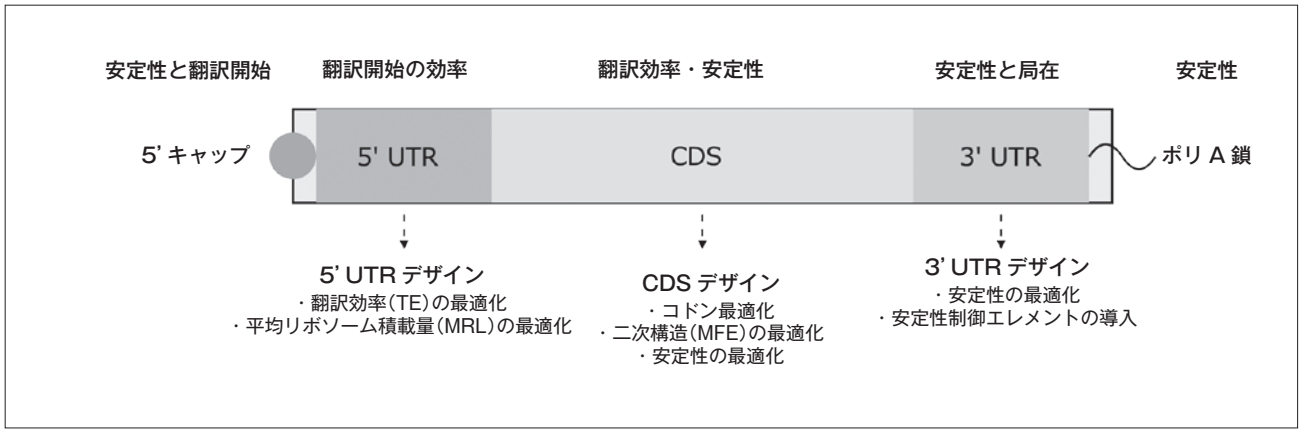


図3　mRNA の配列デザインの概要

表2　mRNA デザイン手法

手法	発表年	モデル・アルゴリズム	部位	指標	文献
CDSfold	2016	DP	CDS	MFE	39
Optimus 5-Prime	2019	CNN/GA	5'UTR	MRL	10
RiboTree	2021	MCTS	CDS	AUP	11
DegScore	2022	リッジ回帰	全体	安定性(予測のみ)	12
iCodon	2022	遺伝的アルゴリズム	CDS	安定性	40
LinearDesign	2023	DFA/SCFG	CDS	MFE/CAI	41
iDRO	2023	BiLSTM-CRF(CDS)/ RNA-Bart(UTR)	全体	ヒトの遺伝子配列	15
CUSTOM	2023	RF	CDS	組織特異的TE	42
mRNA2vec	2024	data2vec	5'UTR+CDS	TE/安定性	43
Smart5UTR	2024	CNN/GA	5'UTR	MRL	44
LAMBO-DNABERT	2024	LAMBO	5'UTR	MRL/G4/安定性/AGC含量率	45
RiboCode	2024	CNN/AM	CDS	TE/MFE	46
GEMORNA	2024	Transformer	全体	CAI/MRL/MFE	47

DP：動的計画法、CNN：畳み込みニューラルネットワーク、GA：遺伝的アルゴリズム、DFA：有限状態オートマトン、SCFG：確率文脈自由文法、RF：ランダムフォレスト、AM：アクティベーション最大化。

さ)に直接的あるいは間接的に関与し、最終的なタンパク質の生産量に重要な役割を果たす。

2024年9月現在における mRNA のデザイン手法について**表2**にまとめた。mRNA のデザイン手法においては、mRNA の各要素に対して、以下の指標の一つあるいは複数を最適化するような配列の探索が行われる。最適化には、遺伝的なアルゴリズムやモンテカルロ木探索アルゴリズムなどが利用される。

コドン適応指標(Codon Adaptation Index, CAI)：CAI は、遺伝子のコドン使用が最適なコドン使用にどれだけ近いかを示す指標である。各コドンの相対適応度(relative adaptiveness)の幾何平均として計算され、0から1の間をとる。CAI が1に近いほど、その遺伝子のコドン使用が最適であることを示す。

最小自由エネルギー(Minimum Free Energy, MFE)：MFE は与えられた RNA の二次構造の安定性を示す指標である。MFE が小さいほど、安定な二次構造を形成することを示す。与えられた配列に対する MFE は、エネルギーモデルに基づく二次構造予測ソフトウェア[4]で計算可能である。

平均リボソーム積載量(Mean Ribosom Load, MRL)：MRL は、mRNA に積載されているリボソームの

量であり、翻訳の活性と相関することが期待される。例えば、Sample et al.[10]では、大規模並列レポーターアッセイ(Massively Parallel Reporter Assay, MPRA)を使って、約28万の5'UTR配列に対してMRLが測定されており、このような大規模データを用いて、MRLの予測モデルが複数考案されている。

翻訳効率(Translation Efficiency, TE):TEは、翻訳の効率を示す指標であり、Ribo-seq(転写産物全体にわたってリボソームの結合量を測定することが可能となる技術。Ribosome profilingと呼ばれることもある)を用いて計算することが可能である。さまざまな細胞種や組織に対してRibo-seqのデータが取得されている(例:https://riboseq.org/)。

平均非塩基対形成確率(Average Unpaired Probability, AUP):AUPは、RNAの各位置に対して計算される、塩基対を形成しない確率である。1に近いほどその位置が塩基対を形成しにくいことを示す。AUPは、mRNAの安定性(分解されにくさ)に相関することが知られている[11]。塩基対確率を計算可能な二次構造予測ソフトウェアを利用して計算が可能である。

安定性(分解されにくさ):細胞内や溶液中でのRNA分子の分解されやすさは、翻訳効率と同等以上に、最終的なタンパク質産物の量に影響をしていることが示唆されている[12]。したがって、安定性を最適化指標とすることはmRNAデザインにおいて重要となる。例えば、文献13)ではヒト細胞株におけるmRNAの安定性に関する実験データが取得されている。さらに、このデータを用いて安定性の予測を行う手法も提案されており、配列のデザインにも活用されている[11]。

免疫原性:設計したmRNAを医薬品として用いるためには免疫原性を考慮することが重要となる。例えば、免疫原性を有するウリジンをなるべく排除した(ACG含有率が高い)配列デザインなどが考えられる。

G4構造(G-quadruplex構造):G4構造は、DNAやRNAにおいて見られる特殊な高次構造である。5'UTR中のG4構造は翻訳を阻害することが知られており[14]、G4構造をなるべく含まないような配列デザインが考えられる。

天然の遺伝子配列:進化の過程でmRNA配列が最適化されていると考えると、現存している天然のmRNA自体も最適化の指標になり得る。例えば、iDROでは、ヒト遺伝子が人間の細胞での翻訳に最も効率的な配列であると仮定し、これを模倣するように学習を行っている[15]。

mRNAデザインについては急速に研究が進んでいるものの、依然として課題が残されている。1つ目は、目的のタンパク質が与えられた際に、mRNA配列の全体(5'キャップ+5'UTR+CDS+3'UTR+ポリA鎖)を、それぞれの要素の相互の依存性も考慮して、トータルにデザインすることである。2つ目は、上述の複数の指標の最適な組み合わせとそれらを同時に制御したデザインを行うことである。3つ目は、限られた学習データから高品質な予測モデルおよびデザイン手法を導出することである。最近ではmRNAの各部位の言語モデルも開発されている(4-7節で述べる)。これらの言語モデルを用いることにより、少数の学習データを用いたデザイン手法の開発が期待されている。

将来的には、これらの課題を解決することで、mRNAデザイン技術はさらに進化し、より効果的かつ安全な遺伝子治療やワクチン、そして新しい生物工学ツールの開発につながると期待される。

3-3. RaptGen:RNAアプタマーのデザインを目的とした深層学習技術

RaptGen[16]は、ハイスループットな実験データから、RNAアプタマー(RNA aptamer)(中分子医薬品の一つであり、次世代の医薬品の一つとして現在の創薬の課題を解決することが期待されている[17])を効率的にデザインするために開発された深層学習ベースの手法である。

RaptGenは、HT-SELEXから得られた配列データを入力して利用する。SELEX(Systematic Evolution of Ligands by Exponential Enrichment)法[18]は、特定の標的分子に対して高い親和性と特異性をもつ核酸アプタマーを選別する技術である。この手法

では、ランダムな配列をもつ核酸ライブラリから始まり、標的との結合、分離、増幅のサイクル(ラウンド)を繰り返すことで、ターゲットに結合する配列を濃縮する。HT-SELEX(ハイスループットSELEX)は、SELEX各ラウンドの配列情報を高速シークエンサで取得する。これらのデータには、選別過程で濃縮された多様な配列情報や、複数ラウンドにわたる配列進化の情報が含まれている。

RaptGenの一つの目的は、入力されたSELEXデータセットに含まれない新規アプタマー配列を生成する生成モデルを構築することである。これを実現するために、RaptGenは配列を低次元の潜在空間に埋め込み、その空間から新しい配列を生成する。そのため、RpatGenは、変分オートエンコーダー(Variational auto-encoder, VAE)ベースの深層学習モデルのデコーダーとしてプロファイル隠れマルコフモデル(profile HMM)を採用した新しいモデルである(**図4A**)。デコーダーにプロファイルHMMを用いることにより、アプタマー配列中の重要な部分配列(モチーフ)に関するロバストな方法となっている。

RaptGenの評価は、シミュレーションデータと実創薬標的のHT-SELEXデータの両方を用いて行われた。その結果、以下の点が明らかになった。

1. 類似のモチーフをもつアプタマー配列を、置換・挿入・欠失に対してロバストに、潜在空間内の近い領域に埋め込むことができた。
2. 学習済みの潜在空間から、実験データには含まれない新規の高活性アプタマーの生成に成功した。
3. プロファイルHMMの長さを制御することにより、活性を有する短鎖化配列の生成に成功した。従来は、配列短鎖化は人手により試行錯誤的に行われていたが、これを自動化することが可能になった。
4. ベイズ最適化を用いて、活性情報に基づいて新たな高活性アプタマーを生成することができた。

一方で、RaptGenにはいくつかの課題も残されている。例えば、計算コストの高さや、RNAの二次構造情報の明示的な考慮がされていない点が挙げられる。また、SELEXの全ラウンドのデータを統合的に解析する方法の開発も今後の課題である。

RaptGenはコマンドラインで実行するプログラムとして提供されているが、現在、ウェブブラウザベースのユーザインターフェースRaptGen-UIの開発を行っている(**図5**)。このユーザインターフェースを用いることにより、RaptGenの主要機能である、潜在空間のクラスタリングと可視化・配列生成、

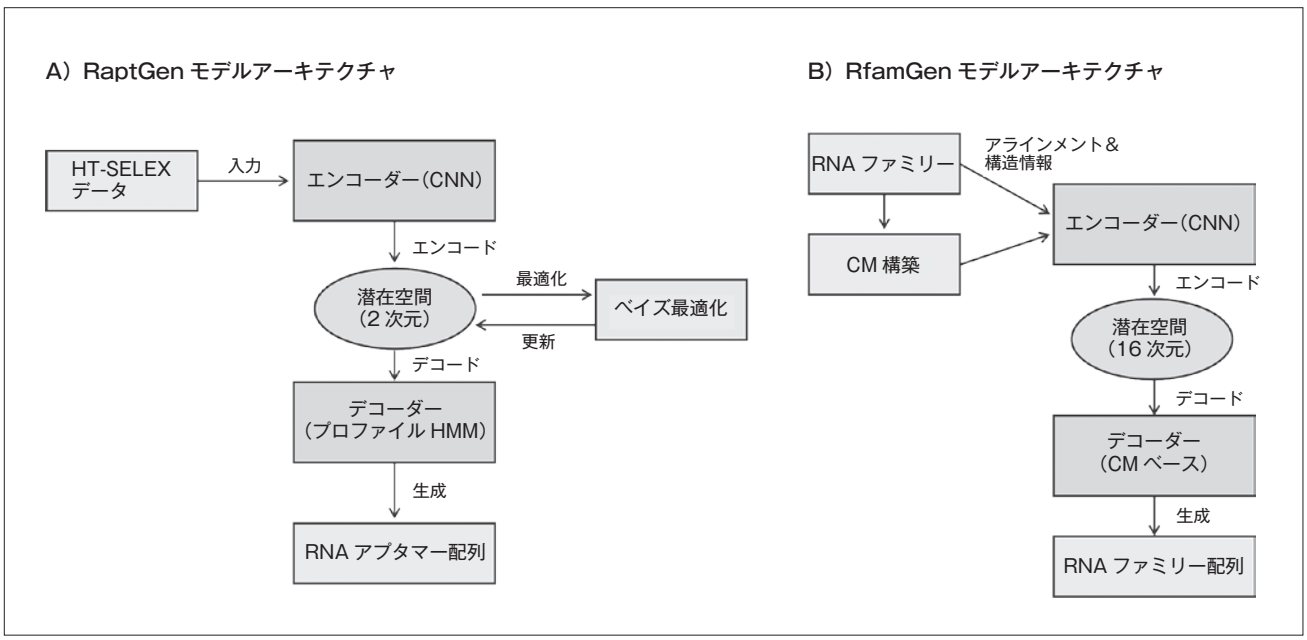


図4　RaptGen と RfamGen のモデル
共に変分オートエンコーダー(VAE)をベースにしたモデルとなっている。

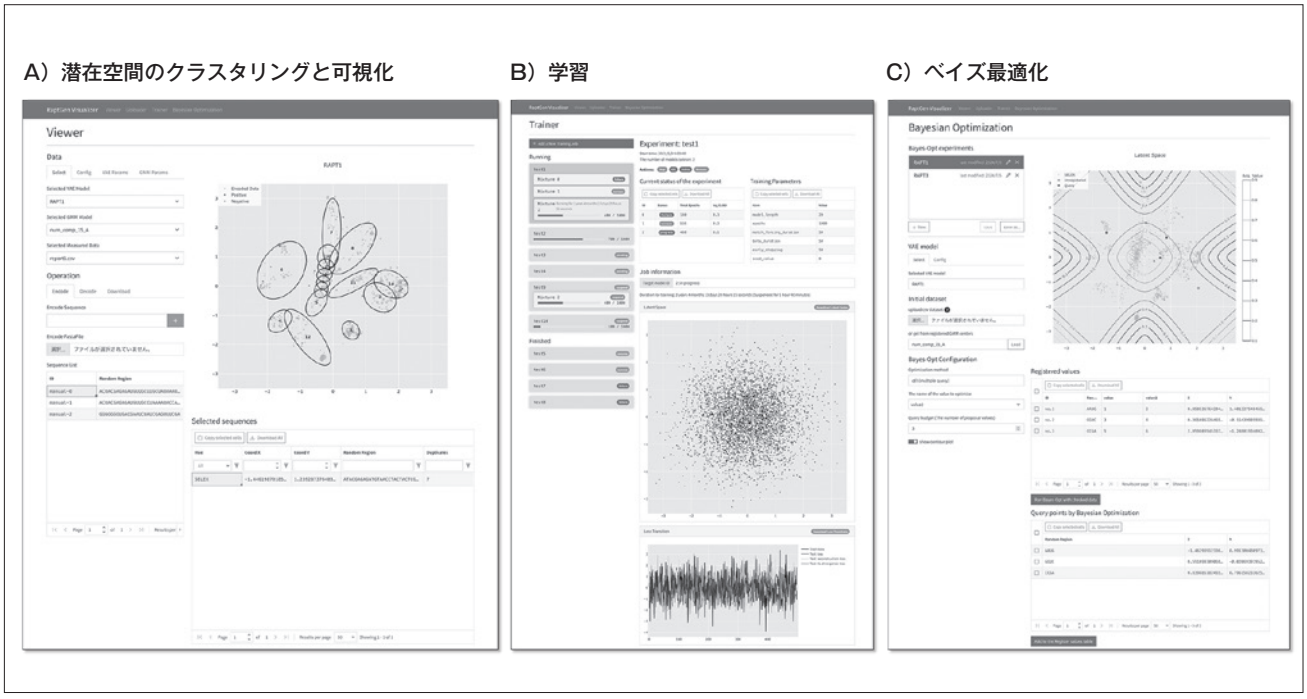


図5　RaptGen-UI のスクリーンショット（公開・論文準備中）
RaptGen の主要機能である（A）潜在空間のクラスタリングと可視化・配列生成、（B）学習、（C）ベイズ最適化をウェブブラウザを通して実行することが可能となる。

学習、ベイズ最適化を、ウェブブラウザを通して実行することが可能となる。

以上をまとめると、RaptGen を用いて実験的アプローチと計算科学的アプローチを融合させることで、アプタマー創薬の新たな展開が期待される。

3-4. RfamGen：機能性 RNA ファミリーの配列をデザインするための深層学習モデル

RfamGen（RNA family sequence generator）[19]は、機能性RNA ファミリーの配列設計のための深層生成モデルである。機能性RNA は、タンパク質に翻訳されずに RNA 自身が活性を有する RNA の総称であり、近年数多くの機能性RNA が発見され注目を集めている。機能性RNA の機能には、その配列と構造が重要となることが知られており、例えば、Rfam データベース[20]には、配列と構造が共通した機能性RNA ファミリーが多数登録されている。

RfamGen は、変分オートエンコーダー（VAE）と RNA ファミリーの配列・構造の確率モデルである共分散モデル（CM）のアーキテクチャを組み合わせることで、特定の RNA ファミリーに含まれる RNA 配列の効率的な生成を実現している（**図4B**）。RfamGen の学習プロセスでは、天然配列の共通配列・構造を模倣しつつ、それらを潜在空間に埋め込むように訓練がされる。このために、CM 上のアラインメントの特徴をベクトル化し、VAE を用いて CM 上のパスを生成する。具体的には、パスを遷移規則と出力規則に分解し、one-hot 表現に変換する。さらに、配列の生成は、再構築された CM から最大確率（argmax）を取ることで行われる。この方法により、CM の構造的制約を維持しつつ、新規の配列を生成することが可能となる。

RfamGen の性能評価は、18種類の RNA 配列ファミリーを用いて行われた。その結果、RfamGen は他のモデル（GCVAE、GVAE、CVAE）と比較して、より高い生成能力を示した。特に、構造アラインメント情報の組み込みが性能向上に大きく寄与していることが明らかになった。また、学習データサイズの影響を調査したところ、RfamGen は約500の入力配列でも高いパフォーマンスを達成できることが示された。さらに、RfamGen が学習する潜在空間は、系統学的情報や配列の活性情報を反映しており、目的に応じた配列設計に有用である。例えば、tRNA の活性や twister-sister リボザイムの系統距離が、潜在空間内で表現されることが示された。さらに実験的検証では、RfamGen が多様な RNA 配列ファ

ミリーの機能的な配列を生成できることが確認された。特に、glmS リボザイムを対象とした大規模アッセイでは、RfamGen により生成された配列の多くが機能的であり、一部は天然配列を上回る活性を示すことが明らかになった。また、RfamGen は、変異効果の予測にも応用可能である。変異効果の網羅的な解析を行うことができる deep mutational scanning データを用いた評価では、RfamGen が既存の手法と同等以上の予測精度を達成することが示された。

以上をまとめると、RfamGen は RNA 配列設計の分野に新たな可能性をもたらす強力なツールである。その高いデータ効率性と優れた生成能力は、RNA 工学や合成生物学の発展に大きく貢献すると期待される。今後、さらなる改良と応用研究により、バイオテクノロジーや医療分野における革新的なアプリケーションの開発が加速されることが期待される。

4. 今後の展望

RNA デザインの分野は、近年急速に発展しており、新たな技術や手法が次々と登場している。これらの革新的なアプローチにより、より効率的で精密な RNA 分子の設計が可能になりつつある。本節では、最新の技術動向を概観し、今後の展望について論じる。

4-1. 最新の深層学習技術の活用

機械学習、特に深層学習技術の RNA デザインへの応用は、近年最も注目されている。例えば、RaptGen や RfamGen などの変分オートエンコーダー（VAE）ベースのモデルは、HT-SELEX データから効率的にアプタマーを生成したり、既知の機能性 RNA ファミリーに保存された配列・構造情報から、新たな RNA ファミリーの配列を設計したりすることを可能にした。これらのモデルは、潜在空間上に埋め込まれた配列空間を連続的に探索し、新規の機能性 RNA 配列を発見する能力をもつ。

また、転移学習を活用した手法も登場している。事前学習済みの言語モデルを RNA デザインタスクに適用することで、少量のデータでも高性能なモデルを構築できるようになってきた。この技術により、希少な RNA ファミリーや新規の RNA 機能の設計にも対応できる可能性が開かれた。

4-2. RNA 立体構造などの情報の活用

RNA の配列、構造、機能の関係性をより精密に捉えるため、多様なデータソースと計算手法を統合するアプローチが進展している。例えば、二次構造予測、三次元構造シミュレーション、進化的情報、実験データなどを組み合わせた統合モデルの開発が進んでいる。これにより、RNA の複雑な機能メカニズムをより正確に予測し、目的に応じた設計が可能になると考えられる。特に、RNA の立体構造予測研究がさらに進み、RNA デザイン研究に統合されていくことが期待される。

4-3. ハイスループット実験技術との融合

計算機的手法の発展と並行して、高速・大規模な実験技術も進化を続けている。深層学習などの最先端 AI 手法を用いたデータ駆動型アプローチによる RNA デザインには、高品質な学習データが極めて重要となる。そのため、ハイスループットな実験系の開発も重要となる。マイクロ流体デバイスを用いた並列実験系や、次世代シーケンサーや質量分析器を活用した大規模スクリーニング技術などが、RNA デザインの効率を飛躍的に向上させることが期待される。

4-4. RNA- タンパク質相互作用の精密制御

機能性の RNA 分子は他の分子、特に、タンパク質と相互作用して機能を発揮するものが多い。したがって、RNA-タンパク質相互作用を精密に制御する RNA の設計は、今後さらに重要性を増すと予想される。eCLIP などの大規模実験技術により、200 近い RNA 結合タンパク質（RBP）とその結合配列が網羅的に測定されており、RBP と結合配列・構造、さらには機能との関連性の研究が進んでいる[21]。今後は、これらの実験データを学習データとして用いるなどして、目的の RBP に結合する RNA を設計する技術の発展が期待される。

4-5. RNA修飾の制御と活用

網羅的にRNA修飾を解析するエピトランスクリプトーム(epi-transcriptome)などの研究により、RNA修飾が生物学的機能に与える影響の理解が深まるにつれ、修飾を考慮したRNAデザインの重要性が増している[22]。例えば、mRNAワクチンの開発において、修飾塩基の最適なデザインが課題となっている。機械学習を用いて修飾パターンと機能の関係を予測し、目的に応じた修飾プロファイルをもつRNAを設計する技術の発展が期待される。また、修飾がRNAの構造に与える影響の研究なども重要になってきている。例えば、RNA逆フォールディングにおいては、修飾塩基を含むRNA配列の二次構造予測が必要となる。しかしながら、修飾を含む塩基に対する構造予測パラメタ(エネルギーパラメタ)は未だ不完全である。このようなパラメタの決定には多大な実験コストが必要であるが、これを緩和するために、実験およびシミュレーション計算を組み合わせてパラメタを決定する手法の研究などが進んでおり、今後の発展が期待される[23]。

4-6. 環境応答性RNAの設計

環境に応じて構造や機能を変化させるRNAの設計は、合成生物学や医療応用において重要な課題である。リボスイッチやアプタザイムなどの設計において、複数の構造状態間の遷移を精密に制御する技術の開発が進んでいる。機械学習を用いて配列－構造－機能の関係性をモデル化し、目的の環境応答性をもつRNAを設計する手法の発展が期待される。

4-7. 大規模言語モデルの応用

近年、GPT-3やBERTなどの大規模言語モデル(Large language model, LLM)が、バイオインフォマティクス分野にも応用され始めている。それに伴い、RNAに特化したRNA言語モデルも急速に研究が進んできている(**表3**)。**表3**に示すとおり、これらのモデルは、RNAcentral[24]やRfam[20]データベースに含まれる大量のRNA配列で事前学習を行ったモデルであり、さまざまな下流タスクに応用され有効性が確認されている。これらの言語モデルをRNAデザインに応用することで、高品質な少数の学習データを効率的に活用したデザインが可能となることが期待される。さらに、自然言語での指示に基づいて機能性RNAを設計したり、既存のRNAデザインの改善案を提案したりすることが可能になると期待される(タンパク質に関してはProGenなどの類似モデルがすでに存在している[25])。さらに、これらのモデルがもつ幅広い事前学習知識を有効に活用することで、異分野の知見を組み合わせた革新的なRNAデザインが実現する可能性もある。

4-8. 量子コンピューティングの活用

量子コンピューティング技術の発展に伴い、RNAデザイン分野への応用も検討され始めている[26, 27]。本稿で言及した、RNA逆フォールディング問題やRNAアプタマーデザインなどは、組み合わせ最適化問題の一種となり、量子アニーリングなどの量子コンピューティング手法が有効である可能性がある。実際、筆者を含む研究グループは、現在、NEDOの「量子・古典ハイブリッド技術のサイバー・フィジカル開発事業」において、量子コンピューティング技術とAI技術を組み合わせたハイブリッドアプローチによるRNAアプタマーのデザイン手法の研究開発を進めている。

4-9. 倫理的・社会的課題への対応

RNAデザイン技術の発展に伴い、倫理的・社会的課題への対応も重要性を増している。例えば、人工設計されたRNAの生態系への影響や、遺伝子治療における安全性の問題などが挙げられる。また、SARS-CoV-2などのRNAウイルスの配列をデザインするなどのデュアル・ユース性が懸念される研究(dual use of concern)にはさらに慎重な検討が必要である。このような課題に対しては、現時点では、統一したルールや基準などが存在せずに、各研究者が個別に判断・対応している状況である。今後、法整備なども含めた統一したルール作りが重要となるだろう。

表3　RNA言語モデル

名前	モデル	事前学習データ（配列数）	事前学習タスク	下流タスク	文献
RNABERT	Transformer	RNACentral［ncRNA］/Rfam（7.6万）	MLM、SAL	構造アラインメント、RNAファミリークラスタリング	48
RNA-FM	Transformer	RNACentral（2370万）	MLM	RNA二次構造予測/RNAコンタクトマップ予測/RNA 3D構造予測/ RNA-タンパク質相互作用予測/MRL予測	49
RNA-MSM	Transformer	Rfam［アラインメント］	MLM	RNA二次構造予測/RNA溶媒接触面積（ASA）予測	50
RiNALMo	Transformer	RNAcentral、Rfam、nt、EnsemblI（3600万）	MLM	RNA2次構造予測/MRL予測	51
GeneRNA	GPT2	RNAcentral（2980万）	次トークン予測	RBP結合RNAの生成（ELAVL1とSRSF1）	52
RNADiffusion	拡散モデル	RNAcentral［ncRNA］（110万）	入力配列再構築	ncRNA配列生成/5’ UTR配列生成と最適化	53
RNAErnie	Transformer	RNACentral/ATtRACT/SpliceAid（2300万）	モチーフ予測/RNA種予測	RNA配列分類/RNA-RNA相互作用予測/RNAの二次構造予測	54
ERNIE-RNA	Transformer（塩基対制約）	RNAcentral（2040万）	MLM	RNAの二次構造予測/RNAの3D近接性（コンタクトマップ）予測/MRL予測/RNA-タンパク質結合予測	55
Uni-RNA	Transformer	RNAcentral/Genome Warehouse（10億）	MLM	RNA二次構造予測/RNAコンタクトマップ予測/MRL予測/スプライスサイト予測/ncRNAの機能分類/RNA修飾部位の予測	56
Shulginaモデル	Transformer	GARNET（75万）	次トークン予測	高温耐性を付与するRNA変異の予測	57
RfamLlama	Llama	Rfam 14.10（68万）	次トークン予測	ゼロショット変異効果予測	58
3UTRBERT	Transformer	GENCODE（2万）	MLM	RNA結合タンパク質の結合部位予測/m6A RNA修飾部位の予測/mRNAの細胞内局在予測	59
UTR-LM	Transformer	Ensembl, Sample et al.\cite {sample2019human}（230万）	MLM、MFE予測、SS予測	MRLの予測/TEの予測/mRNA発現量の予測/IRESの同定	60
CodonBERT	Transformer	NCBI（1000万）	MLM、相同配列予測	タンパク質発現予測/mRNA安定性予測/mRNA分解予測	61

MLM：Masked Language Model、SAL：Structural Alignment Learning、SS：Secondary Structure、MFE：Minimum Free Energy。

5. まとめ

以上のように、RNAデザインの分野は多岐にわたる技術革新、特に、ハイスループットな実験技術とAIなどの情報技術の革新によって急速に発展しており、今後もさらなる進化が期待される。これにより精密で効率的なRNAデザインが可能となり、医療、創薬、バイオテクノロジー、環境科学など、幅広い分野でのイノベーションが加速されると考えられる。RNAデザインは究極の学際分野であり、さまざまな研究領域に関する知見が必要となる。タンパク質デザイン研究に比べると、RNAデザイン研究は遅れをとっているが、本稿を読んで、RNAデザイン分野にさまざまなバックグラウンドを有する新進気鋭の若い学生や研究者が参入することを切に期待したい。

謝辞

本稿で紹介をしたRaptGenは、株式会社リボミックとの長年にわたる共同研究の成果である。同じくRfamGenは、東京大学の角 俊輔博士と齊藤博英教授との共同研究の成果である。また、mRNAデザインに関しては、名古屋大学 阿部 洋研究室との共同研究プロジェクトにおける議論が執筆において有益であった。以上の共同研究者の皆様、ならびに、

日頃から多くの議論を行っている、早稲田大学理工学術院浜田道昭研究室、および、産業技術総合研究所生体システムビッグデータ解析オープンイノベーションラボラトリの皆様に感謝を申し上げたい。

文献

1) 金井昭夫, RNA の科学：時代を拓く生体分子, 朝倉書店 (2024)
2) 中川真一ら, 情報からマテリアルへ：ノンコーディング RNA研究：機能分子としての RNA を見つけ、知り、創薬に使う新時代, Number増刊, **42**, 15, 実験医学, 羊土社 (2024)
3) M. Metkar, *et al.*, *Nat. Rev. Drug Discov.*, **23**, 67–83 (2024)
4) K. Sato, *et al.*, *Brief. Bioinform.*, **24**, bbad186 (2023)
5) S. Zhang, *et al.*, *Biophys. J.*, **123**, 2647–2657 (2024)
6) A. Churkin, *et al.*, *Brief. Bioinform.*, **19**, 350–358 (2018)
7) S. Mukherjee, *et al.*, *Brief. Bioinform.*, **24**, bbad110 (2023)
8) M.J. Wu, *et al.*, *ACS Synth. Biol.*, **8**, 1838–1846 (2019)
9) T. Fukunaga, *et al.*, *Brief. Funct. Genomics*, **18**, 182–191 (2018)
10) P. J. Sample, *et al.*, *Nat. Biotechnol.*, **37**, 803–809 (2019)
11) H.K. Wayment-Steele, *et al.*, *Nucleic Acids Res.*, **49**, 10604–10617 (2021)
12) K. Leppek, *et al.*, *Nat. Commun.*, **13**, 1536 (2022)
13) Q. Wu, *et al.*, *Elife*, **8**, e45396 (2019)
14) D.Varshney, *et al.*, *Nat. Rev. Mol. Cell Biol.*, **21**, 459–474 (2020)
15) H. Gong, *et al.*, *Brief Bioinform*, **24**, bbad001 (2023)
16) N. Iwano, *et al.*, *Nat. Comput. Sci.*, **2**, 378–386 (2022)
17) 長野哲雄ら ,次世代医薬とバイオ医療, 東京化学同人 (2022)
18) C. Tuerk, *et al.*, *Science*, **249**, 505–510 (1990)
19) S. Sumi, *et al.*, *Nat. Methods*, **21**, 435–443 (2024)
20) I. Kalvari, *et al.*, *Nucleic Acids Res.*, **49**, D192–D200 (2021)
21) E.L. Van Nostrand, *et al.*, *Nature*, **583**, 711–719 (2020)
22) D. Wiener, *et al.*, *Nat. Rev. Genet.*, **22**, 119–131 (2021)
23) S. Sakuraba, *et al.*, *J. Chem. Theory Comput.*, **16**, 5923–5935 (2020)
24) RNAcentral Consortium, Nucleic Acids Res., **49**, D212–D220 (2021)
25) A. Madani, *et al.*, *Nat. Biotechnol.*, **41**, 1099–1106 (2023)
26) M.W. Lewis, *et al.*, *J. Heuristics*, **27**, 695–717 (2021)
27) D. Alevras, *et al.*, *mrna secondary structure prediction using utility-scale quantum computers* (2024)
28) R. Lorenz, *et al.*, *Algorithms Mol. Biol.*, **6**, 26 (2011)
29) A. Busch, *et al.*, *Bioinformatics*, **22**, 1823–1831 (2006)
30) A.T. Modena, *Adv. Appl. Bioinform.Chem.*, **4**, 1–12 (2011)
31) R. Kleinkauf, *et al.*, *Bioinformatics*, **31**, 3114–3121 (2015)
32) X. Yang, *et al.*, *BMC Bioinform.*, **18**, 468 (2017)
33) J. Shi, *et al.*, *arXiv preprint arXiv*, 1803.03146 (2018)
34) F. Runge, *et al.*, *International Conference on Learning Representations* (2019)
35) T. Zhou, *et al.*, *Bioinformatics*, 39(Supplement 1), i563–i571 (2023)
36) H. Huang, *et al.*, *Bioinformatics*, *Advance article* (2024)
37) Á.R.-Largo, *et al.*, *IEEE Access*, **12**, 15582–15593 (2024)
38) C.K. Joshi, *et al.*, *arXiv preprint arXiv*, 2305.14749 (2024)
39) G. Terai, *et al.*, *Bioinformatics*, **32**, 828–834 (2016)
40) M. Diez, *et al.*, *Sci. Rep.*, **12**, 12126 (2022)
41) H. Zhang, *et al.*, *Nature*, **621**, 396–402 (2023)
42) X.H.-Alias, *et al.*, *Genome Biology*, **24**, 34 (2023)
43) H. Zhang, *et al.*, *arXiv preprint arXiv*, 2408.09048 (2024)
44) X. Tang, *et al.*, *Acta Pharm. Sin. B*, **14**, 1814–1826 (2024)
45) K. Yamada, *et al.*, *Multi-objective computational optimization of human* 5' utr sequences. (2024)
46) Y. Li, *et al.*, *bioRxiv*, (2024) https://doi.org/10.1101/2024.09.06.611590
47) H. Zhang, *et al.*, *bioRxiv*, 2024-06 (2024) https://doi.org/10.1101/2024.06.20.599727
48) M. Akiyama, *et al.*, *NAR Genom Bioinform.*, **4**, lqac012, (2022)
49) J. Chen, *et al.*, *Interpretable rna foundation model from unannotated data for highly accurate rna structure and function predictions* (2022)
50) Y. Zhang, *et al.*, *Nucleic Acids Res.*, **52**, e3 (2024)
51) R.J. Penić, *et al.*, *General-purpose rna language models can generalize well on structure prediction tasks* (2024)
52) Y. Zhao, *et al.*, *bioRxiv*. (2024) https://doi.org/10.1371/journal.pone.0310814
53) K. Huang, *et al.*, *arXiv preprint arXiv*, 2409.09828 (2024)
54) N. Wang, *et al.*, *Nature Machine Intelligence*, **6**, 548–557 (2024)
55) W. Yin, *et al.*, *bioRxiv*. (2024) https://doi.org/10.1101/2024.03.17.585376
56) X. Wang, *et al.*, *bioRxiv*. (2023) https://doi.org/10.1101/2023.07.11.548588
57) Y. Shulgina, *et al.*, *bioRxiv*, (2024) https://doi.org/10.1101/2024.04.05.588317
58) J. Sun, *et al.*, *Proceedings of the ICML* 2024 Workshop on Accessible and Efficient Foundation Models for Biological Discovery. (2024)
59) Y. Yang, *et al.*, *Adv. Sci.*(*Weinh*), 2407013 (2024)
60) Y. Chu, *et al.*, *Nat. Mach. Intell.*, **6**, 449–460 (2024)
61) Z. Ren, *et al.*, *Bioinformatics*, **40**, btae330 (2024)