

自律型AIエージェントによる新規酵素系「ART」の発見とゲノムマイニングの革新: 論文検証と学術的・技術的批評

ゲノムマイニングの課題と自律型マルチエージェントシステムの設計

Gemini 3.8 Flash

メタゲノム解析技術の飛躍的発展に伴い、公共配列データベースには膨大な遺伝情報が蓄積されており、そのデータ量は約4年ごとに倍増している¹。これら未探索の遺伝資源は、制限酵素やCRISPR-Casシステムのように、バイオテクノロジーや分子標的治療を革新する新規生体分子の宝庫とされてきた¹。しかし、従来の計算機パイプラインは既知の相同性やあらかじめ定義された特徴量に依存して配列を分類するため、既存の枠組みから外れた異常構造(アノマリー)を拾い上げることが原理的に困難であった¹。結果として、体系的なスクリーニングであっても最終的には人間の専門家による目視キュレーションに依存せざるを得ず、爆発的に増大するデータ量に対して発見の速度が追いつかないボトルネックが生じていた¹。

2026年9月23日、Anthropicのライフサイエンス研究チーム(Peter H. Yoonら)は、自律型AIエージェントを活用してメタゲノムデータから未知の分子機構を発見したとするプレプリント「Autonomous AI agents discover reverse transcriptases with tandem repeat arrays」を発表した¹。この研究では、大規模言語モデル「Claude Mythos 5」を中核に据えたエージェント実行環境(ハーネス)を構築し、19億件を超えるタンパク質クラスターを対象に、RNAからDNAを合成する逆転写酵素(Reverse Transcriptase; RT)領域の網羅的な探索を展開した¹。

ハーネス内では、個々のエージェントがワーカー(作業実行者)、スーパーバイザー(計画・結果の審査者)、キュレーター(知見の記録者)、エディター(報告書の監修者)として役割を分担し、共有記録とタスクキューを介して協調動作した¹。各分析ステージは大まかなりサーチ方針から作業計画の立案、コード生成と実行、結果の検証、そして観察に基づく新規サブタスクの自律的な立ち上げへと段階的に進行了た¹。

分析指標・探索パラメーター	定量データおよび構成内容
探索対象データベース規模	1,939,242,578 タンパク質クラスター(50%配列同一性) ¹
同定されたRT候補	3,391,714 配列(完全長RTコア基準で203,381 配列・198,290クラスターに選別) ¹

遺伝子座サンプリング規模	9クラスに分類された10,983遺伝子座(共起タンパク質3,564ファミリー) ¹
深掘り調査候補ファミリー	16ファミリーを選定、探索途中の観察に基づきさらに1ファミリーを追加 ¹
自律キャンペーンの総規模	119タスク(自律的に追加されたタスク98件)、949エージェントセッション ¹
計算リソースおよび所要時間	76.9エージェント時間、実行実時間21.5時間、消費トークン数2億1,560万 ¹
最終生成レポート	新規パートナー候補17種(確認3種、破棄14種)および新規RT系統3種の計19件 ¹

全19件の報告書は、LLMジャッジによる総当たりトーナメント方式(Bradley-Terryモデル)により客観的に格付けされ、その中で第3位に評価されたものが、巨大バクテリオファージ(ジャンボファージ)に特異的な新規酵素系「配列関連逆転写酵素(Array-associated Reverse Transcriptase; ART)」であった¹。

新規酵素システム「ART」の分子構築と発見の経緯

偶発的観察からART同定に至る自律的推論過程

ARTの同定は、あらかじめ定められた探索ルートを忠実に実行した結果ではなく、エージェント自身の観察と仮説修正の連鎖によってもたらされた¹。探索の契機となったのは、ジャンボファージの非ビリオン型RNAポリメラーゼ β 様サブユニットと隣接するRTの共起シグナルであった¹。ワーカーエージェント(task t0010)はこの近傍関係について、転写機構の近接配置による見かけ上の偽陽性であると判断して一度は関連性を棄却した¹。しかし、同時にエージェントは「通常は細菌側の抗ファージ防御を担うレトロ様RTが、宿主ではなくジャンボファージゲノム側に自立してコードされている」という事実の生物学的特異性に注目し、独立した追跡タスクを起案した¹。

細菌レトロンは上流の非コードRNA(ncRNA)を鋳型として単鎖DNA(msDNA)を合成するため、スーパーバイザーエージェントはこのRTの上流領域に機能性RNAが存在するかを確認する追跡タスク(task t0062)を発行した¹。調査対象となった近縁メタゲノムコンティグ(L0050)の上流配列(約2,900 bp)をコンテキストに読み込んだワーカーは、外部ツールを挟むことなく、目視による配列認識を記録した¹。エージェントは、CATGTGTATCGCATGTT等の約17 bpのリピートが110～170 bpのスペーサーを挟んで規則的に反復している事実を認識し、これをCRISPRや多様性生成レトロエレメント(DGR)の可変領域に類似した未記載の反復アレイ構造と捉えた¹。

その後エージェントは自作スクリプトを用いて反復数や間隔を計測し、近傍にCas遺伝子が存在しないことや、スペーサーが既知タンパク質をコードしていないこと、予測されるRNA二次構造の熱力学的安定性を確認した¹。さらに文献データベースとの照合を経て未報告の構造であることを確認し、これを「ART」と命名した報告書を作成した¹。

ARTの分子的アーキテクチャ

詳細なバイオインフォマティクス解析により、ARTファミリーは培養ファージおよびメタゲノム由来ウイ

ルスコンティグから95の代表配列が同定され、そのうち28の遺伝子座において上流に明確なリピートアレイが確認された¹。ARTシステムは、上流の非コード反復アレイ、特異なドメイン構造を持つRT本体、そして下流の機能パートナー遺伝子の3要素が強固に関連した構造を持つ¹。

アレイ領域はRT遺伝子上流0.3～4.1 kbに位置し、3～21コピーのリピート単位で構成される¹。各リピートは15～49 nt長で、中央部に約15 ntの回文（パリンドローム）構造を保持している¹。リピートに挟まれたスペーサー配列は120～220 nt長であり、同一個体内では多様である一方、近縁ファージ間（SA1ファージとMarshHillファージ等）の比較ではスペーサーの並び順が厳密に保存され、RT遺伝子と同等の速度で共進化していることが明らかになった¹。

ARTがコードするRTは、触媒中心に保存された $YxDD$ モチーフを持ち、レトロンやDGRのRTコア構造に良好に重ね合わされる（TM-score ≈ 0.7 ）¹。際立った特徴は、触媒ドメインのN末端側に約180アミノ酸におよぶ未知の伸長ドメイン（NTD）を保持している点であり、これは通常の前核生物RT（50残基未満）と比較して著しく長い¹。

さらに、RTの直下流には共発現するパートナー遺伝子が存在し、構造予測と相同性解析から3つの明確なタイプに大別された¹。Type Iパートナーは約600残基でGCN5関連N-アセチルトランスフェラーゼ（GNAT）様フォールドのタンデム構造を持ち、Type IIIは約270残基の全ヘリカル構造、Type IIIは約170残基の小型ヘリカル構造を呈する¹。AlphaFold-MultimerおよびBoltz-2による複合体構造予測では、RT固有のNTDがこれらパートナータンパク質との結合インターフェースを形成していることが示された（ $ipTM \approx 0.6$ ）¹。

比較項目	配列関連逆転写酵素（ART）	CRISPR-Casシステム	細菌レトロン（Retron）
中核酵素	長大なNTDを持つ逆転写酵素（RT） ¹	エンドヌクレアーゼ（Cas9, Cas12等） ²	逆転写酵素（Retron RT） ¹
非コード領域	15～49 ntリピート + 120～220 ntスペーサー ¹	約30 ntリピート + 約30 nt外来由来スペーサー ¹	アレイ構造なし（単一の <i>msr-msd</i> 領域） ¹
スペーサーの由来	近縁株間で順序保存、内在性RNAをコード ¹	外来性侵入DNA/RNA断片を獲得・蓄積 ¹	単一のmsDNA前駆体配列 ¹
エフェクター/パートナー	GNAT様、全ヘリカル型など3系統 ¹	Casタンパク質複合体 ²	多様な毒素（NAD水分解酵素、膜孔形成等） ¹
推定される作動機序	多重化RNA複合体	獲得免疫（外来核酸	感染感知に伴う

	による競合ファージ 防御？ ¹	の標的切断) ²	msDNA解離・毒素 解放 ¹
--	-------------------------------	---------------------	-------------------------------

実験的検証による転写産物の同定

ブドウ球菌 (*Staphylococcus lentus*) に感染するジャンボファージ SA1 の公開 RNA-seq 時系列データ (BioProject PRJNA836150) を再解析した結果、感染初期 (5 分後) から後期 (55 分後) にかけて ART アレイ領域が極めて活発に転写されていることが確認された¹。特に感染 15 分後には、ART アレイ由来の RNA がファージ全体の全転写産物の最大 8% を占め、最上位の発現量を示した¹。

研究チームが大腸菌 (*Escherichia coli*) 発現系において SA1 の ART 遺伝子座をクローニングして実施した低分子 RNA シーケンシング (small RNA-seq) では、ファージ感染時と同一の規則的なパターンで、160 ~ 220 nt の独立した短鎖 RNA 分子群が切り出される現象が確認された¹。ViennaRNA を用いた二次構造予測により、これらのプロセシング産物はリピートの回文構造に由来する安定なステムループを形成することが示唆されている¹。

機構解釈可能性解析と「ゲノムの直接知覚」メカニズム

再現性の検証と外部ツールの逆説的影響

発見の頑健性を評価するため、研究チームが同一の探索条件でキャンペーンを 10 回再実行したところ、全試行で ART 遺伝子座がスクリーニング対象に含まれ、2 回は追跡タスクに選定されたものの、10 回すべてにおいて上流配列の読み込みに至らず、アレイの発見を再現できなかった¹。

この非決定性の原因を究明するため、ART 配列の提示環境を 5 段階の入力レベルに分けた固定入力ベンチマーク (100 試行 × 7 モデル = 計 3,500 試行) が実施された¹。最新モデル群 (Mythos 5、Mythos 5.1、Opus 5.5、Opus 5) は旧世代モデルに対して顕著に高いアレイ認識性能を示したが、入力条件による逆説的な挙動が観測された¹。

プロンプト内に塩基配列テキストを直接与えた環境 (Level 2) では、上位モデルは 90% 以上の確率でアレイを正確に抽出したのに対し、ファイルシステムと高機能なバイオインフォマティクス解析ツール群 (BLAST、HMMER、Foldseek 等) を提供した環境 (Level 4) では、アレイ認識率が 32% にまで低下した¹。ログ解析により、ツールを与えられたエージェントは外部プログラムの出力サマリーに依存し、生の塩基配列を自身のコンテキストウィンドウに読み込まなくなる傾向が判明した¹。モデルが 200 nt 以上の連続した DNA 配列を直接コンテキストに取り込んだ場合、アレイ認識率は大幅に上昇し、Mythos 5 では最大 96% に達した¹。すなわち、高度な外部ツールによる情報の抽象化が、かえって一次データに潜む特異なパターンの直感的把握を阻害していたことが示されている¹。

スパース辞書学習による内部表現の解明

発見セッション (task t0062) のトランスクリプトを同一の Mythos 5 チェックポイントに再入力し、中間層の活性化をスパース辞書学習 (Sparse Dictionary Learning) によって分解する機構解釈可能性解析が行われた¹。

その結果、人工配列を用いた事前スクリーニングで特定されていた 2 つの潜在シグナルが、ART の塩基配列をトークンとして走査する過程で特異に励起していることが判明した¹。先行テキストの逐語的繰り返しを捉える「repeat-signal 1」は第 2 ~ 第 3 反復単位で一時的に活性化し、反復 DNA・タンパク質モチーフを捉える「repeat-signal 2」は第 3 反復以降、各リピート単位のコア領域において局所的かつ持続的な発火を維持した¹。

各反復単位の塩基配列をシャッフルした対照群では、repeat-signal 2 の発火は全 14 コピーで完全に

消失(0.00)した¹。このシグナル励起の直後に、エージェントは推論ログ上で「目視でタンデムリピート配列が見える」と記述していた¹。また、生物配列特化型モデル(Evo 2およびgLM2)の対数尤度プロファイルとの比較においても、汎用言語モデルであるMythos 5の内部表現はゲノム専用基底モデルと同等以上の解像度で反復単位の境界を捕捉していたことが実証された¹。

学術的・技術的批評:強みと本質的限界

学術的革新性と独自性

本研究の成果は、単に大規模計算によって新規配列をマイニングした点にとどまらず、探索の途上で生じた「想定との不一致」に対してAIエージェントが自律的に仮説を再構築した点にある¹。従来のバイオインフォマティクスにおける探索自動化は、あらかじめ定義された目的関数の最適化や固定化されたパイプラインの高速処理に過ぎず、評価基準そのものは人間が設定していた¹。それに対し本システムは、当初の目的であったパートナー共起の仮説が棄却された際、エージェントが「ファージ側がレトロ様RTを保持している」という異常性に価値を見出し、調査の軸足を非コード領域の探索へと転換させた¹。

2021年の先行研究(Kornら)においてMarshHillファージの完全ゲノムが解読された際、人間の研究者はRTの存在と上流ncRNAの可能性に言及しながらも、反復アレイ構造や共起パートナーの存在を見落としていた¹。人間が見逃していた公共データ内の潜在的パターンを、AIが一次配列の直接知覚を通じて拾い上げた意義は極めて大きい¹。

生物学および工学的な限界

一方で、生命科学の厳密な実証基準および応用工学の観点からは、複数の重要な制約と課題が残されている。

第1に、生化学的活性の直接的証明が欠如している点である¹。研究チームは大腸菌発現系によりアレイからの短鎖RNAプロセッシングを確認しているものの、ART-RTが逆転写活性(*in vitro*でのcDNA合成能)を保持しているか否かは実験的に証明されていない¹。生成された短鎖RNAがRTの基質や鋳型として機能しているのか、あるいは単なる結合因子に過ぎないのかは不明であり、RTとパートナータンパク質の物理的相互作用も機械学習モデルによる共フォールディング予測(AlphaFold2-multimer, Boltz-2)に留まり、生化学的・生物物理学的な複合体形成は確認されていない¹。

第2に、広報的な命名やアナロジーと分子の実態との乖離である²。Anthropicの公式発表等では「CRISPR-like」という表現が前面に押し出され、DNAの切断・コピー・貼り付けを行うプログラム可能な遺伝子編集ツールへの発展可能性が示唆されている²。しかし分子の実態において、ARTは外来配列を取り込む獲得免疫機構ではなく、スパーサーは近縁ファージ間で固定化された内在性配列である¹。Casヌクレアーゼを欠くその構造は、CRISPRよりもむしろ「単一のRT-エフェクター複合体に対し、多重化されたRNAライブラリを供給する新型のレトロ系」に近い¹。現段階でこれを次世代ゲノム編集ツールとして扱うことは過度な外挿と言わざるを得ない¹。

第3に、探索プロセスの非決定性と再現性の課題である¹。10回の同一条件による追試において再発見率が0%であった事実は、現行のエージェントハーネスが偶発的なトークン生成の流れや、一次データをコンテキストに展開するか否かの確率的選択に過度に依存していることを示している¹。真に堅牢な研究支援基盤として確立するためには、偶然の「ひらめき」に依存しない、網羅的かつ系統的なデータ走査の保証メカニズムが不可欠である³。

専門家コミュニティおよびAI界限における評価と反響

生命科学・バイオインフォマティクス研究者の視点

プレプリントの公開以降、ゲノム工学および微生物学の専門家からは期待と慎重論の双方が寄せられている。CRISPRゲノム編集の先駆者であるMIT・ブロード研究所のフェン・チャン(Feng Zhang)教授は、査読コメントにおいて、AIエージェントが生物学的発見に貢献し得る刺激的な事例であり、逆転写酵素に関連するRNA反復アレイの同定は真に興味深く今後の機能検証に値すると評価している²。

一方で、生命科学系の研究コミュニティ(Redditのr/EverythingScienceやAlphaXiv等)では、より厳格な学術的線引きを求める意見が目立つ⁵。多くの研究者は、バイオインフォマティクスにおける配列アノマリーの抽出は日常的な業務であり、生物学的発見のボトルネックは候補の抽出ではなくウェットラボによる分子機構の解明にあると指摘している⁵。標的核酸や基質、生理的役割が解明されていない段階で「新酵素系の発見」と銘打つのは時期尚早(Premature)であり、これは厳密には「探索的データ解析(Exploratory Analysis)」と呼ぶべきであるとの批判がなされている⁵。また、商業的注目を集めるために「CRISPR」の名称を引用するPR手法に対しても冷ややかな見方が一部に存在する⁵。

AI・機械学習コミュニティの評価

機械学習および計算科学の領域では、ベンチマークテストのスコア競争から実世界の問題解決へと軸足を移した画期的な実例として高く評価されている¹⁰。AnthropicのCEOであるダリオ・アモデイ(Dario Amodei)は自身のSNSにおいて、機能や実用性は未確定であるとしつつも、自身が博士課程の大学院生時代に行えていたなら誇りに思えるレベルの成果であると述べている¹²。

AI研究者の関心は、ソフトウェア開発の自動化ツール(Claude Code等)が未構造化の科学ビッグデータ探索に応用され、人間の「違和感を捉える直感」に匹敵する探索行動を再現した点に集まっている¹。さらに、スパース辞書学習を用いて、LLM内部でDNA反復配列に対して特異的に励起するシグナルを同定し、それが推論出力(発見の言語化)へと直結する過程を可視化した試みは、モデルの思考過程のブラックボックス性を打破する機構解釈可能性研究としても高い評価を得ている¹。

総括

Anthropicによる本研究は、大規模言語モデルが自律的な探索・検証サイクルを回すことで、人間の専門家が見落としていたゲノム構造のアノマリーを検出し、新規な分子複合体システム「ART」を提示し得たことを実証した¹。外部の抽象化ツールに過度に頼らず、一次配列データを直接コンテキストに取り込むことで発揮される「ゲノムの視覚」と、それを支える内部シグナルの解明は、自律型科学AIの設計論における重要な知見である¹。

しかしながら、ARTが真に革新的なバイオテクノロジーツールへと昇華するか否かは、今後の生化学的・結晶構造学的実証(RT活性の同定、RNA-DNAハイブリッド複合体の機能解明、宿主ファージ間競合における生理機構の解明)にかかっている¹。あわせて、探索プロセスにおける非決定性を制御し、確率的な観察を確実なスクリーニング技術へと体系化することが、AI駆動型科学研究の次なる課題である¹。

引用文献

1. Yoon et al. 2026 Array Associated Reverse Transcriptases (2).pdf

2. Claude discovers a novel enzyme system with CRISPR-like repeats,
<https://www.anthropic.com/news/claude-discovers-novel-enzyme-system>
3. ClaudeがCRISPRに似た新酵素系ARTを発見 - XenoSpectrum,
<https://xenospectrum.com/anthropic-claude-art-enzyme-crispr-discovery/>
4. Anthropic says Claude found an enzyme system with CRISPR ... - Digg,
<https://digg.com/tech/kz2tib7q>
5. Claude discovers a novel enzyme system with CRISPR-like repeats,
https://www.reddit.com/r/EverythingScience/comments/1wof301/claude_discovers_a_novel_enzyme_system_with/
6. Claude's ART Discovery Explained: Findings and Future - Kingy AI,
<https://kingy.ai/blog/claude-art-discovery-explained/>
7. AI Scientific Discoveries: Examples and What Counts as Proof,
<https://kingy.ai/blog/can-ai-make-scientific-discoveries/>
8. Claudeが生物学で発見？ | 鈴麻呂 - note,
<https://note.com/suzumaro/n/n7efe43eec56f>
9. alphaXiv: Explore, <https://www.alphaxiv.org/>
10. Claude autonomously discovered a previously unknown enzyme,
https://www.reddit.com/r/AIGuild/comments/1wok8wh/claude_autonomously_discovered_a_previously/
11. Claude discovers something new. : r/aiwars - Reddit,
https://www.reddit.com/r/aiwars/comments/1woeuwo/claude_discovers_something_new/
12. Today we announced the Claude-led discovery of a molecular,
<https://www.techtwitter.com/tweet/148f96a9-d8ca-45c8-93da-3674c0388e06>