Journal Club (2025 年 4 月 7 日) まとめ

発表論文：

Brixi, G et al., Genome modeling and design across all domains of life with Evo 2.

bioRxiv [Preprint]. Posted: December 18, 2024.

doi: https://doi.org/10.1101/2024.12.17.628962

研究目的および概要:

真核生物を含む全生物ドメインのゲノム配列から生物学的複雑性を学習し、多様な予測・設計タスクに応用可能な汎用ゲノム基盤モデル「Evo 2」を開発することを目的とした (図 1)。筆者らは、9.3 兆塩基対という大規模ゲノムデータセット「OpenGenome2」を用いて、70 億 (7B) および 400 億 (40B) パラメータ、最大 100 万トークンコンテキストを持つモデルを学習させた。アーキテクチャには長コンテキスト処理に優れた StripedHyena 2 を採用した。Evo 2 は、(1)変異影響予測（特に非コード領域や indel で高い性能）、(2)ゲノムスケールでの配列生成（ミトコンドリア、細菌、酵母など）、(3)推論時探索による制御可能なエピゲノム特性（クロマチンアクセシビリティ）の設計、(4)Sparse Autoencoder を用いたモデル内部表現の解釈、といった多様なタスクで高い能力を示した。本研究の成果として、モデルパラメータ、学習・推論コード、データセットの全てがオープンソースで公開されている。

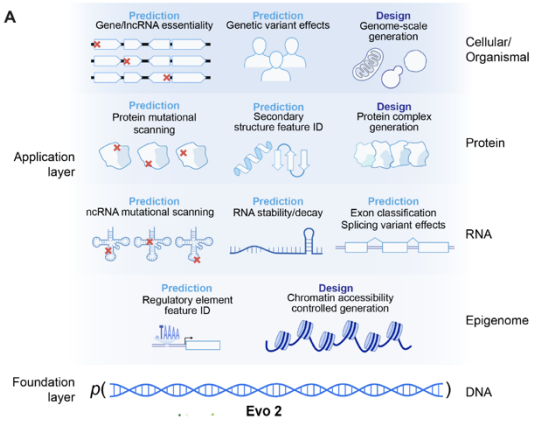

図 1 Evo 2 の全体像

<u>先行研究と比べて何がすごい？ 技術やアプローチのキモはどこ？:</u>

1. **圧倒的なスケールと網羅性:** 学習データ量（9.3 兆 bp）、モデルサイズ（最大 40B）、コンテキスト長（最大 1M トークン）のいずれも既存のゲノムモデルを凌駕する。また、単一モデルで**全生物ドメイン（真核生物を含む）**をカバーする初の試みである。
2. **高い汎用性:** 特定のタスクに特化せず、変異影響予測、ゲノムスケール生成、制御生成、特徴解釈といった多様な下流タスクに適用可能な**「基盤モデル」**である点。
3. **非コード/non-SNV 予測能力:** 特に、これまで評価が難しかった**非コード領域の変異や挿入・欠失（indel）の影響予測**において、最先端の性能（SOTA）を達成した点。
4. **制御可能な配列設計:** 推論時探索と外部予測モデルを組み合わせることで、**クロマチンアクセシビリティのような複雑な生物学的特性を制御しながら DNA 配列を設計**する「生成的エピゲノミクス」の概念実証に成功した点。
5. **技術的基盤:** 長コンテキスト処理と計算効率を両立する新アーキテクチャ **StripedHyena 2** の採用と、それを大規模に学習させるための基盤技術。
6. **完全なオープンソース:** 大規模 AI モデルとしては異例とも言える、モデル・データ・コードの**完全なオープンソース化**を実現し、研究コミュニティへの貢献と透明性を重視している点。

どうやってこの手法/仮説の有効性を検証したのか:

- **変異影響予測:** ClinVar、SpliceVarDB、BRCA1/2 データセットなどを用いて、病原性予測精度を既存手法と比較（ゼロショットおよび教師あり）。Deep Mutational Scanning (DMS) データを用い

てタンパク質・ncRNA の適応度予測性能を評価。遺伝子必須性予測タスクも実施。既存の特化型モデルと同等以上の性能を示した。

- **ゲノム生成:** プロンプト応答による遺伝子配列補完タスクでアミノ酸配列の再現率を評価。ミトコンドリア、細菌 (M. genitalium)、酵母 (S. cerevisiae) のゲノム/染色体スケールでの配列生成を行い、生成物の遺伝子構造、相同性、タンパク質特性（pLDDT、二次構造など）を評価。真核生物の訓練データが大半を占める中、原核生物に対する性能が Evo1 に劣らず、真核生物での性能が大幅に向上した。
- **制御生成 (エピゲノミクス):** Enformer と Borzoi モデルをスコアリング関数として利用したビームサーチアルゴリズムを実装。指定したクロマチンアクセシビリティパターン（ピーク位置・幅、モールス信号）を持つ配列が設計できることを実証し、計算量と設計成功率のスケーリング則を確認。
- **モデル解釈:** Sparse Autoencoder (SAE) を用いてモデルの中間層表現から解釈可能な特徴（プロファージ、ORF、二次構造、TF モチーフ、エクソン/イントロンなど）を抽出し、生物学的アノテーションとの対応を確認。学習データ外の種（マンモス）ゲノムでも特徴が機能することを検証。

この研究をさらに発展させるとしたら:

- **特定タスクへの最適化:**
  - ノンコーディング変異の遺伝子発現への影響解析など、特定の生物学的課題に対するファインチューニング。
  - Few-shot learning や強化学習といった手法を用いた応用展開。
- **ジェノタイプ-フェノタイプ関連解析:** 患者ゲノム情報と臨床情報を統合し、疾患感受性や薬物応答性などを予測するモデルへの応用。
- **モデル解釈の深化と応用:** モデル内部で学習された特徴（Figure 4）と疾患メカニズムを結びつけ、新たな生物学的知見を得る。
- **7B モデルの活用:** 計算資源が比較的少なくても扱える 7B モデルをベースとした、よりアクセスしやすい応用研究の推進（例：LoRA を用いたファインチューニングなど）。
- **オープンソース性の活用:** 公開されたモデル・コード・データを活用した、研究コミュニティ全体での多様な応用と改良。

生物学・医学的課題解決への貢献: モデル開発だけでなく、具体的な生物学的・医学的な問いに答えるための応用研究を強化する。

その他、議論した内容 (ネガティブコメントや limitation もあれば):

- **特定タスクでの性能:** ヒトのミスセンス変異予測など特定のタスクにおいては、GPN-MSA や AlphaMissense のような特化型モデルに性能でわずかに劣る場合がある。
- **Non-coding 領域での性能:** Figure 3 において coding 領域と比較すると性能が良いのは興味深い。
- **生成物の機能:** 生成されたゲノムや配列が実際に機能を持つかどうかは、この研究では実験的に検証されていない。
- **デュアルユース懸念:** 安全対策は講じられているものの、強力なゲノム生成・設計能力を持つモデルの悪用リスクについては継続的な注意が必要。
- **解釈性の限界:** SAE による解釈は進んでいるが、モデル内部の全ての動作原理が完全に解明されたわけではない。