

ゲノム言語モデル
genome Language Model

バイオ生成AI (gLM) の研究動向 Evo2と関連研究

東 光一
国立遺伝学研究所

Evo2：全生物ドメインのゲノム基盤モデル

New Results

 [Follow this preprint](#)  [Previous](#)

Genome modeling and design across all domains of life with Evo 2

Posted February 21, 2025.

Garyk Brix, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li,  Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta,  Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke,  Hani Goodarzi,  Patrick D. Hsu,  Brian L. Hie

 [Download PDF](#)

 [Print/Save Options](#)

Subject Area

パラメータサイズ：7B/40B
コンテキストサイズ：1Mbp

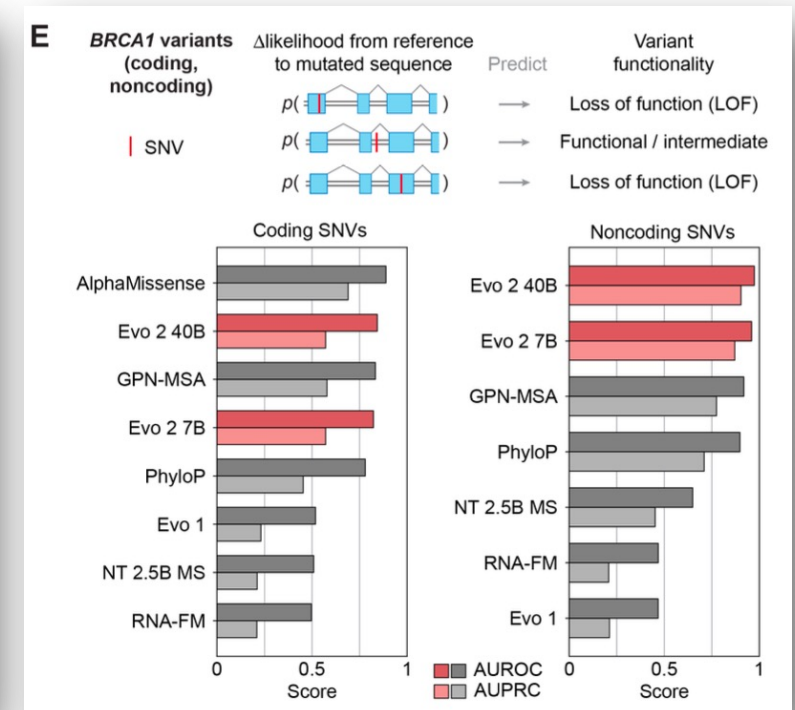
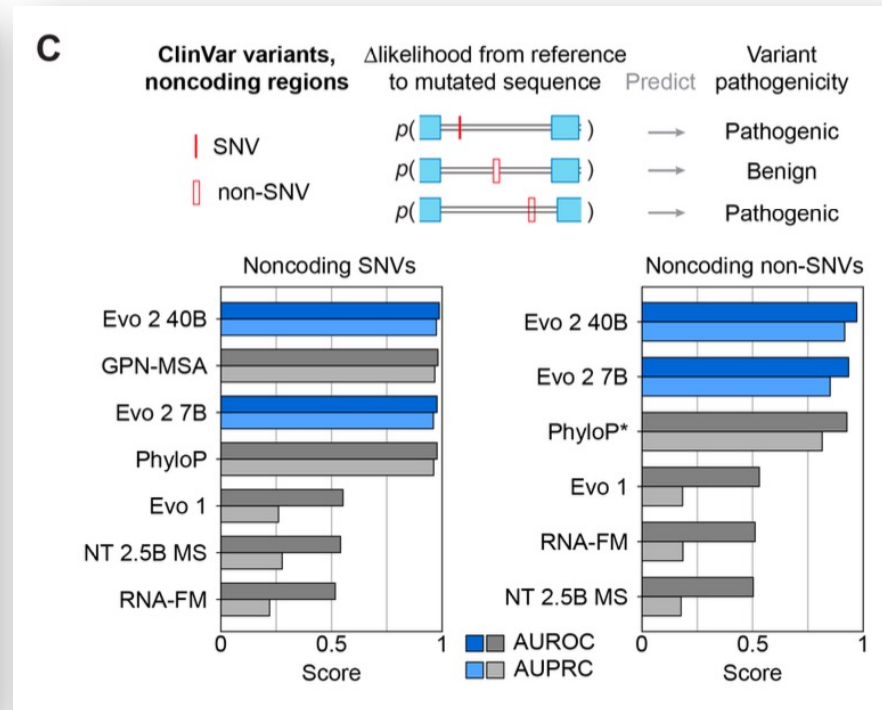
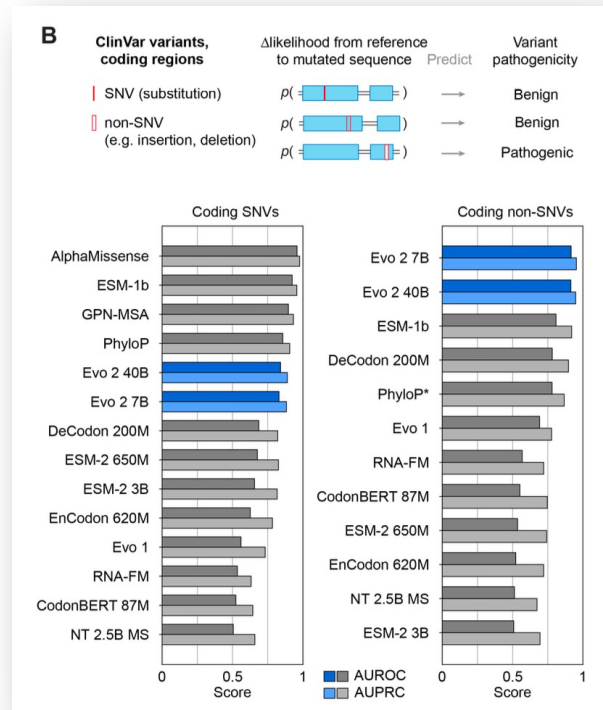
Evo1同様のHyenaアーキテクチャを
ベースに改良

学習データ：OpenGenome2 dataset

- Bacteria/Archaea: GTDB代表ゲノム約11万配列
- Eukaryotes: NCBI Genome. Mash距離でクラスタリングして代表ゲノムを抽出。
追加のフィルタリングで最終的に15,032ゲノム
- Metagenome: NCBI, JGI IMG, MGnify, Tara Oceansなどのコンティグ、MAG
- オルガネラゲノム：NCBI Organella. 32,240ゲノム
- Evo1同様、配列先頭に系統タグを付加
- 合計8.84Tbp

NVIDIA DGX Cloudで数ヶ月の学習。
H100 GPU 2,000枚以上に相当。

Evo2：結果・ヒトバリエーション影響予測



追加学習や回帰モデルの接続をいっさい使わない、**ゼロショットの予測性能**を評価。

つまり、DNA配列の尤度の数値のみで良性の変異か、有害な変異かを予測させる。（対数尤度差分をスコアとして利用）

バリエーション効果予測に特化したモデル（AlphaMissenseなど）に匹敵する性能。non-SNV（indel等）、Noncoding領域の予測性能はトップ。

GWASの変異データセットで学習したわけではない。ヒトの配列は学習データに一本しかないはずなのに、こんなことができるのは不思議。

→ 学習データの配列多様性・進化的多様性を進化的制約の代理変数として利用し、適応度地形を高精度に近似している？

あくまで進化史全体における期待値にすぎない点に注意（現存配列で議論する進化解析共通の課題）

BRCA1変異効果については、Functional/Intermediate/LOFといったクラス分類の追加学習を実施（Evo埋め込みを利用したシンプルなFFNN）

Evo2: “Likelihood = Fitness” について

From Likelihood to Fitness: Improving Variant Effect Prediction in Protein and Genome Language Models

Charles W. J. Pugh^{1,2,3} Paulina G. Nuñez-Valencia^{1,2,3}

Mafalda Dias^{1,2,3,*} Jonathan Frazer^{1,2,3,*}

¹Centre for Genomic Regulation (CRG),
The Barcelona Institute of Science and Technology,
Dr. Aiguader 88, Barcelona 08003, Spain

²Barcelona Collaboratorium for Modelling and Predictive Biology

³Universitat Pompeu Fabra (UPF), Barcelona, Spain

{mafalda.dias, jonathan.frazer}@crg.eu

<https://www.biorxiv.org/content/10.1101/2025.05.20.655154v1>

複数種オーソログの対数尤度変化を平均するだけで
ヒト変異分類の精度が上がる

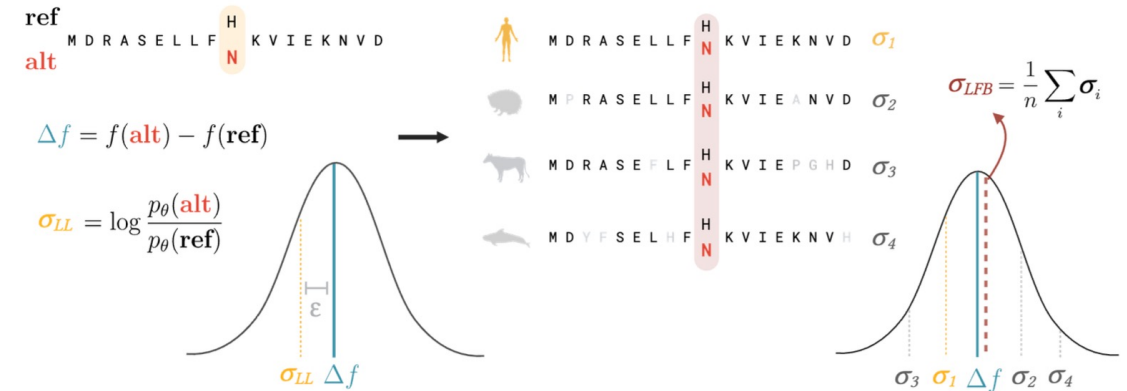
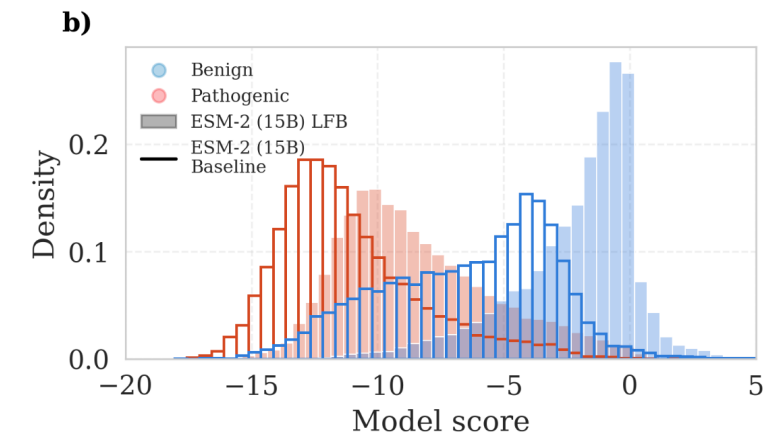
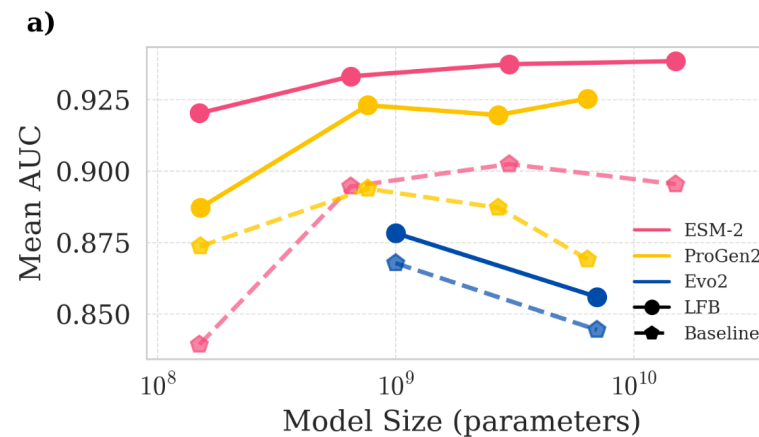


Figure 1: Overview of likelihood-fitness bridging (LFB) procedure. We propose that the log-likelihood ratio of alternative and reference sequences is a noisy estimate of the change in fitness Δf .

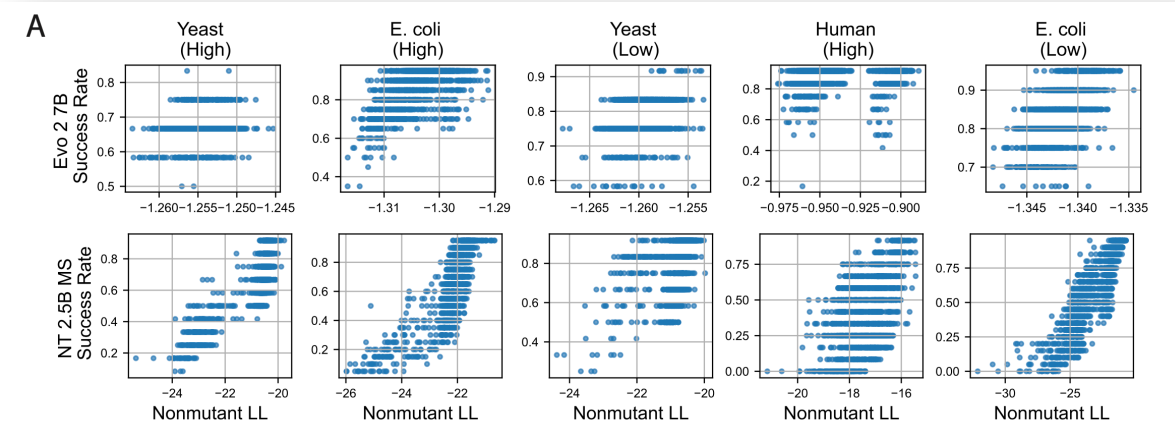
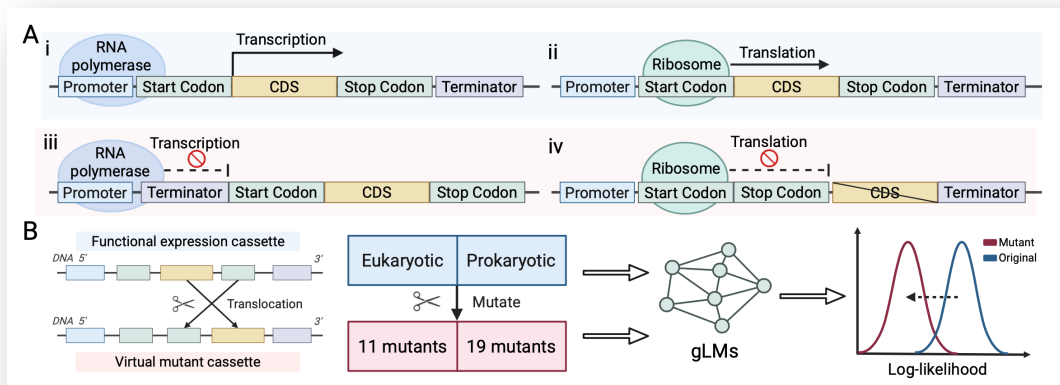


ベンチマーク構成の困難：

天然の生物配列を使ったベンチでは、同一の（または相同・類似した）配列が学習データに含まれてしまうため、（合成生物学で必要になる）「自然界に実在しない配列」の生成能力や適応度評価を計測できない。

NULLSETTESデータセット：

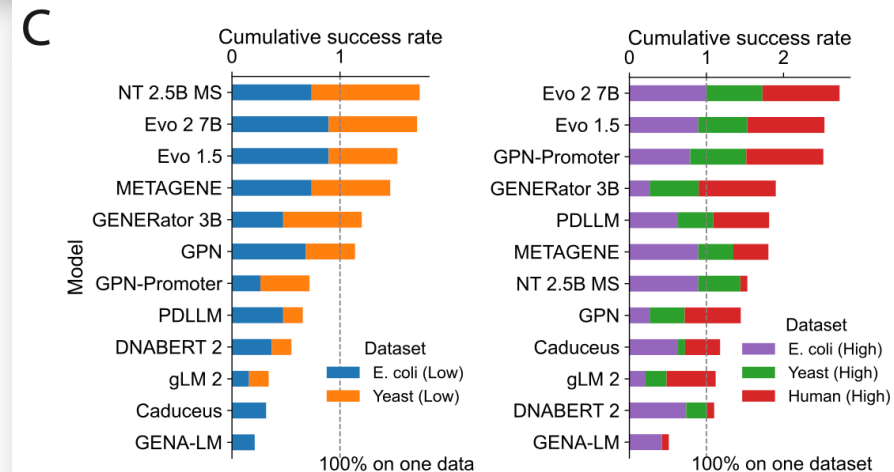
合成プロモータライブラリのデータセットを土台にベンチマークデータセットを構築。



遺伝子の構成要素を入れ替えたりして、ひとつの遺伝子領域について原核生物の場合11の擬似的変異配列を構成。それらがちゃんと「元配列と比較して対数尤度が下がるか」を評価する。

発見：

変異効果予測精度は「元の配列の尤度」と相関する。もともと尤度が低い（モデルが見慣れていない）配列では、変異によるLOFの予測精度が悪くなる。



BEND (Benchmark for DNA language models)

「生物学的タスクを、ヒトゲノム上で難易度・長さ・ラベル密度などについて網羅的に揃える」ことを目的に設計。

Published as a conference paper at ICLR 2024

BEND: BENCHMARKING DNA LANGUAGE MODELS ON BIOLOGICALLY MEANINGFUL TASKS

**Frederikke I. Marin^{1,2*}, Felix Teufel^{3,4*}, Marc Horlacher⁵, Dennis Madsen³,
Dennis Pultz¹, Ole Winther^{4,6}, Wouter Boomsma²**

¹Bioinformatics & Design, Enzyme Research, Novozymes A/S

²Department of Computer Science, University of Copenhagen

³Digital Science & Innovation, Novo Nordisk A/S

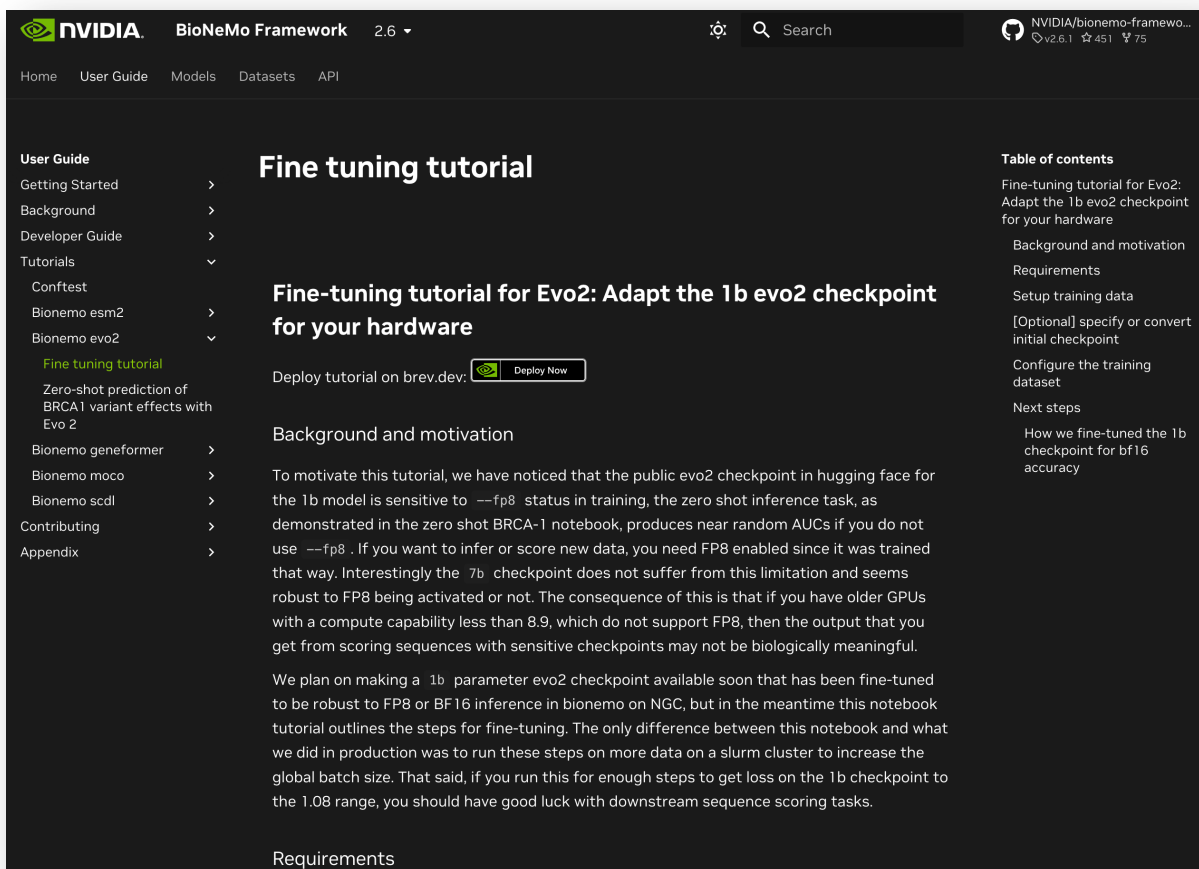
⁴Department of Biology, University of Copenhagen

⁵Computational Health Center, Helmholtz Center Munich

⁶DTU Compute, Technical University of Denmark

1. Gene finding
 - エクソン／イントロン／スプライスサイトなど 9 クラスを塩基単位で推定。
2. Enhancer annotation
 - 遺伝子 TSS 周辺 100 kb の領域で、実験的に機能が実証されたエンハンサーを検出。
3. Chromatin accessibility
 - DNase-Iアクセシビリティを 512 bp ウィンドウごとに多値予測。
4. Histone modification
 - 18 種ヒストンマークの有無。
5. CpG methylation
 - CpGメチル化を判定。
6. Non-coding variant effects – expression (eQTL)
 - 変異とeQTLによる遺伝子発現変化の関係。
7. Non-coding variant effects – disease (ClinVar)
 - ClinVarの良性／病原性変異

Evo2のファインチューニング



The screenshot shows the NVIDIA BioNeMo Framework documentation page for the "Fine tuning tutorial". The page is dark-themed and includes a sidebar with navigation links, a main content area with the tutorial title and description, and a table of contents on the right.

Navigation Links: Home, User Guide, Models, Datasets, API

User Guide: Getting Started, Background, Developer Guide, Tutorials, Conftest, Bionemo esm2, Bionemo evo2, Fine tuning tutorial, Zero-shot prediction of BRCA1 variant effects with Evo 2, Bionemo geneformer, Bionemo moco, Bionemo scdl, Contributing, Appendix

Fine tuning tutorial

Fine-tuning tutorial for Evo2: Adapt the 1b evo2 checkpoint for your hardware

Deploy tutorial on brev.dev: 

Table of contents

- Fine-tuning tutorial for Evo2: Adapt the 1b evo2 checkpoint for your hardware
- Background and motivation
- Requirements
- Setup training data
- [Optional] specify or convert initial checkpoint
- Configure the training dataset
- Next steps
- How we fine-tuned the 1b checkpoint for bf16 accuracy

Background and motivation

To motivate this tutorial, we have noticed that the public evo2 checkpoint in hugging face for the 1b model is sensitive to `--fp8` status in training, the zero shot inference task, as demonstrated in the zero shot BRCA-1 notebook, produces near random AUCs if you do not use `--fp8`. If you want to infer or score new data, you need FP8 enabled since it was trained that way. Interestingly the 7b checkpoint does not suffer from this limitation and seems robust to FP8 being activated or not. The consequence of this is that if you have older GPUs with a compute capability less than 8.9, which do not support FP8, then the output that you get from scoring sequences with sensitive checkpoints may not be biologically meaningful.

We plan on making a 1b parameter evo2 checkpoint available soon that has been fine-tuned to be robust to FP8 or BF16 inference in bionemo on NGC, but in the meantime this notebook tutorial outlines the steps for fine-tuning. The only difference between this notebook and what we did in production was to run these steps on more data on a slurm cluster to increase the global batch size. That said, if you run this for enough steps to get loss on the 1b checkpoint to the 1.08 range, you should have good luck with downstream sequence scoring tasks.

Requirements

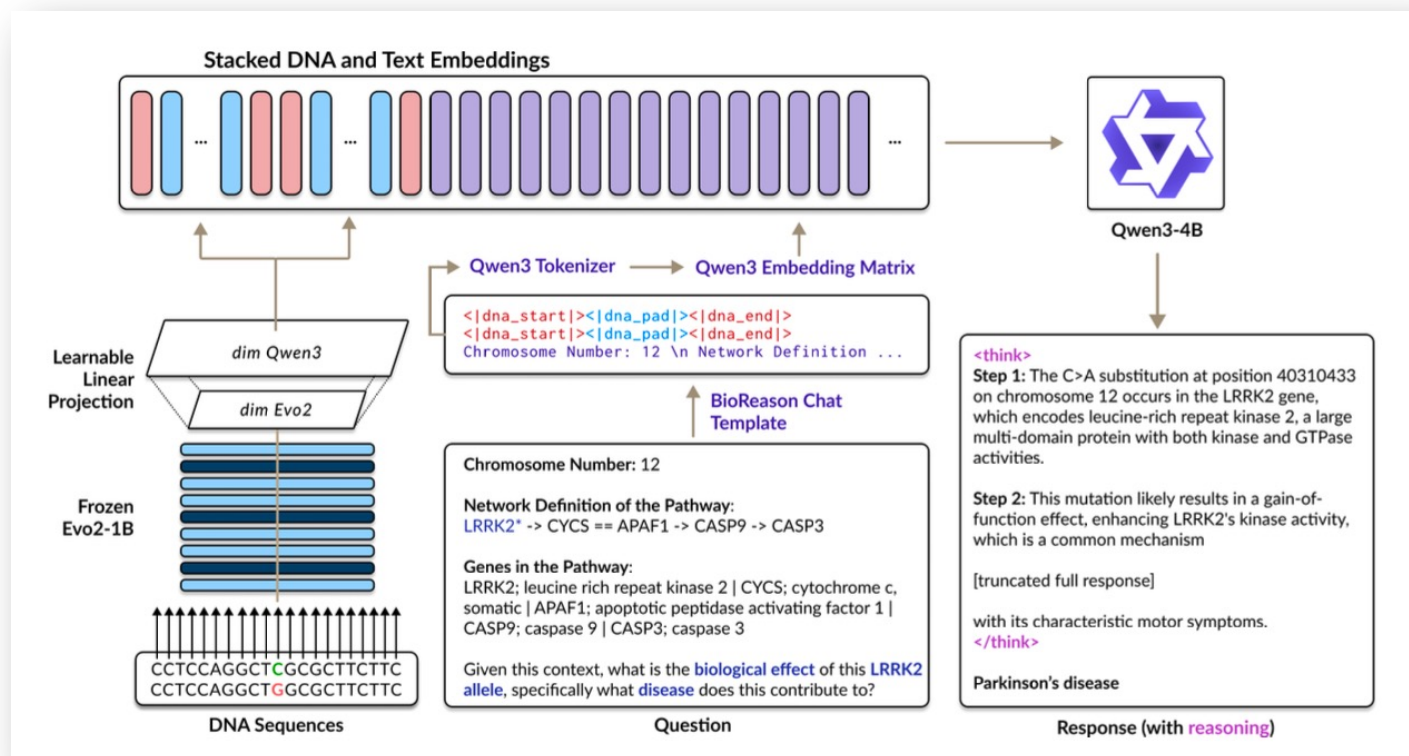
Evo2はHugging Faceでも NVIDIA NeMO/BioNeMo（NVIDIA版のHF transformersのようなフレームワーク）でも公開されている。
NVIDIAクラウド環境でファインチューニング可能。

ただし、現状はまだEvoを凍結してLoRAで学習する簡単な仕組みはない。
構造がTransformerではなく、どのレイヤー・ブロックがLoRAアダプテーションに最適なのかもよくわからない。

したがっていまのところ、ファインチューニングにあたっては、

1. Evo2全体を、手持ちのFASTAファイルとBioNeMoでファインチューニング
2. 適当なレイヤーのResidual streamを取り出して、FFNNなどで教師つき学習のどちらか。

<https://docs.nvidia.com/bionemo-framework/2.6/user-guide/examples/bionemo-evo2/fine-tuning-tutorial/>



KEGGを学習データとして、
変異と疾患の間のパス（作用機序など）の
推論モデルをトレーニング。

Evoのパラメータは固定。

Evoのトークン（塩基）埋め込みをQwenのトークン埋め込みの空間に線形にプロジェクションする。
（線形変換のパラメータは学習する）

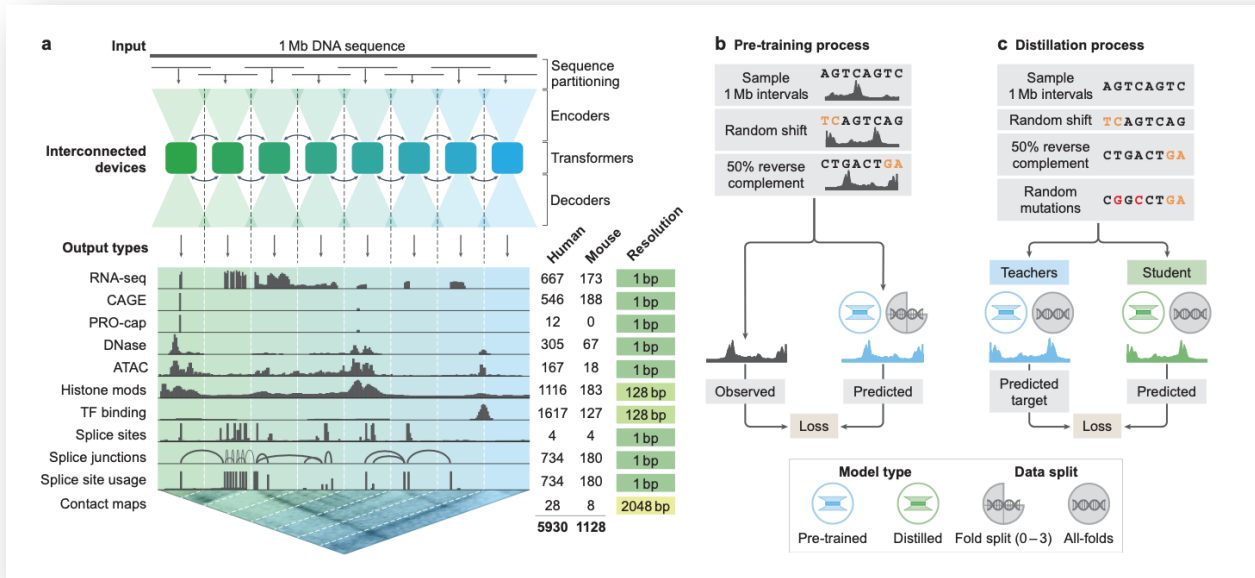
これによって、自然言語とDNA配列が同一の空間における
系列として扱える。

あとはふつうの推論モデルのように、推論ステップを
GRPO (Group relative policy optimization) で強化学習。

KEGGは1,500程度のエントリ。
ClinVarデータなどでデータを水増ししているが、それらは
KEGGのようなアノテーションを持たないため、フリーテキストとして損失計算に含めない。

Evo/Evo2以外のDNAモデル：AlphaGenome

<https://deepmind.google.com/science/alphagenome/>



生成モデルではない。

Enformerなどと同様、DNA配列入力 → エピジェネティクスパターン出力の教師あり学習。

現状、ヒトとマウス限定。

アテンションは使っているが、U-Net風のアーキテクチャ。

1MbpのDNA 配列をCNNで 128bp までダウンサンプリングし、アテンションで長距離依存性を捉えて、U-Net風デコーダで 1bp 解像度に戻してから、各種ゲノムトラックを出力する。

トレーニング後、蒸留によって、1Mbp入力のエピジェネパターンを数秒で予測可能。

予測内容は、

遺伝子発現: RNA-seq、CAGE、PRO-cap

スプライシング: スプライス部位など

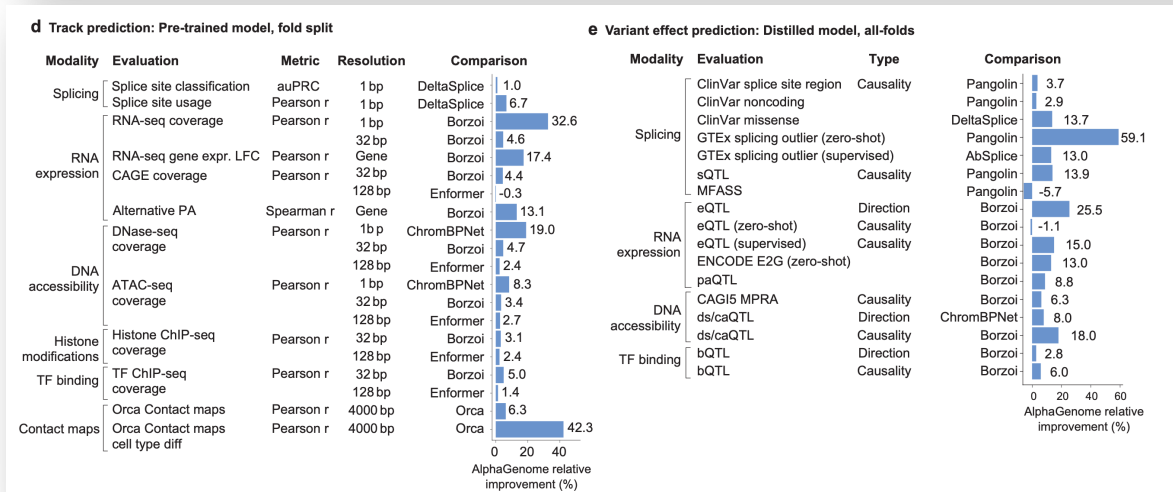
クロマチン状態: DNase、ATAC-seq、ヒストン修飾、転写因子結合

クロマチン接触マップ: Hi-C/micro-C

の11種類のゲノムトラックについて、細胞処理条件や細胞型の違い、コンソーシアムの違い（ENCODE, 4D Nucleomeなど）を、それぞれ別々に予測する数千本のゲノムトラック。

なので、特定の細胞型、

たとえばHeLa細胞のH3K27acパターンが知りたい、といった場合、いったん該当領域の数千本のゲノムトラック全部を予測させて、得られた数千本の出力トラックから該当するものを選んでくる、という感じになる。



所感・課題など

- ・ベンチマークの不足
- ・そもそもベンチマークの構成そのものが難しい
- ・本当に「生成」が必要か。
 - ・ 埋め込みの利用とエピ予測には、必ずしも生成能力は必要ではない。（AlphaGenome）
 - ・ 対数尤度計算は生成モデル特有。
 - ・ だが適応度との関連についてまだじゅうぶんに検証できていない。