ゲノム言語モデル研究動向と開発進捗状況

東　光一
国立遺伝学研究所

最近提案されたモデル①　https://arxiv.org/abs/2507.19229

arXiv > cs > arXiv:2507.19229

Computer Science > Computational Engineering, Finance, and Science

[Submitted on 25 Jul 2025]

TrinityDNA: A Bio-Inspired Foundational Model for Efficient Long-Sequence DNA Modeling

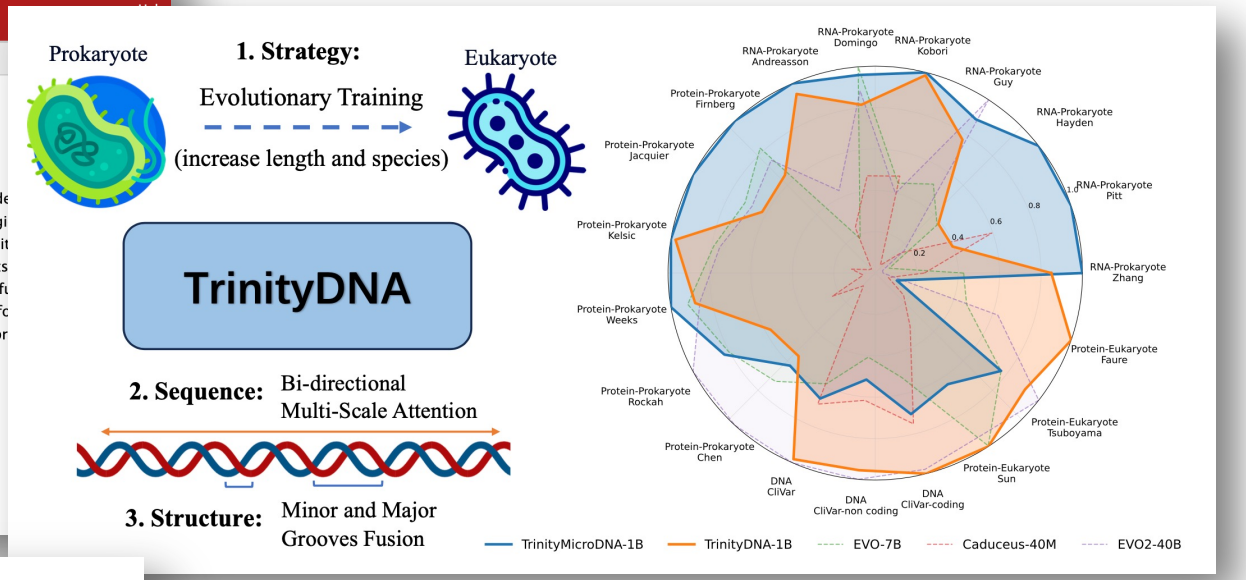
Qirong Yang, Yucheng Guo, Zicheng Liu, Yujie Yang, Qijin Yin, Siyuan Li, Shaomin Ji, Linlin Chao, Xiaoming Zhang, Stan Z. Li

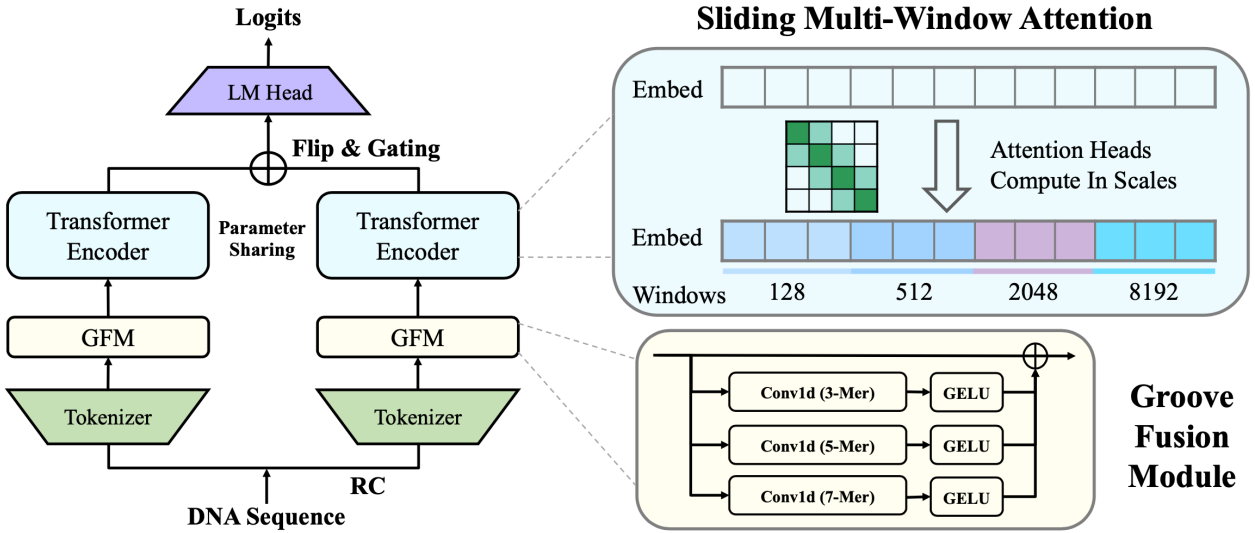
The modeling of genomic sequences presents unique challenges due to their length and structural complexity. Traditional sequence models struggle to capture long-range de
biological features inherent in DNA. In this work, we propose TrinityDNA, a novel DNA foundational model designed to address these challenges. The model integrates biologi
components, including Groove Fusion for capturing DNA's structural features and Gated Reverse Complement (GRC) to handle the inherent symmetry of DNA sequences. Addit
a multi-scale attention mechanism that allows the model to attend to varying levels of sequence dependencies, and an evolutionary training strategy that progressively adapts
prokaryotic and eukaryotic genomes. TrinityDNA provides a more accurate and efficient approach to genomic sequence modeling, offering significant improvements in gene fu
regulatory mechanism discovery, and other genomics applications. Our model bridges the gap between machine learning techniques and biological insights, paving the way fo
analysis of genomic data. Additionally, we introduced a new DNA long-sequence CDS annotation benchmark to make evaluations more comprehensive and oriented toward pr

Subjects: **Computational Engineering, Finance, and Science (cs.CE)**; Genomics (q-bio.GN)
Cite as: arXiv:2507.19229 **[cs.CE]**
(or arXiv:2507.19229v1 **[cs.CE]** for this version)
https://doi.org/10.48550/arXiv.2507.19229

Submission history

From: Zicheng Liu [view email]
[v1] Fri, 25 Jul 2025 12:55:30 UTC (1,753 KB)





進化史をトレースした学習で
学習効率・性能が上がる？

最近提案されたモデル②　https://www.arxiv.org/abs/2508.04739

arXiv > q-bio > arXiv:2508.04739

Quantitative Biology > Genomics

[Submitted on 6 Aug 2025]

CodonMoE: DNA Language Models for mRNA Analyses

Shiyi Du, Litian Liang, Jiayi Li, Carl Kingsford

Genomic language models (gLMs) face a fundamental efficiency challenge: either maintain separate specialized models for each biological modality (DNA and RNA) or develop large multi-modal architectures. Both approaches impose significant computational burdens – modality-specific models require redundant infrastructure despite inherent biological connections, while multi-modal architectures demand massive parameter counts and extensive cross-modality pretraining. To address this limitation, we introduce CodonMoE (Adaptive Mixture of Codon Reformative Experts), a lightweight adapter that transforms DNA language models into effective RNA analyzers without RNA-specific pretraining. Our theoretical analysis establishes CodonMoE as a universal approximator at the codon level, capable of mapping arbitrary functions from codon sequences to RNA properties given sufficient expert capacity. Across four RNA prediction tasks spanning stability, expression, and regulation, DNA models augmented with CodonMoE significantly outperform their unmodified counterparts, with HyenaDNA+CodonMoE series achieving state-of-the-art results using 80% fewer parameters than specialized RNA models. By maintaining sub-quadratic complexity while achieving superior performance, our approach provides a principled path toward unifying genomic language modeling, leveraging more abundant DNA data and reducing computational overhead while preserving modality-specific performance advantages.

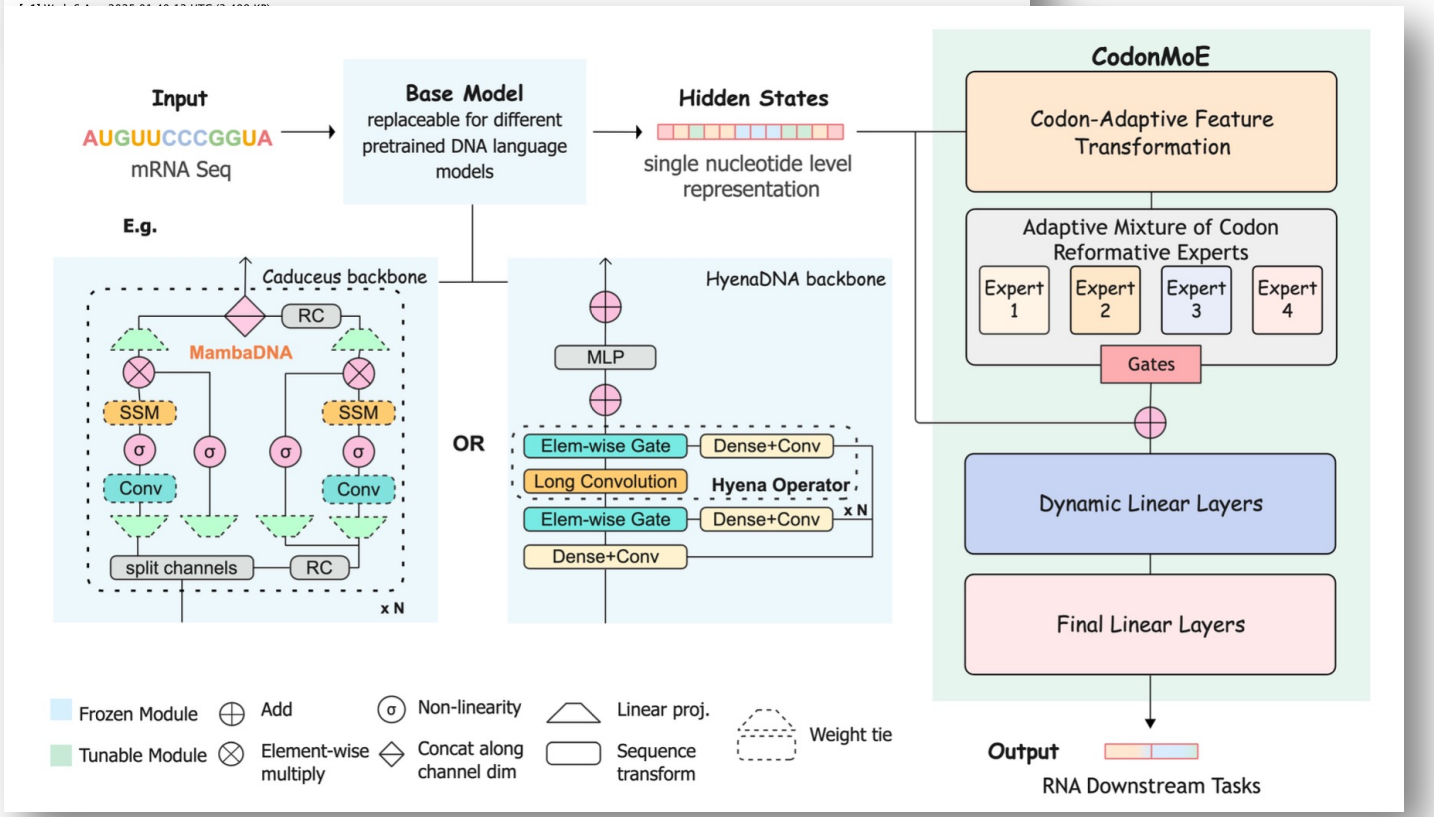
Subjects: **Genomics (q-bio.GN)**; Machine Learning (cs.LG)
Cite as: arXiv:2508.04739 [q-bio.GN]
(or arXiv:2508.04739v1 [q-bio.GN] for this version)
https://doi.org/10.48550/arXiv.2508.04739

Submission history

From: Shiyi Du [view email]

	mRFP		Vaccine degradation	
	GPN-MSA	HyenaDNA	GPN-MSA	HyenaDNA
MLP	0.330	0.439	0.572	0.695
XGBoost	0.479	0.512	0.750	0.711

	mRFP		Vaccine degradation	
	GPN-MSA	HyenaDNA	GPN-MSA	HyenaDNA
CodonMean	0.740	0.765	0.729	0.789
CodonMoE	0.790	0.837	0.770	0.812
CodonMoE-pro	**0.808**	**0.878**	**0.823**	**0.844**



基盤モデルというより、タスク適応のあり方についてのアイディア。

mRNAの発現量や分解予測タスクについて、生物学的知識（コドン）を明示的に構造化したモデリング（帰納バイアス）が効果的。

CodonMoE: コドンの希少性、GC含量、翻訳速度などを異なるExpertに分担して処理させる
CodonMoE-pro: Di-codon, Tri-codonの畳み込みを追加

開発体制

モデリング側の実働部隊（敬称略）

引き続き、ご興味ある方は東（ khigashi@nig.ac.jp ）までご連絡ください。
特に、アプリケーションに近い領域に関心ある方お待ちしてます。

国立遺伝学研究所 黒川顕研究室
　　東　光一　（助教）
国立遺伝学研究所 中村保一研究室
　　望月　孝子（特任研究員）
東京科学大学 山田拓司研究室
　　廣田　佳亮（博士後期課程一年）
　　松本　淳弥（修士課程一年）
　　豊田　大樹（学部四年）
　　中居　風雅（学部四年）
東京科学大学 佐藤健吾研究室
　　築山　翔（助教）
東京大学 津田宏治研究室
　　張　一鳴（修士課程一年）
筑波大学 天笠俊之研究室
　　鈴木　翔介（博士後期課程一年）

https://sc.ddbj.nig.ac.jp/guides/hardware/hardware2025/

アクセラレータ最適化ノード

アクセラレータ最適化ノード Type 1 (4台)

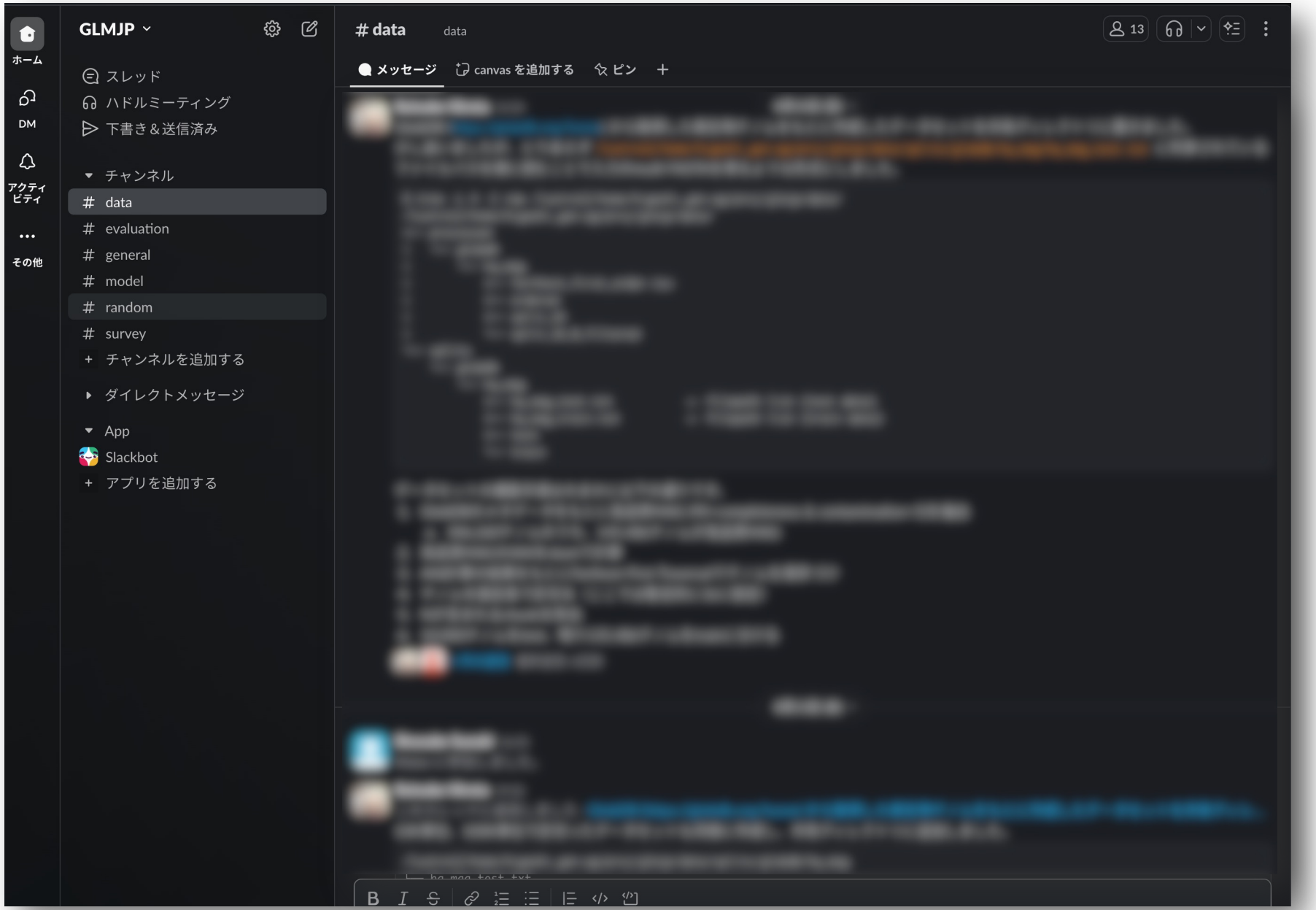
NVIDIA B200 GPU を各ノードに 8基搭載した計算ノードです。AI用の計算に適したGPU搭載計算ノードです。

NVIDIA DGX B200

構成要素	型番	員数	ノードあたりの性能など
CPU	Intel Xeon Platinum 8570 (56 cores) Base 2.1GHz, Max 4.0GHz, 1.97TFlops	2	合計 112 コア, 3.94TFlops
Memory	合計2TB		合計 2TB (CPU コアあたり 17.9GB)
GPU	NVIDIA Blackwell B200	8	
Storage (OS)	1.9TB NVMe SSD	2	合計3.8TB
Storage (Data)	3.84TB NVMe SSD	8	合計30.7TB
Network	InfiniBand NDR	1	400Gbps

開発環境：
遺伝研スパコン NVIDIA DGX B200 4台
（B200 1枚あたりVRAM 192GB, 4ノード合計32枚）

開発メンバーSlack



開発項目

データチーム
・データセット（DNA配列）収集、整理
・原核ゲノム、真核ゲノム、メタゲノム、プラスミド
・ウイルス配列使用の是非（バイオセキュリティ）
・リピート配列の影響調査（データセット冗長性）、マスキングあるいはダウンサンプリング
・配列多様性の影響（ANI、進化系統）
・学習スケジュール、進化史をトレースした学習で学習効率が上がる？

評価チーム
・計算（学習・推論）時間計測
・性能評価
・ベンチマーク収集
・現状、埋め込み（後段SFTタスク）の評価がほとんど
・「生成」を評価するベンチの収集・評価
・既存モデル（HyenaDNA, Evo2, Caduceus等）の推論実行・評価
・応用との接続

モデルチーム
・モデル構築
・要素技術検証
・実装
・学習・推論実行
・最新技術キャッチアップ

進捗状況

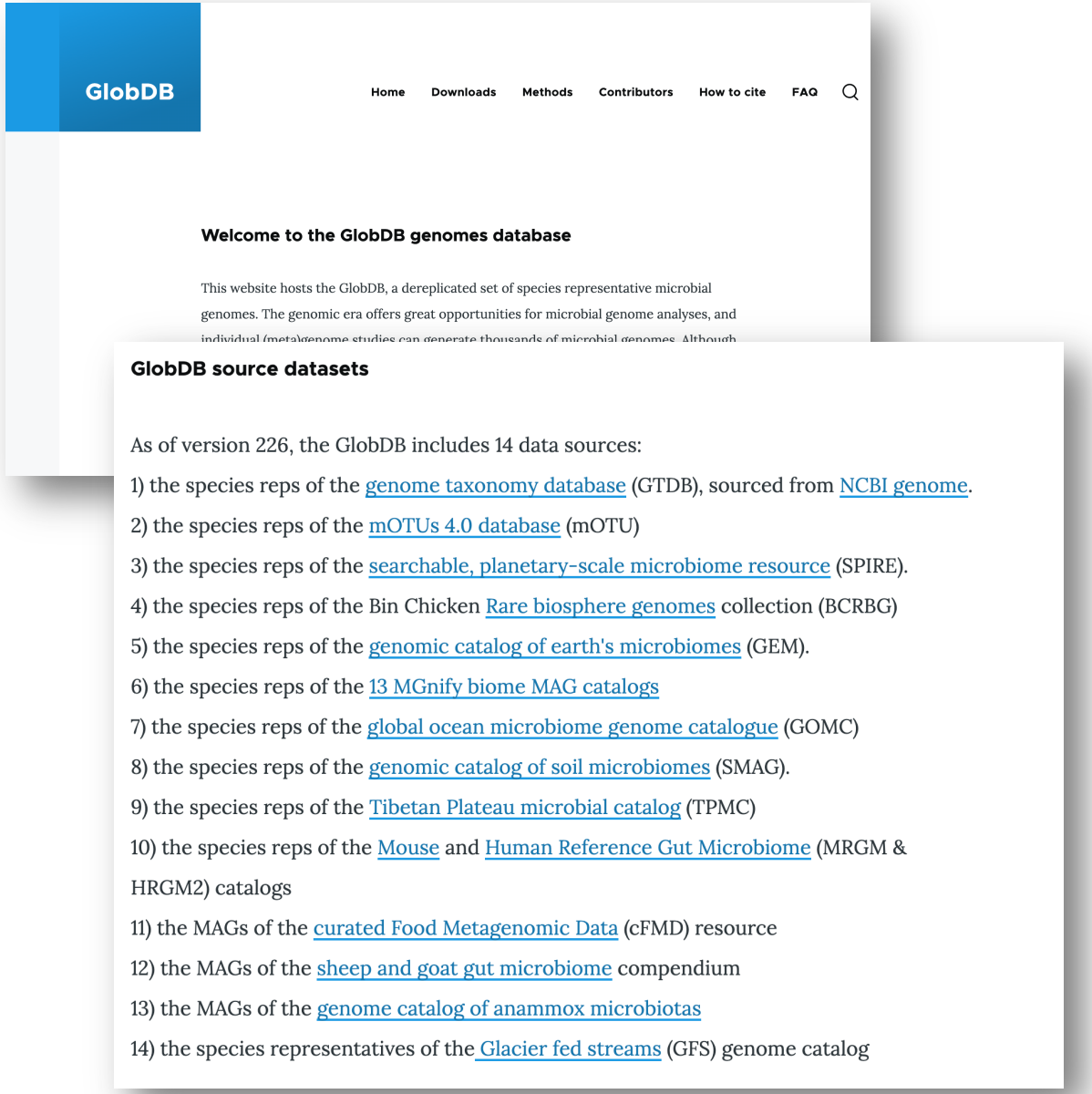
- 環境構築
  - 遺伝研スパコン個人ゲノム解析区画のアカウント申請
  - 計算環境はすべてApptainer (Singularity)コンテナ上に構築
  - 既存モデル（HyenaDNA, Evo1/Evo2）推論実行コンテナイメージの開発
  - Slurmでジョブ管理
  - DNA配列のロードと尤度計算をApptainerで実行するSlurmジョブ開発
- データセット構築（暫定的な性能評価データセット）
- 既存モデルの性能評価
- HyenaDNAトレーニングの実行と評価

データセット構築

MAG（Metagenome-assembled genomes）データの利用

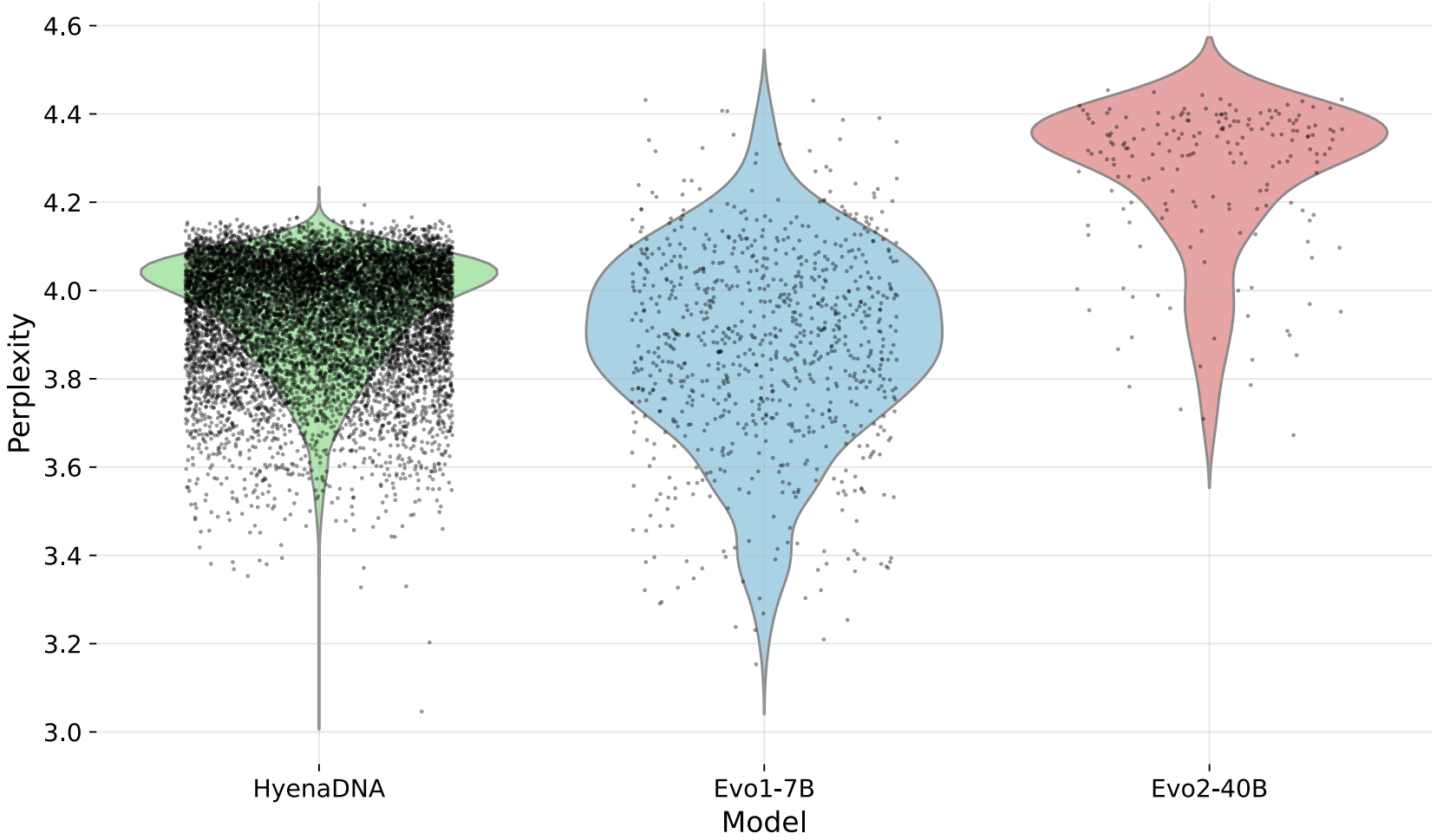
1. GlobDBのメタデータをもとに高品質MAG (90>completeness & contamination<5)を抽出
   - 306,260ゲノムのうち、145,486ゲノムが高品質MAG
2. 高品質MAGのANIをskaniで計算
3. ANI計算の結果をもとにFarthest-First Traversalでゲノムを順序づけ
4. ゲノムを固定長で区切る
   - ここでは暫定的に1kに設定
5. Nが含まれるchunkを除去
6. 10,000ゲノムをtest、残り135,486ゲノムをtrainに分ける

https://globdb.org



既存モデルのテストセット予測性能評価

テストセットのゲノム配列それぞれについて、1kbp コンテキストのPerplexityを計算。



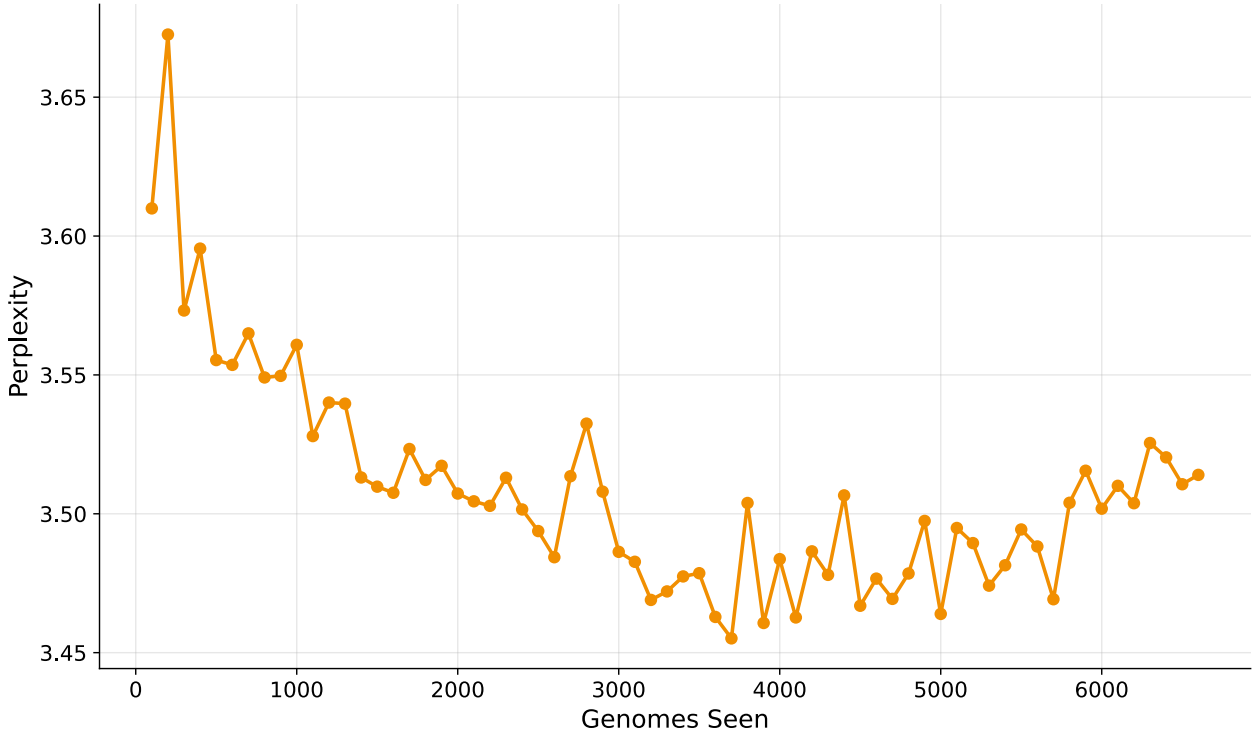
※Evo1/Evo2については系統タグ（Greengenes-style lineages strings）未使用時の計算。

HyenaDNAトレーニングの実行

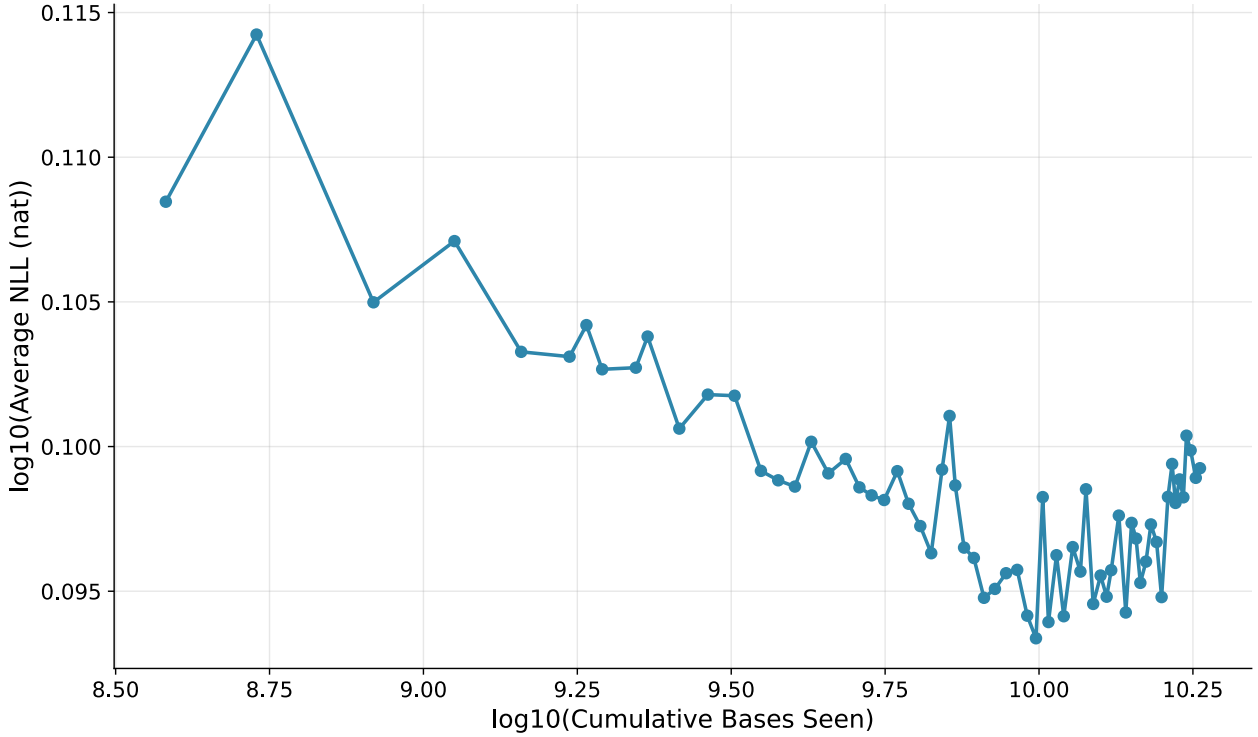
HyenaDNAのアーキテクチャだけ拝借してランダム初期値からトレーニング。

100ゲノム学習するごとにチェックポイントを保存してテストデータを予測した。

観測ゲノム数 vs. Perplexity



観測トークン（塩基）数 vs. NLL



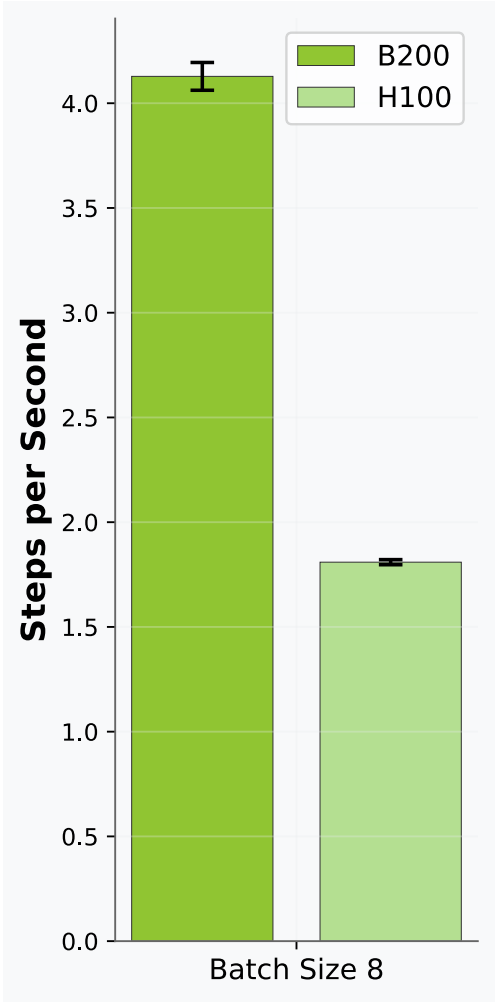
参考 ・学習データ総トークン数

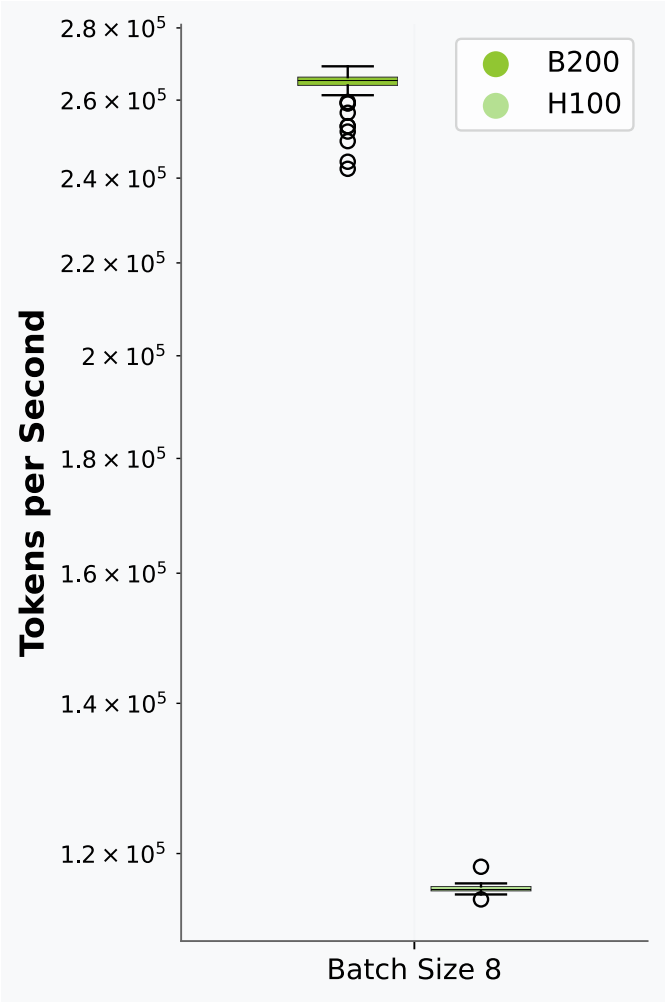
Evo1: 300B

Evo2: 8.8T

B200 計算速度評価

同一モデル、同一データ、同一バッチサイズで、HyenaDNAをトレーニング。
H100 GPUとB200 GPUの処理速度を比較した。





今後の方針

- トレーニング、性能評価の土台はできた
- 小規模モデルの要素技術を検証して性能向上を目指す
- 学習データ（の冗長性）の検証
- ベンチマーク検討（テストセットは適切か？NULLSETSなど合成配列の利用）
- 応用タスクの設定と性能評価