

ゲノム言語モデル
genome Language Model

バイオ生成AI（gLM）の研究動向 および今後の開発方針

東 光一
国立遺伝学研究所

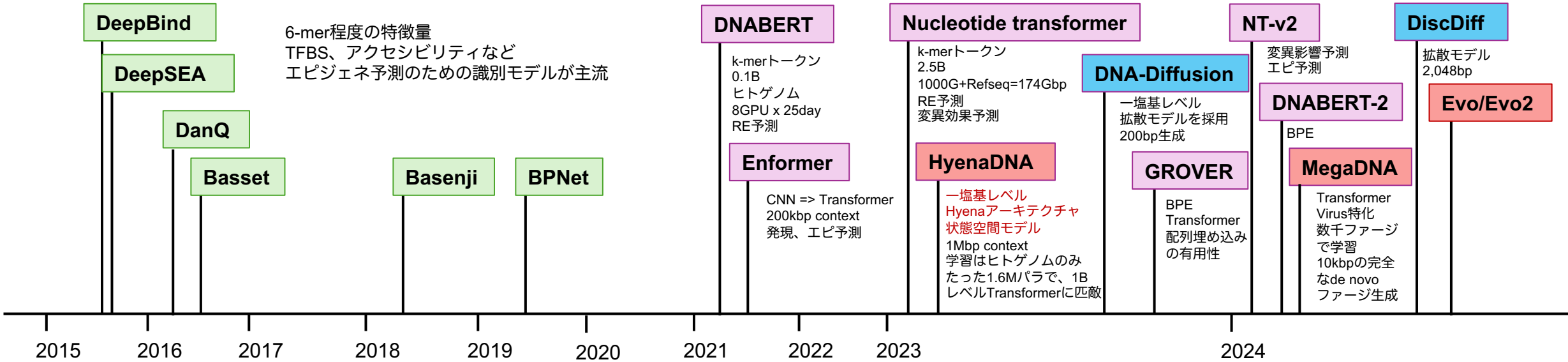
gLM開発動向（～2024）

CNN（畳み込みNN）識別モデル

Encoder-only Transformer（識別モデル）

自己回帰型生成モデル

拡散モデル



- RNNが主流
 - word2vecなど単語埋め込みの研究
 - Attentionの初提案（Bahdanau 2014）
 - seq2seqによるニューラル翻訳
 - AttentionベースのRNN
 - **Attention is all you need論文**
 - Transformerの提案
- BERT
 - 大規模コーパスによる
基盤モデル
 - GPT-2発表
 - GPT-3
 - **Language models are
few-shot learners論文**
 - **スケーリング則論文**
 - 基盤モデルの万能性
- 画像生成における
拡散モデルの流行
 - GPT-3.5
 - ChatGPT
 - GPT-4
 - LLMの爆発的な普及
- マルチモーダル
 - 推論モデルの登場
（OpenAI o1/o3,
DeepSeek-R1）

同時期のNLP（自然言語）モデルの状況

gLM開発動向（2024/2025）

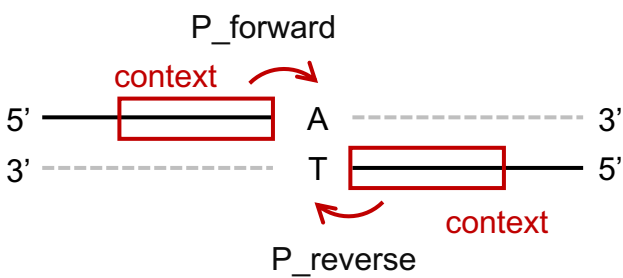
Date	Model name	Tokenization	Model size	Context	Training data	Architecture	Architecture補足	Benchmark	応用等
Jun-23	HyenaDNA	塩基レベル	~ 6.6M	~ 1Mbp	hg38	SSM	Hyena SSM	Genomic benchmarks, NTBench	
Feb-24	Evo		7B	~ 131kbp	OpenGenome (原核300B)	SSM	StripedHyena SSM		CRISPR-cas生成、Transposon生成、全ゲノム生成、尤度評価
Feb-24	DNA-Diffusion		90M	200bp	ENCODE ATAC peaks	Diffusion	U-net like. DDPM		Enhancer生成
Feb-24	DiscDiff			2kbp	EPD-GenDNA (15種)	Diffusion	VAE潜在表現をU-net like DDPMでモデリング。Absorb-Escape手法（学習後、生成配列のARモデルによるRefine, 10.48550/arXiv.2410.21345）の開発。		
Feb-24	Dirichlet Flow Matching			500bp	E.coli promoters, Human ATAC, Fly enhancers	Flow matching	Dirichlet Flow Matching. Simplex上確率フローのモデリング。		RE活性予測
Mar-24	Caduceus	塩基レベル	770M	~ 131kbp	hg38	SSM	BiMamba. Reverse Complement対応。	Genomic benchmarks, NTBench	
Feb-25	Evo2	塩基レベル	7B, 40B	~ 1Mbp	OpenGenome2 (8.8T)	SSM	StripedHyena2 SSM		
Feb-25	HybriDNA	塩基レベル	300M, 3B, 7B	~ 131kbp	845種ゲノム、160B	SSM+Transformer	Decoder-only transformerとMamba2のハイブリッド。埋め込みはEcho embedding（文脈を二回繰り返して埋め込む方法）	GUE, BEND, LRB	CRE生成（Human cell type specific enhancers, Yeast promoters）
Feb-25	GENERator	6-mer	1.2B	~ 98kbp	RefSeq真核368B	Transformer	Llama系Decoder-only Transformer	Genomic benchmarks , NTBench, 独自ベンチ	Enhancer活性予測、新規タンパク質
Mar-25	Shortlisting (SLM)	塩基レベル		~ 数千bp		Diffusion	Simplex DDPM. Simplex上centroidの移動パスを表現する独自設計の拡散モデル。		Promoter, Enhancerデザイン
May-25	JanusDNA			~ 1Mbp	hg38	SSM+Transformer	Bidirectional. 自己回帰型とMasked LMのハイブリッド、それぞれの損失を同時最適化。MoEの有効性主張。	Genomic benchmarks (8RE分類), NTBench, LRB(eQTL task)	
May-25	D3		92M	500bp	RE dataset (~100k seqs)	Diffusion	離散拡散モデル。クラスラベルによる条件付き生成。	独自構成ベンチ（生成 vs. 自然スコアリング）	Cell type specific RE生成。
Jun-25	eccDNAMamba	BPE			630k eccDNAs	SSM	Mamba2, 両方向モデル		eccDNA vs. chromosomal予測、疾患予測
Jun-25	GENERanno	6-mer	500M	8kbp	RefSeq原核715B	Encoder-only Transformer	GENERatorのファインチューニング。メタゲノムアノテーションに特化。	メタゲノム遺伝子アノテーション	GeneMark, Prodigalを凌駕。Pseudo geneをゼロショット検出。ARG推定、系統分類
Jun-25	Frequency-Blended Diffusion (FBD)	塩基レベル		~ 数千bp		Diffusion	Frequency-blended DDPM. 塩基頻度の周波数領域データを通常のDDPMにブレンド。周期性パターンを高精度に再現。		スプライシングシグナルを高精度に再現した配列生成

gLM開発動向（2024/2025）

Date	Model name	Tokenization	Model size	Context	Training data	Architecture	Architecture補足	Benchmark	応用等
Jun-23	HyenaDNA	塩基レベル	~ 6.6M	~ 1Mbp	hg38	SSM	Hyena SSM	Genomic benchmarks, NTBench	
Feb-24	Evo	塩基レベル	7B	~ 131kbp	OpenGenome (原核300B)	SSM	StripedHyena SSM		CRISPR-cas生成、Transposon生成、全ゲノム生成、尤度評価
Feb-24	DNA-Diffusion	塩基レベル	90M	200bp	ENCODE ATAC peaks	Diffusion	U-net like. DDPM		Enhancer生成
Feb-24	DiscDiff	塩基レベル		2kbp	EPD-GenDNA (15種)	Diffusion	VAE潜在表現をU-net like DDPMにインテグ。Absorb-Escape手法（学習後、生成配列のARモデルによるRefinement） 10.48550/arXiv.2410.21335		
Feb-24	Dirichlet Flow Matching	塩基レベル		500bp	E.coli promoters, Human ATAC, Fly enhancers	Flow matching	Dirichlet Flow Matching のモデリング。	Simplex上確率フロー	
Mar-24	Caduceus	塩基レベル	770M	~ 131kbp	hg38	SSM	BiMamba. Reverse Completion	Genomic benchmarks	
Feb-25	Evo2	塩基レベル	7B, 40B	~ 1Mbp	OpenGenome2 (8.8T)	SSM	StripedHyena2 SSM		
Feb-25	HybriDNA	塩基レベル	300M, 3B, 7B	~ 131kbp	845種ゲノム、160B	SSM+Transformer	Decoder-only transformerとMamba2のハイブリッド。埋め込みはEchob embedding（文脈を二回繰り返して埋め込む方法）	CRE生成（Human cell type specific promoters）	
Feb-25	GENERator	6-mer	1.2B	~ 98kbp	RefSeq真核368B	Transformer	Llama系Decoder-only Transformer	Genomic benchmarks, Enhancer活性予測、新規タンパク質	
Mar-25	Shortlisting (SLM)	塩基レベル		~ 数千bp		Diffusion	Simplex DDPM. Simplexを表現する独自設計の拡散モデル。		Promoter, Enhancerデザイン
May-25	JanusDNA	塩基レベル		~ 1Mbp	hg38	SSM+Transformer	Bidirectional. 自己回帰型とMasked LMのハイブリッド、それぞれの損失を同時最適化。MoEの有効性主張。	Genomic benchmarks (8RE分類), NTBench, LRB(eQTL task)	
May-25	D3	塩基レベル	92M	500bp	RE dataset (~100k seqs)	Diffusion	離散拡散モデル。クラスラベルによる条件付き生成。	独自構成ベンチ（生成 vs. 自然スコアリング）	Cell type specific RE生成。
Jun-25	eccDNAMamba	BPE			630k eccDNAs	SSM	Mamba2, 両方向モデル		eccDNA vs. chromosomal予測、疾患予測
Jun-25	GENERanno	6-mer	500M	8kbp	RefSeq原核715B	Encoder-only Transformer	GENERatorのファインチューニング。メタゲノムアノテーションに特化。	メタゲノム遺伝子アノテーション	GeneMark, Prodigalを凌駕。Pseudo geneをゼロショット検出。ARG推定、系統分類
Jun-25	Frequency-Blended Diffusion (FBD)	塩基レベル		~ 数千bp		Diffusion	Frequency-blended DDPM. 塩基頻度の周波数領域データを通常のDDPMにブレンド。周期性パターンを高精度に再現。		スプライシングシグナルを高精度に再現した配列生成

- トークン：塩基レベル
- SSM（状態空間モデル）
- 生成を評価するベンチの不足
- 拡散モデルは試みられているが、文脈を広くとれない
- タスク特化（ecDNA, メタゲノム）、相補鎖表現、周期性表現など、様々なアイデアが試されている

Evo/Evo2の問題点：演算の方向性

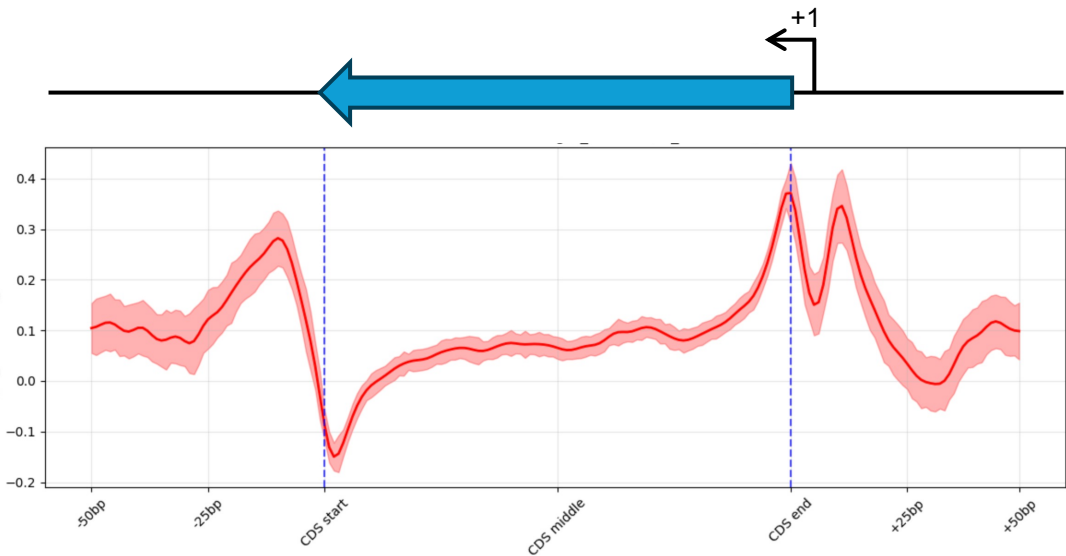
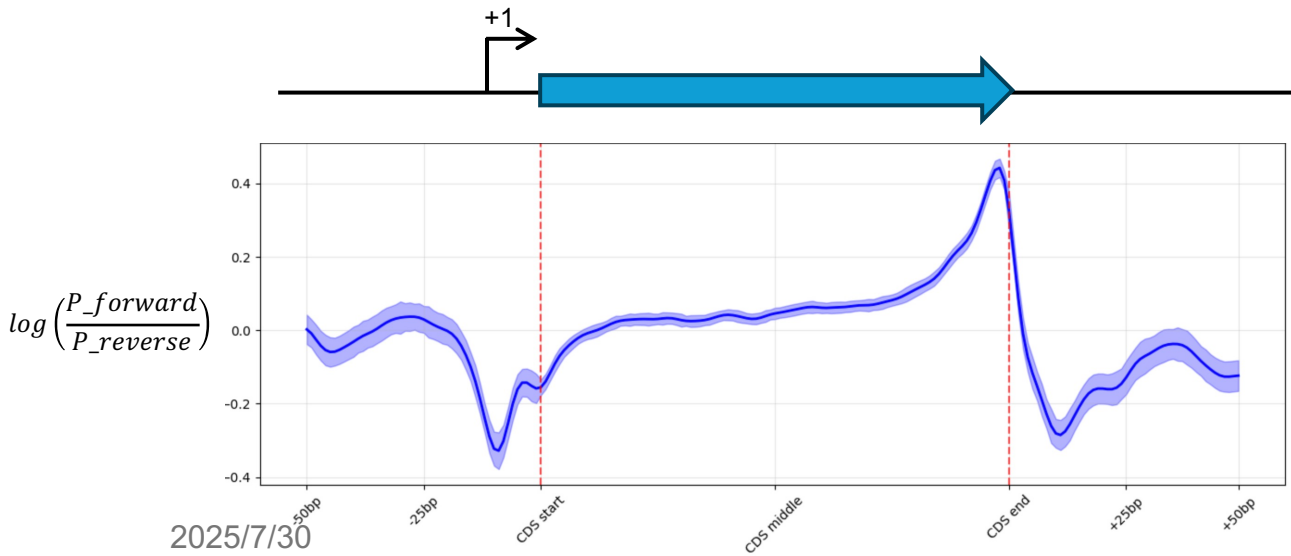
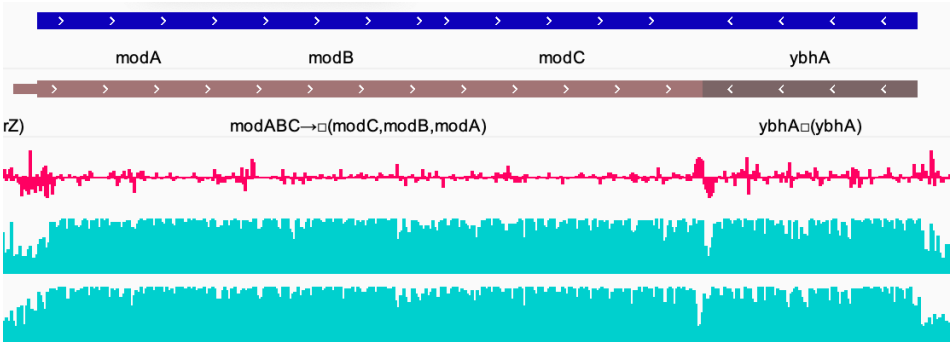
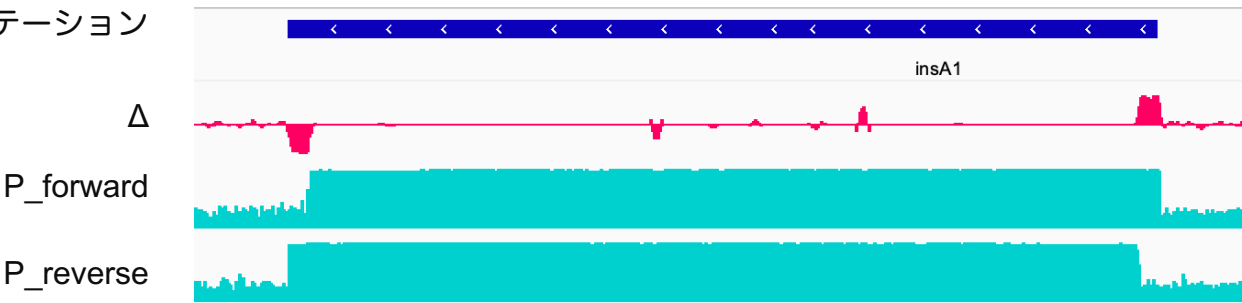


$$\Delta = \log(P_forward) - \log(P_reverse)$$

$\Delta > 0$: 左方向に制約された塩基

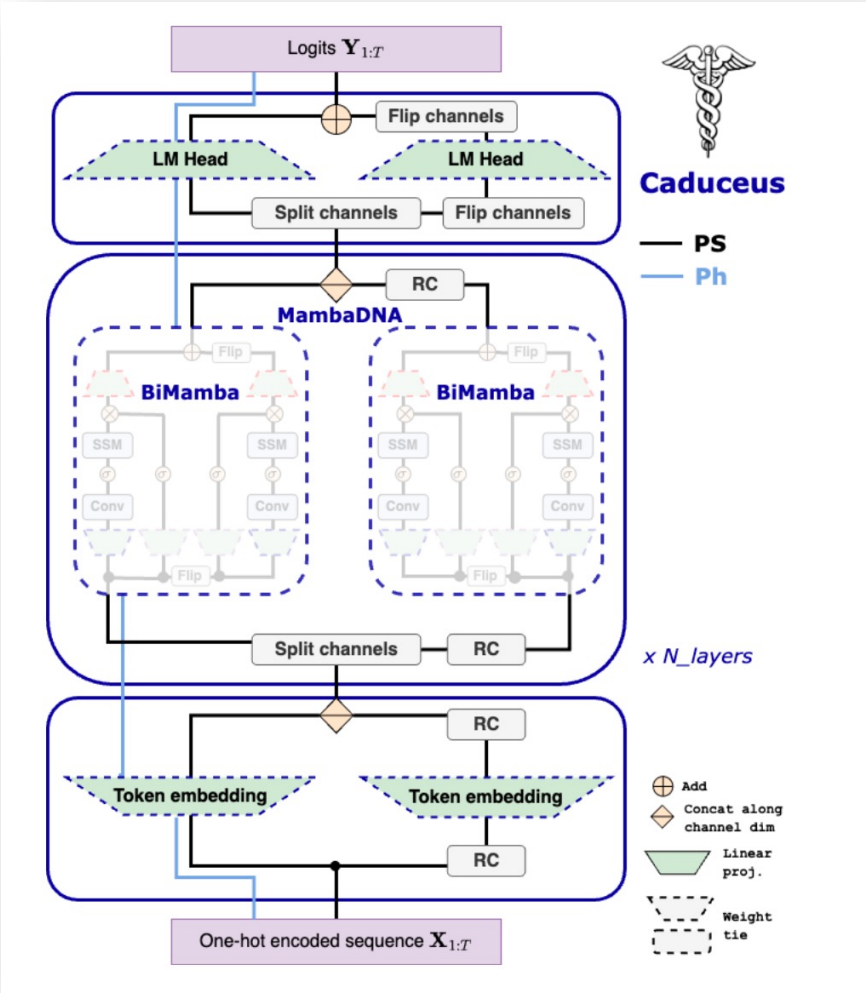
$\Delta < 0$: 右方向に制約された塩基

アノテーション

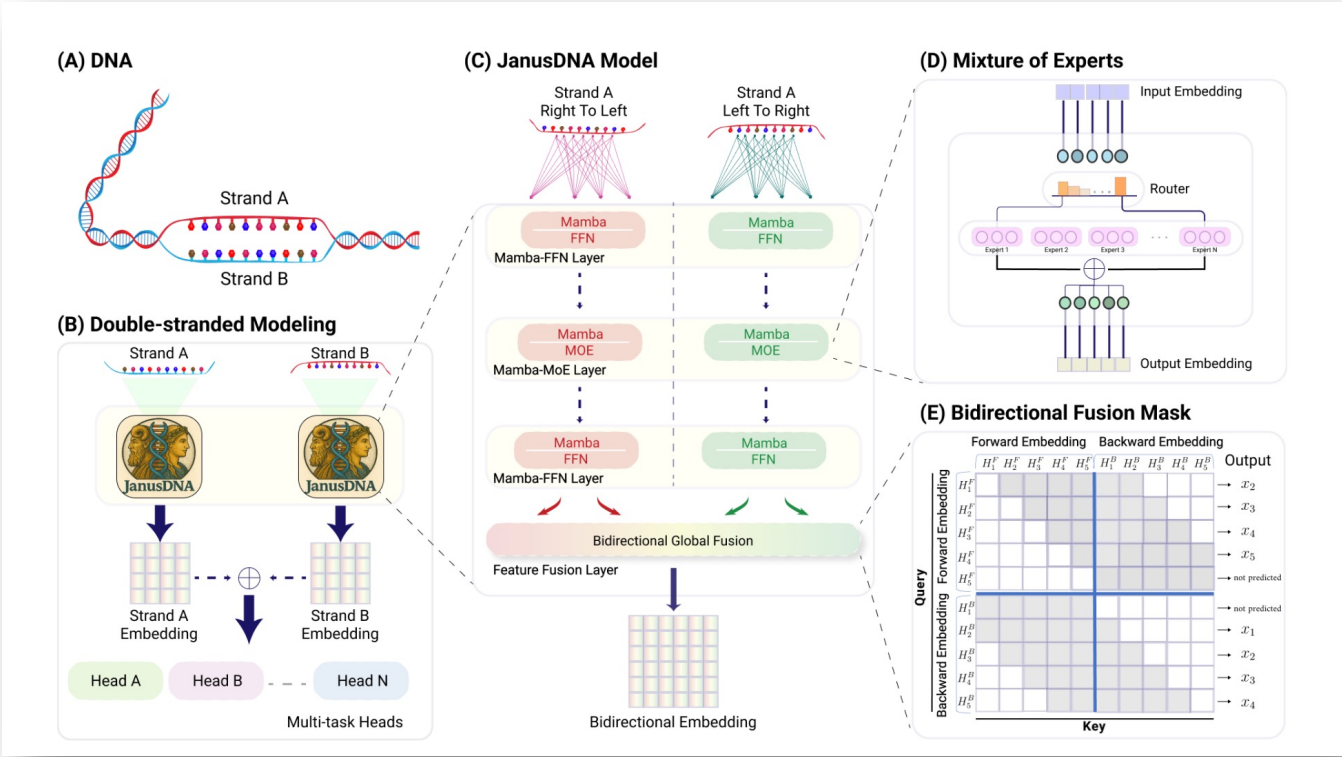


DNA分子特有の条件（相補鎖）を考慮したモデル

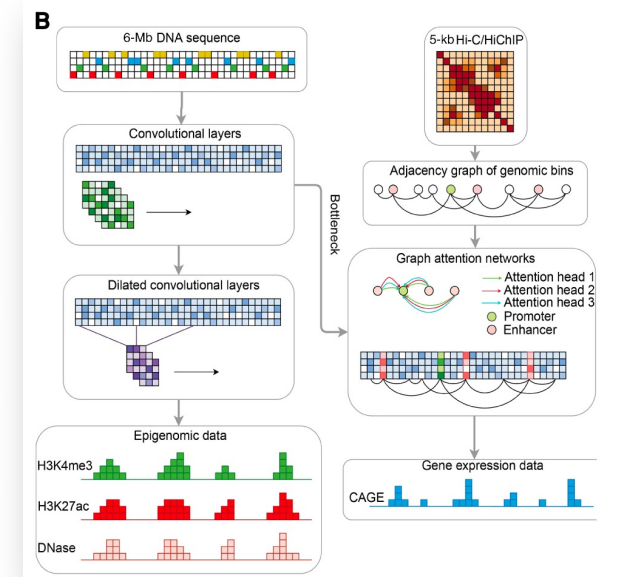
Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling
<https://doi.org/10.48550/arXiv.2403.03234>



JanusDNA: A Powerful Bi-directional Hybrid DNA Foundation Model
<https://doi.org/10.48550/arXiv.2505.17257>

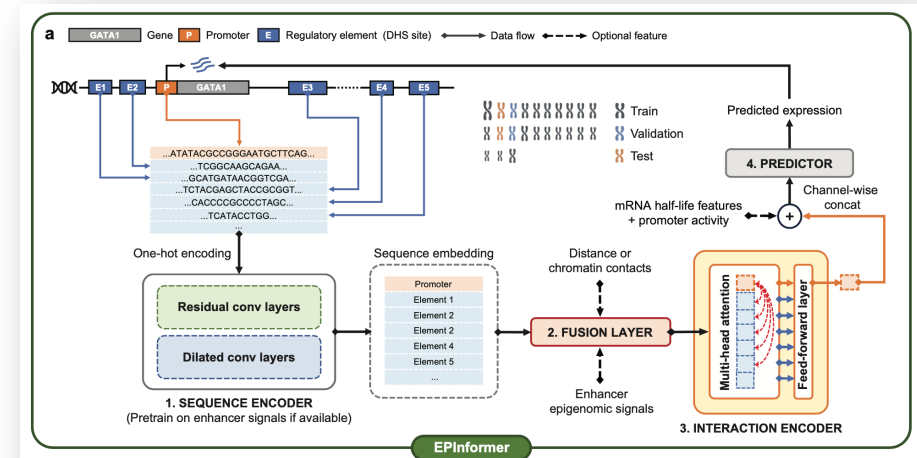


- モデルの構造にDNA特有の情報を入れ込む
 - 最低限、双方向性、相補鎖
- 他の生物学的情報を取り込むべきか？ (cf. GraphReg, EPInformer)
 - 他種ゲノムへの汎用性が課題
- MoEなど、いろんな要素を足したり引いたりしながら模索。
 - ベースラインと評価指標の固定が最優先。



- アプリケーション側要望との連携
- メガ塩基スケールのモデリングには現状、SSMベースのモデル一択。
 - 拡散モデルの動向を注視しつつ
- ベンチマーク、現状、生成評価に適したものがない。
 - NULLSETSの利用？さしあたってTest set perplexity

EPInformer: a scalable deep learning framework for gene expression prediction by integrating promoter-enhancer sequences with multimodal epigenomic data
<https://doi.org/10.1101/2024.08.01.606099>



開発体制

コアメンバー会議（7/9, 7/16, 7/23）

東、森下先生、浅井先生、笠原先生、黒川先生、津田先生

モデリング側の実働部隊：

東、学生4名（東京科学大学）

一緒に開発していただける方は東（khigashi@nig.ac.jp）までご連絡ください。

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）

モデル比較、要素技術検証、

データセット準備・検証（反復配列マスキングなど）

ベンチマーク設定、評価

「自然言語インターフェースの開発」
広く利用されることを目的とした場合必須。

<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

LLM-jp, 統合データベースとの連携

データベースが蓄積してきた文献情報、配列情報、アノテーション情報の活用

ゲノム表現と知識グラフとの連携
cf. BioReasonモデル

LLMエージェントの開発と運用
どのようにデータベース情報を
参照させるべきか？（MCP?）

検討課題

バイオセキュリティ対応。
機微情報でファインチューニングする仕組み。
基盤モデルと共に、プライベートデータで
SFTするシステムを同時提供するなど。
連合学習などでセキュアに可能か？

<https://rdfportal.org/>

