Generalized Biological Foundation Model with Unified Nucleic Acid and Protein Language

Yong He[1*], Pan Fang[1], Yongtao Shan[2], Yuanfei Pan[3], Yanhong Wei[4], Yichang Chen[5], Yihao Chen[6], Yi Liu[1], Zhenyu Zeng[1], Zhan Zhou[5], Feng Zhu[7], Edward C. Holmes[8], Jieping Ye[1], Jun Li[9], Yuelong Shu[10,11], Mang Shi[12*], Zhaorong Li[1*]

[1]Apsara Lab, Alibaba Cloud Intelligence, Alibaba Group, Hangzhou, Zhejiang, China.

[2]Centre for Infection and Immunity Study(CIIS), School of Medicine(Shenzhen), Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, Guangdong, China.

[3]Ministry of Education Key Laboratory of Biodiversity Science and Ecological Engineering, School of Life Sciences, Fudan University, Shanghai, China.

[4]Alibaba Health Information Technology, Alibaba Group, Hangzhou, Zhejiang, China.

[5]State Key Laboratory of Advanced Drug Delivery and Release Systems & Innovation Institute for AI in Medicine, College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang, China.

[6]Institute of Pathogen Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China.

[7]College of Pharmaceutical Sciences, The Second Affiliated Hospital, Zhejiang University School of Medicine, Zhejiang University, Hangzhou, Zhejiang, China.

[8]Sydney Institute for Infectious Diseases, School of Medical Sciences, The University of Sydney, Sydney, Australia.

[9]Department of Infectious Diseases and Public Health, Jockey Club College of Veterinary 33 Medicine and Life Sciences, City University of Hong Kong, Hong Kong, China.

[10]Key Laboratory of Pathogen Infection Prevention and Control(Peking Union Medical College, Ministry of Education), State Key Laboratory of Respiratory Health and Multimorbidity, National Institute of Pathogen Biology of Chinese Academy of Medical Science(CAMS)/Peking Union Medical College(PUMC), Beijing, China.

[11]School of Public Health(Shenzhen), Sun Yat-sen University, Shenzhen, Guangdong, China.
[12]State Key Laboratory for Biocontrol, the Centre for Infection and Immunity Studies, School of Medicine, Shenzhen Campus of Sun Yat-sen University, Sun Yat-sen University, Shenzhen, Guangdong, China.

*Corresponding author(s). E-mail(s): sanyuan.hy@alibaba-inc.com; shim23@mail.sysu.edu.cn; zhaorong.lzr@alibaba-inc.com;

Abstract

In recent years, significant advancements have been observed in the domain of Natural Language Processing(NLP) with the introduction of pre-trained foundational models, paving the way for utilizing similar AI technologies to interpret the language of biology. In this research, we introduce ”LucaOne”, a novel pre-trained foundational model designed to integratively learn from the genetic and proteomic languages, encapsulating data from 169,861 species encompassing DNA, RNA, and proteins. This work illuminates the potential for creating a biological language model aimed at universal bioinformatics application. Remarkably, through few-shot learning, this model efficiently learns the central dogma of molecular biology and demonstrably outperforms competing models. Furthermore, in tasks requiring inputs of DNA, RNA, proteins, or a combination thereof, LucaOne exceeds the state-of-the-art performance using a streamlined downstream architecture, thereby providing empirical evidence and innovative perspectives on the potential of foundational models to comprehend complex biological systems.

Keywords: Computational biology and bioinformatics, Computational Models, Machine Learning, Foundation Models

Main

From the discovery of DNA to the sequencing of every living form, the faithful rule-based flow of biological sequence information from DNA to RNA and protein has been the central tenet of life science. These three major information-bearing biopolymers carry out most of the work in the cell and then determine the structure, function, and regulation of diverse living organisms[1, 2].

The basic information in the threes is presented in a linear order of letters: four nucleotides for DNA or RNA and 20 standard and several non-standard amino acids for proteins. Their secondary or higher structure also contains information attributed to biological functions and phenotypes. This genetic principle resembles the human linguistic system. Darwin wrote in his ***The Descent of Man***: “The formation of different languages and distinct species, and the proofs that both have been developed through a gradual process, are curiously the same.”[3]. Various studies have testified to these parallels ever since, promoting the understanding and decoding of biological language[4–6].

Given the rapid advancements in machine learning technologies for human language processing, our efforts to decode biological language are bound to accelerate by leveraging insights from the former. The recent development of transformer architecture showed the superior capability of generalizing massive sequence-based knowledge from large-scale labeled and unlabeled data, which empowered language models and achieved unprecedented success in natural language processing(NLP) tasks. By pre-training on large datasets, foundational models learn the general characteristics of biological sequences. These models compute the input sequence into an embedding, a numerical representation that succinctly captures its semantic or functional properties. On this basis, various biological computation problems can be addressed through direct prediction, embedding analysis, or transfer learning[7]. In life science, substantial efforts have been put into adopting such language models, especially in protein tasks(ProTrans[8], ProteinBERT[9], ESM2[10], Ankh[11]), such as structure prediction[10, 12] and function annotation[13, 14]. In the realm of nucleic acid-focused tasks, several models have been introduced within niche areas(DNABert2[15], HyenaDNA[16], ScBert[17]). However, a broadly applicable, foundational model for nucleic acids remains elusive in widespread adoption across various disciplines.

Therefore, We have opted for a more fundamental and universal approach and developed a pre-trained, biological language semi-supervised foundation model, designated as ”LucaOne”, which integrates both nucleic acid(DNA and RNA) and protein sequences for concurrent training. This methodology allows the model to process and analyze data from nucleic acids and proteins concurrently, facilitating the extraction of complex patterns and relationships inherent in the processes of gene transcription and protein translation[18, 19].

We further examined that LucaOne exhibits an emergent understanding of the central dogma in molecular biology: the correlation between DNA sequences and their corresponding sequences of amino acids, which supports the notion that the concurrent training of nucleic acid and protein sequences together yields valuable insights[20]. To illustrate LucaOne’s practical effectiveness, we present seven distinct bioinformatics computational scenarios. These examples highlight LucaOne’s ease of use in real-world applications and demonstrate its superior performance to state-of-the-art(SOTA) models and other existing pre-trained models.

Results

LucaOne as a Unified Nucleic Acid and Protein Pre-trained Foundation Model

LucaOne was designed as a biological language foundation model through extensive pre-training on massive datasets, enabling the extraction of generalizable features for effective adaptation to various downstream tasks, therefore allowing researchers to efficiently employ pre-trained embeddings from LucaOne for a diverse range of bioinformatics analysis, even when there is limited training data, thereby significantly enhancing their performance. This model leverages a multifaceted computational training strategy that concurrently processes nucleic acids(DNA and RNA) and protein data from 169,861 species(only those that possess a minimum of 10 sequences within the training dataset are counted). Consequently, LucaOne possesses the capability to interpret biological signals and, as a foundation model, can be guided through input data prompts to perform a wide array of specialized tasks in biological computation.

Fig. 1 depicts the LucaOne framework, which adopts and enhances the Transformer-Encoder[21](**Methods A**). LucaOne's vocabulary consists of 39 unique tokens representing nucleotides and amino acids(**Methods B**). To make deep networks easier to train, pre-layer normalization supersedes post-layer normalization. Rotary Position Embedding(RoPE) replaces traditional absolute positional encoding for inferring longer sequences. Additionally, the model with mixed training distinguishes nucleotides and amino acids by utilizing token-type encoding, assigning 0 to nucleotides and 1 to amino acids.

To comprehensively assimilate the patterns and structures pervasive in universal biological language and the inherent knowledge these patterns convey, we have compiled an extensive collection of nucleic acid and protein datasets as the foundational pre-training material. RefSeq provided nucleic acid sequences, including DNA and RNA, and annotations for eight selected Genome region types and their order-level taxonomy. Protein data included sequences, annotations(from InterPro, UniProt, and ColabFoldDB), and tertiary structures(from RCSB-PDB and AlphaFold2)(**Fig. 2-a, Supplementary Table S1, Supplementary Fig. S1, and Supplementary Fig. S2**). A semi-supervised learning[19] approach was employed to enhance its applicability in biological language modeling. So, our pre-training tasks have been augmented with eight foundational sequence-based annotation categories. These annotations complement the fundamental self-supervised masking tasks, facilitating more effective learning for improved performance in downstream applications(**Fig. 2-b, Supplementary Table S2**). Overall, LucaOne comprised 20 transformer-encoder blocks with an embedding dimension of 2,560 and a total of 1.8 billion parameters. The downstream task utilized a model checkpoint at 5.6M. To illustrate the benefits of mixed training for nucleic acids and proteins, we trained the two additional models(LucaOne-Gene/LucaOne-Prot) separately using nucleic acids and proteins individually, and made a comparison using the same checkpoint in the central dogma of molecular biology task. Details of the pre-training data, pre-training tasks, and pre-training details refer to **Methods C, D, and E**, respectively.

We utilized t-distributed stochastic neighbor embedding(T-SNE) to visualize the embeddings from two distinct datasets: a nucleic acid dataset(S1), comprising sequences from 12 marine species, and a protein dataset(S2), consisting of sequences from 12 clans(Pfam clans are groups of protein families that are evolutionarily related and share similar structures and functions.). This visualization was compared to the results obtained using the Multi-OneHot, DNABert2[15], and ESM2-3B[10] embedding approaches. The outcomes, as illustrated in **Fig. 2(c∼i)**, revealed that the embeddings produced by LucaOne were more densely clustered, indicating that this method may encapsulate additional contextual information beyond the primary sequence data. (Dataset S1 and S2 details are in **Methods F**, and the embedding clustering metrics are in **Supplementary Table S3**).

Learning Central Dogma of Molecular Biology

Our additional objective was to account for known gene and protein sequences occupying a minuscule yet biologically active niche within their respective design spaces, with a subset of these sequences exhibiting correspondence based on the central dogma. Consequently, throughout the training phase of the LucaOne model, we refrained from explicitly encoding the interrelation between these two data types into the model, seeking to test whether the model inherently grasped the correlation between the genetic and protein data[22, 23].

We designed an experimental task to assess LucaOne's ability to recognize the inherent link between DNA sequences and their corresponding proteins. We have constructed a dataset comprising DNA and protein matching pairs derived from the NCBI-RefSeq database, with a proportion of 1:2 between positive and negative samples(**Fig. 3-a, 3-b**, **Methods G**). The samples were then randomly allocated across the training, validation, and test sets in a ratio of 4:3:25, respectively.

The study employed a simple downstream network to evaluate LucaOne's predictive capacity(**Fig. 3-c**). LucaOne encoded nucleic acid and protein sequences into two distinct, fixed embedding matrices(Frozen LucaOne). Then, each matrix was processed through pooling layers(either Max Pooling or Value-Level Attention Pooling[24]) to produce two separate vectors. The vectors were concatenated and passed through a dense layer for classification.

We compared the performance of different modeling approaches, including One-hot with a transformer, a transformer model with the random initialization, nucleic acid embeddings from DNABert2, protein embeddings from ESM2-3B, as well as two separate versions of the LucaOne foundation model trained independently on nucleic acid and protein sequences(LucaOne-Gene and LucaOne-Prot), and the unified training foundational version of LucaOne(**Fig. 3-d**). The findings indicated that modeling methods lacking pre-trained elements(One-hot and random initialization, see **Fig. 3-d**) were unable to acquire the capacity for DNA-protein translation in this dataset. In contrast, LucaOne's embeddings were able to effectively learn this capacity with limited training examples, and significantly surpassed both the amalgamation of the other two pre-trained models(DNABert2 + ESM2-3B) and the combined independent nucleic acid and protein LucaOne models using the same dataset, architecture, and checkpoint. This suggests that pre-trained foundational models can provide additional information beyond the specific task samples for such biological computation tasks. Moreover, LucaOne's unified training approach for nucleic acids and proteins enabled it to learn within a single framework, thereby capturing the fundamental intrinsic relationships between these two categories of biological macromolecules to some extent.

The analysis of LucaOne's performance across various sub-datasets unveiled several noteworthy findings. Firstly, both LucaOne and DNABert2 + ESM2-3B exhibit a gradual decline in both F1 score and accuracy as the number of exons in the dataset increases(more untranslated fragments). Notably, LucaOne demonstrates resilience to the increased exon count, suggesting a potentially enhanced capability to grasp fundamental principles of genetic information flow, such as codon-to-amino acid mapping and splice site identification(**Fig. 3-f**). Furthermore, when evaluating performance across datasets from different species, both models show consistent results, except for a notable decrease in performance with *Ciona intestinalis*. This deviation can largely be attributed to its unique codon usage patterns, significantly differing from other species in the study(**Fig. 3-e, 3-g**). Given the minimal sample size for this species in the dataset and with only 16% designated for training, it is likely that the models were unable to adequately learn the specific rules of the central dogma under these codon preferences, even though the analysis was conducted under the rule of The Standard Code. The observed divergence in codon preference suggests that *Ciona intestinalis* may possess more distinctive translation mechanisms from genetic material to proteins, which could be attributed to its unique evolutionary trajectory and selective pressures[25]. Based on this, it is inferred that with an expanded training data size encompassing a wider array of central dogma rules, LucaOne has the potential to more thoroughly assimilate the syntactical rules associated with genetic information processing, enabling its application to a more diverse set of scenarios.

LucaOne as a Foundation Model to Provide Embeddings for a Wide Range of Biological Computation Tasks

To ascertain the LucaOne model's capacity to provide effective embeddings for a variety of downstream tasks, we conducted validation studies across seven distinct downstream tasks, which include single-sequence tasks such as prediction of genus taxon(GenusTax), classification of ncRNA families(ncRNAFam), and the prediction of protein subcellular localization(ProtLoc) as well as the assessment of protein thermostability(ProtStab). For homogeneous paired-sequence tasks, we predicted influenza hemagglutination assays based on a pair of nucleic acid sequences(InfA) and assessed protein-protein interactions(PPI) utilizing pairs of protein sequences. Additionally, we forecasted the interactions between ncRNA and proteins(ncRPI) for heterogeneous sequences task. (Full task descriptions in **Methods H** and **Supplementary Table S4**).

For each task, we performed two types of comparative analyses: one against the state-of-the-art(SOTA) results and another using the same downstream network to assess LucaOne embeddings against the widely used nucleic acid and protein pre-trained language models, DNABert2 and ESM2-3B, respectively. These comparative analyses are instrumental in elucidating the incremental contributions of foundation models when addressing related analytical tasks and in evaluating the specific effectiveness of the embeddings generated by LucaOne with DNABert2 and ESM-3B.

Similarly, we used a simple downstream network to facilitate the processing of these tasks and illustrated the capacity of trained and frozen LucaOne to analyze nucleic acid(DNA and RNA) and protein sequences. **Fig. 4(a∼c)** displays the network architectures for three distinct input types. For tasks requiring paired inputs, a concatenation step is necessary to merge the output vectors of the pairs into a single extended vector. Finally, a fully connected(FC) layer was utilized for the ultimate output, which could be for classification or regression purposes.

Fig. 4(d∼k) displays a comparative analysis of performance on seven distinct biomedical tasks, revealing that LucaOne demonstrates superior representational capabilities over competing models in the GenusTax, ProtStab, ncRNAFam, InfA, and PPI evaluations, and comparable performance on the other two: ProtLoc and ncRPI. Notably, within the nucleic acid-centric GenusTax and ncRNAFam, LucaOne's accuracy has risen by 0.05 and 0.026, respectively, indicating a marked improvement over DNABert2. In the InfA task, LucaOne excelled with an exceptional accuracy of 1.0, reflecting its outstanding ability to represent this task data. For the ProtStab task, it surpassed ESM2-3B with a 0.015 increase in Spearman's Rank Correlation Coefficient(SRCC), and similarly showed a slight improvement in the protein-protein interaction(PPI) evaluation. Compared with DeepLocPro[26] in the task of ProtLoc, LucaOne was competitive with ESM2-3B and demonstrated a 0.025 accuracy improvement. Although LucaOne did not outperform the elaborate network model ncRPI-LGAT[27] in the ncRPI evaluation, it still exceeded the combined abilities of DNABert2 and ESM2-3B. LucaOne's effectiveness was particularly noteworthy in processing tasks involving heterogeneous sequences of nucleic acids and proteins; employing a unified representation model is advantageous compared to using separate models. The outcomes of these tasks underscored LucaOne's robust representational capabilities for both nucleic acid and protein sequences. LucaOne could enhance performance across a spectrum of downstream tasks, streamline networks for downstream tasks, and reduce computational resource demands(More hyperparameter comparison experiments results in **Methods I** and **Supplementary Table S5**).

Discussion

The attempt to build a universal biological language model is to develop a sophisticated cataloging and retrieval system for ”The Library of Mendel” - the genetic version of ”The Library of Babel.”[28, 29]. The diversity of genetic variations presents a vast ”design space” that is arguably as rich as the entirety of human literature, if not more so, given the far longer history of life on Earth compared to our record of literature. However, in stark contrast, the proportion of genetic sequences we have successfully identified and cataloged remains significantly smaller than the volume of documented human languages. Moreover, the growth of our understanding and documentation of this “biological language” is unlikely to occur suddenly or rapidly[30, 31]. Our endeavor herein offers a computational model that posits the potential to represent the paradigm of biological language. However, we must temper our expectations regarding this model’s rapid and seamless refinement toward an idealized state of perfection.

In developing the LucaOne model, we utilized deep learning frameworks and techniques from natural language processing. However, we observed systemic discrepancies when applying these models, which were highly successful in natural language contexts, to genomics language[32]. The architecture of BERT-based pre-trained language models focuses on understanding context but may not efficiently capture biological sequences’ unique attributes and characteristics[33, 34]. Furthermore, the functions and expressions of biological sequences are not solely determined by their genetic sequences but also by the environment in which they are expressed - a factor for which there is presently no practical modeling approach. Standardized methods for processing annotated or phenotypic data are lacking, which can lead to inaccuracies and omissions[35, 36]. Moreover, the continual learning and scalability aspects have yet to be fully explored in this study, primarily due to resource constraints. As a result, the complexities of the model’s learning capabilities have not been thoroughly examined at this point, highlighting the primary area of research for the subsequent phase[37]. In terms of application, due to the diversity of contexts, a robust evaluation system is absent for generalizability and domain adaptability, with small, specialized models occasionally outperforming large pre-trained models in conjunction with downstream tasks in certain areas[32, 38].

In light of these considerations, researchers may need to develop specialized pre-trained models tailored to genomic language to improve encoding and comprehension of biological data, ensuring adaptability across a broader spectrum of computational biology tasks. Promising directions include architectural innovations in pre-training models, such as incorporating genetic programming concepts into Large Language Models(LLMs)[39, 40]. Another avenue is harmonizing multimodal data, encompassing sequences, feature annotations, experimental outcomes, images, and phenotypical information to understand better biological systems beyond unsupervised sequence data learning[41, 42]. Additionally, employing more transparent algorithms may enhance model interpretability, facilitating better integration with existing biological research frameworks and model development[43, 44]. Lastly, given the necessity for pre-trained models to efficiently fine-tune or apply to downstream tasks, paradigms need to expedite model adaptation to new tasks and broader application contexts[32].

To conclude, this paper documented our effort to build a comprehensive large model to represent the intricacies of the biological world. The capabilities demonstrated by LucaOne showed considerable promise and highlighted several areas that necessitate substantial advancements. Such multimodal pre-trained foundational models, grounded in bioinformatics, will prove immensely valuable in accelerating and enhancing our comprehension of biological phenomena.

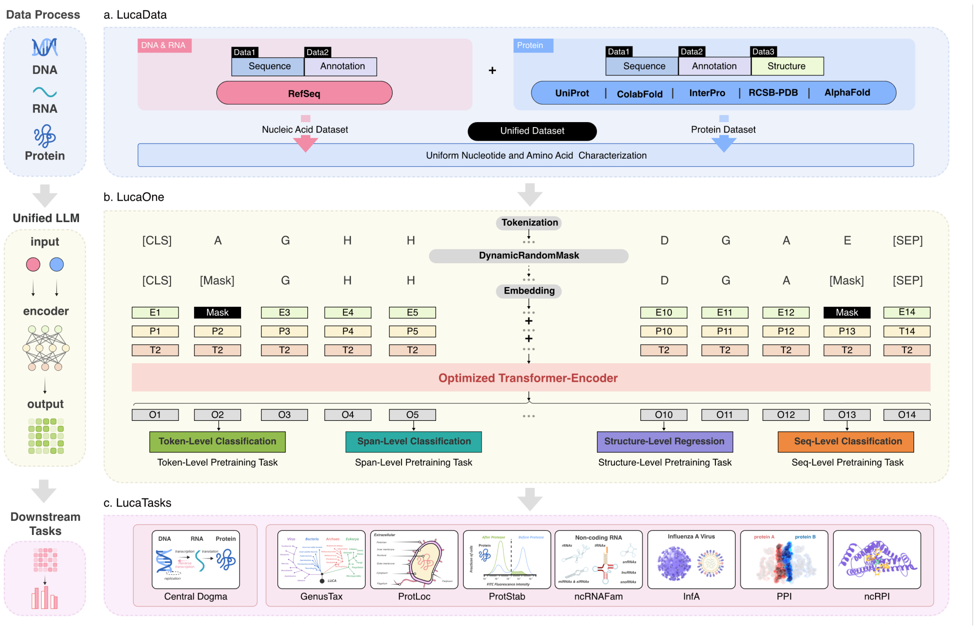

Data Process
DNA
RNA
Protein
a. LucaData
DNA & RNA
Data1
Sequence
Data2
Annotation
RefSeq
+
Protein
Data1
Sequence
Data2
Annotation
Data3
Structure
UniProt | ColabFold | InterPro | RCSB-PDB | AlphaFold
Nucleic Acid Dataset
Unified Dataset
Protein Dataset
Uniform Nucleotide and Amino Acid Characterization
Unified LLM
input
encoder
output
b. LucaOne
Tokenization
[CLS] A G H H D G A E [SEP]
DynamicRandomMask
[CLS] [Mask] G H H D G A [Mask] [SEP]
Embedding
E1 Mask E3 E4 E5 E10 E11 E12 Mask E14
P1 P2 P3 P4 P5 P10 P11 P12 P13 T14
T2 T2 T2 T2 T2 T2 T2 T2 T2 T2
Optimized Transformer-Encoder
O1 O2 O3 O4 O5 ... O10 O11 O12 O13 O14
Token-Level Classification
Token-Level Pretraining Task
Span-Level Classification
Span-Level Pretraining Task
Structure-Level Regression
Structure-Level Pretraining Task
Seq-Level Classification
Seq-Level Pretraining Task
Downstream Tasks
c. LucaTasks
Central Dogma
GenusTax
ProtLoc
ProtStab
ncRNAFam
InfA
PPI
ncRPI

Fig. 1: **The workflow of LucaOne. a.** Data source and processing for pre-training. The nucleic acid data was from RefSeq and included sequences and annotations, which consisted of order-level taxonomy and eight selected genome region types. Protein encompasses sequences(from Uniref50, UniProt, and ColabFoldDB-metagenomic protein collection(i.e. ColabFoldDB), where Uniref50 is clustered set of sequences from the UniProt with at least 50% sequence identity to enhance the learning of these representative sequences), annotations(order-level taxonomy from UniProt and ColabFoldDB, keywords from UniProt, and features such as sites, homologous superfamilies, and domains from InterPro), and tertiary structures(experimentally-determined structure from RCSB-PDB and predicted structure from AlphaFold2-Swiss-Prot). **b.** Pre-training model architecture and pre-training tasks. The Encoder is an improved transformer encoder. Based on two self-supervised mask tasks, an additional eight semi-supervised pre-training tasks were introduced to enhance the model's understanding of the data through annotations in the sequences. **c.** Downstream tasks for validation based on LucaOne embedding. The representational capabilities of LucaOne were verified using eight downstream tasks, whose inputs include DNA, RNA, proteins, and their interrelated pairs.

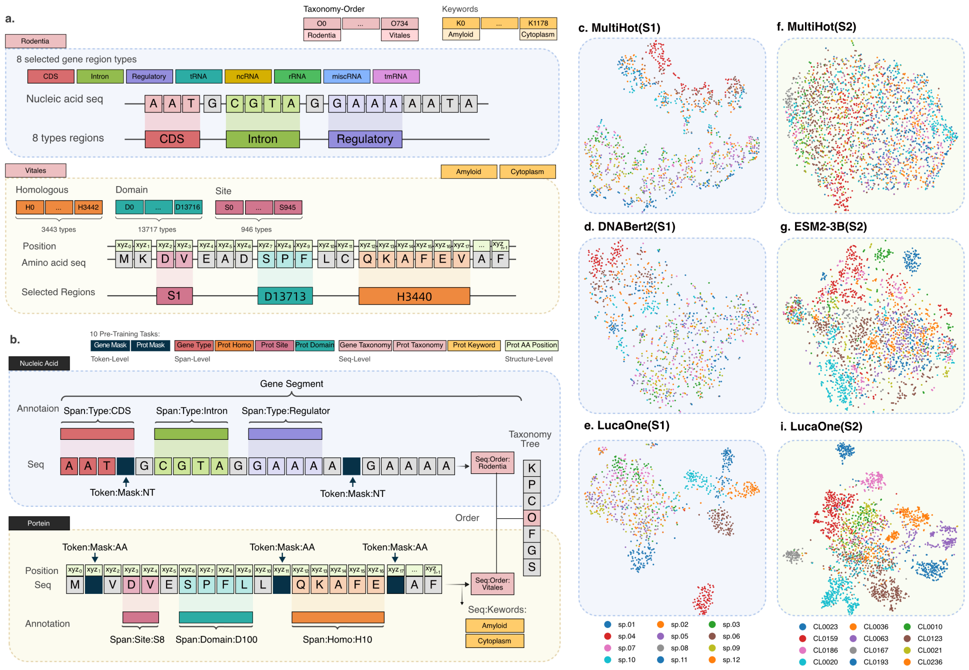
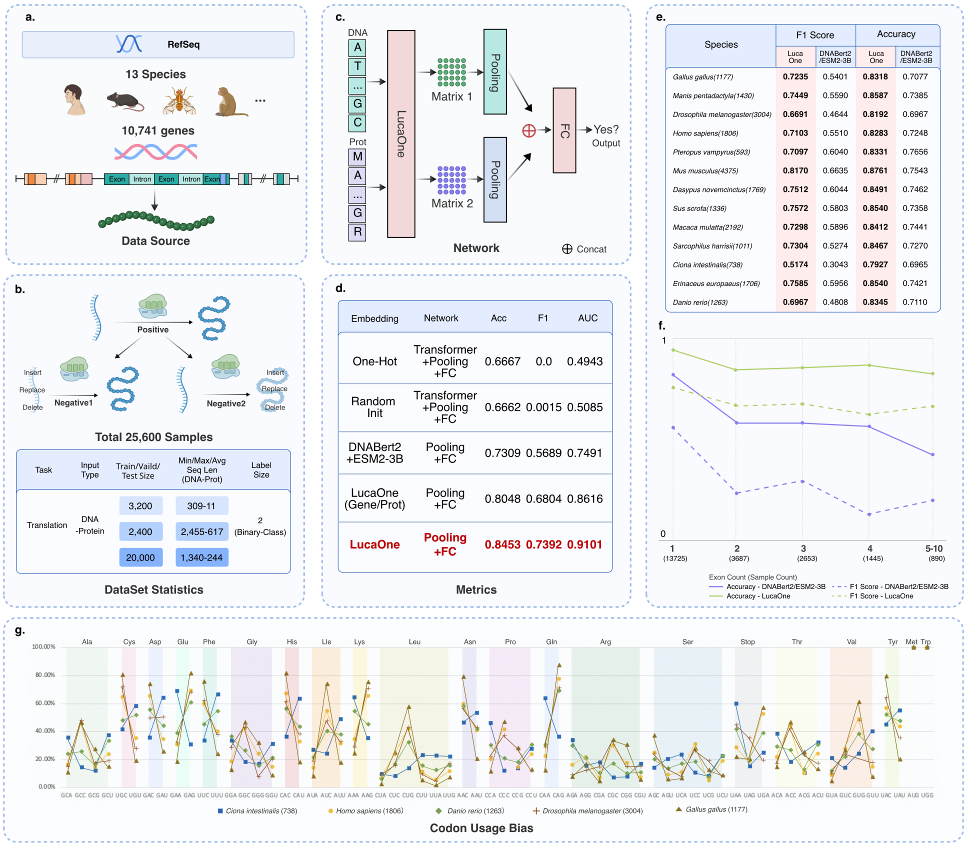


Fig. 2: **The data and tasks for pre-training LucaOne, and T-SNE on four embedding models. a.** Details of pre-training data. Nucleic acids included sequence and two kinds of annotation. The protein consisted of sequence, five types of annotation, and tertiary structure coordinates. **b.** Details of pre-training tasks. The pre-training tasks included two self-supervised mask tasks and eight semi-supervised tasks. **c.**∼**i.** T-SNEs of the four embedding methods on the S1-nucleic acid contigs with 12 species from the CAMI2 database, and S2-protein sequences across 12 clan categories from the Pfam database. The results show that LucaOne's representation has better clustering on these two datasets(nucleic acid sequences of the same species should be clustered because of high sequence similarity, and protein sequences of the same Pfam clan should be clustered of similar structures and functions). (**sp.01**: unclassified *Pseudomonas* species, **sp.02**: *Aeromonas salmonicida*, **sp.03**: unclassified *Vibrio* species, **sp.04**: *Streptomyces albus*, **sp.05**: *Aliivibrio salmonicida*, **sp.06**: unclassified *Brevundimonas* species, **sp.07**: *Vibrio anguillarum*, **sp.08**: *Aliivibrio wodanis*, **sp.09**: *Moritella viscosa*, **sp.10**: unclassified *Enterobacterales* species, **sp.11**: unclassified *Tenacibaculum* species, **sp.12**: unclassified *Aliivibrio* species).



Task	Input Type	Train/Vaild/ Test Size	Min/Max/Avg Seq Len (DNA-Prot)	Label Size
Translation	DNA -Protein	3,200	309-11	2 (Binary-Class)
		2,400	2,455-617	
		20,000	1,340-244	

Embedding	Network	Acc	F1	AUC
One-Hot	Transformer +Pooling +FC	0.6667	0.0	0.4943
Random Init	Transformer +Pooling +FC	0.6662	0.0015	0.5085
DNABert2 +ESM2-3B	Pooling +FC	0.7309	0.5689	0.7491
LucaOne (Gene/Prot)	Pooling +FC	0.8048	0.6804	0.8616
LucaOne	**Pooling +FC**	**0.8453**	**0.7392**	**0.9101**

Species	F1 Score		Accuracy	
	Luca One	DNABert2 /ESM2-3B	Luca One	DNABert2 /ESM2-3B
Gallus gallus(1177)	**0.7235**	0.5401	**0.8318**	0.7077
Manis pentadactyla(1430)	**0.7449**	0.5590	**0.8587**	0.7385
Drosophila melanogaster(3004)	**0.6691**	0.4644	**0.8192**	0.6967
Homo sapiens(1806)	**0.7103**	0.5510	**0.8283**	0.7248
Pteropus vampyrus(593)	**0.7097**	0.6040	**0.8331**	0.7656
Mus musculus(4375)	**0.8170**	0.6635	**0.8761**	0.7543
Dasypus novemcinctus(1769)	**0.7512**	0.6044	**0.8491**	0.7462
Sus scrofa(1336)	**0.7572**	0.5803	**0.8540**	0.7358
Macaca mulatta(2192)	**0.7298**	0.5896	**0.8412**	0.7441
Sarcophilus harrisii(1011)	**0.7304**	0.5274	**0.8467**	0.7270
Ciona intestinalis(738)	**0.5174**	0.3043	**0.7927**	0.6965
Erinaceus europaeus(1706)	**0.7585**	0.5956	**0.8540**	0.7421
Danio rerio(1263)	**0.6967**	0.4808	**0.8345**	0.7110

Fig. 3: **The workflow of the central dogma of molecular biology task. a.** Dataset from 13 species with 10,471 genes in RefSeq. **b.** Prepared 8,533 positive samples and 17,067 negative samples and took a specific sample dividing strategy to test the model performance in this task(training set: validation set: test set=4:3:25). **c.** Based on different embedding methods of DNA-Protein pair sequences, a simple downstream network was used for modeling and illustrating their representational ability. **d.** LucaOne was not only compared with several existing embedding methods but also with itself, which was trained using nucleic acids and proteins separately(LucaOne-Gene/LucaOne-Prot)). LucaOne, with mixing training, gave the best performance. **e.** Comparative Performance Analysis(Validation + Testing Dataset) of the models across diverse species datasets(Sample counts in brackets). **f.** Comparative performance analysis across different exon count subsets. **g.** One species for each Class was selected to undergo a codon usage bias analysis, which adheres to the conventions of the standard genetic code; this entails comparing the relative usage frequencies of different codons for each amino acid, ensuring that the total adds up to 100%. The species *Ciona intestinalis* exhibits a codon usage bias that is markedly distinct from that of other species - overall lower GC content. Details of the **Dataset Construction** and **Analysis of Misclassified Samples** are in **Methods G**.

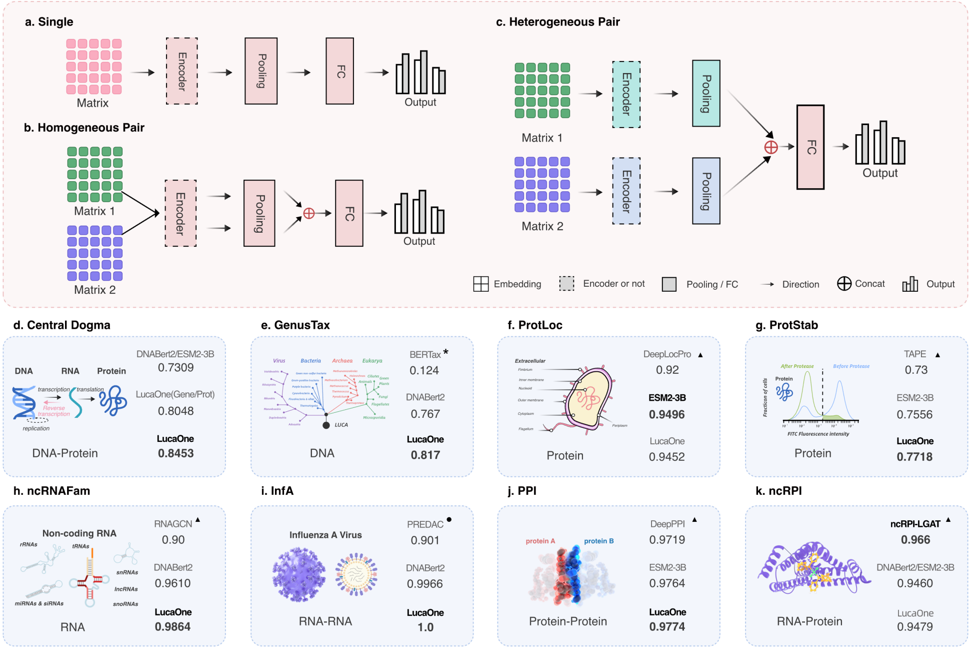


Fig. 4: **Downstream task networks with three input types and results comparison of 8 verification tasks.** Based on the embedding matrix, three types of inputs in the downstream task are corresponding networks: **a.** A single sequence, including GenusTax, ncRNAFam, ProtLoc, and ProtStab; **b.** Two same-type sequences, including InfA and PPI; **c.** Two heterogeneous sequences: Central Dogma and ncRPI. **d.**∼**k.** Comparison results of 8 downstream tasks. The Spearman Correlation Coefficient(SRCC) was employed for the ProtStab regression task, and Accuracy(Acc) was used for other tasks. Comparative methods include the SOTA, DNABert2-based(for nucleic acids), ESM2-3B-based(for proteins), and LucaOne-based. The top left ★ indicates inference using the trained method, the top left ▲ indicates direct use of the results in its paper, and the top left ● indicates repetition using its method and is better than the results in the paper.

Methods

A. Model Architecture

Fig. 1-b illustrates the design of LucaOne, which utilizes the Transformer-Encoder[21] architecture with the following enhancements:

1) The vocabulary of LucaOne comprises 39 tokens, including both nucleotide and amino acid symbols(refer to **Methods B**);

2) The model employs Pre-Layer Normalization over Post-Layer Normalization, facilitating the training of deeper networks[45];

3) Rotary Position Embedding(RoPE[46]) is implemented instead of absolute positional encoding, enabling the model to handle sequences longer than those seen during training;

4) It incorporates mixed training of nucleic acid and protein sequences by introducing token-type embeddings, assigning 0 for nucleotides and 1 for amino acids;

5) Besides the pre-training masking tasks for nucleic acid and protein sequences, eight semi-supervised pre-training tasks have been implemented based on selected annotation information(refer to **Methods D**).

B. Vocabulary

The vocabulary of LucaOne consists of 39 tokens. Due to the unified training of nucleic acid and protein sequences, the vocabulary includes 4 nucleotides('A', 'T', 'C', 'G') of nucleic acid('U' compiled with 'T' in RNA), 'N' for unknown nucleotides, 20 amino acids of protein(20 uppercase letters excluding 'B', 'J', 'O', 'U', 'X', and 'Z'), 'X' for unknown amino acids, 'O' for pyrrolysine, 'U' for selenocysteine, other 3 letters('B', 'J', and 'Z') not used by amino acids, 5 special tokens: '[PAD]', '[UNK]', '[CLS]', '[SEP]', '[MASK]', and 3 other characters:('.', '-', and '*'). Due to the amino acid letters overlapping with the nucleotide letters, the use of '1', '2', '3', '4', and '5' instead of 'A', 'T', 'C', 'G', and 'N', respectively.

C. Pre-training Data Details(Supplementary Fig. S1, Supplementary Fig. S2, and Supplementary Table S1)

Nucleic Acid

The nucleic acid was collected from Refseq, involving 297,780 assembly accessions. The molecular types included DNA and RNA(**Fig. 2-a**). DNA sequence, DNA selected annotation, RNA sequence, and RNA selected annotation were obtained from the format files 'genomic.fna', 'genomic.gbff', 'rna.gbff', and 'rna.fna', respectively.

DNA reverse strand: The DNA reverse strand also contains a lot of annotation information, so the DNA dataset expanded reverse strand sequences with their annotation. A total of 23,095,687 DNA reverse-strand sequences were included.

Genome region types: Eight important genome region types in nucleic acids were selected, including 'CDS', 'intron', 'tRNA', 'ncRNA', 'rRNA', 'miscRNA', 'tmRNA', and 'regulatory'. Each nucleotide in the sequence had a label index of 8 categories(0$\sim$7) or *-100* when it did not belong to these 8 categories.

Order-level taxonomy: The order-level label of the taxonomy tree was selected as the classification label of the nucleic acid sequence. Each sequence had a label index of 735 categories(0∼734) or *-100* without the order-level taxonomy.

Segmentation: Due to the limited computing resources, each nucleic acid sequence was segmented according to a given maximum length. The fragmentation included the following three scenarios(**Supplementary Fig. S3**).

Case 1: Only the genome regions were selected in the sequence for processing. If the length of consecutive genome regions did not exceed the maximum, merge them. It was segmented for genome regions exceeding the maximum length, and the fragment position should be ensured as far as possible in non-7 important regions(excluding intron regions). If a fragment was less than the maximum length, the fragment was expanded left and right simultaneously.

Case 2: The entire sequence without genome regions was processed. If the length exceeded the maximum, it was fragmented in the same way as in **case 1**.

Case 3: For sequences without genome regions and 8 essential regions, the entire sequence was fragmented at no more than the specified maximum length.

Protein

As shown in **Fig. 2-a**, compared with nucleic acids, proteins contained more information, including sequence, taxonomy, keywords, sites, homology regions, domains, and tertiary structure. Sequence, taxonomy, and keywords were collected from UniProt and Colab-FoldDB. Sites, domains, and homology regions were extracted from Interpro. The tertiary structure was derived from RCSB-PDB and AlphaFold2-Swiss-Prot.

Sequence: The *right truncation* strategy was applied when the sequence exceeded the maximum length.

Order-level taxonomy: Order-level classification information is used as the protein sequence taxonomy. There were 2,196 categories; each sequence had a label index(0∼2,195) or *-100* if its order-level information was missing.

Site: Four types of site regions('Active site', 'Binding site', 'Conserved site', and 'PTM') with 946 categories were included. For each amino acid in a sequence, if it was a site location, there was a label index(0∼945); otherwise, it was marked with *-100.*

Homology: A homologous superfamily is a group of proteins that share a common evolutionary origin with a sequence region, reflected by similarity in their structure. There were 3,442 homologous region types; each amino acid in these regions had a label index(0∼3,441) corresponding to its type, and the other amino acids were labeled *-100.*

Domain: Domain regions are distinct functional, structural, or sequence units that may exist in various biological contexts. A total of 13,717 domain categories were included; each amino acid in these regions had a label index(0∼13,716) corresponding to its category, and the other amino acids were marked with *-100.*

Keyword: Keywords are generated based on functional, structural, or other protein categories. Each sequence was labeled as a label-index set with 1,179 keywords or *-100* without keywords.

Structure: The spatial coordinates of the C_α*-atom* were used here as the coordinates of the amino acids. Each amino acid was labeled by a three-dimensional coordinate normalized within the protein chain. Amino acids at missing locations were labeled *(-100, -100, -100).*

We obtained the experimentally-determined structure in the RCSB-PDB and the predicted structure by AlphaFold2 of UniProt(Swiss-Prot) and preferentially selected the structure in RCSB-PDB. In total, only about half a million protein sequences had structural information.

D. Pre-training Tasks Details

LucaOne has employed a semi-supervised learning approach to enhance its applicability in biological language modeling. Unlike the traditional natural language machine learning domain, where tasks involve input and output of the same textual modality, bioinformatics analysis often involves different modalities for input and output data. Most bioinformatics downstream tasks extend from understanding nucleic acid or protein sequences, so our pre-training tasks have been augmented with eight foundational sequence-based annotation categories. These annotations complement the fundamental self-supervised masking tasks, facilitating more effective learning for improved performance in downstream applications. The selection criteria for these annotations focused on universality, lightweight design, and a high confidence level; consequently, only a subset of the data possesses such annotations. According to **Supplementary Table S2**, there are 10 specific pre-training tasks at four levels.

Token Level Tasks: Gene Mask and Prot Mask tasks randomly mask nucleotides or amino acids in the sequence following the BERT[47] masking scheme and predict these masked nucleotides or amino acids based on the sequence context in training.

Span Level Tasks: The model is trained to recognize some essential regions based on the sequence context. For nucleic acid sequences, eight essential genome region types are learned. For protein sequences, three kinds of regions are labeled: site, homology, and domain regions.

Seq Level Tasks: Gene-taxonomy, Prot-taxonomy, and Prot-keyword are the order-level taxonomies of nucleic acid, protein, and protein-tagged keywords, respectively. They are all sequence-level learning tasks.

Structure Level Tasks: Since the structure of a protein determines its function, we use a small amount of protein data with a tertiary structure for simple learning in the pre-training phase. Instead of fine-tuned atomic-level spatial position learning, the spatial position of the amino acids is trained(using the position of C_α*-atom* as the position of the amino acid).

Data Processing: All data processing tasks were conducted on the Alibaba Cloud MaxCompute platform.

E. Pre-training Information

On the dimensions of the embedding, the research conducted by Elnaggar et al. [11] demonstrates that the ESM2-3B model, with an embedding dimension of 2,560, notably enhances performance compared to its counterpart, ESM2-650M, harbors an embedding dimension of 1,280. However, when the parameter count escalates to 15 billion with an embedding dimension of 5120, there is no significant increment in performance. It is also observed that the ESM2-15B model requires substantially more time than ESM2-3B and ESM2-650M concerning the relationship between input sequence length and computational time. For the relationship between model size and training data size, Hoffmann et al. suggest that a minimum of 20.2 billion tokens is essential to train a 1B-sized model adequately[48].

Taking into account these insights, along with the consideration of available computational resources and the volume of pre-training data, some of the critical hyperparameters we adopted are as follows: the architecture of LucaOne consists of 20 transformer-encoder blocks with 40 attention heads each, supports a maximal sequence length of 1,280, and features an embedding dimension of 2,560. The model is composed of a total of 1.8 billion parameters. We employed 10 distinct pre-training tasks, allocating an equal weight of 1.0 to Gene-Mask, Prot-Mask, Prot-Keyword, and Prot-Structure tasks while assigning a reduced weight of 0.2 to the remaining tasks to equilibrate task complexity. The learning rate was 2e-4, which leverages a warm-up approach throughout the updates. For the model's training regimen, we utilized a batch size of 8 coupled with a gradient accumulation step of 32. The model underwent training on 8 Nvidia A100 GPUs spanning 120 days. A model checkpoint of 5.6 million(5.6M) was selected for downstream validation tasks, aligning with ESM2-3B in terms of the volume of data trained for comparison.

To elucidate the advantages of mixed training involving both nucleic acids and proteins, we further conducted experiments with two supplementary models, LucaOne-Gene and LucaOne-Prot, trained exclusively with nucleic acids and proteins, respectively. Their performances in the central dogma of the biology task were evaluated with the same checkpoint(5.6M) of the two models.

F. Details of T-SNE Datasets

Dataset Construction: The S1 dataset was curated from marine data available in CAMI2[49], selecting contigs with lengths ranging from 300 to 1,500 nucleotides. We focused on species that were identifiable and possessed at least 200 contigs. Each species' contigs were deredundant by MMSeqs, employing a coverage threshold of 80% and sequence identity of 95%, culminating in a collection of 37,895 nucleic acid contigs from 12 species. We randomly selected 100 samples from each species, totaling 1,200 items for visualization. The S2 dataset originated from clan data within Pfam, maintaining clan categories with a minimum of 100 Pfam entries, resulting in 189,881 protein sequences across 12 clan categories. For visualization, we randomly selected one sample for each Pfam entry under every clan, amounting to 2,738 samples.

G. Details of Central Dogma Dataset

Dataset Construction: We devised an experimental task to determine whether LucaOne has established the intrinsic association between DNA sequences and their corresponding proteins. A total of 8,533 accurate DNA-protein pairs were selected from 13 species in the NCBI-RefSeq database, with each DNA sequence extending to include an additional 100 nucleotides at both the 5' and 3' contexts, preserving intron sequences within the data. In contrast, we generated double the amount of negative samples by implementing substitutions, insertions, and deletions within the DNA sequences or altering amino acids in the protein sequences to ensure the resultant DNA sequences could not be accurately translated into their respective proteins. For more details, see Section - **Data and Code Availability**

Analysis of Misclassified Samples: According to **Fig. 3-d**, an acceptable accuracy(84.53%) was achieved by learning from fewer samples and testing more samples. We analyze the misidentified samples from two perspectives: sequence and embedding. **Supplementary Fig. S4-a and S4-b** analyzed these misclassified samples using sequence similarity. We constructed two negative samples for each positive sample by inserting, replacing, and deleting nucleotide and amino acid letters. The global alignment algorithm was applied to calculate the similarity(normalized to 0$\sim$1) between the edited nucleic acid sequence and the original nucleic acid sequence, and the same calculation was performed for the protein. **Supplementary Fig. S4-a** shows that the average similarity of the nucleic acid of false negatives(FN) is close to that of the correctly classified samples. **Supplementary Fig. S4-b** shows that the average similarity of proteins in false positives(FP) and false negatives(FN) is higher than in the correctly classified samples. These indicate that the sequence differences of nucleic acids or proteins in negative samples are too subtle to classify. **Supplementary Fig. S4-c and S4-d** analyzed these misclassification samples by the embedding. **Supplementary Fig. S4-c** shows that the Euclidean distance between the nucleic acid and protein was calculated using LucaOne's embedding. Compared with the samples that can be accurately classified, the distance gap between positive and negative samples in the false positives(FP) and false negatives(FN) samples is minimal, making these samples challenging to classify. **Supplementary Fig. S4-d** shows that the two-dimensionally reduced vectors(2560 $\rightarrow$ 128) output by the pooler layer in the downstream network(Pooling + FC) are calculated for cosine similarity. The cosine similarity distinction of incorrect samples is also tiny, indicating that these samples are difficult to classify. Therefore, more data should be adopted to allow the model to learn these more subtle differences. For more details, see Section -**Data and Code Availability**.

H. Downstream Tasks Details

Genus Taxonomy Annotation(GenusTax): This task is to predict which genus(taxonomy) the nucleic acid fragment may come from, which is very important in metagenomic analysis. A comparative dataset was constructed utilizing NCBI RefSeq, comprising 10,000 nucleotide sequences, each extending 1,500 nucleotides, and annotated with labels corresponding to 157 distinct genera. The dataset was randomly segregated into training, validation, and test sets, adhering to an 8:1:1 partitioning ratio. This investigation constitutes a multi-class classification challenge.

Prokaryotic Protein Subcellular Location(ProtLoc): This task is to predict the subcellular localization of proteins within prokaryotic cells, which has garnered substantial attention in proteomics due to its critical role[50]. It involves classifying proteins into one of six subcellular compartments: the cytoplasm, cytoplasmic membrane, periplasm, outer membrane, cell wall and surface, and the extracellular space. Our approach adopted the same dataset partitioning strategy as DeeplocPro[26], a model based on experimentally verified data from the UniProt and PSORTdb databases. Our research undertakes a multi-class classification challenge, categorizing proteins based on their distinct subcellular localizations.

Protein Stability(ProtStab): The evaluation of protein stability is paramount for elucidating the structural and functional characteristics of proteins, which aids in revealing the mechanisms through which proteins maintain their functionality in vivo and the circumstances predisposing them to denaturation or deleterious aggregation. We utilized the same dataset from TAPE[51], which includes a range of denovo-designed proteins, natural proteins, mutants, and their respective stability measurements. It is a regression task; each protein input(x) correlates with a numerical label($y \in \Re$), quantifying the protein's intrinsic stability.

Non-coding RNA Family(ncRNAFam): Non-coding RNA(ncRNA) represents gene sequences that do not code for proteins but have significant functional and biological roles. The objective is to assign ncRNA sequences to their respective families based on their characteristics. For this purpose, we utilize the dataset from the nRC[52], which is consistent with the data employed in the RNAGCN[53] study. Our methodology adheres to the same data partitioning into training, validation, and test sets as done in these previous studies, enabling direct comparison of results. This project involves a multi-class classification challenge encompassing 88 distinct categories.

Influenza A Antigenic Relationship Prediction(InfA): One of the foremost tasks in influenza vaccine strain selection is monitoring Hemagglutinin(HA) variant emergence, which induces changes in the virus's antigenicity. Precisely predicting antigenic responses to novel influenza strains is crucial for developing effective vaccines and preventing outbreaks. The study utilizes data from the PREDAC[54] project to inform vaccine strain recommendations. Each data pair in this study comprises two RNA sequences of the HA fragment from distinct influenza strains, accompanied by corresponding antigenic relationship data. The objective is framed as a binary classification task, determining the antigenic similarity or difference between virus pairs.

Protein-Protein Interaction(PPI): The forecasting of protein-protein interaction networks represents a significant area of research interest. Our study utilized the DeepPPI[55] database, whose positive dataset samples were sourced from the Human Protein Reference Database after excluding redundant interactions, leaving 36,630 unique pairs. This dataset was randomly partitioned into three subsets: training(80%), validation(10%), and testing(10%). The primary objective of this research is to perform binary classification of protein-protein interaction sequences.

ncRNA-Protein Interactions(ncRPI): An increasing number of functional non-coding RNAs (ncRNAs), such as snRNAs, snoRNAs, miRNAs, and lncRNAs, have been discovered. ncRNAs play a crucial role in many biological processes. Experimentally identifying ncRNA-protein interactions(ncRPI) is typically expensive and time-consuming. Consequently, numerous computational methods have been developed as alternative approaches. For comparison, we have utilized the same dataset as the currently best-performing study, ncRPI-LGAT[27]. It is a binary classification task involving pairs of sequences.

I. Comparison Result Detaills

We conducted a series of comparative experiments. According to **Fig. 4**, for all embedding methods, we compare whether the transformer encoder and two pooling strategies(max pooling and value-level attention pooling) were used on the model. At the hyperparameter level, we compared the number of encoder layers with the number of heads(4 layers with 8 heads and 2 layers with 4 heads), the maximum learning rate of the Warmup strategy(1e-4 and 2e-4), and the batch size(8 and 16). **Supplementary Table S5** shows the result of comparing whether the encoder was used and which pooling method was used accordingly.

In the **ProtLoc** task, LucaOne's accuracy is very close to that of the ESM2-3B; In the **ncRPI** task, the accuracy of the simple network with LucaOne's embedding matrix is less than that of ncRPI-LGAT[27] but higher than that of DNABert2 + ESM2-3B; In **the other five tasks**, we achieved the best results. It is better not to use an encoder for **ProtLoc**, **InfA**, **PPI**, and **ncRPI** tasks. Using the Max Pooling strategy straightforwardly for the **ncRNAFam** and **GenusTax** tasks can obtain better results. We extended two tasks, four superkingdoms, and 180 species prediction tasks for the genus classification task with the same sequence data. LucaOne's accuracy improved by 0.1 and 0.054, respectively. In particular, LucaOne is more effective than other large models in embedding sequences without an encoder.

Acknowledgment

We thank J. Wang, D.-C. Ma, and D.-Z. Shi for computational resources coordination. We thank H.-W. Zhang for maintaining the computational resources and optimizing specific computing tasks at Yunqi Academy of Engineering(Hangzhou). We thank X.-J. Du, W.-C. Wu, J.-Y. Yang, and S.-Q. Mei from Sun Yat-sen University(Shenzhen) for a valuable conversation on the development of the method, especially on understanding the downstream tasks. Last but not least, we thank Charles Darwin, Richard Dawkins, Steven Pinker, and Daniel Dennett for their profound insights that led to the early conceptual foundations of this study and guided its development pathway.

Computational Resources

The data processing and training operations for LucaOne were executed on Alibaba Cloud Computing Co., Ltd. Additionally, several tasks related to further processing or downstream computing were carried out on alternative computing platforms: Yunqi Academy of Engineering(Hangzhou, China) and Zhejiang Lab(Hangzhou, China).

Author Contributions

Conceptualization: Y. H., Z.-R. L., and M. S.; Model development and data preparation for LucaOne: Y. H., Y.-H. W., Y.-F. P., and Y.-C. C.; Downstream tasks understanding and models training: Y. H., P. F., Y.-T. S., and Y.-H. C.; Original draft: Y. H., Z.-R. L., and P. F.; Writing - Review and Editing: All authors; Graphic presentation design: Y. L. and Y. H.; Engineering leadership and resource acquisition: Z.-Y. Z. and J.-P. Y.; Science leadership and resource acquisition: J. L., E.-C. H., Z. Z., F. Z., and Y.-L. S.; Supervision: Y. H., M. S., and Z.-R. L..

Competing Interests

Y. H., Z.-R. L., P. F., and J.-P. Y. have filed an application for a patent covering the work presented. The other authors declare no competing interests.

Data and Code Availability

The LucaOne's model code is available at: LucaOne Github or LucaOne. The trained-checkpoint files are available at: TrainedCheckPoint. LucaOne's representational inference code is available at: LucaOneApp Github or LucaOneApp. The project of 8 downstream tasks is available at: LucaOneTasks Github or LucaOneTasks. The pre-training dataset of LucaOne is opened at: PreTrainingDataset. The datasets of downstream tasks are available at: DownstreamTasksDataset. Other supplementary materials are available at: Others.

Table S1: **The statistics on pre-training data.** The inclusion of Uniref50, a subset of Uniprot, is to enhance the learning of representative sequences. The explanation of terms(Genome Region Type, Order-level Taxonomy, Keyword, Site, Domain, Homology, and Structure) is in Methods C.

Data Type/Source	Seq Count	Seq Length (Min/Max/Mean/Median)	Seq Count in which Annotation Exist
DNA/Refseq	1,181,133,873	3/1,280/987/1,280	Genome Region Type: 1,124,090,054
			Order-level Taxonomy: 1,180,785,987
RNA/Refseq	136,311,178	3/1,280/1,035/1,280	Genome Region Type: 106,918,452
			Order-level Taxonomy: 135,933,074
Protein/Uniref50	62,150,523	13/45,356/286/194	Order-level Taxonomy: 58,537,832
			Keyword: 35,188,544
			Site: 3,341,857
			Domain: 21,981,013
			Homology: 24,246,265
			Structure: 190,861
Protein/UniProt	252,170,925	3/45,356/352/281	Order-level Taxonomy: 242,706,004
			Keyword: 192,226,173
			Site: 37,449,976
			Domain: 141,276,963
			Homology: 156,909,761
			Structure: 569,034
Protein/ColabFoldDB	208,966,064	4/36,993/182/147	Order-level Taxonomy: 5,389,742
			Keyword: 3,117,716
			Site: 343,344
			Domain: 1,972,216
			Homology: 2,136,440
			Structure: 6,715

Table S2: **The pre-training tasks.** The detailed description of all pre-training tasks is in Methods D.

Task Level	Task	Task Type	Label Size	Description
Token Level	Gene Mask	Multi-Class	39	Nucleotide mask
	Prot Mask	Multi-Class	39	Amino acid mask
Span Level	Genome Region Types	Multi-Class	8	Selected genome region types
	Prot Site	Multi-Class	946	Site region
	Prot Homo	Multi-Class	3,443	Homology region
	Prot Domain	Multi-Class	13,717	Domain region
Seq Level	Gene Taxonomy	Multi-Class	735	Order-level taxonomy
	Prot Taxonomy	Multi-Class	2,196	Order-level taxonomy
	Prot Keyword	Multi-Label	1,179	Protein keyword
Structure Level	Prot Structure	Regression	-	Amino-acid level position

Table S3: **Clustering metrics of the four embedding methods on the S1 and S2 datasets.**

DataSet	Embedding	ARI	AMI	V-measure	FMI
12-Marine Species(S1)	Multi-OneHot	0.1106	0.2704	0.2861	0.1908
	DNABert2	0.1450	0.2624	0.2782	0.2218
	LucaOne	**0.3445**	**0.5425**	**0.5525**	**0.4078**
12-Clan Pfam(S2)	Multi-OneHot	0.1042	0.1765	0.1854	0.2330
	ESM2-3B	0.0549	0.1057	0.1140	0.1451
	LucaOne	**0.2313**	**0.4060**	**0.4116**	**0.3084**

Table S4: **Details on downstream validation tasks.**

Task	Task Type	Input Type	Train/Valid/ Test Size	Seq Length (Max/Min/Median)
Central Dogma	Binary-Class(2)	DNA-Protein	3,200/2,400/20,000	2,455-617/ 309-11/ 1,273-260
GenusTax	Multi-Class(157)	DNA	8,000/1,000/1,000	1,500/1,500/1,500
ProtLoc	Multi-Class(6)	Protein	9,915/1,991/1,131	5,627/8/396
ProtStab	Regression	Protein	53,614/2,512/12,851	50/43/43
ncRNAFam	Multi-Class(88)	RNA	105,864/17,324/25,342	200/24/114
InfA	Binary-Class(2)	RNA-RNA	4,645/581/581	1,690-1,690/ 984-984/ 1,095-1,095
PPI	Binary-Class(2)	Protein-Protein	59,766/7,430/7,425	33,423-33,423/ 24-24/ 465-437
ncRPI	Binary-Class(2)	RNA-Protein	16,658/-/4,166	3,999-3,678/ 52-49/ 1,858-414

Table S5: Detailed results on downstream validation tasks(results of the better pooling method for each task with or without encoder). The top left ★ indicates inference using the trained method, the top left ▲ indicates direct use of the results in its paper, and the top left ● indicates repetition using its method and higher than the results in the paper.

Task	Input	Method	Encoder	Better Pooling	Acc/SRCC
SpeciesTax	DNA	BERTax★[56]	-	-	-
		DNABert2	W/O Encoder	Attention	0.519
			Encoder	Max	0.696
		LucaOne	W/O Encoder	Attention	0.713
			Encoder	Attention	**0.750**
GenusTax	DNA	BERTax★[56]	-	-	0.124
		DNABert2	W/O Encoder	Attention	0.551
			Encoder	Max	0.767
		LucaOne	W/O Encoder	Attention	0.765
			Encoder	Max	**0.817**
SupKTax	DNA	BERTax★[56]	-	-	0.816
		DNABert2	W/O Encoder	Attention	0.805
			Encoder	Attention	0.848
		LucaOne	W/O Encoder	Attention	0.940
			Encoder	Attention	**0.947**
ProtLoc	Protein	DeepLocPro▲[26]	-	-	0.92
		ESM2-3B	W/O Encoder	Attention	**0.9496**
			Encoder	Max	0.9408
		LucaOne	W/O Encoder	Attention	0.9452
			Encoder	Max	0.9310
ProtStab	Protein	TAPE▲[51]	-	-	0.73
		ESM2-3B	W/O Encoder	Attention	0.7556
			Encoder	Attention	0.7102
		LucaOne	W/O Encoder	Attention	0.7512
			Encoder	Attention	**0.7718**
ncRNAFam	RNA	RNAGCN▲[53]	-	-	0.90
		DNABert2	W/O Encoder	Attention	0.9036
			Encoder	Max	0.9610
		LucaOne	W/O Encoder	Attention	0.9743
			Encoder	Max	**0.9864**
InfA	RNA-RNA	PREDAC●[54]	-	-	0.9010
		DNABert2	W/O Encoder	Attention	0.9966
			Encoder	Attention	0.9966
		LucaOne	W/O Encoder	Attention	**1.0**
			Encoder	Attention	0.9983
PPI	Protein-Protein	DeepPPI▲[55]	-	-	0.9719
		ESM2-3B	W/O Encoder	Attention	0.9764
			Encoder	Attention	0.9745
		LucaOne	W/O Encoder	Attention	**0.9774**
			Encoder	Attention	0.9751
ncRPI	RNA-Protein	ncRPI-LGAT▲[27]	-	-	**0.966**
		DNABert2 + ESM2-3B	W/O Encoder	Attention	0.9460
			Encoder	Attention	0.9332
		LucaOne	W/O Encoder	Attention	0.9479
			Encoder	Attention	0.9380

Table S6: **Terms & Abbreviations Definitions**

Terms & Abbreviation	Definition
NLP	Natural Language Processing
SOTA	State of the Art
Transformer	a Popular Deep Learning Architecture
BERT	Bidirectional Encoder Representations from Transformers
RoPE	Rotary Positional Encoding
CDS	Coding Sequence
Intron	a Segment of DNA not Translated into Protein
tRNA	Transfer RNA
ncRNA	Non-Coding RNA
rRNA	Ribosomal RNA
miscRNA	Miscellaneous RNA
tmRNA	Transfer-Messenger RNA
Regulatory	the Regulation of Gene Expression
MMSeqs	Many-against-Many Sequence Searching
OneHot	Convert a Unique Category Value into a Binary Vector
Multi-OneHot	Convert Multiple Categories into a Binary Matrix based on OneHot
Taxonomy	a Hierarchical Classification System in Biology
Site	Protein Site Region
Homology	a Common Evolutionary Origin in Proteins
Domain	a Distinct Functional, Structural, or Sequence Unit
Keyword	Keywords for Protein Annotation
Structure	Three-Dimensional Arrangements of Atoms within a Protein
C_α*-atom*	the Carbon that is Next to a Functional Group
RefSeq	NCBI Reference Sequence Database
UniProt	the Universal Protein Knowledgebase
UniRef	Comprehensive and Non-redundant UniProt Reference Clusters
ColabFoldDB	MMseqs2 Expandable Profile Databases of Proteins
InterPro	Classification of Protein Families
RCSB-PDB	Protein Data Bank
AlphaFold2	AlphaFold2 Protein Structure Database
Pfam	a Database with Large Collection of Protein Families
CAMI2	Critical Assessment of Metagenome Interpretation II
ESM2-3B	a Protein Language Model with 3B Parameters
DNABert2	a Transformer-based Genome Foundation Model
Pooling	Downsampling Operation that Reduces the Dimensionality of the Feature Map
Max Pooling	Each Feature Retains the Maximum Value in Pooling
Value-Level Attention Pooling	Using the Attention Mechanism in Pooling
FC	Fully Connected Layer
Warm-up	Gradually Increasing the Learning Rate During the Initial Stages of Training
TP/TN/FP/FN	Four Metrics in the Confusion Matrix
SRCC	Spearman's Rank Correlation Coefficient
T-SNE	t-Distributed Stochastic Neighbor Embedding
ARI	Adjusted Rand Index Score
AMI	Adjusted Mutual Information Score
V-measure	the Harmonic Mean between Homogeneity and Completeness
FMI	the Fowlkes-Mallows Index Score

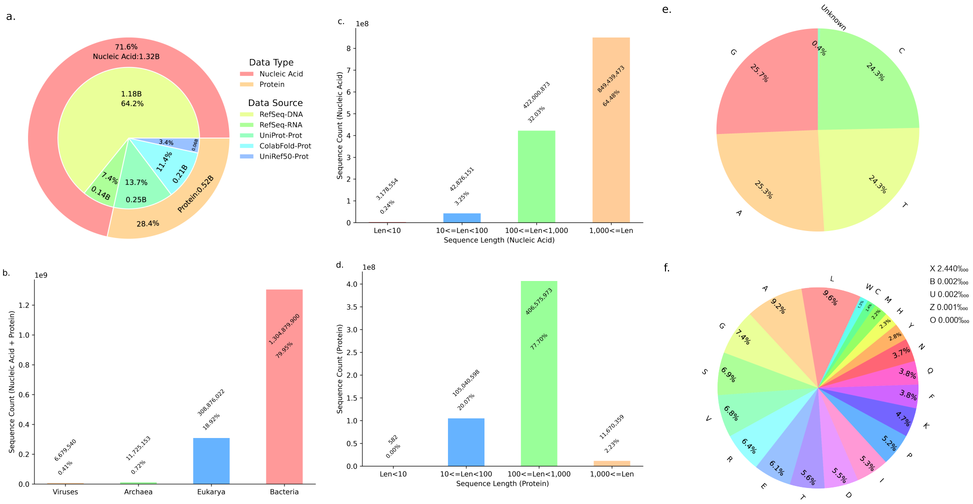


Fig. S1: **Overall statistics on pre-training data of LucaOne. a.** Sequences(DNA, RNA, and proteins) were derived from RefSeq, UniProt, ColabFoldDB, and UniRef50. **b.** The data(nucleic acids and proteins) involved four superkingdom types: Viruses, Archaea, Eukarya, and Bacteria, of which Bacteria accounted for the most. **c.** The sequence length distribution of nucleic acids, with the most being more than 1,000. **d.** The sequence length distribution of proteins, with the maximum length ratio between 100 and 1,000. **e.** The proportion of five nucleotides('A', 'T', 'C', 'G', and 'Unknown') in nucleic acid sequences('U' compiled with 'T' in RNA) and the four identified nucleotides were close in proportion. **f.** The proportion of the 20 standard amino acid letters and 5 other letters (including 4 non-standard amino acids and 'X' for unknown amino acid) in the protein sequence, and Leucine has the highest proportion.

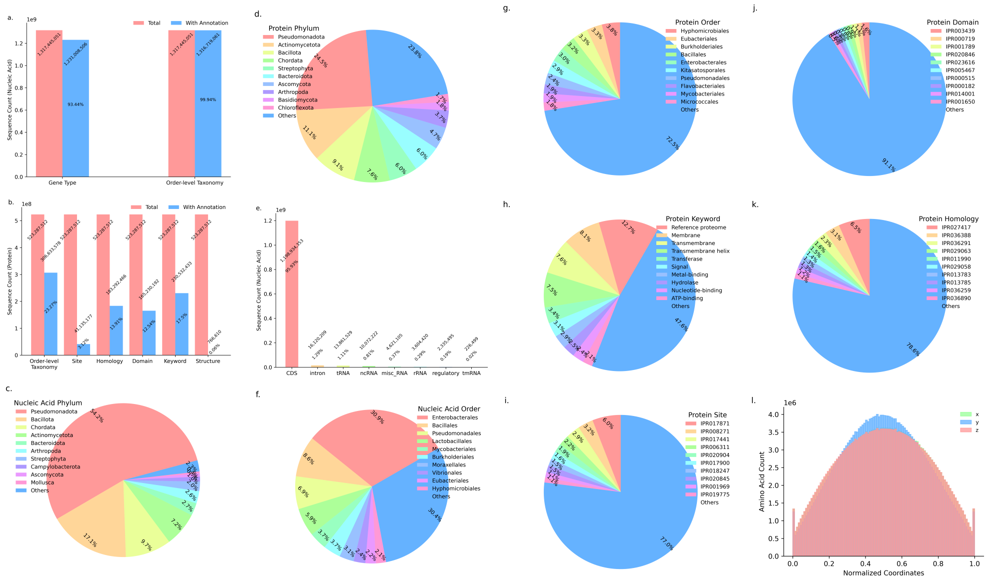


Fig. S2: **Annotation statistics on pre-training data of LucaOne. a.** The proportion of genome region types and order-level taxonomy in nucleic acid. Most sequences have both types of annotation information. **b.** The proportion of the count of sequences with each of the selected six annotations, including order-level taxonomy, keyword, site, domain, homology, and tertiary structure, of which the proportion of sequence count with tertiary structure is tiny. **c. and d.** The proportion of sequence counts in the top 10 phylum-level taxonomy of nucleic acids and proteins, respectively. **e.** The distribution of eight selected genome region types in nucleic acids, of which the CDS region is the most. **f. and g.** The proportion of sequence counts in the top 10 order-level taxonomy(total 2,196 categories) of nucleic acids and proteins, respectively. **h.**∼**k.** The proportion of protein sequence counts in the top 10 keywords(total 1,179 categories), the top 10 site types(total 946 categories), the top 10 domain types(total 13,717 categories), and the top 10 homology types(total 3,442 categories), respectively. **l.** The *coord-(x, y, z)* distribution of C_α*-atom* position(local normalization within a protein chain). It is very similar to the normal distribution. The distribution has a long tail in **c.**∼**f.**. The distribution is ladder decreasing in **g.**∼**k.**.

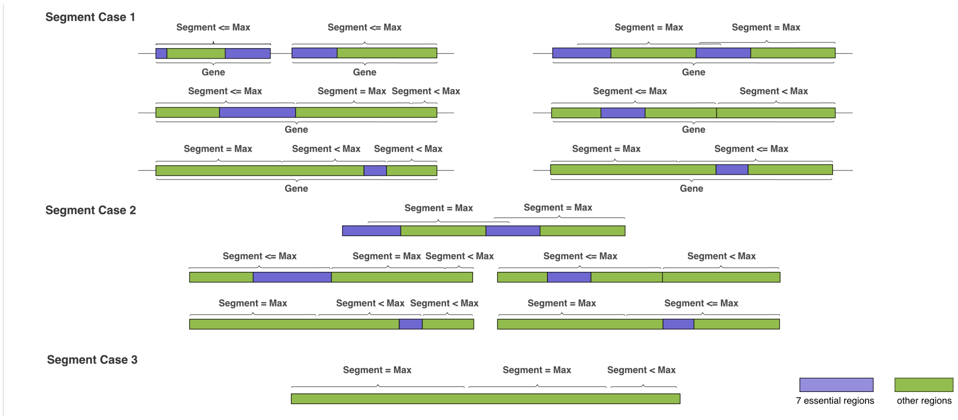


Fig. S3: **Three cases of segmentation.** 7 essential regions include: ’CDS’, ’regulatory’, ’tRNA’, ’ncRNA’, ’rRNA’, ’miscRNA’, and ’tmRNA’. Other regions include the genome region types not mentioned above, such as ’intron’.

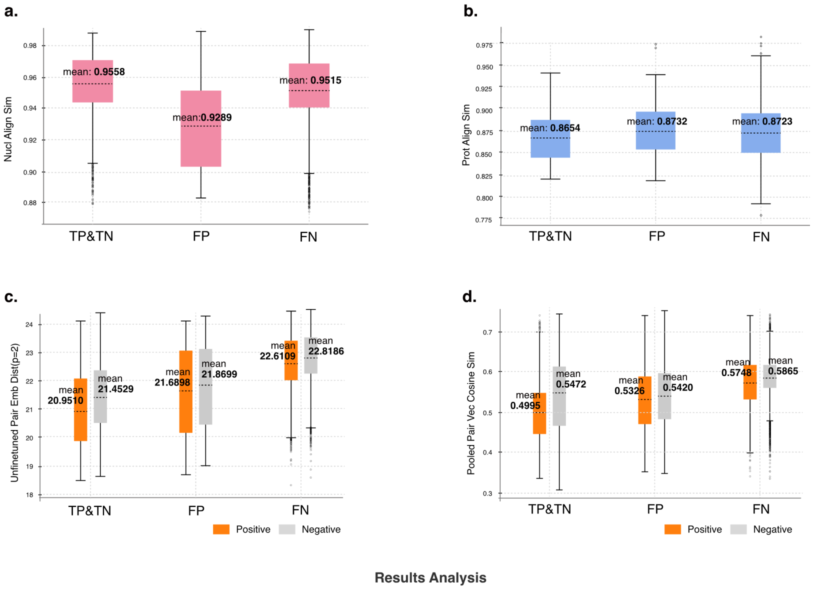


Fig. S4: **Results Analysis of Central Dogma of Molecular Biology Task. a.**∼**d.** We analyze the misidentified samples from two perspectives: **Sequence** and **Embedding**. **a. and b.** Analyze these misclassified samples using the sequence alignment similarity. **c. and d.** Analyze these misclassification samples by the distance and similarity of the embedding vector. From these two perspectives, it can be concluded that the misclassified samples themselves are naturally challenging to classify. Therefore, more data should be adopted to allow the model to learn these more subtle differences.

References

[1] Crick, F., Barnett, L., Brenner, S., Watts-Tobin, R.J., *et al.*: General nature of the genetic code for proteins. Nature **192**(4809), 1227–1232 (1961)

[2] Searls, D.B.: The language of genes. Nature **420**(6912), 211–217 (2002)

[3] Darwin, C. (ed.): The Descent of Man: and Selection in Relation to Sex. John Murray, Albemarle Street., ??? (1888)

[4] Gimona, M.: Protein linguistics—a grammar for modular protein assembly? Nature Reviews Molecular Cell Biology **7**(1), 68–73 (2006)

[5] Barbieri, M. (ed.): The Organic Codes: an Introduction to Semantic Biology. Cambridge University Press, ??? (2003)

[6] Pinker, S. (ed.): The Language Instinct: How the Mind Creates Language. Penguin uK, ??? (2003)

[7] Simon, E., Swanson, K., Zou, J.: Language models for biological research: a primer. Nature Methods **21**(8), 1422–1429 (2024)

[8] Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., *et al.*: Prottrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence **44**(10), 7112–7127 (2021)

[9] Brandes, N., Ofer, D., Peleg, Y., Rappoport, N., Linial, M.: Proteinbert: a universal deep-learning model of protein sequence and function. Bioinformatics **38**(8), 2102–2110 (2022)

[10] Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., *et al.*: Evolutionary-scale prediction of atomic-level protein structure with a language model. Science **379**(6637), 1123–1130 (2023)

[11] Elnaggar, A., Essam, H., Salah-Eldin, W., Moustafa, W., Elkerdawy, M., Rochereau, C., Rost, B.: Ankh: Optimized protein language model unlocks general-purpose modelling. arXiv preprint arXiv:2301.06568 (2023)

[12] Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., Kinch, L.N., Schaeffer, R.D., *et al.*: Accurate prediction of protein structures and interactions using a three-track neural network. Science **373**(6557), 871–876 (2021)

[13] Hou, X., He, Y., Fang, P., Mei, S.-Q., Xu, Z., Wu, W.-C., Tian, J.-H., Zhang, S., Zeng, Z.-Y., Gou, Q.-Y., et al.: Using artificial intelligence to document the hidden rna virosphere. bioRxiv, 2023–04 (2023)

[14] Yu, T., Cui, H., Li, J.C., Luo, Y., Jiang, G., Zhao, H.: Enzyme function prediction using contrastive learning. Science **379**(6639), 1358–1363 (2023)

[15] Zhou, Z., Ji, Y., Li, W., Dutta, P., Davuluri, R., Liu, H.: Dnabert-2: Efficient foundation model and benchmark for multi-species genome. arXiv preprint arXiv:2306.15006 (2023)

[16] Nguyen, E., Poli, M., Faizi, M., Thomas, A., Wornow, M., Birch-Sykes, C., Massaroli, S., Patel, A., Rabideau, C., Bengio, Y., et al.: Hyenadna: Long-range genomic sequence modeling at single nucleotide resolution. Advances in neural information processing systems **36** (2024)

[17] Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., Yao, J.: scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. Nature Machine Intelligence **4**(10), 852–866 (2022)

[18] Nguyen, E., Poli, M., Durrant, M.G., Thomas, A.W., Kang, B., Sullivan, J., Ng, M.Y., Lewis, A., Patel, A., Lou, A., et al.: Sequence modeling and design from molecular to genome scale with evo. bioRxiv, 2024–02 (2024)

[19] Li, Q., Hu, Z., Wang, Y., Li, L., Fan, Y., King, I., Song, L., Li, Y.: Progress and opportunities of foundation models in bioinformatics. arXiv preprint arXiv:2402.04286 (2024)

[20] Crick, F.: Central dogma of molecular biology. Nature **227**(5258), 561–563 (1970)

[21] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)

[22] Koonin, E.V.: Why the central dogma: on the nature of the great biological exclusion principle. Biology direct **10**, 1–5 (2015)

[23] Yockey, H.P.: Information theory, evolution and the origin of life. Information Sciences **141**(3-4), 219–225 (2002)

[24] He, Y., Wang, C., Zhang, S., Li, N., Li, Z., Zeng, Z.: Kg-mtt-bert: knowledge graph enhanced bert for multi-type medical text classification. arXiv preprint arXiv:2210.03970 (2022)

[25] Delsuc, F., Brinkmann, H., Chourrout, D., Philippe, H.: Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature **439**(7079), 965–968 (2006)

[26] Moreno, J., Nielsen, H., Winther, O., Teufel, F.: Predicting the subcellular location of prokaryotic proteins with deeplocpro. bioRxiv, 2024–01 (2024)

[27] Han, Y., Zhang, S.-W.: ncrpi-lgat: Prediction of ncrna-protein interactions with line

graph attention network framework. Computational and Structural Biotechnology Journal **21**, 2286–2295 (2023)

[28] Robbins, J.W.: Darwin's Dangerous Idea: Evolution and the Meanings of Life. JSTOR (1996)

[29] Chomsky, N.: Three factors in language design. Linguistic inquiry **36**(1), 1–22 (2005)

[30] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)

[31] Liu, J., Yang, M., Yu, Y., Xu, H., Li, K., Zhou, X.: Large language models in bioinformatics: applications and perspectives. arXiv preprint arXiv:2401.04155 (2024)

[32] Sapoval, N., Aghazadeh, A., Nute, M.G., Antunes, D.A., Balaji, A., Baraniuk, R., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., *et al.*: Current progress and open challenges for applying deep learning across the biosciences. Nature Communications **13**(1), 1728 (2022)

[33] Vig, J., Madani, A., Varshney, L.R., Xiong, C., Socher, R., Rajani, N.F.: Bertology meets biology: interpreting attention in protein language models. arXiv preprint arXiv:2006.15222 (2020)

[34] Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J.R., Grabska-Barwinska, A., Taylor, K.R., Assael, Y., Jumper, J., Kohli, P., Kelley, D.R.: Effective gene expression prediction from sequence by integrating long-range interactions. Nature methods **18**(10), 1196–1203 (2021)

[35] Nakano, F.K., Lietaert, M., Vens, C.: Machine learning for discovering missing or wrong protein function annotations: a comparison using updated benchmark datasets. BMC bioinformatics **20**, 1–32 (2019)

[36] Alharbi, W.S., Rashid, M.: A review of deep learning applications in human genomics using next-generation sequencing data. Human Genomics **16**(1), 26 (2022)

[37] Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)

[38] Whalen, S., Schreiber, J., Noble, W.S., Pollard, K.S.: Navigating the pitfalls of applying machine learning in genomics. Nature Reviews Genetics **23**(3), 169–181 (2022)

[39] Banzhaf, W., Machado, P., Zhang, M. (eds.): Handbook of Evolutionary Machine Learning. Springer, ??? (2023)

[40] Blanchard, A.E., Shekar, M.C., Gao, S., Gounley, J., Lyngaas, I., Glaser, J., Bhowmik, D.: Automating genetic algorithm mutations for molecules using a masked language

model. IEEE Transactions on Evolutionary Computation **26**(4), 793–799 (2022)

[41] Ebrahim, A., Brunk, E., Tan, J., O'brien, E.J., Kim, D., Szubin, R., Lerman, J.A., Lechner, A., Sastry, A., Bordbar, A., *et al.*: Multi-omic data integration enables discovery of hidden biological regularities. Nature communications **7**(1), 13091 (2016)

[42] Vahabi, N., Michailidis, G.: Unsupervised multi-omics data integration methods: a comprehensive review. Frontiers in genetics **13**, 854752 (2022)

[43] Han, H., Liu, X.: The challenges of explainable ai in biomedical data science. BMC bioinformatics **22**(Suppl 12), 443 (2022)

[44] Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**(4), 1312 (2019)

[45] Xiong, R., Yang, Y., He, D., Zheng, K., Zheng, S., Xing, C., Zhang, H., Lan, Y., Wang, L., Liu, T.: On layer normalization in the transformer architecture. In: International Conference on Machine Learning, pp. 10524–10533 (2020). PMLR

[46] Su, J., Ahmed, M., Lu, Y., Pan, S., Bo, W., Liu, Y.: Roformer: Enhanced transformer with rotary position embedding. Neurocomputing **568**, 127063 (2024)

[47] Kenton, J.D.M.-W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT, pp. 4171–4186 (2019)

[48] Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)

[49] Meyer, F., Fritz, A., Deng, Z.-L., Koslicki, D., Lesker, T.R., Gurevich, A., Robertson, G., Alser, M., Antipov, D., Beghini, F., *et al.*: Critical assessment of metagenome interpretation: the second round of challenges. Nature methods **19**(4), 429–440 (2022)

[50] Xu, Q., Hu, D.H., Xue, H., Yu, W., Yang, Q.: Semi-supervised protein subcellular localization. BMC bioinformatics **10**, 1–10 (2009)

[51] Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., Song, Y.: Evaluating protein transfer learning with tape. Advances in neural information processing systems **32** (2019)

[52] Noviello, T.M.R., Ceccarelli, F., Ceccarelli, M., Cerulo, L.: Deep learning predicts short non-coding rna functions from only raw sequence data. PLoS computational biology **16**(11), 1008415 (2020)

[53] Rossi, E., Monti, F., Bronstein, M., Liò, P.: ncrna classification with graph convolutional networks. arXiv preprint arXiv:1905.06515 (2019)

[54] Du, X., Dong, L., Lan, Y., Peng, Y., Wu, A., Zhang, Y., Huang, W., Wang, D., Wang, M., Guo, Y., *et al.*: Mapping of h3n2 influenza antigenic evolution in china reveals a strategy for vaccine strain recommendation. Nature communications **3**(1), 709 (2012)

[55] Sun, T., Zhou, B., Lai, L., Pei, J.: Sequence-based prediction of protein protein interaction using a deep-learning algorithm. BMC bioinformatics **18**, 1–8 (2017)

[56] Mock, F., Kretschmer, F., Kriese, A., Böcker, S., Marz, M.: Bertax: taxonomic classification of dna sequences with deep neural networks. BioRxiv, 2021–07 (2021)