Contrastive omics pre-training for multimodal embeddings of genomic sequences

Lyam Baudry *University of Lausanne* lyam.baudry@unil.ch
Cyril Matthey-Doret *EPFL, Lausanne* cyril.matthey-doret@epfl.ch
Amelie Fritz *DTU, Copenhagen* ameli@dtu.dk
Celia Raimondi *Sanger Institute, Cambridge* celia.raimondi@avrafa.com

Abstract

The exponential growth of genomic data presents a significant challenge in bridging the gap between raw DNA sequences and their complex biological functions. We introduce CLOP (Contrastive Learning for Omics Pre-training), a model designed to learn joint representations of DNA sequences and their textual annotations. Inspired by the success of CLIP in vision-language tasks, CLOP employs a dual-encoder architecture trained with a contrastive objective to map DNA k-mers and descriptive text into a shared, multimodal embedding space. In this paper, we present a proof of concept based on a preliminary training run on genomic sequences from various species. Despite the limited training (5 epochs) and small sample size, our results demonstrate that the learned embeddings qualitatively capture meaningful biological information. The embedding space shows clear clustering of sequences by functional category. These early results highlight the potential of contrastive learning to build foundational models for genomics.

Keywords: genomics, deep learning, contrastive learning, multi-modal embeddings, bioinformatics

1 Introduction

Modern sequencing technologies have produced an immense volume of genomic data, yet interpreting the functional role of these sequences remains a central challenge in biology. A key goal of computational biology is to develop models that can automatically infer function from sequence, effectively learning the "language" of DNA. Models like CLIP [Radford et al., 2021] have shown remarkable success in learning rich, joint representations of images and text without direct supervision, enabling powerful zero-shot capabilities. We hypothesize that a similar approach can be applied to genomics, treating DNA sequences and their corresponding biological annotations (e.g., gene biotype, species, functional descriptions) as two distinct modalities.

In this work, we introduce CLOP (Contrastive Learning for Omics Pre-training), a CLIP-inspired framework for learning a shared embedding space between DNA sequences and plain-text genomic annotations. Our goal is to create a model that learns semantically

meaningful representations of DNA, where sequences with similar biological roles are located close to each other in the embedding space. Such a model could serve as a foundation for various downstream tasks, including automated gene annotation, function prediction, and cross-modal search (e.g., finding DNA sequences that match a textual description). This paper presents an early-stage proof of concept, demonstrating the feasibility and potential of this approach.

2 The CLOP Model

2.1 Model Architecture

CLOP utilizes a dual-encoder architecture, a standard design for multimodal contrastive learning. The code is available at `https://github.com/baudrly/clop`. A demo using pre-set embeddings can be viewed at `https://baudrly.github.io/clop`.

- **DNA Encoder:** This module processes raw DNA sequences. The sequence is first tokenized into overlapping 6-mers (k=6). These k-mer tokens are fed into an embedding layer followed by a bidirectional long short-term memory (LSTM) network for simplicity. The final hidden states are pooled to produce a single fixed-size vector representation for the entire DNA sequence.
- **Text Encoder:** This module processes the textual annotations associated with each DNA sequence. The text is tokenized into words, which are passed through a separate embedding layer and a bidirectional LSTM to produce a vector representation of the description.

Both encoders project their respective inputs into a shared-dimensional embedding space. The final embedding vectors from both modalities are L2-normalized.

2.2 Contrastive training

The model is trained using a contrastive loss function. Given a batch of N (DNA sequence, text annotation) pairs, the encoders produce N DNA embeddings and N text embeddings. We then compute an $N \times N$ matrix of cosine similarities between all possible DNA-text embedding pairs.

The training objective is to maximize the similarity of the N correct (DNA, text) pairs on the diagonal of this matrix while simultaneously minimizing the similarity of the $N^2 - N$ incorrect pairs. This is implemented as a symmetric cross-entropy loss over the similarity scores, guided by a learnable temperature parameter that scales the logits before the loss calculation. This process encourages the model to pull representations of corresponding DNA sequences and their descriptions together in the embedding space while pushing all non-matching pairs apart.

2.3 Experimental Setup

For this initial proof-of-concept study, we used a dataset of 78,319

DNA sequences extracted from a FASTA file. The annotations were derived directly from FASTA headers, which contained information on the species (e.g., *Homo sapiens,* Mus musculus) and biological biotype (e.g., 'protein coding', 'lncRNA', 'snRNA'). The dataset was split into a 90% training set and a 10% validation set.

The model was configured with an embedding dimension of 32 and a hidden dimension of 256 for the LSTM encoders. We trained the model for only 5 epochs using the AdamW optimizer, which is sufficient to demonstrate initial learning but far from convergence.

3 Preliminary results

Given the early stage of this research, we focus on qualitative results and simple downstream tasks that probe the semantic structure of the learned embedding space, rather than state-of-the-art performance on established benchmarks.

3.1 Embedding space visualization

The most direct way to assess the model's learning is to visualize the DNA embeddings. Figure 1 shows a 2D projection of the DNA embeddings from our validation set, generated using the t-SNE algorithm and colored by biotype.

Even after only five epochs, a clear structure has begun to emerge. The model has learned to separate different functional categories into distinct regions of the embedding space. The two largest classes, 'protein coding' and 'lncRNA', form several overlapping clouds (with some separation). More encouragingly, smaller and functionally specific classes such as 'rRNA' (ribosomal RNA) and 'snRNA' (small nuclear RNA) form tight, well-defined clusters. This qualitative result is a strong indicator that the model is not merely memorizing data but is learning fundamental patterns within the DNA sequences that correspond to their biological function.

3.2 Zero-Shot classification via k-NN

To quantify the quality of the learned representations, we performed a k-Nearest Neighbors (k-NN) classification task on the DNA embeddings. We used the embeddings from our validation set to classify the sequence's biotype without any further training or fine-tuning. This 'zero-shot' evaluation tests whether the embedding space is structured in a way that is immediately useful for downstream tasks.

Using this simple k-NN approach, we achieved an overall accuracy of **61.5%** on biotype classification. While not state-of-the-art, this result is significantly better than random chance across 14 classes and demonstrates that the embeddings contain a strong, separable signal corresponding to biological function. The macro F1-score was lower (0.42), indicating, as expected, that the model performs better on well-represented classes.

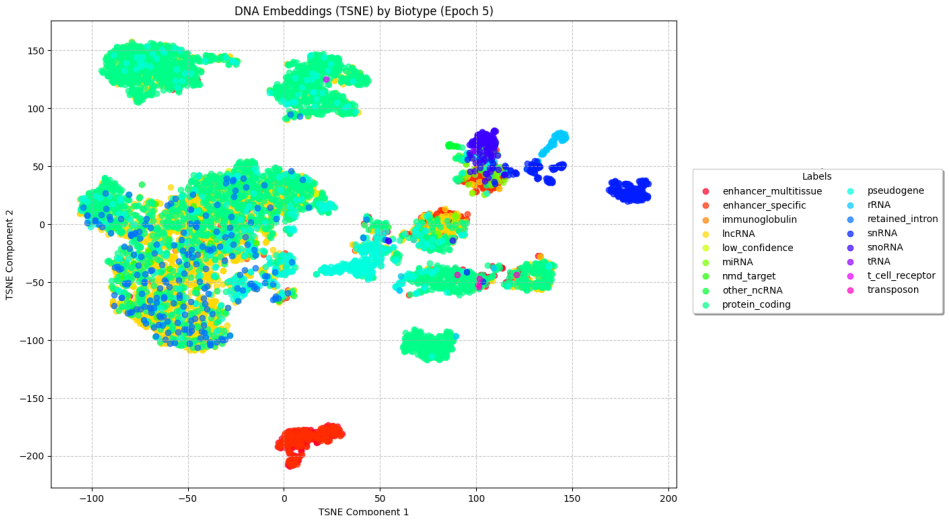

Figure 1: t-SNE visualization of DNA sequence embeddings from the validation set, colored by biotype. The model begins to group sequences by functional type, with large categories like ‘protein coding‘ and ‘lncRNA‘ forming distinct masses, and smaller, specific classes like ‘rRNA‘ and ‘snRNA‘ forming tighter, more localized clusters. This demonstrates that the training successfully imparts biologically relevant structure onto the embedding space.

3.3 Limitations

This preliminary work has notable limitations. The retrieval metrics (e.g., Mean Reciprocal Rank and Recall), which measure the ability to find the correct text description for a given DNA sequence, were very low. This suggests that while the embedding space has a coherent global structure (as seen in the t-SNE plot), the fine-grained alignment required for precise one-to-one retrieval has not yet been learned. Similarly, clustering metrics like the Silhouette score were poor, indicating that the clusters are not yet compact and well-separated. We attribute these limitations directly to the short training duration and small model size.

4 Conclusions

Our preliminary proof-of-concept experiment shows that even with minimal training, this approach can produce DNA sequence embeddings that are surprisingly rich in biological meaning. The model successfully learns to cluster sequences by their functional biotype in an unsupervised manner, and these representations can be used for zero-shot classification with promising accuracy.

Future efforts will focus on scaling uptraining for more epochs on larger and more diverse datasets with richer textual descriptions. We will also explore more powerful encoder architectures, such as StripedHyena Nguyen et al. [2024], to bet-

ter capture long-range dependencies in DNA. Our results are in line with the broader goal of applying multimodal AI to accelerate scientific discovery.

References

E. Nguyen, M. Poli, M. G. Durrant, A. W. Thomas, B. Kang, J. Sullivan, M. Y. Ng, A. Lewis, A. Patel, A. Lou, S. Ermon, S. A. Baccus, T. Hernandez-Boussard, C. Ré, P. D. Hsu, and B. L. Hie. Sequence modeling and design from molecular to genome scale with evo. *bioRxiv*, 2024. doi: 10.1101/2024.02.27.582234.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.