**Academic Project**

**Course: ECE648-P – Machine Learning**

**Spring 2025**

Attention-Based LSTM for DNA Sequence Modeling Using k-mers

Fabrisio Braulio Ponte Vela

**University of Miami**  
**College of Engineering**  
**Department of Electrical and Computer Engineering**

**1. Introduction**

DNA sequence modeling is a fundamental task in bioinformatics with implications for gene prediction, disease diagnosis, and evolutionary biology. The sequential nature of DNA, composed of four nucleotides (A, T, G, C), shares structural similarities with natural language, making it amenable to deep learning methods initially developed for text processing. Traditional models often rely on fixed-window approaches or Markov chains, which can fail to capture long-range dependencies and contextual relationships that are crucial in genomic sequences.

Recent advancements in deep learning, particularly Recurrent Neural Networks (RNNs) and their variants such as Long Short-Term Memory (LSTM) networks, have shown promise in modeling sequential biological data. However, LSTMs alone can struggle with capturing the varying importance of each element in a sequence. To address this limitation, attention mechanisms have been introduced to help models focus on the most relevant parts of the sequence during prediction, improving both accuracy and interpretability.

In this study, we present a hybrid model that combines LSTM networks with an attention mechanism to predict the next k-mer in a DNA sequence. A k-mer is a subsequence of length *k*, and its prediction is crucial for tasks like motif discovery and genomic annotation. We trained our model using the *Escherichia coli* genome (strain GCF\_003018035.1), a well-studied bacterial organism with a genome size suitable for deep learning experimentation.

We evaluate the model’s performance based on prediction accuracy and analyze the learned attention weights to interpret which parts of the sequence the model deems most informative. This work aims to contribute to the growing body of research applying machine learning techniques to biological data, providing a reproducible and interpretable framework for genomic sequence modeling.

**2. Design and Implementation**

**2.1 Data Preprocessing**

* **Input**: Raw DNA sequence from the E. coli genome in FASTA format.
* **Cleaning**: FASTA headers and newline characters are stripped.
* **K-mer Tokenization**:
* DNA sequences are tokenized into overlapping k-mers (k=4). For instance, "ATCGAGGATC" is transformed into:

["ATCG", "TCGA", "CGAG", "GAGG", "AGGA", "GGAT", "GATC"]

* Each unique k-mer is assigned a numerical index, forming a vocabulary.
* **Encoding**:
* The full genome sequence is converted into a sequence of indices based on the k-mer vocabulary.
* The data is split 80/20 into training and testing sets.

**2.2 Model Architecture**

**a) Encoder**

* **Embedding Layer**: Maps k-mer indices into dense vectors.
* **LSTM Layer**: Processes embedded sequences to capture temporal dependencies.

**b) Attention Mechanism**

* **Inputs**: Final hidden state and all encoder outputs.
* **Process**: Computes alignment scores, applies softmax to get attention weights, and generates a weighted context vector.
* **Purpose**: Enables the model to focus on the most relevant parts of the sequence for prediction.

**c) Decoder**

* **LSTM**: Takes current input and attention-generated context to produce decoded states.
* **Fully Connected Layer**: Maps LSTM outputs to vocabulary logits for final prediction.

**2.3 Training Process**

* **Loss Function**: Cross-entropy loss for multi-class classification of k-mers.
* **Optimizer**: Adam optimizer with a learning rate of 0.001.
* **Batch Size**: 64
* **Sequence Length**: 50 k-mers
* **Epochs**: 1000, with early stopping based on validation loss.

**3. Experiments**

To evaluate the effectiveness of our attention-based LSTM model in predicting DNA sequences, we conducted experiments using the complete genome of E. coli. The genome was tokenized into overlapping 4-mers to form a sequence suitable for language modeling. The dataset was split into training and testing subsets to ensure reliable evaluation. Model training was carried out using PyTorch, leveraging GPU acceleration where possible. We carefully selected model parameters to balance complexity and performance, with an emphasis on capturing both short- and long-range dependencies within the DNA. Below are the detailed experimental settings:

**3.1 Dataset**

* **Genome**: *E. coli* strain (GCF\_003018035.1)
* **Split**: 80% training, 20% testing
* **Tools**: Preprocessing and data loading handled via Python and PyTorch

**3.2 Model Parameters**

* Embedding Dimension: 256
* LSTM Hidden Size: 512
* K-mer Size: 4
* Learning Rate: 0.001

**3.3 Training Setup**

* **Framework**: PyTorch
* **Device**: GPU acceleration (if available)
* **Monitoring**: Training loss printed every 100 epochs

**4. Results**

The attention-based LSTM model effectively learned patterns in the E. coli DNA sequence. Throughout training, the loss steadily decreased, indicating successful convergence. The model demonstrated high accuracy in predicting the next 4-mer in a sequence, and evaluation on a held-out test set confirmed that it generalized well to unseen sequences. The attention mechanism helped focus on specific input positions, leading to better performance than a plain LSTM. Visualization of attention weights revealed that the model prioritized biologically meaningful motifs, such as promoter-like regions and repetitive nucleotide patterns, which are critical in genomic sequence modeling.

**4.1 Training Performance**

* **Observation**: Training loss showed steady decrease with minimal overfitting due to early stopping.
* **Impact of Attention**: Enhanced the model’s ability to capture long-range nucleotide dependencies.

**4.2 Testing Performance**

* **Generalization**: The model successfully predicted unseen k-mers on the test set.
* **Attention Visualization**: Attention weights highlighted key sequence regions influencing predictions.

**4.3 Example Predictions**

* **Initial Input**: First 20 nucleotides of the genome
* **Predicted Output**:
  + 5 k-mers following the input sequence
* **Biological Insight**: The predicted k-mers often corresponded to biologically plausible subsequences.

**5. Discussion**

The use of attention in conjunction with LSTM proved to be beneficial for sequence modeling. It allowed the model to weigh more important parts of the input, enhancing interpretability and performance. Compared to standard LSTM models, the attention-based approach required more memory but yielded more biologically plausible predictions. One of the key insights was that attention maps often highlighted regions known to regulate gene expression, suggesting the model captured functional DNA structure without explicit labels. Despite this, the model struggled with very long dependencies and fixed-length k-mer encoding, which limits adaptability to varying motif sizes. Nonetheless, this architecture provides a solid foundation for more advanced genomics models.

**5.1 Key Findings**

* The attention mechanism substantially improved predictive performance.
* K-mer tokenization effectively modeled local sequence motifs.
* The model learned biologically relevant patterns without explicit feature engineering.

**5.2 Advantages**

* Scalable to long sequences
* Interpretable via attention weights
* Adaptable to other genomics tasks (e.g., motif detection, mutation prediction)

**5.3 Limitations**

* Memory consumption grows with vocabulary size and sequence length
* Long training time for large genomes
* Limited by fixed k-mer size in this implementation

**5.4 Future Work**

* Experiment with variable or larger k-mer sizes
* Incorporate bidirectional LSTM for full context
* Add positional encoding (as in Transformers)
* Replace LSTM with a Transformer model for better scaling
* Apply to other organisms and tasks like promoter prediction or CRISPR off-target effects

**6. Reproducibility**

Reproducing this study is straightforward with the provided code and standard bioinformatics resources. The process involves downloading a public genome sequence, preparing it into a machine learning–ready format, and training a deep learning model using PyTorch. This modular pipeline allows easy experimentation with different sequence lengths, k-mer sizes, and model architectures. The model can be evaluated for both predictive accuracy and biological interpretability using attention visualizations. Below are the key steps:

**Step 1: Data Preparation**

* Download the *E. coli* genome (FASTA format).
* Run the script to clean the sequence, generate 4-mers, encode them as numerical indices, and create training/testing splits.

**Step 2: Model Setup**

* Install necessary Python libraries (e.g., PyTorch, NumPy).
* Load the AttentionLSTM model which combines embeddings, LSTM, and attention for sequence prediction.

**Step 3: Training**

* Train the model for up to 1000 epochs using the specified hyperparameters.
* Use early stopping to prevent overfitting and save the best-performing model.

**Step 4: Evaluation**

* Use a sample DNA input to generate predicted k-mers.
* Examine attention weights to gain insights into the learned sequence dependencies.

**7. Conclusions**

The results of this study demonstrate the effectiveness of integrating attention mechanisms into LSTM models for DNA sequence prediction. By treating DNA as a series of overlapping k-mers, the model captures both local and global dependencies that are often missed by simpler sequence models. The attention layer plays a crucial role in identifying which parts of the input sequence are most relevant for predicting the next k-mer, resulting in improved context awareness and interpretability.

This architecture outperforms traditional LSTM models in handling long-range dependencies, which is especially valuable in biological sequences where distant nucleotide relationships often encode functional or structural significance. The model’s ability to generalize well to unseen data suggests that it has successfully learned meaningful patterns in the DNA structure rather than memorizing sequences. Moreover, attention weight visualizations provide interpretability, helping to identify regions within the DNA that contribute most to specific predictions—offering potential insights into genomic regions of interest for further biological investigation.

The choice of using 4-mers provided a balance between vocabulary size and pattern resolution, although this parameter can be adjusted to optimize for different types of genomic data. The model architecture is flexible and can be extended to more complex tasks, such as identifying motifs, enhancer regions, or predicting mutations. Despite its strengths, the model is computationally intensive, especially when dealing with large genomes or longer sequences, due to the attention mechanism and recurrent layers. This can be mitigated in future iterations by experimenting with bidirectional LSTMs or transformer-based architectures.

In conclusion, this attention-based LSTM approach provides a powerful and interpretable framework for DNA sequence modeling. It demonstrates strong performance on a real-world genomic dataset, confirms the biological plausibility of its predictions, and opens the door for advanced applications in genomics, such as regulatory sequence identification or variant effect prediction. With further optimization and domain-specific tuning, this approach holds promise for broader adoption in computational biology and genomics research.