COMPARING BIOLLM AND GFM PERFORMANCES FOR CLASSIFICATION OF RNA-SEQ AS TUMOR OR NON TUMOR

Raphael HIERSO- - IGLESIAS

BACKGROUND

RNA sequencing allows for transcriptomic comparison between tumor and non-tumor samples. Nevertheless, the classification process requires a lot of time and resources. A strategy in bio-informatics is to represent sequencing data as de Bruijn Graphs, where k-mers[1] form nodes and overlaps define edges. This method allows for the compaction of the data. It enables scalable assembly and downstream analysis through Eulerian walks (reconstruction of the transcript or further classification). At the same time, these graphs are challenging to store and traverse efficiently [1]. Tools such as Logan [2] exploit the properties of compacted De Bruijn graphs to encode datasets while preserving interesting information for later investigation. This makes such structures a privileged starting point for graph-based learning on RNA data.

In parallel, biological foundation models[2] have emerged as effective sequence encoders for genomics. Recent DNA foundation models, such as Caduceus [5], build on architectures (such as Mamba) that can capture long-range dependencies between bases. These models are designed to treat a sequence and its reverse-complement in an equivalent way, a property named reverse-complement equivariance. These models handle DNA and RNA as token sequences and learn contextual embeddings that capture local and long-range dependencies. These elements provide a base for BioLLM to classify patient sequencing data.

More recently, Graph Foundation Models (GFMs) aim to extend the Foundation Model paradigm from textual resources and images to more general graph-structured data [6]. GFMs are pre-trained on datasets of heterogeneous graphs and are adapted to graph related tasks (node, edge, graph classification), with the goal of achieving transfer across different fields. The work "Equivariance Everywhere All At Once" [3] suggests that such models must respect multiple symmetries of graph data : node permutation equivariance, label permutation equivariance and feature permutation invariance. Furthermore, it provides a recipe for building GFMs that would be universal approximators under these conditions, demonstrating strong zero-shot prediction performance on unseen graphs. This is very interesting for bio-informatics, where graphs often come from different experiments.

Recent progress related to neural positional encoding shows that it could allow a model to approximate faster solutions to the Traveling Salesman Problem,with up to ten thousand cities [4]. This highlights the influence of positional encoding on a model's ability to reason over graph structures.

Together, these developments interrogate the RNA-based cancer classification. How do sequence-only bioLLMs compare to GFMs that operate on De Bruijn graphs, and can neural positional encoding enhance graph-level representation for distinguishing tumor from non-tumor patients ?

AIM

The primary aim of this study is to compare two modeling strategies for RNA-based cancer prediction. Firstly, a sequence-based biological language model is applied directly to RNA sequences from NCBI data, followed by a Graph Foundation Model operating on RNA-derived graphs (such as de Bruijn graphs).

We want to assess, under similar experimental conditions, whether there is a gain from using an explicit graph-level reasoning model for classifying patients.

As a secondary objective, we want to investigate the role of neural positional encoding within the GFM by studying how different encoding choices later influence the graph representations, classification performance, and the interpretability of the resulting subgraphs.

METHODS

Dataset

We use a dataset of RNA sequences from the NCBI online database. Each sample corresponds to one patient and is labeled as tumor or non-tumor based on the original clinical metadata. From the Logan index, we extract patient-specific RNA-sequences (as unitigs) together. The samples are randomly assigned to the training and validation sets. The experiments use the same splits.

[1] Sequence of a length of k genetic bases (A, T, C or G for DNA)

[2] Foundation models are large machine learning models pre-trained on massive amounts of specific data such as texts, pictures, graphs; these models are then reused and adapted to perform tasks for which they were not specifically trained

The training dataset consists of tissue samples from lung cancer. The data size is 139 G bases, corresponding to approximately 53 GB of sequencing data.

BioLLM

We use a pre-trained RNA foundation model (such as Caduceus [5]) as a bioLLM. Each sequence is tokenized as nucleotides, chunked if necessary , and passed through the frozen bioLLM to obtain contextual embeddings. We aggregate the embeddings and train a lightweight classifier to predict the label.

GFM

For the graph-based approach, we use the compacted de Bruijn Graph representation from Logan [2]. Nodes correspond to compacted unitigs, and edges encode adjacency in the de Bruijn graph . Node features are initialized from statistics (relative abundance).

We then adapt a pretrained Graph Foundation Model for graph-level classification (Anygraph, for example, is used in Finkelshtein work[3]).

To study the impact of structural information, we propose using a GFM with standard positional encoding (Laplacian) followed by a GFM modified with neural positional encodings inspired by AliBi [4].

Metrics

To exploit the results, we will use the accuracy (Accuracy $= \frac{\text{Correct Classifications}}{\text{Total Classifications}}$) which gives the proportion of correct classifications. Because, from a medical standpoint, the most costly case would be a False Negative (tumor not detected), we will use the recall, which represents the proportion of true positive results correctly classified (Recall $= \frac{\text{True Positive}}{\text{True Positive + False Negative}}$). To analyze the precision, we use the F1-score, which is the harmonic mean between precision and recall (F1 $= \frac{2\text{TP}}{2\text{TP+FP+FN}}$).

EXPECTED RESULTS

The implementation has not been realized yet, but we expect both modeling strategies to correctly predict the labels between tumor and non-tumor patients in other clinical experiments. For the graph-based approach, we anticipate that the GFM operation on the de Bruijn graph could match or surpass the bioLLM, thanks to its ability to exploit graph structures such as recurrent motifs and relationships between nodes. Regarding positional encoding, we expect that learned encodings will produce better representations and improve classification performance. Finally, we expect the GFM to offer more fine-grained interpretability than the bioLLM by highlighting and adding context to the classification results. If these expectations are confirmed, the study will show when it is preferable to use a simple bioLLm-based pipeline versus when it is worth constructing RNA-derived graphs and using a GFM.

References

[1] R. Chikhi, A. Limasset, S. Jackman, J. Simpson, and P. Medvedev. On the representation of de Bruijn graphs, Oct. 2014. arXiv:1401.5383 [q-bio].

[2] R. Chikhi, B. Raffestin, A. Korobeynikov, R. Edgar, and A. Babaian. Logan: Planetary-Scale Genome Assembly Surveys Life's Diversity, July 2024. Pages: 2024.07.30.605881 Section: New Results.

[3] B. Finkelshtein, Ceylan, M. Bronstein, and R. Levie. Equivariance Everywhere All At Once: A Recipe for Graph Foundation Models, Aug. 2025. arXiv:2506.14291 [cs].

[4] P. Pereira, F. Giroire, and E. Natale. Solving the Traveling Salesman Problem with Positional Encoding. Oct. 2025.

[5] Y. Schiff, C.-H. Kao, A. Gokaslan, T. Dao, A. Gu, and V. Kuleshov. Caduceus: Bi-Directional Equivariant Long-Range DNA Sequence Modeling, June 2024. arXiv:2403.03234 [q-bio].

[6] Z. Wang, Z. Liu, T. Ma, J. Li, Z. Zhang, X. Fu, Y. Li, Z. Yuan, W. Song, Y. Ma, Q. Zeng, X. Chen, J. Zhao, J. Li, M. Jiang, P. Lio, N. Chawla, C. Zhang, and Y. Ye. Graph Foundation Models: A Comprehensive Survey, May 2025. arXiv:2505.15116 [cs].

When exceeding the maximum context length
Successor and predecessor relations