**C.3. Aim 3: Developing a biological-informed token pruning method to accelerate training.**

**图片包含 图形用户界面

AI 生成的内容可能不正确。C.3.1: Rationale**

Biological data underpins our understanding of genomic function, yet much of the genome can be considered “noise” from a regulatory standpoint. Researchers commonly define enhancers by the presence of histone modifications such as H3K4me1 and H3K27ac, and promoters by H3K4me351. Crucially, the sequences that matter most for gene regulation are often concentrated within these marked regions. However, standard modeling approaches tend to feed entire genomic segments into the model, forcing it to expend considerable capacity on filtering out uninformative regions.

*Figure 8. A snapshot of biological-informed token pruning.*

To address this inefficiency, we can leverage histone mark signals and other regulatory annotations for causal inference, pinpointing the genuinely important regions52. By restricting our input to these high-confidence regulatory elements, we minimize noise and reduce the computation burden on the model. This strategy not only streamlines computation but also enhances interpretability, as the model is then focused on the most biologically relevant sequences for downstream tasks.

**C.3.2: Research Plan**

**C.3.2.1: Preliminary Data**

To simplify the experimental setup, we begin by selecting a biologically relevant dataset aimed at predicting gene expression levels53. This choice reflects a common and impactful task in genomics, as understanding how sequence variants or regulatory elements influence gene expression is crucial for unraveling the mechanisms of gene regulation. By focusing on gene expression prediction, we can leverage existing chromosome accessibility like ATAC-seq54 data as our guide signal.

**C.3.2.2: Naïve Approach to Masking Tokens**

A straightforward method for reducing input complexity involves using biological signals (e.g., histone marks or other regulatory annotations) to identify less relevant regions of the genome. A global threshold is applied to these signals, and any segment falling below that threshold is considered non-essential. Instead of removing these regions entirely, we replace them with a special “SEP” token that also encodes the distance between the core, unmasked sequences. For instance, if a 500-base pair stretch is removed, the mask token might carry information indicating the approximate length or distance of the masked region. This strategy ensures that the model remains aware of potential spacing or contextual relationships, even if the specific nucleotides in that region are deemed unimportant (Figure 8).

**C.3.2.3: Deep Learning Model for Causal Inference and Token Reduction**

Building on the naive approach, a more sophisticated alternative is to train a small deep learning model dedicated to deciding whether a particular region should be masked. Conceptually, we aim to learn a probability function according to following formula:

To model , we use a deep learning model to output the parameters and ​ of a Beta distribution:

and then sample a probability for masking region from:

By learning which signals are most predictive of regulatory activity, this model can dynamically mask out regions that are unlikely to contribute to the final prediction, thereby improving efficiency and focusing the downstream model on the most informative portions of the sequence.

**C.3.2.4: Acceleration of Training Process**

图片包含 图形用户界面

AI 生成的内容可能不正确。In general, we propose a general framework for accelerating model training by focusing on the most informative regions of the sequence. First, we identify a “core” sequence—either through a naive threshold-based method or a small deep learning model designed to pinpoint functionally relevant segments (Figure 9A). During training, we retain these core regions while randomly sampling flanking regions as background sequence. This approach encourages the model to concentrate on the core sequence where most of the regulatory signals are found, yet still incorporates a subset of surrounding nucleotides for contextual information. By reducing the overall sequence length in each training batch, we can train with a larger batch size and improve computational efficiency. Moreover, by filtering out uninformative segments upfront, the model converges more rapidly, leading to faster and potentially more robust learning (Figure 9B).

*Figure 9. A snapshot of training acceleration framework and expected training curve.*

**C.3.3: Expected Outcomes**

From a performance perspective, masking part of sequences is not beneficial for the deep learning model performance because some important region may didn’t annotated by our chosen signal. However, by filtering out uninformative regions at the outset, the model can devote more capacity to learning from genuinely relevant sequences, potentially keeping predictive power.

In terms of efficiency, replacing large sections of non-critical sequence with a single mask token—augmented with distance information—dramatically reduces the number of tokens fed into the model. This compression can allow for much longer genomic contexts to be processed at once, potentially doubling or tripling the effective input length without a corresponding increase in computational cost. Furthermore, once this initial pruning is performed, the techniques outlined in Aims 1 and 2 can be applied to further streamline token usage, ultimately enabling entire chromosomes to be processed if desired.

Regarding interpretability, this approach offers a middle ground. While it may not be as straightforward as entropy-based token merging (Aim 1) or model-based pruning (Aim 2), the deep learning model responsible for masking decisions can still provide insight into which biological signals triggered the pruning. Observing how certain histone modifications or regulatory markers lead to masking decisions can still reveal which genomic features most strongly influence the downstream task.

**C.3.4: Potential Pitfalls and Alternative Approaches**

A key limitation of this method is its reliance on biological data. If such data are unavailable or incomplete, it becomes more challenging to determine which regions should be masked. In these cases, other proxies for functional relevance may be necessary. One possibility is to use conservation scores derived from multiple sequence alignments (MSAs)55,56, as conserved regions57 often correlate with regulatory or functional importance in both protein and non-coding genomic contexts. Another option is to rely on annotations from a closely related cell type or organism, using chromatin state58 or similar regulatory profiles to guide masking decisions. By combining these alternative signals, researchers can still benefit from token reduction strategies, even when direct biological data are limited.