Learning Mutational Contexts Beyond Fixed *k*-mers

Seungho Oh[1], Sangwoo Kim[2]

[1]Graduate School of Medical Science, Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, Republic of Korea

[2]Department of Biomedical Systems Informatics and Brain Korea 21 Project, Yonsei University College of Medicine, Seoul, Republic of Korea

Abstract

Fixed-length *k*-mer contexts underpin most mutational-signature pipelines yet under-specify biologically meaningful sequence structure and long-range dependencies. We explore a representation-learning approach that (i) learns adaptive, variable-length/shape tokenization of mutation-centered DNA and (ii) uses learned sequence embeddings for downstream prediction, with sample-level aggregation currently via a simple voting baseline (a permutation-invariant multiple instance learning model, MIL, is a work in progress but not reported here).

Data and pretraining. We assembled 11,670,722 PASS somatic variants from four PCAWG cancer types—Skin-Melanoma (n=111), Eso-AdenoCA (n=142), Breast-AdenoCA (n=98), and CNS-Medulloblastoma (n=71); total n=422 samples. For each variant we extracted 512-bp flanking sequences. A ~139M-parameter Transformer in the MxDNA family was pretrained on a down-sampled subset (1,000,000 sequences) using a 15% masked-token objective on 2×H100 GPUs.

Fine-tuning and current baseline. We fine-tuned a per-sequence classification head to predict cancer type from mutation-centered sequence alone. In this sequence-level setting we observe ~60% accuracy and macro-F1 across the four classes under heterogeneous per-sample sequence counts. Sample-level labels are presently obtained by majority vote over per-sequence predictions; this serves as the baseline aggregator.

Qualitative interpretability. A visualization pipeline overlays **learned tokens** on sequences stratified by single-sample assignments (SigProfilerSingleSample). In Breast-AdenoCA, token “hotspots” qualitatively align with SBS1-like N[G>A]TC (GA[C>T]N, in reverse complement) patterns, suggesting that adaptive tokens recover signature-relevant motifs beyond canonical 3-mer encodings.

Scope and outlook. The present work establishes (a) a scalable corpus and compute recipe for mutation-centered sequence pretraining, (b) a functioning per-sequence classifier with a sample-level baseline, and (c) a token-visualization workflow linking model saliency to signature annotations. Ongoing work (not included here) implements a MIL module for direct sample-level exposure estimation and a prototype of mutagen mapping algorithm to link tumor genomes with experimental exposure signatures, including controlled mixtures. Together, these pieces lay the groundwork for replacing fixed *k*-mer contexts with a data-driven system and for systematically evaluating how learned tokens map onto mutational processes.