Input sequence generations

Tokenization methods

[CLS] CTCGCT GTTGTA TGCCGA A A [SEP]
[CLS] GGGCTC GGTACC GTATCA G T [SEP]
[CLS] CGGCTC CCGCAT GACTGG A T [SEP]

The non-overlapping 6-mer

[CLS] CTCG CTGT TGTAT GCCGA [SEP] [PAD]
[CLS] GGGCT CGGTA CCGT ATCA [SEP] [PAD]
[CLS] CGG CTCCC GCAT GACT GG [SEP]

BPE

Sampling strategies

Reference Sampling

CTCGCTGTTGTATGCCGAAA
GGGCTCGGTACCGTATCAGT
CGGCTCCCGCATGACTGGAT

Pan-genome Sampling

TTATCGAGTCCCCTGGAAAC
ACCATCTTGTTCATGCGTTT
AACAGCCAAGGGCCAAACAG

Variation Sampling

TGGGGACTTATATGAATTTT
GTCGTCCACCAAGAATAACT
CCGGGACGTCGAAGCTTAAC

Datasets

The human reference genome

Based on the genome assembly GRCh38

The human pan-genome

Based on the 1000 Genomes Project