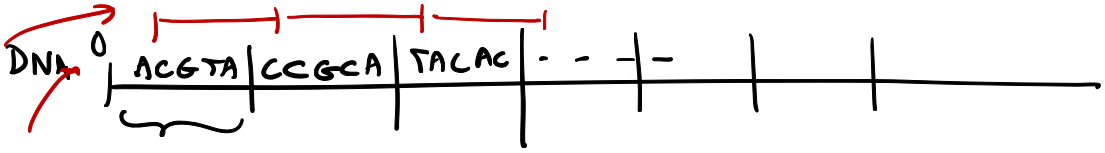
DNA/genome : 3 billion base pairs

Protein sequence : Uniprot ~ 100 million known sequence

k-mers
n-grams



k=5

offset

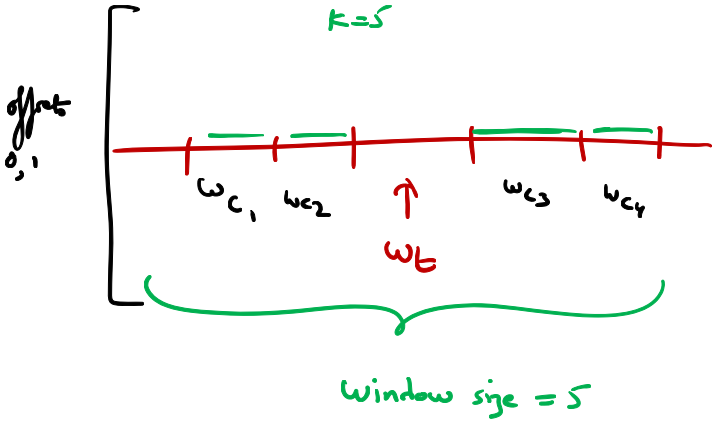
I) Need 3 ORF → 0, 1, 2

Open reading Frame ← transcription / translation

II) for k-mer

use all k offsets

k = 0, 1, 2, . . , k-1



Window size = 5