



Session 7

K-mers

-> [Table of Contents](#) <-

1	Extract K-mers of a seq
1	What is a k-mer
1	Extract a K-mer
2	Why So Popular in Bioinformatics
2	Applications of k-mer Counting
3	Find & Count a given K-mer
3	Finding
3	Counting
4	Forces Affecting K-mer Freq
5	Most Popular K-mers
5	Most Popular K-mers with (X) Mismatches

◀ Extract K-mers of a seq

◆ What is a k-mer

- A *k-mer* is just a sequence of *k* characters in a string (or nucleotides in a DNA sequence). Now, it is important to remember that to get *all k-mers* from a sequence you need to get the first *k* characters, then move just a single character for the start of the next *k-mer* and so on. Effectively, this will create sequences that overlap in $k-1$ positions.

◆ Extract a K-mer

- by way of example the next sequence
the Seq("ATCGATCAC")
One with 3-mer & the other with 4-mer

- (*k-mer of size 3*) => ATCGATCAC

Sequence: ATCGATCAC

3-mer #0: ATC

3-mer #1: TCG

3-mer #2: CGA

3-mer #3: GAT

3-mer #4: ATC

3-mer #5: TCA

3-mer #6: CAC

- (*k*-mer of size 4) => ATCGATCAC

Sequence: ATCGATCAC

4-mer #0: ATCG

4-mer #1: TCGA

4-mer #2: CGAT

4-mer #3: GATC

4-mer #4: ATCA

4-mer #5: TCAC

◆ Why So Popular in Bioinformatics

- Decomposing a sequence into its *k*-mers for analysis allows this set of fixed-size chunks to be analysed rather than the sequence, and this can be more efficient.
- *K*-mers are very useful in sequence matching (string matching with [n-grams](#) has a rich history), and set operations are faster, easier, and there are a lot of readily-available algorithms and techniques to work with them.
- A simple example: to check if a sequence *S* comes from organism *A* or from organism *B*, assuming the genomes of *A* and *B* are known and sufficiently different, we can check if *S* contains more *k*-mers present in *A* or in *B*. Yes, there are many tools that do just that.
- Basically, using *k*-mers simplifies bioinformatics to counting and comparing whether things are there or not.

◆ Applications of k-mer Counting

- Genome assembly, Sequence alignment, Sequence clustering, Error correction of sequencing reads, Genome size estimation, Repeat identification.

◀ Find & Count a given K-mer

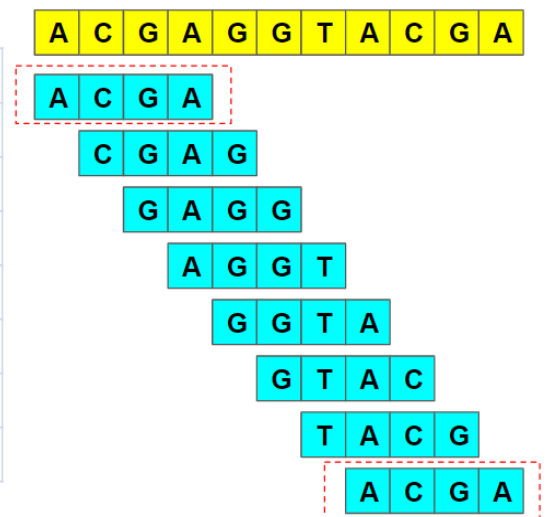
◆ Finding

- We Can Find the K-mers in the sequence as a movable imaginary box consisting of a constant number that the search is based on.
- For Ex. suppose we have a seq("ATTCGAACT") and (k-mer) of size (4). Then the K-mers will be: (ATTC), (TTCG), (TCGA), (CGAA), (GAAC), (AACT). Then there is a 6 k-mers of the sequence with the size 4.

◆ Counting

- The *Total Count* is simply how many times each k-mer has appeared in the given sequence. Except for ACGA, which has appeared twice, the rest of the 4-mers have appeared only once.
- (k-mer of size 4) => ACGAGGTACGA

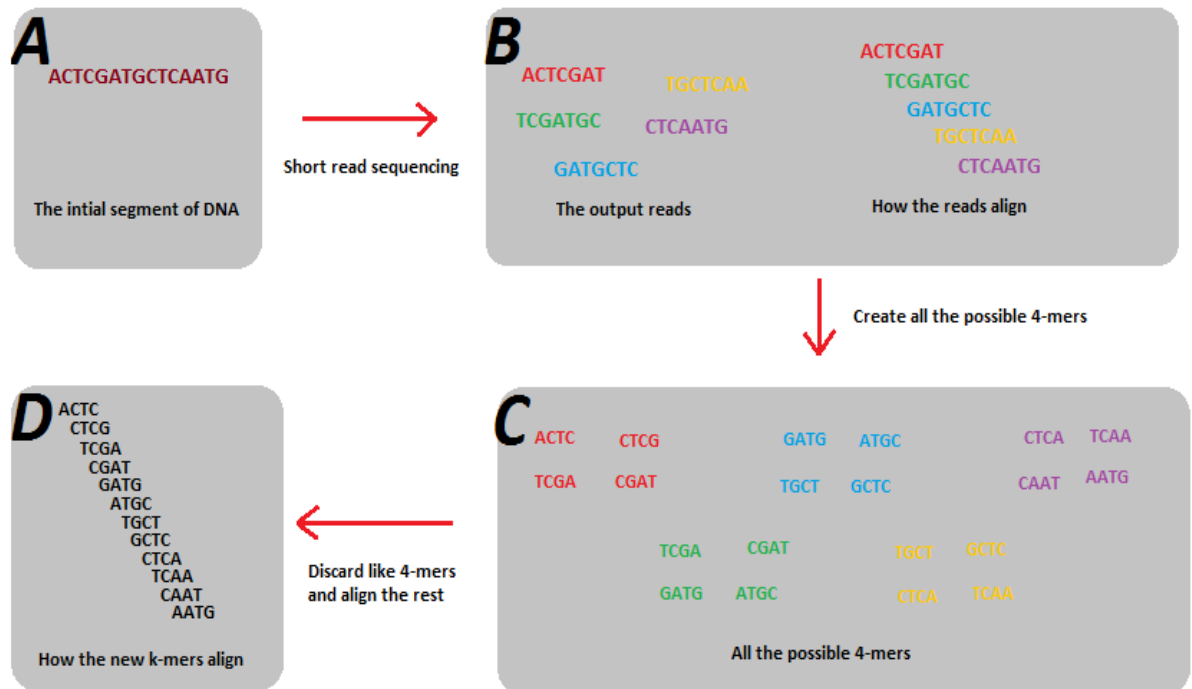
	Total	Distinct	Unique
ACGA	2	1	0
CGAG	1	1	1
GAGG	1	1	1
AGGT	1	1	1
GGTA	1	1	1
GTAC	1	1	1
TACG	1	1	1



- The distinct k-mers are counted only once regardless of how many times they appear. For example, even though ACGA has appeared twice, we count it only once. The *Distinct Count* provides us the information of if a k-mer has appeared or not (not how many times it has appeared).

- The unique k-mers are those which appear only once. In the above example, since ACGA has appeared twice, its unique count is zero. The *Unique Count* reveals which k-mers have appeared only once.

- The next Ex. shows all the possible (6 & 4 K-mers) in a Sequence:



◆ Forces Affecting K-mer Freq

- The frequency of *k*-mer usage is affected by numerous forces, working at multiple levels, which are often in conflict.
- It is important to note that *k*-mers for higher values of *k* are affected by the forces affecting lower values of *k* as well.
- For example, if the 1-mer A does not occur in a sequence, none of the 2-mers containing A (AA, AT, AG, and AC) will occur either, thereby linking the effects of the different forces.

◆ Most Popular K-mers

- We say that k-mer is popular or not depending on how often it is repeated in the Sequence.



Here we see that there are Popular K-mers: (CGG), (GGC), (GCA)
Each of them repeated 2 times in the Sequence so they are the most Popular in the Sequence.

◆ Most Popular K-mers with (X) Mismatches

- Here Our Sequence is: “**ATGTCGATCATTATG**” and we have already a popular K-mer but with num of (x) Mismatches.
- That means if we Suppose there is a popular K-mer which is: “ATG” and we want to find if there is (1) Mismatches of this popular K-mer, It could be:

(TTG), (CTG), (GTG),
(AAG), (ACG), (AGG),
(ATA), (**ATT**), (**ATC**)

These are all the Possible tries to get all the ((1) Mismatches) of “ATG”

- Then we return to the Sequence and Compare all its K-mers with these Probabilities of mismatches.
- When we Compare Both of Them, we will find that the ((1) Mismatches) of “ATG” which is popular is: (**ATC**), (**ATT**). so we have 2 Mismatches.