_k_-mer Group Activity Sheet V1 – Class #8 - Software Tools – U Guelph

Activity adapted from Sally Adamowicz; last updated September 29, 2024

Learning Outcomes: By the end of this short group activity, you should be able to:

1. describe what is a *k*-mer
2. calculate *k*-mer frequencies
3. compare results of DNA sequence comparisons from *k*-mer frequencies vs. alignment

Context: Calculating simple features from biological sequences is an important approach in bioinformatics, which is used for a wide range of applications (including taxonomic classification, gene identification, and pathogen detection). This approach involves calculating or extracting simple measures for each individual sequence. This approach does not rely on aligning sequences to one another and hence is more computationally efficient than alignment for making sequence comparisons. Calculating sequence features by hand for simple examples enables us to understand this idea, and then we will move into using software tools to perform these calculations for us.

What is a *k*-mer? A *k*-mer is a word of length *k*. Therefore, a 1-mer consists of one character, a 2-mer of two characters, a 3-mer of three characters, etc. Calculating a *k*-mer profile for a given sequence (i.e. string) involves counting how many instances there are of every unique word of length *k* in that string. This is easiest to understand through an example. We will be working with hypothetical DNA sequences of total length 10 with four characters (A, C, G, T).

1-mer counts (a.k.a. nucleotide frequencies)

After looking at the example, fill in the remainder of this table with the count for each 1-mer.

No.	Sequence	A	C	G	T
1	ACTGACTGAC	3	3	2	2
2	ACACACACAC	5	5	0	0
3	AAAAACCCCC	5	5	0	0
4	ACTGACTGGC	2	3	3	2

Question: On the basis of nucleotide frequency alone, which pair of sequences is the most similar in this data set?

Answer: 2 & 3

2-mer counts (a.k.a. dinucleotide frequencies)

After looking at the example, fill in the remainder of this table with the count for each 2-mer.

No.	Sequence	AA	AC	CT	TG	GA	CA	CC	GC	GG
1	ACTGACTGAC	0	3	2	2	2	0	0	0	0
2	ACACACACAC	0	5	0	0	0	4	0	0	0
3	AAAAACCCCC	4	1	0	0	0	0	4	0	0
4	ACTGACTGGC	0	2	2	2	1	0	0	1	1

Question: On the basis of dinucleotide frequency, which pair of sequences is the most similar in this data set?

Answer: 1 & 4

Comparison with alignment-based approach

Now, look at the sequences against one another instead of individually. Calculate the p-distance for each pair of aligned sequences (i.e. the proportion of positions that differ). Look at the example and fill in the half of the table below the diagonal. Along the diagonal, the p-distances are zero, as each sequence is identical to itself (i.e. no nucleotide positions differ in the pairwise alignments).

		1 ACTGACTGAC	2 ACACACACAC	3 AAAAACCCCC	4 ACTGACTGGC
1	ACTGACTGAC	--	0.4	0.6	0.1
2	ACACACACAC	0.4	--	0.4	0.5
3	AAAAACCCCC	0.6	0.4	--	0.6
4	ACTGACTGGC	0.1	0.5	0.6	--

Question: Which pair of sequences is the most similar?

Answer: 1 & 4

Question: Was the answer based upon aligned sequences more similar to the answer from nucleotide frequencies or dinucleotide frequencies?

Answer: The answer was closer to the dinucleotide frequencies.