CS4330: Combinatorial Methods in Bioinformatics

Overview on K-mers in genomics

Wong Limsoon

What K-mers are

K-mers are short DNA sequences of length K

They are crucial for various genomic analyses due to their ability to capture essential information about DNA sequences

By breaking down the genome into these short fragments, researchers can gain valuable insights into genomic structure, variations, and functional elements

K-mers in a DNA sequence

K-mers are length-k substrings contained in the DNA sequence

# of K-mers in a length-L DNA sequence = L – K + 1

Example

Length-10 DNA sequence = `AGCTCAGCTA`

4-mers = `AGCT x 2, GCTC, CTCA, TCAG, CAGC, GCTA`

Confusing terminologies

Consider `AGCTCAGCTA` and K = 4

K-mer species, unique K-mers, distinct K-mers, K-mers:

`AGCT, GCTC, CTCA, TCAG, CAGC, GCTA`

K-mer occurrences, distinct K-mers, K-mers:

`AGCT``, GCTC, CTCA, TCAG, CAGC, `**`AGCT`**`, GCTA`

Most of the time, people use the term “K-mers” to refer to both senses of the word. You have to figure out which sense is intended

K-mers from sequencing projects

Dataset	Genome size	Read length	Coverage	No. paired-end reads	Input size (fastq)
D1	2.8M	101	46.3×	1 294 104	280M
D2	4.6M	101	33.6×	766 646	446M
D3	88.3M	101	38.3×	16 757 120	9.4G
D4	249.2M	124	150.8×	303 118 594	92G
D5	3121.8M	101	27.6×	854 084 773	442G

Data	$\|k\text{-mer}_1\|$(m)	$\|k\text{-mer}_{2-1000}\|$(m)	MU_1(G)	MU_{2-1000}(G)	MR_1	MR_{2-1000}	MR_{all}
D1	35.67	3.49	0.056	0.021	20.53	5.599	15.82
D2	54.13	5.91	0.085	0.035	20.52	5.689	15.67
D3	372.09	99.92	0.593	0.626	20.22	5.378	12.45
D4	4643.11	543.89	7.418	3.227	20.17	5.678	15.72
D5	4171.45	2748.5	6.665	17.16	20.17	5.396	9.368

Note: MU, memory usage; MR, memory-saving ratio. The subscript '1' means *k*-mers having frequency of 1, while the subscript '2–1000' represents *k*-mers having frequency lager than 1 but less or equal to 1000. Note, the number of *k*-mers having frequency larger than 1000 is negligible, and their frequency is usually replaced by 1000 by default for *k*-mer counters.

D1 – S. aureus; D2 – R. sphaeroides;
D3 – human chr 14; D4 – B. impatiens;
D5 – HapMap NA12878

Jiang et al., *Bioinformatics*, 35(23):4871-4878, 2019

K-mer counting

Counting the occurrences of K-mers in a genome forms the basis for understanding the genomic landscape and extracting meaningful patterns

Many applications, e.g.:

Genome assembly

- K-mer counts are used in de Bruijn graph-based algorithms for genome assembly.
- The frequencies of K-mers help identify overlaps between sequences, aiding in reconstructing the complete genome from short DNA fragments.

Error correction

- K-mer counting is employed in error correction of sequencing data.
- By identifying rare or erroneous K-mers, researchers can enhance the accuracy of genomic data, particularly in next-generation sequencing experiments.

Comparative genomics

- Comparative studies utilize K-mer counting to compare the genomic content of different organisms.
- Differences in K-mer frequencies can indicate evolutionary relationships and genetic variations.

Challenges & considerations

Memory requirements

Large genome & K-mer length $\Rightarrow$ big memory needed

Trade-offs in K-mer length

Long K-mers are more specific but miss subtle variations

Short K-mers can capture variations but are error prone

Repetitive regions

Complicate K-mer-based analysis

Difficult to have unique K-mers in repetitive regions

Choosing K

Choose K such that:

Big enough so that most K-mers are "unique"

Small enough to reduce computational resource needed for K-mer counting

GenomeScope's recommendation

K = 21

Minimum sequencing coverage > 25x; otherwise, genome size estimation is not accurate

Vurture et al, "GenomeScope", *Bioinformatics* 33(14):2202-2204, 2017

Popular tools - check these out yourself

1. **Jellyfish:**
   - Jellyfish is a versatile and efficient tool for K-mer counting.
   - It supports various operations for analyzing and manipulating K-mer data.
2. **KMC (K-mer Counter):**
   - KMC is a high-performance tool for K-mer counting and analysis.
   - It offers speed and memory efficiency, making it suitable for large-scale genomic datasets.
3. **K-merGenie:**
   - K-merGenie is used for optimal K-mer size selection in genome assembly.
   - It helps researchers choose an appropriate 'K' value for their specific genomic data.

Good to read

K-mers according to Wikipedia

https://en.wikipedia.org/wiki/K-mer#:~:text=The%20frequency%20of%20a%20set,in%20alignment%2Dfree%20sequence%20analysis