COM SCI C121 Week 3

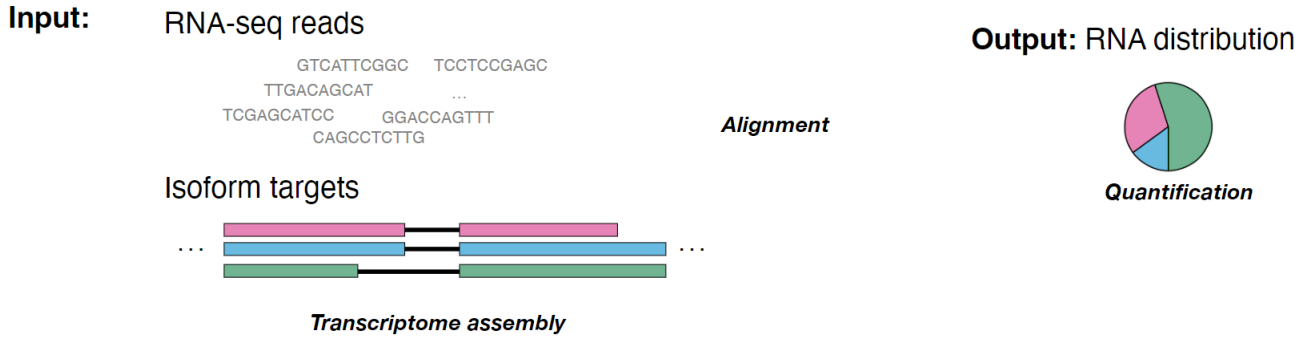
Aidan Jan

April 23, 2024

Pseudoalignment

RNA-seq quantification is a computational problem.

- **Goal:** given a known set of isoform targets (genes) and RNA-seq fragments, recover the distribution of RNA molecules.



Alignment is the (in)exact matching of a subsequence to a reference.

- In simplest terms:
  - chromosome 1, position 342,215
  - transcript A, position 32; transcript B, position 3, ...
- Also possible:

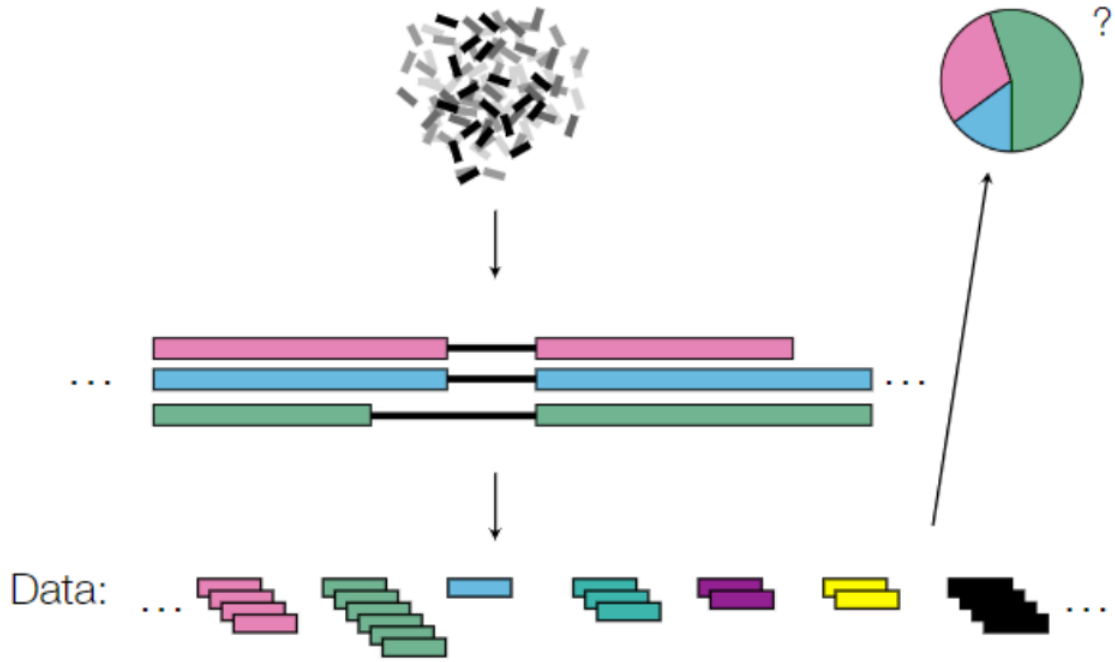
```
TACGGGCCCGCTA-C
TA---G-CC-CTATC
```

A Fundamental Problem: Alignment and Counting

- Classical approaches for exact matching are too slow.
  - Boyer-Moore $O(|R| + |T|)$
- Contemporary methods use heuristics
  - Seed and extend
- Our approach: use the *redundancy* and structure of the target sequences

A Fundamental Problem: Counting and Quantification

- **Quantification:** given many alignments to a reference transcriptome, what is the likely *relative* abundance of each isoform?
  - Complication: most reads will give many, many transcripts



Equivalence classes are sufficient for quantification.

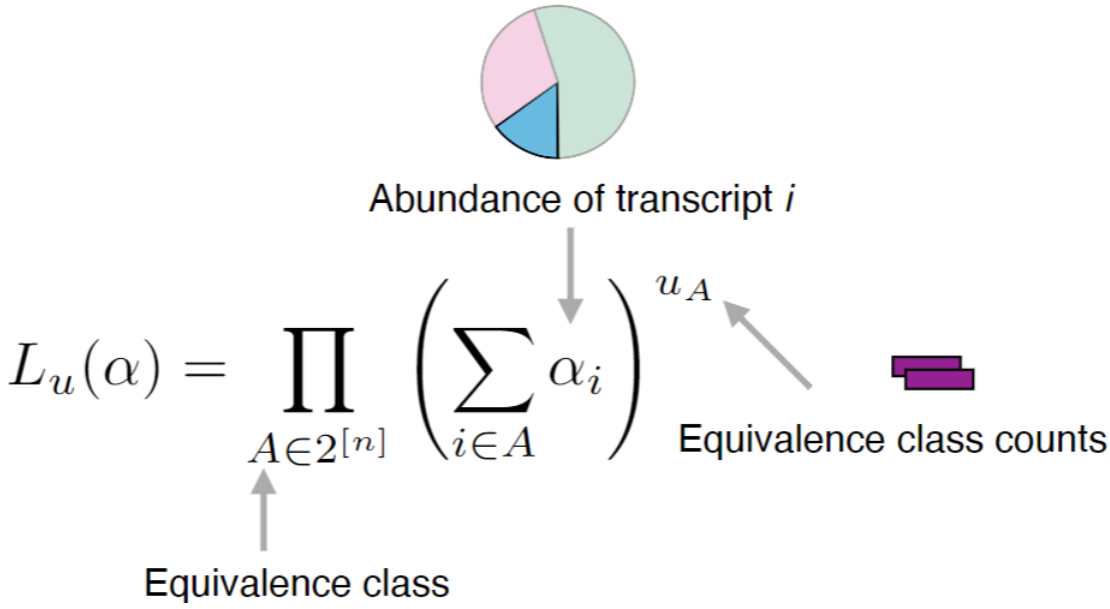
Quick Aside: k-mer

A **k-mer** is a sequence of length k that is a substring of a longer sequence

Consider 'ACGGT':

- k-mers of length 3: ACG, CGG, GGT
- k-mers of length 4: ACGG, CGGT

The Linear Allocation Problem Likelihood



All observed equivalence classes easily fit into the memory on a laptop.

- 100M kmers
- <1M equivalence classes