11.12.2020

RNA Sequencing

- Expression analysis

(genome) → mRNA → protein

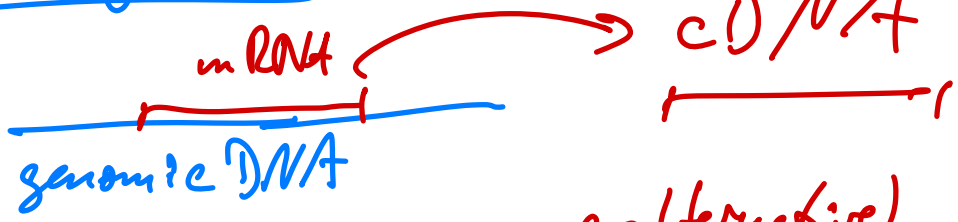
transcription — translation

reverse transcription

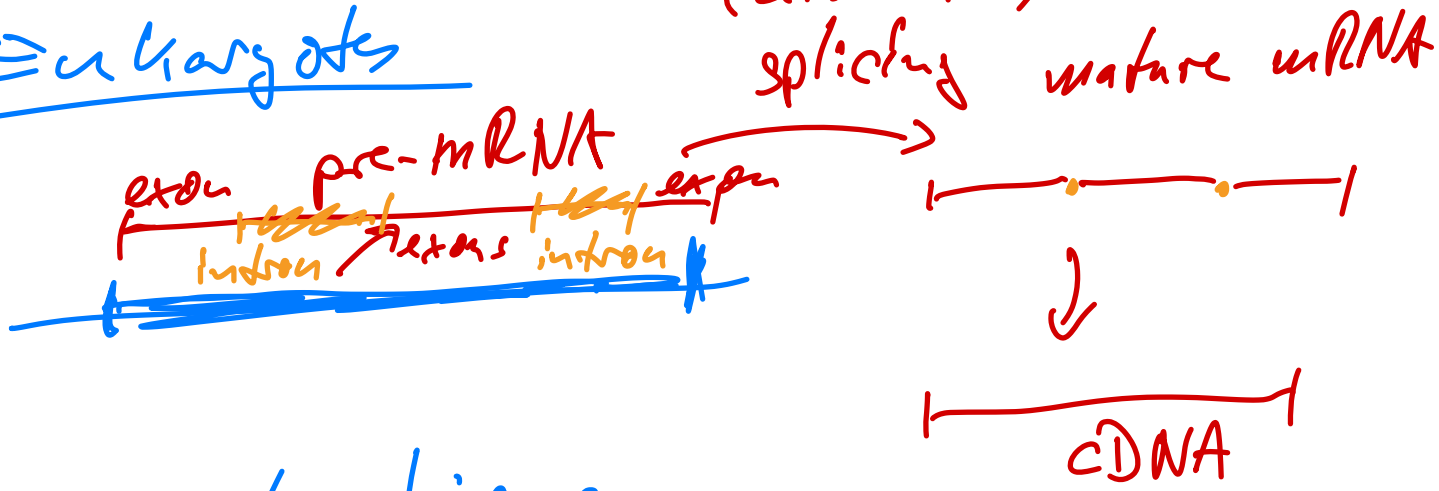
cDNA → RNA-Seq

sequencing

Prokaryotes:



Eukaryotes



Two applications

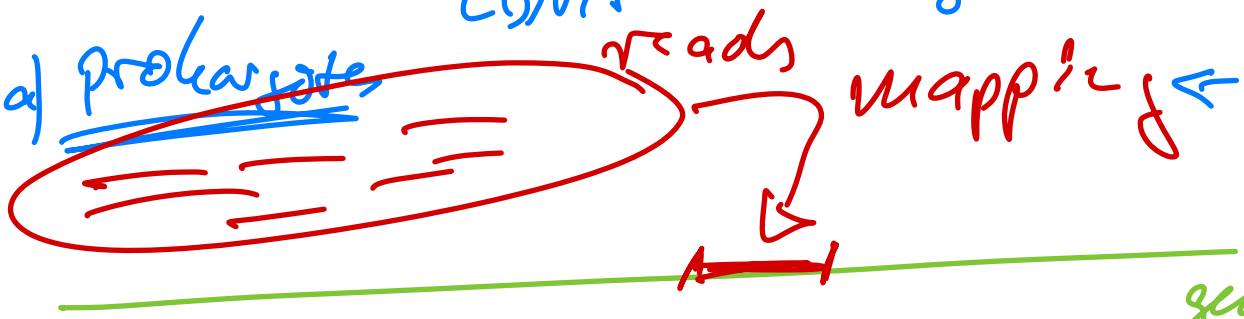
1) De-novo gene identification
2) Quantitative transcriptomics

(1) De-novo gene identification by RNA-Seq

given: RNA-Seq output = cDNA
& genomic sequence

wanted: physical location of the cDNA in the genome

(a) prokaryote



← approx. string matching

genome

Algorithms

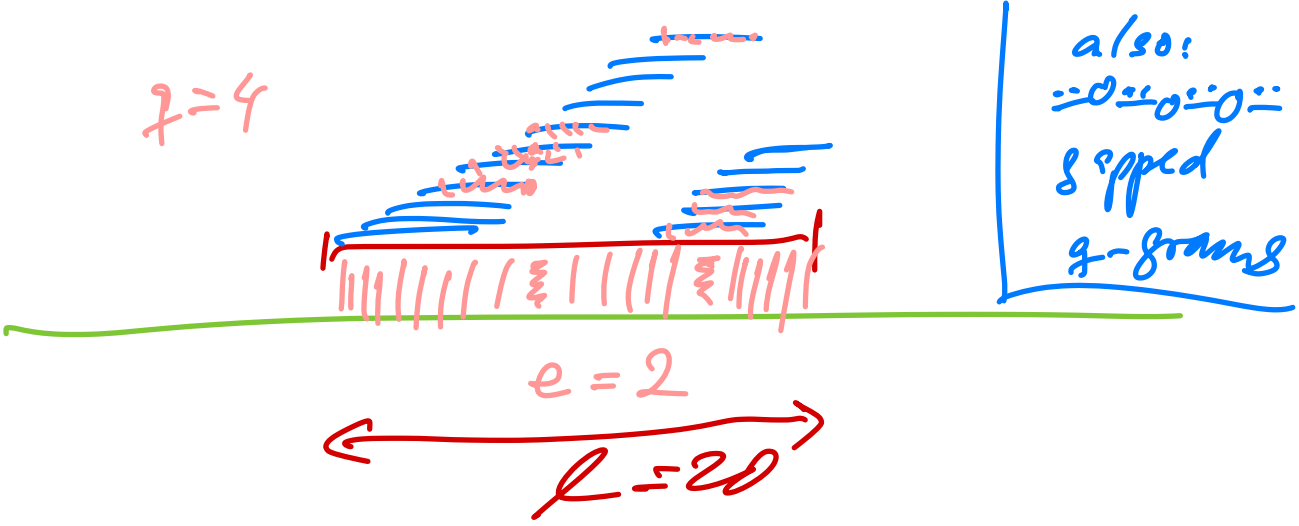
- Sellers' algorithm = semi-global alignment

  $O(n \cdot m)$ n = genome length, m = read length

  ⇒ might be too slow.

- Alternative algorithms for read mapping use

  - **q-gram lemma:**
    at least $(l+1) - q(e+1)$ hits of q-grams exist in a region of length l with e errors



matching q-grams: $l-q+1-e\cdot q$

$= l+1-q(e+1)$

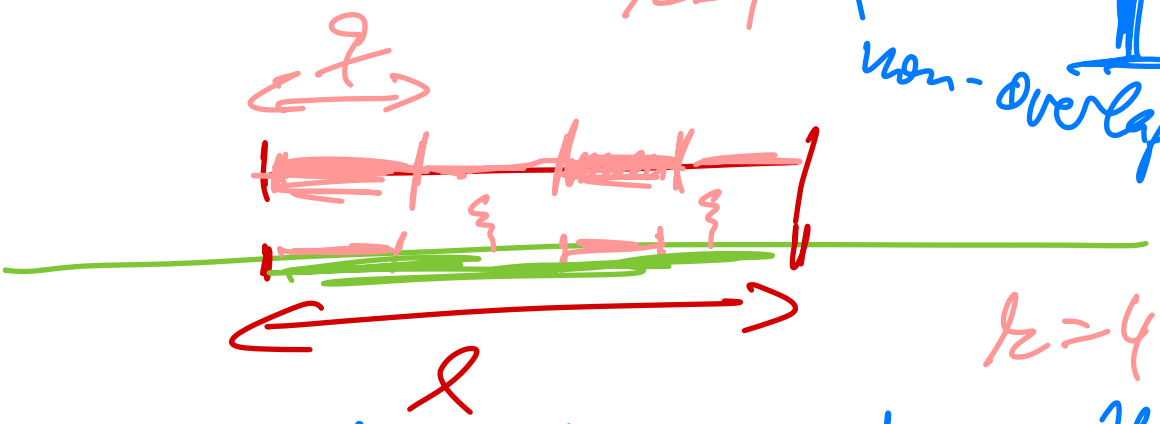
here: $20+1-4(3)=9$

- <u>pigeonhole principle</u>

e errors, k pieces

k=4

non-overlapping

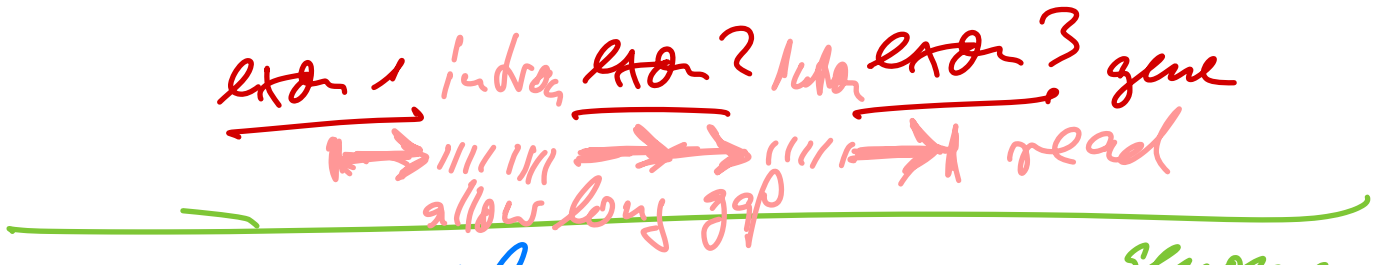


split read in k pieces, then at least k-e pieces must match with the corresponding genome region

- to identify matching q-grams between read and genome, use efficient data structures like
  - q-gram hash
  - suffix array
  - BWT

× software: BWA, BowTie, ...
ELAND, SOAP, ...

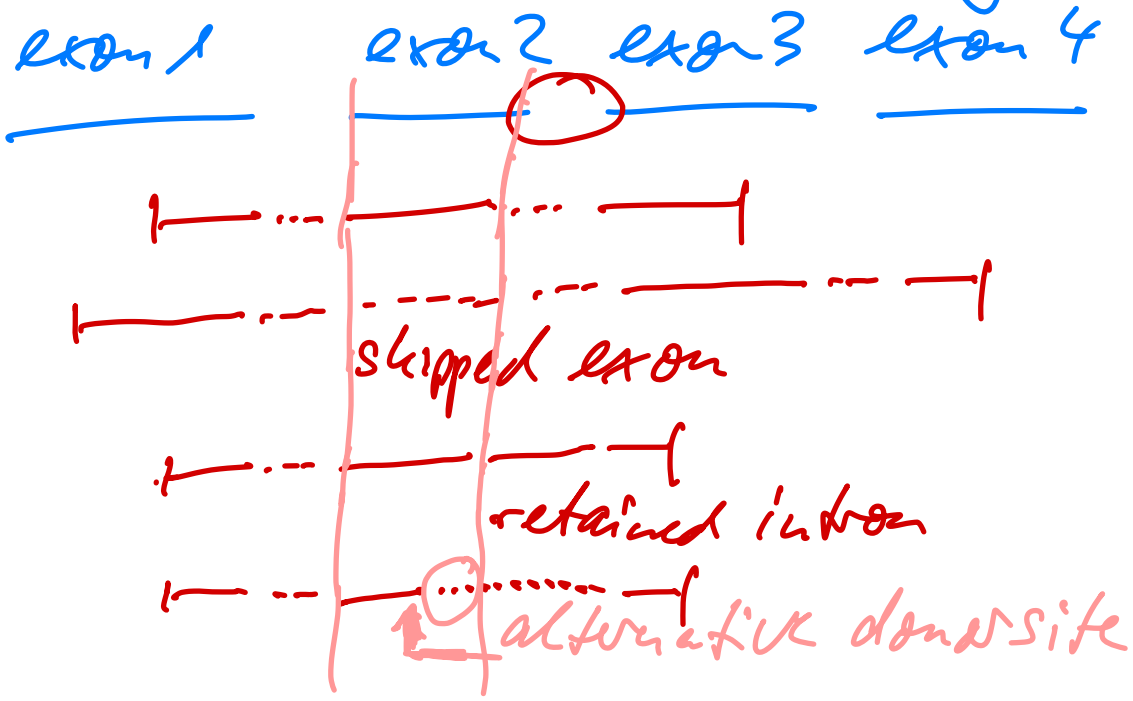
(b) en kargotes



adapt the read mapper

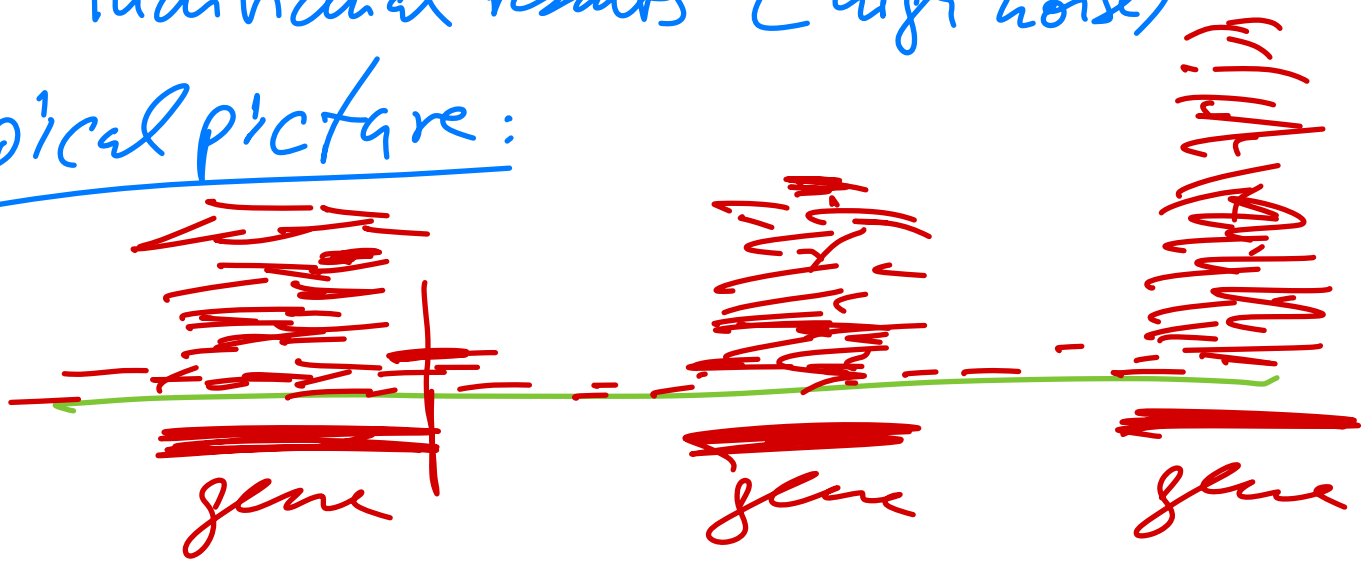
for example, search left-to-right

be careful with alternative splicing



overall result is valid only as the sum/average of many individual results (high noise)

typical picture:

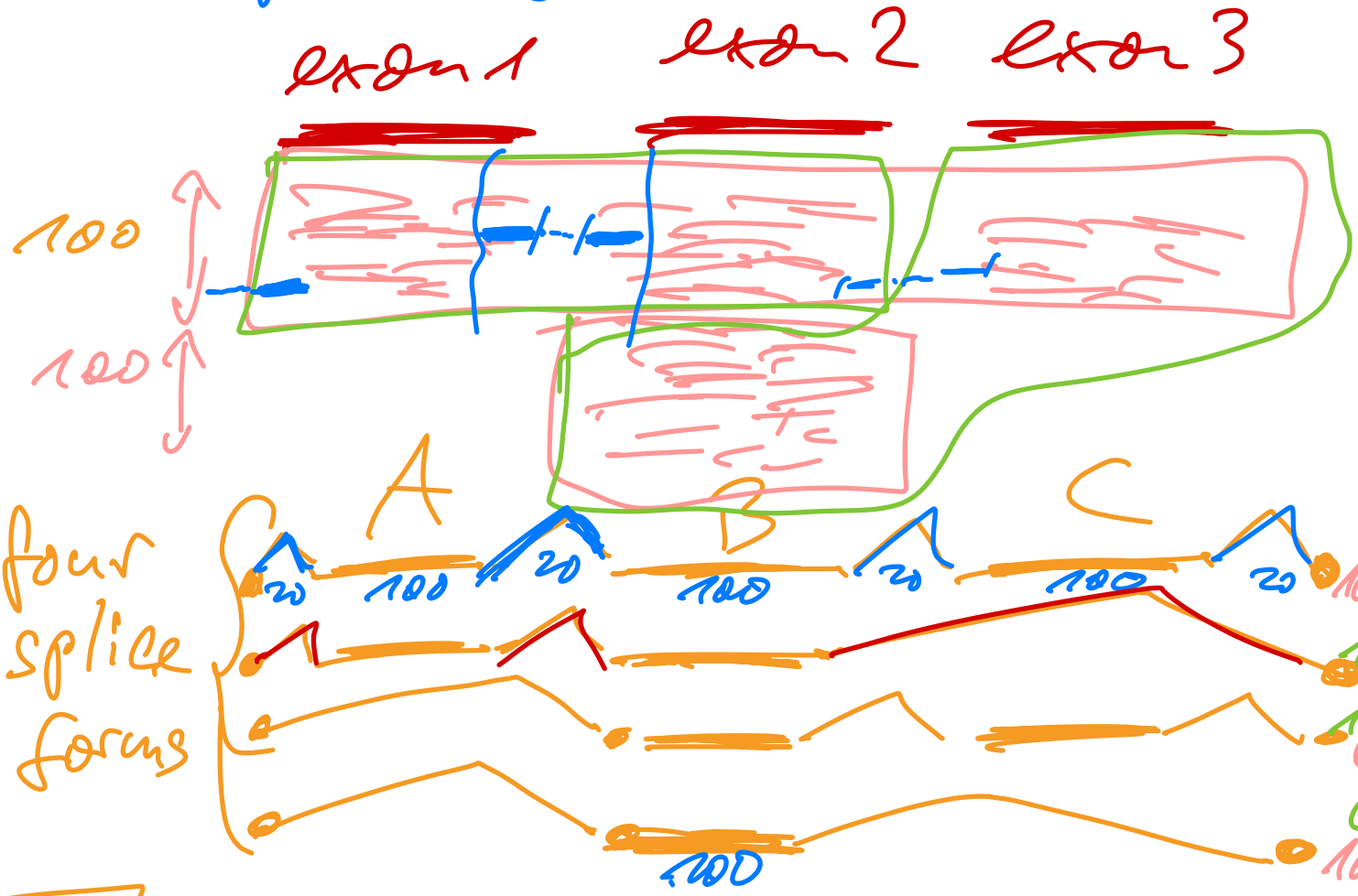


often: use statistics

DeSeq ?

2 Quantitative transcriptomics

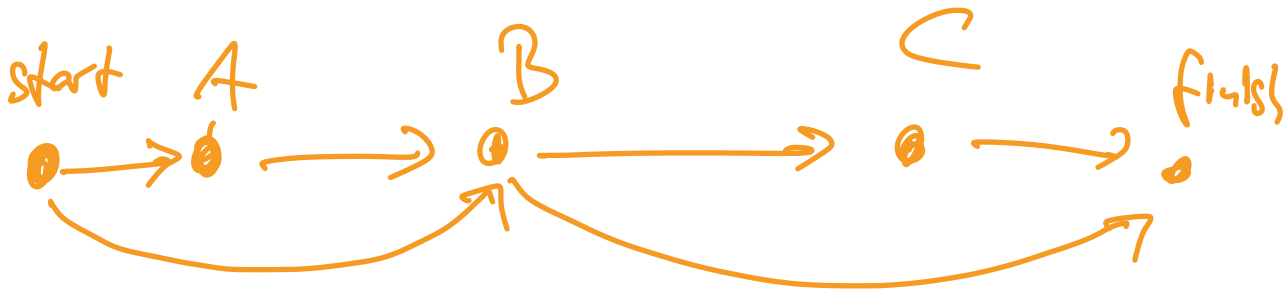
here focus on alternative splicing in eukaryotes



OR

as a splicing graph

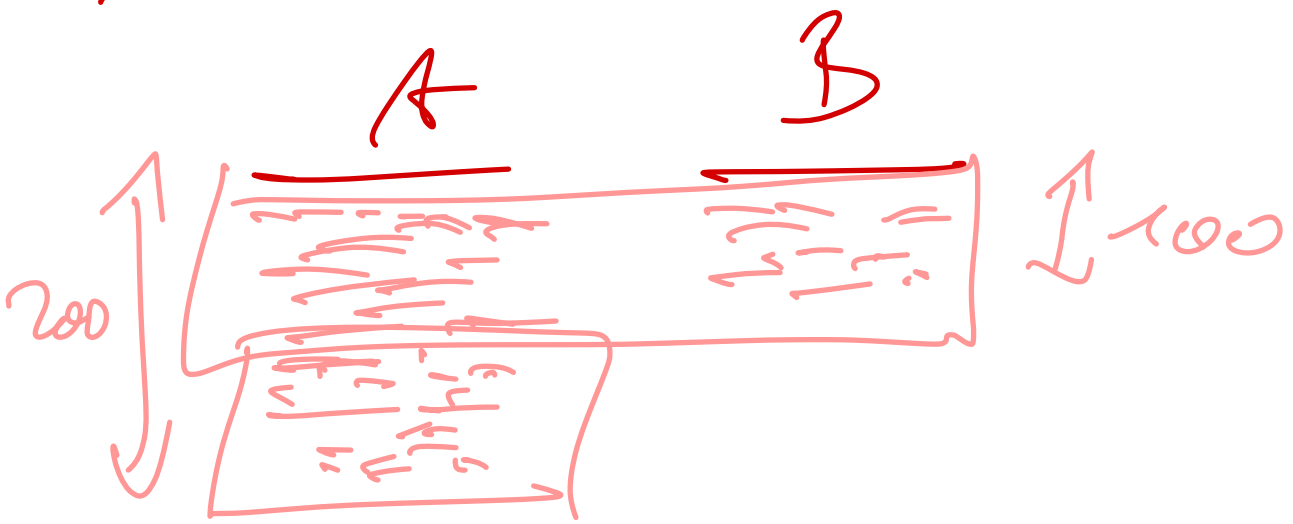
vertices = exons, edges = introns



OR

{ABC, AB, BC, B}

simple pictures



splice forms: AB, A

$\Rightarrow$ A : 100
AB : 100

Lemma: Given the abundances of splice junctions, the abundance of a splice variant (or combination of splice variants) can be predicted if and only if the (combination of) splice variant(s) has unique set of splice junctions.

exons only

1 start A B C finish

2 A B

3 B C

4 B

1 A B C

4 B

2 A B

3 B C

---

1 sA AB BC Cf

2 sA AB Bf

3 sB BC Cf

4 sB Bf

splice junctions

1 sA AB BC Cf

4 sB Bf

2 sA AB Bf

3 sB BC Cf