Mapping

Lecture 10
Sept 23, 2016

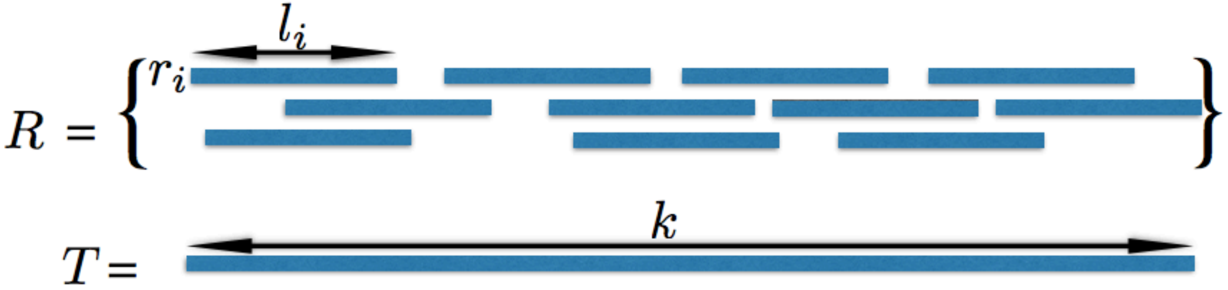
Announcements

- Codes??
- No class on Wednesday
- Practice launching AWS instance!!

What is the alignment problem?

Given: A collection of sequencing reads, and some target sequence (e.g. a genome)

Find: For each read, all locations where the read is within edit distance **ϵ** of the reference, and the edits that achieve this distance.



Edit Distance

Given: Two strings

$$a = a_1a_2a_3a_4\ldots a_m$$
$$b = b_1b_2b_3b_4\ldots b_n$$

where a_i, b_i are letters from some alphabet, Σ, like {A,C,G,T}.

Compute how similar the two strings are.

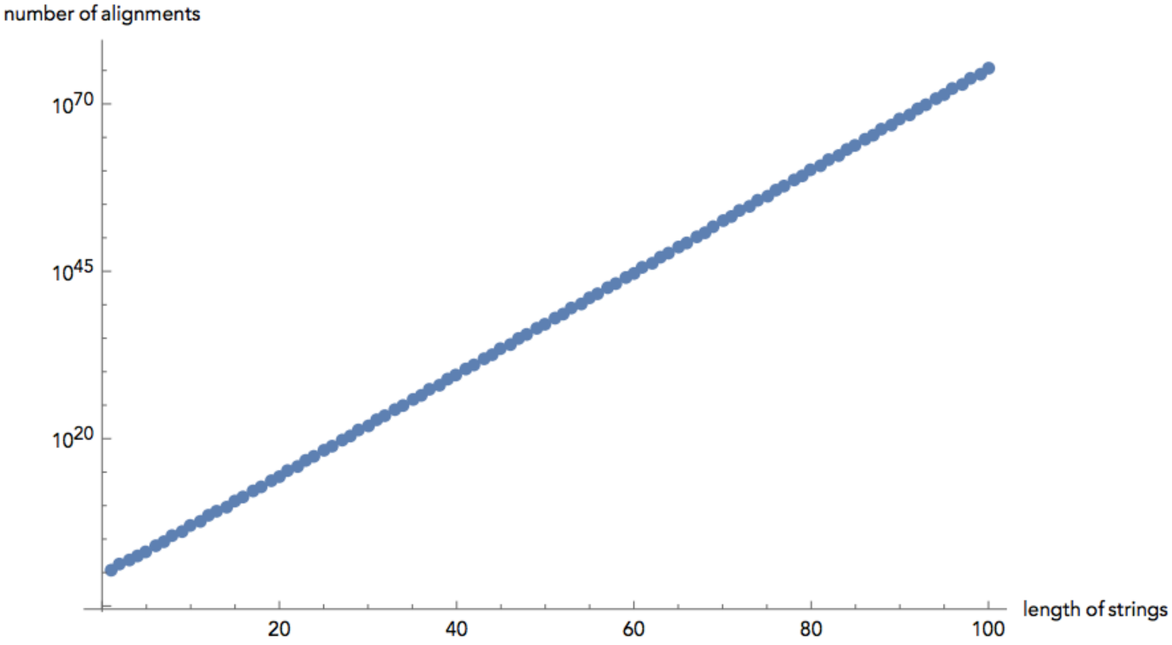
What do we mean by “similar”?

Edit distance between strings a and b = the smallest number of the following operations that are needed to transform a into b:

- mutate (replace) a character
- delete a character
- insert a character

$$\text{riddle} \xrightarrow{\text{delete}} \text{ridle} \xrightarrow{\text{mutate}} \text{riple} \xrightarrow{\text{insert}} \text{triple}$$

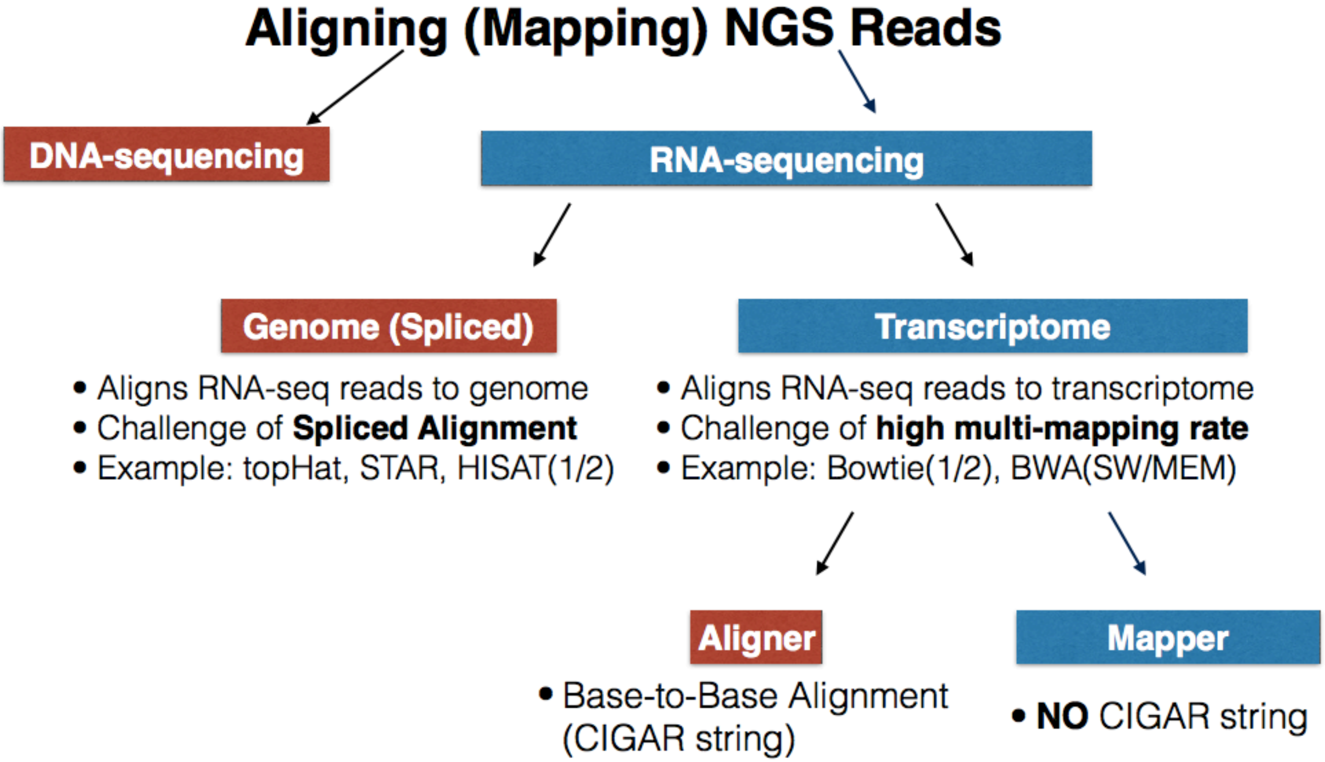
Can't we just test and choose the best?



$$f(n,m) = \sum_{k=0}^{\min(m,n)} 2^k \binom{m}{k} \binom{n}{k}$$

Andrade, Helena, et al. "The number of reduced alignments between two DNA sequences." BMC bioinformatics 15.1 (2014): 94.

Phylogeny of Read-Alignment



Aligning (Mapping) NGS Reads

DNA-sequencing

RNA-sequencing

Genome (Spliced)

- Aligns RNA-seq reads to genome
- Challenge of **Spliced Alignment**
- Example: topHat, STAR, HISAT(1/2)

Transcriptome

- Aligns RNA-seq reads to transcriptome
- Challenge of **high multi-mapping rate**
- Example: Bowtie(1/2), BWA(SW/MEM)

Aligner

- Base-to-Base Alignment (CIGAR string)

Mapper

- **NO** CIGAR string

RNA-Seq Read Alignment

Given an RNA-seq read, where *might* it come from?

Two main “regimes”

Align to transcriptome

Align reads directly to txps

No “split” alignments — transcripts contain spliced exons directly.

Typically *a lot* of multi-mapping (80-90% of reads may map to multiple places)

Does *not* require target *genome*

Can be used in *de novo* context (i.e. after *de novo* assembly)

Align to genome

Align reads to target genome

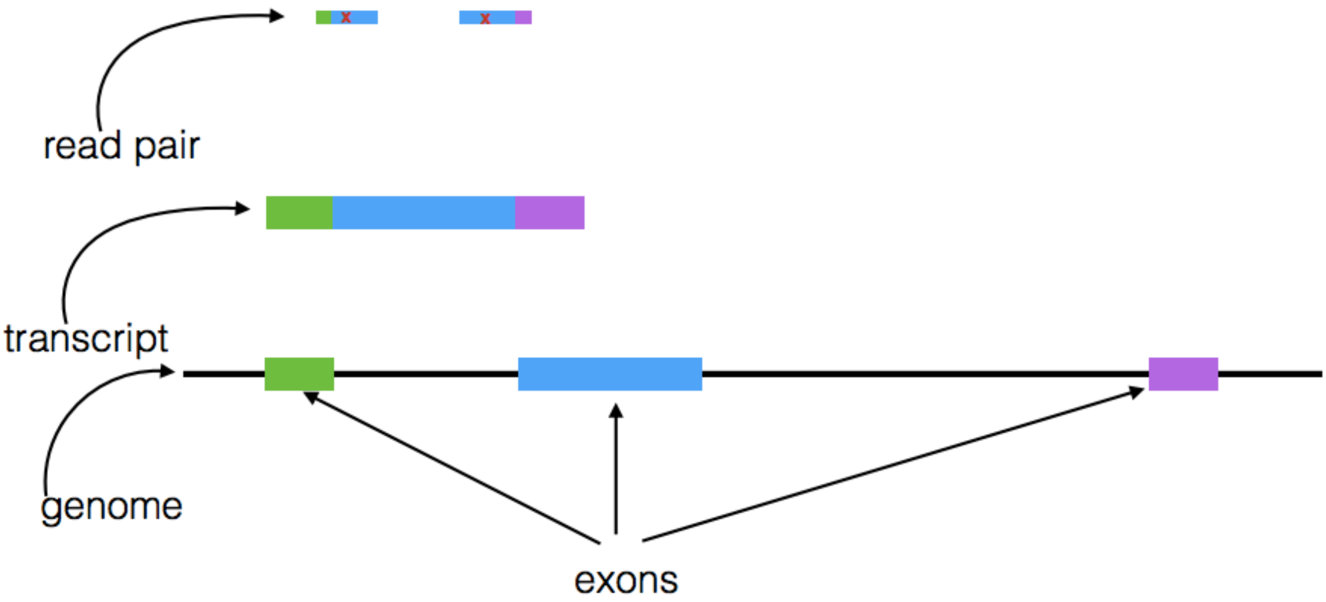
Reads spanning exons will be “split” (gaps up to 10s of kb)

Typically little multi-mapping (most reads have single genomic locus of origin)

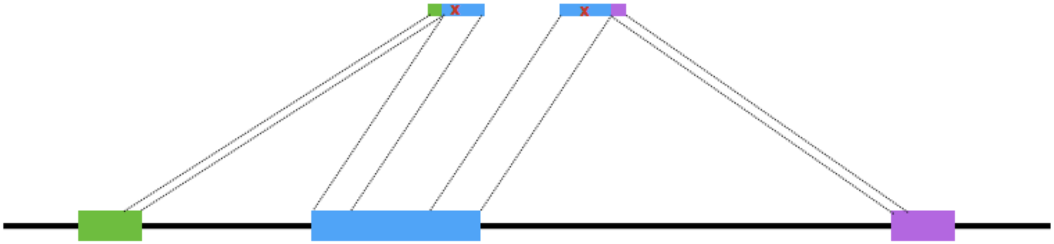
Requires target *genome*

Can be used to find new transcripts

Spliced Alignment



Spliced Alignment



Splice junctions might be known, or *unknown.*

Overlap of read with exon may be *very short*, sequence is ambiguous (e.g. 10 bases).

Sequence of read might be repetitive in the genome.

Aligning reads to a Transcriptome

Consider the following scenario:

Transcripts

Read