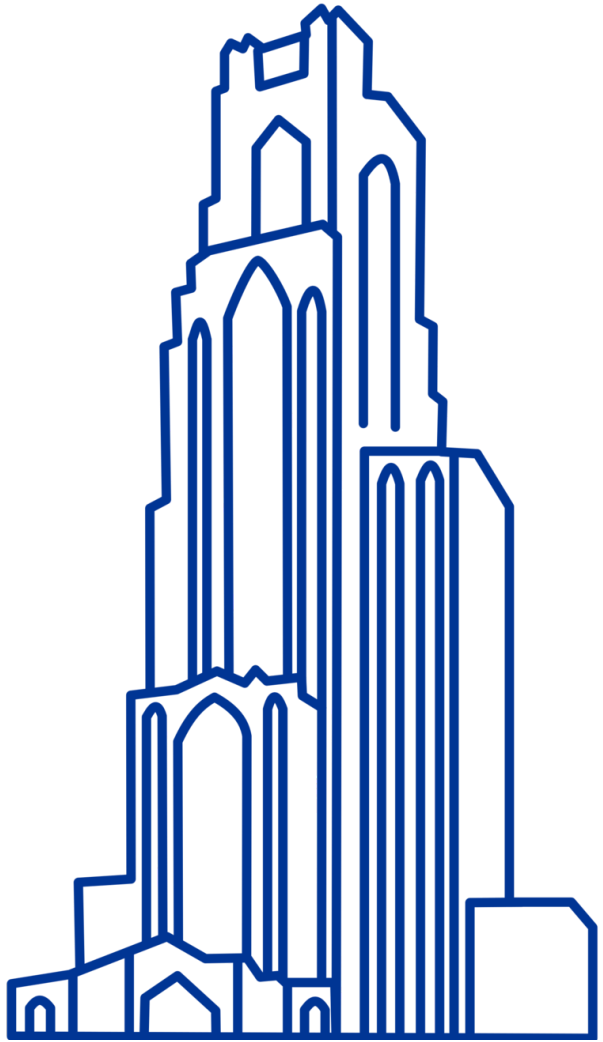


Computational Biology

(BIOSC 1540)

Lecture 07B

Quantification

Methodology

Feb 20, 2025

Announcements

Assignments

- P02A is due March 14 (Q01 will be released tomororw)

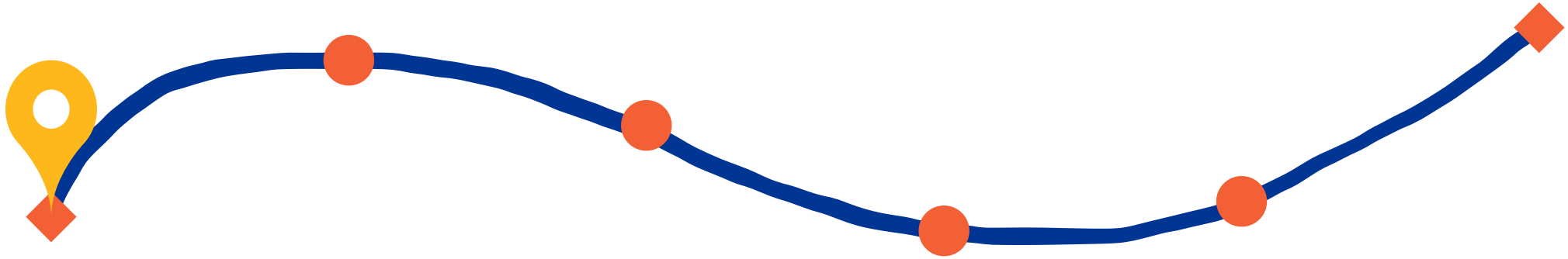
Quizzes

- Quiz 03 is on Mar 18 and will cover L06B to L08B

CBits

- César will provide optional Python recitations on Fridays from 2 - 3 pm (Located in Clapp Hall, room TBD).

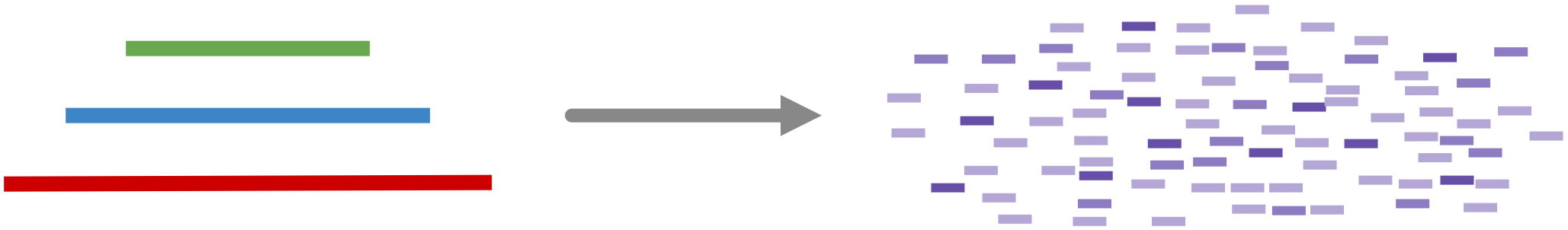
After today, you should have a better understanding of



RNA quantification problem formulation

The RNA quantification problem statement

Given the sequencing reads that were sampled from these transcripts



Transcriptome

Unknown quantity

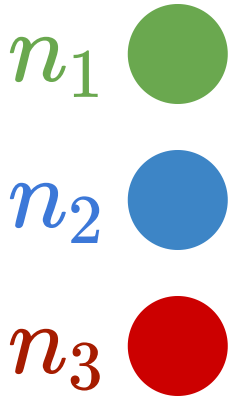
Reads/Fragments

Experimental
biases and errors

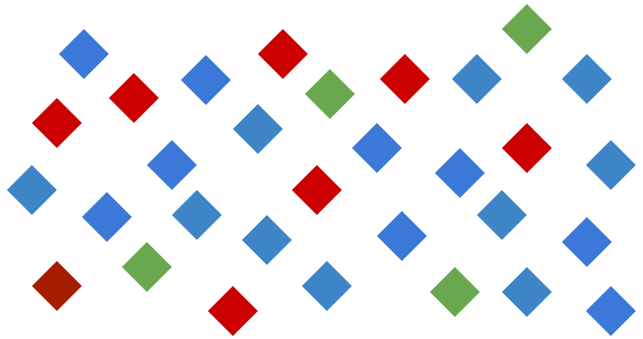
How many copies of each transcript were in my original sample?

We need to maximize the probability that our generative model and parameters explain our observations

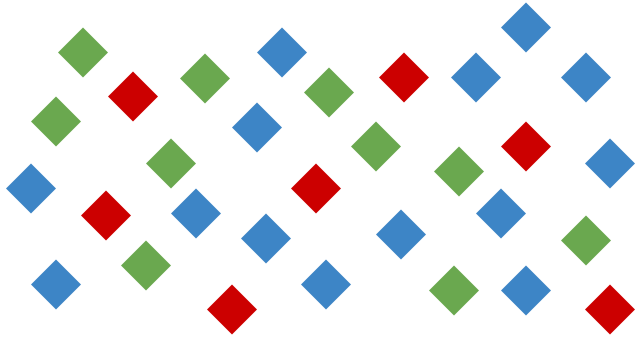
1. Estimate transcript abundance



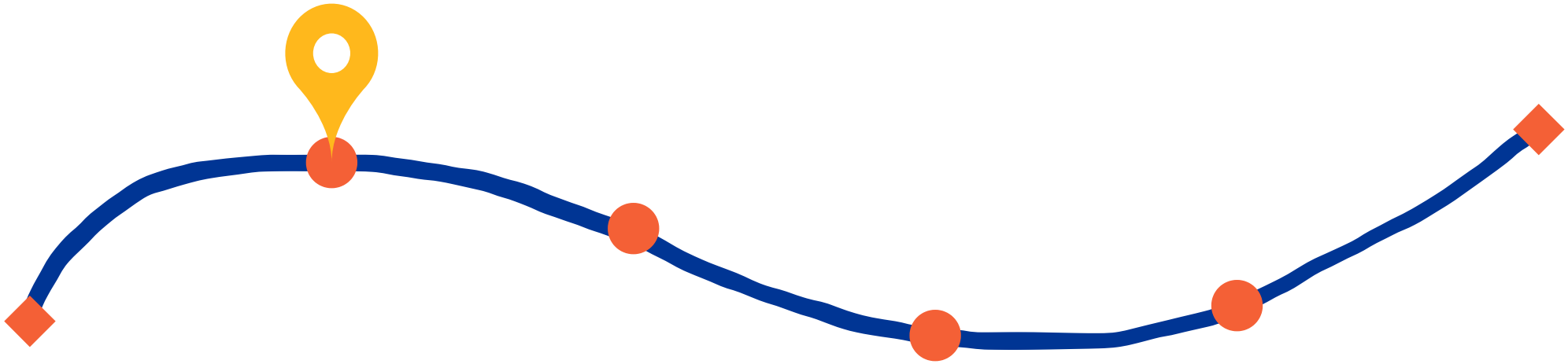
2. Randomly sample *n* fragments



We iteratively optimize our transcript abundances until our generated reads look very similar to our observed reads



After today, you should have a better understanding of



Generative models for RNA quantification

Salmon's mathematical definition of a transcriptome

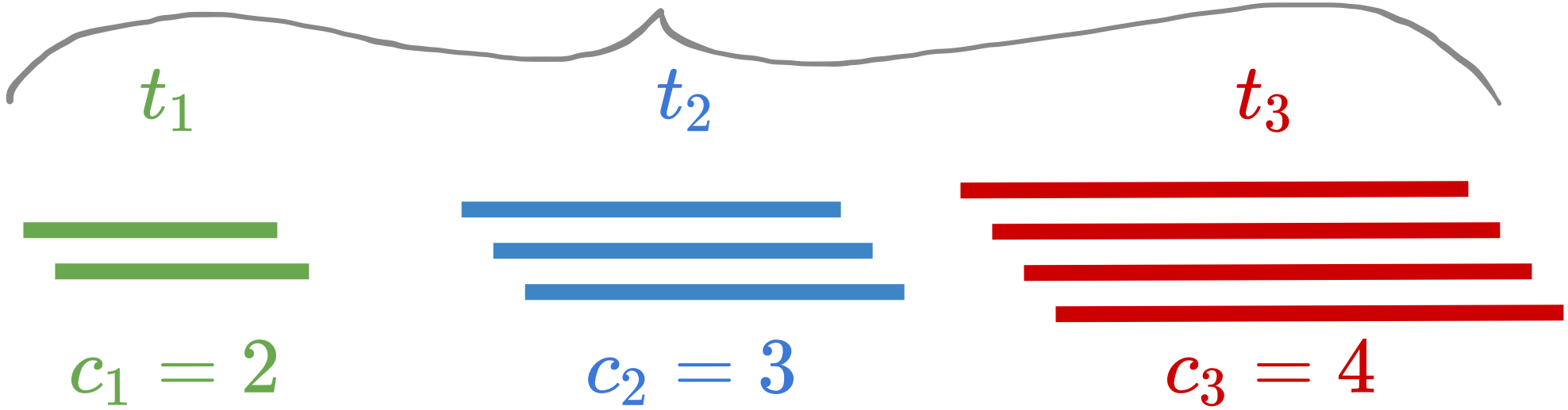
Individual transcripts

Transcript counts

Our whole transcriptome

$$T = \{(t_1, \ldots, t_M), (c_1, \ldots, c_M)\}$$

T $M = 3$

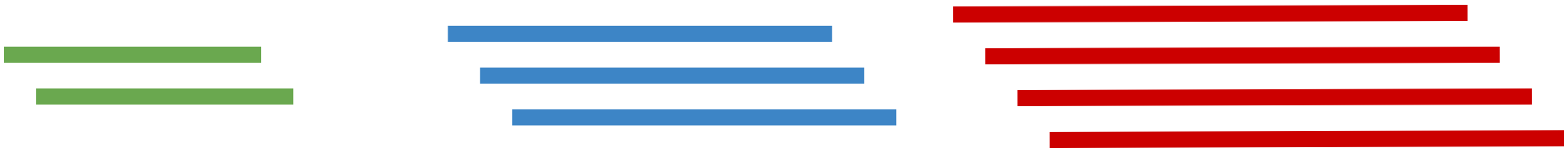


Salmon's formulation of **transcript abundance**

$c_1 = 2$
$\tilde{l}_1 = 200$

$c_2 = 3$
$\tilde{l}_2 = 300$

$c_3 = 4$
$\tilde{l}_3 = 400$



So far, we have been talking about transcript fractions

$$f_i = \frac{c_i}{\sum_j^M c_j}$$

$$\eta_i = \frac{c_i \tilde{l}_i}{\sum_j^M c_j \tilde{l}_j} \qquad \eta = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \eta_3 \end{bmatrix}$$

We can also take nucleotide fractions by taking into account the effective length of each transcript

This tells us how much of the total RNA pool comes from each transcript

I will explain the effective length later. For now, think of it as a "corrected" length

Converting to relative abundances

τ_i The transcript fraction normalizes nucleotide fraction by the effective length

$$\tau_i = \frac{\frac{\eta_i}{\tilde{l}_i}}{\sum_{j=1}^{M} \frac{\eta_j}{\tilde{l}_j}}$$

Adjusts for the fact that longer transcripts generate more reads

This gives the relative abundance of each transcript *i*

$$\mathrm{TPM}_i = \tau_i \cdot 10^6$$

The **transcript fraction** tells us the proportion of total RNA molecules in the sample that come from transcript *i*

TPM is "Transcripts per million"

Transcript-Fragment Assignment Matrix

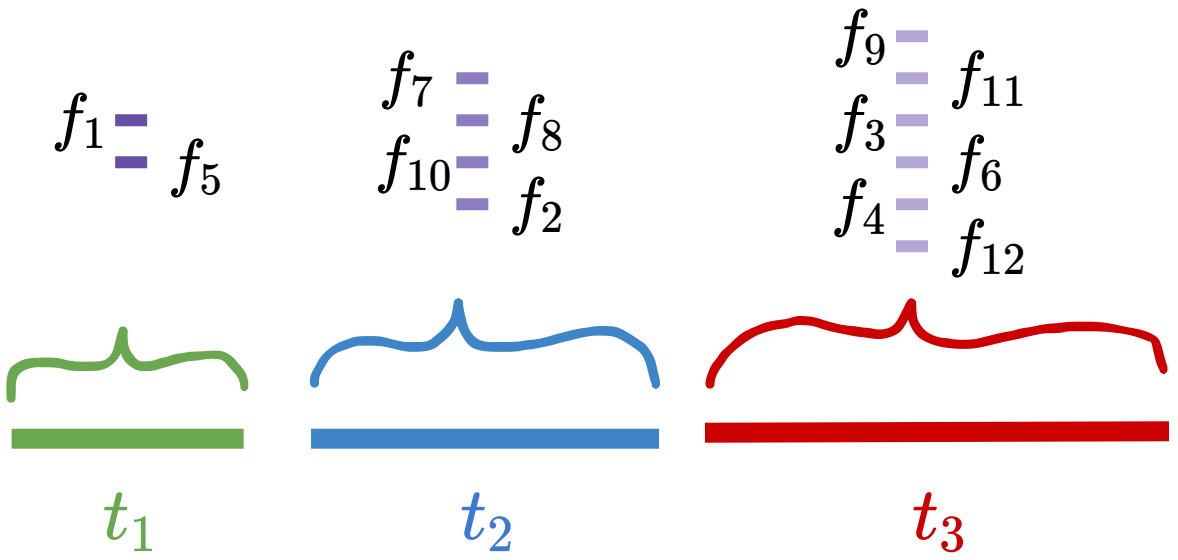
Z is a binary matrix (i.e., all values are 0 or 1)
of *M* transcripts (rows) and *N* fragments (columns)

$$Z = \begin{array}{c} \begin{matrix} \textbf{Fragment 1} & \textbf{Fragment 2} & \textbf{...} & \textbf{Fragment N} \end{matrix} \\ \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & \dots & Z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{M1} & Z_{M2} & \dots & Z_{MN} \end{bmatrix} \end{array} \begin{array}{l} \textbf{Transcript 1} \\ \textbf{Transcript 2} \\ \\ \textbf{...} \\ \\ \textbf{Transcript M} \end{array}$$

$Z_{i,j} = 1$ if fragment *j* is assigned to transcript *i*

Z example

Suppose we have 3 transcripts and 12 fragments

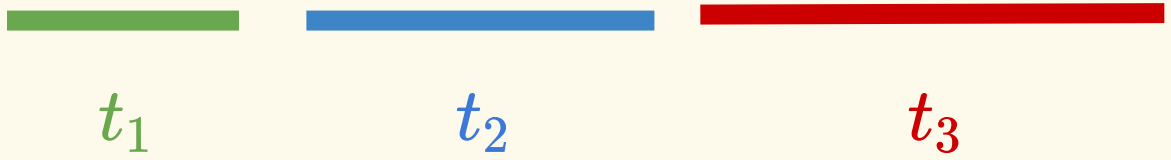


Z is just how we computationally assign fragments to transcripts

$$Z = \begin{array}{c} \begin{matrix} f_1 & f_2 & f_3 & f_4 & f_5 & f_6 & f_7 & f_8 & f_9 & f_{10} & f_{11} & f_{12} \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \end{array} \begin{matrix} t_1 \\ t_2 \\ t_3 \end{matrix}$$

Generative model inference

Known from organism and experiment



Given these inputs, generate a distribution of fragments

Transcript-fragment assignment

$$Z = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & \dots & Z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{M1} & Z_{M2} & \dots & Z_{MN} \end{bmatrix}$$

Transcript abundance

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_M \end{bmatrix}$$

N and *M* are same as experiment

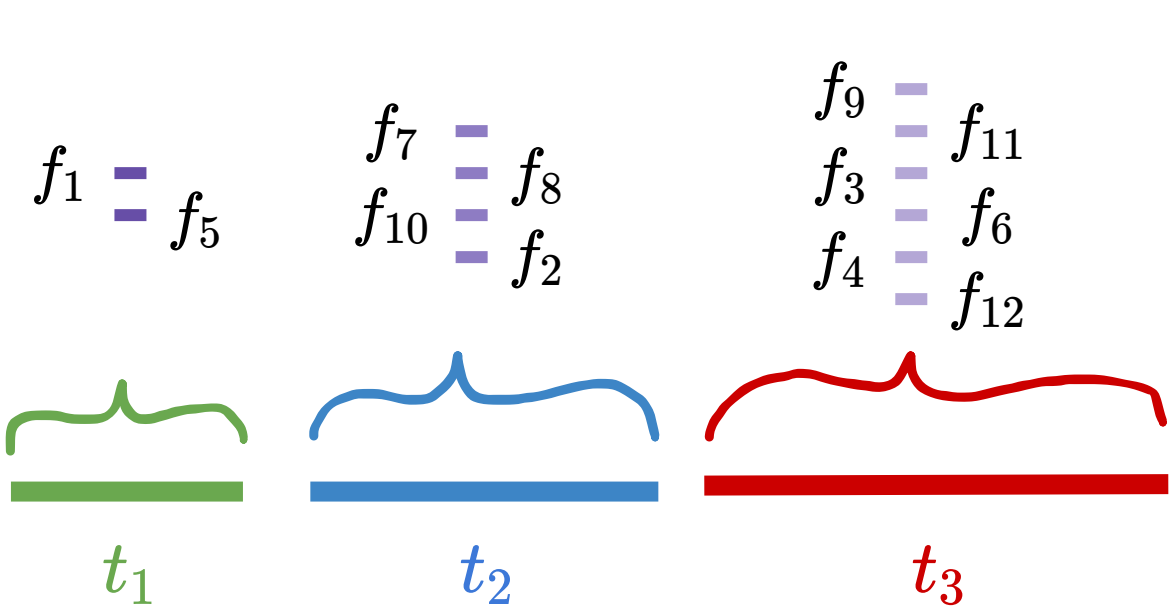
Run 1

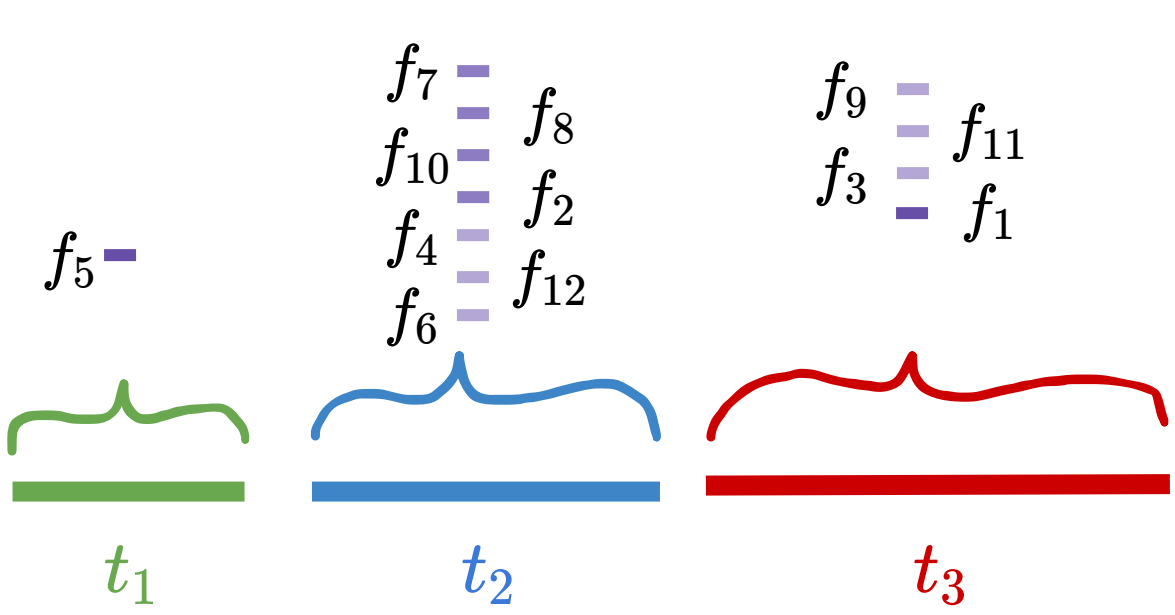
Run 2

Probability of observing the sequence fragments

Which scenario is more likely, given our generative model?

We can use probabilistic methods to find
parameters that explain our observed distirbution

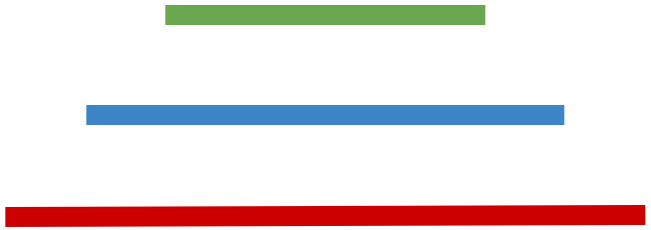




Probability of observing the sequenced fragments

$$P(F|T,\eta,Z)$$

Available transcripts



Transcript-fragment assignment

$$Z = \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1N} \\ Z_{21} & Z_{22} & \dots & Z_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{M1} & Z_{M2} & \dots & Z_{MN} \end{bmatrix}$$

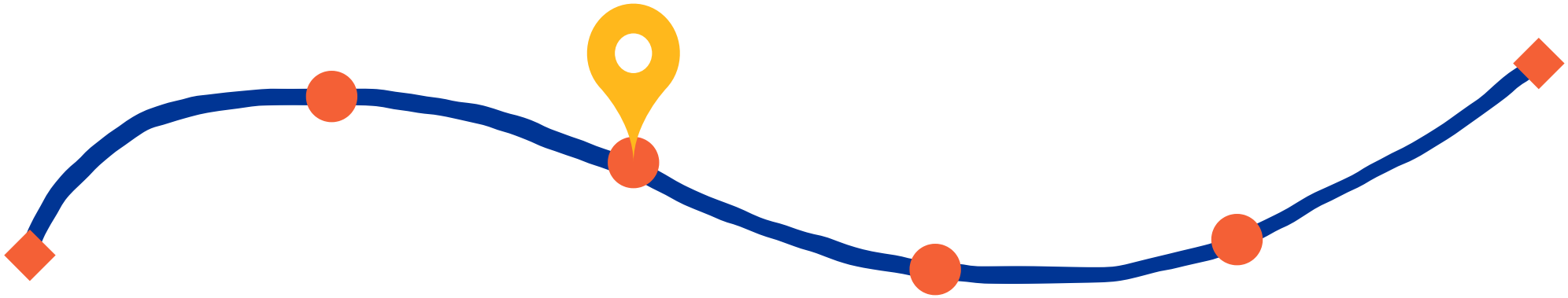
Transcript abundance

$$\eta = \begin{bmatrix} \eta_1 \\ \vdots \\ \eta_M \end{bmatrix}$$

Given these **parameters**, how probable is it that our experiment generated these observed reads?

Optimize these values until we get the highest probability

After today, you should have a better understanding of



Probability optimization instead of generation

Probability of observing the sequenced fragments

We can now compute the probability of observing: Set of fragments F

Given:

Transcriptome T

Transcript assignment Z Transcript abundance η

$$P(F|\eta, Z, T) = \prod_{j=1}^{N} \sum_{i=1}^{M} \eta_i P(f_j|t_i)$$

$P(f_j|t_i)$

Probability of observing fragment f_j

given that it comes from transcript t_i

This expression accounts for all possible transcripts a fragment might come from, weighted by how likely that fragment is to come from each transcript

Fragment probabilities

$P(f_j|t_i)$ is a conditional probability that depends on the **position** of the fragment within the transcript, the **length** of the fragment, and any technical biases

In Salmon's quasi-mapping approach, this probability is approximated based on transcript compatibility rather than exact positions.

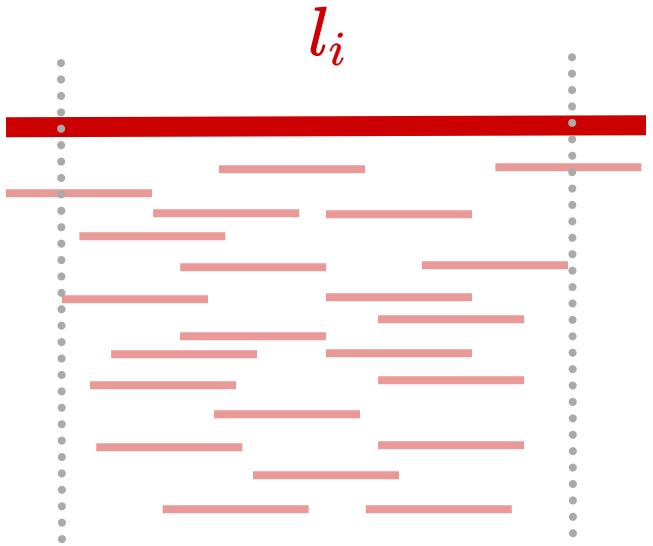
$$P(f_j|t_i) = P(\text{fragment length}, \text{position}, \text{GC content}, \ldots)$$

Positional bias

Fragments that include transcript ends might be too short

Fragments from central regions are more likely to be of optimal length for sequencing reads

A transcript's **effective length** adjusts for the fact that fragments near the ends of a transcript are less likely to be sampled

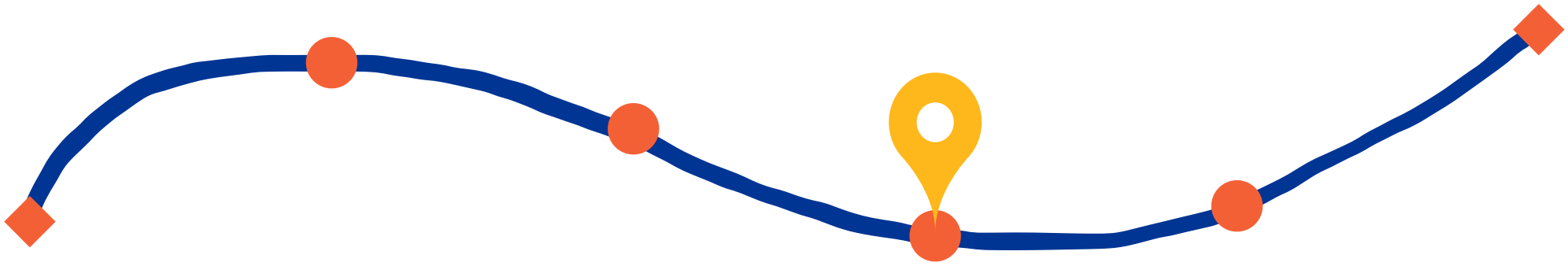


$$\tilde{l}_i = l_i - \mu_i \qquad \tilde{l}_i < l_i$$

μ_i **Mean of the truncated empirical fragment length distribution**

$$\eta_i = \frac{c_i \tilde{l}_i}{\sum_i c_i \tilde{l}_i}$$

After today, you should have a better understanding of



Probability maximization with inference

Two-phase inference in salmon

Inference refers to the process of estimating transcript abundances from observed RNA-seq reads using statistical models.

Salmon processes reads in **two stages**

Online phase

Makes fast, initial estimates of transcript abundances as the reads are processed

Offline phase

Refines these initial estimates using more complex optimization techniques

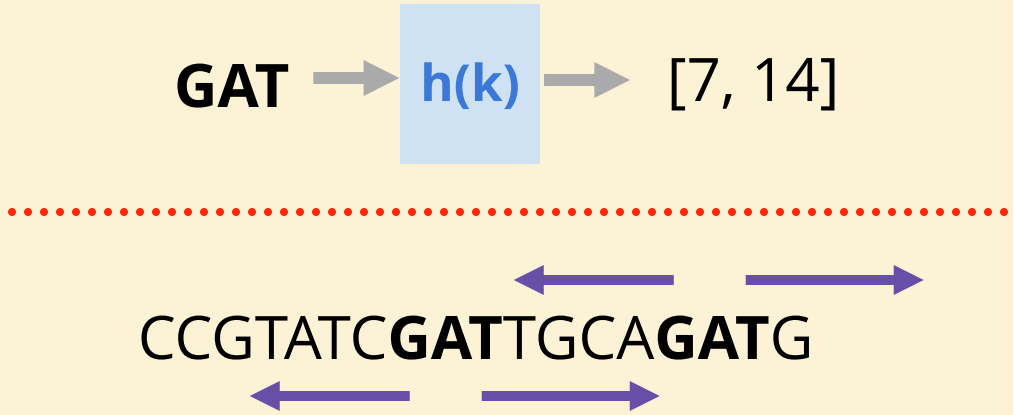
This two-phase approach balances **speed** (in the online phase) with **accuracy** (in the offline phase)

Online phase:
Stochastic variational inference

Initial estimates using quasi-mapping

Quasi-mapping is A fast, lightweight technique used to associate RNA-seq fragments with possible transcripts

Read mapping



Identify seeds, then extend and compute base-by-base alignment

Essentially early stopping of read mapping

Alignment is expensive, so quasi-mapping stops after identify seeds

This is what initializes compatible transcripts and abundance

$$\eta_t \approx \frac{\text{Number of fragments mappting to } t}{\text{Total number of fragments}}$$

Iteratively update parameters based on mini batches

Take current parameters

Mini-batch 1

Compute derivatives from batch

Update parameters (i.e., abundances)

Repeat for each batch

Mini-batch 2

Mini-batch 3

Offline Phase:
Expectation-Maximization (EM) algorithm

Offline phase fine tunes transcript abundance

After the online phase, Salmon refines the estimates using a more complex optimization method, typically based on the **Expectation-Maximization (EM) algorithm**

This phase ensures the accuracy of abundance estimates, incorporating the bias corrections learned during the online phase

Likelihood of the Data

The **likelihood** function is central to the inference process in Salmon:

$$\mathcal{L}\{\alpha|F,Z,T\} = \prod_{j=i}^{N}\sum_{i=1}^{M}\hat{\eta}_i Pr\{f_j|t_i\}$$

This is the probability of observing the entire set of fragments F, given the transcriptome T and nucleotide fractions η

Optimize the estimates of α, a vector of the estimated number of reads originating from each transcript

$$\hat{\eta}_i = \frac{\alpha_i}{\sum_j \alpha_j}$$

The goal is to **maximize this likelihood** to infer the most likely values of η, which correspond to the relative abundances of the transcripts

Maximum Likelihood Estimation (MLE)

The goal of **maximum likelihood** is to find the parameters (transcript abundances) that **maximize the probability** of the observed data (sequenced reads)

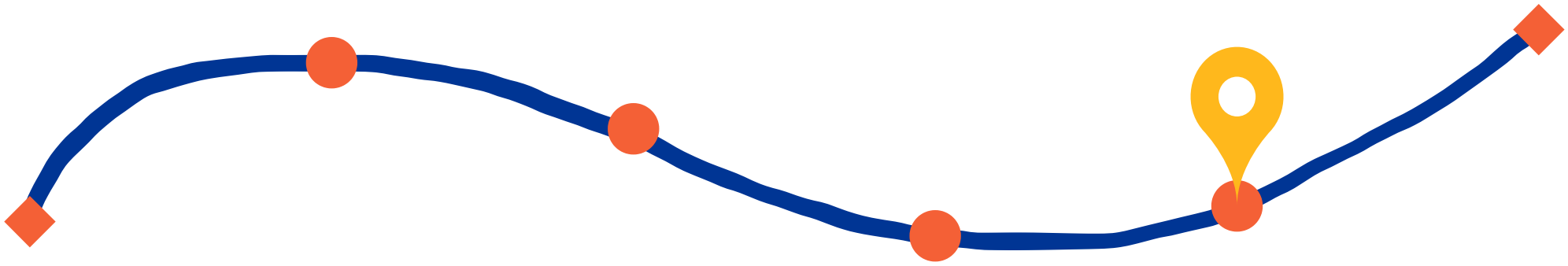
The **likelihood** function is central to the inference process in Salmon:

$$\mathcal{L}\{\alpha|F,Z,T\} = \prod_{j=i}^{N}\sum_{i=1}^{M}\hat{\eta}_i Pr\{f_j|t_i\}$$

Optimize the estimates of α, a vector of the estimated number of reads originating from each transcript

Given α, η can be directly computed.

After today, you should have a better understanding of



Methodology with Python a implementation

Before the next class, you should

Lecture 07B:

Quantification - Methodology

Lecture 08A:

Differential gene expression - Foundations

Today

Tuesday

- Work on P02A (due Mar 14)