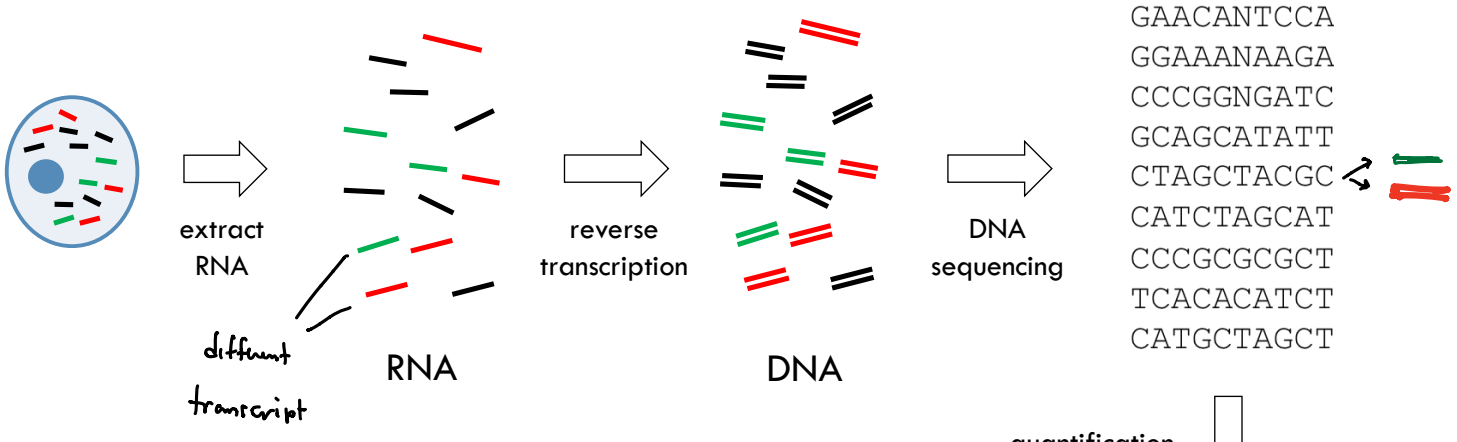
Lecture 7: Expectation-Maximization (EM) Algorithm

ECE 365

Announcements:

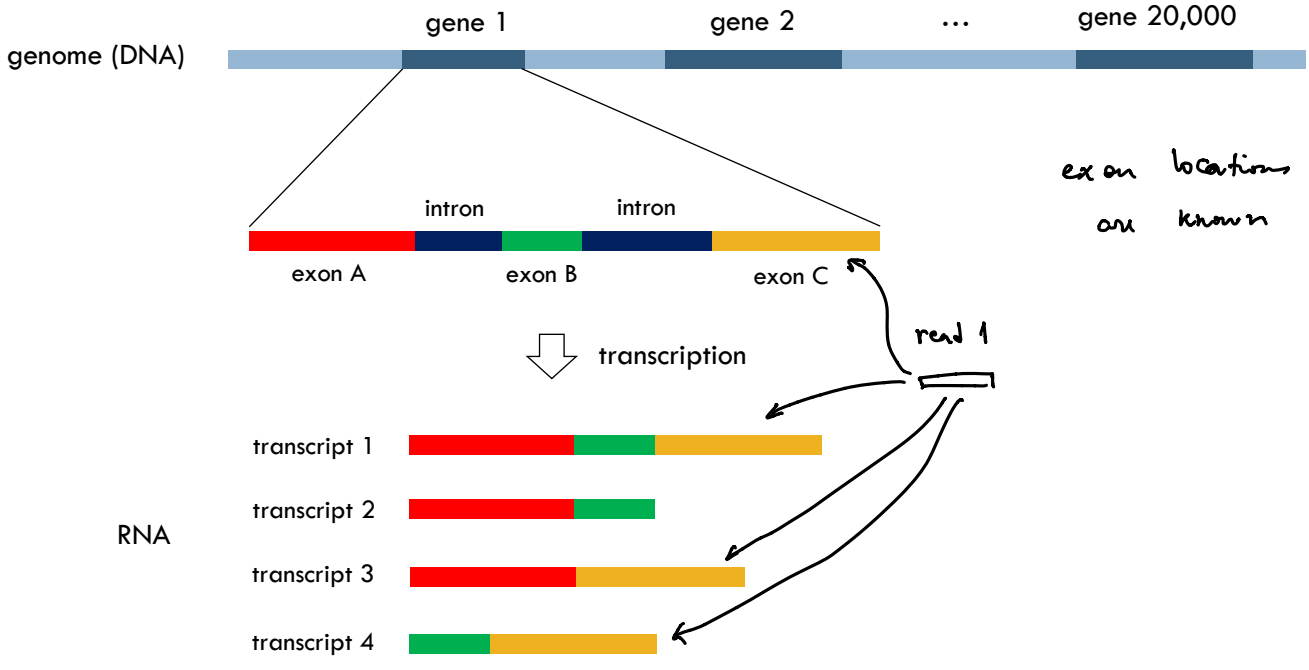
- Lab 3 due on Friday
- Lab 4 released tonight (due April 2)

From previous lecture: RNA quantification



Why is this hard? Because transcripts may "look alike"

From genes to transcripts



General RNA quantification problem

- Input: **aligned read data**

relative abundances

	tr. 1	tr 2	⋯		tr. K
read 1	0	1	1	0	0	0
read 2	0	0	0	1	0	0
	0	1	1	0	0	0
⋮	0	0	0	0	1	0
	1	0	0	0	0	0
	1	0	0	0	1	0
read N	0	1	0	0	0	1

means: read N maps to tr. 2

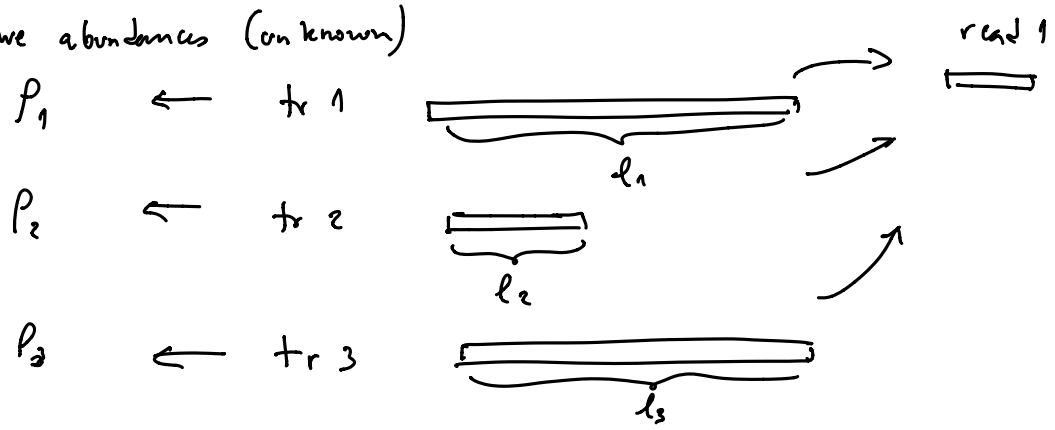
$\Rightarrow$ $\hat{\rho}_1 = 0.01$, $\hat{\rho}_2 = 0.04$, ⋮, $\hat{\rho}_K = 0.02$ } add up to 1

General approach to solve this:

Expectation - Maximization Algorithm (EM)

Model for read data generation:

relative abundances (unknown)



$$\theta_k = P(\text{read comes from tr. } k) = \frac{\rho_k \ell_k}{\sum_{j=1}^{K} \rho_j \ell_j}$$

True read origin/assignment (unknown)

$$Z_{ik} = \begin{cases} 1 & \text{if read } i \text{ comes from transcript } k \\ 0 & \text{otherwise} \end{cases}$$

form a matrix

K transcripts

N reads

$$\begin{bmatrix} 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

(one 1 per row)

↘ "cleaned up" version of data matrix

EM strategy: Estimate $\hat{\rho}_k$ and $\hat{Z}_{ik}$ in an alternating fashion and recursively

Initialize : $\hat{\rho}_k = \frac{1}{K}$ for $k = 1, \ldots, K$

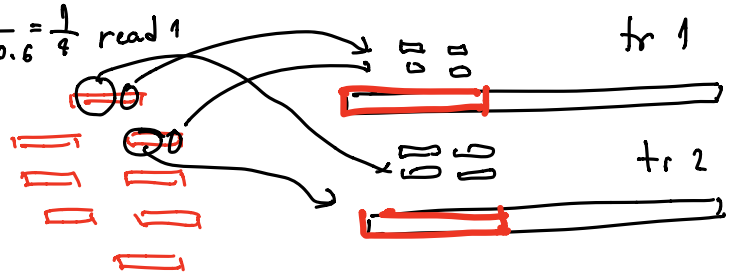
(*) Based on $\hat{\rho}_k$'s, estimate $\hat{z}_{ik}$'s :

in the middle of the algorithm

$\hat{z}_{11} = \frac{0.2}{0.2+0.6} = \frac{1}{4}$ read 1

$\hat{z}_{12} = \frac{3}{4}$

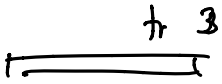
$\hat{z}_{13} = 0$



$\hat{\rho}_1 = 0.2$

$\hat{\rho}_2 = 0.6$

$S_1 = \{1, 2\}$



$\hat{\rho}_3 = 0.2$

Let S_i be the set of transcripts read i maps to

$$\hat{z}_{ik} = \begin{cases} \dfrac{\hat{\rho}_k}{\sum_{j \in S_i} \hat{\rho}_j} & \text{if } k \in S_i \\ 0 & \text{if } k \notin S_i \end{cases}$$

E-step

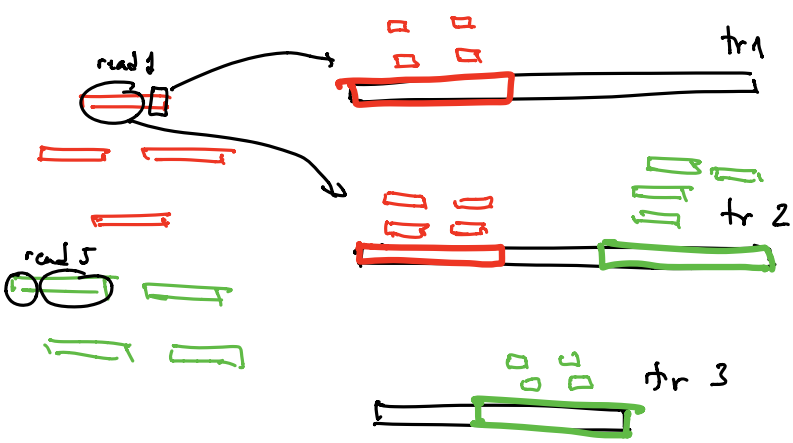
Based on $\hat{z}_{ik}$, we want to update $\hat{\rho}_k$'s :

$\hat{z}_{11} = 0.2$

$\hat{z}_{12} = 0.8$

$\hat{z}_{52} = 0.6$

$\hat{z}_{53} = 0.9$



$\hat{\theta}_1 = \frac{0.2+0.2+0.2+0.2}{8} = 0.1$

$\hat{\theta}_2 = \frac{4 \times 0.8 + 4 \times 0.6}{8}$

$\hat{\theta}_3 = \frac{4 \times 0.9}{8}$

First, estimate probability θ_k :

$$\hat{\theta}_k = \frac{1}{N} \sum_{i=1}^{N} \hat{z}_{ik} \quad , \qquad \theta_k \propto \rho_k \ell_k$$

$$\hat{\rho}_k = \frac{\frac{\hat{\theta}_k}{\ell_k}}{\sum_{j=1}^{K} \frac{\hat{\theta}_j}{\ell_j}}$$

M-step

Repeat. Go back to ⊛

EM algorithm for RNA quantification

- Initialize: $\hat{\rho}_k^{(1)} = \frac{1}{K}$
- For $t = 1, 2, \ldots,$

E-step

- $$\hat{Z}_{ik}^{(t)} = \begin{cases} \frac{\hat{\rho}_k^{(t)}}{\sum_{j \in S_i} \hat{\rho}_j^{(t)}} & \text{for } k \in S_i \\ 0 & \text{otherwise} \end{cases}$$

M-step

- $$\hat{\rho}_k^{(t+1)} = \frac{\frac{\theta_k^{(t+1)}}{\ell_k}}{\sum_{j=1}^{K} \frac{\theta_j^{(t+1)}}{\ell_j}} \quad \text{where} \quad \theta_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \hat{Z}_{ik}^{(t)}$$

EM algorithm for RNA quantification

- Does it converge? Yes.
- If so, does it converge to the correct values?

For RNA quant. problem, $\hat{\rho}_1, \ldots, \hat{\rho}_K$ converge to the maximum likelihood parameters.

- Let's look at a concrete example in Excel