Transcript Abundance Estimation Using Expectation-Maximization Algorithm

Mir Imtiaz Mostafiz (0417052041)
S.M.Farabi Mahmud (0417052068)

Introduction

Transcript Abundance Estimation is a problem which finds relative abundances of different RNA-transcripts from a set of reads, where the reads themselves were generated from those transcripts.
Expectation-Maximization Algorithm is a systematic way to deduce the unknown parameters of a system, from some known observations and some known parameters of a system, through multiple iterations of expectations and maximizations.
In this note, a bridge will be constructed between the concepts stated above.

Prerequisites

To understand this topic, you need to have some basic knowledge about:

- **Basic of Probability and Distributions**: Different kinds of distributions and their mean, variance, likelihood; Sampling, Events.
- **Central Dogma of Genetics**: Interactive relations between DNAs, RNAs and Proteins.
- **Some Advanced Concepts of Computational Biology**: Short Read Mapping, Alternative Splicing, *-seq, RNA-seq etc.

1 Expectation-Maximization Algorithm

1.1 Prelude

Every system, or experimental setup, consists of three basic parts: system parameters, event observations and unknown quantities(observations/parameters). In ideal and most favorable situations, we usually use the parameters and observations to deduce the unknowns. A such experimental setup , credited to [2], is described below.

Let's assume, there are two coins, A and B. The probability that the head side is up after tossing coin A is θ_A (so the probability that the tail side is up after tossing coin A is $1 - \theta_A$). Similarly, the probability that the head side is up after tossing coin B is θ_B (so the probability that the tail side is up after tossing coin B is $1 - \theta_B$). The probability of picking a coin is uniformly distributed, i.e. probability of picking coin A is $P(A) = 0.50$ and so is the probability of picking coin B, $P(B)$. To determine θ_A and θ_B, we will follow this procedure:

- **Step 1**: Pick a coin (A or B).
- **Step 2**: Toss it 10 times and tally the head/tail events.
- **Step 3**: Run the steps 1 & 2 for 5 times.
- **Step 4**: Determine θ_A and θ_B from the tallied event counts using the law of maximum-log-likelihood.

So, the system can be described like this in table 1 (the observations are shown as an example):

We know that the population mean of an observation set sampled from Bernouli distribution is simply their average. From table 1, the tallies of coin flips can be shown in table 2.

So, as per **Maximum Likelihood Estimation (MLE)** rules, ${\theta_A}' = \frac{\text{Heads from A}}{\text{Heads from A + Tails from A}} = \frac{24}{24+6} = 0.8$ and ${\theta_B}' = \frac{\text{Heads from B}}{\text{Heads from B + Tails from B}} = \frac{9}{9+11} = 0.45$.
So, given the parameters and observations of a system, we could deduce some of the unknown parameters using MLE. But what, if, some of the necessary parameters, or some of the necessary observation details(like from which coins the observations were generated) are missing? **Expectation-Maximization Algorithm** is the answer.

Table 1: Experimental Setup (Example)

Known Parameters	$P(A) = 0.5$
	$P(B) = 0.5$
Observed Events	Coin B: HTTTHHTHTH
	Coin A: HHHHTHHHHH
	Coin A: HTHHHHHTHH
	Coin B: HTHTTTHHTT
	Coin A: THHHHHTHTH
Unknown Quantities	θ_A and θ_B

Table 2: Coin Flip Tallies (Example)

Observations	Coin A	Coin B
HTTTHHTHTH		5H,5T
HHHHTHHHHH	9H,1T	
HTHHHHHTHH	8H,2T	
HTHTTTHHTT		4H,6T
THHHHHTHTH	7H,3T	
Total	24H,6T	9H,11T

1.2 Expectation Maximization Algorithm Explanation

To explain **Expectation Maximization Algorithm** (EM), we will manipulate the experimental setup described previously. We will assume that the information about which coins were tossed when will not be present.
At first, let's brief the basic steps of EM in short:

- Step 1: Assume the unknown parameters.
- Step 2: Using the current parameters, both known and hidden (assumed), deduce the expected value of different quantities needed for unknown parameter computation.
- Step 3: Compute the new values of unknown parameters.
- Step 4: Repeat steps 2 & 3 until convergence.

Now, we will explain this process using the converted example. Table 3 shows the current setup, without mentioning the origin of observations (which coins generated which tosses). As we have to determine the value of unknown parameters θ_A and θ_B, we will assume some values of them, i.e. ${\theta_A}^{(0)} = 0.60$ and ${\theta_B}^{(0)} = 0.50$ initially. Also, as the origins are not mentioned, we cannot compute the heads and tails in each sets of observations straightforward. Rather, we need to compute the expected number of heads and tells for each type of coins. This expected value calculation is explained below:

Let n tosses of coin X generated events $x_1, x_2, \ldots, x_n$, where each $x_i \in \{H, T\}$ and $X \in \{A, B\}$. The posterior probability of X being A can be calculated from priors and likelihoods as,

$$P(A|x_1, x_2, \ldots, x_n) = \frac{P(x_1, x_2, \ldots, x_n|A)P(A)}{P(x_1, x_2, \ldots, x_n)}$$

$$\implies P(A|x_1, x_2, \ldots, x_n) = \frac{P(x_1|A)P(x_2|A)\ldots P(x_n|A)P(A)}{P(x_1, x_2, \ldots, x_n)} \quad \text{[Because the coin tosses are independent]}$$

Similarly,

$$P(B|x_1, x_2, \ldots, x_n) = \frac{P(x_1|B)P(x_2|B)\ldots P(x_n|B)P(B)}{P(x_1, x_2, \ldots, x_n)}$$

.

Combining and dividing both equations, we get new equation,

$$\frac{P(A|x_1, x_2, \ldots, x_n)}{P(B|x_1, x_2, \ldots, x_n)} = \frac{P(x_1|A)P(x_2|A)\ldots P(x_n|A)P(A)}{P(x_1|B)P(x_2|B)\ldots P(x_n|B)P(B)}$$

$$\implies \frac{P(A|x_1, x_2, \ldots, x_n)}{P(B|x_1, x_2, \ldots, x_n)} = \frac{P(x_1|A)P(x_2|A)\ldots P(x_n|A)}{P(x_1|B)P(x_2|B)\ldots P(x_n|B)} \quad \text{[Because } P(A) = P(B) = 0.5\text{]}$$

If $x_i = H$, then $P(x_i|A) = \theta_A$ and if $x_i = T$, then $P(x_i|A) = 1 - \theta_A$. Similarly, if $x_i = H$, then $P(x_i|B) = \theta_B$ and if $x_i = T$, then $P(x_i|B) = 1 - \theta_B$.
If there are k heads, then there will be $n - k$ tails, and the equation will look like,

$$\frac{P(A|x_1, x_2, \ldots, x_n)}{P(B|x_1, x_2, \ldots, x_n)} = \frac{{\theta_A}^k(1 - \theta_A)^{n-k}}{{\theta_B}^k(1 - \theta_B)^{n-k}}$$

Table 3: Experimental Setup 2 (Example)

Known Parameters	$P(A) = 0.5$
	$P(B) = 0.5$
Observed Events	HTTTHHTHTH
	HHHHTHHHHH
	HTHHHHHTHH
	HTHTTTHHTT
	THHHHHTHTH
Unknown Quantities	θ_A and θ_B
Assumed Quantities	${\theta_A}^{(0)} = 0.60$ and ${\theta_B}^{(0)} = 0.50$

$$\implies \frac{P(A|x_1, x_2, \ldots, x_n)}{P(A|x_1, x_2, \ldots, x_n) + P(B|x_1, x_2, \ldots, x_n)} = \frac{{\theta_A}^k(1-\theta_A)^{n-k}}{{\theta_A}^k(1-\theta_A)^{n-k} + {\theta_B}^k(1-\theta_B)^{n-k}}$$

$$\implies P(A|x_1, x_2, \ldots, x_n) = \frac{{\theta_A}^k(1-\theta_A)^{n-k}}{{\theta_A}^k(1-\theta_A)^{n-k} + {\theta_B}^k(1-\theta_B)^{n-k}}$$

Similarly,

$$P(B|x_1, x_2, \ldots, x_n) = \frac{{\theta_B}^k(1-\theta_B)^{n-k}}{{\theta_A}^k(1-\theta_A)^{n-k} + {\theta_B}^k(1-\theta_B)^{n-k}}$$

.

Now, the expected number of H and T from $\{x_1, x_2, \ldots, x_n\}$ from type A if there are k H's and $n-k$ T's are respectively,

$$|H|^A_{expected} = \text{Probability that the observations are from A} \times \text{number of H}$$

$$\implies |H|^A_{expected} = P(A|x_1, x_2, \ldots, x_n) \times k$$

$$\implies |H|^A_{expected} = \frac{{\theta_A}^{(0)^k}(1-{\theta_A}^{(0)})^{n-k}}{{\theta_A}^{(0)^k}(1-{\theta_A}^{(0)})^{n-k} + {\theta_B}^{(0)^k}(1-{\theta_B}^{(0)})^{n-k}} \times k$$

For first set of observations,

$$\implies |H|^{A^{(1)}}_{expected} = \frac{(0.6)^5(1-0.6)^{10-5}}{(0.6)^5(1-0.6)^{10-5} + (0.5)^5(1-0.5)^{10-5}} \times 5$$

$$\implies |H|^{A^{(1)}}_{expected} = 0.45 * 5 = 2.25$$

Similarly, for first observations, we can deduce,

$$|T|^{A^{(1)}}_{expected} = 0.45 * 5 = 2.25$$

$$|H|^{B^{(1)}}_{expected} = 0.55 * 5 = 2.75$$

$$|T|^{B^{(1)}}_{expected} = 0.55 * 5 = 2.75$$

For each set of observations, we deduce this quantities and calculate the estimated MLE from table 4,

Table 4: MLE Computation (Example)

Observations	Coin A	Coin B
HTTTHHTHTH	2.25 H, 2.25 T	2.75 H, 2.75 T
HHHHTHHHHH	7.2 H, 0.8 T	1.8 H, 0.2 t
HTHHHHHTHH	5.9 H, 1.5 T	2.1 H, 0.5 T
HTHTTTHHTT	1.4 H, 2.1 T	2.6 H, 3.9 T
THHHHHTHTH	4.5 H, 1.9 T	2.5 H, 1.1 T
Total	21.3 H, 8.6 T	11.7H, 8.4 T

$${\theta_A}^{(1)} = \frac{|H|^{A^{(total)}}_{expected}}{|H|^{A^{(total)}}_{expected} + |T|^{A^{(total)}}_{expected}} = \frac{21.3}{21.3 + 8.6} = 0.71$$

$${\theta_B}^{(1)} = \frac{|H|^{B^{(total)}}_{expected}}{|H|^{B^{(total)}}_{expected} + |T|^{B^{(total)}}_{expected}} = \frac{11.7}{11.7 + 8.4} = 0.58$$

So, starting with assuming values 0.60 and 0.50, we ended with 0.71 and 0.58 after first iteration. The values will converge after some iterations to 0.80 and 0.52 respectively.

The step where we calculated the expected numbers of heads and tails is called the **Expectation** step. The step where we calculated the new estimated value of MLE is known as the **maximization** step. We repeat these two steps one after another in EM algorithm.

We can summarize the uses and steps of EM algorithm in table 5.

Table 5: When's and how's of EM

When to apply	How to apply
Some parts of the data is hidden	Initially guess the parameters of the model
If we know the parameters to be estimated, we can calculate probabilities of the hidden variables	**Expectation (E) step** : Use current parameters (and observations) to reconstruct the hidden structure
If we know the values of hidden variables, we can estimate the parameters	**Maximization (M) step** : Use that hidden structure (and observations) to re-estimate parameters

2 Transcript Abundance Estimation

Before getting a detailed view about **Transcript Abundance Estimation (TAE)**, we will revisit the *Central Dogma* of genetics: a workflow assumption based on which the whole genetics is based.

2.1 Central Dogma

Statement: ***Instructions on DNA are transcribed onto messenger RNA. Ribosomes are able to read the genetic information inscribed on a strand of messenger RNA and use this information to string amino acids together into a protein.*** [1].
Explanation: The central dogma of molecular biology describes the flow of genetic information in cells from DNA to messenger RNA (mRNA) to protein. It states that genes specify the sequence of mRNA molecules, which in turn specify the sequence of proteins. In technical terms, the central dogma consists of two steps:

Transcription The information stored in DNA is so central to cellular function, the cell keeps the DNA protected and copies it in the form of RNA. An enzyme adds one nucleotide to the mRNA strand for every nucleotide it reads in the DNA strand [1].

Translation The translation of information to a protein is more complex because three mRNA nucleotides correspond to one amino acid in the polypeptide sequence [1].

2.2 Some related terms to know

2.2.1 DNA sequencing

DNA sequencing is the process of determining the precise order of nucleotides within a DNA molecule. It includes any method or technology that is used to determine the order of the four basesadenine, guanine, cytosine, and thyminein a strand of DNA [9].

2.2.2 Transcript

A primary transcript is the single-stranded ribonucleic acid (RNA) product synthesized by transcription of DNA, and processed to yield various mature RNA products such as mRNAs, tRNAs, and rRNAs. The primary transcripts designated to be mRNAs are modified in preparation for translation. For example, a precursor messenger RNA (pre-mRNA) is a type of primary transcript that becomes a messenger RNA (mRNA) after processing [10].

2.2.3 Alternative Splicing

Alternative splicing, or differential splicing, is a regulated process during gene expression that results in a single gene coding for multiple proteins. In this process, particular exons of a gene may be included within or excluded from the final, processed messenger RNA (mRNA) produced from that gene [6] (see fig 3).

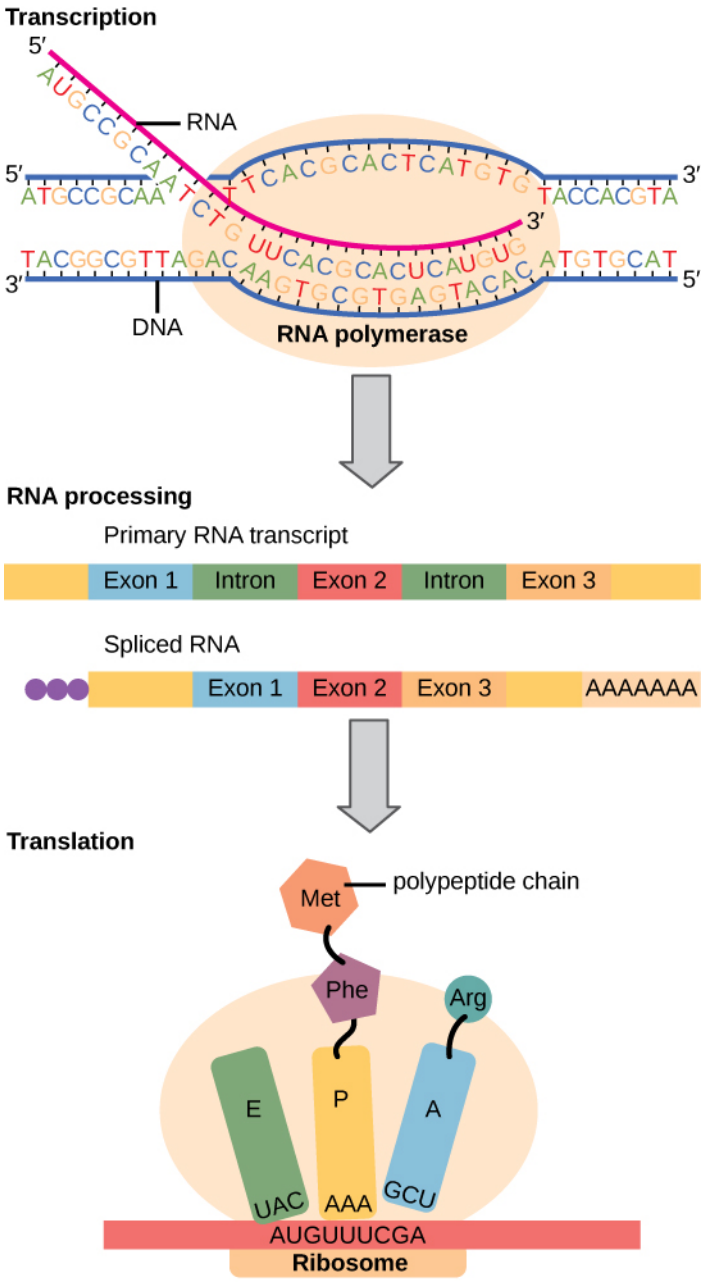

Figure 1: Central Dogma [1]

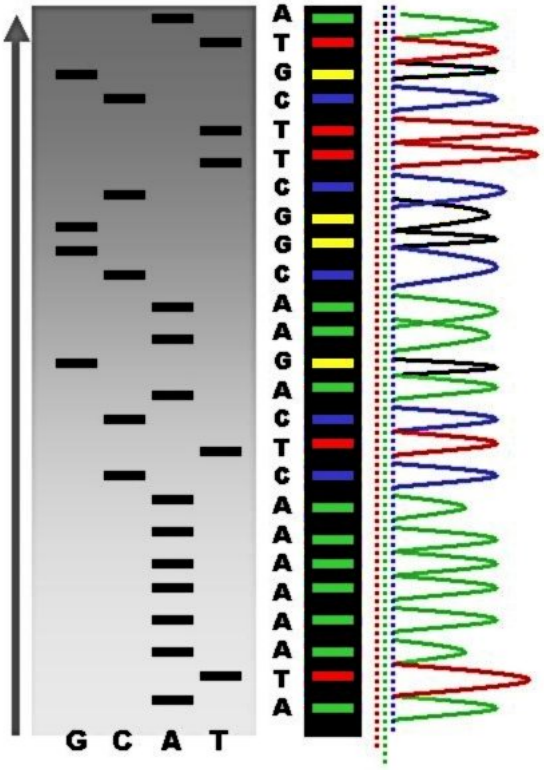

Figure 2: An example of the results of automated chain-termination DNA sequencing [9].

2.2.4 Protein Isoform

A protein isoform, or "protein variant"[1] is a member of a set of highly similar proteins that perform the same or similar biological roles. A set of protein isoforms may be formed from alternative splicings or other post-translational modifications of a single gene [7] (see figure 3).

2.2.5 Short Read Alignment

Short read alignment is the process of figuring out where in the genome a sequence is from. By transcription process we get the mRNAs from DNA. By putting them through a reverse transcription and sequencing pipeline, we can get the sequences corresponding to the RNA transcript. A huge task after this is finding the place in the genome from where that gene sequence generated. This is done by short read alignment algorithms.

2.2.6 *-seq

*-seq is a generalized term used to denote the class of high-throughput selection algorithms. The full list of *-seq algorithms can be found at `https://liorpachter.wordpress.com/seq/`.

2.2.7 RNA-seq

RNA-Seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time. The process of RNA-seq has been described in [8] (see figure 4.

2.3 Relative Transcript Abundance Estimation

Sequencing RNAs are easier than sequencing proteins. In genetics, there might often be the necessity of knowing the proportion of proteins present in a cell. However, as protein sequencing is replaced by RNA-sequencing as a proxy, we are interested in finding the *Relative Transcript Abundance* of the RNA-transcripts: which deals with the percentage amount of the RNA-transcripts present in a set of cell or genome reads.

2.3.1 Challenges and motivation

Usually, in theoritical term, to compute the relative abundances, it is obvious to count the amount of reads from a transcript and deduce the parameters thereby. But there are some challenges. Due to the presence of isoform, it can happen that a single sequence read may be mapped to multiple transcripts while short read mapping. This multiple mapping pushes us to use the expected number of reads from each transcripts rather than using their exact numbers

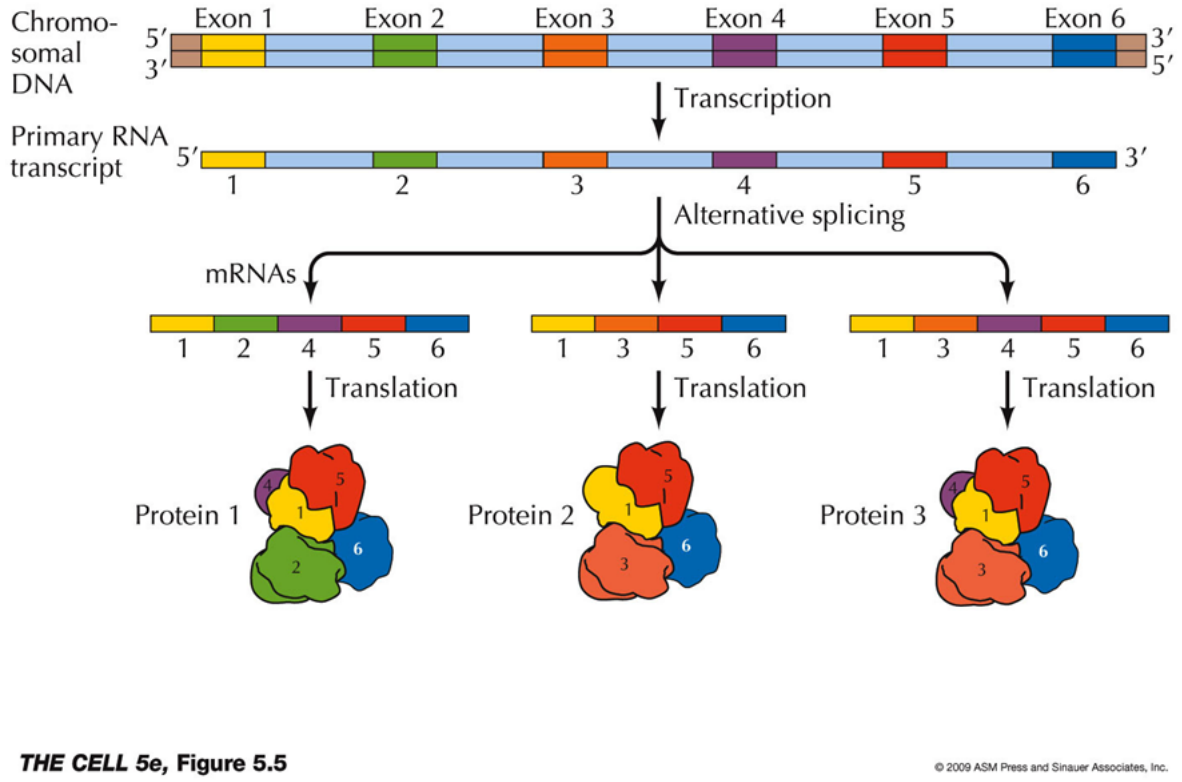

Figure 3: Protein A, B and C are isoforms encoded from the same gene through alternative splicing [6]

(remember using the expected number of heads and tails from coins A & B instead of using the exact counts because we did not know which coins were tossed?).
But to use expected value, we need to know the relative abundance of the transcripts again. This poses a deadlock like situation as the coin bias problem stated in previous section, where to know the biases of the coins, we needed to know the expected numbers of heads and tails from different coins, which in turn needed to use coin biases. We solved that problem using EM algorithm. This abundance estimation problem won't be the exception too.

In the next section, we will discuss a mathematical model of relative abundance estimation and try to solve it using EM algorithm.

3 How to use EM Algorithm to solve Transcript Abundance Estimation

The proof is credited to [5].
Let T be the set of transcripts from a RNA-sequencing samples and let l_t be the length of transcript $t \in T$.
We define $\rho = \{\rho_t\}_{t \in T}$ be the relative abundances of transcripts such that $\sum_{t \in T} \rho_t = 1$
Let F be the set of reads and F_t be the set of short reads from transcript $t \in T$ ($F_t \subseteq F$).
Let there are $|F|$ reads of length m on average. So, for a transcript $t \in T$, a read could start from $l_t - m + 1$ possible locations. This will be denoted as the *effective length* of transcript t, $\tilde{l}_t = l_t - m + 1$.

3.1 Expectation step

As per generative model assumptions, the probability of choosing a transcript t for read is,

$$\frac{\rho_t \tilde{l}_t}{\sum_{r \in T} \rho_r \tilde{l}_r}$$

There are $\tilde{l}_t$ possible places for a read from t to start at. So the probability that a read was chosen from transcript t is equal to the product of the probabilities of choosing t and choosing a place from t to start at,

$$\frac{\rho_t \tilde{l}_t}{\sum_{r \in T} \rho_r \tilde{l}_r} \times \frac{1}{\tilde{l}_t}$$

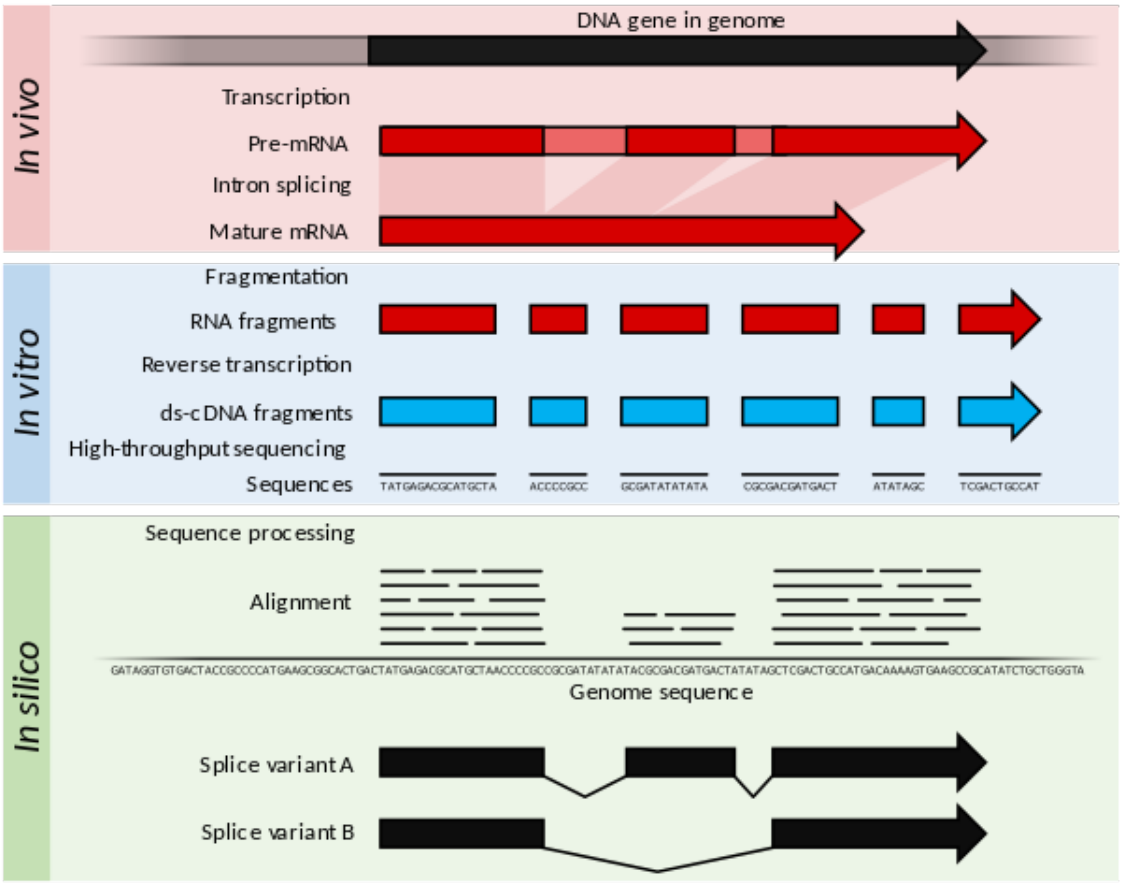

Figure 4: Summary of RNA-Seq. Within the organisms, genes are transcribed and spliced (in eukaryotes) to produce mature mRNA transcripts (red). The mRNA is extracted from the organism, fragmented and copied into stable ds-cDNA(blue). The ds-cDNA is sequenced using high-throughput, short-read sequencing methods. These sequences can then be aligned to a reference genome sequence to reconstruct which genome regions were being transcribed. This data can be used to annotate where expressed genes are, their relative expression levels, and any alternative splice variants [8].

So, we will start the estimation process by assuming some initial values of hidden variables ρ.

3.2 Maximization step

So, for the set of reads from t, the number of reads is $X_t = |F_t|$.
To compute the likelihood of the parameter settings ρ, we need to compute the probabilities of $f \in F_t$ being generated from transcript t,

$$(\frac{\rho_t \tilde{l_t}}{\sum_{r \in T} \rho_r \tilde{l_r}} \times \frac{1}{\tilde{l_t}})^{X_t}$$

So, the likelihood for parameter settings ρ is,

$$\mathbf{L}(\rho) = \prod_{t \in T} (\frac{\rho_t \tilde{l_t}}{\sum_{r \in T} \rho_r \tilde{l_r}} \times \frac{1}{\tilde{l_t}})^{X_t}$$

Let, $\alpha_t = \frac{\rho_t \tilde{l_t}}{\sum_{r \in T} \rho_r \tilde{l_r}}$.
So,

$$\mathbf{L}(\rho) = \prod_{t \in T} (\frac{\alpha_t}{\tilde{l_t}})^{X_t}$$

After applying log-likelihood and taking derivatives to find the maxima, we get,

$$\hat{\alpha_t} = \frac{X_t}{\sum_{t \in T} X_t}$$

,

As $\sum_{r \in T} \rho_r \tilde{l_r}$ will be equal for each $t \in T$, we can write,

$$\hat{\alpha_t} \propto \hat{\rho_t} \tilde{l_t} \implies \hat{\rho_t} \propto \frac{\hat{\alpha_t}}{\tilde{l_t}}$$

From this calculation, we can re-estimate the value of ρ and use it in next iteration of EM until convergence.

4 Some related works

[3] and [4] are two notable works regarding abundance estimation problem.

References

[1] Boundless. The Central Dogma: DNA Encodes RNA and RNA Encodes Protein, 2017.

[2] Do, C. B., and Batzoglou, S. What is the expectation maximization algorithm? *Nature biotechnology 26*, 8 (2008), 897–899.

[3] Li, B., and Dewey, C. N. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC bioinformatics 12*, 1 (2011), 323.

[4] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. Rna-seq gene expression estimation with read mapping uncertainty. *Bioinformatics 26*, 4 (2009), 493–500.

[5] Pachter, L. Models for transcript quantification from rna-seq. *arXiv preprint arXiv:1104.3889* (2011).

[6] Wikipedia. Alternative splicing — wikipedia, the free encyclopedia, 2017. [Online; accessed 29-July-2017].

[7] Wikipedia. Alternative splicing — wikipedia, the free encyclopedia, 2017. [Online; accessed 29-July-2017].

[8] Wikipedia. Alternative splicing — wikipedia, the free encyclopedia, 2017. [Online; accessed 29-July-2017].

[9] Wikipedia. Dna sequencing — wikipedia, the free encyclopedia, 2017. [Online; accessed 29-July-2017].

[10] Wikipedia. Primary transcript — wikipedia, the free encyclopedia, 2017. [Online; accessed 29-July-2017].

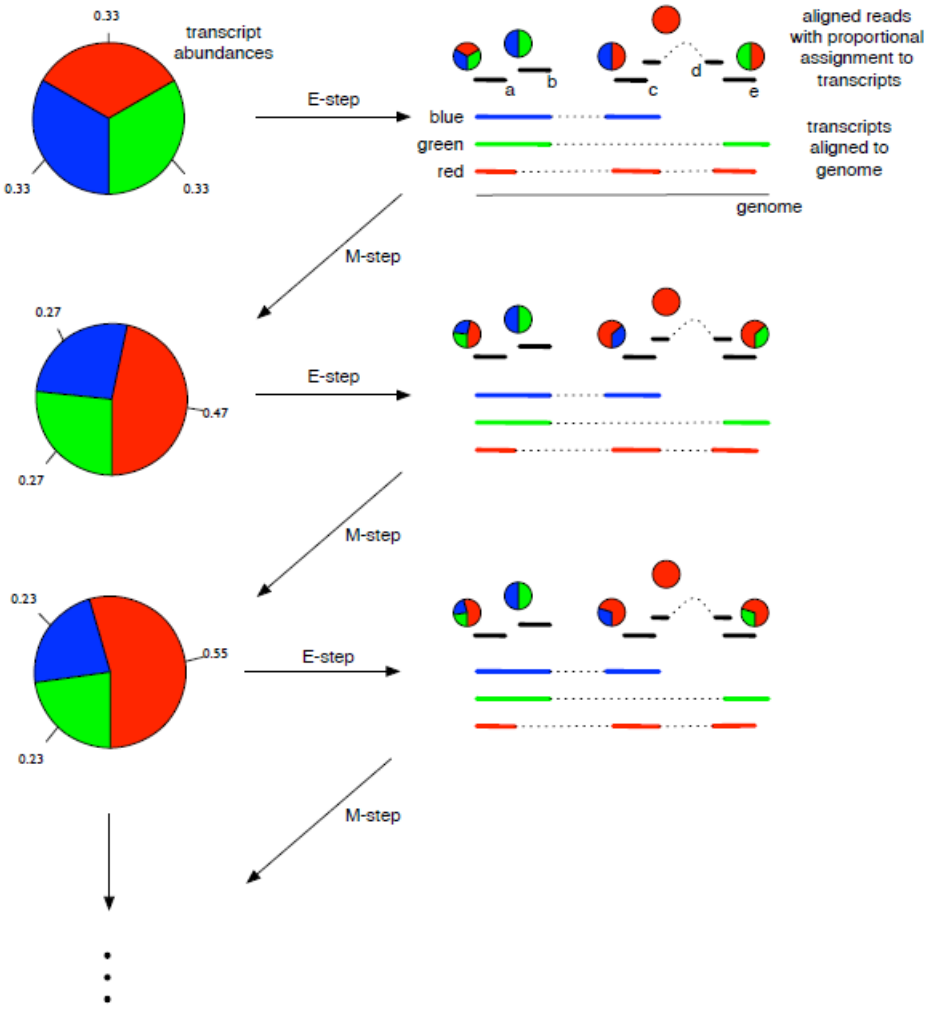

Figure 5: EM in Transcription Abundance Estimation [5]