Lesson 5 - mRNA-seq Analysis Using Transcript Annotations

26/04/17

Transcriptome

- The transcriptome is the set of all messenger RNA molecules(mRNA, rRNA, tRNA, ncRNA) in one cell or a population of cells.

Why should we bother to characterize the transcriptome?
Most of the regulation happens in the transcription phase

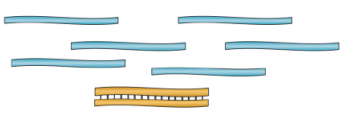
- Snapshot of the "internal state" of the cell
- In contrast to the genome which is static the transcriptome dynamics helps us understand different states and study them
- We can answer questions like - what genes are active?
- Find genes expression levels.

How it is done?
Used to be RNA →cDNA library →hybridization to microarray.
Today we are using **RNA - sequencing.**

RNA - Sequencing

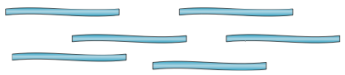
Today there are many methods to perform RNA-sequencing, here is a general scheme of the process :

1. RNA is isolated from tissue



2. Remove DNA with DNase.

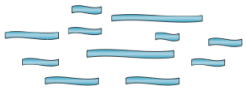
   We can also filter just the mRNA out of the total RNA by using polyA primer for example.



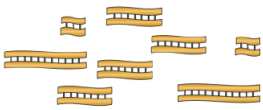
3. Fragment RNA -

   A specific length range is needed in order to sequence properly.

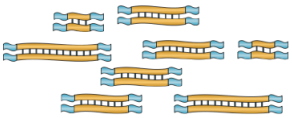
   Fragmentation of the RNA also reduces 5'/3' bias of randomly primed-reverse transcription and the influence of primer binding sites.



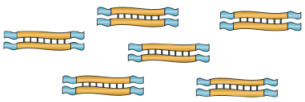
4. Build cDNA library - RNA is reverse transcribed to cDNA



5. Ligate sequence adaptors



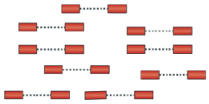
6. Using PCR to amplify, select range of sizes



7. Sequence cDNA library , with Illumina for example.

   Sequencing can be done by:

   - Single end sequencing - sequence only from one end of the fragment
   - Paired-end sequencing - sequence from both ends of a fragment



Analysis of mRNA-seq Data

The Main challenge with mRNA-seq

- Many short reads (typical read length - 50-75bp) originating from long transcripts (range from 500 to 10000bps , 2000bps on average)
- As a result of RNA instability we keep it as cDNA which leads to loss of data, such as orientation.
- We do not know how much of the total RNA we covered.
- As a result of RNA alternative splicing we might have a few different isoforms of the same transcript. This is handled using a unique part(if there is) from each isoform a part of the input to a probabilistic analysis (EM,Rescue).

Analysis Methods

1. Using full transcriptome annotations - focus of this lecture.
2. Using genome sequence - this approach relies on the fact that we have the genome as a reference but we don't know what are the isoform expressed and their expression level. Will be discussed next lecture.
3. De novo transcriptome assembly - this approach does not require a reference genome to reconstruct the transcriptome, and is typically used if the genome is unknown or incomplete. Will be disscussed next lecture.

mRNA-seq Analysis using Full Transcript Annotations

Estimating Transcript Abundance

- Assume for each gene we get all of its transciptions. Even in that case, we can't know for sure from which isoform it came from.

  Input -

  (1) A full qualitative catalogue of transcriptome

  (2) Set of sequenced fragments (reads)

Output -

ρ - Estimated abundance (expression level) of each transcript

Example

- Figure 1 - the first reads part can belong to either the blue or red transcript, but if you look at the next group of reads we see match only the blue transcript.

  That is why in this case we assume that those reads probably originated from the blue transcript $\rho(blue) > \rho(red)$.
- Figure 2 - the first reads part can belong to either the blue or red transcript, but if you look at the next group of reads we see match only the red transcript.

  That is why in this case we assume that those reads probably originated from the red transcript $\rho(red) > \rho(blue)$.

But looking at figure 1 we might miss the red small part in the mRNA-seq process, so how can we be sure?

Figure 1: 

Figure 2: 

Simple Generative Model (multinomial)

The assumption is that all reads uniquely mapped to a single transcript

- T- set of transcripts (isoforms)
  - l_t - length of transcript t
  - ρ_t - relative abundance of transcript t, s.t $\sum_{t \in T} \rho_t = 1$
  - F_t - set of reads mapping to transcript t(a set of all the reads matching isoform t)
  - $\tilde{l_t}$- (l_t - m + 1), effective length of transcript t, where m is the read length

- $\alpha_t := P(f \in t) = \frac{p_t \tilde{l_t}}{\Sigma_{r \in T} p_r \tilde{l_r}}$, probability of a read being from transcript t.

  We can see that $\alpha_t \propto p_t \tilde{l_t}$ (division by a constant $\Sigma_{r \in T} p_r \tilde{l_r}$)$therefore$ we can say $\rightarrow \rho_t \propto \frac{\alpha_t}{\tilde{l_t}}$

Example

In figure 3 we see two different reads with the same abundance $\rho_1 = \rho_2$, but different length $2l_1 = l_2$, therefore we get $\alpha_2 = 2\alpha_1$

Figure 3:

- In order to calculate the abundance $\rho_t$for a specific transcript t ($t \in T$), we go through all it's reads(read from F_t) and for each multiply the probability of a read being from transcript t ($\alpha_t = \frac{p_t \tilde{l_t}}{\Sigma_{r \in T} p_r \tilde{l_r}}$) with $\frac{1}{\tilde{l_t}}$ which represent the probability of that read to start at one of the$\tilde{l_t}$ places in t.
- Therefore the likelihood of the data(sequencing output) as a function of ρ is : $\mathcal{L}(\rho) = \prod_{t \in T} \prod_{f \in F_t} (\frac{p_t \tilde{l_t}}{\Sigma_{r \in T} p_r \tilde{l_r}} \cdot \frac{1}{\tilde{l_t}})$
- Let's define X_t as the total number of reads mapped to transcript t, hence, $X_t = |Ft|$
  - $\mathcal{L}(\rho) = \prod_{t \in T} (\frac{p_t \tilde{l_t}}{\Sigma_{r \in T} p_r \tilde{l_r}} \cdot \frac{1}{\tilde{l_t}})^{X_t}$
  - $\prod_{t \in T} (\frac{p_t \tilde{l_t}}{\Sigma_{r \in T} p_r \tilde{l_r}} \cdot \frac{1}{\tilde{l_t}})^{X_t} = \prod_{t \in T} (\alpha_t \cdot \frac{1}{\tilde{l_t}})^{X_t} = \mathcal{L}(\alpha)$
- Therefore $\Rightarrow \mathcal{L}(\alpha) = \prod_{t \in T} (\frac{\alpha_t}{\tilde{l_t}})^{X_t} \propto \prod_{t \in T} (\alpha_t)^{X_t}$(multinomial distribution)

  *We assume we know $\tilde{l_t}$

ML Estimator of ρ_t

- Since we just showed $\mathcal{L}(\alpha)$ has a multinomial distribution, $\hat{\alpha_t} = \frac{X_t}{N}$
- $\mathcal{L}(\rho) = \prod_{t \in T} (\frac{\alpha_t}{\tilde{l_t}})^{X_t}$ (since $\rho_t \propto \frac{\alpha_t}{\tilde{l_t}}$)
- $N = \sum_{t \in T} X_t$ (from $X_t$definition)

$\Rightarrow \hat{\rho_t} = \frac{\frac{\alpha_t}{\tilde{l_t}}}{\Sigma_{r \in T} \frac{\alpha_r}{\tilde{l_r}}} \propto \frac{X_t}{N \cdot \tilde{l_t}} \propto \frac{X_t}{\frac{N}{10^6} \cdot \frac{\tilde{l_t}}{10^3}}$

- That is often called **FPKM** - fragments per kilobase of transcript per million reads

  We normalize the FPKM because libraries can have different sizes.

Multi - reads

In real life, many reads can map to more than one transcript, for several reasons:

- Overlapping isoforms
- Repeats in the genome.
- Sequencing errors - less likely since usually even 20bp are enough to uniquely map.
- Abundance estimation using incomplete data.

Is ignoring multi reads a bad idea ?

- Loss of information
- Biased estimation

Parameter Estimation using the "Rescue" Method

Step_0 Discard all multi-reads - reads which map to a few transcripts.

Calculate "unique length" of each transcript, refers to the unique parts in each transcripts, parts to which only a single read is mapped to.

Moreover a transcript might have a overlap with another read and then the reads in this region are considered multi reads but the rest of the transcript might map uniquely and that remaining part will be considered in calculating the "unique length".

Estimate abundances ρ_t using remaining reads.

Step_1 Estimated expected contribution of f to each t :

Divide each multi-read between $y_{f,t}$ transcripts proportionally to their abundances ρ_t ($\sum_{t'} y_{f,t'} = 1$)

Step_2 Recompute ρ_t abundances based on the updated counts X_t for each transcript t

Comments on "Rescue" Method

- Why stop after last step?
- Initialization by unique reads could bias results.

  For example if we have 2 isoforms like in figure 4 and in our sample we have 50% blue isoform and 50% green.

  At step zero we will discard the blue isofrom since it has no unique parts. As a result we won't find any match read ($\rho_t = 0$) .
- How to regularize initial estimation of ρ_t ? In order not to get a biased result.
- What about isoforms with no unique reads? like the blue in the figure 4.

  ρ_{blue} will be initialized to zero , in order to overcome this problem we can use uniform

  initialization.

Figure 4:
Two different isoforms of the same transcript

Define Compatibility Matrix Y

$$y_{f,t} = \begin{cases} 1 & read \quad f \quad aligns \quad to \quad transcript \quad t \\ \\ 0 & otherwise \end{cases}$$

- $\mathcal{L}(\alpha) = \prod_{t \in T} (\frac{\alpha_t}{\tilde{l}_t})^{X_t} \Rightarrow \mathcal{L}(\alpha) = \prod_f (\sum_t y_{f,t} \frac{\alpha_t}{\tilde{l}_t})$
- Where the probability for each f is accumulated over all transcripts t that match f , since $y_{f,t} = 1 \iff$ read f aligns to transcript t

Parameter Estimation using EM

Init Uniformly/by random/ Otherwise

M-step Compute expected contributions of f to each t - divide each multi-read between $y_{f,t}$ transcripts proportionally to their abundances ρ_t

E-step Compute expression values (transcript abundances) ρ_t based on updated X_t counts

Comparison Between the Two Algorithms

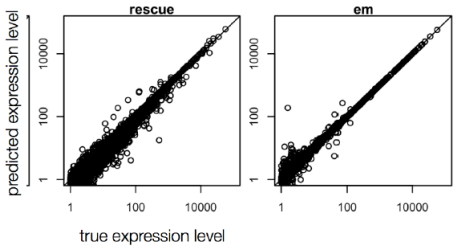
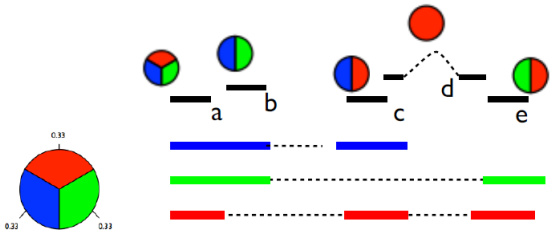


Figure 5:

From figure 5 we see that both of the algorithms get good correlation between the true expression level and the predicted expression level, but it is clear that the EM gets a stronger correlation.

Example to the EM Algorithm



$$\mathbf{Y} = \begin{array}{c} \\ \texttt{red} \\ \texttt{green} \\ \texttt{blue} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{pmatrix} 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Figure 6:

- **Init-** each small pie chart represents the probability a read belongs to a transcript($P(f \in t)$).The big pie chart represents the abundance of each transcript(ρ_t). Here we started with uniform initialization and we fill the matrix according to $y_{f,t}$ formula; 1 in every transcript a read might belong to.

- **M-step**

  Recomputing X_t -

  X_{red}= 0.33 + 0.5 + 1 + 0.5 = $2\frac{1}{3}$

  X_{green} =X_{blue}= 0.33+0.5+0.5 = $1\frac{1}{3}$r

- **E-step**

  Compute transcript abundances ρ_t based on updated X_t counts

  $\rho_{red} = \frac{X_{red}}{X_{red}+X_{blue}+X_{green}} = 0.47$

  $\rho_{blue} = \rho_{green} = \frac{X}{X_{red}+X_{blue}+X_{green}} = 0.27$

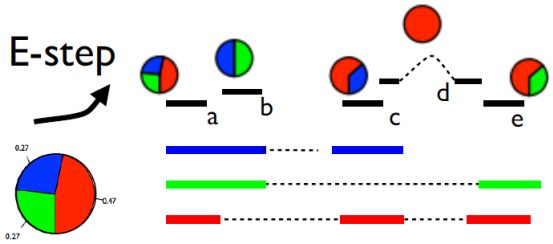


Figure 7:

What You Should Worry About? - You Don't

- Sequencing errors, remember $y_{f,t}$ since it is binary if a read won't perfectly match$\rightarrow$ $y_{f,t} = 0$(causes loss of data)

  A suggested solution would be to replaced by P(f |t), this way we will get a probability instead of binary classification.

- EM - different initialization might lead to different results.
- EM - convergence into global optimum is not guaranteed, we might get to a local optimum (solution - few restarts)
- If we don't have enough data $\Rightarrow$ Regularization is needed