FEture STructure (FEST) model

February 7, 2017

Chapter 1

Introduction of FEST model

RNA-seq is a high-throughput technology based on next-generation sequencing (NGS). It has been widely applied in a variety of studies in genetic biology. RNA-seq uses the large amount of short sequence reads to interrogate the transcriptome, the existence and expression level of RNAs. Recently, the significant biases have been confirmed affecting the accuracy of RNA-seq expression by computational estimates in many different models, which results from the factors, such as GC content, reads mapping uncertainly, degradation of fragments et cetera. In the current methods, these problems are commonly addressed by developing statistical models. Specifically, approaches using in models can be mainly categorized into two groups. First, mixture models are developed to estimate the target parameters based on the experimental data by applying Expectation Maximization (EM) algorithm. The representative researches are published by Trapnell and Salzberg et al., 2010, for the details of Cufflinks measurement tool [11]; Li and Dewey et al., 2011, for RSEM measurement tool [5] and so on. Second, models are not only simulated by EM, but also considered the regression methods to deal with data. Robustness analysis of specific parameters is also applied. Ref to Jiang and Salzman, 2015 [3]. Love, Hogenesch and Irizarry, 2016 [6]. However, most of the publications about RNA-seq only focus on one or two factors which affect the accuracy of measurement of target molecules (like isoforms, transcripts and genes). In this paper, we consider GC content, the degradation of fragments and uncertain mapping reads as the factors which cause the bias on the exon expression level.

Before introducing our model, we will summarize the current popular methods in the correction of different bias problems.

- RSEM
- FPKM
-

We propose the program in an alternative RNA-seq model development, FEature STructure (FEST) model. We intend to implement the FEST model and compare the results to other methods to determine whether FEST is equivalent to or better than the previously described counterparts in terms of accuracy, computational memory requirements, and speed. Specifically, in FEST model, we will address the GC content bias problem in read counts, a probability for "mean read count" on individual locations of each exon will be obtained, only the uniquely mappable reads are considered in GC content bias model. The uncertainly mappable reads will be used in redundant bias model to estimate the expression level for each exon by finding one "best" nucleotide "position" through EM algorithm. Finally, on each exon, the results from GC content bias model and redundant bias.

Chapter 2

GC content bias on fragment abundance

GC content bias reflect the connection between fragment abundance (read depth) and GC content in the sequenced data. These bias are not consistent between samples, and there is no best method to remove it in a single sample so far [1]. Due to a lack of modelling sample-specific bias in present methods, reads coverage is commonly amplified or lost and is related to sequence features like fragment GC content and GC stretches [6]. In the current published work, some researches have been investigated on the GC bias that affecting accuracy of the quantification results. For example, a model in Chip-seq technology has been proposed by Teng M and Irizarry R A, which improves the peak calling algorithms by estimating "effective GC-content" to correct the GC content bias [10]. The spline function is wildly used [6, 10]. However, In the FEST model, for the efficiency of the computation, only the estimation of poisson distribution rate is considered. We categorize the entire genome into several classes based on the different GC content through the given size windows. A statistical model is developed by using the "mean read count", which follows the poisson distribution. Here, the "mean read count" is defined for each GC content classes refers to the "signal position model" as proposed by Benjamin Y and Speed T P [1]. However, we are applying this model in RNA-seq to obtain the specific GC content bias for each read rather than the different GC content comparison between samples in DNA-seq, which makes different implementations. Furthermore, the local region is used to evaluate the GC bias in FEST model, instead of the individual position in the "signal position model", which reflects the targets of the work are distinct.

There is a consensus showing that the GC content affects the fragmentation of RNA-seq, and therefore influences the read coverage in the regions of genome. GC content is related to the stability of DNA. Because of its three hydrogen bonds, GC-rich area usually gives a lower probability of a fragmentation point occur in the genome. For every given tissue sample, the GC content of entire

genome is fixed, as well as the GC content bias is certain. However, in the different parts of genome, due to the different number GC base pairings, the GC content bias is not the same, which causes the different read coverages at each nucleotide position. We can imagine that for the most ideal case of GC content bias, the read coverages distribute evenly on the entire genome, which means there is no GC content bias if uniform distribution of read occurs. In order to find the GC content bias for each read, the "global mean read count" of the whole transcriptome based on the GC content is required to be investigated. For each region of the transcriptome, the "difference" between the "mean read count" and the "global mean read count" reflects the GC content bias of that region. The results are used in poisson distribution to evaluate the probability of this "difference". This probability here, we define it as the GC content bias index for the given region. Moreover, as the reads derive from the genome, we assume that the reads and their derived regions have the same GC content bias index. A general example of this calculation is introduced in Figure 2.1.

In order to find the specific GC content bias index for each read , we need to estimate the rate λ_{gc} (i.e. the "mean read count" for different GC classes for each read). Only the uniquely mappable reads are kept here, because the distribution of multi-position mapping reads has a significant influence on abundance of reads, it disturbs the results whether the bias greatly depends to the GC content or multi-position mappabilities.

We consider Read GC content Bias (RGCB) model which links read count to the GC content. The number of r read are randomly picked up on the entire transcriptome, instead of on fragments or reads. We assume that for a given transcriptome, the different GC content bias is fixed on the different local region. The expected counted read starts on a local region depending on the GC content, and this local region is defined as the size of window, h, which is the average length of reads. We use $G(h)$ to represent the GC count of the h bp window. The S_{gc} denotes the total number of reads starting on the windows which GC contents are all equal to $G(h)$. Thus, the rate λ_{gc} is obtained

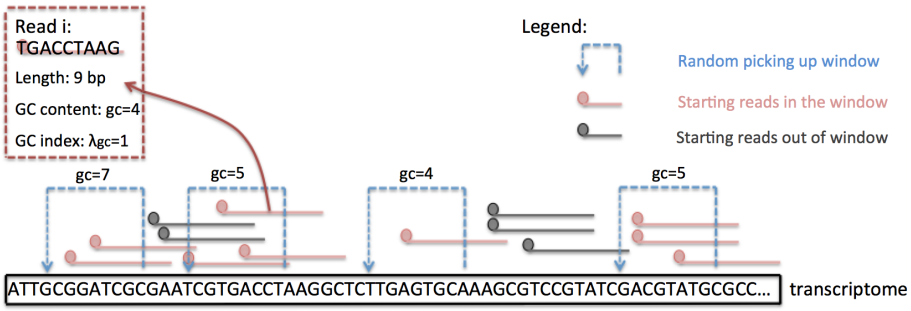
$$\lambda_{gc} = \frac{S_{gc}}{R_{gc}}, \tag{2.1}$$

where R_{gc} is the number of windows which GC content is equal to $G(h)$. Additionally, for the h bp window, the model contains parameters $G(h) = \{0, 1, 2, ..., h\}$. Note that $S = \sum_{gc=0}^{h} \lambda_{gc} R_{gc}$. Therefore, the "global mean read count" $\tilde{\lambda} = S/r$. For the given read i, we consider that on the derived transcriptome region, read i has the corresponding λ_{gc}^{i} based on its GC content. Hence, we apply poisson distribution to represent the RGCB showing as follows

$$P(K = \lambda_{gc}^{i}) = \frac{\tilde{\lambda}^{K} e^{-\tilde{\lambda}}}{K!}. \tag{2.2}$$

The above equation means "the probability of read i which has GC content equal to gc is affected by the entire transcriptome GC content."

Figure 2.1: A general example to introduce how to estimate GC index of given read. The blue arrow indicates the random samples chosen at that position on the transcriptome. The red circle represents a read starting the position in the window. The **black** circle gives a read starting the position out of the window.



GC content (gc)	0	1	2	3	4	5	6	7	8	9
Window #	-	-	-	-	1	2	-	1	-	-
Reads #	-	-	-	-	1	6	-	2	-	-
GC index (λgc)	-	-	-	-	1	3	-	2	-	-

For a given section of transcriptome, we have the sequenced nucleotides of A, T, G, C. Figure 2.1 gives general example that four random windows are picked up on a piece of transcriptome. Each of the window has 9 bp length. Note that the length of window is chosen by the same length of read. Different windows contain different read coverages (the red circle line). The GC content of each window can be calculated by counting G, C bases. Because the length of window is 9 bp here, we thus have (9+1=) 10 classes categorized from $gc = 0$ to $gc = 9$. For each class, we measure the total number of windows which belongs to that class (the blue arrow) and the total number of reads which starts the position in their corresponding windows (the read circle). Therefore, the GC index rate λ_{gc} of the specific classes are estimated by dividing total reads number by the total windows number. More details of the calculation is defined in Equation 2.2. Additionally, since just a general example is given here, some classes, such as $gc = 0, 1, 2, ...$, have no GC index expression. However, in realistic cases, the random samples are picked up sufficiently, and the distribution between GC index and GC content are nonlinear curve showing in some publications [1].

Chapter 3

Degradation of RNA-seq

In most of RNA-seq samples, the RNA degradation occurs and it depends on the factors such as the specimen collection, the storage conditions [2, 7, 12]. RNA degradation affects the read counts in a genetic manner, and thus has a significant influence on the entire gene expression [13]. Currently, several metric methods are applied to address the RNA degradation problem. One of the most wildly used approach is called RNA Integrity Number (RIN) [2,7,9,12]. However, the RIN measurement has its own weakness that limits the application in the RNA sequence data analysis [13]. Therefore, a novel method named Transcript Integrity Number (TIN) is taken into consideration in our case. TIN is developed based on information theory to measure not only the RNA integrity at transcriptome level but also the each signal transcript degradation [13]. The core of TIN algorithm is to estimate the uniform coverage of a transcript by Shannon's entropy (uncertainty or information content). Here, we are applying it to estimate the percentage of a given read that has uniform distribution on the mapping exon.

As previously mentioned, the reads are associated with their derived transcripts' sequences in some common features, such as the similar nucleotides orders, the GC content, the degradation "potential" and et cetera. We assume that the reads and their mapping location has the same degradation rate, and this rate can be investigated by the following steps.

Assuming that the degradation of a exon would cause the curved read depths. Therefore, we used TIN metric to uniform the read coverage for the exon. For the given exon, the length of it is n nucleotides long and the read coverage at each position is R_k, where $k = 1, 2, ..., n$. The proportion coverage (P_k) for each nucleotide site is calculated by:

$$P_k = \frac{R_i}{\sum_{k=1}^{n} R_k}.$$

Note that $\sum P_k = 1$.

By Shannon's entropy, the read coverage evenness of a given exon is measured

as:

$$H = -\sum_{k=1}^{n} P_k \log_b P_k,$$

where b is the base of logarithm function used. Common values of b are 2, Euler's number e, and 10, and the units of entropy with different base value $b = 2, e, 10$ are *shannon*, *nat*, *hartley* respectively. Here the base value is $b = e$.

We define the case that when there is no read coverage at some particular positions (i.e. $P_k = 0$), the entropy $H = 0$. The maximized H is obtained when the read coverage perfectly uniformly distribute on the entire exon. Now, for each exon, we have the proportion read coverage P. The mappable reads from derived exon can estimate its Shannon's entropy by picking up the value of proportion coverage at each nucleotide. Let the length of given read i be l, we have

$$H_i = -\sum_{k=1}^{l} P_k \ln P_k.$$

The entropy gives the uncertainty of the species in a sample, but it does not tell the number of species in the community [4]. Though the Shannon's entropy is a profound and useful index to capture uniformity of species, its value gives uncertainty instead of diversity, which is difficult to implement and interpret in biology [4]. Therefore, a new concept is introduced to measure the "real uniformity" proposed by Jost et al. [4], which is given as:

$$\tilde{H}_i = \exp(H_i) = \exp(-\sum_{k=1}^{l} P_k \ln P_k),$$

where $\tilde{H}_i \in [0, l]$.

The biological meaning of $\tilde{H}_i$ can be interpreted as the number of nucleotides that have the uniform coverage for the given read i. We define the degradation rate of given read i as:

$$de_i = \frac{\tilde{H}_i}{l}, \tag{3.1}$$

where $de_i \in [0, 1]$.

The degradation rate de_i is considered as the probability of read i that distributes uniformly.

Chapter 4

Redundant reads expression

4.1 Problem statement

Due to the reads mapping uncertainly in RNA-Seq, the mapped reads may be repetitive sequences (redundant) from other locations of the genome. In this chapter, we mainly focus on solving the redundant problem of gene expression.

The measure of relative expression is the proportion of nucleotides of the genome made up by the give exon. For exon j, we denote the quantity expressions by u_j. The problem addressed here is that of using the given RNA-Seq data to estimate the values of u_j. Specifically, we are trying to find the "best" nucleotide position to represent the exon expression level. The definition of "best" means the most counts of mappable reads after dealing with redundant.

4.2 Methods

4.2.1 Statistical probability model

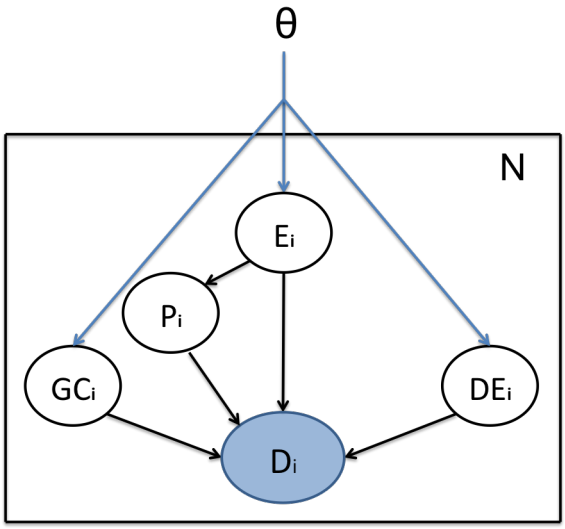
The statistical probability model is estimated to solve the redundant of exon expressions based on the Bayesian network. Model contains N reads of length L_i. For the given read i, the i takes a integer value in $[1, N]$, it is associated with four hidden random variables E_i, P_i, GC_i, DE_i, which represent the derived exon, starting position(s), the GC content and the degradation level, respectively. We consider C_j as the length of the given exon j, which j is a integer number from $[1, M]$. We assume that all given M exon could be found in the genome. The read sequence is given by the observed data D_i. The parameters space of this model is defined as matrix Θ, which is related to the exon

expression level. Probability matrix Θ is given:

$$\Theta = \begin{pmatrix} \theta_{11} & \theta_{12} & \theta_{13} & \cdots & \theta_{1M} \\ \theta_{21} & \theta_{22} & \theta_{23} & \cdots & \theta_{2M} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \theta_{N1} & \theta_{N2} & \theta_{N3} & \cdots & \theta_{NM} \end{pmatrix}.$$

The entry θ_{ij} of matrix Θ represent the priori probability of mapped read i which starts the position on exon j with its own GC content and degradation rate. Note that probability matrix Θ is sparse matrix, its most of entries are zeros. Additionally, the sum of each row of matrix Θ is equal to one, i.e. $\sum_{j=1}^{M} \theta_{ij} = 1$.

Figure 4.1: Bayesian graphical model for RNA-seq data. The notations of $\theta, E_i, P_i, GC_i, DE_i, D_i, N$ represent the objective parameters, the derived exon of read i, the starting position on exon of read i, the GC content of read i, the degradation level of read i, the observed data and the total number of reads from RNA-seq respectively.



Model based on the Bayesian Network follows Figure 4.1. We define the observed data likelihood function for the model as follows:

$$P(e_i, p_i, gc_i, de_i, d_i|\ \Theta) = \prod_{i=1}^{N} P(e_i, gc_i, de_i|\ \theta) P(p_i|\ e_i) P(d_i|\ e_i, p_i, gc_i, de_i) \quad (4.1)$$

$$= \prod_{i=1}^{N} P(e_i, gc_i, de_i|\ \theta) P(p_i|\ e_i) P(d_i|\ e_i, p_i) P(d_i|\ gc_i) P(d_i|\ de_i)$$

Let e_i, p_i, d_i be a specific sample of E_i, P_i, D_i. Note that the model defined here is similar model as proposed by Bo L. and Colin N. D. [5], but we explicitly

model that the probabilities θ for each read are different, which is not the same θ defined by Bo L. and Colin N. D [5]. Moreover, the Expectation Maximization algorithm will be used to find the best parameters θ, which is introduced lately.

We let $P(E_i = j|\, \theta) = \theta_{ij}$, which gives the probability of read i distribute on exon j.

The variable P_i is a random integer from $[1, \max_j C_j]$. We interpret $P(P_i = k|\, E_i = j)$ as the distribution of read i that starts at the position k on exon j. Assuming that the reads distribute uniformly on exon across genome, we have $P(P_i = k|\, E_i = j) = 1/C_j$. When the lengths of reads are considered, we let $P(P_i = k|\, E_i = j) = L_i/C_j$, where L_i is the length of read i. In the assumption of that mRNA-seq have poly(A) tails, which allow the reads to start the position from the end of exon, it gives $P(P_i = k|\, E_i = j) = 1/(C_j - L_i)$ [5]. The non-uniform distribution will be discussed in the following.

Remark that D_i represents the observed data of read i for the model. We use the following equation to present the conditional probability of sequenced read starting position on the given exon. We let $P(D_i = d_i|\, E_i = j, P_i = k) = \prod_{t=1}^{L_i} V_{xt}(r_{k+t-1}, d_t)$ for the case of that read i does not overlap between different exon, where V is called position probability matrix (PPM) [8]. In the matrix V, it has 4 rows for nucleotides of A, T, G, C, and It also has the columns for each position of the given exon in our model. The element of PPM is defined as

$$V_{xk} = \frac{1}{N} \sum_{i=1}^{N} I(U_{ik} = X),$$

where k is defined previously, which is a integer value from $[1, C_j]$, interpret as the nucleotides' position of exon j. N is the number of reads. Furthermore, when nucleotide U_{ik} of read i at position k has same alphabet of X, function $I(U_{ik} = X) = 1$, otherwise, it is zero. Moreover, for the case of that read i mapped on different exon, we will combine the different exon to build the bigger matrix V (for example, if a read mapped to exon i, q, s. The size of above matrix V increased to $4 \times C_j + C_q + C_s$, and the value of k in the above equation varying to $[1, C_j + C_q + C_s]$). Then, the calculation of $P(D_i = d_i|\, E_i = j, P_i = k) = \prod_{t=1}^{L_i} V_{xt}(r_{k+t-1}, d_t)$ will be done by using the new matrix V. Additionally, $X = \{A, T, G, C\}$ and corresponding $x = \{1, 2, 3, 4\}$. The size of position probability matrix V is $4 \times C_j$, where C_j is the length of exon j.

We consider that the GC contents and the degradation level of reads are associated with observed data, which help the development of the probability functions of reads. We let the correction function of GC content bias for the read i be: $P(D_i = d_i|GC_i = gc_i) = P(K = \lambda_{gc}^i) = \frac{\tilde{\lambda}^K e^{\tilde{\lambda}}}{K!}$, where $K, \tilde{\lambda}$ are defined in Equation 2.2 in Chapter 2. Furthermore, the degradation rate function of read i is given as: $P(D_i = d_i|\, DE_i = de_i) = \frac{\tilde{H}_i}{l}$, where $\tilde{H}_i, l$ are detailed in Equation 3.1 in Chapter 3. Remark that the values of $\frac{\tilde{\lambda}^K e^{\tilde{\lambda}}}{K!}$, $\frac{\tilde{H}_i}{l}$ are between 0 and 1.

As previously mentioned, our goal is finding the "best" nucleotide unit to represent belonging exon expression level. For each exon j, We set up a matrix

W^j

$$W^j = \begin{pmatrix} W_{11} & W_{12} & W_{13} & \cdots & W_{1C_j} \\ W_{21} & W_{22} & W_{23} & \cdots & W_{2C_j} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ W_{N1} & W_{N2} & W_{N3} & \cdots & W_{NC_j} \end{pmatrix}.$$

In the above matrix, the entry W^j_{ik} is considered as the probability of read i distribute to exon j at nucleotide position k. Because we are estimating the exon expression level which is defined as two equations as follow
1. If $p_i + L_i < C_j$, means the read i completely mapped on exon j.

$$W^j_{ik} = \begin{cases} 0 & \text{when read } i \text{ does not map on exon } j. \\ \theta_{ij} & k \in [p_i, p_i + L_i], \text{ when read } i \text{ map on exon } j. \\ 0 & k \in [1, p_i] \cup [p_i + L_i, C_j], \text{ when read } i \text{ map on exon } j. \end{cases}$$

2. If $p_i + L_i > C_j$, assume that the read i mapped on several exon $\{j, q, s\}$ (the idea that read i mapped on more than three exon are very similar).

$$W^j_{ik} = \begin{cases} 0 & k \in [1, p_i]. \\ \theta_{ij} & k \in [p_i, C_j]. \end{cases}$$

$$W^q_{ik} = \begin{cases} \theta_{ij} & k \in [1, C_q]. \end{cases}$$

$$W^s_{ik} = \begin{cases} \theta_{ij} & k \in [1, L_i - C_j - C_q + P_i]. \\ 0 & k \in [L_i - C_j - C_q + P_i, C_s]. \end{cases}$$

Note that k is integer number, which can be interpreted as nucleotide "position" of exon j. Reminding that p_i, L_i, represent the starting position and the length of read i respectively, C_j is the length of exon j. The objective function, u_j, of expression level for exon j is obtained

$$u_j = \max \left[\sum_{i=1}^{N} W^j_{i1}, \sum_{i=1}^{N} W^j_{i2}, \cdots, \sum_{i=1}^{N} W^j_{ik}, \cdots, \sum_{i=1}^{N} W^j_{iC_j} \right].$$

Bibliography

[1] **Benjamini Y** and **Speed T P**. *"Summarizing and correcting the GC content bias in high-throughput sequencing"*. University of California, Berkeley. 2011.

[2] **Botling J, Edlund K, Segersten U, Tahmasebpoor S, Engstrm M** and **Sundstrm M**. *"Impact of thawing on RNA integrity and gene expression analysis in fresh frozen tissue."*. Diagn Mol Pathol. 18: 44-52 2009.

[3] **Jiang H** and **Salzman J**. *"A penalized likelihood approach for robust estimation of isoform expression"*. Stat Interface. 8(4): 437-445. 2015.

[4] **Jost L.** *"Entropy and diversity"*. Oikos. 113(2): 363-75. 2006.

[5] **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): 493-500. 2010.

[6] **Love M I, Hogenesch J B** and **Irizarry Rafael A.** *"Analysis of chemical modification of RNA from formalin-fixed samples and optimization of molecular biology applications for such samples"*. Nature Biotechnology. 34: 1287-1291. 2016.

[7] **Masuda N, Ohnishi T, Kawamoto S, Monden M** and **, Okubo K**. *"Modeling of RNA-seq fragment sequence bias reduces systematic error in transcript abundance estimation"*. Nucleic Acids Res. 27: 4436-43. 1992.

[8] **Stormo G D, Schneider T D, Gold L** and **Ehrenfeucht A**. *"Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli"*. Nucleic Acids Research. 10(9): 2997-3011. 1982.

[9] **Schroeder A, Mueller O, Stocker S, Salowsky R, Leiber M** and **Gassmann M**. *"The RIN: an RNA integrity number for assigning integrity values to RNA measurements"*. BMC Mol Biol. 7: 3. 2006.

[10] **Teng M** and **Irizarry R A**. *"Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq peak callers"*. doi: http://dx.doi.org/10.1101/090704. 2016.

[11] **Trapnell C, Williams B A, Pertea G, Mortazavi A, Kwan G, Baren M J V, Salzberg S L, Wold B J** and **Pachter L**. *"Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms"*. Nat Biotechnol. 28(5): 511-515. 2010.

[12] **von Ahlfen S, Missel A, Bendrat K** and **Schlumpberger M**. *"Determinants of RNA quality from FFPE samples"*. PLoS One. 2:e1261. 2007.

[13] **Wang L, Nie J, Sicotte H, Li Y, Echel-Passow J E, Dasari S, Vedell P T, Barman Poulami** and **Kocher J-P A**. *"Measure transcript integrity using RNA-seq data"*. BMC Bioinformatics. 10.1186/s12859-016-0922-z. 2016.