Expression quantification II

Helene Kretzmer

15.05.2012

Pipeline

(i) RNA isolation from sample
(ii) RNA transcription to cDNA and fragmentation
(iii) sequencing
(iv) **mapping reads to reference genome**
(v) using read counts for expression level estimation

Mapping Problems

- unknown isoforms
- sequencing non-uniformity
- **read mapping uncertainty**

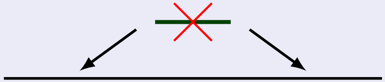
Read Mapping Uncertainty

- paralogous genes
- low-complexity regions
- high sequence similarity
- reference sequence errors
- sequencing errors

$$\Rightarrow \text{multireads} \begin{cases} \text{gene mulitreads} \\ \text{isoform multireads} \end{cases}$$

Mapping Strategies

(a) discard mulitreads



(b) rescue mulitreads

(c) em - a statistical model

Measures of Expression - isoform *i*

- τ_i .. fraction of transcripts

  percentage of isoform *i* of all transcripts in the sample

- ν_i .. fraction of nucleotides

  percentage of isoform *i* of all nucleotides in the sample

  ℓ_i .. length of isoform *i* in nucleotides

$$\tau_i = \mathrm{RPKM}_i \cdot 10^{-9} \sum_j \tau_j \ell_j$$

Measures of Expression - isoform *i*

- τ_i .. fraction of transcripts

$$\tau_i = \frac{\nu_i}{\ell_i}\left(\sum_j \frac{\nu_j}{\ell_j}\right)^{-1}$$

- ν_i .. fraction of nucleotides

$$\nu_i = \frac{\tau_i \ell_i}{\sum_j \tau_j \ell_j}$$

ℓ_i .. length of isoform *i* in nucleotides

$$\tau_i = \mathrm{RPKM}_i \cdot 10^{-9} \sum_j \tau_j \ell_j$$

EM-Model

Generative Model

- N reads
- all of length L

Assumptions

- M isoforms
- isoform sequence is known
- additional noise isoform
- uniformly distributed reads: $\frac{\text{\# reads of isoform } i}{N} \longrightarrow \nu_i$

R_n .. sequence of read n

R_n .. sequence of read n

G_n .. isoform of read n

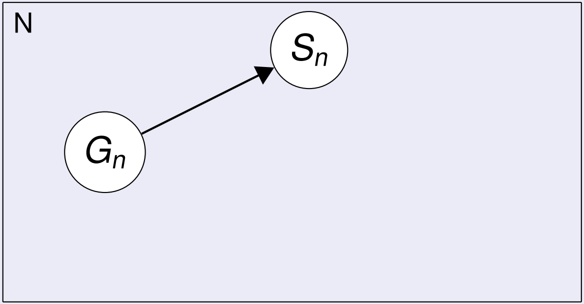
R_n .. sequence of read n

G_n .. isoform of read n

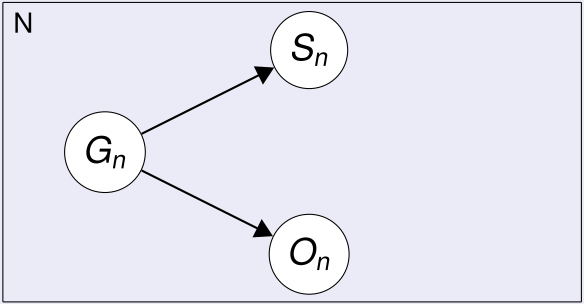
S_n .. start position of read n

R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n

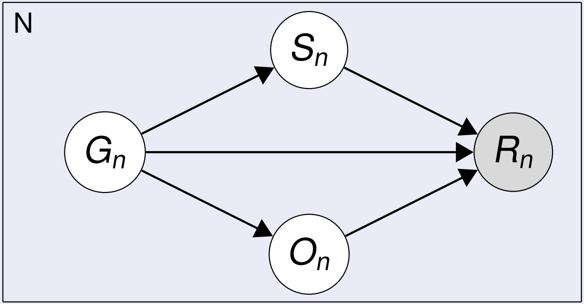
R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n



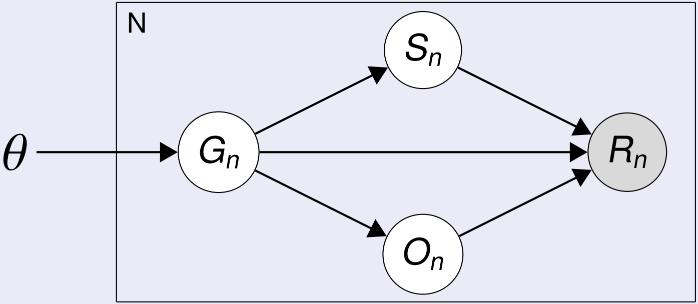
R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n



R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n



R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n
$\theta = [\theta_0, \ldots, \theta_M]$ .. expression levels of the isoforms $0, \ldots, M$



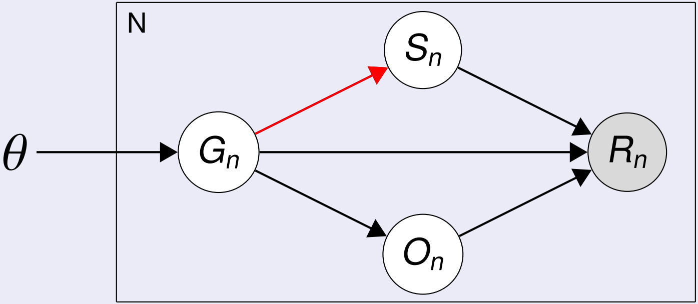
R_n .. sequence of read n

G_n .. isoform of read n

S_n .. start position of read n

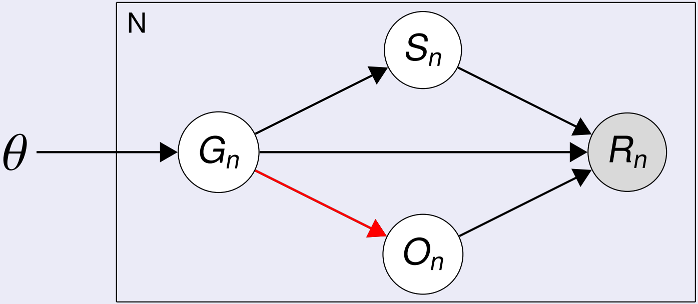
O_n .. orientation (strang) of read n

$\theta = [\theta_0, \ldots, \theta_M]$ .. expression levels of the isoforms $0, \ldots, M$



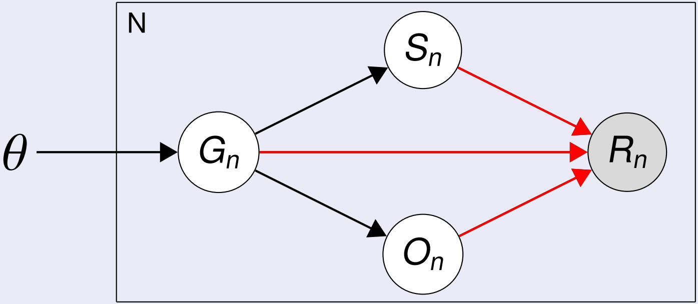
$$P(s_n | g_n)$$

R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n
$\theta = [\theta_0, \ldots, \theta_M]$ .. expression levels of the isoforms $0, \ldots, M$



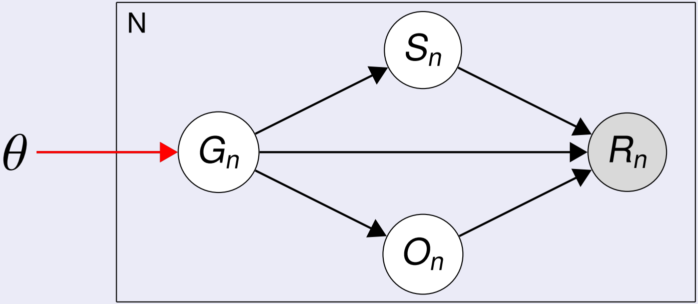
$$P(s_n|g_n)P(o_n|g_n)$$

R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n
$\theta = [\theta_0, \ldots, \theta_M]$ .. expression levels of the isoforms $0, \ldots, M$



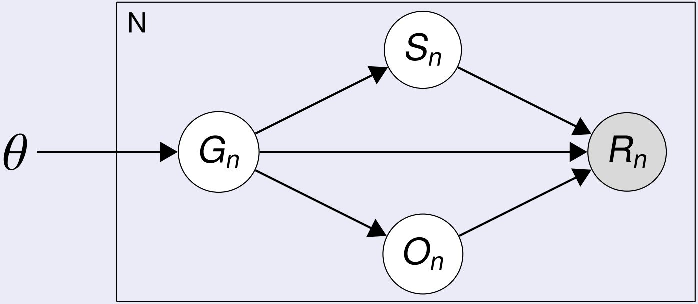
$$P(s_n|g_n)P(o_n|g_n)P(r_n|g_n, s_n, o_n)$$

R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n
$\theta = [\theta_0, \ldots, \theta_M]$ .. expression levels of the isoforms $0, \ldots, M$



$$P(g_n|\theta)P(s_n|g_n)P(o_n|g_n)P(r_n|g_n, s_n, o_n)$$

R_n .. sequence of read n
G_n .. isoform of read n
S_n .. start position of read n
O_n .. orientation (strang) of read n
$\theta = [\theta_0, \ldots, \theta_M]$ .. expression levels of the isoforms $0, \ldots, M$



$$P(g, s, o, r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(s_n|g_n)P(o_n|g_n)P(r_n|g_n, s_n, o_n)$$

Summary

- $P(G_n = i|\theta)$ .. probability that read n maps to isoform i given the expression levels $\theta_0, \ldots, \theta_M$
- $P(O_n = 0|G_n \neq 0)$ .. probability that read n has the same orientation as its template given that it is not from the noise isoform
- $P(S_n = j|G_n = i)$ .. probability that read n starts at position j given that it is from isoform i
- $P(R_n = \rho|G_n = i, S_n = j, O_n = 0)$ .. probability that read n has sequence ρ given it is from isoform i, starts at position j and has the same orientation as its template

Isoform G_n

$$P(g, s, o, r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(o_n|g_n)P(s_n|g_n)P(r|g_n, s_n, o_n)$$

$P(G_n = i|\theta)$

$G_n \in [0, M]$ — 0 noise isoform
$1, \ldots, M$ known isoforms

$P(G_n = i|\theta) = \theta_i$ and $\sum_i \theta_i = 1$

Orientiation O_n

$$P(g, s, o, r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(o_n|g_n)P(s_n|g_n)P(r|g_n, s_n, o_n)$$

$P(O_n = 0|G_n \neq 0)$

$$O_n = \begin{cases} 1, & \text{reverse complement} \\ 0, & \text{same orientation as its template} \end{cases}$$

$$P(O_n = 0|G_n \neq 0) = \begin{cases} 1, & \text{strand specific sequencing} \\ 0.5, & \text{not strand specific sequencing} \end{cases}$$

Startposition S_n

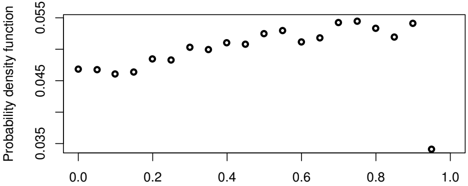
$$P(g,s,o,r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(o_n|g_n)P(s_n|g_n)P(r|g_n,s_n,o_n)$$

$P(S_n = j|G_n = i)$

$S_n \in [1,\ldots,\max_i \ell_i]$ $\qquad$ ℓ_i .. length of isoform i

$$P(S_n = j|G_n = i) = \begin{cases} \frac{1}{\ell_i}, & \text{uniform read start distribution} \\ f(\frac{j}{\ell_i}) - f(\frac{j-1}{\ell_i}), & \text{non-uniform read start distribution} \end{cases}$$

f .. empirical cumulative density function over $[0,1]$



Sequence R_n

$$P(g,s,o,r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(o_n|g_n)P(s_n|g_n)P(r|g_n,s_n,o_n)$$

$P(R_n = \rho | G_n = i, S_n = j, O_n = k)$

- strand specific protocol, known isoforms:

$$P(R_n = \rho | G_n = i, S_n = j, O_n = 0) = \prod_{t=1}^{L} \omega_t(\rho_t, \gamma^i_{j+t-1})$$

$\omega_t(a,b) = P(\text{read}[t] = a|\text{isoform}[j+t-1] = b)$
γ^i .. sequence of isoform i

Sequence R_n

$$P(g, s, o, r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(o_n|g_n)P(s_n|g_n)P(r|g_n, s_n, o_n)$$

$P(R_n = \rho|G_n = i, S_n = j, O_n = k)$

- strand specific protocol, known isoforms:

$$P(R_n = \rho|G_n = i, S_n = j, O_n = 0) = \prod_{t=1}^{L} \omega_t(\rho_t, \gamma^i_{j+t-1})$$

Alignment of read and isoform:

			C	G	A	T			
A	T	C	C	G	A	A	T	C	G

$$P(R_n = \rho|G_n = i, S_n = j, O_n = 0) = \omega_1(C, C)\omega_2(G, G)\omega_3(A, A)\omega_4(T, A)$$

Sequence R_n

$$P(g, s, o, r|\theta) = \prod_{n=1}^{N} P(g_n|\theta)P(o_n|g_n)P(s_n|g_n)P(r|g_n, s_n, o_n)$$

$P(R_n = \rho|G_n = i, S_n = j, O_n = k)$

- strand specific protocol, known isoforms:

$$P(R_n = \rho|G_n = i, S_n = j, O_n = 0) = \prod_{t=1}^{L} \omega_t(\rho_t, \gamma^i_{j+t-1})$$

$\omega_t(a, b) = P(\text{read}[t] = a|\text{isoform}[j + t - 1] = b)$
γ^i .. sequence of isoform i

- strand specific protocol, noise isoform 0:

$$P(R_n = \rho|G_n = 0, S_n = j, O_n = 0) = \prod_{t=1}^{L} \beta(\rho_t)$$

β .. background distribution

Estimation of Expression Levels

Given: N reads of length L and M known isoforms

Assumption: reads are uniformly sampled from the transcriptome

EM Algorithm: find $\theta = [\theta_0, \ldots, \theta_M]$ that maximizes $P(r|\theta)$

$$P(r|\theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \theta_i \frac{1}{\ell_i} \sum_j P(r_n | g_n = i, s_n = j)$$

$$\nu_i \approx \frac{\theta_i}{1 - \theta_0}$$

Esti

Give

Ass

EM

EM-Algorithm: iteratively optimization of θ

latent variables: G_n, S_n, O_n

E-step:

$$E[G_n = i, S_n = j, O_n = k] = P(G_n = i, S_n = j, O_n = k | r, \theta^t)$$

M-step:

$$\theta^{t+1} = \arg\max_{\theta} E[log(P(r, g_n, o_n, s_n|\theta))|r, \theta^t]$$

$$\nu_i \approx \frac{\theta_i}{1 - \theta_0}$$

Estimation of Expression Levels

Given: N reads of length L and M known isoforms

Assumption: reads are uniformly sampled from the transcriptome

EM Algorithm: find $\theta = [\theta_0, \ldots, \theta_M]$ that maximizes $P(r|\theta)$

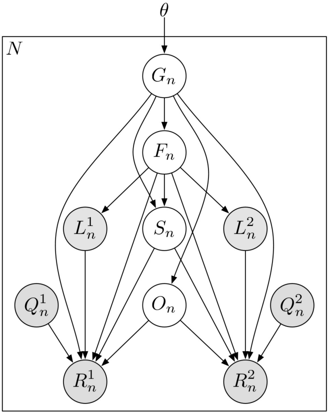
$$P(r|\theta) = \prod_{n=1}^{N} \sum_{i=0}^{M} \theta_i \frac{1}{\ell_i} \sum_{j} P(r_n|g_n = i, s_n = j)$$

$$\nu_i \approx \frac{\theta_i}{1 - \theta_0}$$

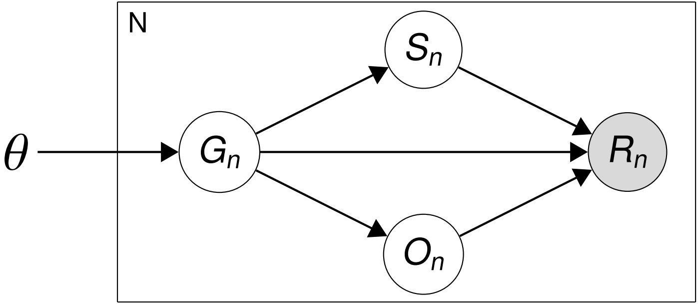
(a)

(b)

Gene expression estimates (y-axis) vs. sample values (x-axis) for the simulated mouse (a) and maize (b) RNA-Seq data sets. Comparisons are given for ν.



θ
N
G_n
F_n
L_n^1
S_n
L_n^2
Q_n^1
O_n
Q_n^2
R_n^1
R_n^2



Thank you for your attention!