A Sequential Monte Carlo Method for estimating Transcriptional Landscape at Basepair level from RNA-Seq data

Bogdan MIRAUTA1, Pierre NICOLAS2 and Hugues RICHARD1

1 Génomique des Microorganismes, UPMC UMR7238, 75005 Paris, France
{bogdan.mirauta,hugues.richard}@upmc.fr
2 Mathématique, Informatique et Génome, INRA UR1077, 78350 Jouy-en-Josas, France
pierre.nicolas@jouy.inra.fr

Keywords RNA-Seq, Sequential Monte Carlo, Hidden Markov model.

1 Introduction

Sequencing technologies applied to transcriptome interrogation (RNA-Seq) permit a high precision in the inference of transcript localization and relative expression level. Namely, RNA-Seq produces millions of reads that, once aligned to the reference genome, provide counts that reflect the transcriptional landscape. This landscape is, even in the presence of post transcription processes, directly correlated to the expression activity. Read counts can serve to estimate, with an existing annotation, gene expression level up to the isoform level [1] assuming homogeneous distribution of the reads inside predefined genome segments. Transcript boundaries can also be identified from abrupt variations in the local abundance of reads using sliding windows. However, realistic probabilistic models that simultaneously account for transcript boundaries and expression levels are still not available to describe RNA-Seq data. This precludes the use of a parametric inference framework to obtain estimate expression at the basepair level.

In this work, we design a strand specific model that includes the changes in expression level along the genome and randomness due to the read sampling. Through Hidden Markov Model formalism this problem reduces to estimating the hidden path for the unobserved variable \( u_t \) - the expression level at basepair \( t \). \( u_t \) is defined as the product between the relative abundance of the position \( t \) and the total number of reads. We then use a Sequential Monte Carlo (SMC) approach [2] to infer hidden expression level \((u_t)_{1:T}\) along a genome sequence of length \( T \) from observed read counts \((y_t)_{1:T}\). Model parameters are estimated in Bayesian framework. The SMC approach allows us to specify a reasonable model that fully exploits the complexity of RNA-Seq data.

2 Model and SMC Algorithm

The Hidden Markov model includes two main dependencies: the hidden chain transition kernel that describes changes in expression level and the emission function that relates counts to expression levels. Our choice for the transition between the hidden states \((u_t)\) aims at including abrupt changes characteristic of transcript boundaries, and smooth variations, arising from technological biases or biological processes. Biological effects could be partial termination or degradation, which generate progressive changes in transcription levels. In keeping with [3] we refer these two types of changes as the shifts and the drifts. The transition kernel writes:

\[
\pi(u_{t+1}, u_t) = I_{\{u_t = 0\}} \cdot \left[ (1 - \eta) \cdot \delta_0(u_{t+1}) + \eta \cdot \frac{1}{c} e^{-\frac{1}{c} u_{t+1}} \right] \\
+ I_{\{u_t > 0\}} \cdot \left[ \alpha \cdot \delta_{u_t}(u_{t+1}) + \beta \cdot \frac{1}{c} e^{-\frac{1}{c} u_{t+1}} + \beta_0 \cdot \delta_0(u_{t+1}) \\
+ \gamma_u \cdot 1_{\{u_{t+1} > u_t\}} \cdot \frac{\lambda_u}{u_t} \cdot e^{-\frac{\lambda_u}{u_t} (u_{t+1} - u_t)} + \gamma_d \cdot 1_{\{u_{t+1} < u_t\}} \cdot \frac{u_t}{u_{t+1}} \cdot \frac{\lambda_d}{u_{t+1}} \cdot e^{-\frac{\lambda_d}{u_{t+1}} (u_t - u_{t+1})} \right]
\]

and is best understood as a mixture of several move types. In expressed regions, expression remains at the same level with probability \( \alpha \); jumps to a non expressed region with probability \( \beta_0 \); changes to a different level exponentially distributed with probability \( \beta \); and can finally drift upward or downward with probabilities \( \gamma_u \) and \( \gamma_d \). Small increases or decreases caused by drifts have percentual amplitudes exponentially distributed
with parameters $\lambda_u$ and $\lambda_d$. A jump from a not expressed position into an expressed region has probability $\eta$. Read counts ($y_t$) are considered independent given ($u_t$). This assumption holds when counting only the first position of the reads. We model the read count $y_t$ as a mixture between a Poisson with expectation $u_t$ [4] and a distribution accounting for technological outliers.

The reconstruction of expression levels is based on a Monte Carlo approach where estimation of $u_{1:T}$ given the observation $y_{1:T}$ relies on the sampling of particles (trajectories) from the target distribution $\pi(u_{1:T} \mid y_{1:T})$. In our context this distribution is complex and $T$ is large. An Importance Sampling scheme is used where sampling is done from a proposal whose outcomes are reweighted. For long sequences, a good importance proposal $q(u_{1:T})$ to approximate $\pi(u_{1:T} \mid y_{1:T})$ is impossible to define directly but the problem remains tractable using a sequential approach [2]. Briefly, at each position $t$ we obtain a sample from $\pi(u_{1:t} \mid y_{1:t})$ using the sample from $\pi(u_{1:t-1} \mid y_{1:t-1})$ obtained at position $t-1$ by drawing $u_t$ and updating importance weights according to the formula $w_t^i = \pi(u_t \mid u_{t-1}^i)\pi(y_t \mid u_t^i)/q_t(u_t^i; u_{t-1}^i)w_{t-1}^i$ where $i$ is the particle index. As $t$ increases, weights $w_t^i$ tend to degenerate leading to a poor estimation. A resampling step is performed avoid this behaviour. The choice of the importance proposal can improve the number of used particles, the resampling frequency and thus the performance of the algorithm. Our proposal $q_t(u_t; u_{1:t-1})$ was designed to approximate $\pi(u_t \mid u_{1:t-1}, y_t)$; this algorithm is a filtering algorithm that provides sample approximating $\pi(u_{1:t} \mid y_{1:t})$ for each $t$ and therefore from our target distribution $\pi(u_{1:T} \mid y_{1:T})$ at time $T$. For large $T$, when looking backward, the trajectories of the particles coalesce due to resampling and thus make impossible a good estimation of $u_t$ for $t \ll T$. To tackle with this problem a backward smoothing is implemented. From the backward sample we can compute both point estimates and credibility intervals of the expression level $u_t$ for each $t$.

3 Results and Discussion

Relevance of this method is illustrated in Fig. 1 by the application on simulated data for low expressed regions. We used a Gibbs algorithm for the estimation of the parameters.

![Figure 1](image)

**Figure 1.** Expression level inference on simulated data. Left panel: simulated expression level (thin line) and read counts (dots). Right panel: estimated expression level (black line) and 95% credibility interval (gray area).

This methodology extends previous work on tiling array data [3]: it introduces a model adapted to RNA-Seq data and it presents an SMC algorithm for estimation of underlying expression levels that overcomes the need for discretization. Better description of the RNA-Seq data is an important step towards disentangling technological artifacts from subtle biological signals. A software is in preparation and will be made available.

References


