**Parseq: reconstruction of microbial transcription landscape from RNA-Seq read counts using state-space models**

Bogdan Mirauta\(^1,\)*, Pierre Nicolas\(^2,\)\(^\dagger\) and Hugues Richard\(^1,\)\(^\dagger\)

\(^1\)Biologie Computationnelle et Quantitative, UPMC and CNRS UMR7238, Paris, France and \(^2\)Mathématique Informatique et Génome, INRA UR1077, Jouy-en-Josas, France

Received on May 3, 2013; revised on November 30, 2013; accepted on January 21, 2014

**1 INTRODUCTION**

Sequencing technologies play an increasing role in the investigation of gene expression [RNA sequencing (RNA-Seq)]. The most common RNA-Seq strategy consists of random shearing, amplification and high-throughput sequencing of the RNA fraction. Methods to analyze transcription level variations along the genome from the read count profiles generated by the RNA-Seq protocol are needed.

Motivation: The most common RNA-Seq strategy consists of random shearing, amplification and high-throughput sequencing of the RNA fraction. Methods to analyze transcription level variations along the genome from the read count profiles generated by the RNA-Seq protocol are needed.

Results: We developed a statistical approach to estimate the local transcriptional levels and to identify transcript borders. This transcriptional landscape reconstruction relies on a state-space model to describe transcription level variations in terms of abrupt shifts and more progressive drifts. A new emission model is introduced to capture not only the read count variance inside a transcript but also its short-range autocorrelation and the fraction of positions with zero counts. The estimation relies on a particle Gibbs algorithm whose running time makes it more suited to microbial genomes. The approach outperformed read-overlapping strategies on synthetic and real microbial datasets.

Availability: A program named Parseq is available at: http://www.lgm.upmc.fr/parseq/

Contact: bodgan.mirauta@upmc.fr

Supplementary information: Supplementary data are available at Bioinformatics online.

Received on May 3, 2013; revised on November 30, 2013; accepted on January 21, 2014
composition on the priming step (Li et al., 2010) and by other pre-sequencing procedures that can introduce biases in read coverage (Griebel et al., 2012; Wu et al., 2010). To tackle these issues, we present a probabilistic model of RNA-Seq count data, which integrates transcription level variation as well as a generic description of the longitudinal variability induced by the sequencing protocol.

This modeling approach builds on previous works, originally motivated by the analysis of comparative genomic hybridization and transcription level variation as well as a generic description of the longitudinal variability induced by the sequencing protocol.

2 A PROBABILISTIC MODEL FOR TRANSCRIPTION LEVELS AND READ COUNTS

2.1 The SSM framework

Throughout this work, we refer to the transcription level at position $t$ of the genome as $u_t$. This level is scaled such that it corresponds to the expectation of the count $y_t$ of reads whose $5’$-ends map at position $t$; it is thus also proportional to the total number of reads sequenced. It cannot be directly equated to the read count $y_t$ because of the randomness of the selection of the sequenced reads and to local variability artifacts. Our aim is to reconstruct the trajectory $u = (u_t)_{t<1}$ from the sequence of read counts $y = (y_t)_{t<1}$. For this purpose, we consider an SSM where $u_t$ is a hidden variable taking values on the real half line $[0; +\infty)$ whose distribution depends on $u_{t-1}$ via a Markov transition kernel and $y_t$ is an observation whose emission distribution depends on $u_t$. This framework allows accounting for the longitudinal dependency between the $u_t$’s and provides great flexibility in the modeling of $y_t$ given $u_t$.

2.2 Longitudinal model of transcriptional level

The Markov transition kernel $k(u_{t+1}; u_t)$ that we use distinguishes expressed ($u_t > 0$) and non-expressed ($u_t = 0$) regions and assigns a positive probability for unchanged transcription level between $t$ and $t + 1$. The allowed changes of transcription levels between $t$ and $t + 1$ breaks down into distinct types: jump from between expressed and non-expressed regions as well as changes of transcription level within transcribed regions—accounting for transcription initiation and termination sites in presence of overlapping transcription units. Following the work on tiling array data (Nicolas et al., 2009), changes within transcribed regions further subdivide into two types that differ by their amplitudes and are referred as shifts (large amplitude) and drifts (small amplitude). Coexistence of shifts and drifts is designed to pull apart well-defined initiation or termination sites internal to transcribed regions from smoother changes in measured transcriptional levels that can have a biological origin (e.g. random termination events) or can reflect technical artifacts (e.g. longitudinal bias caused by messenger RNA capture and fragmentation protocols).

The Markov transition kernel $k(u_t; u_{t-1})$ for transcriptional level writes

$$\begin{align*}
I_{[u_{t-1} = 0]}[(1 - \eta)\delta_0(u_t) + \eta f(u_t)] + I_{[u_{t-1} > 0]}[a\delta_{u_{t-1}}(u_t) + \beta f(u_t) + \beta_0 \delta_0(u_t)] + \gamma_d g_d(u_t; u_{t-1}, \lambda) + \gamma_d^0 g_d(u_t; u_{t-1}, \lambda)],
\end{align*}$$

where $I_{\cdot}$ denotes the indicator function that serves to indicate whether $t - 1$ is an expressed or non-expressed position, and $\delta_x$ denotes the Dirac delta function with mass at point $x$ that gives a non-zero probability for unchanged transcription level and for jumping to 0 between $t - 1$ and $t$. The parameters $\eta \in (0, 1)$ and $(a, \beta, \beta_0, \gamma_d, \gamma_d^0) \in (0, 1)^3$ with $a + \beta + \beta_0 + \gamma_d + \gamma_d^0 = 1$ define the probabilities of the different types of moves. The terms $f(u_t; \cdot)$, $g_d(u_t; u_{t-1}, \lambda)$ and $g_d^0(u_t; u_{t-1}, \lambda)$ are probability densities for the transcription level $u_t$ at the beginning of a transcribed region (occurring with probability $\eta$ when $u_{t-1} = 0$) or after a shift (probability $\beta$ when $u_{t-1} > 0$), after an upward drift (probability $\gamma_d$ when $u_{t-1} > 0$) and after a downward drift (probability $\gamma_d^0$ when $u_{t-1} > 0$), respectively. The density $f(u_t; \cdot)$ corresponds to an exponential distribution of rate $\lambda$ (mean $1/\lambda$) and the parameter $\lambda > 0$ defines the average relative change caused by drifts: $(u_t - u_{t-1})/u_{t-1}$ if upward drift or $(u_{t-1} - u_t)/u_{t-1}$ if downward drift.

2.3 Distribution of read counts in real datasets

The variability of read counts observed when resequencing the same library has been described as almost compatible with a Poisson distribution (Marioni et al., 2008). However, when compared between samples (or even replicate libraries), it exhibits overdispersion, and the negative binomial (NB) distribution is often used to accommodate this behavior (Anders and Huber, 2010; Robinson et al., 2010). Initially, we planned to rely also on the NB to account for read counts overdispersion between positions inside each transcript. It seems required to involve a mixed Poisson distribution to account simultaneously for the incompressible variance of the final sampling by sequencing (Poisson) and for the extra-variability introduced by randomness in library preparation and by position-specific biases that can be introduced at all steps of the protocols. In this context, the NB is viewed as a gamma–Poisson mixture $\gamma \sim \text{Poisson}(\lambda z); z \sim \text{Gamma}(1, \varphi)$ stands as the most tractable model (Karlis and Xekalaki, 2005).

Based on two real datasets, we examined the distribution of read counts inside regions expected to be homogeneous in terms of expression level. Namely, we asked whether the NB could respond to the expectation of the count distribution on the priming step (Li et al., 2010) and by other pre-sequencing procedures that can introduce biases in read coverage (Griebel et al., 2012; Wu et al., 2010). To tackle these issues, we present a probabilistic model of RNA-Seq count data, which integrates transcription level variation as well as a generic description of the longitudinal variability induced by the sequencing protocol.
discrepancy between the data and the NB is with respect to the zero counts: given the mean and the variance of the empirical distribution, the fraction of positions with zero counts under the NB assumption tends to be too low for low expression levels and too high for high expression levels.

The usual parametrization of the NB with overdispersion parameter $\phi$ mentioned above is also contradicted by the data. The variance increases markedly faster than the mean $u$ even for low expression level, in sharp contrast with the prediction that the variance should write $u + \phi u^2$. In the Poisson-mixture context, breaking these relationships that arise from the law of total variance should write variance increases markedly faster than the mean.

To circumvent these problems, we used a recently described SMC method known as particle Gibbs (PG) that makes it possible to obtain exact answers in the context of classical HMMs need to be substituted by particle filtering algorithms build on sequential Monte Carlo (SMC) principles whose results are only approximate for finite numbers of particles (Doucet and Johansen, 2008). Parameter inference that relies heavily on hidden trajectory reconstruction in this category of models is also directly impacted. Here, the existence of a second hidden variable $s(t)$ and the sequence lengths ranging in millions of base pairs increase the difficulty.

3 TRANSCRIPTIONAL LANDSCAPE RECONSTRUCTION

3.1 Markov chain Monte Carlo with particle Gibbs

In SSMs, the reconstruction of the hidden trajectory, given the parameter values and the observed data (here the characterization of $u(t, \Theta)$), is more challenging than in a classical HMM where only discrete values are considered for the hidden variable. The forward–backward recursions that provide exact answers in the context of classical HMMs need to be substituted by particle filtering algorithms build on sequential Monte Carlo (SMC) principles whose results are only approximate for finite numbers of particles (Doucet and Johansen, 2008). Parameter inference that relies heavily on hidden trajectory reconstruction in this category of models is also directly impacted. Here, the existence of a second hidden variable $s(t)$ and the sequence lengths ranging in millions of base pairs increase the difficulty.

To circumvent these problems, we used a recently described SMC method known as particle Gibbs (PG) that makes it possible to obtain exact (but correlated) joint samples of the hidden trajectory and parameters, given the data (Andrieu et al., 2010). PG is based on a modified SMC step, the conditional SMC (CSMC), that is integrated into more general Markov chain Monte Carlo (MCMC) algorithms for Bayesian inference of the parameters. In this setup, this also allowed to combine the reconstructions of $u(s, y, \Theta)$ and $s(u, y, \Theta)$ to obtain a joint
reconstruction of \( u, s | y, \theta \) and to extract the marginal of interest \( u | y, \theta \). We also implemented an additional PG step intended to preserve \( s \times u \) by rescaling \( s \) when updating \( u \). The problem posed by sequence length was properly handled within the PG framework by successive partial (block) CSMC updates of the hidden trajectories. To validate the implementation of our PG algorithm, we extended the algorithm to sample the joint \((s, u, y, \theta)\) distribution and verified that we could retrieve the priors. Detailed descriptions of the parameter priors, MCMC and validation procedures used in this work are provided in Supplementary Sections S2 and S3.

3.2 The Parseq workflow

In theory, our PG algorithm permits to tackle parameter estimation and transcriptional landscape reconstruction simultaneously, but our software Parseq subdivides the problem in three successive steps for practical reasons (Fig. 2). The parameters of the read count emission model are estimated, and the emission density corresponding to the different values of \( u, s, y \) are tabulated (step 1). PG iterations are too time-consuming to be performed on a single CPU for genomes of moderate sizes such as the yeast *Saccharomyces cerevisiae* (~12 Mb). To distribute computation on independent CPUs, we decided to subdivide each chromosome in fragments (~1 Mb each), to perform parameter estimation separately on these fragments, and then to select a common set of parameters based of the obtained results (step 2). Posterior sampling of transcriptional landscape trajectories \( u \) is then carried out on a different CPU for each genome fragment, but with common parameters (step 3). With an Intel Core i7-3610QM CPU @ 2.30 GHz, each complete sweep of the MCMC algorithm was recorded to take ~1 min for 1 Mb using 150 particles in each CSMC update. In this study, we used 2200 sweeps, including 200 burn-in sweeps, for parameter estimation (step 2), and 2200 sweeps for making predictions at fixed parameters (step 3). Thus, on multi-CPU computers, the complete procedure takes slightly <3 days for each dataset with this algorithm setup, which we currently recommend for applications.

The output of the algorithm is a sample of transcriptional landscape trajectories drawn from \( u \sim y, \theta \) that conveys rich information about the actual transcriptional landscape. Here, these trajectories served to estimate the expected value of \( u \), the 95% credibility interval of \( u \), and the probability of \( u > 0 \) (transcribed position), together with the probability of the different types of breakpoints along the sequence. Because of the posterior uncertainty on the exact position of each breakpoint, we further aggregated the breakpoint probabilities at adjacent positions into small regions with high cumulative probabilities using a local-score approach. The weighted center of each small region was taken as a point estimate of the position of the breakpoint and the cumulative probability served as a confidence measure. According to the direction of the change in expression level, the breakpoints were identified as upshifts or downshifts. To better distinguish genuinely expressed regions from (biological or technological) background noise, we also realized the relevance of computing the probability for \( u \) to be above a selected cutoff and to predict the breakpoints that lead the trajectory \( u \) above this cutoff. Details on the workflow, including parameter estimation and post-processing, are provided in Supplementary Sections S4. The accuracy of transcriptional landscape reconstruction was assessed from two different standpoints: the number of transcribed positions that can be correctly called based on the estimated value of \( u \), the number of transcript 5' and 3'-ends at <50 bp of an identified upshift and downshift, respectively. To establish the lists of predictions, we used a probability cutoff set to 0.5 for both the probability of \( u > 0 \) and the cumulative probability shift in the small region delineated by local-score approach. When comparing the predictions with a reference annotation, we needed to take into account that Parseq models the distribution of the 5'-end of the reads. For this reason, the regions predicted as transcribed by Parseq were extended of \( l_s \) bp on their 3'-ends, and the same correction needs to be applied to the predicted downshifts before comparing with transcript 3'-ends (adjusted to 50 bp for the simulated dataset). To report results in terms of sensitivity and positive predictive values (PPV) we computed the fraction of the true positives that could be matched to a prediction and the fraction of the predictions that could be matched to a true positive. Parseq predictions were systematically compared with the results of Cufflinks v2.1.1 (Trapnell et al., 2010), a method for transcript assembly that is based on read overlapping.
The results obtained on synthetic data are summarized in Figure 4. While both Parseq and Cufflinks perform well when the depth of sequencing exceeds an average of 0.12 reads/bp, below this level the differences between the two methods become evident. Even though they do not have the same sensitivity-specificity trade-off, it appears clearly that the results obtained by Parseq are better. The model-based approach adopted in Parseq makes it possible to extrapolate transcription across coverage gaps and this results in a better calling of transcribed positions (not shown) and transcript borders. The mechanistic interpretation of our new emission model is also well supported by the results: Parseq estimation of the amplification coefficient \( \mu_c \) distinguishes remarkably well the two scenarios considered in our simulations where sequencing depth increases either as a consequence of higher amplification or as a consequence of higher number of initial molecules sampled.

4.2 Evaluation on real data

On synthetic data, both the model-based approach of Parseq and the read-overlapping approach of Cufflinks perform well at detecting transcribed positions and transcript borders once the sequencing depth becomes high enough (0.12 reads/bp in our simulations). However, despite the efforts made on the simulation pipeline to mimic the different types of artifacts, the synthetic data do not have the complexity of a real dataset.

For evaluation on real data, we chose strand-specific single-end datasets from two major model microorganisms: the yeast *S.cerevisiae* and the bacterium *Escherichia coli*. The *S.cerevisiae* dataset was sequenced on a SOLiD platform (Short Read Archive identifier SRR121907) and published in a study on regulatory non-coding RNAs (van Dijk et al., 2011). It has a read length of 50 bp and a sequencing depth of 1.6 reads/bp after mapping. The *E.coli* dataset (SRR794838) was sequenced on an Illumina platform and published together with the
presentation of the Rockhopper workflow for bacterial RNA-Seq data processing (McClure et al., 2013). It has a read length of 100 bp and a sequencing depth of 2.4 reads/bp after mapping.

As a reference annotation for the transcribed positions in \textit{S.cerevisiae}, we relied on the 5874 coding sequences (CDSs) found in the \textit{S.cerevisiae} database SGD (Cherry et al., 2012) and lists of untranslated regions (UTRs) mapped from RNA-Seq experiments in Yassour et al. (2009) (5200 5'UTRs and 5295 3'UTRs). To better assess the accuracy of the prediction of transcripts 5'- and 3'-ends, we also included comparison with experimental data that aimed at mapping precisely these sites: 4393 transcriptional start sites (TSSs) (Zhang, 2005), and 7977 polyadenylation sites (pAs) (Ozsolak et al., 2010). For \textit{E.coli}, we used annotations available in the RegulonDB database (Salgado et al., 2013) (2438 promoters and 2647 operons) and also the sequence-based predictions of 2260 rho-independent transcription terminators obtained with Petrin software (d’Aubenton Carafa et al., 1990).

Table 1 presents a detailed breakdown of the results according to the different sets of reference annotations, which could be considered to assess accuracy. In this context, we found that the probability of $\omega_i > 0$ (expression cutoff 0°) is not necessarily the most relevant to compare the prediction of transcribed positions with a reference annotation. The best trade-offs are obtained near 0.1 reads/bp on the \textit{S.cerevisiae} dataset, and 0.25 reads/bp on the \textit{E.coli} dataset. These values are in agreement with the presence of a large number of positions associated with low expression level, resembling a background noise (Supplementary Figs S6 and S7). The accuracy of the detection of transcribed position is remarkable (e.g. 83% sensitivity, 90% PPV with the 0.1 reads/bp expression cutoff on \textit{S.cerevisiae}), but similar to Cufflinks (Table 1). In keeping with our observations on synthetic data, this suggests that detecting transcribed positions is easy at high sequencing depth, and consequently, the model-based approach implemented in Parseq provides only small benefits.

The accurate identification of transcript borders is by far more challenging. For instance, on \textit{S.cerevisiae} 5'-ends, with the same 0.1 reads/bp expression cutoff, the sensitivity reaches 64% and the PPV 48%. On \textit{E.coli}, PPVs remain acceptable, but sensitivity values are much lower. This could be due to a combination of the following: lower quality of the data ($\mu_A$ estimated to 6.15 in \textit{E.coli} versus 1.18 in \textit{S.cerevisiae}), adjusted $l_3$ is 50 bp for \textit{S.cerevisiae} versus 160 bp for \textit{E.coli}; lower quality of the annotation taken as reference (e.g. Petrin predictions are expected to contain substantial numbers of false positives and false negatives); higher proportion of genes with low or no expression and thus for which promoters and terminators cannot be detected (with the 0° expression cutoff, sensitivity for detection of transcribed regions is only 0.81 in \textit{E.coli} versus 0.91 in \textit{S.cerevisiae}). On both datasets and for 5'-ends and 3'-ends alike, Parseq results are consistently better than the ones obtained by Cufflinks, particularly in terms of sensitivity. This confirms our expectations, as Cufflinks reconstruction ignores the possibility of overlapping transcripts and thus overlooks transcript-ends in these configurations. We also included in our comparison the predictions made on \textit{E.coli} by Rockhopper (Table 1).

As we were interested here in \textit{de novo} predictions but this software could not run without annotations, we discarded successively the annotation on one-tenth of the genome and recorded the predictions on it. Parseq and Cufflinks provide results markedly better than Rockhopper in this comparison setup.

### 4.3 Importance of drift and local scaling

Transcript borders are detected on the basis of significant changes in read counts. Therefore, high variability in read counts can lead to breakpoint overpredictions resulting in a loss of specificity when not properly incorporated in the model. We pulled this need by introducing two different components in our model: a drift term on the transition kernel for progressive variations as opposed to the abrupt changes modeled by shifts,

### Table 1. Detection of transcribed positions and transcript borders on \textit{S.cerevisiae} (SRR121907) and \textit{E.coli} (SRR794838) datasets

<table>
<thead>
<tr>
<th>Features</th>
<th>\textit{S.cerevisiae}</th>
<th>\textit{E.coli}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Reference</td>
<td>Parseq</td>
</tr>
<tr>
<td>Transcripts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sensitivity</td>
<td>CDSs and UTRs</td>
<td>0.83 (0.91)</td>
</tr>
<tr>
<td>PPV</td>
<td>CDSs and UTRs</td>
<td>0.90 (0.68)</td>
</tr>
<tr>
<td>5' End Number</td>
<td></td>
<td>6689 (8353)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>TSSs</td>
<td>0.64 (0.65)</td>
</tr>
<tr>
<td>PPV</td>
<td>TSSs and 5'UTRs</td>
<td>0.48 (0.4)</td>
</tr>
<tr>
<td>3' End Number</td>
<td></td>
<td>6287 (7440)</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>pAs</td>
<td>0.60 (0.62)</td>
</tr>
<tr>
<td>PPV</td>
<td>pAs and 3'UTRs</td>
<td>0.57 (0.51)</td>
</tr>
</tbody>
</table>

Predictions and reference data were matched based on a ±50 bp distance cutoff (for a ±25 bp distance cutoff, see Supplementary Table S3). Outside parentheses: results obtained after applying a stricter expression cutoff. \textit{S.cerevisiae}: 0.1 reads/bp for Parseq, 100 fragments per transcript for Cufflinks. \textit{Escherichia coli}: 0.25 reads/bp cutoff for Parseq, 200 fragments per transcript for Cufflinks, $z = 0.2$ for Rockhopper. Between parentheses: 0° reads/bp for Parseq, 5 fragments per transcript for Cufflinks and $z = 0.01$ for Rockhopper.
and a local scaling Markov-dependent variable $s$ intended to capture short-range autocorrelations. By monitoring the accuracy in terms $5'$- and $3'$-ends detection, we assessed the effect of these two model components on the quality of the inference. The results are reported in Table 2 and confirm that taken individually the drift and the local scaling improve the results. Moreover, the results also demonstrate that the two terms are complementary rather than redundant, as their combination leads to further improvements.

5 CONCLUSION

This article presents a model-based approach for analyzing the RNA-Seq read count profiles along the genome. The model aims for an account of artificial longitudinal variability's sources, via a new model of overdispersion able to capture not only the variance versus mean relationships but also the fraction of zero counts and the short-range autocorrelations. From a methodological standpoint, our work also demonstrates the feasibility of analyzing genome-scale data within the framework of SSMs. The recently described PG algorithm (Andrieu et al., 2010) was instrumental in this success. Running time does not depend on the depth of the sequencing, but is proportional to genome length, which makes it more suited to microbial genomes.

The method outperforms a read assembly approach at low sequencing depth, and shows a clear improvement on real data even for high sequencing depth. We believe that the availability of such a tool will become increasingly useful as the use of RNA-Seq becomes more popular. In particular, the availability of confidence scores and credibility intervals will be relevant to build a reference annotation from a compendium of experiments as done from tiling array data (Nicolas et al., 2012) and also to compare global RNA-Seq profiles with the results of protocols targeting more specifically the sequencing of transcript ends (Lin et al., 2013; Pelechano et al., 2013). In principle, one could also envision to extend the model for situations where data would be collected with distinct RNA-Seq protocols on the same biological sample.

ACKNOWLEDGEMENTS

The authors are grateful to Adeline Samson and Claude Thermes for initial discussions, respectively, on the inference framework and on the RNA-Seq technology. The authors also thank Alessandra Carbone for her support and the reviewers for their stimulating comments.

Funding: Ministère de l'Enseignement Supéérieure et de la Recherche, doctoral fellowship and teaching assistantship (B.M.).

Conflict of Interest: none declared.

REFERENCES


