Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
category

OXFORD

Subject

Investigation of RNA metabolism using pulseR

Uvarovskii Alexey [1,2]*, Christoph Dieterich [1,2]

[1] Section of Bioinformatics and Systems Cardiology Klaus Tschira Institute for Integrative Computational Cardiology Department of Internal Medicine III University Hospital Heidelberg, Im Neuenheimer Feld 669 69120 Heidelberg, and [2] German Center for Cardiovascular Research (DZHK), Im Neuenheimer Feld 669 69120 Heidelberg

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Results: Availability: Contact: alexey.mipt@gmail.com **Supplementary information:** Supplementary data are available at *Bioinformatics* online.

1 Introduction

Gene expression is a dynamic process. Experiments which aim to measure kinetics of expression levels in time can help to reveal mechanisms of gene regulation. With new experimental set-ups, new types of generated data occur and one needs to develop new methods and frameworks to be able to extract the knowledge out of the results.

Pulse-chase experimental approach allow to track changes of mRNA due to synthesis of nascent RNA molecules and degradation of the old ones. The idea of the method is to introduce a label, which incorporates in the newly synthesised RNA molecules and can be traced later. 4sU labelling introduced by [Dölken et al., 2008] allowed to estimate kinetic rates of RNA metabolism in a number of studies up to date [Sabò et al., 2014, Rabani et al., 2011, Miller et al., 2011, Schwanhäusser et al., 2011, Eser et al., 2016, Schwalb et al., 2016, Marzi et al., 2016, Zhang et al., 2016, Neymotin et al., 2014, Mukherjee et al., 2016].

reviewed in [Wachutka and Gagneur, 2016]

2 Methods

2.1 Kinetic model

First-order reaction kinetics is one of approaches to describe gene expression [Kærn et al., 2005]. Given

- constant synthesis rate s and
- degradation rate d,

RNA concentration r follows the ordinary differential equation

$$\dot{r} = s - dr, \tag{1}$$

where $\dot{r}$ stands for the time derivative of the r.

During synthesis, a new RNA molecule incorporates labelled uridine bases **?**. For zero initial condition $r_L(0) = 0$, the solution is

$$r_{\mathrm{L}}(t) = \frac{s}{d}\left(1 - e^{-dt}\right). \tag{2}$$

With time, the labelled fraction tends to the steady state concentration level μ,

$$\lim_{t\to\infty} r_{\mathrm{L}}(t) = \frac{s}{d} = \mu = r_{\mathrm{T}}, \tag{3}$$

where r_{T} stands for the total fraction. In contrast, the unlabelled molecules are only being degraded during the *pulse*-experiment. Hence, assuming initial level of unlabelled RNA to be the steady-state one, $r_U = \mu$, the the amount of unlabelled fraction at a time t is

$$r_{\mathrm{U}}(t) = \mu e^{-dt}. \tag{4}$$

The example model includes only two parameters and does not consider RNA maturation and existence of several isoforms. For more complex approaches we refer to **?**.

For completeness we provide the formulas, which describe expression levels for *chase*-experiments. In this case, we assume that no synthesis of labelled RNA occurs after the labelling period t_L:

$$r_{\mathrm{T}} = \mu \tag{5}$$

$$r_{\mathrm{L}} = \mu\left(1 - e^{-dt_{\mathrm{L}}}\right)e^{-dt_{\mathrm{C}}} \tag{6}$$

$$r_{\mathrm{U}} = \mu\left(1 - \left(1 - e^{-dt_{\mathrm{L}}}\right)e^{-dt_{\mathrm{C}}}\right), \tag{7}$$

where t_{C} stands for the longitude of the chase period. If t_{L} is assumed to be long enough, the "chase"-equations above are reduced to

$$r_{\mathrm{T}} = \mu \tag{8}$$

$$r_{\mathrm{L}} = \mu e^{-dt_{\mathrm{C}}} \tag{9}$$

$$r_{\mathrm{U}} = \mu\left(1 - e^{-dt_{\mathrm{C}}}\right) \tag{10}$$

similar to the "pulse"-equations.

2.2 Statistical model

In RNA-seq experiments, expression level is represented by read number. To model such data, we use the negative binomial (NB) distribution, because it was shown to successfully describe over-dispersed RNA-seq data [Robinson and Smyth, 2007]. This type of distribution has two parameters, mean m and dispersion parameter α. Hence, a read number of gene i in a sample j reads

$$K_{ij} \sim \mathrm{NB}(m_{ij}, \alpha). \tag{11}$$

Here we assume, that the dispersion parameters α is shared between all samples and genes, because replicate numbers used in practice are very small and it's not possible to infer about several parameters from only 2 or 3 points.

2.2.1 Normalisation

We normalise the samples in two stages:

1. inside one fraction, e.g. between samples corresponding to "labelled, 12 hr" measurement
2. between the fractions, e.g. how read numbers in the total fraction relate to the read numbers in "labelled, 12hr".

We perform the inside-fraction normalisation as described in [Anders and Huber, 2010]. The normalisation sample-specific coefficient s_j is the median of ratios gene read number to the geometric mean along samples in this fraction F:

$$s_j = \underset{i \in \text{genes}}{\text{median}} \frac{K_{ij}}{\underset{f \in F}{\text{geometric mean}}\, K_{if}}. \tag{12}$$

If spike-in molecules are present in the samples, then one computes s_j using spike-in read numbers:

$$s_j = \underset{i \in \text{spike-ins}}{\text{median}} \frac{K_{ij}}{\underset{f \in F}{\text{geometric mean}}\, K_{if}}. \tag{13}$$

No between-fraction normalisation is needed in this case, since RNA quantities in different fractions can be related via spike-in data only.

For between-fraction normalisation, we introduce fraction-specific coefficients n_f as model parameters, e.g. n_{total} and $n_{\text{label-12hr}}$. The parameter values are identified during the fitting procedure (next section).

Finally, the mean m_{ij} of the read number distribution is factorised as:

$$m_{ij} = [\text{inside-fraction}] \times [\text{between-fractions}] \times r_{if} = s_j n_f r_{if}, \tag{14}$$

where r_{if} stands for the expression level of the gene i in a fraction f, e.g. r_L from the eq. 2.

An additional relations between fractions can be introduced in the form of cross-contamination coefficient. For example, if the labelled fraction r_{L} is assumed to be contaminated with the unlabelled one r_{U}, the expected read number is modified as

$$r^*_{\mathrm{L}} = (1 - c) r_{\mathrm{L}} + c r_{\mathrm{U}} \tag{15}$$

Such normalisation approach may look overcomplicated from one hand. However, the inside-fraction normalisation helps to reduce the number of model parameters.

2.2.2 Fitting

In order to estimate parameters of the model, we use the maximum-likelihood method. We separated the fitting procedure into several simpler steps:

1. fitting of gene-specific parameters (e.g. degradation rate and expression level);
2. fitting of parameters shared between fractions (e.g. cross-contamination coefficient);
3. fitting of the between-sample normalisation factors n_f (if no spike-ins are provided) and estimation of the dispersion parameter α.

We repeat the steps 1-3 until user-specified convergence criteria is not met. Since gene-gene interactions are not considered by the model, it is possible to fit this parameters independently in parallel.

To optimise the likelihood functions, we use implementation of the L-BFGS-U method [Byrd et al., 1995] available in the `stats` R package [R Core Team].

In the paper, we visualise fitting results using `ggplot2` package [Wickham, 2009].

3 Results

3.1 Implementation

We implement a framework in order to simplify parameter estimation for pulse-chase RNA-seq experiments. The source code is available at https://github.com/dieterich-lab/pulseR. We use *R* programming environment, because it allows very flexible handling of user-specified models, includes implementations a variety of commonly used procedures and has powerfull plotting features [R Core Team].

A user needs to provide

- a count table with the raw read numbers
- a condition matrix (to infer sample fractions in the count table)
- formulas for the mean read numbers, e.g. $r_{\mathrm{L}} \sim \mu e^{-dt}$, (section 2.1).
- spike-ins, if relevant

The package utilises so-called *computing on the language*, a feature of the *R* language [R Core Team, 2000]. This allows to handle arbitrary formulas, which the user supplies for the mean gene expression levels. Hence, no manual implementation for the likelihood functions is necessary and the formulas are needed to defined only once.

3.2 Evaluation using simulated data

To demonstrate the analysis workflow, we generated a data set on the basis of the pulse-model, introduced in the section 2.

$$r_{\mathrm{i,T}} = \mu_i \tag{16}$$

$$r_{\mathrm{i,L}} = \mu_i \left(1 - e^{-d_i t_{\mathrm{L}}}\right) \tag{17}$$

$$r_{\mathrm{i,U}} = \mu_i e^{-d_i t_{\mathrm{L}}}, \tag{18}$$

where $t_{\mathrm{L}} = 12$ hr is the time amount of pulse-labelling, $i \in 1..100$ is a gene index, μ_i and d_i are gene-specific parameters, which are sampled from random number generator. Usually concentration of RNA in the samples is normalised before the sequencing procedure []. This introduces a bias, which we model by fraction coefficients in section 2.2.1. For this data set, the fraction-specific normalisation factors are n_{total}=1 (reference fraction), n_{label}=2, n_{unlab}=3.

On the basis of this parameters, we sampled read counts from the NB distribution with the dispersion parameter $\alpha = 0.01$, in 3 replicates for every fraction ((total, label, unlab) $\times$ (replicates 1,2,3)). Using this simulated count data, we evaluated the performance of the developed `pulseR` package.

To improve the fitting procedure, we provided the initial guess for the expression level on the basis of the median read count in the total fraction. The initial values for the degradation rates were chosen as random numbers.

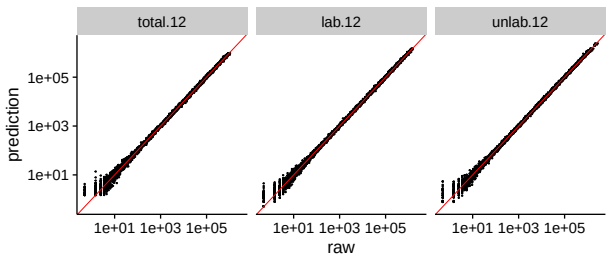

Fig. 1. Comparison of the model predictions to the generated raw counts for 3 different fractions (total, labelled, unlabelled) at after 12 hr of pulse-experiment.

Fig. 2. Comparison of the fitted and true parameter values (A, B) and relative errors in the parameter fit (C, D) for the expression level μ and the degradation rate d.

The predictions for every gene in three different fractions demonstrate good quality of fitting, fig. 1. The relative errors of the parameter fits is worse for the lower expressed genes, fig. 2. That is expected to occur, because the dispersion of the NB-distributed values is higher for the data with a lower mean.

4 Discussion

4.1 Comparison to existing approaches

The published so far approaches are different in terms of data normalisation, statistical model and underlying mathematical model of the RNA metabolism.

Normalisation

The normalisation step aims to uncover relations between samples, which can be different in sequencing depth, represent different fractions and were subjected different protocols of preparation. Spike-in molecules, which one adds before fraction separation step, allow to simplify the fitting procedure. In this case, estimation of fraction-specific coefficients and cross-contamination rates is separated from fitting of gene-specific parameters [Schwalb et al., 2016, Neymotin et al., 2014].

However, it comes with the price of complication of the exprerimental procedure, because one needs to synthesis labelled spike-ins molecules []. Alternative approach is to determine fraction- or sample- specific normalising coefficients on the basis of the gene expression data only. In most cases, the analysis is performed on transformed raw read counts, i.e. to RPKMs (Reads Per Kilobase of transcript per Million).

- **normalisation to RNA quantities:** the total amount of RNA is measured for every fraction. This ratios of RNA quantity between different fractions are used as fraction-specific coefficients [Rabani et al., 2011].
- **normalisation via regression:** Schwanhäusser et al. [2011] used an approach which was previously applied to the microarray data in [Dölken et al., 2008]. The total fraction T is related to the labelled L and unlabelled U ones as $T = aU + bL$, where a and b are unknown. Ordinary regression recovers the coefficients a and b. However, such procedure results in unrealistic fraction ratios, and Schwanhäusser et al. [2011] had to exclude certain genes from the analysis.
- CHANGE ME total least-squares [Miller et al., 2011]
- **normalisation coupled with model fitting:** coefficients can be introduced as model parameters. Their values are recovered by MLE together with other parameters [Eser et al., 2016, De Pretis et al., 2015]. Although being harder to implement, this approach allows parameters to influence on each other, and hence, can result in better estimations.

Our package can be used either with spike-in or spike-in free data. Additional known or unknown shared parameters can be introduced by a user as well, which allows high flexibility in the experiment analysis.

Statistical model

The NB distribution is known to describe well the count data in RNA-seq experiments [Robinson and Smyth, 2007]. This encourages to use it over normal distribution assumption, which was applied earlier to the microarray data sets [Miller et al., 2011]. In our package, we implement NB assumptions, which were successfully applied to short pulse-labelling in [Eser et al., 2016, Schwalb et al., 2016].

The normal distribution is implemented in `INSPEcT` R package [De Pretis et al., 2015], `DRiLL` software [Rabani et al., 2014], `HALO` Java framework [Friedel et al., 2010] and `DTA/cDTA` R packages [Schwalb et al., 2012].

- no delays in the model (transcr-transl)

Acknowledgements

Funding

money

References

Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Stefano De Pretis, Theresia Kress, Marco J Morelli, Giorgio EM Melloni, Laura Riva, Bruno Amati, and Mattia Pelizzola. Inspect: a computational tool to infer mrna synthesis, processing and degradation dynamics from rna-and 4su-seq time course experiments. *Bioinformatics*, 31(17):2829–2835, 2015.

Lars Dölken, Zsolt Ruzsics, Bernd Rädle, Caroline C Friedel, Ralf Zimmer, Jörg Mages, Reinhard Hoffmann, Paul Dickinson, Thorsten Forster, Peter Ghazal, et al. High-resolution gene expression profiling for simultaneous kinetic parameter analysis of rna synthesis and decay. *Rna*, 14(9):1959–1972, 2008.

Philipp Eser, Leonhard Wachutka, Kerstin C Maier, Carina Demel, Mariana Boroni, Srignanakshi Iyer, Patrick Cramer, and Julien Gagneur. Determinants of rna metabolism in the schizosaccharomyces pombe genome. *Molecular systems biology*, 12(2):857, 2016.

Caroline C Friedel, Stefanie Kaufmann, Lars Dölken, and Ralf Zimmer. HaloâŁ"a java framework for precise transcript half-life determination. *Bioinformatics*, 26(9):1264–1266, 2010.

Mads Kærn, Timothy C Elston, William J Blake, and James J Collins. Stochasticity in gene expression: from theories to phenotypes. *Nature Reviews Genetics*, 6(6):451–464, 2005.

Matteo J Marzi, Francesco Ghini, Benedetta Cerruti, Stefano de Pretis, Paola Bonetti, Chiara Giacomelli, Marcin M Gorski, Theresia Kress, Mattia Pelizzola, Heiko Muller, et al. Degradation dynamics of microrna revealed by a novel pulse-chase approach. *Genome research*, 26(4):554–565, 2016.

Christian Miller, Björn Schwalb, Kerstin Maier, Daniel Schulz, Sebastian Dümcke, Benedikt Zacher, Andreas Mayer, Jasmin Sydow, Lisa Marcinowski, Lars Dölken, et al. Dynamic transcriptome analysis measures rates of mrna synthesis and decay in yeast. *Molecular systems biology*, 7(1):458, 2011.

Neelanjan Mukherjee, Lorenzo Calviello, Antje Hirsekorn, Stefano de Pretis, Mattia Pelizzola, and Uwe Ohler. Integrative classification of human coding and noncoding genes through rna metabolism profiles. *Nature Structural & Molecular Biology*, 2016.

Benjamin Neymotin, Rodoniki Athanasiadou, and David Gresham. Determination of in vivo rna kinetics using rate-seq. *Rna*, 20(10): 1645–1652, 2014.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `https://www.R-project.org`.

R Core Team. R language definition. *Vienna, Austria: R foundation for statistical computing*, 2000.

Michal Rabani, Joshua Z Levin, Lin Fan, Xian Adiconis, Raktima Raychowdhury, Manuel Garber, Andreas Gnirke, Chad Nusbaum, Nir Hacohen, Nir Friedman, et al. Metabolic labeling of rna uncovers principles of rna production and degradation dynamics in mammalian cells. *Nature biotechnology*, 29(5):436–442, 2011.

Michal Rabani, Raktima Raychowdhury, Marko Jovanovic, Michael Rooney, Deborah J Stumpo, Andrea Pauli, Nir Hacohen, Alexander F Schier, Perry J Blackshear, Nir Friedman, et al. High-resolution sequencing and modeling identifies distinct dynamic rna regulatory strategies. *Cell*, 159(7):1698–1710, 2014.

Mark D Robinson and Gordon K Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

Arianna Sabò, Theresia R Kress, Mattia Pelizzola, Stefano De Pretis, Marcin M Gorski, Alessandra Tesi, Marco J Morelli, Pranami Bora, Mirko Doni, Alessandro Verrecchia, et al. Selective transcriptional regulation by myc in cellular growth control and lymphomagenesis. *Nature*, 511(7510):488–492, 2014.

Björn Schwalb, Daniel Schulz, Mai Sun, Benedikt Zacher, Sebastian Dümcke, Dietmar E Martin, Patrick Cramer, and Achim Tresch. Measurement of genome-wide rna synthesis and decay rates with dynamic transcriptome analysis (dta). *Bioinformatics*, 28(6):884–885, 2012.

Björn Schwalb, Margaux Michel, Benedikt Zacher, Katja Frühauf, Carina Demel, Achim Tresch, Julien Gagneur, and Patrick Cramer. Tt-seq maps the human transient transcriptome. *Science*, 352(6290):1225–1228, 2016.

Björn Schwanhäusser, Dorothea Busse, Na Li, Gunnar Dittmar, Johannes Schuchhardt, Jana Wolf, Wei Chen, and Matthias Selbach. Global quantification of mammalian gene expression control. *Nature*, 473 (7347):337–342, 2011.

Leonhard Wachutka and Julien Gagneur. Measures of rna metabolism rates: Toward a definition at the level of single bonds. *Transcription*, (just-accepted):00–00, 2016.

Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. ISBN 978-0-387-98140-6. URL `http://ggplot2.org`.

Yang Zhang, Wei Xue, Xiang Li, Jun Zhang, Siye Chen, Jia-Lin Zhang, Li Yang, and Ling-Ling Chen. The biogenesis of nascent circular rnas. *Cell reports*, 15(3):611–624, 2016.