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

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.

© The Author 2015. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

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

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

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].

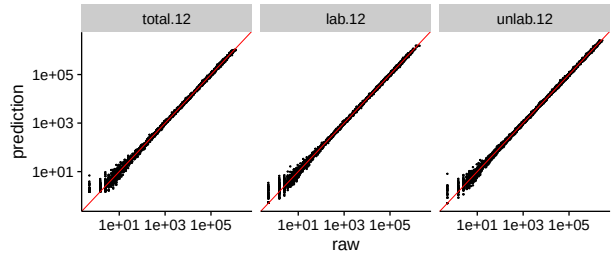

Fig. 1. Model predictions.

3 Results

3.1 Implementation

We implement a framework in order to simplify parameter estimation for pulse-chase RNA-seq experiments. 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. $r_{\text{L}} \sim \mu e^{-dt}$, (section 2.1).
- spike-ins, if relevant

The package utilizes so-called *non-standard evaluation*, 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 on simulated data

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

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

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

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

where $t_{\text{L}} = 12$ 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 $\alpha = 100$, , 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 `pulseR` package.

4 Discussion

- no delays in the model (transcr-transl)

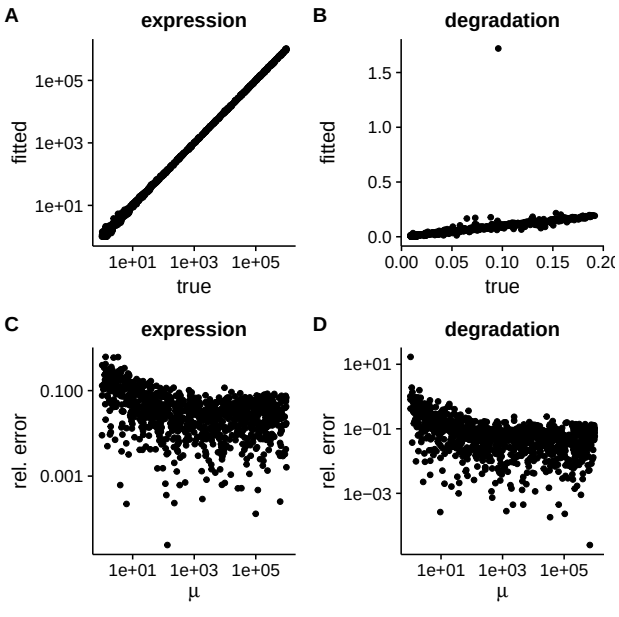

Fig. 2. Parameter misprediction.

Acknowledgements

thank you

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.

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.

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.

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