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

Subject

Investigation of RNA metabolism using pulseR

Uvarovskii Aleksei [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,

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

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.

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

$$K_{ij} \sim \mathrm{NB}(m_{ij}, \alpha). \tag{8}$$

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 fraction, e.g. how read numbers in the total fraction relate to the read numbers in "labelled, 12hr".

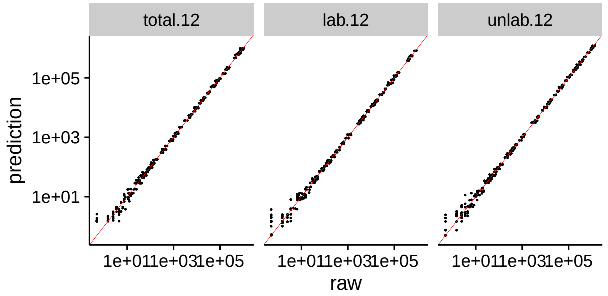

Fig. 1. Model predictions.

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:

$$s_j = \underset{i \in \text{genes}}{\text{median}} \frac{K_{ij}}{\underset{f \in F}{\text{geometric mean}}\, K_{if}}. \tag{9}$$

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

$$s_j = \underset{i \in \text{spike-ins}}{\text{median}} \frac{K_{ij}}{\underset{f \in F}{\text{geometric mean}}\, K_{if}}. \tag{10}$$

For between-fraction normalisation we introduce fraction-specific normalisation coefficients n_f as a model parameters, e.g. n_{total} and $n_{\text{label-12hr}}$. Finally, the mean m_{ij} of the read number distribution is as:

$$m_{ij} = [\text{inside-fraction}] \times [\text{between-fractions}] \times r_{if} = s_j n_f r_{if}, \tag{11}$$

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

2.2.2 Fitting

3 Results

results

4 Discussion

- no delays in the model (transcr-transl)

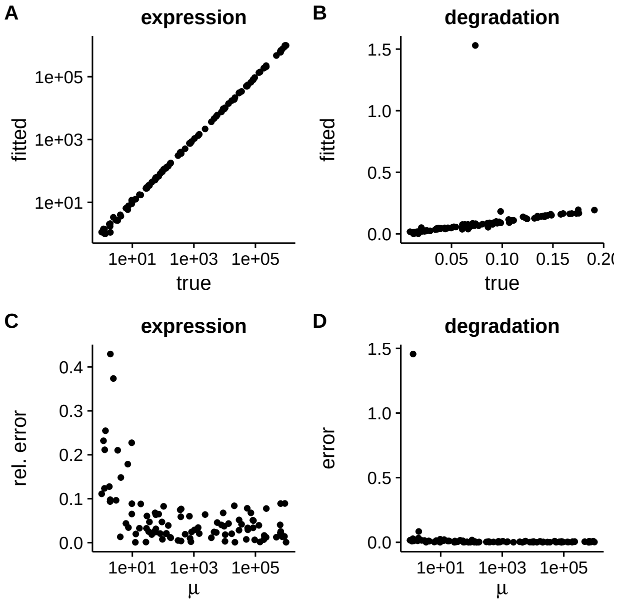

Fig. 2. Parameter misprediction.

Acknowledgements

thank you

Funding

money

References

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

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.

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