THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Power series method for solving TASEP-based models of mRNA translation

Citation for published version:
Scott, S & Szavits Nossan, J 2019, 'Power series method for solving TASEP-based models of mRNA translation', *Physical Biology*, vol. 17, no. 1. https://doi.org/10.1088/1478-3975/ab57a0

Digital Object Identifier (DOI):
10.1088/1478-3975/ab57a0

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Physical Biology

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 06. Aug. 2024

Physical Biology

OFFICIAL JOURNAL OF
Sociedad de Biofísicos Latino Americanos

ACCEPTED MANUSCRIPT

Power series method for solving TASEP-based models of mRNA translation

To cite this article before publication: Simon Scott *et al* 2019 *Phys. Biol.* in press https://doi.org/10.1088/1478-3975/ab57a0

Manuscript version: Accepted Manuscript

Accepted Manuscript is "the version of the article accepted for publication including all changes made as a result of the peer review process, and which may also include the addition to the article by IOP Publishing of a header, an article ID, a cover sheet and/or an 'Accepted Manuscript' watermark, but excluding any other editing, typesetting or other changes made by IOP Publishing and/or its licensors"

This Accepted Manuscript is **© 2019 IOP Publishing Ltd**.

During the embargo period (the 12 month period from the publication of the Version of Record of this article), the Accepted Manuscript is fully protected by copyright and cannot be reused or reposted elsewhere.
As the Version of Record of this article is going to be / has been published on a subscription basis, this Accepted Manuscript is available for reuse under a CC BY-NC-ND 3.0 licence after the 12 month embargo period.

After the embargo period, everyone is permitted to use copy and redistribute this article for non-commercial purposes only, provided that they adhere to all the terms of the licence https://creativecommons.org/licences/by-nc-nd/3.0

Although reasonable endeavours have been taken to obtain all necessary permissions from third parties to include their copyrighted content within this article, their full citation and copyright line may not be present in this Accepted Manuscript version. Before using any content from this article, please refer to the Version of Record on IOPscience once published for full citation and copyright details, as permissions will likely be required. All third party content is fully copyright protected, unless specifically stated otherwise in the figure caption in the Version of Record.

View the article online for updates and enhancements.

This content was downloaded from IP address 129.215.73.92 on 18/11/2019 at 12:24

Power series method for solving TASEP-based models of mRNA translation

S Scott[1], **J Szavits-Nossan**[1]

[1] SUPA, School of Physics and Astronomy, University of Edinburgh, Peter Guthrie Tait Road, Edinburgh EH9 3FD, United Kingdom

E-mail: jszavits@staffmail.ed.ac.uk

Abstract. We develop a method for solving mathematical models of messenger RNA (mRNA) translation based on the totally asymmetric simple exclusion process (TASEP). Our main goal is to demonstrate that the method is versatile and applicable to realistic models of translation. To this end we consider the TASEP with codon-dependent elongation rates, premature termination due to ribosome drop-off and translation reinitiation due to circularisation of the mRNA. We apply the method to the model organism *Saccharomyces cerevisiae* under physiological conditions and find excellent agreements with the results of stochastic simulations. Our findings suggest that the common view on translation as being rate-limited by initiation is oversimplistic. Instead we find theoretical evidence for ribosome interference and also theoretical support for the ramp hypothesis which argues that codons at the beginning of genes have slower elongation rates in order to reduce ribosome density and jamming.

Keywords: protein synthesis, messenger RNA, translation, exclusion process, TASEP, steady state, power series
Submitted to: *Phys. Biol.*

Accepted Manuscript

1. Introduction

Translation of a mRNA sequence into a protein is central to normal cell function. How is this process carried out and controlled in the cell is a topic of major interest not only from the standpoint of understanding protein function and regulation, but also for the possibility of making adjustments to the genetic code that would improve yields of foreign and synthetic proteins.

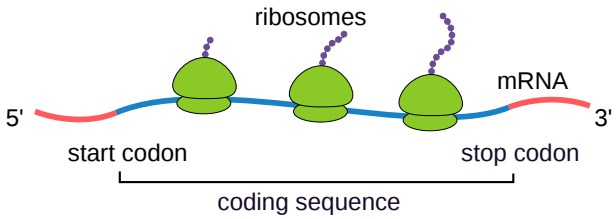

Figure 1. A schematic picture of mRNA translated by ribosomes in the $5' \to 3'$ direction.

Translation is performed by ribosomes that move along the mRNA from the 5' end to the 3' end (Figure 1). The process can be split into three main stages: initiation, elongation and termination. During initiation, the ribosome assembles on a portion of the mRNA before the coding sequence and moves to the start codon where the first amino acid is added to the ribosome. Elongation begins when the ribosome moves to the second codon with a newly amino acid attached to the protein chain. This process is repeated codon by codon until the ribosome encounters the stop codon and detaches itself from the mRNA along with a newly produced protein.

Mathematical modelling of translation has a long history in mathematics, physics and biology. Most of the models that are in use today are based on a model introduced by MacDonald, Gibbs and Pipkin in 1968 [1, 2] and independently by Spitzer in 1970 [3]. Spitzer, who was interested in a much broader class of interacting random walks, is also responsible for naming the model the exclusion process due to excluded-volume interactions between the random walkers. The full name of the process relevant to mRNA translation is the totally asymmetric simple exclusion process or TASEP; "totally asymmetric" means that random walkers (ribosomes) move unidirectionally on a discrete lattice (mRNA) and "simple" means that they move one lattice site (codon) at a time.

In physics, the TASEP is one of the simplest models belonging to a broad class of *driven diffusive systems* [4]. These systems are of great interest because they do not attain thermal equilibrium, even when they settle in the steady state. The question of how to describe nonequilibrium steady states is one of the biggest open questions in statistical physics. For the TASEP in which each random walker occupies one lattice site and moves forward at a constant speed this problem was solved in full by Derrida, Evans, Hakim and Pasquier [5] and Schütz and Domany [6], both in 1993. The exact solution described in detail the nature of phase transitions previously discovered by Krug [7], which sparked a great interest in the model.

Unfortunately, most TASEP-based models which are of interest to modelling translation cannot be solved using techniques developed in Refs. [5, 6]. These models account for the correct ribosome length (approximately the length of 10 codons) [8], variable ribosome speed that depends on the codon being translated [9], elongation consisting of several intermediate steps [10], nonsensical errors such as premature termination [11, 12], translation reinitiation due to mRNA circularisation [11, 13–15] and many more (for a recent review see Ref. [16]).

A fundamental question in molecular biology is how the mRNA codon sequence affects the translation process and in particular the rate of protein production [17, 18]. In the TASEP the rate of protein production corresponds to the current of ribosomes leaving the stop codon. If we assume that each of 61 codon types‡ is translated at a different speed, this leaves us with 61 parameters describing elongation and two parameters describing initiation and termination, and that is only for the basic model. Using stochastic simulations alone in order to understand how these parameters affect the translation process is a difficult, if not a formidable task. A different approach is needed.

In previous work [19], Szavits-Nossan, Ciandrini and Romano developed a mathematical method for solving the TASEP with codon-dependent elongation that accounted for tRNA delivery and ribosome translocation [20]. The main idea was to express the steady-state solution as a power series expansion in the translation initiation rate.

‡ The remaining three codons are stop codons that do not code for an amino acid.

Accepted Manuscript

It is a common view in cell biology that translation initiation is the rate-limiting step in protein production under nutrient-rich growth conditions. Existing estimates of translation initiation and elongation rates in various organisms support this view. For example, translation initiation rate in *Escherichia coli* has been estimated to 1 initiation every 3 seconds [21], compared to the elongation rates of individual codons that are in the range of $4-40$ amino acids per second [22]. Translation initiation rates of *Saccharomyces cerevisiae* genes were found to be in the range of 1 initiation every $0.2-200$ seconds with the median value of 1 initiation every 11 seconds. Hence most of the genes initiate translation at a much slower rate than the elongation rates of individual codons, which were estimated to be in the range of $1-35$ amino acids per second.

These numbers may change significantly under low-nutrient conditions such as amino acid starvation leading to much smaller elongation rates, in which case the power series method may not be applicable. We will discuss this point in more detail later in the text.

In the present study, we apply the power series method to the TASEP that accounts for codon-dependent elongation rates, premature termination due to ribosome drop-off and translation reinitiation due to mRNA circularisation. The last two features of the model were chosen to show that the method is versatile and practical to use for studying more realistic models of translation and more features may be added to this model in the future. We test the method on the model organism *Saccharomyces cerevisiae* and find an excellent agreement with the results of stochastic simulations.

2. Methods

2.1. The basic TASEP-based model of translation

We model mRNA as one-dimensional lattice consisting of L codons labelled from 1 (start codon) to L (stop codon) that code for $L-1$ amino acids. We assume that each ribosome occupies $\ell = 10$ codons [25] and that the ribosome P and A sites are positioned at the fifth and sixth codon respectively, measured from the ribosome's trailing end.

Translation initiation is a multi-step process which is different in prokaryotic and eukaryotic cells. We model translation initiation as a one-step process occurring at rate α in which a new ribosome is recruited at the start codon so that its P-site and A-site are positioned at the first and second codon, respectively. This one-step process thus encompasses both prokaryotic and eukaryotic translation initiation mechanisms.

During elongation, a ribosome at codon i receives an amino acid from the corresponding tRNA and translocates to the next codon at rate ω_i, provided there is no ribosome at codon $i+\ell$. Translation terminates once a ribosome A-site reaches the stop codon, releases the polypeptide chain and unbinds from the mRNA at rate β. For each codon $i = 2, \ldots, L$ we define the corresponding ribosome occupancy number $\tau_i \in \{0, 1\}$,

$$\tau_i = \begin{cases} 1 & \text{if codon } i \text{ is occupied by a ribosome A-site} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

These numbers uniquely determine the configuration of the system which we denote by $C = \{\tau_2, \ldots, \tau_L\}$. Using this notation, kinetic steps in translation can be summarised as:

$$\text{(initiation): } \tau_2 = 0 \xrightarrow{\alpha} 1 \text{ if } \tau_2 = \ldots = \tau_{\ell+1} = 0 \quad (2a)$$

$$\text{(elongation): } \tau_i, \tau_{i+1} = 1, 0 \xrightarrow{\omega_i} 0, 1 \text{ if } \tau_{i+\ell} = 0 \quad i = 2, \ldots, L-1 \quad (2b)$$

$$\text{(termination): } \tau_L = 1 \xrightarrow{\beta} 0. \quad (2c)$$

Equations $(2a)$-$(2c)$ constitute the basic model of mRNA translation proposed by MacDonald, Gibbs and Pipkin in 1968 [1].

2.2. More realistic models

In addition to the basic model we also consider premature termination by ribosome drop-off and translation reinitiation due to mRNA circularisation. Ribosome drop-off is a translational error which results in the ribosome being released from the mRNA along with a non-functional polypeptide that is targeted for degradation. We model ribosome drop-off as a one-step process in which a ribosome at codon $i = 2, \ldots, L-1$ unbinds from the mRNA at rate μ,

$$\text{(ribosome drop-off): } \tau_i = 1 \xrightarrow{\mu} 0 \quad (2d)$$

for $i = 2, \ldots, L-1$. Translation reinitiation is a mechanism by which the ribosome that just finished translation may pass directly from the 3' end to the 5' and initiate another round of translation (see [15] and references therein). This is made possible by interactions between the two ends of the mRNA resulting in a mRNA circularisation [26], also known as the closed-loop model (for a recent review see Ref. [27]. In the lack of more details about the exact translation reinitiation mechanism, we consider the simplest one-step process in which a ribosome recognises the stop codon, releases the polypeptide chain and reinitiates translation at rate γ,

$$\text{(translation reinitiation): } \tau_2, \tau_L = 0, 1 \xrightarrow{\gamma} 1, 0 \quad \text{if } \tau_2 = \ldots \tau_{\ell+1} = 0. \quad (2e)$$

A schematic picture of the steps $(2a)$-$(2e)$ is presented in Fig. 2.

Accepted Manuscript

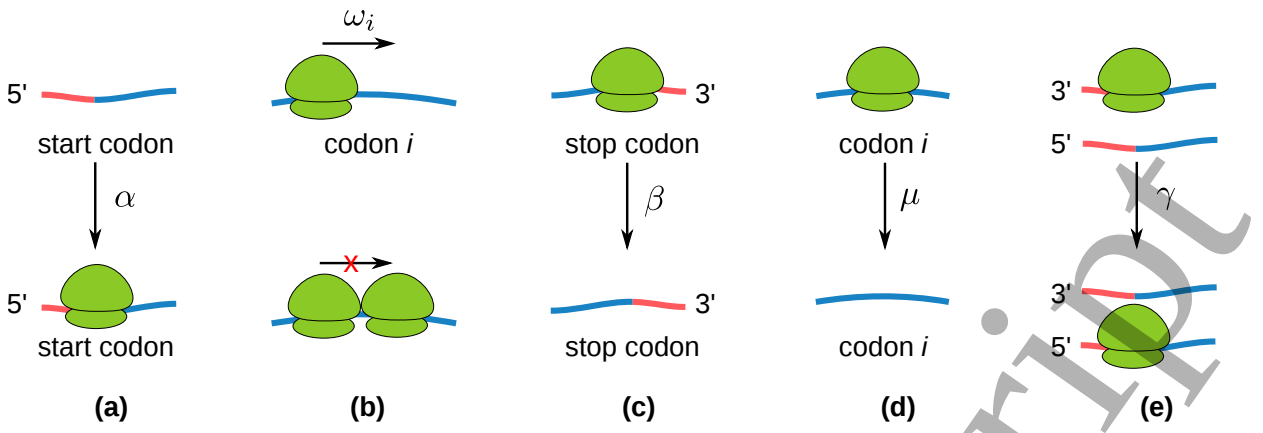

Figure 2. A schematic picture of all the kinetic steps included in the model along with their corresponding rates: (a) initiation (rate α), (b) elongation (codon-specific rate ω_i), (c) termination (rate β), (d) ribosome drop-off (rate μ) and (e) reinitiation (rate γ).

There are other details that we do not consider here. For example, we consider translation initiation to be a one-step process and thus we do not discriminate between prokaryotic and eukaryotic translation. This can be corrected by including more steps involved in translation initiation, which would help to elucidate determinants of translation initiation rate, for example its dependence on initiation factors and mRNA secondary structures upstream of the start codon [28]. Unfortunately, the rate constants involved in translation initiation steps are in general not known and have to be inferred from experiments, which is a difficult task [29].

Another possible extension of the basic model is to include more steps involved in the elongation cycle [30–32]. The power series method is applicable to such models, which was demonstrated in Ref. [19] on a two-step elongation cycle that accounts for tRNA delivery to the ribosome A-site followed by translocation [20, 23]. The method presented in this paper can be also applied to a recently proposed mechanism of premature termination caused by ribosome collisions [33, 34].

An important detail of mRNA translation that we do not consider here are mRNA secondary structures downstream of the start codon. All mRNA secondary structures must be unfolded by the ribosome, which can slow or even stop its progress along the mRNA. Some of these pauses are programmed by 'slippery' sequences such as AAAAAAG leading to beneficial frameshifting [35]. Recent experiments have greatly elucidated the mechanism by which a ribosome passes through the mRNA secondary structures [36–38], which could serve as a basis for building more realistic models of translation (for an early model that accounted for translation of mRNA secondary structures see Ref. [39]). In principle such details can be studied with the present method but the calculations may become cumbersome due to large number of parameters.

2.3. Ribosome current and density

Our goal is to compute the rate of protein synthesis J and ribosome (A-site) density ρ_i. The rate of protein synthesis J is equal to the total current of ribosomes leaving the stop codon,

$$J = \beta\langle\tau_L\rangle + \gamma\left\langle \tau_L \prod_{i=2}^{\ell+1}(1-\tau_i)\right\rangle. \tag{3}$$

Here the first term is due to termination and the second term is due to translation reinitiation. Th current J is not conserved across the coding mRNA (unless we ignore premature termination) and is different from the current of ribosomes initiating translation

$$J_{\rm in} = \alpha\left\langle \prod_{i=2}^{\ell+1}(1-\tau_i)\right\rangle + \gamma\left\langle \tau_L \prod_{i=2}^{\ell+1}(1-\tau_i)\right\rangle. \tag{4}$$

For the rest of the codons the ribosome current (number of ribosomes moving from codon i to codon $i+1$ per second) is given by

$$J_i = \omega_i\left\langle \tau_i \prod_{j=i+1}^{i+\ell}(1-\tau_j)\right\rangle,\ i = 2,\ldots,L-1. \tag{5}$$

Other important observables are the ribosome (A-site) density ρ_i at codon i and the average density ρ defined as

$$\rho_i = \langle\tau_i\rangle, \tag{6}$$

$$\rho = \frac{1}{L-1}\sum_{i=2}^{L}\rho_i. \tag{7}$$

The averaging $\langle\ldots\rangle$ in Eqs. (3)-(7) is taken with respect to the steady-state probability $P(C)$ to find the

Accepted Manuscript

Table 1. List of TASEP parameters for *S. cerevisiae*.

Parameter	Variable	Value	Reference
number of codons	L	25–4093	Ref. []
ribosome size	ℓ	10 codons	Ref. [25]
initiation rate	α	0.005–4 s^{-1}	Ref. [23]
elongation rate	ω_i	1–16 s^{-1}	Ref. [23]
termination rate	β	35 s^{-1}	-
drop-off rate	μ	$1.4 \cdot 10^{-3}$ s^{-1}	Ref. [40]
reinitiation rate	γ	-	-
reinitiation efficiency	η	0–1	-

system in a configuration C,

$$\langle \ldots \rangle = \sum_C (\ldots) P(C) = \tag{8}$$

$$= \sum_{\tau_2=0,1} \cdots \sum_{\tau_{L+1}} (\ldots) P(\tau_2, \ldots, \tau_{L+1}). \tag{9}$$

The steady-state probability $P(C)$ satisfies a master equation,

$$0 = \sum_{C'} W(C' \to C) P(C') - \sum_{C'} W(C \to C') P(C), \tag{10}$$

where $W(C \to C')$ denotes the rate of transition from configuration $C = \{\tau_2, \ldots, \tau_L\}$ to $C' = \{\tau_2', \ldots, \tau_L'\}$.

2.4. Model parameters

In this paper we study *S. cerevisiae* as a model organism using model parameters presented in Table 1.

Translation initiation rates were obtained in Ref. [23] by matching a theoretical prediction for the total density to the density obtained from polysome profiling experiments [24]. We note that the TASEP-based model used to estimate initiation rates in Ref. [23] is different from the TASEP-based models we consider here. Because our main goal here is to assess the applicability of the power series method, we use the same values for initiation rates as in Ref. [23], but note that these may be different from the true (physiological) values. Codon-specific translation elongation rates ω_i were computed according to

$$\omega_i = \frac{k_i r_{\text{trans}}}{k_i + r_{\text{trans}}}, \tag{11}$$

where k_i is the tRNA delivery rate for the amino acid corresponding to codon i and $r_{\text{trans}} = 35$ codons/s is the rate of ribosome translocation [41]. The values of k_i are assumed to be proportional to tRNA gene copy numbers and were taken from Ref. [23]. The rate of termination is assumed to be large and not limiting for translation; for that purpose we set $\beta = \gamma = 35$ s^{-1}. The rate of ribosome drop-off is assumed to be the same as for *E. coli*, whose value was estimated at $1.4 \cdot 10^{-3}$ s^{-1} in Ref. [40]. We are not aware of any estimates of the reinitiation rate γ in the literature. Instead we introduce a new parameter $0 \le \eta \le 1$ that we call reinitiation efficiency,

$$\eta = \frac{\gamma}{\gamma + \beta}, \quad \gamma = \frac{\eta\beta}{1 - \eta} \tag{12}$$

which measures the value of γ relative to the total termination rate $\gamma + \beta$. For example, $\eta = 0$ and $\eta = 1$ correspond to $\gamma = 0$ and $\gamma \to \infty$, respectively.

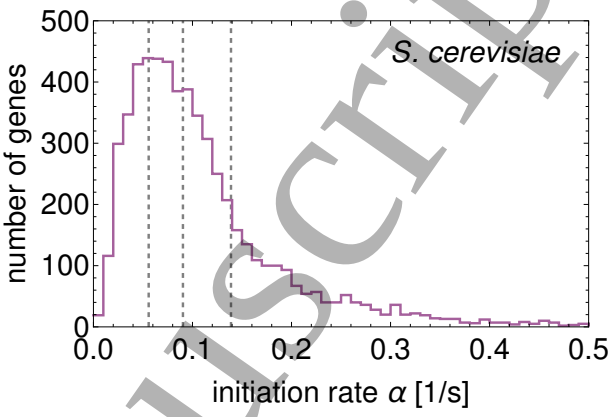

Figure 3. Distribution of translation initiation rates for the *S. cerevisiae* genome taken from Ref. [23]. Vertical dashed lines are quartile values 0.05578, 0.09037 and 0.13889.

2.5. Power series method

The power series method, previously developed in Refs. [19, 42], represents $P(C)$ as a power series in the translation initiation rate α,

$$P(C) = \sum_{n=0}^{\infty} c_n(C) \alpha^n, \tag{13}$$

where $c_n(C)$ are unknown coefficients that depend on configuration C and other rates. Here we summarise the main idea behind this expansion. We first note that the master equation (10) is a linear system of equations in which the variables are the steady-state probabilities $P(C)$,

$$M\mathbf{P} = \mathbf{0}, \tag{14}$$

where $\mathbf{P}$ is a column vector made of all $P(C)$ and M is a square matrix whose matrix elements $M(C, C')$ are given by

$$M(C, C') = \begin{cases} W(C' \to C), & C \neq C' \\ -\sum_{C'' \neq C} W(C \to C''), & C = C'. \end{cases} \tag{15}$$

The solution of this system is given by the following expression

$$P(C) = \frac{\det M^{(C,C)}}{\sum_{C'} \det M^{(C',C')}}, \tag{16}$$

Accepted Manuscript

where $M^{(C,C)}$ is a matrix derived from M by removing the row and the column that correspond to the position of $P(C)$ in the column vector $\mathbf{P}$. For example, if $P(C)$ is the second element in the column vector $\mathbf{P}$ for a given C, then we obtain $M^{(C,C)}$ by removing the second row and the second column from M. Next, we note that M is made of all the transition rates α, all ω_i, β, γ and δ. Therefore $\det M^{(C,C)}$ is a *multivariate polynomial* in all these rates and so is the denominator in (16). This means that $P(C)$ is a multivariate rational function of all the rates α, all ω_i, β, γ and δ and as such can be expanded around zero in any of these rates. However, the power series method will be useful only if the following two criteria are met: (1) the expansion parameter is small so that we can approximate the series with the first few terms and (2) these terms are easy enough to find. We argued in the Introduction that the translation initiation rate α is a good candidate for the first criterion, due to a common view in molecular biology that the translation initiation is rate-limiting for translation. Later in this Section we show that the second criterion is also met.

What happens if we expand $P(C)$, for example, in the drop-off rate δ? This rate is also thought to be small, and therefore meets the first criterion. However, it does not meet the second criterion, because the zero-order term ($\delta = 0$) in the expansion of $P(C)$ is unknown - it corresponds to the basic model with codon-dependent elongation and reinitiation but without premature termination. So although we can expand $P(C)$ around $\delta = 0$, we cannot compute the zero-order term and therefore the method is not useful for finding $P(C)$.

Another possibility is to expand $P(C)$ in one of the elongation rates, say ω_i. For example, if we starve the cell with an amino acid that corresponds to the i-th codon, the elongation rate ω_i may become smaller than the initiation rate α. This makes ω_i a better choice for the expansion parameter than α. However, expanding $P(C)$ in ω_i does not meet the second criterion. The zero-order term ($\omega_i = 0$) is easy to find, which corresponds to a long queue upstream of the i-th codon. However, finding higher-order terms turns out to be a difficult problem. Our method is not applicable to such conditions. In the rest of the paper we expand $P(C)$ in α and assume that all the elongation rates $k_2, \ldots, k_{L-1}$ and the termination rate β are larger than the initiation rate α:

$$\text{(assumption): } k_2, \ldots, k_{L-1}, \beta < \alpha. \tag{17}$$

From the fact that all $P(C)$ must sum to 1, we immediately get that

$$\sum_C c_n(C) = \begin{cases} 1, & n = 0 \\ 0 & n \geq 1. \end{cases} \tag{18}$$

Assuming the initiation rate α to be small allows us to approximate series expansion of $P(C)$ by the first K terms (13)

$$P(C) \approx c_0(C) + c_1(C)\alpha + \ldots + c_K(C)\alpha^K. \tag{19}$$

It needs to be emphasised that keeping only a finite number of terms may lead to significant errors when the rate of initiation is high. This in turn may lead to non-physical values of $P(C) < 0$ or $P(C) > 1$. Of course if that happens the method is not applicable for that choice of α and one has to compute higher-order terms.

In order to find $c_n(C)$, we insert the power series (13) back into the master equation (10) and collect all the terms that contain α^n. These terms must all sum to zero because the left hand side of the stationary master equation (10) is equal to zero. Before we write down a general expression for $c_n(C)$ we need to distinguish between $W(C \to C') = \alpha$ and $W(C \to C') \neq \alpha$. For that purpose we introduce an indicator function $I_{C,C'}$ defined as

$$I_{C,C'} = \begin{cases} 1 & C \to C' \text{ is an initiation event} \\ 0 & \text{otherwise.} \end{cases} \tag{20}$$

This allows us to write $W(C \to C')$ as

$$\begin{aligned} W(C \to C') &= \alpha I_{C,C'} + W(C \to C')(1 - I_{C,C'}) \\ &= \alpha I_{C,C'} + W_0(C \to C') \end{aligned} \tag{21}$$

where $W_0(C \to C') = (1 - I_{C,C'})W(C \to C')$. Inserting $P(C)$ from (13) into (10) and equating the sum of all terms containing α^n to 0 gives the following equation for $c_n(C)$ for $C \neq \emptyset$

$$\begin{aligned} c_n(C) = \frac{1}{e(C)} \Bigg(&\sum_{C'} W_0(C' \to C) c_n(C') \\ &+ \sum_{C'} c_{n-1}(C') I_{C',C} - c_{n-1}(C) \sum_{C'} I_{C,C'} \Bigg), \end{aligned} \tag{22}$$

where $e(C)$ is the total exit rate from C excluding initiation

$$e(C) = \sum_{C'} W_0(C \to C'). \tag{23}$$

For $C = \emptyset$ we can use Eq. (18) instead which gives

$$c_n(\emptyset) = \delta_{n,0} - \sum_{C' \neq \emptyset} c_n(C'). \tag{24}$$

The equation (22) applies to $n \geq 1$. For $n = 0$ the equation is simpler and reads

$$e(C)c_0(C) = \sum_{C'} W_0(C' \to C) c_0(C') \tag{25}$$

Notice that (25) is the same as the original master equation in which the rate of initiation is set to 0. If

Accepted Manuscript

there is no initiation then $c_0(C) = 1$ if $C = \emptyset$ and is 0 otherwise,

$$c_0(C) = \begin{cases} 1, & C = \emptyset \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

The power series method can be understood as a perturbation theory in which translation initiation events can be seen as a small "perturbation" of the empty lattice.

An important consequence of (26) is that any $c_n(C)$ for which the index n is smaller than the total number of ribosomes $N(C)$ in C is equal to zero, or alternatively

$$c_n(C) \neq 0 \text{ only if } n \geq N(C) = \sum_{i=2}^{L} \tau_i. \tag{27}$$

This result is not obvious but follows from the Markov chain tree theorem [43] (also known as Schnakenberg network theory in physics [44]). The theorem states that the steady-state probability $P(C)$ in Eq. (16) can be interpreted as a sum over mathematical objects known as spanning trees. For a given graph G consisting of vertices connected by directed edges, a spanning tree rooted at vertex C is a subgraph of G that contains all the vertices of G but only a subset of edges such that there is a *unique* path from C to any other vertex C' of G. For the Markov jump process described by Eq. (10), the vertices of G are configurations C and the directed edges are possible transitions between two configurations C and C', weighted by the corresponding transition rate $W(C \to C')$.

According to the Markov chain tree theorem, $P(C)$ can be written as

$$P(C) = \frac{\sum_{T(C)} w(T(C))}{\sum_{C'} \sum_{T(C')} w(T(C'))}, \tag{28}$$

where $T(C)$ is a spanning tree of G rooted at configuration C and $w(T(C))$ is the product of all the transition rates corresponding to the directed edges contained in the spanning tree $T(C)$. Let us now consider a configuration C and let this configuration has N ribosomes. We may ask: What spanning trees $T(C)$ rooted at C contribute to $P(C)$? By the definition of $T(C)$ there must be a unique path from *any* C' to C; we choose $C' = \emptyset$, which is the empty mRNA. In order to get from $C' =$ to C that has N ribosomes, we have to make at least N initiation transitions (we can make more than N initiations because some ribosome may terminate, either at the stop codon or prematurely). Consequently, the weight $w(T(C))$ cannot have terms with α^n where $n < N$. In other words, $P(C)$ is of order of α^N where N is the number of ribosomes in C, which is equivalent to the claim in Eq. (27). For more details we refer the reader to Ref. [45] in which we proved (27) for the standard TASEP with particles of size $\ell = 1$, but the same arguments pertain to the models studied in this paper.

The result in (27) simplifies the calculation of $c_n(C)$ considerably. For $n = 1$, we only have to consider configurations with one ribosome ($C = 1_i$ for $i = 2, \ldots, L$) or less ($C = \emptyset$). For $n = 2$, only configurations with two ribosomes ($C = 1_i 1_j$, $i = 2, \ldots, L-\ell$, $j = i+\ell, \ldots, L$) or less ($C = 1_i$ for $i = 2, \ldots, L$ and $C = \emptyset$) need to be studied and so on. This simplification is central to the success of the power series method, allowing us to solve many TASEP-based models for which no exact solution is known.

2.5.1. First-order approximation According to (27) we can ignore all configurations with more than one ribosome. Using (22) we get

$$c_1(1_2) = \frac{1}{\omega_2 + \mu} + \frac{\gamma}{\omega_2 + \mu} c_1(1_L) \tag{29a}$$

$$c_1(1_i) = \frac{\omega_{i-1}}{\omega_i + \mu} c_1(1_{i-1}),\ i = 3, \ldots, L-1 \tag{29b}$$

$$c_1(1_L) = \frac{\omega_{L-1}}{\beta + \gamma} c_{L-1}(1_{L-1}) \tag{29c}$$

$$c_0(\emptyset) = \sum_{i=2}^{L-1} \mu c_1(1_i) + \beta c_1(1_L). \tag{29d}$$

Here we adopted a shorter notation in which 1_i denotes a configuration with ribosome at codon i, and the rest of the mRNA is empty. First we solve equations (29*b*) and (29*c*) recursively yielding coefficients $c_1(1_i)$ for $i = 3, \ldots, L$ that depend on $c_1(1_2)$. After that we insert $c_1(1_L)$ back into equation (29*a*) and find $c_1(1_2)$. Once we have found $c_1(1_2)$ we solve the rest of the equations recursively. Altogether the solution is

$$c_1(1_i) = \frac{\prod_{j=2}^{i} \frac{\omega_j}{\omega_j + \mu}}{\omega_i \left(1 - \frac{\gamma}{\beta+\gamma} \prod_{j=2}^{L-1} \frac{\omega_j}{\omega_j + \mu}\right)}, 2 \leq i \leq L-1 \tag{30a}$$

$$c_1(1_L) = \frac{\prod_{j=2}^{L-1} \frac{\omega_j}{\omega_j + \mu}}{(\beta + \gamma)\left(1 - \frac{\gamma}{\beta+\gamma} \prod_{j=2}^{L-1} \frac{\omega_j}{\omega_j + \mu}\right)} \tag{30b}$$

$$c_1(\emptyset) = -\sum_{i=2}^{L} c_1(1_i). \tag{30c}$$

In the last expression we used the property in (18) which says that all first-order coefficients must sum to zero.

2.5.2. Second-order approximation For the second order, $c_2(C) \neq 0$ only if C contains at most two particles. The equations for $c_2(C)$ are more complicated than for $c_1(C)$ and must be solved numerically.

Accepted Manuscript

Before we write the equations, we first introduce Kronecker delta function δ_{ij} and unit step function $\theta[i]$ defined as

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad \theta[i] = \begin{cases} 1 & i \geq 0 \\ 0 & i < 0 \end{cases}. \tag{31}$$

These two functions allows us to write the equations for $c_2(C)$ in a compact form which reads

$$\begin{aligned} c_2(1_i1_j) = &\frac{\delta_{i,2}}{e(1_i1_j)} c_1(1_j) + \frac{\theta[i-3]\omega_{i-1}}{e(1_i1_j)} c_2(1_{i-1}1_j) \\ &+ \frac{\theta[j-i-\ell-1]\omega_{j-1}}{e(1_i1_j)} c_2(1_i1_{j-1}) \\ &+ \frac{\delta_{i,2}\theta[L-\ell-j]\gamma}{e(1_i1_j)} c_2(1_j1_L), \end{aligned} \tag{32}$$

where $e(1_i1_j)$ is the total exit rate from configuration 1_i1_j excluding initiation,

$$\begin{aligned} e(1_i1_j) = &\theta[j-i-\ell-1]\omega_i + (1-\delta_{j,L})\omega_j + \delta_{j,L}\beta \\ &+ \theta[i-\ell-2]\delta_{j,L}\gamma + 2\mu. \end{aligned} \tag{33}$$

Without reinitiation ($\gamma = 0$), $c_2(1_i1_j)$ depends only on $c_2(1_{i-1}1_j)$ and $c_2(1_i1_{j-1})$, except for $i = 2$ for which it also depends on the known coefficient $c_1(1_j)$. The equation (32) for $\gamma = 0$ can be thus solved recursively starting from $i = 2$ and $j = 2+\ell$, for which $c_2(1_21_{\ell+2}) = c_1(1_{\ell+2})/(\omega_{\ell+2} + 2\mu)$, and iterating over $i = 2, \ldots, L-\ell$ and $i+\ell \leq j \leq L$.

This procedure cannot be immediately applied to the model with reinitiation (in which $\gamma > 0$), because $c_2(1_21_j)$ also depends on $c_2(1_j1_L)$ for $\ell+2 \leq j \leq L-\ell$. Instead, the idea is to find coefficients $c_2(1_j1_L)$ independently and insert them back into Eq. (32), which can be then solved as before.

To this end, we start from $i = 2$ and $j = \ell+2$ in which case $c_2(1_21_{\ell+2})$ is a linear combination of $c_1(1_{\ell+2})$ and $c_2(1_{\ell+2}1_L)$,

$$\begin{aligned} c_2(1_21_{\ell+2}) = &\frac{1}{e(1_21_{\ell+2})} c_1(1_{\ell+2}) \\ &+ \frac{\gamma}{e(1_21_{\ell+2})} c_2(1_21_{\ell+2}) \end{aligned} \tag{34}$$

Next, we iterate Eq. (32) over $\ell+3 \leq j \leq L$ for fixed $i = 2$, which can be done explicitly yielding

$$c_2(1_21_j) = \sum_{m=\ell+2}^{j} \left[F_{2,j}^{(m)} c_2(1_m1_L) + G_{2,j}^{(m)} c_1(1_m) \right], \tag{35}$$

where $F_{2,j}^{(m)}$ and $G_{2,j}^{(m)}$ are given by

$$F_{2,j}^{(m)} = \gamma\theta[L-\ell-m] G_{2,j}^{(m)}, \tag{36a}$$

$$G_{2,j}^{(\ell+2)} = \frac{1}{e(1_21_{\ell+2})} \prod_{k=\ell+2}^{j-1} B_k \tag{36b}$$

$$G_{2,j}^{(m)} = \frac{1}{e(1_21_m)} \frac{\prod_{k=\ell+2}^{j-1} B_k}{\prod_{k=\ell+2}^{m-1} B_k}, \; m = \ell+3, \ldots, L, \tag{36c}$$

and $B_k = \omega_k / e(1_21_{k+1})$. If we now choose $j = L$ we get what we were looking for – an equation that contains coefficients $c_2(1_m1_L)$ and $c_1(1_m)$. We can now repeat this procedure for $i = 3$ by iterating over j until we get the equation for $c_2(1_31_L)$, which will again contain $c_2(1_m1_L)$ and $c_1(1_m)$ and so on. At the end of this procedure we will have a linear system of $L-\ell-1$ equations for $L-\ell-1$ coefficients $c_2(1_21_L), \ldots, c_2(1_{L-\ell}1_L)$ that can be solved numerically using standard techniques. Once these coefficients are computed, we can then proceed to iterate Eq. (32) as we did before for the model without reinitiation.

Once all two-particle second order coefficients are computed, we can easily compute the remaining one-particle coefficients $c_2(1_i)$ from the following equations,

$$\begin{aligned} c_2(1_2) = &\frac{1}{\omega_2+\mu} c_1(\emptyset) + \beta c_2(1_21_L) + \frac{\gamma}{\omega_2+\mu} c_2(1_L) \\ &+ \mu \sum_{j=\ell+2}^{L-1} c_2(1_21_j) \end{aligned} \tag{37a}$$

$$\begin{aligned} c_2(1_i) = &\frac{\omega_{i-1}}{\omega_i+\mu} c_2(1_{i-1}) + \theta[L-\ell-i]\beta c_2(1_i1_L) \\ &+ \mu \sum_{j=2}^{i-\ell} c_2(1_j1_i) + \mu \sum_{j=i+\ell}^{L-1} c_2(1_i1_j) \\ &- \theta[i-\ell-2] c_1(1_i), \; i = 3, \ldots, L-1 \end{aligned} \tag{37b}$$

$$\begin{aligned} c_2(1_L) = &\frac{\omega_{L-1}}{\beta+\gamma} c_{L-1}(1_{L-1}) - c_1(1_L) \\ &+ \mu \sum_{j=2}^{L-\ell} c_2(1_j1_L). \end{aligned} \tag{37c}$$

Finally, we can compute $c_2(\emptyset)$ using Eq. (18), which completes the procedure of finding all second-order coefficients $c_2(C)$.

2.5.3. Higher-order approximations. In principle, we can use Eq. (22) to compute $c_n(C)$ for any order n. In practice, we are limited by the amount of computer memory we need for storing these coefficients, which is the only limitation if we ignore translation reinitiation. In the model with translation reinitiation, we are further limited by the size of the linear system that can be solved numerically. In the present work we computed ribosome density up to the fourth order in the model without reinitiation and up to the second order in the model with reinitiation.

2.6. Monte Carlo simulations

All Monte Carlo simulation were performed using the Gillespie algorithm. In the first part of the simulation we checked the total density ρ every $100 \cdot L$ updates until the percentage error between two values of the total density ρ was less than 0.1%. After that we ran

Accepted Manuscript

the simulation for further $M = 10^4 \cdot L$ updates during which we computed the time average of ρ_i defined as

$$\rho_i = \frac{1}{T}\sum_{k=1}^{M} \tau_i^{(k)} \Delta t^{(k+1)}, \tag{38}$$

where $\tau_i^{(k)}$ is the value of τ_i (1 if codon i is occupied by the ribosome's A-site and 0 otherwise) at k-th update in the simulation, $\Delta t^{(k)} = t^{(k)} - t^{(k-1)}$, $t^{(k)}$ is the time of the k-th update, $t^{(0)} = 0$ and $T = t^{(M)}$.

3. Results

3.1. First-order approximation does not account for ribosome interference

Using (26) and (30a)-(30b) we can compute ribosome density ρ_i and protein synthesis rate J up to the linear order in α,

$$\rho_i \approx \frac{\alpha}{\omega_i} \frac{\prod_{j=2}^{i} \frac{\omega_j}{\omega_j+\mu}}{\left(1 - \frac{\gamma}{\beta+\gamma}\prod_{j=2}^{L-1}\frac{\omega_j}{\omega_j+\mu}\right)}, \; 2 \le i \le L-1 \tag{39a}$$

$$\rho_L \approx \frac{\alpha}{\beta+\gamma} \frac{\prod_{j=2}^{L-1} \frac{\omega_j}{\omega_j+\mu}}{\left(1 - \frac{\gamma}{\beta+\gamma}\prod_{j=2}^{L-1}\frac{\omega_j}{\omega_j+\mu}\right)} \tag{39b}$$

$$J \approx \frac{\alpha \prod_{j=2}^{L-1} \frac{\omega_j}{\omega_j+\mu}}{\left(1 - \frac{\gamma}{\beta+\gamma}\prod_{j=2}^{L-1}\frac{\omega_j}{\omega_j+\mu}\right)}. \tag{39c}$$

These results are similar to the ones obtained by Gilchrist and Wagner using a deterministic model of mRNA translation that includes codon-specific elongation rates, ribosome drop-off and mRNA circularisation but ignores ribosome interference [11]. This similarity is not a coincidence but comes from the fact that first order includes configurations with only one ribosome.

Another interesting prediction from the first order is that the impact of reinitiation strongly depends on the rate of premature termination. That is expected because reinitiation due to mRNA circularisation can only happen if the ribosome has not terminated translation prematurely. The strongest effect is thus when premature termination does not occur, i.e. when $\mu = 0$. In that case the products in Eqs. (39a)-(39c) are equal to 1 and the resulting ribosome density and current read

$$\rho_i \approx \frac{\alpha(1+\gamma/\beta)}{\omega_i}, \; i = 2, \ldots, L-1 \tag{40a}$$

$$\rho_L \approx \frac{\alpha}{\beta} \tag{40b}$$

$$J \approx \alpha\left(1 + \frac{\gamma}{\beta}\right). \tag{40c}$$

From here we conclude that in the first-order approximation reinitiation has the same effect as increasing initiation rate from α to $\alpha(1+\gamma/\beta)$.

In principle, the first order is a good approximation of the steady state provided the ribosomes on the mRNA are well separated, so that the ribosome collisions are negligible. In practice, that means that all the ratios α/ω_i and γ/β are much smaller that 1, so that the overall ribosome density is small. What also matters, according to Eq. (48a), is that the ratio of $k_i/k_{i+\ell}$ is close to 1 or smaller, otherwise there will be significant contributions to the second order, as we show later in the paper. This occurs if there is a 'fast' codon at site i and a 'slow' codon at site $i+\ell$ causing a traffic jam, in which case the ribosomes are not well separated and the first order may not be a good approximation in that part of the mRNA.

3.2. Second-order approximation accounts for ribosome interference

In the Methods we described in detail how to find all second-order coefficients. This allows us to compute local density ρ_i and current J up to the second order in α,

$$\rho_i = \rho_i^{(1)}\alpha + \rho_i^{(2)}\alpha^2 \tag{41}$$

$$J = J^{(1)}\alpha + J^{(2)}\alpha^2, \tag{42}$$

where linear coefficients $\rho_i^{(1)}$ and $J^{(1)}$ are given in Eqs. (39a) and (39c), respectively, and the second-order coefficients $\rho_i^{(2)}$ and $J^{(2)}$ read

$$\rho_i^{(2)} = c_2(1_i) + \sum_{j=2}^{i-\ell} c_2(1_j 1_i) + \sum_{j=i+\ell}^{L} c_2(1_i 1_j) \tag{43}$$

$$J^{(2)} = (\beta+\gamma)c_2(1_L) + \beta\sum_{j=2}^{L-\ell} c_2(1_j 1_L) + \gamma\sum_{j=2+\ell}^{L-\ell} c_2(1_j 1_L). \tag{44}$$

Figure 4 shows ribosome density (first 50 codons) for two genes of *S. cerevisiae*, YDR223W and YDR233C, computed using the model without reinitiation. These two genes have translation initiation rate smaller than the first quartile and larger than the third quartile of all initiation rates, respectively (see Figure 4). On the left are density profiles computed using the first order and compared with the results of Monte Carlo simulations. As expected, the agreement is worse for the gene that has a larger value of α. On the right are density profiles obtained using the second order, which agree well with the results of Monte Carlo simulations.

In Figure 5 we show ribosome current J_i across the mRNA, computed from Eq. (5) for the same two genes as before and using the model without reinitiation. Unlike the density, the first-order approximation of the current already shows a significant discrepancy

Accepted Manuscript

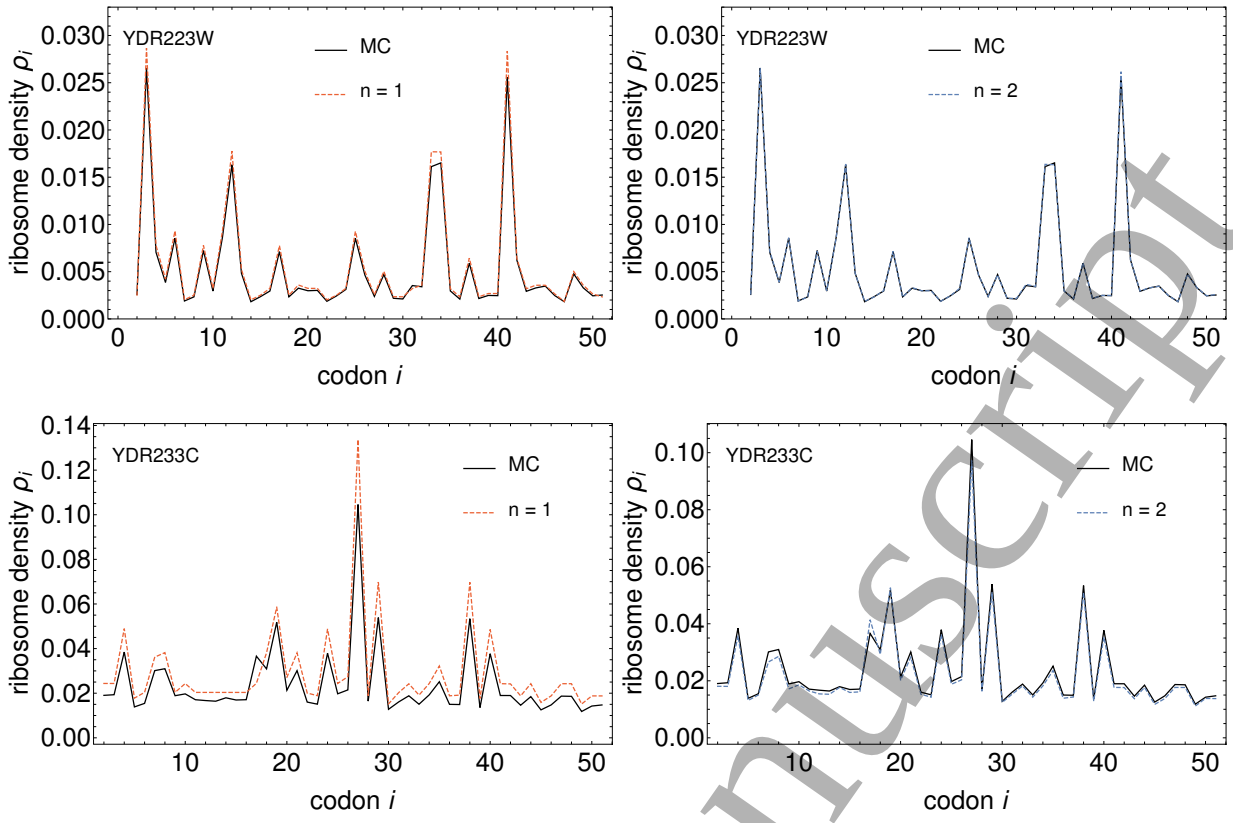

Figure 4. Density profiles (first 50 codons) for *S. cerevisiae* genes YDR233W and YDR233C. On the left and right are density profiles computed using the first and second order, respectively, and compared to the results of Monte Carlo (MC) simulations. Translation initiation rates are 0.02846 for YDR223W and 0.21425 for YDR233C. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1} and no translation reinitiation ($\gamma = 0$).

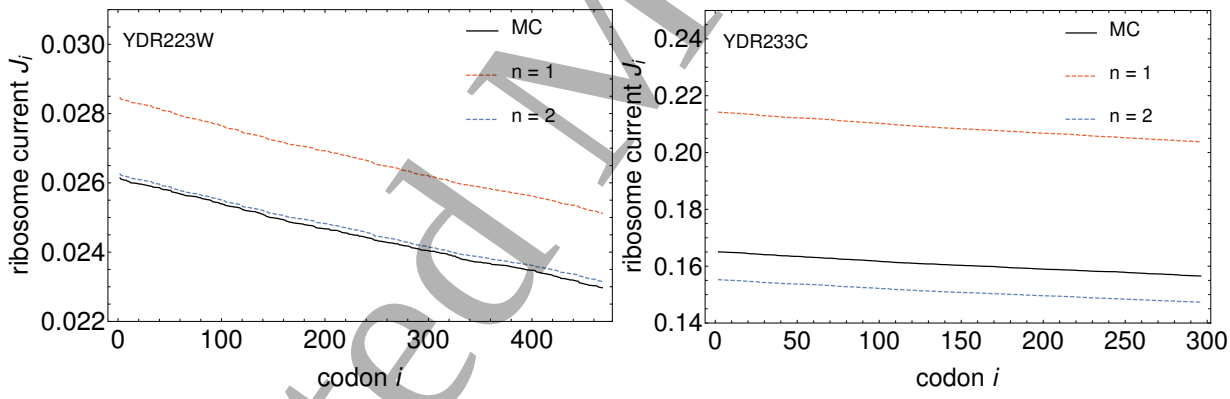

Figure 5. Ribosome current J_i across the mRNA for *S. cerevisiae* genes YDR233W and YDR233C, computed from Eq. (5). Solid black line is the result of stochastic simulations, while red and blue dashed lines represent first-order and second-order approximation, respectively. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1} and no translation reinitiation ($\gamma = 0$).

compared to Monte Carlo simulations for both genes. As expected, the discrepancy is reduced when using second-order approximation.

3.3. Effect of ribosome interference on second-order coefficients

Because the second order must be computed numerically, how exactly the second-order coefficients are affected by ribosome interference is not immediately obvious. If we imagine a mathematical model in which ribosome interference is ignored, we would expect $P(C)$

Accepted Manuscript

to be a product of single-particle weights $c_1(1_i)\alpha$

$$P(C) = \frac{1}{Z_L}\prod_{j=1}^{N(C)} \alpha c_1(1_{X(j)}) = \frac{1}{Z_L}\prod_{i=2}^{L}\left[\tau_i c_1(1_i)\alpha + (1-\tau_i)\right], \quad (45)$$

where $N(C)$ is the number of particles in C, $X(j)$ is the position of the j-th particle and $Z_L = \prod_{i=2}^{L}(1 + c_1(1_i)\alpha)$ is the normalisation (see Ref. [19] for more details in which we termed this approximation the independent particle approximation or IPA). Taking $C = 1_i 1_j$ and expanding $P(C)$ in α up to the quadratic order we get

$$c_2(1_i 1_j) \overset{\text{IPA}}{=} c_1(1_i)c_1(1_j). \quad (46)$$

Going back to the model with exclusion, we can write $c_2(1_i 1_j)$ as

$$c_2(1_i 1_j) = c_1(1_i)c_1(1_j)g_2(1_i 1_j). \quad (47)$$

where $g(1_i 1_j)$ measures the deviation from the IPA (for which $g(1_i 1_j) = 1$), i.e. the effect of exclusion. The equations for $g_2(1_i 1_j)$ for $i \neq 2$ and $j \neq L$ read

$$g_2(1_i 1_{i+\ell}) = \frac{e(1_i)}{e(1_{i+\ell})} g_2(1_{i-1} 1_{i+\ell}),\ i \neq 2 \quad (48a)$$

$$g_2(1_i 1_j) = \frac{e(1_i)}{e(1_i)+e(1_j)} g_2(1_{i-1} 1_j) + \frac{e(1_j)}{e(1_i)+e(1_j)} g_2(1_i 1_{j-1}),\ i \neq 2, j \neq L, \quad (48b)$$

where $e(1_i) = (1-\delta_{i,L})(\omega_i + \mu) + \delta_{i,L}\beta$. We notice that Eq. (48*b*) could be solved by setting all g_2 to 1, however that would violate the initial equation (48*a*). On the other hand, both $e(1_i)/(e(1_i) + e(1_j))$ and $e(1_j)/(e(1_i) + e(1_j))$ in Eq. (47) are strictly less than 1, which means that any deviation of g_2 from 1 in Eq. (48*a*) will be attenuated by subsequent iterations of Eq. (48*b*). Therefore we expect to find $g_2(1_i 1_j) \approx 1$ when codons i and j are far apart, i.e.

$$c_2(1_i 1_j) \approx c_1(1_i)c_1(1_j) \quad \text{for } |i-j| \gg \ell. \quad (49)$$

Certainly, the effect of exclusion is strongest when the ribosomes are next to each other, i.e. for $j = i + \ell$. In that case there is either a magnification ($e(1_i) > e(1_{i+\ell})$) or reduction ($e(1_i) < e(1_{i+\ell})$) in $g_2(1_i 1_j)$ compared to the IPA that is carried over to the surrounding codons.

In Figure 6 we plot $g_2(1_i 1_j)$ for YDR233C gene as a function of j for several values of i. As predicted, the deviation of $g_2(1_i 1_j)$ from 1 is the largest at $j = i + \ell$ and eventually decays to 1 as j gets away from i.

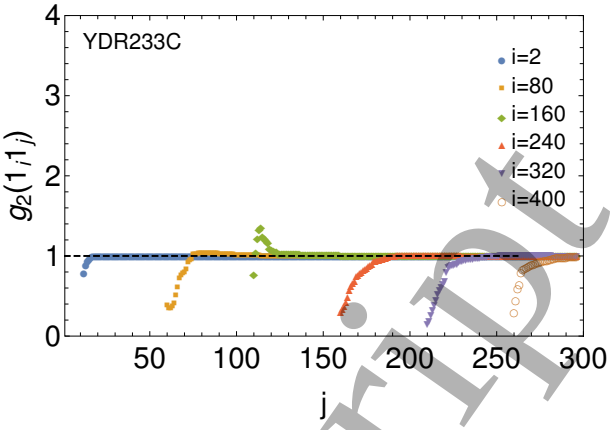

Figure 6. The coefficient $g_2(1_i 1)j)$ for YDR233C as a function of j for several values of i and assuming no translation reinitiation.

3.4. High-order approximations are needed for genes with high initiation rates

As the rate of initiation increases, using the first-order or second-order approximation may lead to significant errors. In Figure 7 we demonstrate this for gene YOR045W, which has a relatively large value of $\alpha = 0.35423$ and total ribosome density $\rho = 0.03256$ (approximately 33% of the maximum theoretical density $1/\ell = 0.1$). On the left are density profiles computed using first-order and second-order approximation and compared to the results of Monte Carlo simulations. On the right is the density profile obtained using the fourth-order approximation, which agrees well with the results of Monte Carlo simulations. Similar conclusions can be made for the ribosome current J_i across the mRNA., see Figure 8.

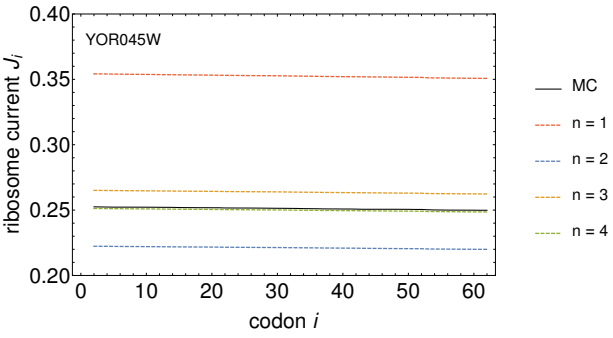

Figure 8. Ribosome current J_i across the mRNA for *S. cerevisiae* gene YOR045W, computed from Eq. (5). Solid black line is the result of stochastic simulations, while red, blue, orange and green dashed lines represent first-order, second-order, third-order and fourth-order approximation, respectively. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1} and no translation reinitiation ($\gamma = 0$).

Accepted Manuscript

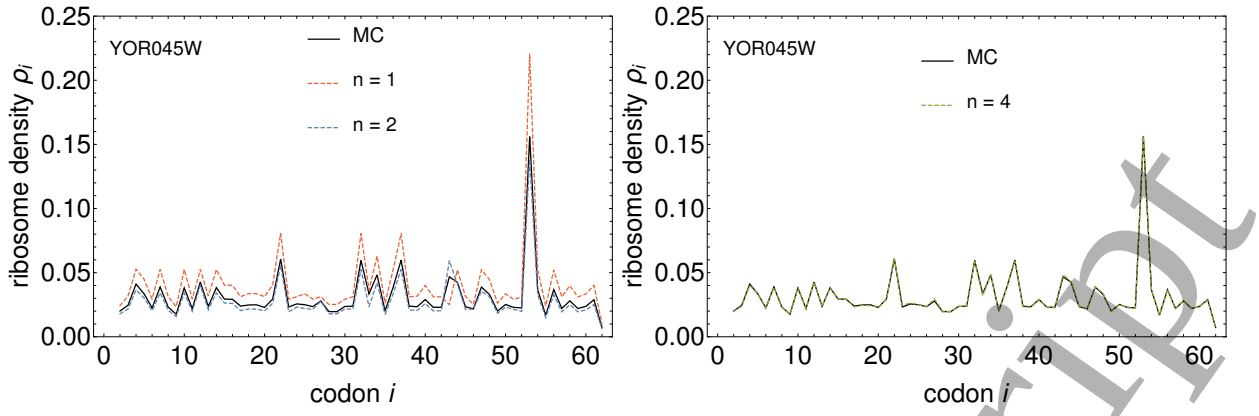

Figure 7. Ribosome density profiles for *S. cerevisiae* gene YOR045W. On the left and right are density profiles computed using the second and fourth order, respectively, and compared to the results of Monte Carlo (MC) simulations. Translation initiation rate is 0.35423. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1} and no translation reinitiation ($\gamma = 0$).

3.5. Translation reinitiation has the same effect as increasing initiation rate

In Figure 9 we present density profiles for two genes, YDR223W and YDR233C, obtained using a model with translation reinitiation with reinitiation efficiency set to $\eta = 0.2$.

For gene YDR223W, which has a small value of α, the agreement between the second-order approximation and results of Monte Carlo simulations is excellent. On the other hand, there is a visible discrepancy between the second-order approximation and results of Monte Carlo simulations for gene YDR233C, which has a relatively large value of α. This result is expected because translation reinitiation increases the number of ribosomes that initiate translation, which in turn may require more terms in the series expansion. Therein lies the problem–computing higher-order terms in the model with translation reinitiation is not as straightforward as without reinitiation, because it involves solving a linear system of equations. For example, to compute the third order we need to solve a system of roughly L^2 equations for the unknown coefficients $c_2(1_i 1_j 1_L)$. For a typical gene of $L = 300$ codons that is 90000 equations. Thus the power series method may not be a feasible approach for solving TASEP-based models of translation with reinitiation beyond the second order.

This problem motivates to ask if the model with translation reinitiation can be replaced with an effective model without reinitiation but in which the rate of translation initiation is set to a higher value $\alpha_{\text{eff}} > \alpha$. This value must be such that

$$\alpha_{\text{eff}} = \frac{J_{\text{in}}}{\left\langle \prod_{i=2}^{\ell+1}(1-\tau_i) \right\rangle} = \alpha + \gamma \frac{\left\langle \tau_L \prod_{i=2}^{\ell+1}(1-\tau_i) \right\rangle}{\left\langle \prod_{i=2}^{\ell+1}(1-\tau_i) \right\rangle} = \alpha + \frac{J - \beta\langle\tau_L\rangle}{\left\langle \prod_{i=2}^{\ell+1}(1-\tau_i) \right\rangle} \tag{50}$$

where J_{in} is the total influx of ribosomes initiating translation, Eq. (4), and the denominator is the probability that the first $\ell = 10$ codons are not occupied by another ribosome's A-site. From here we can express the effective initiation rate α_{eff} in terms of J_{in}, J and $\langle\tau_L\rangle$ as

$$\alpha_{\text{eff}} = \frac{\alpha}{1 - (J - \beta\langle\tau_L\rangle)/J_{\text{in}}}. \tag{51}$$

We can check that $\alpha_{\text{eff}} = \alpha$ when $\gamma = 0$ in which case $J = \beta\langle\tau_L\rangle$.

In order to test whether we can replace the model with reinitiation ($\gamma > 0$) with a model without reinitiation ($\gamma = 0$) but with an effective initiation rate α_{eff}, we first use stochastic simulations to compute the values of J_{in}, J and $\langle\tau_L\rangle$ for the model with reinitiation. We then use Eq. (51) to find the effective initiation rate α_{eff}, and use that rate in stochastic simulations of the model without reinitiation. So at this point we are not using the power series method at any point, we are only testing if we can replace the original model with a simpler one.

In Figure 10 we present density profiles for genes YDR233C and YOR045W obtained using Monte Carlo simulations of the model with reinitiation $\eta = 20\%$ and the effective model without reinitiation. For both genes we find an excellent agreement between the two models. Next, we consider $\eta = 90\%$ for YOR045W and YKL036C. Both of these two genes have high initiation rates belonging to the last quartile in Figure 3. In fact, YKL036C has the largest initiation rate of all *S. cerevisiae* genes estimated at the value of $\alpha = 4.1$ initiations/s, for which the power series method is inapplicable. In Figure 11 we present density profiles obtained using Monte Carlo simulations of the model with reinitiation $\eta = 90\%$ and the effective

Accepted Manuscript

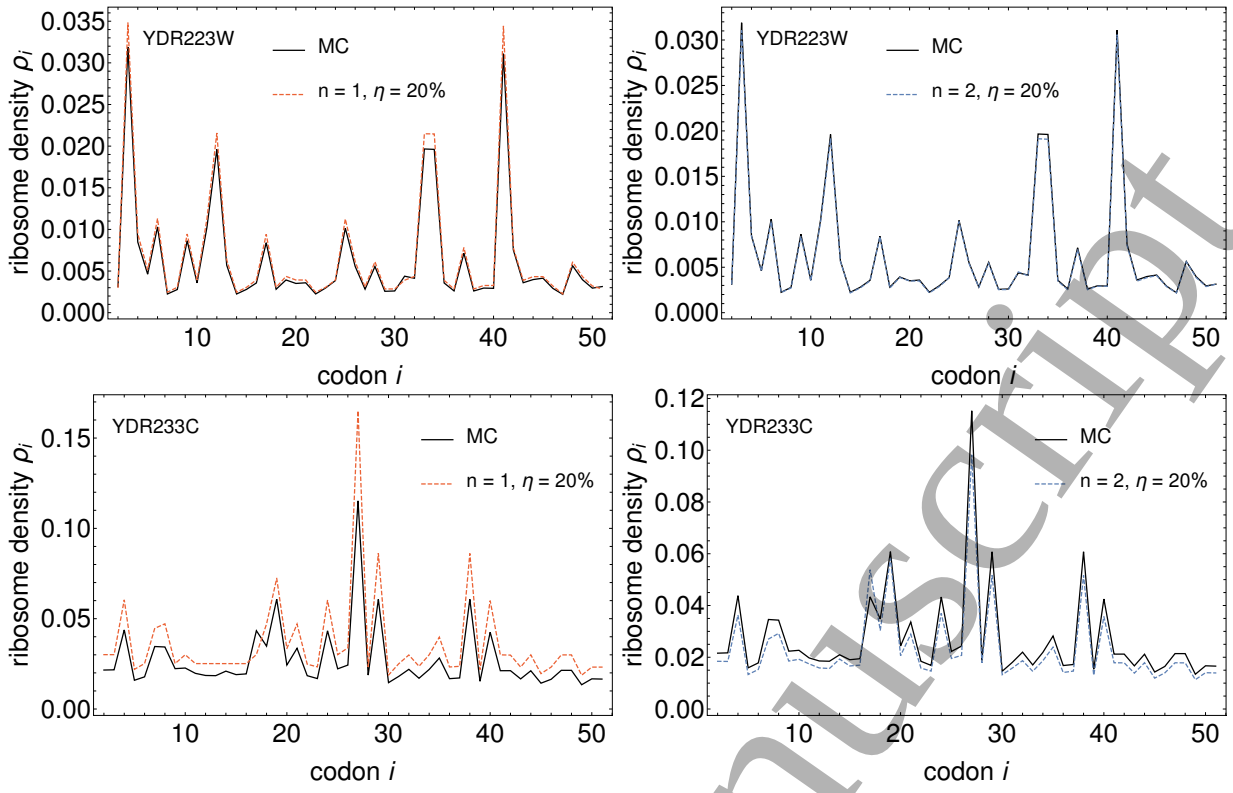

Figure 9. Density profiles (first 50 codons) for *S. cerevisiae* genes YDR233W and YDR233C. On the left and right are density profiles computed using the first and second order, respectively, and compared to the results of Monte Carlo (MC) simulations. Translation initiation rates are 0.02846 for YDR223W and 0.21425 for YDR233C. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1} and translation reinitiation with $\eta = 0.2$.

model without reinitiation. Again, we find an excellent agreement between the two models.

This result has two important implications. The first one is technical–we can apply the power series method to the effective model (provided the effective initiation rate α_{eff} is smaller than any of the elongation and termination rates) and avoid the problem of solving a linear system of equations. The second one is biological. In experiments the rates of the model are typically unknown and have to be inferred from the data. For example, if we want to estimate the rate of initiation α by matching theoretical density $\rho(\alpha)$ to the experimental density from polysome profiling experiments, as it was done in Ref. [23], we cannot truly distinguish reinitiation from *de novo* initiation. In other words, the evidence for translation reinitiation may be very difficult to find experimentally because the effect of translation reinitiation is the same as *de novo* initiation at a higher rate.

4. Discussion

Our first main result is that the power series method is applicable to the TASEP with ribosome drop-off and translation reinitiation. This complements previous work in which the method was applied to the TASEP with multi-step elongation [19]. We tested the method on *Saccharomyces cerevisiae* under physiological conditions and found that the model-predicted ribosome density and current are faithfully described by the second-order approximation for most of the genes. Interestingly, second order is the lowest order at which ribosome interference occurs, suggesting that ribosome interference does have an effect on translation. This is clearly visible for genes with high initiation rates belonging to the last quartile in Figure 3, for which higher-order approximations are needed to describe the data. In that sense the statement often found in biology that initiation is rate-limiting for translation is true [46], but incomplete–translation elongation does have an effect on translation.

Our second main result is an iterative algorithm that computes ribosome density and current up to any order. This is a significant improvement over previous work that considered only second order [19]. However, computing orders beyond the second is limited to the model without translation reinitiation. The problem is that reinitiation does not allow for the coefficients

Accepted Manuscript

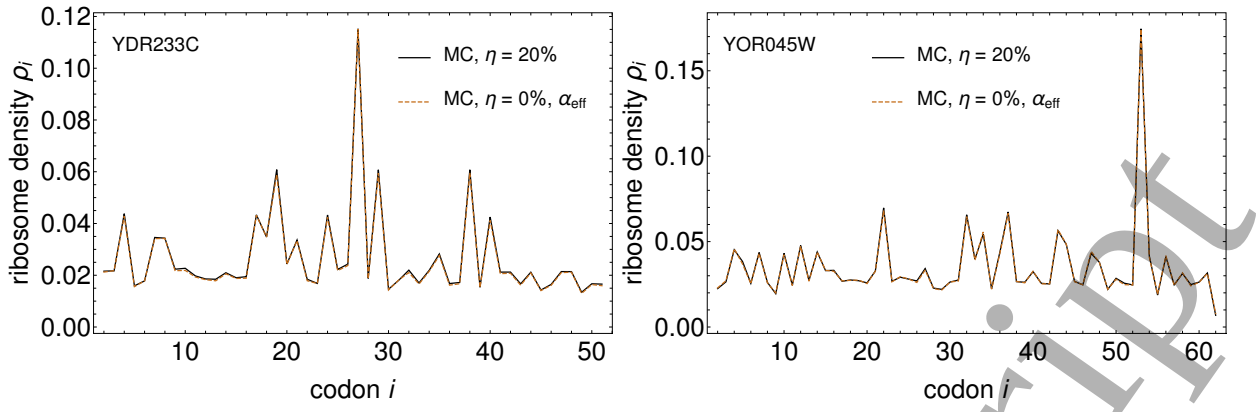

Figure 10. Density profiles for *S. cerevisiae* genes YDR233W (first 50 codons) and YDR233C (all codons), obtained using Monte Carlo simulations of the model with reinitiation ($\eta = 0.2$) and the effective model without reinitiation ($\eta = 0$). Translation initiation rates are $\alpha = 0.21425$ for YDR233C and 0.35423 for YOR045W. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1}.

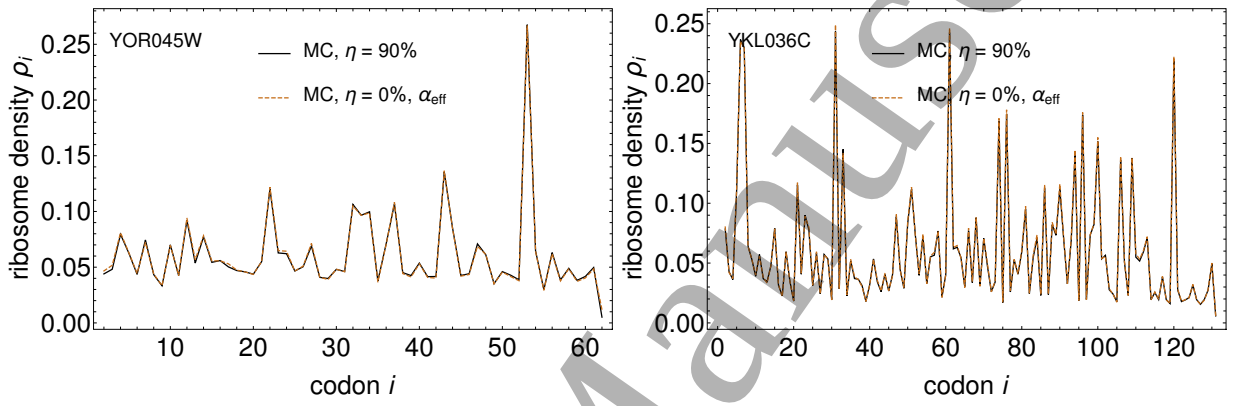

Figure 11. Density profiles for *S. cerevisiae* genes YOR045W and YKL036C, obtained using Monte Carlo simulations of the model with reinitiation ($\eta = 0.9$) and the effective model without reinitiation ($\eta = 0$). Translation initiation rates are $\alpha = 0.40965$ for YOR045W and 4.1966 for YKL036C. All results were obtained assuming ribosome drop-off rate $\mu = 1.4 \cdot 10^{-3}$ s^{-1}.

$c_n(C)$ in Eq. (13) to be found recursively starting from a configuration with all ribosomes stacked to the left, as it is the case in the model without reinitiation. Instead one must first solve a closed linear system of equations for the coefficients $c_n(C)$ with n-th ribosome at the last codon site (the stop codon). The number of such coefficients is of the order of L^{n-1}, which becomes too large for $n > 2$ given a typical gene length in hundreds of codons. Another serious limitation is that the number of all configurations contributing to n-th order is of order of L^n. This is a problem because the coefficients are computed recursively and need to be stored during the recursion process in Eq. (22), which limits how large n and L can be. In practice, we expect memory shortage for computations beyond the third order for typical mRNAs consisting of hundreds of codons. It may be possible to compute high-order terms for short genes though, such as the fourth order that we computed for YOR045W ($L = 61$ codons) in Figure 7.

As with any perturbation theory in physics, such as the power series method presented here, computing high-order terms becomes progressively more difficult. We can then ask what is the advantage of the power series method compared to stochastic simulations and what biological insight we can get from it?

The power series method was developed in Ref. [19] in order to understand how the protein production rate depends on the initiation and elongation rates, which is a long-standing problem in molecular biology, especially in the context of codon optimisation [47]. This problem is too difficult to study with stochastic simulations alone, because there are too many parameters that can be varied.

When the power series method was applied to the basic TASEP with two-step elongation in Ref. [19], it revealed that the protein production rate per mRNA is predominantly determined by the rate of initiation and the rate of elongation of the first 10 codons, which is the size of the ribosome (in codons). In this paper the power series method further elucidated how ribosome jamming may occur at codons for which

Accepted Manuscript

the ratio $\omega_i/\omega_{i+\ell}$ is much larger than 1, i.e. when a 'slow' codon is $\ell = 10$ codons downstream of a 'fast' codon. Furthermore, the first-order expression for the ribosome current in Eq. (40*c*), $J = \alpha(1+\gamma/\beta)$ prompted us to replace the model with reinitiation with a simpler model without an effective initiation rate.

Another useful application of the power series method is when the model parameters are not known and thus have to be inferred from the experimental data. For example, the data can be ribosome density profiles obtained from ribosome profiling (Ribo-seq) experiments [25]. The problem is then to match the density profile of the model to the experimental data, which amounts to solving a system of L nonlinear equations. There are nonlinear optimisation algorithms that are built for this problem, however what is computationally expensive is to run stochastic simulations for every iteration of the algorithm. Instead, even the third-order calculations are fast (assuming no reinitiation), allowing the optimisation to finish in a reasonable time. This approach was recently developed in Ref. [48] and successfully applied to *S. cerevisiae* Ribo-seq data.

TASEP-based models of translation are usually studied using stochastic simulations (generated by the Gillespie algorithm) or using mathematical approximations (called mean-field approximations) that ignore correlations between two neighbouring ribosomes [1,2,8]. Power series method is the only mathematical method available that can account for these correlations, other that the stochastic simulations. Put differently, the only approximation in the derivation of $P(C)$ is that the power series is approximated by a polynomial. This is markedly different from the mean-field approximations of Refs. [1, 2, 8] that explicitly ignore correlations between two neighbouring ribosomes (see Ref. [45] in which this difference was demonstrated for the TASEP with particles of size $\ell = 1$).

In this work we studied the effect of ribosome-ribosome correlations on the second-order coefficients $c_2(1_i 1_j)$ for the TASEP without translation reinitiation. The strongest correlations were found for ribosomes that are next to each other ($j = i + \ell$), with the strength of correlations depending on the ratio $(\omega_i + \mu)/(\omega_{i+\ell} + \mu)$. For $(\omega_i + \mu)/(\omega_{i+\ell} + \mu) < 1$ ($(\omega_i + \mu)/(\omega_{i+\ell} + \mu) > 1$), the density at codon i is smaller (larger) than it would be on a mRNA composed of only one ribosome. Taking this further, if we could arrange codons in a sequence such that

$$\omega_2 < \omega_3 < \ldots < \omega_L, \tag{52}$$

then according to the second-order approximation, the total ribosome density for that sequence would be minimal compared to the same choice of codons arranged in a different sequence. Since ribosomes are highly costly in terms of cellular energy, it makes sense for the cell to reduce ribosome density and avoid ribosome queuing. This explanation is also known as the ramp hypothesis and may explain why the preference for slower codons is typically found at the beginning of the mRNA [49–51]. Our hypothetical arrangement in Eq. (52), which could be considered as a perfect ramp, is unlikely to occur in real codon sequences due to other evolutionary factors driving codon usage. Nevertheless, our findings may provide the first step in understanding the origin of the ramp from a mathematical point of view.

5. Conclusions

We have presented a versatile method for studying TASEP-based models of translation that account for several mechanistic details of the translation process: codon-dependent elongation, premature termination and mRNA circularisation. We have applied our method to the model organism *Saccharomyces cerevisiae* using realistic estimates for the model's parameters under physiological conditions (except for the value of the reinitiation rate which is, to the best of our knowledge, unknown).

In the model without reinitiation, we find an excellent agreement for the ribosome density and current with the results of stochastic simulations using approximations up to the fourth order of the power series expansion. In order to obtain these results we devised an algorithm that can, in principle, compute any order of the power series expansion. In practice, the program is limited only by the amount of memory used for storing the coefficients.

Once the reinitiation is introduced in the model, the power series method becomes too cumbersome to do beyond the second order. However, the first order calculation revealed that the effect of reinitiation is similar to setting the rate of reinitiation to zero but increasing the rate of *de novo* initiation. We tested this hypothesis for several genes and found that the simpler model without reinitiation correctly describes the model with reinitiation. In the biological context this result suggests that the effect of reinitiation on translation cannot be easily distinguished from *de novo* initiation, for example if the only available experimental data are ribosome density profiles.

The main advantage of the presented method is that it is robust, in the sense that it can be applied to many realistic models of translation. The case of multi-step elongation was already studied in Ref. [19]. Interactions between ribosomes that are more complex than the simple exclusion can also be included in the model, for example premature termination caused by ribosome collisions [33, 34].

There are also general limitations to the power

Accepted Manuscript

series method that we wish to emphasise. The method is applicable only to the initiation-limited regime, in which the rate of initiation is smaller than the elongation rates of individual codons and the termination rate. In *S. cerevisiae* under physiological conditions there are few genes with very high initiation rates that do not meet this criterion. Our approach also excludes the case of amino acid starvation that may cause the elongation rates to be smaller than the initiation rate. Finally the method is applicable on to the steady state and thus cannot take into account finite lifetime of the mRNA.

While the TASEP as a model for translation has been proposed half a century ago, it has only recently become a common tool in computational biology. Our goal for the future is to make the presented method a standard tool for analysing biological data e.g. from ribosome profiling experiments, which would give us a better understanding of the translation process and allow us to address open questions in the cell biology.

Acknowledgments

JSN was supported by the Leverhulme Trust Early Career Fellowship under grant number ECF-2016-768.

References

[1] MacDonald C T, Gibbs J H and Pipkin A C 1968 *Biopolymers* **6** 1–25
[2] MacDonald C T and Gibbs J H 1969 *Biopolymers* **7** 707–25
[3] Spitzer, F 1970 *Advances in Mathematics* **5**(2) 246–290
[4] Schmittmann B and Zia R K P 1995 *Statistical mechanics of driven diffusive systems* (*Phase Transitions and Critical Phenomena* vol 17) ed C Domb and J L Lebowitz (London:Academic Press)
[5] Derrida B, Evans M R, Hakim V and Pasquier V 1993 *J. Phys. A: Math. Gen.* **26** 1493
[6] Schütz G and Domany E 1993 *J. Stat. Phys.* **72** 277
[7] Krug J 1991 *Phys. Rev. Lett.* **67** 1882
[8] Shaw L B, Sethna J P and Lee K H 2004 *Phys. Rev. E* **70** 021901
[9] Varenne S, Buc J, Lloubes R and Lazdunski C 1984 *J. Mol Biol.* **180** 549
[10] Sorensen M A, Kurland C G and Pedersen S 1989 *J. Mol. Biol.* **207** 365–377
[11] Gilchrist M A and Wagner A 2006 *J. Theor. Biol.* **239** 417–34
[12] Bonnin P, Kern N, Young N T, Stansfield I and Romano M C 2017 *PLoS Comput. Biol.* **13** e1005555
[13] Chou T 2003 *Biophys. J.* **85** 755–773
[14] Sharma A K and Chowdhury D 2011 *J. Theor. Biol.* **289** 36–46
[15] Marshall E, Stansfield I and Romano M C 2014 *J. R. Soc. Interface* **11** 20140589
[16] Zur H and Tuller T 2016 *Nucleic Acids Res.* **44** 9031–9049
[17] Gingold H and Pilpel Y 2011 *Mol. Syst. Biol.* **7** 481
[18] Brule C E and Grayhack E J 2017 *Trends Genet.* **33** 283–297
[19] Szavits-Nossan J, Ciandrini L and Romano M C 2018 *Phys. Rev. Lett.* **120** 128101
[20] Ciandrini L, Stansfield I and Romano M C 2010 *Phys. Rev. E* **81** 051904
[21] Kennell D and Riezman H 1977 J. Mol. Biol. **114** 1–21
[22] Rudorf S and Lipowsky R 2015 *PLoS ONE* **10**(8) e0134994
[23] Ciandrini L, Stansfield I, Romano M C 2013 *PLOS Computational Biology* **9**(1) e1002866
[24] MacKay V L *et al* 2004 *Mol. Cell. Proteomics.* **3** 478–489
[25] Ingolia N T, Ghaemmaghami S, Newman J R and Weissman J S 2009 *Science* **324**(5924) 218–23
[26] Wells S E, Hillner P E, Vale R D and Sachs A B 1998 *Mol. Cell.* **2**(1) 135–40
[27] Vicens Q, Kieft J S and Rissland O S 2018 *Molecular Cell* **72** 805–812
[28] Borujeni A E and Salis H M 2016 *J. Am. Chem. Soc.* **138** 7016-7023
[29] Dimelow R J and Wilkinson S J 2009 *J. Roy. Soc. Interface* **6** 51–61
[30] Fluitt A, Pienaar E and Viljoen H 2007 *Comput. Biol. Chem.* **31** 335–346
[31] Basu A and Chowdhury D 2007 *Phys. Rev. E* **75** 021902
[32] Zouridis H and Hatzimanikatis V 2007 *Biophysical Journal* **92** 717–730
[33] Ferrin M A and Subramaniam A R 2017 *eLife* **6** e23629
[34] Park H and Subramaniam AR 2019 *PLoS Biol* **17**(9) e3000396
[35] Schuller A P and Green R 2018 *Nat. Rev. Mol. Cell. Biol.* **19**(8) 526–541
[36] Wen J-D, Lancaster L, Hodges C, Zeri A-C, Yoshimura S H, Noller H F, Bustamante C and Tinoco I 2008 *Nature* **452** 598–603
[37] Qu X, Wen J-D, Lancaster L, Noller H F, Bustamante C and Tinoco I 2011 *Nature* **475** 118–121
[38] Chen C, Zhang H, Broitman S L, Reiche M, Farrell I, Cooperman B S and Goldman Y E 2013 *Nat. Struct. Mol. Biol.* **20** 582–588
[39] Von Heijne G, Nilsson L and Blomberg C 1977 *J. Theor. Biol.* **68**(3) 321–329
[40] Sin C, Chiarugi D and Valleriani A 2016 *Nucleic Acids Res.* **44**(6) 2528–2537
[41] Savelsbergh A, Katunin V I, Mohr D, Peske F, Rodnina M V and Wintermeyer W 2003 *Mol. Cell.* **11**(6) 1517-23
[42] Szavits-Nossan J 2013 *J. Phys. A: Math. Theor.* **46** 315001
[43] Chaiken S and Kleitman D J 1978 *J. Comb. Theory, Ser. A* **24** 377
[44] Schnakenberg J 1976 *Rev. Mod. Phys.* **48** 571
[45] Szavits-Nossan J, Romano M C and Ciandrini L 2018 *Phys. Rev. E* **97** 052139
[46] Shah P, Ding Y, Niemczyk M, Kudla G and Plotkin J B 2013 *Cell* **153** 1589–1601
[47] Plotkin J B and Kudla G 2011 *Nature Reviews Genetics* **12** 32–42
[48] Szavits-Nossan J and Ciandrini L 2019 bioRxiv 719302
[49] Bulmer M 1988 *J. Theor. Biol.* **133** 67–71
[50] Mitarai N, Sneppen K and Pedersen S 2008 *J. Mol. Biol.* **382** 236–245
[51] Tuller T *et al* 2010 *Cell* **141**(2) 344–354

Accepted Manuscript