Available online at www.sciencedirect.com
ScienceDirect
Journal of Multivariate Analysis 99 (2008) 2082–2095
Journal of Multivariate Analysis
www.elsevier.com/locate/jmva

Limit theorems for hybridization reactions on oligonucleotide microarrays

Grzegorz A. Rempala[a,b,*], Iwona Pawlikowska[c]

[a] *Department of Mathematics, University of Louisville, United States*
[b] *Center for Genetics and Molecular Medicine, University of Louisville, United States*
[c] *Department of Mathematics, University of Silesia, Katowice, Poland*

Received 4 April 2007
Available online 13 February 2008

Abstract

We derive herein the limiting laws for certain stationary distributions of birth-and-death processes related to the classical model of chemical adsorption–desorption reactions due to Langmuir. The model has been recently considered in the context of a hybridization reaction on an oligonucleotide DNA-microarray. Our results imply that the truncated-gamma- and beta-type distributions can be used as approximations to the observed distributions of the fluorescence readings of the oligo-probes on a microarray. These findings might be useful in developing new model-based, probe-specific methods of extracting target concentrations from array fluorescence readings.
Published by Elsevier Inc.

AMS 2000 subject classifications: 60F05; 60G35

Keywords: Birth-and-death process; Density-dependent Markov process; Oligonucleotide microarray; Hybridization reaction; Gamma distribution; Langmuir adsorption–desorption model

1. Introduction

High density oligonucleotide microarrays are a widely used modern bio-technology tool enabling the simultaneous testing for the presence as well as quantification of large numbers of genes in prepared target RNA samples. For a general introduction to this technology we refer

* Corresponding address: Department of Mathematics, University of Louisville, Louisville, KY 40292, United States.
E-mail address: g.rempala@louisville.edu (G.A. Rempala).

0047-259X/$ - see front matter. Published by Elsevier Inc.
doi:10.1016/j.jmva.2008.02.014

the reader to the celebrated paper [14] or to [13] for a more recent overview. Among several competing types of oligonucleotide microarrays, the Affymetrix GeneChip design appears to be currently one of the most common ones. GeneChip arrays consist of a substrate onto which short single strand DNA oligonucleotide probes have been synthesized using a photolithographic process. A chip surface is divided into some hundreds of thousands of regions typically tens of microns in size, with the DNA probes within each region being synthesized to a specific nucleotide sequence. The target-RNA sample is hybridized onto the chip to form probe-target duplexes, and the chip is scanned to obtain fluorescence intensity readings from dyes incorporated during the laboratory procedures. In principle, with suitable calibration, intensity readings are intended as a 'proxy' measure of the concentration of matching target RNA in the sample. However, due to optical noise, non-specific hybridization, probe-specific effects, and measurement error, the empirical measures of expression (i.e., the scanner-measured fluorescence) that summarize probe intensities can often lead to imprecise and inaccurate results (see, e.g., [16]).

It seems that some potentially significant improvement in relating the scanner readings of the probe intensities to the target genes concentrations could be obtained by using a model-based approach accounting for the physical processes driving hybridization. Recently, some authors have begun to address these issues by appealing to the dynamic adsorption models well known in the physical chemistry literature (see, [6] or [1]). Such models stemming from the physics of the chemical reactions involved are especially valuable as they could also help us in understanding better the physical processes driving hybridization and lead to improvements in both microarray design and performance.

One of the most popular adsorption models considered in the context of microarrays (cf. e.g., [5] or [1]) is the so-called *Langmuir model* (see the next section) which in its simplest deterministic form describes the relationship between concentration and fluorescence levels of probe-target complexes by means of a hyperbolic function. In the context of microarrays (in particular, GeneChips) in order to properly account for the effects of multiple simultaneous hybridizations as well as the cross-hybridization due to competition between similarly sequenced targets for the same probe regions, its seems that the stochastic version of the Langmuir model is needed. The analysis of such a model was carried out recently for instance in [1] or earlier in [12, 10] by means of adopting the general results of [2] on the fluctuations of the stochastic diffusion equations around their stable equilibrium points.

The model for the stochastic fluctuations of the equation described by Dennis and Patil was cast as a boundary-free problem and intended to provide a continuous diffusion-type approximation to the behavior of large biological systems as typically encountered in population dynamics problems. With no natural boundary restrictions it was argued in [2] that the fluctuations around stable equilibria are approximately distributed as a gamma random variable. Based on this argument the gamma model for gene expressions was since adopted by several authors in the context of analyzing microarray data (cf. e.g., [12,11,1]).

The simple extension of the Dennis and Patil results to microarray setting, albeit appealing, seems to require further justification since the microarray hybridization models are neither continuous nor boundary-free. Whereas the continuous approximation to the large discrete system seems easily justifiable, it is not entirely clear what discrete system is being approximated by the boundary-free diffusion model (see (2)).

The purpose of the current paper is to formally derive some simple closed-form stochastic laws approximating the equilibrium distributions of the discrete stochastic hybridization reactions under the explicit assumptions on the random noise terms which are consistent with the stochastic Langmuir model but, unlike the latter, are not boundary-free. The idea for the

derivation is very simple. We start by noticing that the reaction rate equation of the deterministic Langmuir model may be considered as the usual approximation to the set of two coupled stochastic chemical reactions on the finite-state space (i.e., probe region size for GeneChips). We then add the stochastic forcing of the Langmuir equation as an additional term to the original birth-and-death rates of this discrete chemical system. It turns out that the analysis of the equilibrium distribution of this adjusted birth-and-death process (we refer to it below as the Langmuir BD process) when the number of states is large leads to the stochastic laws which are mostly consistent with the Dennis and Patil [2] approach as well as Burden et al. [1] results. However, due to the fact that we had based our analysis on a finite discrete system, unlike previous approaches ours gives more insight into the boundary behavior of the underlying discrete stationary process and its approximations. In particular, when considering discrete model of hybridization reaction it becomes obvious that an adjustment for the saturation effect is needed in the form of a Dirac-delta probability distribution at the boundary of the state space. This leads to an interesting consequence that the limiting stochastic law is not absolutely continuous as in the Dennis and Patil result but rather has an atom at the state-space boundary. We give the formal details of these findings in Section 3.

Beyond the current introductory section the paper is organized as follows. In the next section (Section 2) we offer a brief overview of some of the results in the literature related to the classical Langmuir adsorption–desorption model in our context. Our main theorem on the limiting stochastic law for the stationary distributions of the Langmuir birth-and-death process in large state space along with some discussion is presented in Section 3. We conclude with some final remarks in Section 4.

2. The Langmuir model

In 1916 Irving Langmuir devised a simple model involving a thermodynamic equilibrium to predict the fraction of solid surface covered by an adsorbate as a function of its gas pressure [8]. The model was later extended to liquid systems, where the equilibrium involved concentrations in solution. In the Langmuir model adsorbate and solvent molecules compete to adsorb on sites on the surface of the powder and each site must be occupied by either a solvent molecule or an adsorbate molecule. For the hybridization reaction in oligonucleotide DNA-microarrays the same principle is applied in order to represent competing adsorption and desorption of RNA molecules to form probe-target complexes at the chip surface (see, e.g., [4]).

Let $u = u(t) \in (0, 1)$ be a fraction of sites within a probe region occupied by probe-target complexes at time t after the commencement of hybridization, and d_1 and d_2 be the forward adsorption and backward desorption rate constants, respectively. The forward adsorption reaction is assumed to occur at a rate $d_1 x(1 - u)$, proportional to the RNA-target concentration x and fraction $(1 - u)$ of unoccupied probe sites. The backward reaction (desorption) is assumed to occur at a rate $d_2 u$, proportional to the fraction of occupied probe sites. In a deterministic setting, the fraction of probe sites occupied by probe-target complexes is then given by the reaction rate equation known as the *Langmuir equation*

$$\frac{\mathrm{d}u}{\mathrm{d}t} = d_1 x - (d_1 x + d_2)u. \tag{1}$$

The corresponding equation incorporating the stochastic noise associated with both target and non-target-specific hybridization is given by the following stochastic version of (1) herein referred to as the *stochastic Langmuir equation*. It has the form

$$\frac{\mathrm{d}u}{\mathrm{d}t} = d_1 x - (d_1 x + d_2)u + \sqrt{g(u)}Z_t, \tag{2}$$

where $g(u) \geq 0$ is a known function of u and Z_t is a Gaussian white noise process with unit variance. The model described by the above equation is known in the literature as the *stochastic Langmuir model* and is a special case of a diffusion model considered e.g., by [2] in their study of stochastic fluctuations of populations about their stable equilibria. We note that in the present context of the Langmuir adsorption–desorption model, the solution of the stochastic equation (2) is no longer bounded and thus (2) suffers an obvious drawback in the fact that the function u has no physical interpretation for $u > 1$.

We also note that (2) is concerned with a single DNA (oligo) probe only, with the effect of other probes replaced by a random noise term (stochastic forcing). In the context of modeling a reaction network of simultaneous hybridization reactions on a DNA-microarray (2) is therefore one of the simplest examples of a complex system *decoupling* and *stochastic excitation* (see, e.g., [9,7]).

In general, under mild regularity conditions on g, one can argue that the stationary solution of (2) is approximately distributed as a gamma random variable around the deterministic system *steady state* [2]. However, as noted by Burden et al. [1] for a linear choice of g, namely

$$g(u) = Cd_1 xu \tag{3}$$

with $C > 0$ the Eq. (2) has an *exact* stationary gamma solution. That fact may be easily inferred from the corresponding Fokker–Planck equation (see, e.g., the monographs by van Kampen [15] or Ethier and Kurtz [3] and the references therein) which written in terms of the density of u, say $\psi(u, t)$, is given by[1]

$$\frac{\partial \psi(u,t)}{\partial t} = \frac{\partial}{\partial u}\left\{[(d_1 x + d_2)u - d_1 x]\psi(u,t) + \frac{Cd_1 x}{2}\frac{\partial[u\psi(u,t)]}{\partial u}\right\}. \tag{4}$$

Solving for the steady-state density, say $\psi_0(u)$, gives

$$\psi_0(u) \propto u^{2/C-1}\exp\left(-2\frac{d_1 x + d_2}{Cd_1 x}u\right) \quad \text{for } u \in [0, \infty) \tag{5}$$

and zero otherwise.

From the above considerations we see that adopting the stochastic model (2) with no additional assumptions may result in a stationary solution $\psi_0(u)$ being an absolutely continuous distribution with positive support on $(0, \infty)$. This finding is, however, not consistent with the experimental data which suggests that, at least for some values of the parameters, the saturated state $u = 1$ should have positive probability. Additionally, we once again note an apparent lack of physical interpretation for the values $u > 1$ in the context of the original Langmuir model.

An alternative approach to modeling the dynamics of hybridization (or absorption–desorption) reaction is to analyze directly an underlying discrete stochastic system which (2) intends to approximate. We note that in our setting we have a simple one-dimensional BD process described by one chemical species *Cmpx* i.e., the amount of probe-target complex or, in other words, the

[1] Herein we are primarily concerned with modeling the *internal* fluctuations of the stochastic system modeled by (2). Accordingly, we interpret the stochastic equation (2) in the sense of Itô calculus. For the discussion of an alternative Fokker–Planck equation (4) using the Stratonovich calculus, see e.g., [2] or, for more details, [15].

number of occupied nucleotides in the probe region. Hence, we consider a system of two coupled chemical reactions

$$\begin{aligned} &\emptyset \xrightarrow{b(\cdot)} Cmpx \\ &Cmpx \xrightarrow{d(\cdot)} \emptyset, \end{aligned} \tag{6}$$

where $b(\cdot)$ and $d(\cdot)$ are system-state-dependent birth-and-death rates, respectively. In order to describe our approach we need to specify the form of these rate functions. To this end we shall define a discrete, finite-state version of the stochastic Langmuir adsorption–desorption.

Definition 2.1 (*LBD Process*). Langmuir BD process is any BD process with the set of states $\{0, \ldots, N\}$ and the birth-and-death rates of the form

$$\begin{aligned} b(k) &= c_1(N-k) + \mathcal{C}(k, N) \\ d(k) &= c_2 k + \mathcal{C}(k, N) \end{aligned}$$

for $k = 0, \ldots, N$. Here c_1, c_2 are some positive real constants and the function $\mathcal{C}(\cdot, N)$ is assumed to be of the form

$$\mathcal{C}(k, N) = \frac{N^2}{2} g(k/N) \tag{7}$$

for $0 < k < N$ and to satisfy the boundary conditions ensuring the finiteness of the system space, i.e. $\mathcal{C}(0, N) = \mathcal{C}(N, N) = 0$.

In the LBD process the terms $c_1(N-k)$ and $c_2 k$ are linear rates of birth and death as suggested by the deterministic Langmuir model (1). The additional term $\mathcal{C}(\cdot, N)$ introduced into $b(k)$ and $d(k)$ is intended to model the noise of the *non-target* adsorption and desorption. For instance on a GeneChip $\mathcal{C}(\cdot, N)$ accounts for the competition for the same RNA targets between different probe regions with similar nucleotide sequences. Assumption (7) implies the "density-dependent" form for the rates $b(\cdot)$, $d(\cdot)$ (see, e.g., [3] chapter 11) with the noise term $\mathcal{C}(\cdot, N)$ being of higher order than the terms $c_1(N-k)$ and $c_2 k$.

Note that the (infinite) BD process with the boundary-free rates (i.e., $d(k)$ and $b(k)$ given as in Definition 2.1 but without requiring that $\mathcal{C}(0, N) = \mathcal{C}(N, N) = 0$) may be approximated by the solution of the Langmuir stochastic equation (2) for large N. Indeed, this may be informally argued as follows. Let k_t be the state of the system (6) at $t \geq 0$, described as a difference of two independent unit Poisson processes, say Y_1, Y_{-1}, with random time changes (see, e.g., [3] chapter 6)

$$k_t = k_0 + Y_1\left(\int_0^t b(k_s)\mathrm{d}s\right) - Y_{-1}\left(\int_0^t d(k_s)\mathrm{d}s\right). \tag{8}$$

Since for any unit Poisson process Y and large N we have $N^{-1/2}(Y(Nv) - Nv) \approx W(v)$ for any real v with $W(v)$ being the standard Brownian motion (SBM), the Poisson processes Y_1, Y_{-1} may be approximated for large N by independent SBM processes, say W_1, W_{-1}. Denoting $u(t) = k_t/N$ this diffusion approximation of (8) is

$$\begin{aligned} u(t) = {} & N^{-1}k_0 + N^{-1/2}W_1\left(\int_0^t \left[c_1(1-u(s)) + \frac{N}{2}g(u(s))\right]\mathrm{d}s\right) \\ & - N^{-1/2}W_{-1}\left(\int_0^t \left[c_2 u(s) + \frac{N}{2}g(u(s))\right]\mathrm{d}s\right) + \int_0^t [c_1(1-u(s)) + c_2 u(s)]\mathrm{d}s, \end{aligned}$$

which is distributionally equivalent to

$$u(t) = \int_0^t [c_1(1 - u(s)) + c_2 u(s)]\mathrm{d}s + \int_0^t \sqrt{g(u(s))}\mathrm{d}W(s) + o_P(1) \tag{9}$$

and hence in the limit to the integral version of (2) (see [3] chapter 11 for details).

Of course, depending on the form of $\mathcal{C}(\cdot, N)$ we shall have different forms of the LBD process. In order to cast our results somewhat parallel to the model (2) under a linear form of g in (3), we consider only $\mathcal{C}(k, N)$ given by the functions $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$ defined below, with the corresponding models henceforth referred to as (M_1), (M_2), (M_3), respectively.

$$\mathcal{C}_1(k, N) = c_3 N k \quad \text{for } 0 \le k < N \text{ and } \mathcal{C}_1(N, N) = 0 \tag{M_1}$$

$$\mathcal{C}_2(k, N) = c_3 N(N - k) \quad \text{for } 0 < k \le N \text{ and } \mathcal{C}_2(0, N) = 0 \tag{M_2}$$

$$\mathcal{C}_3(k, N) = c_3 k(N - k) \quad \text{for } 0 \le k \le N. \tag{M_3}$$

Note that if we disregard the boundary condition $\mathcal{C}_1(N, N) = 0$ then the model (M_1) is a discrete analogue of (2) with the choice of g given by (3) in the sense that BD process (6) (or (8)) may be approximated by (9) for large N, leading to the Fokker–Planck equation given in (4) and, consequently, to (5) with

$$\begin{aligned} c_1 &= d_1 x \\ c_2 &= d_2 \\ c_3 &= C d_1 x/2. \end{aligned} \tag{10}$$

This casting of the Eq. (2) as an approximation to (M_1) gives also additional insight into the somewhat mysterious choice of the form of function g in (3). Considering the rates in (M_1) it becomes clear that the choice of (3) is a reflection of two implicit assumptions concerning the microarray hybridization reactions, namely that (i) the level of the target-specific signal in the probe region has lower magnitude than the level of non-specific signal (i.e., signal noise) and (ii) the non-specific signal noise is proportional to the total system (i.e., probe region) size as well and the current system state and the target concentration.

Note that the model (M_2) is simply a 'reflection' of (M_1) obtained by considering the amount of unoccupied probe region $N - Cmpx$ instead of the amount of $Cmpx$. Model (M_2) is thus not concerned with saturation but rather with an *empty probe* or a *threshold* effect of the probe adsorption. This phenomena occurs when the LBD process attains an empty state with positive probability.

Note also that both (M_1) and (M_2) rate functions have discontinuities at the boundary. This is in contrast with the model (M_3) which enjoys smooth boundary conditions with no discontinuities. In general such discontinuities in rate functions for BD processes prevent the direct application of an approximation of the form (9), however it turns out that for an LBD processes (M_1)–(M_3) their stationary distributions may be approximated more directly.

3. Limit theorem

In this section we state and prove the main result of the paper, namely the limit theorem for the stationary distributions of the LBD processes under the models (M_1)–(M_3). The proof we give herein is quite elementary and is based on the fact that for one-dimensional birth-and-death processes with bounded-state space and polynomial rates, the moments of their limiting

distributions must be uniquely determined by the corresponding detailed balance (reversibility) conditions. At this point it is perhaps also worth noticing that even though herein we have restricted ourselves only to the models (M_1)–(M_3), it is not difficult to see that the method of the proof allows one to extend the result to any LBD processes with polynomial-type birth-and-death rates. This, at least in principle, then allows us to obtain limit theorems for the discrete versions of (2) with any function g continuous on $(0, 1)$ and continuously extendable to $[0, 1]$ where it may be always uniformly approximated by polynomials. However, such considerations go beyond the scope of our current investigation.

In order to state the theorem we shall need some additional notation. For $z, \gamma > 0$ denote the incomplete gamma function by $\Gamma(z, \gamma) = \int_0^\gamma s^{z-1} \exp(-s)\mathrm{d}s$ and for any $\alpha, \beta > 0$ let $IG(\alpha, \beta, 1)$ denote an incomplete gamma random variable with the density function $f_{\alpha,\beta}(x) = \Gamma(\alpha, \beta)^{-1}\beta^\alpha x^{\alpha-1}\exp(-x\beta)$ for $x \in (0, 1)$ and zero otherwise. Let us denote by $F_{\alpha,\beta}$ the distribution function of $IG(\alpha, \beta, 1)$. We introduce the following definition.

Definition 3.1 (*LIG Distribution*). We say that the random variable has the *Langmuir-incomplete gamma* (LIG) distribution with parameters α, β satisfying $\beta > \alpha > 0$ if its distribution function is given by the mixture

$$G = (1 - \pi_{\alpha,\beta})F_{\alpha,\beta} + \pi_{\alpha,\beta}\delta_1,$$

where $F_{\alpha,\beta}$ is the distribution function for $IG(\alpha, \beta, 1)$ random variable, δ_1 is the distribution function of a degenerate random variable with mass concentrated at one and

$$\pi_{\alpha,\beta} = \frac{\beta^\alpha}{\beta^\alpha + \Gamma(\alpha, \beta)(\beta - \alpha)\exp(\beta)}. \tag{11}$$

Below we denote the Langmuir-incomplete gamma distribution with parameters α, β by $LIG(\alpha, \beta)$. We shall also denote by $Beta(\alpha, \beta)$ the usual beta distribution with parameters $\alpha, \beta > 0$ and the density $h(x) = \Gamma(\alpha + \beta)\Gamma(\alpha)^{-1}\Gamma(\beta)^{-1}\int_0^1 x^{\alpha-1}(1 - x)^{\beta-1}\mathrm{d}x$ for $x \in (0, 1)$ and zero otherwise. We have the following

Theorem 1 (*Limit Theorem for a Stationary Distribution of an LBD Process*). *Let $X_N^{(i)}$ be the stationary distributions of LBD Process M_i for $i = 1, 2, 3$, and let $a = c_1/c_3$ and $b = (c_1 + c_2)/c_3$, as well as $Y_N^{(i)} = X_N^{(i)}/N$. Then, as $N \to \infty$ we have the weak convergence*

$$Y_N^{(i)} \xrightarrow{D} \mathcal{Z}_i \quad i = 1, 2, 3,$$

where the limiting random variables $\mathcal{Z}_i$ are as follows

(i) *$\mathcal{Z}_1$ is $LIG(a, b)$,*
(ii) *$\mathcal{Z}_2$ is such that $1 - \mathcal{Z}_2$ is $LIG(b - a, b)$,*
(iii) *$\mathcal{Z}_3$ is $Beta(a, b - a)$.*

Before discussing the proof of this result, some remarks are perhaps in order.

- As it shall become clear from the proof, it turns out that all LBD processes (hence, also M_1–M_3) have the correct Langmuir mean given by the stationary solution of the deterministic equation (1) with the c_1 and c_2 constants as in (10). In that sense an LBD process may be viewed as a discrete analogy of continuous models (1) and (2) with specific functions g in the latter related to LBD via (7). One should stress, however, the fundamental difference between the approach to approximating a stochastic equilibrium of a discrete system (6)

offered by Theorem 1 and that based on analyzing the equilibrium distribution of the diffusion approximation (2) outlined in (8) and (9). In Theorem 1 one considers a sequence of stationary distributions of LBD processes indexed by the size of the state space N and derives its limit as N increases. The approximation via (2) is based on approximating the entire discrete process (not just its equilibrium distribution) for large N and then deriving a stationary distribution of the approximation.

- Despite the very different models behind them, if $\pi_{a,b} \approx 0$ then (i) of Theorem 1 specializes to the result on the stationary density (5) obtained via (4). Thus when $\pi_{a,b} \approx 0$ our theorem formally justifies the use of gamma approximation for modeling hybridization reactions under the boundary-free model described by (2) and the assumption (3).
- If the condition $\pi_{a,b} \approx 0$ is not satisfied then there may be a significant difference between the stationary distribution obtained from the boundary-free-type analysis via the stochastic equation (2) and the LBD process analysis. This is due to the fact that the LBD analysis takes properly into account the discontinuities in the rate functions whereas the continuous model (2) does not.
- The theorem indicates that both (M_1) and (M_2) models which incorporate the linear noise term $\mathcal{C}(\cdot, N)$ into their rate functions are amenable to the gamma-type approximation of their stationary distributions perhaps after some adjustment for the bounded state space. In contrast, the model (M_3) with the quadratic and 'boundary symmetric' noise term yields a different type of stationary distribution (i.e, beta) with no boundary effects.

It seems that there are several ways of arriving at the result of the theorem. Herein we have chosen the method of the proof which is perhaps slightly convoluted but on the other hand almost completely elementary and thus fully accessible to readers without extensive background in stochastic processes theory.

In order to provide a proof of Theorem 1 we shall need two auxiliary results stated as Lemmas 1 and 2. The first one of them concerns some elementary properties of the moments of a LIG distribution.

Lemma 1. *Let Z be a random variable distributed according to LIG(α, β). Then for any integer $m \geq 0$ we have*

$$EZ^{m+1} = \frac{m+\alpha}{\beta} EZ^m - \frac{m}{\beta}\pi_{\alpha,\beta}. \tag{12}$$

Proof of Lemma 1. Let W be a random variable distributed according to $IG(\alpha, \beta, 1)$. Elementary calculation based on the integration by parts shows that for any integer $m \geq 0$

$$EW^{m+1} = \frac{m+\alpha}{\beta} EW^m - \frac{\beta^{\alpha-1}\exp(-\beta)}{\Gamma(\alpha,\beta)}.$$

In view of the above and by the definition of Z we have for any integer $m \geq 0$

$$\begin{aligned} EZ^{m+1} &= (1-\pi_{\alpha,\beta})EW^{m+1} + \pi_{\alpha,\beta} \\ &= (1-\pi_{\alpha,\beta})\left[\frac{m+\alpha}{\beta}EW^m - \frac{\beta^{\alpha-1}\exp(-\beta)}{\Gamma(\alpha,\beta)}\right] + \pi_{\alpha,\beta} \\ &= \frac{m+\alpha}{\beta}\left[(1-\pi_{\alpha,\beta})EW^m + \pi_{\alpha,\beta}\right] + \left[1 - \frac{m+\alpha}{\beta}\right]\pi_{\alpha,\beta} \end{aligned}$$

$$
\begin{aligned}
&- (1-\pi_{\alpha,\beta})\frac{\beta^{\alpha-1}\exp(-\beta)}{\Gamma(\alpha,\beta)}\\
&= \frac{m+\alpha}{\beta}\left[(1-\pi_{\alpha,\beta})EW^m + \pi_{\alpha,\beta}\right] - \frac{m}{\beta}\pi_{\alpha,\beta} - \frac{\beta-\alpha}{\beta}\pi_{\alpha,\beta}\\
&\quad - (1-\pi_{\alpha,\beta})\frac{\beta^{\alpha-1}\exp(-\beta)}{\Gamma(\alpha,\beta)}\\
&= \frac{m+\alpha}{\beta}EZ^m - \frac{m}{\beta}\pi_{\alpha,\beta}. \quad \square
\end{aligned}
$$

Our second lemma is as follows.

Lemma 2. *Let $\alpha, \beta > 0$ be arbitrary. Consider $N \to \infty$. For any non-increasing real sequence $\alpha_N \downarrow \alpha > 0$ satisfying $(\alpha_N - \alpha)\log N \to 0$ and any real sequence $\beta_N \to \beta > 0$ we have*

$$
N^{-\alpha_N}\sum_{k=0}^{N}\frac{\Gamma(\alpha_N+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k \to \int_0^1 x^{\alpha-1}\mathrm{e}^{-\beta x}\mathrm{d}x = \beta^{-\alpha}\Gamma(\alpha,\beta).
$$

Proof of Lemma 2. Assume first that $\alpha_N \equiv \alpha$. For given $\alpha, \beta > 0$ define $k(N) = [\delta \log N]$ where δ is a fixed positive number such that $\delta < \alpha$ and $[x]$ denotes the largest integer not greater than x. Write

$$
\begin{aligned}
N^{-\alpha}\sum_{k=0}^{N}\frac{\Gamma(\alpha+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k &= N^{-\alpha}\sum_{k=0}^{k(N)}\frac{\Gamma(\alpha+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k\\
&\quad + N^{-\alpha}\sum_{k=k(N)+1}^{N}\frac{\Gamma(\alpha+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k\\
&= \mathrm{(I)} + \mathrm{(II)}.
\end{aligned}
$$

We first show $(I) \to 0$ as $N \to \infty$. To this end, note

$$
\begin{aligned}
\mathrm{(I)} &\le N^{-\alpha}\sum_{k=0}^{k(N)}\frac{\Gamma(\alpha+k)}{k!} \le N^{-\alpha}\sum_{k=0}^{k(N)}\frac{(\alpha+k)^k}{k!}\\
&\le N^{-\alpha}\sum_{k=0}^{k(N)}\frac{(\alpha+k(N))^k}{k!} \le N^{-\alpha}\mathrm{e}^{(\alpha+k(N))} \to 0 \quad \text{as } N \to \infty.
\end{aligned}
$$

Now we argue that

$$
\mathrm{(II)} \to \beta^{-\alpha}\Gamma(\alpha,\beta) \quad \text{as } N \to \infty. \tag{13}
$$

To this end, recall the following version of the Gauss formula

$$
\frac{\Gamma(\alpha+k)}{k!k^{\alpha-1}} \to 1 \quad \text{as } k \to \infty. \tag{14}
$$

In view of the above it follows that for any given $\varepsilon \in (0,1)$ and N sufficiently large we have

$$
(1-\varepsilon)N^{-\alpha}\sum_{k=k(N)+1}^{N}k^{\alpha-1}\mathrm{e}^{-k\beta/N} \le \mathrm{(II)} \le (1+\varepsilon)N^{-\alpha}\sum_{k=k(N)+1}^{N}k^{\alpha-1}\mathrm{e}^{-k\beta/N}.
$$

Since the expression $N^{-\alpha}\sum_{k=k(N)+1}^{N} k^{\alpha-1}\mathrm{e}^{-k\beta/N} = N^{-1}\sum_{k=k(N)+1}^{N}(k/N)^{\alpha-1}\mathrm{e}^{-k\beta/N}$ is seen to be the Riemann sum for $\beta^{-\alpha}\Gamma(\alpha,\beta)$ taking $N\to\infty$ gives

$$(1-\varepsilon)\beta^{-\alpha}\Gamma(\alpha,\beta) \le \lim_{N}(\mathrm{II}) \le (1+\varepsilon)\beta^{-\alpha}\Gamma(\alpha,\beta).$$

Since ε may be taken arbitrarily close to zero, the relation (13) follows and yields the assertion of the lemma with $\alpha_N \equiv \alpha$. To complete the proof for an arbitrary sequence α_N note that we need in essence only to argue that $N^{\alpha_N-\alpha}\to 1$ as $N\to\infty$ (this follows by assumption) and that the formula (14) holds with α replaced by α_k (since α_k is monotone). By the continuity of gamma function, $\Gamma(\alpha)/\Gamma(\alpha_N)\to 1$ as $N\to\infty$. This and (14) entail

$$\frac{\Gamma(\alpha_k)\alpha(\alpha+1)\cdots(\alpha+k-1)}{k!k^{\alpha-1}} \to 1 \quad \text{as } k\to\infty.$$

The relationship (14) with α replaced by α_k will now follow if we can argue that

$$\prod_{s=0}^{k}\frac{\alpha+s}{\alpha_k+s} \to 1 \tag{15}$$

as $k\to\infty$. To this end, note that

$$\log\left(\prod_{s=0}^{k}\frac{\alpha+s}{\alpha_k+s}\right) = \sum_{s=0}^{k}\log\left(\frac{\alpha_k+s}{\alpha+s}\right) \le (\alpha_k-\alpha)\sum_{s=0}^{k}\frac{1}{s+\alpha} \le 2(\alpha_k-\alpha)\log k \to 0$$

by our assumption on α_k and thus (15) follows. This, however, yields the assertion of the lemma, since for sufficiently large N

$$\begin{aligned} N^{-\alpha}\sum_{k=k(N)}^{N}\frac{\Gamma(\alpha+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k &\le N^{-\alpha}\sum_{k=k(N)}^{N}\frac{\Gamma(\alpha_N+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k \\ &\le N^{-\alpha}\sum_{k=k(N)}^{N}\frac{\Gamma(\alpha_k+k)}{k!}\left(1-\frac{\beta_N}{N}\right)^k \end{aligned}$$

and we have just shown that the first and the last of the expressions above tend to $\beta^{-\alpha}\Gamma(\alpha,\beta)$ as $N\to\infty$. □

Having established the assertions of the lemmas above, we are finally in a position to prove the result given in Theorem 1.

Proof of Theorem 1 (*Part* (i)). Denote by X the random variable $X_N^{(1)}$ and set $P(X=k)=p(k)$ for $k=0\ldots,N$. Let $m\ge 0$ be an integer. Multiplying by k^m the detailed balance equation

$$p(k+1)d(k+1) = p(k)b(k) \tag{16}$$

and then summing over $k=0,\ldots,N-1$ we obtain under the (M_1) model

$$\begin{aligned} &(c_2+c_3N)\sum_{k=0}^{N-1}k^m(k+1)p(k+1) - c_3(N-1)^l p(N)N^2 \\ &\quad = \sum_{k=0}^{N-1}k^m\left[p(N)(c_1(N-k)+c_3Nk)\right]. \end{aligned}$$

Expanding now $k^m = (k+1-1)^m$ on the right-hand side we get

$$(c_2 + c_3 N) \sum_{k=0}^{N-1} \sum_{s=0}^{m} \binom{m}{s} (-1)^{m-s} (k+1)^{s+1} p(k+1) \; - c_3 (N-1)^m N^2 p(N)$$
$$= \sum_{k=0}^{N} [c_1(N-k) + c_3 N k] k^m p(k) - c_3 N^{m+2} p(N).$$

Denoting $\tilde{\mu}_m(N) = EX^m$ we may rewrite the above relationship as

$$(c_2 + c_3 N) \sum_{s=0}^{m} \binom{m}{s} (-1)^{m-s} \tilde{\mu}_{m+1}(N) - c_3 (N-1)^m N^2 p(N)$$
$$= c_1 N \tilde{\mu}_m(N) + (c_3 N - c_1) \tilde{\mu}_{m+1}(N) - c_3 N^{m+2} p(N).$$

We set $\mu_m = \lim_{N\to\infty} \tilde{\mu}_m(N)/N^m$. Note that dividing both sides by N^{m+1} and taking $N \to \infty$ give

$$(c_1 + c_2)\mu_{m+1} = (m c_3 + c_1)\mu_m - c_3 m p^*,$$

where

$$p^* = \lim_{N\to\infty} p(N). \tag{17}$$

Assuming for a moment that p^* exists and is finite (this follows from (20)) we see that these considerations give the following recursive relationship for the limiting moments of $Y_N^{(1)}$

$$\mu_{m+1} = \frac{m + c_1/c_3}{(c_1 + c_2)/c_3} \mu_m - \frac{m p^*}{(c_1 + c_2)/c_3} \quad \text{for } m = 0, 1 \ldots$$

or in terms of a, b

$$\mu_{m+1} = \frac{m + a}{b} \mu_m - \frac{m}{b} p^* \quad \text{for } m = 0, 1 \ldots . \tag{18}$$

We note that in view of $\mu_0 = 1$ the solution to the above recursive equation is unique. Moreover, we note that since the support of the sequence $\{Y_N^{(1)}\}_{N=1}^{\infty}$ is contained in the closed interval $[0, 1]$ then (i) the corresponding sequence of probability measures is tight and (ii) any of its weak limits must be a probability measure whose moments satisfy (18). Since the probability distributions on bounded intervals are uniquely determined by their moments it follows that as $N \to \infty$

$$Y_N^{(1)} \xrightarrow{D} \mathcal{Z}_1 \tag{19}$$

for some random variable $\mathcal{Z}_1$ with moments μ_m given by (18). To complete the proof of part (i) we need to show only that $\mathcal{Z}_1$ is a $LIG(a, b)$ random variable as given in Definition 3.1. Since $\mathcal{Z}_1$ is identified completely by its moments, it suffices to show that the moments of the random variable $LIG(a, b)$ satisfy the recursive relation (18). This follows by Lemma 1 provided that

$$p^* = \pi_{a,b}, \tag{20}$$

where $\pi_{a,b}$ is given by (11).

In order to argue (20) we again consider the detailed balance equation (16)

$$p(N) = p(N-1)b(N-1)/d(N) = \frac{p(N-1)b(N-1)/d(N)}{p(N-1)b(N-1)/d(N) + \sum_{k=0}^{N-1} p(k)}$$
$$= \frac{p(N-1)b(N-1)/d(N)}{p(N-1)b(N-1)/d(N) + \sum_{k=0}^{N-1} p(k)} = \frac{\Delta_N}{\Delta_N + 1}, \tag{21}$$

where we define

$$\Delta_N = \frac{p(N-1)b(N-1)/d(N)}{\sum_{k=0}^{N-1} p(k)}.$$

We note that again by (16) we get under the model (M_1) the following form of Δ_N

$$\Delta_N = \frac{c_2 + c_3 N}{c_2 N!}\left(\frac{c_3 N - c_1}{c_2 + c_3 N}\right)^N \frac{\prod_{s=0}^{N-1}(s + \frac{c_1 N}{c_3 N - c_1})}{\sum_{k=0}^{N-1} \frac{1}{k!}\left(\frac{c_3 N - c_1}{c_2 + c_3 N}\right)^k \prod_{s=0}^{k-1}\left(s + \frac{c_1 N}{c_3 N - c_1}\right)}$$
$$= \frac{c_2 + c_3 N}{c_2 N!}\left(\frac{c_3 N - c_1}{c_2 + c_3 N}\right)^N \frac{\Gamma(N + \frac{c_1 N}{c_3 N - c_1})}{\sum_{k=0}^{N-1} \frac{1}{k!}\left(\frac{c_3 N - c_1}{c_2 + c_3 N}\right)^k \Gamma\left(k + \frac{c_1 N}{c_3 N - c_1}\right)}.$$

Denote

$$a_N = \frac{c_1 N}{c_3 N - c_1} \qquad b_N = \frac{(c_1 + c_2)N}{c_3 N + c_2}$$

then

$$\Delta_N = \frac{(b - a + N)}{(b - a)N} \frac{\Gamma(N + a_N)}{N! N^{a_N - 1}} \frac{(1 - b_N/N)^N}{N^{-a_N} \sum_{k=0}^{N-1} \frac{1}{k!}(1 - b_N/N)^k \Gamma(k + a_N)}.$$

Applying now the Gauss formula (14) and using the result of Lemma 2 we conclude that

$$\lim_N \Delta_N = \frac{b^a}{\exp(b)\Gamma(a, b)}$$

and hence (20) follows by (21) which completes the proof of part (i) of the theorem.

Part (ii). The result follows by applying part (i) to the random variable $N - X_N^{(2)}$.

Part (iii). For the proof of the last part of the theorem we denote now by X the random variable $X_N^{(3)}$ and otherwise retain the notation from part (i). Multiplying (16) by k^m and summing as before, we obtain under (M_3)

$$\sum_{k=0}^{N} k^m p(k+1)[c_2(k+1) + c_3(k+1)(N-k-1)]$$
$$= \sum_{k=0}^{N} k^m p(k)[c_1(N-k) + c_3 k(N-k)].$$

The above, by an argument similar to the one used in (i), gives the relationship

$$(c_2 + c_3 N) \sum_{s=0}^{m} (-1)^{m-s} \binom{m}{s} \tilde{\mu}_{s+1} - c_3 \sum_{s=0}^{m} (-1)^{m-s} \binom{m}{s} \tilde{\mu}_{s+2}$$
$$= c_1 N \tilde{\mu}_m - c_1 \tilde{\mu}_{m+1} + c_3 N \tilde{\mu}_{m+1} - c_3 \tilde{\mu}_{m+2}.$$

Denoting $\mu_m = \lim_{N\to\infty} \tilde{\mu}_m(N)/N^m$, dividing both sides by N^{m+1} and taking $N \to \infty$ give a somewhat simpler formula than the (18) recursion formula for the limiting moments, namely

$$\mu_{m+1} = \frac{m + c_1/c_3}{m + (c_1 + c_2)/c_3} \mu_m \quad \text{for } m = 0, 1 \ldots.$$

Writing the above in terms of a, b

$$\mu_{m+1} = \frac{m + a}{m + b} \mu_m \quad \text{for } m = 0, 1 \ldots$$

we obtain the familiar relationship between the moments of $Beta(a, b - a)$ distribution. This, along with the tightness of measures argument similar to the one used in (i) completes the proof of part (iii). □

4. Conclusions

Herein we have derived a limit theorem for stationary distributions of some special birth-and-death processes related to the Langmuir dynamic adsorption–desorption model. Such a model is of interest in the context of the microarray hybridization reactions if one may assume that the fluorescence signal on the array is approximately a realization of a chemical Langmuir equilibrium of the adsorption and desorption reactions between the target mRNA molecules and the DNA probes. Whereas this assumption may be questionable for long (hundred basis or more) probes, it seems reasonable for the short ones, like e.g., the 25-mers used on many Affymetrix chips. Indeed, in the context of Affymetrix GeneChip arrays, the gamma-type approximation to the gene expression data based on an ad hoc Langmuir-like equilibria argument has been proposed in the literature as a way of enhancing the data analysis. Our current result gives a rigorous justification of the use of truncated-gamma- and beta-type distributions in order to approximate the fluorescence readings of the probe-RNA complexes obtained in the course of an Affymetrix microarray experiment. It also explains some experimentally observed behavior of these readings like e.g. the signal saturation and the signal thresholding phenomena.

The potential usefulness of our approximation results stems also from the fact that they allow one to describe the theoretical means of measured fluorescence intensity readings by three parameter hyperbolic response functions which can be obtained as solutions of the corresponding deterministic Langmuir equations. In general, these response functions for specific probes shall be only sequence dependent and could be therefore used universally in all experiments involving a particular probe sequence. Our results imply also that the fold changes in RNA-target concentration are not linearly related to fold changes in fluorescence intensity readings, as is often generally assumed.

As pointed out by some authors (cf. [16]) the formidable challenge in microarray experiments is to establish a reliable algorithm for extracting the true RNA concentration measurements from the probe fluorescence intensity readings. We believe that the results of this paper could perhaps take us a step closer to that goal.

Acknowledgments

This research was partially sponsored by the National Science Foundation under grant DMS0553701 as well as by the Center for Environmental Genomics and Integrative Biology at the University of Louisville which receives funding from the National Institute of Environmental Health Sciences under grant 1P30ES014443.

The authors wish to acknowledge an anonymous referee whose comments helped them in improving the original manuscript.

References

[1] C. Burden, Y. Pittelkow, S. Wilson, Statistical analysis of adsorption models for oligonucleotide microarrays, Statistical Applications in Genetics and Molecular Biology 3 (35) (2004).
[2] B. Dennis, G.P. Patil, The gamma distribution and weighted multimodal gamma distributions as models of population abundance, Mathematical Biosciences 68 (1984) 187–212.
[3] S.N. Ethier, T.G. Kurtz, Markov processes, in: Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons Inc., New York, 1986, Characterization and convergence.
[4] J.E. Forman, I.D. Walton, D. Stern, R.P. Rava, M.O. Trulson, Thermodynamics of duplex formation and mismatch discrimination on photolithographically synthesised oligonucleotide arrays, in: N.B. Leontis, J. SantaLucia (Eds.), Molecular Modeling of Nucleic Acids, in: ACS Symposium Series, vol. 682, Am. Chem. Soc., Washington, DC, 1998.
[5] D. Hekstra, A.R. Taussig, M. Magnasco, F. Naef, Absolute mrna concentrations from sequence-specific calibration of oligonucleotide arrays, Nucleic Acids Research 31 (2003) 1962–1968.
[6] G.A. Held, G. Grinstein, Y. Tu, Modeling of dna microarray data by using physical properties of hybridization, Proceedings of the National Academy of Science 100 (2003) 7575–7580.
[7] M.A. Katsoulakis, A.J. Majda, V. Vlachos, Coarse–grained stochastic processes and monte carlo simulations in lattice systems, Journal of Computational Physics 186 (2003) 250–278.
[8] I. Langmuir, The constitution and fundamental properties of solids and liquids. Part I. Solids, Journal of the American Chemical Society 38 (1916) 2221–2295.
[9] A.J. Majda, I. Timofeyev, E. Vanden-Eijnden, Systematic strategies for stochastic mode reduction in climate, Journal of the Atmospheric Sciences 60 (2003) 1705–1722.
[10] M. Newton, A. Noueiry, D. Sarkar, P. Ahlquist, Detecting differential gene expression with a semiparametric hierarchical mixture method, Biostatistics 5 (2004) 155–176.
[11] M.A. Newton, C.M. Kendziorski, Parametric empirical Bayes methods for microarrays, in: The Analysis of Gene Expression Data, in: Stat. Biol. Health, Springer, New York, 2003, pp. 254–271.
[12] M.A. Newton, C.M. Kendziorski, C.S. Richmond, F.R. Blattner, K.W. Tsui, On differential variability of expression ratios: Improving statistical inference about gene expression changes from microarray data, Journal of Computational Biology 8 (2001) 37–52.
[13] I. Pigeot, K. Bammann, A. Reineke, N. Wawro, A. Zierer, Statistical methods in genetics: from microarrays to genetic epidemiology—an overview, Jahresbericht der Deutschen Mathematiker Vereinigung 106 (1) (2004) 3–38.
[14] B. Sinclair, Everything's great when it sits on a chip — a bright future for dna arrays, The Scientist 13 (11) (1999) 18–20.
[15] N. van Kampen, Stochastic processes in physics and chemistry, Elsevier Science, Amsterdam, The Netherlands, 1992.
[16] Z. Wu, R. Irizarry, Stochastic models inspired by hybridization theory for short oligonucleotide arrays, Journal of Computational Biology 12 (6) (2005) 882–893.