Scientific Research Publishing

Applied Mathematics, 2017, 8, 1128-1139
http://www.scirp.org/journal/am
ISSN Online: 2152-7393
ISSN Print: 2152-7385

Novel Quantitative Approach for Predicting mRNA/Protein Counts in Living Cells

Henri C. Jimbo[1*], Seraphin I. Ngongo[2], Achille Mbassi[3], Nicolas G. Andjiga[4]

[1]Department of Applied Mathematics and Statistics, Waseda University, Tokyo, Japan
[2]Department of Applied Mathematics, University of Paris 1, Pantheon Sorbonne, Paris, France
[3]Department of Urology, Central Medical Hospital, CEMAC Region, Yaounde, Cameroon
[4]Department of Mathematics, Higher National School (ENS), Yaounde, Cameroon
Email: *jimbo_maths@yahoo.conm

How to cite this paper: Jimbo, H.C., Ngongo, S.I., Mbassi, A. and Andjiga, N.G. (2017) Novel Quantitative Approach for Predicting mRNA/Protein Counts in Living Cells. *Applied Mathematics*, **8**, 1128-1139.
https://doi.org/10.4236/am.2017.88085

Received: April 21, 2017
Accepted: August 18, 2017
Published: August 21, 2017

Abstract

One of the most complex questions in quantitative biology is how to manage noise sources and the subsequent consequences for cell functions. Noise in genetic networks is inevitable, as chemical reactions are probabilistic and often, genes, mRNAs and proteins are present in variable numbers per cell. Previous research has focused on counting these numbers using experimental methods such as complex fluorescent techniques or theoretical methods by characterizing the probability distribution of mRNAs and proteins numbers in cells. In this work, we propose a modeling based approach; we build a mathematical model that is used to predict the number of mRNAs and proteins over time, and develop a computational method to extract the noise-related information in such a biological system. Our approach contributes to answering the question of how the number of mRNA and proteins change in living cells over time and how these changes induce noise. Moreover, we calculate the entropy of the system; this turns out to be important information for prediction which could allow us to understand how noise information is generated and expanded.

Keywords

Applied Mathematics, Embedded Control System, Genetic Algorithm, Optimization, Biodynamics, Stochastic Modelling, Simulations, Stochasticity, Bioengineering and Medicine

1. Introduction

Randomness, or noise, in biological systems has long been predicted from basic physical principles [1]-[8] and later on by observations of phenotype heteroge-

neity [7]. But the confirmation came later with [9], [10] and [11] who showed that mRNA and protein variability may lead to important a source of noise in biology. Researches in [12], [13], [14] and [15] have reported that the number of proteins translated from an mRNA obeys a geometric distribution but the distribution describing the number of protein remaining once mRNA is degraded will no longer be geometric. Various techniques have so far been used to monitor and capture those numbers among which fluorescent probes or green fluorescent protein variants which allow the quantification of protein levels in living cells by flow cytometry or fluorescence microscopy [16] [17]. The first quantitative study collectively examines the noise associated with the principal step of central dogma of molecular biology in replication, gene activation, transcription, translation and the enslaving intracellular environment, and suggested that autorepression of replication and transcriptions suppresses noise. This then leads to examination (by analysis, modelling and simulation) of the role of noise in biology relying on the similarity between biological and engineering systems—see [7], [10] and [18]. In general, noise may be considered either intrinsic or extrinsic to a specific gene circuit, and within a specific gene circuit there are three different effects of noise: i) noise is negligible with little or no influence over function; ii) noise is detrimental to function and gene circuit; iii) noise is important for circuit function, and by using simple assumptions, it is possible to evaluate these effects. The assumption we use in this paper is dynamic correlation between the noise level of molecules (mRNA/protein) and the change in the probability of having those molecules in given interval of time. Our paper is organised as follows. In Section 2, we introduce our model of the dynamic of the number of mRNA and proteins after a brief review of previous models. In Section 4, we present our method and algorithm for solving the (mRNA and protein) prediction problem. In Section 5, we present the simulation results, followed by a discussion of those results, and end this work in Section 6 with a short conclusion.

2. Somes Examples and Motivations

2.1. Birth-Death Model

To understand noise in biological systems, biochemical circuits and genetic networks are often used as the measured noise properties to elucidate the structure and the function of the underlying gene circuit [6] [8]. Also recent researches [13] and [14] have clearly established the existence of dynamic correlation between genetic network and mRNA/protein variability. In the next section we will present previous models with their strengths and weaknesses. The preliminary model used was a simple birth-death Markov process which captures noise in a biochemical process. This model showed that noise in the population was a consequence of the change in the parameters of the system over time and was used to explore the temporal change of the number of proteins in a biological system. The time course of the number of proteins was modelled consequently by the equation

$$\frac{\mathrm{d}n(t)}{\mathrm{d}t} = \alpha - \gamma n(t) \tag{1}$$

with parameters α representing the rate of production and γ the rate of decay of number of proteins $n(t)$. However, such continuous time formulation neglects the discrete nature of proteins and the random timing of molecular transition [17] because the actual time evolution may follow any one of a number of trajectories, and hence sufficiently many trajectories have to be examined to obtain statistics that converge. In the next section a probabilistic approach using the extended versions of Kolmogorov's equations is used to explore randomness in the system.

2.2. Kolmogorov's Equations Based Model

In general, the Kolmogorov's equations are used as master equation to capture the distribution of chemical components of the gene circuit over time. The state of the system is defined by a vector $n(t) = (n_1, n_2, \cdots, n_N)$, where $n_i(t)$ represents the i-th component of molecule n at time t; a_i and v_i are internal parameters representing respectively the propensity of the dynamic and the actual change in x_i, resulting from the change in the previous state The probability, $p(n,t)$, that the system evolves into the state $n(t) = (n,t)$ at time t is described by the following partial differential equation:

$$\frac{\partial p(n,t)}{\partial t} = \sum_{j=1}^{N} a_j (n - v_j)\, p(n - v_j) - a_j(n)\, p(n) \tag{2}$$

This equation makes sense only if we assume that the probability for two or more reactions to occur in the time interval dt is negligible compared to the case when only one reaction occurs. In addition, (2) can only be solved numerically for relatively simple systems. In a recent work by [15], a similar mathematical model was used for gene expression and an approximate solution was proposed to the PDE; the model was based on the assumption that gene expressions are Brownian motions. They considered a two-stage model of gene expression, assuming that the promoter was always active and so had two stochastic variables (the number of mRNA and the number of proteins). The probability of having m mRNA and n proteins at time t was given by the following master equation:

$$\begin{aligned} \frac{\partial P_{m,n}}{\partial t} = {} & v_0 \left(P_{m-1,n} - P_{m,n} \right) + v_1 \left(P_{m,n-1} - P_{m,n} \right) \\ & + d_0 \left[(m+1) P_{m+1,n} - m P_{m,n} \right] \\ & + d_1 \left[(n+1) P_{m,n+1} - n P_{m,n} \right] \end{aligned} \tag{3}$$

The meanings of the rates in (3) are: v_0 is the probability per unit time of transcription, v_1 the probability per unit of translation, d_0 the probability per unit time of degradation of an mRNA, and d_1 the probability per unit time of degradation of a protein. The authors use a particular generating function and transform (3) into a first order PDE which is solved using a simple approxima-

tion. However, this model works only on a single cell, and all rates ν_0, ν_1, d_0 and d_1 are fixed over time. Further, by assuming that the protein synthesis occurs in bursts $(m=0)$, the authors derive the Kolmogorov (master) equation for gene expression that considers only proteins, by implicitly including mRNAs (since n and m seem to be correlated over time). In the next section, we shall re-examine this model and propose a new one in order to overcome the above limitations.

3. The New Model

Our setup is motivated by the necessity to overcome the limitations from the previous models by increasing the cell numbers and relaxing the restriction on constant parameters. We propose a new, flexible, and more general, model for a population of N cells. This model is an extended version of the previous Kolmogorov's equation with additional cell-dependent constraints.

$$\begin{cases} \dfrac{\partial p(m,t/m_0,t_0)}{\partial t} = \displaystyle\sum_{j=1}^{N} a_j(m-\nu_j)\, p(m-\nu_j,t/m_0,t_0) - a_j(m)\, p(m,t/m_0,t_0) \\ \dfrac{\partial p(n,t/n_0,t_0)}{\partial t} = \displaystyle\sum_{j=1}^{N} a_j(n-d_j)\, p(n-d_j,t/n_0,t_0) - a_j(n)\, p(n,t/n_0,t_0) \end{cases} \tag{4}$$

The parameters of the model have an autoregressive form:

$$\begin{cases} a_j(m) = \vartheta_1 a_{j-1}(m) + \theta_1 \\ a_j(n) = \vartheta_1 a_{j-1}(n) + \theta_1 \end{cases} \tag{5}$$

The transcription, translation and degradation rates are assumed to vary from one cell to another as

$$\nu_j = \nu_0 \mathrm{e}^{-0.005j} \quad \text{and} \quad d_j = d_0 \mathrm{e}^{-0.001j} \tag{6}$$

We assume for $k \in [0,N]$, the first $\nu_1,\cdots,\nu_k$ are sequences of transcription rates and the late $\nu_{k+1},\cdots,\nu_N$ are sequences of translation rates with ν_0 being the fixed initial rate. Our model, which is composed of the Equations (4)-(6), is well adapted to various real biological promoter change. We shall notice that Equation (4) is a system of 200 equations with 100 by 2 unknowns, which is likely to be only numerically solvable after some good approximations. To efficiently predict the number of mRNAs and proteins over time, we shall rely on the following assumptions.

***Proposition* 1.** Over time the number of mRNAs/Proteins is perfectly correlated with the probability mass functions of mRNAs $m(t)$ and proteins $n(t)$ respectively. That is, $m(t) = p(m,t)m_0$ and $n(t) = p(n,t)n_0$, where m_0 and n_0 are initial measurements.

Proof: The proof follows from our algorithm and solution in this paper. □

***Proposition* 2.** Let $n = n(t)$ be the number of proteins and $\eta = \eta(n,\Delta t)$ be the noise generated by n proteins (or m for mRNAs) in the same time interval Δt. Then there exists a unique constant C such that $\eta(n,\Delta t) = Cp(n,\Delta t)$ which means that noise is cells is proportionally correlated to the probability distribu-

tion of protein and mRNA numbers.

Proof:

Let $\eta = \eta(n, \Delta t)$, $\Delta n(t) = n(t) - n(t-1)$ be respectively the noise and the number of proteins in a cell. By the simple decomposition of numbers of mRNA/proteins, $p(\Delta n(t)) = p(n, \Delta(t))$ and $p(\Delta n(t)) = p(n(t)) - p(n(t-1))$, (by the additivity property of probability distribution. We also have, using the definition, that $\eta = \frac{\Delta n}{\Delta t}$, $p(\Delta n) = \frac{\Delta n}{N}$ and $\sum n = N$. This implies that $p(\Delta n(t)) = \frac{\Delta n}{N} \Rightarrow Np(\Delta n(t)) = \Delta n(t) = n(t) - n(t-1)$ multiplying the right side of above with $\frac{\Delta t}{\Delta t}$ and we obtain $N \times p(\Delta n(t)) = \frac{n(t) - n(t-1)}{\Delta t} \Delta t$.

since $\eta(n, \Delta t) = \frac{n(t) - n(t-1)}{\Delta t}$

and $N \times p(\Delta n(t)) = \eta(n, \Delta t) \times \Delta t$

thus $\frac{N}{\Delta t} \times p(\Delta n(t)) = \eta(n, \Delta t)$

leading to $C \times p(n, \Delta(t)) = \eta(n, \Delta t)$

Finally we conclusion that $C = \frac{N}{\Delta t}$. □

Here we put $\Delta t = \frac{T}{N_0}$ where T is the total time, N_0 the total number of points in the simulation and N is the total number of mRNA or proteins in a single cell. In the next section we introduce our method and algorithm for solving Equations (4)-(6).

4. Method and Justification

We propose a straightforward method of solving the above problem based on numerical approximation via the following algorithm. As the analytical solution to Equation (4) is (at least) hard to obtain, even for a "reasonable" number of cells, a numerical algorithm using an adapted stochastic simulation approach is proposed in this paper. In our algorithm, two random variables $m(t)$ and $n(t)$ determine the temporal evolution of the system. The variable τ_k is the time for the next event to occurs, the probability density of an event (appearance of $m(t)$ or $n(t)$) is evaluated based upon our model (4), so as to give a better flexibility and applicability to the approach in comparison with previous ones. The main purpose of creating such an algorithm is to simultaneously simulate the process noise, while predicting the online *probability mass function* (≈probability density) of each event over time. An important assumption here is that the hypothetical probability distribution functions (p.m.fs) of the translation and transcription rates are of the form $\nu_j \sim N(2, 0.05)$ and the mRNA and protein degradation rates are $d_j \sim N(2, 0.05)$. This is in line with the exis-

tence of a one-to-one relation between the dynamic and distribution for predictable dynamic systems. We will present our algorithm in the next section of our work.

Our Algorithm

Input: Initial data m_0, n_0

Outputs: P_m, P_n

1. Set $a_0(m) := m_0, a_0(n) := n_0$.
2. For $j = 1{:}k$ do [k = number of iterations]
   a. Let $\nu_j \sim N(2, 0.05), d_j \sim N(2, 0.05)$ be the changes associated to a single event;
   b. Compute $a_j(m) = \theta a_{j-1}(m) + \vartheta, a_j(n) = \theta a_{j-1}(n) + \vartheta$;
   c. Compute $\alpha_m(j) = \vartheta \alpha_m(j-1) + \theta, \alpha_n(j) = \vartheta \alpha_n(j-1) + \theta$;
   d. Compute

$$P_m(j) = \frac{-\nu_j(\alpha_m(j) - \nu_j) + \xi_1(j)}{1+\nu_j}, P_n(j) = \frac{-d_j(\alpha_n(j) - d_j) + \xi_2(j)}{1+d_j},$$

   where $\xi_1(j) \sim Po(10), \xi_2(j) \sim Po(10)$;

   e. Normalize P_m, P_n.
3. Output P_m, P_n.

End

5. Simulations and Results

The initial data here is a matrix of randomly generated numbers between one and fifty for mRNAs and between one and forty for proteins. The rows represent the cell numbers and the columns are the number of mRNAs/proteins counted at each time interval. Therefore we have 100 cells (population) and 50 samples taken at a time interval of one unit, and the total time of 50 time units in the entire population; (a unite could be second, minute or hour depending on the experiment). The bar and image pots of the initial data are shown in Figure 1.

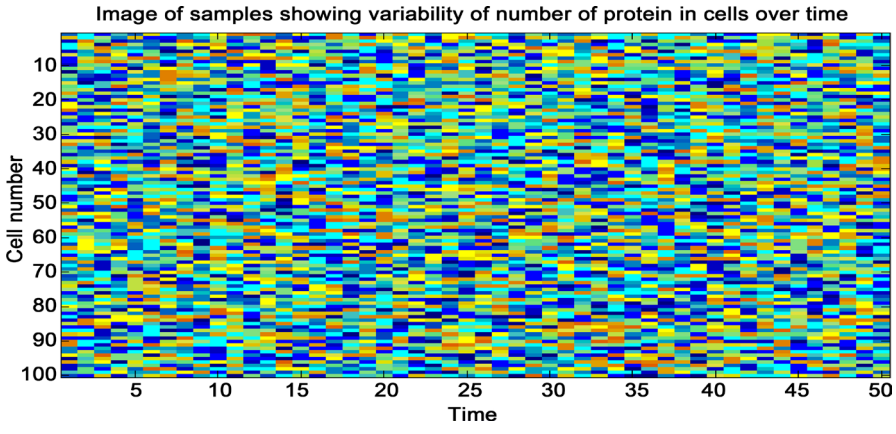

Figure 1. These subfigures give plots of the initial randomly generated data with 50 samples for proteins. A sample is the number of protein in cells for a fixed interval of time. This image plot support the presence of various level of noise in the biodynamic system.

5.1. Results

Our results will show various figures related to our solutions. We first plot the variability of the number of protein in cells over time for a sample of 50. Next, we plot the solutions of (4) over time and explain their relevance for our work. pmf (probability mass function) of the mRNA and proteins in separate graphs for each sample, and further we plot the histograms of the distribution and finally the scatter plot of P_n against P_m. Our observations are presented in the caption of each figure.

It can be seen that all probability values are between 0.1 and 0.9 and do not overlap in most of the cases; this is an indication that mRNAs and proteins number may be dynamically dependent, and therefore correlated. Next, we predict the number of mRNAs and proteins m_j, n_j using a straightforward probabilistic concept which states that "a good value of m (or n) depends on a good guess of p". The prediction for the number of mRNA and Protein m_j, n_j (for iteration $j = 1,2,3,\cdots,100$) are then given by the following Markov equations.

$$m_j = \begin{cases} P_m(0) * m_0; & \text{if } j = 0 \\ P_m(j) * m_{j-1}; & \text{if } j > 0 \end{cases} \tag{7}$$

$$n_j = \begin{cases} P_n(0) * n_0; & \text{if } j = 0 \\ P_n(j) * n_{j-1}; & \text{if } j > 0 \end{cases} \tag{8}$$

Leading to the following results for mRNA

5.2. Entropy Distribution

To measure the uncertainty associated with each sample of mRNA or proteins count, we introduce the concept of entropy over a population, which is calculated as follows:

$$(for\ mRNAs)\quad H(m) = -\sum_{j=1}^{N} p(m_j)\log\left(p(m_j)\right) \tag{9}$$

$$(for\ Proteins)\quad H(n) = -\sum_{j=1}^{N} p(n_j)\log\left(p(n_j)\right) \tag{10}$$

Computational results are shown in the figures below in the discussion section.

6. Discussion

We have shown (Figure 2) that, one may calculate the distribution of the number of mRNAs and proteins during gene expression, according to our model in Section 3. Based on these distributions (Figure 3 and Figure 4) we were able to predict the number of proteins and mRNAs over time. We use two main assumptions: i) The initial number of mRNA and proteins must be known; and ii) all cells must present similarity (functional, structural, architectural and/or dynamical). Our results show that both the protein and mRNA distributions are typically non-symmetric and may not be unimodal (Figure 5, Figure 6 and Figure 7). Consequently the mean and the mode are significantly different, and

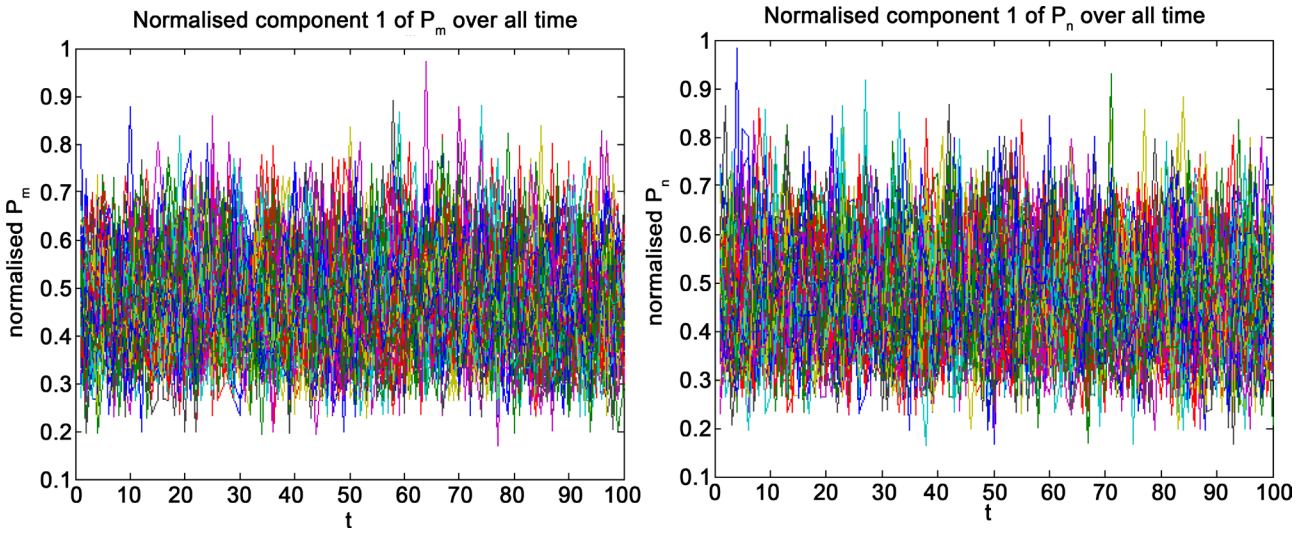

Figure 2. These figures are a recorded solution of the main PDE, showing a continuous random change in the probability distribution of mRNA (left) and proteins (right) in all 100 cells. It can be seen that most probability values are between 0.3 and 0.7, this indicates that very low and/or very high probability values are rare and our prediction approach is suitable to this problem.

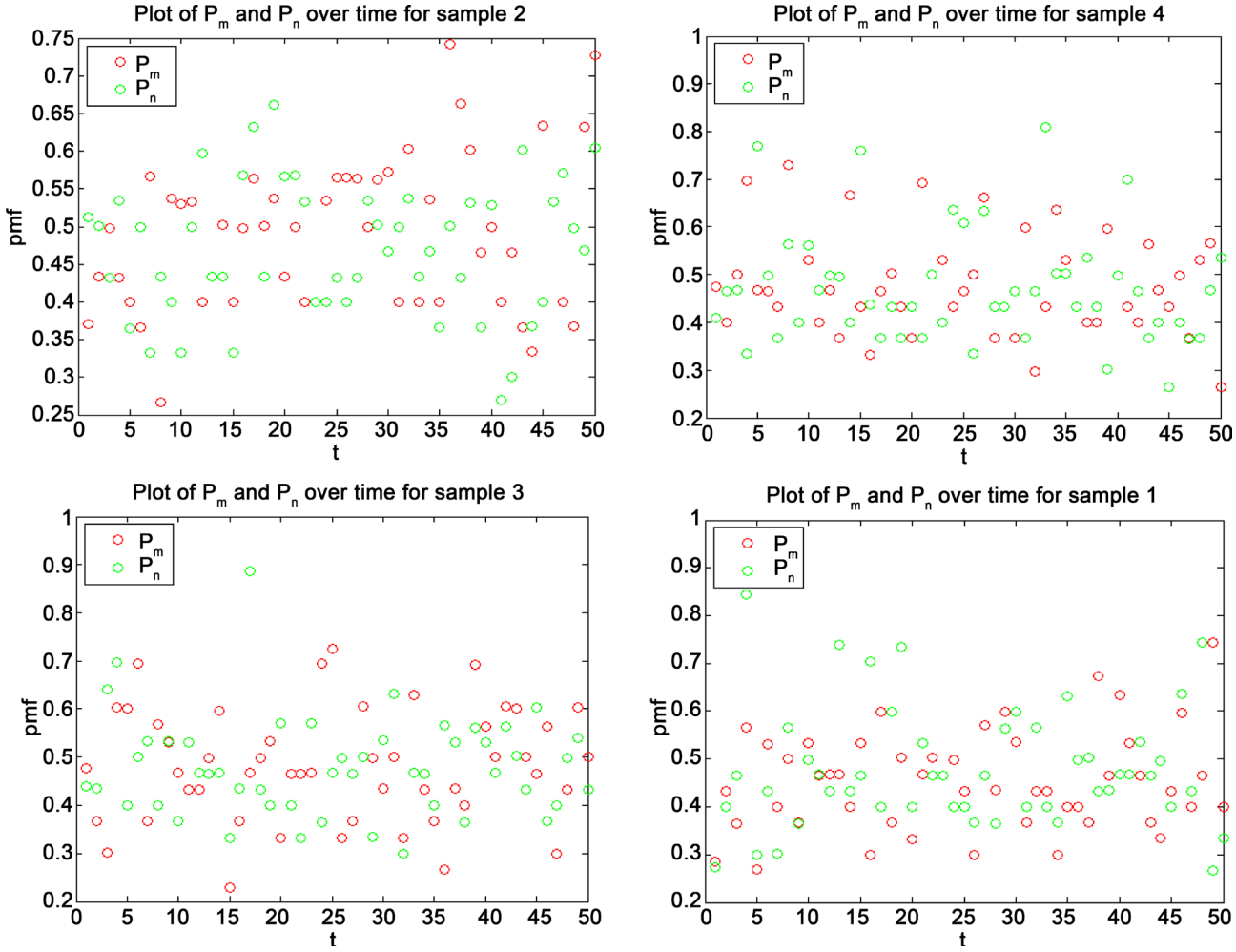

Figure 3. Shows scatter plots of the probability distribution of both number of mRNA and proteins over time for four different samples, plotted on the same graph. Each column represents the number of mRNA or proteins in all cells at a specific time period.

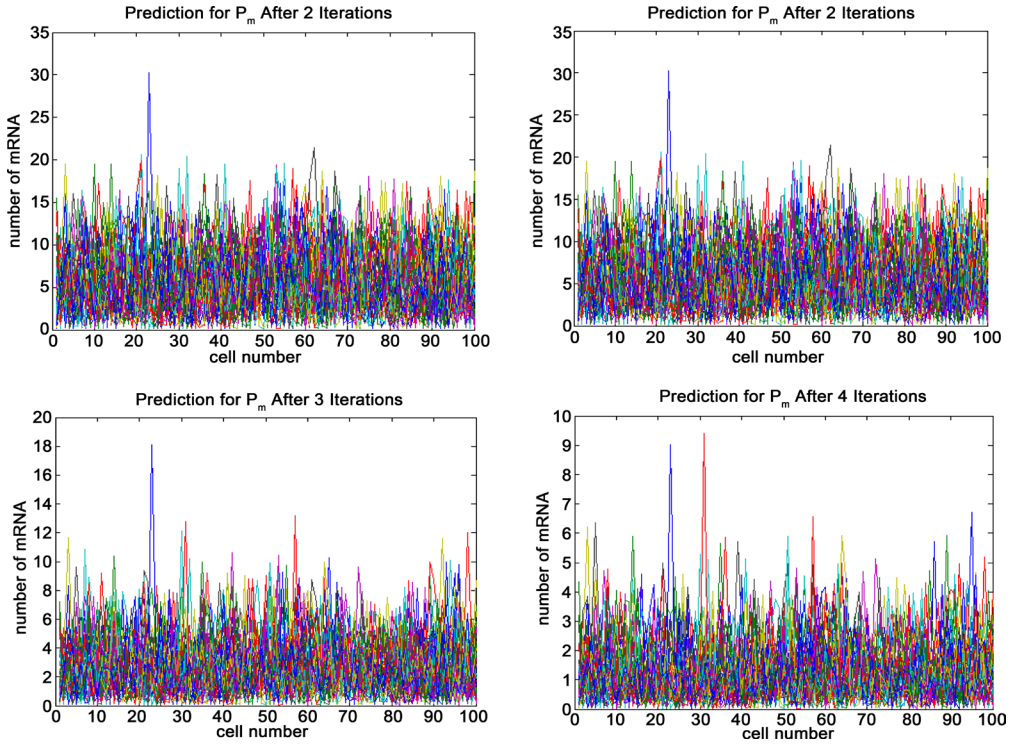

Figure 4. Four steps ahead prediction of mRNA numbers in all cells using Equations ((7) and (8)), for $j = 1, 2, 3, 4$. This result shows a fast decrease of probability values of mRNAs in cells iteration, indicating that mRNAs have a short life time, which is in accordance with biological evidence.

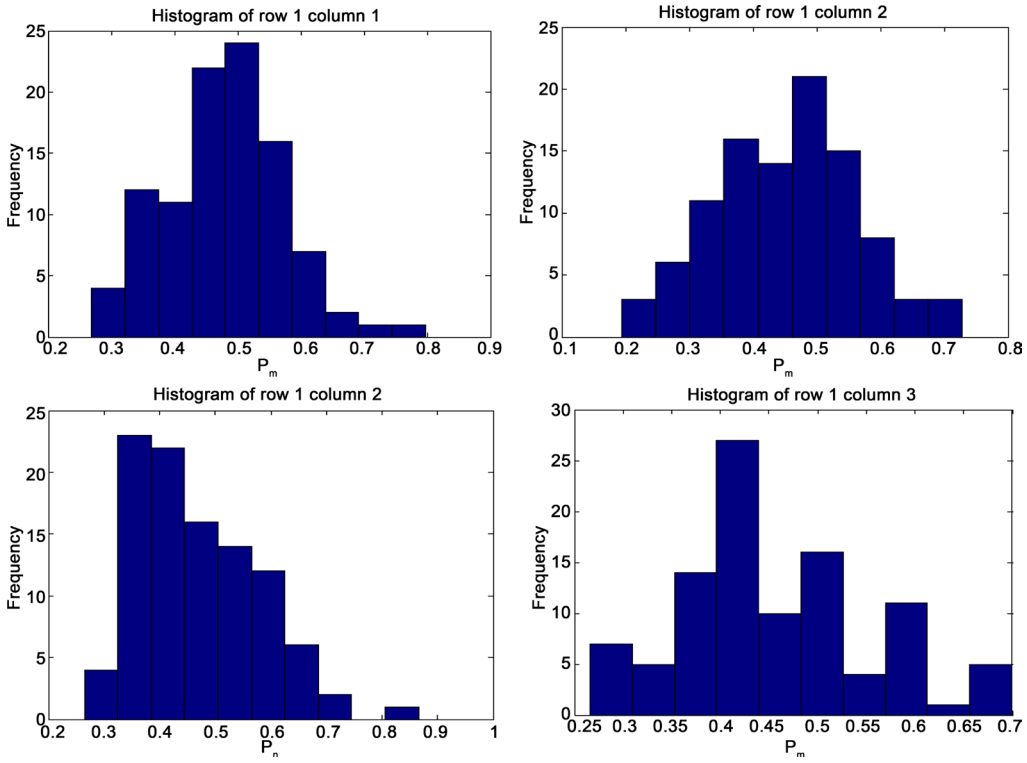

Figure 5. Shows the frequency of proteins related to P_n values in four generations. The optimal for proteins is around 0.4 (except generation 3), this indicates that on average 40% of probability level will give a better proteins count over time. The frequency distribution shows in all cases an asymmetric distribution, which indicates protein numbers are not normally distributed.

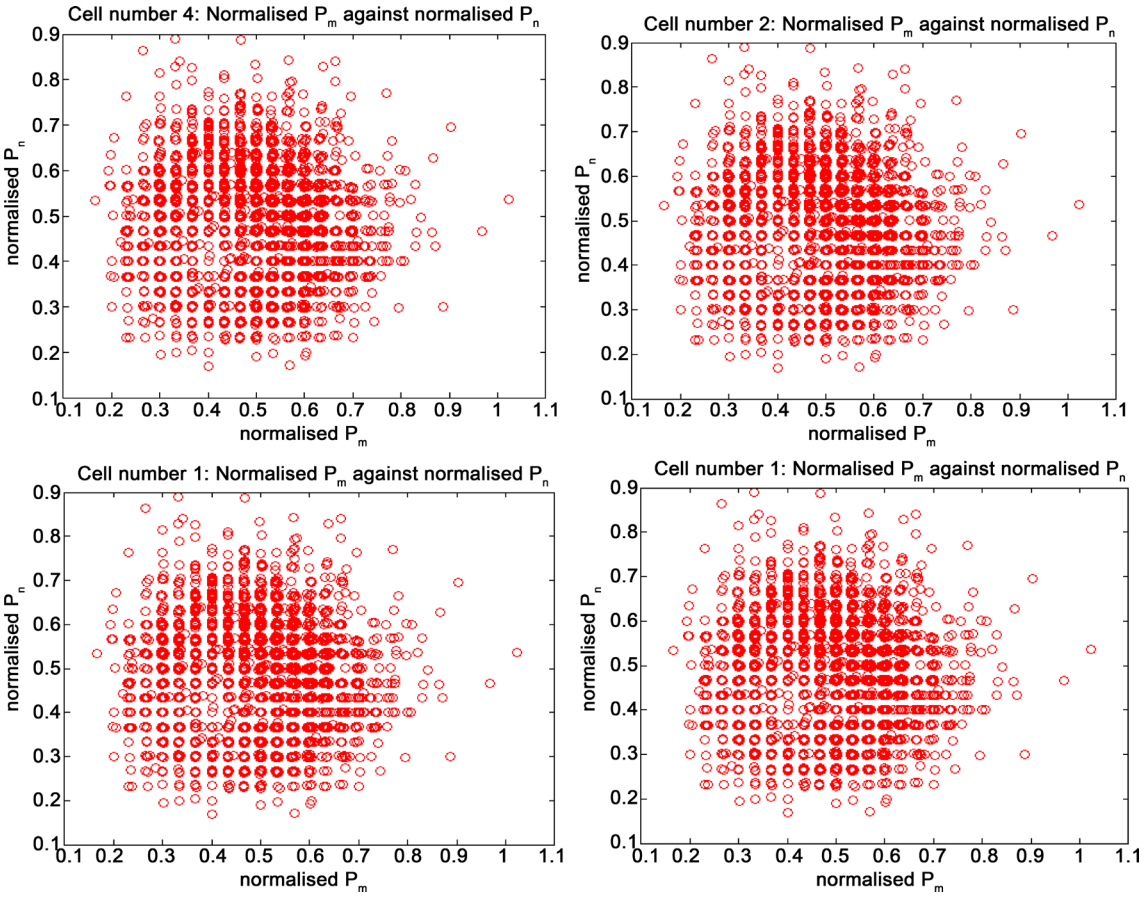

Figure 6. Shows the online scatter plots of P_n against P_m in four different cells. These figures confirm that both processes are strongly correlated over time in each of the cells, indicating that mRNA and protein dynamics (count) are depend over time. This will also hold for m and n since probabilities and mRNA/proteins are correlated.

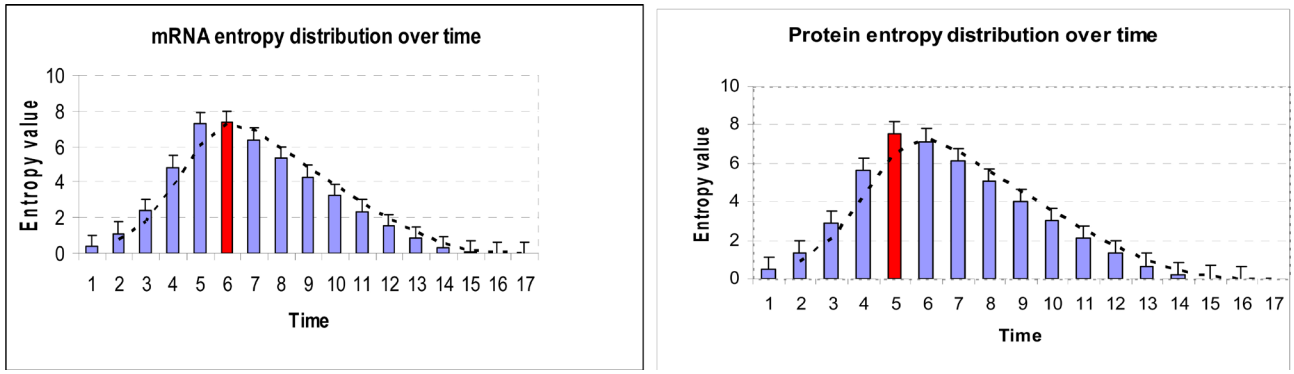

Figure 7. Shows the plot of entropy distributions over time in a chosen cell. It can be seen that the maximum entropy reached earlier for proteins and that the right tail is also longer in protein compared to that of mRNA. This may suggest that proteins have a longer life time, compared to mRNA. This evidence is in line with biological knowledge. The next section gives some discussion of the results.

the standard deviation is clearly not constant over time. Such distributions are poorly characterized by Gaussian characteristics. This paper was primarily designed to promote a modelling culture among noise biologists, modellers and to cope with the noise source and consequences in cell development.

7. Conclusion

The advantage of counting single molecules (mRNAs or proteins) is that, one obtains the probability distribution of molecules corresponding to each stage of the "central dogma" of molecular biology for each single gene. The mathematical model developed here differs from those that cellular biologists are accustomed to encountering [3] [5]. Instead of having a continuous and deterministic model of kinetic behavior, the mathematics of gene expression may be described by discrete stochastic models that take into account the numbers of molecules involved at both the mRNA and protein levels variability. **Figure 7** shows the plot of entropy distributions over time in a chosen cell. We have found that the maximum entropy reached earlier for proteins in comparison to mRNAs, the right tail density is also longer in protein in comparison of that of mRNA. This result clearly suggests that proteins have a longer life time, compared to mRNA. This evidence is in line with biological principles.

Acknowledgements

The authors would like to thank Dr. Sakumura for dedicating his precious time, giving many insightful comments and suggestions. This work was supported by the GCOE International Senior Research Fellowship, NAIST and the Grant of Aid from the Ministry of Education, Culture, Science and Technology (MEXT) Japan. We also thank the Universities of Sorbonne (France) and AUAF for their generous support and collaboration.

References

[1] Elowitz, E., Levine, A., Siggla, E. and Swain, P. (200) Stochastic Gene Expression in a Single Cell. *Science*, **297**, 1183-1186. https://doi.org/10.1126/science.1070919

[2] Fraser, H., Hirsh, A., Glaever, G., Kumm, J. and Eiseen, M. (2004) Noise Minimization in Eukaryotic Gene Expression. *PLOS Biology*, **2**, 834-838. https://doi.org/10.1371/journal.pbio.0020137

[3] Gillepsie, D.T. (1997) Exact Stochastic Simulation of Coupled Chemical Reactions. *The Journal of Physical Chemistry*, **81**, 2340-2361.

[4] Jimbo, H.C. and Craven, M.C. (2011) Unconstrained Optimization in a Stochastic Cellular Automata System. *Journal of Nonlinear Analysis and Optimization*, **1**, 103-110.

[5] Mc Adams, H.H. and Arkin, K. (1997) Stochastic Mechanisms in Gene Expression. *Proceedings of the National Academy of Sciences of the United States*, **94**, 814-819. https://doi.org/10.1073/pnas.94.3.814

[6] Newman, J.R., Ghaemmaghami, S., Ihmels, J., Breslow, D., Noble, M., DeRisi, J. and Weissman, J. (2006) Single-Cell Proteomic Analysis of S. Cerevisiae Reveals the Architecture of Biological Noise. *Nature*, **441**, 840-846. https://doi.org/10.1038/nature04785

[7] Novick, A. and Weiner, M. (1957) Enzyme Induction as an All-or-None Phenomenon. *Proceedings of the National Academy of Sciences of the United States*, **43**, 533-566. https://doi.org/10.1073/pnas.43.7.553

[8] Ozbudak, E., Thattai, M., Grossman, A. and Van Oudenaarden, A. (2002) Regulation of Noise in the Expression of a Single Gene. *Nature Genetics*, **31**, 69-73. https://doi.org/10.1038/ng869

[9] Raser, J.M. and O'Shea, E.K. (2004) Control of Stochasticity in Eukaryotic Gene Expression. *Science*, **304**, 1811-1814. https://doi.org/10.1126/science.1098641

[10] Thattai, M. and Van Oudenaarden, A. (2001) Intrinsic Noise in Gene Regulatory Networks. *Proceedings of the National Academy of Sciences of the United States*, **98**, 8614-8619. https://doi.org/10.1073/pnas.151588598

[11] Thattai, M. and Van Oudenaarden, A. (2004) Stochastic Gene Expression in Fluctuating Environments. *Genetics*, **167**, 523. https://doi.org/10.1534/genetics.167.1.523

[12] Peccoud, J. and Ycart, B. (1995) Markovian Modelling of Gene Products Synthesis. *Theoretical Population Biology*, **48**, 222-234. https://doi.org/10.1006/tpbi.1995.1027

[13] Pedraza, J.M. and Paulsson, J. (2008) Effects of Molecular Memory and Bursting on Fluctuations in Gene Expression. *Science*, **319**, 339-343. https://doi.org/10.1126/science.1144331

[14] Rosenfeld, N., Young, J., Alon, U., Swain, P. and Elowitz, M. (2005) Gene Regulation at the Single-Cell Level. *Science*, **307**, 1962-1965. https://doi.org/10.1126/science.1106914

[15] Shahrezaei, *et al.* (2008) Colored Extrinsic Noise Fluctuations and Stochastic Gene Expression. *Molecular Systems Biology*, **4**, 1-19. https://doi.org/10.1038/msb.2008.31

[16] Bengtsson, M., Hemberg, M., Rorsmsn, P. and Stahlberg, A. (2008) Quantification of mRNA in Single Cells and Modelling of RT-qPCR Induced Noise. *BMC Molecular Biology*, **9**, 63. https://doi.org/10.1186/1471-2199-9-63

[17] Schrodinger, E. (1994) What is Life? Cambridge University Press, Cambridge.

[18] Pedraza, J.M. and Van Oudenaarden (2005) Noise Propagation in Gene Networks. *Science*, **307**, 1965-1969. https://doi.org/10.1126/science.1109090