Modeling noise in gene expression

*Ivan Mura

Fecha de recepción: 17 de abril de 2013
Fecha de aprobación: 2 de mayo de 2013
Pag. 151 a 165

* Ph.D. University of Pisa, *Ingegneria Elettronica, Informatica e delle Telecomunicazioni.* Magister University of Pisa, *Scienze dell´Informazione*. Magister George Washington University School of Business Information Technology Project Management.

Abstract

This study deals with the modeling of the basic steps of gene expression in biological systems. It approaches the problem of determining the steady-state variance of the messenger RNA gene products, through a discrete-space; continuous-time stochastic modeling based on Stochastic Petri Nets. The model is solved at steady-state using both the analytical and simulation approaches. The results obtained for the variance of gene products indicate that the relative speed in the different gene expression processes may determine very noisy conditions, which are likely to affect many cellular phenomena

Keywords

Biological Systems
Messenger RNA Genes

Modelando ruido en expresión de genes

Resumen

Este estudio trata sobre la modelización de los pasos básicos de la expresión de genes en los sistemas biológicos. Se aborda el problema de la determinación de la variabilidad de estado estacionario de los productos de los genes ARN mensajeros, a través de un espacio discreto, el modelado estocástico de tiempo continuo con base en redes de Petri estocásticas. El modelo se resuelve en el estado estacionario, utilizando los enfoques analíticos y de simulación. Los resultados obtenidos para la variabilidad de los productos de los genes, indica que la velocidad relativa en los diferentes procesos de expresión génica, puede determinar las condiciones muy ruidosas, afectando muchos fenómenos celulares

Palabras clave

Sistemas biológicos
Genes ARN mensajeros

1. Introduction

Gene expression is the biological process by which information encoded in the genome is used to synthesize functional molecules such as proteins, enzymes, ribosomal RNA and many other basic components of cellular machinery.

The basic mechanisms of gene expression are common to most life forms (Clancy and Brown, 2008). Genes in the cell nuclear DNA encode the structure of functional molecules in a purely digital way, through a 4-symbol alphabet whose elements are the four DNA nucleotides Adenine, Guanine, Cytosine and Thymine (A, G, C and T). The gene linear sequence of nucleotides is read by the RNA polymerase macromolecule, which makes a copy of the gene content as a linear sequence of nucleotides to produce a molecule called messenger RNA (mRNA). This gene expression sub-process is known as gene transcription, as the mRNA is again a string of four characters; it has just a slight difference in the alphabet with respect to DNA in that the Uracil nucleotide is used to replace the Thymine in the sequence. The mRNA is termed messenger because of its ability to travel from the nucleus of the cell to its cytoplasm. In eukaryotic cells, mRNA molecules are exported from the nucleus through an extrusion process via the nuclear pore complexes that are embedded in the nuclear envelope (Carmody and Wente, 2009). The mRNA molecules found in the cytoplasm get engaged by macromolecules called ribosomes, which again read the string of nucleotides. This sub-process is called mRNA translation, because the ribosome interprets each substring of 3 nucleotides as a key to map one specific amino acid in a set of 20. Each mRNA molecule is commonly used for multiple translations. Multiple ribosomes attach to the mRNA and concurrently read it creating more copies of the gene product. The degradation of the mRNA molecule is regulated through various cellular mechanisms (Beelman and Parker, 1995). The amino acid

sequence resulting from mRNA translation is called polypeptide, and upon production in the cytosol, it loses its linear structure by folding into a 3-dimensional molecule (Alberts et al, 2002). This conformational change is regulated by the atomic forces among the subunits of the sequence, and it is often assisted by molecules called chaperones, which ultimately determines the biochemical properties of the molecule. The correct three-dimensional structure is essential for its function; Failure to fold into native structure generally produces inactive proteins, but in some instances misfolded proteins have modified or toxic functionality, as it happens for instance in the Creutzfeldt-Jakob disease (Chiti and Dobson, 2006).

The few processes outlined above provide a very simplified view of the biological phenomena occurring in gene expression. In this paper, we just sketch the basic features of a complex, regulated production process that uses a single copy template information (the gene), to generate a number of active products used by the cell to implement its functions. Regulation of the sub-processes is exerted a multiple steps in the process. For instance, the binding of RNA polymerase to DNA (the initiation step in gene transcription), is regulated by the availability of various classes of molecules called transcription factors. Each gene has its specific transcription factors, which are the product of the expression of other genes.

A fascinating aspect of cell biology is the achievement of a nearly deterministic system response obtained as a result of inherently noisy processes. In a matter of few minutes, all cells in a tissue exhibit nearly identical behaviors in terms of their gene expression when responding to the same stimuli. This is even more surprising if one considers that at the level of gene expression the response is determined by the biochemical interaction of a limited number of molecules (1 gene, few tens of mRNA copies) and thus large number effects that mask noisy behaviors are much less effective than in other biological processes (for instance metabolism) where thousands of copies of the same reactions happen per millisecond.

The objective of this work is to build simple quantitative models of the gene expression process to evaluate the sources and the propagation of noise along some basic steps of gene transcription. In this document we consider di-

fferent models of a gene regulation scheme based on the availability of transcription factors, and we solve the models in order to make predictions about the steady-state levels of the final mRNA product of the gene.

The paper is organized as follows: in Section 2, we provide a description of the basic mechanisms of gene expression that will be considered in this study. We do it by describing informal preliminary models of the system, in the style that is usually adopted by biologists. Then, in Section 3, we translate these informal models into a formal model with fixed semantics, specifically a Stochastic Petri Net. In Section 4 we conduct an analytical study of the mode, by determining some average measures of the steady-state population and two asymptotic results for the gene product noise . In Section 5, which contains a prospective analysis of results, we formulate a conjecture about the noise levels in the system, which is supported by preliminary simulation results. This same section provides some hints for future research on the topic.

2. A simple gene regulation model

Let us consider a gene whose transcription rate is dependent on the abundance of a transcription factor. A production/degradation process determines the concentration of the transcription factor, and this process is assumed to be an independent driver of the gene transcription. We suppose molecules of a species TF (Transcription Factor) are produced by a zero order process (that is, independent of the amount of TF already present) at a rate given by α following a process which details are not modeled here. The TF molecules are degraded by a process that we model as a simple first order process (that is, whose total speed or rate is proportional to the amount of TF present in the system) of rate β. These modeling abstractions can be represented in the language of chemical reactions, as follows:

$\emptyset \xrightarrow{\alpha} T$

$TF \xrightarrow{\beta} \emptyset$

The first reaction is modeling the fact that TF molecules are entering the system at rate α from the outside of it (the environment, represented by the empty set symbol ∅), whereas the second reaction models the destruction of TF molecules at rate α and hence their return to the model environment.

The gene transcription process is modulated (regulated) by the availability of TF molecules. Each mRNA molecule is produced from the environment of the system, i.e. the cell, with a rate that is proportional to a rate constant λ (the speed) of transcription and to the number of TF molecules. Each mRNA molecule in the system is degraded at a rate μ, and to hence the total mRNA degradation process is a first order process operating at rate μ. In the language of chemical reactions, these behaviors of the system are encoded as follows:

$\text{TF+Gene} \xrightarrow{\lambda} \text{TF+Gene+ mRNA}$

$\text{mRNA} \xrightarrow{\mu} \emptyset$

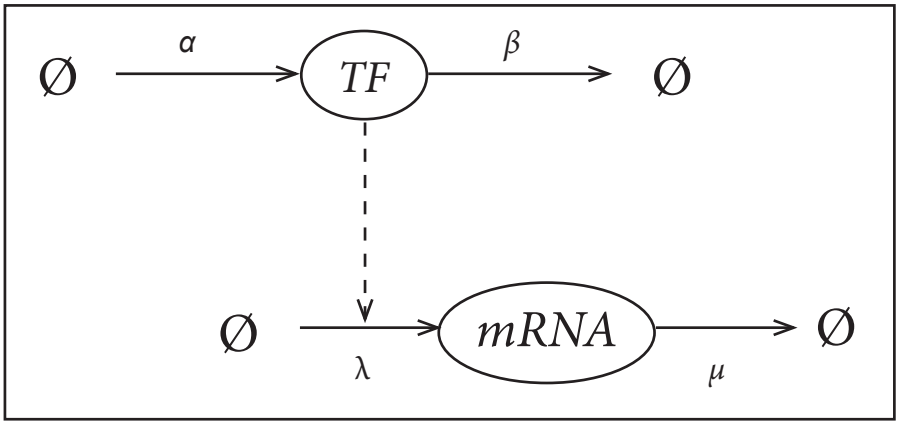
The first reaction above is representing the gene transcription process. A gene molecule (*Gene*, which exists in a single instance) engages a TF molecule in a reaction that as a result produces a new mRNA molecule, and gives back a free TF molecule and the unaltered Gene. The second reaction models the degradation of the mRNA molecules.

This kind of modeling based on chemical reactions is quite common in biology: it is detailed enough as far as the reactants (species at the left in a reaction) and products (species at the right in a reaction) of reactions are considered, and allows an easy representation of reaction stoichiometry (multiplicity of reactants and products in reactions).

A graphical representation of this system is provided in Figure 1. This diagrammatic representation of the model is based on the same information used to define the chemical reactions, but it slightly moves the focus of the modeling, by abstracting reactants that do not change their state. Thus, the mRNA production process is represented in the diagram by abstracting the Gene molecule, in which is not affected by the reaction itself. The TF participation in the reaction is represented by the influence arrow, which indicates that the rate of mRNA production depends (in some way) on the abundance of TF molecules.

It is worthwhile observing that both classes of models are not formal and that they do not have a precise semantics. Therefore, to conduct a quantitative predictive analysis of the models they first need to be translated into an unambiguous formalism.

Figure 1. A graphical model of a regulated synthesis reaction



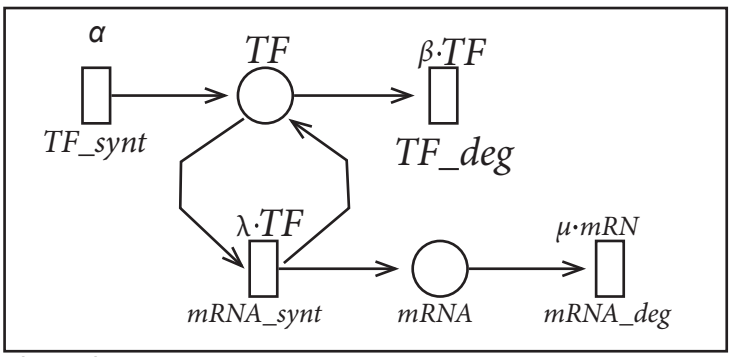
Source. by the author.

3. An SPN model of the system

Here, we consider a stochastic model of the system, and in particular, we assume that the occurrence times of all reaction events follow a negative exponential distribution, as considered in Gillespie's model of stochastic chemical kinetics (Gillespie, 1977).

We can encode the model into a Stochastic Petri Net (SPN), as shown in Figure 2. For a short introduction to the use of SPNs for biological modeling, see (Mura, 2010). Each biochemical event in the system is modeled through a transition in the net (the bar), and each variable is encoded as a discrete number of tokens contained in a place (the circle) of the net. The model contains four transitions: TF_synt and TF_deg model the synthesis and degradation of *TF* molecules, respectively, mRNA_synt and mRNA_deg model the synthesis and degradation of the mRNA. The arcs in the net show the flow of tokens as a result of transitions firing (occurrence of the associated events). For instance, each firing of TF_synt will add one token to the *TF* place and each firing of TF_deg will remove one token from that same place. The arcs that go from the *TF* place to the mRNA place, model the regulation effect exerted by the transition of the TF factor on the mRNA synthesis. The mRNA is produced at a rate that depends

Figure 2. An SPN model for the gene expressions reactions



Source. by the author.

on the number of TF molecules, but the mRNA synthesis does not affect the number of TF molecules. Therefore, the transition of mRNA_synt removes and puts back one token in the TF place.

To compute the steady-state value of the number of TF and mRNA tokens in the SPN model, we first notice that the steady-state distribution of the sub-model of the TF synthesis and degradation follows a Poisson distribution of parameter $\rho=\alpha/\beta$. In other words, if X_1 is the number of TF molecules at steady-state, the following equation holds for the probability of X1 being equal to m:

$$P[X_1=m] = \frac{p^m}{m!} e-\rho, \qquad m=0,1,2,\ldots$$

The value of the steady-state average TF is then easily determined from the distribution, and in particular from the results to be equal to $E[X_1]=\rho$. The steady-state distribution of the number of mRNA molecules, which we denote by X_2, is dependent upon the process of the TF molecules variation. However, its average production rate is known to be equal to $\lambda E[X_1]$ and therefore by the application of Little's law (Little, 1961), we get for the equilibrium that $E[X_2]=\lambda E[X_1]/\mu$.

4. Noise analysis

It is important to notice that in the SPN model the steady-state average value only depends on the ratio between the synthesis and degradation processes, and is therefore insensitive to any simultaneous scaling of the two parameters.

We shall now look at the noise in the population of species, expressed as their variance. The steady-state variance $VAR[X_1]$ of the TF sub-model is also known from the poisson distribution, and takes the same value as the mean, that is $VAR[1]=\rho$. The steady-state variance $VAR[X_2]$ of the $mRNA$ population is however unknown.

In this section we shall initiate a characterization of $VAR[X_2]$ as a function of the driving TF population statistics. As we mentioned, the steady-state distribution of TF (and thus its steady-state average and variance, as well as any higher moment) is insensitive to the simultaneous scaling of the two parameters α and β defining the model. The same applies to the average value of the $mRNA$ $E[X_2]$. However, $VAR[X_2]$ has a dependence on the specific value of any scaling factor $k>0$, such that $\rho=(\alpha k)/(\beta k)$. Informally, the scaling factor k defines the speed of variation of the TF process, and the variance of the $mRNA$ production process is influenced by that speed. In the following, we shall demonstrate two asymptotic values for $VAR[X_2]$, specifically the following ones:

$$\lim_{k\to\infty} VAR[X_2] = \rho\varphi$$

$$\lim_{k\to 0^+} VAR[X_2] = \rho\varphi$$

where for the sake of conciseness we denote $\varphi=\lambda/\mu$.

Theorem 1. *When the rates of the TF synthesis and degradation process approach infinity, the variance of the number of $mRNA$ approaches $\rho\varphi$.*

When $k\to\infty$, we expect the number of TF to quickly move across the support values of the distribution before any two consecutive $mRNA$ synthesis events. This is as to say that, in each state with a given

number of *mRNA* molecules, the next *mRNA* molecule is synthesized at a rate given by the following average:

$$\sum_{i=1}^{\infty} i\,\lambda P[X_1 = i] = \lambda E[X_1] = \lambda\rho$$

Thus, in the limit for $k\to\infty$ the distribution of X_2 is Poisson of parameter $\lambda\rho/\mu=\rho\varphi$, with an average value (as already anticipated) given by $\rho\varphi$ and a variance again equal to $\rho\varphi$.

Theorem 2. *When the rates of the TF synthesis and degradation process approach zero, the variance of the number of mRNA approaches* $\rho\varphi(\varphi+1)$.

Consider the distribution of X_2, which by applying the total probability law can be expressed as follows:

$$\sum_{m=0}^{\infty} P[X_2 = n, X_1 = m] = \sum_{m=0}^{\infty} P[X_2 = n|X_1 = m]P[X_1 = m]$$

Assume now that X_2 is much faster than X_1, which is the case when $k\to 0^+$, so that X_2 goes quickly to its steady-state distribution, before any change in the number of *TF* molecules. Under this assumption, the distribution of X_2 conditioned to the event $X_1=m$ is known to be Poisson, with parameter $m\varphi$. Therefore:

$$P[X_2 = n] = \sum_{m=0}^{\infty} \frac{(m\varphi)^n}{n!} e^{-m\varphi} \frac{\rho^m}{m!} e^{-\rho} = e^{-\rho} \frac{\varphi^n}{n!} \sum_{m=0}^{\infty} \frac{m^n (e^{-\varphi}\rho)^m}{m!}$$

To close the last summation, we recall that the summation

$$e^{-x} \sum_{k=0}^{\infty} \frac{k^n x^k}{k!}$$

is the generating function for the Bell polynomial of grade n in the variable x, which we denote by $B_n(x)$. Therefore, we have the following equality:

$$\sum_{m=0}^{\infty} \frac{m^n (e^{-\varphi}\rho)^m}{m!} = e^{e^{-\varphi}\rho} B_n(e^{-\varphi}\rho)$$

Thus, we can finally write the following expression for the probability distribution of X_2:

$$P[X_2 = n] = e^{-\rho} \frac{\varphi^n}{n!} e^{e^{-\varphi}\rho} B_n(e^{-\varphi}\rho) = e^{-\rho(1-e^{-\varphi})} \frac{\varphi^n}{n!} B_n(e^{-\varphi}\rho)$$

For the sake of clarity let us rewrite the last expression as follows:

$$P[X_2 = n] = \theta \frac{\varphi^n}{n!} B_n(\sigma)$$

Where $\theta=e^{-\rho(1-e^{-\varphi})}$ and $\sigma=e^{-\varphi}\rho$. From the steady-state distribution we can compute the probability generating function, as follows:

$$G_{X_2}(z) = \sum_{n=0}^{\infty} \theta \frac{\varphi^n}{n!} B_n(\sigma)\, z^n$$

Because the following property holds of the Bell polynomials,

$$\sum_{k=0}^{\infty} \frac{B_k(x)}{k!} t^k = e^{(e^t-1)x}$$

We can rewrite the probability generating function as follows:

$$G_{X_2}(z) = \theta \sum_{n=0}^{\infty} \frac{(\varphi z)^n}{n!} B_n(\sigma) = \theta e^{(e^{\varphi z}-1)\sigma}$$

With some algebraic manipulations we further get:

$$G_{X_2}(z) = e^{-\rho(1-e^{-\varphi})} e^{(e^{\varphi z}-1)e^{-\varphi}\rho} =$$

$$e^{-\rho} e^{\rho e^{-\varphi}} e^{\rho e^{\varphi z} e^{-\varphi}} e^{-\rho e^{-\varphi}} = e^{-\rho} e^{\rho e^{\varphi(z-1)}}$$

The first moment of the distribution can be obtained from the first derivative of the probability generating function, as follows:

$$E[X_2] = G'_{X_2}(1^-) =$$

$$e^{-\rho} \lim_{z \to 1^-} \frac{d}{dz} e^{\rho e^{\varphi(z-1)}} =$$

$$e^{-\rho} \lim_{z \to 1^-} \left[e^{\rho e^{\varphi(z-1)}} \rho e^{\varphi(z-1)} \varphi \right] =$$

$$= e^{-\rho} e^{\rho} \rho\varphi = \rho\varphi,$$

recovering again the already known result that the steady-state average value of X_2 is indeed independent of the driving process speed. To compute the variance of X_2, we use the following equality involving the first and second derivatives of the probability generating function:

$$VAR[X_2] = G''_{X_2}(1^-) + G'_{X_2}(1^-) - \left[G'_{X_2}(1^-)\right]^2,$$

and we finally get:

$$VAR[X_2] = (\rho\varphi)^2 + \rho\varphi^2 + \rho\varphi - (\rho\varphi)^2 = \rho\varphi(\varphi + 1)$$

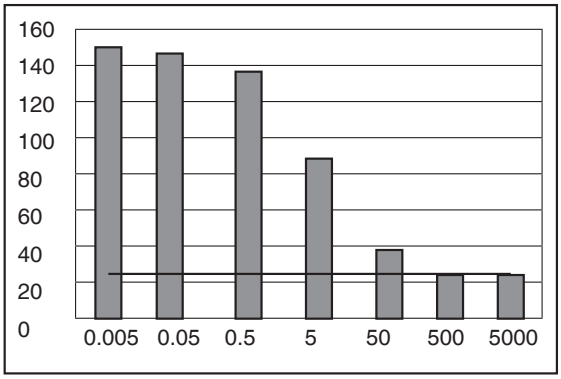
5. Conclusions and speculations

We showed in the previous section that the speed of the TF synthesis and degradation processes has an effect on the variance of gene products. This is quite an interesting result, because the noise in the population of species may have profound effects on the evolution of living systems. As an example, consider the cell cycle process, which ordered process is ruled by the level of concentration of cyclins (Csikász-Nagy and Mura, 2010).

A discrete stochastic modeling of the system allowed the identification of an interesting interaction effect between process speeds. This effect would be totally invisible in any continuous deterministic model.

We speculate that the two asymptotic values computed for the variance of the *mRNA* gene product establish indeed bounds for the noise. More specifically, we make the conjecture that the variance of X_2 is a monotonic decreasing function of k. We show the results of simulation study supporting this conjecture. Figure 3 depicts the average variance of X_2, the number of *mRNA* gene products, in an simulation experiment where α= 0.05 and β= 0.01, which turns out in an average number of TF molecules X_1=5, $\lambda = k \bullet 5$, $\mu = k \bullet 1$, which results in an average number of *mRNA* molecules of $X_2 = 25$. In the simulations, the parameter k varies in the range [0.005,5000], thus causing the exploration of a large range of ratios between the relative speeds of the *TF* and *mRNA* relative processes of synthesis and degradation. The simulations were conducted using the Möbius tool (Graham et al., 2001). We used 20,000 runs of simulation for each sampling value of k, and we computed confidence intervals at 98% confidence levels for two measures: the average number of *mRNA* molecule, which we used as a control measure to ascertain the correctness and accuracy of simulation results, and the variance of the *mRNA,* the measure objective of the study.

Figure 3. simulated varience of the mRNA steady-state population. The straight line is used as a control of simulation accuracy, and provides the average steady-state number of mRNA molecules



source. By the authors

The simulation results in Figure 3 show that the average number of $mRNA$ molecules X_2, the black straight line, is not affected by the variations of the scaling factor and constantly takes its theoretically expected value of 25.

On the other hand, the variance of X_2 is monotonically decreasing with k, as conjectured. It achieves its maximum value for values of k that tend to zero, and the simulated maximum value is exactly the theoretical one $VAR[X_2]=\rho\varphi(\varphi+1)=150$. The simulated minimum value is achieved for k that goes to infinity and again, it perfectly matches the theoretical value stated in Theorem 1, $VAR[X_2]=\rho\varphi=25$. Notice that confidence intervals are not reported in Figure 3, as they are within 1% of the estimated measure and therefore too small to be graphically appreciated.

These results support the conjecture. We intend to carry on with this study along two different directions: the first one is to prove the conjecture analytically, and the second one is to explore the effect that the variance of the mRNA gene products has on the further steps of the gene expression process, specifically on the translation and on the average number of final gene products.

6. Referencias

Alberts, B. Alexander, J. Lewis, J. Raff, M. Roberts, K. y Walters, P. (2002). *The Shape and Structure of Proteins, Molecular Biology of the Cell*, Fourth Edition. New York and London: Garland Science. ISBN 0-8153-3218-1.

Beelman C.A. y Parker, R. (1995). *Degradation of mRNA in eukaryotes*, Cell 81(2), pp. 179-183.

Carmody, S.R. y Wente, S.R. (2009). *mRNA nuclear export at a glance*, Journal of Cell Science 122, pp. 1933-1937

Chiti, F. y Dobson, C. (2006). *Protein misfolding, functional amyloid, and human disease*, Annual review of biochemistry 75, pp. 333–366.

Clancy, S. y Brown, W. (2008). *Translation: DNA to mRNA to protein*. Nature Education 1(1).

Clark, G. Courtney, T. Daly, D. Deavours, D. Derisavi, S. y et al. (2001). The Möbius Modeling Tool, in Proceedings of the 9th international Workshop on Petri Nets and Performance Models (PNPM'01) (PNPM '01). IEEE Computer Society, Washington, DC, USA, 241-250.

Csikász-Nagy, A. y Mura, I. (2010). *Role of mRNA gestation and senescence in noise reduction during the cell cycl*e, In Silico Biology, 10, pp. 81-88.

Gillespie, D.T. (1977). *Exact stochastic simulation of coupled chemical reactions*, The Journal of Physical Chemistry 81(25), pp. 2340–2361.

Little, J.D.C. (1961). *A proof for the queuing formula: L = λW*, Operations Research 9(3), pp. 383–387.

Mura, I. (2010). *Stochastic Modeling*, in I. Koch, W. Reisig and F. Schreiber (eds.), Modeling in Systems Biology, The Petri Net Approach, Computational Biology 16, ISBN: 978-1-84996-473-97, Springer-Verlag London Limited, chapter 7.