On the Impact of Hidden Modeling Assumptions on Living Systems- Predictive Dynamics

Ivan Mura *

Fecha de recepción: 28 de febrero 2014
Fecha de aprobación: 05 de marzo 2014
Pp. 85 - 106

Abstract:

This paper aims at showing the predictive modeling of living systems, particularly some commonly structured modeling assumptions which simplify the behavior of living systems. It also takes into account the stochastic modeling of basic gene expression mechanisms, such as transcription and translation, reaffirming the effect that simplifications have on the predictive behavior of living systems. These mechanisms rely on the basis of most gene expressions, signaling pathways and protein-protein interaction network models. This paper states that the usage of naïve modeling abstractions may result in predictive behaviors that are quite far from reality.

Key words

Living systems, network models, predictive behaviors.

* *Ph.D. en Ingeniería Electrónica, Informática y de Telecomunicaciones, Universidad de Pisa. Magister en Ciencias de la Información, Universidad de Pisa y en Information Technology Project Management, George Washington University School of Business.*

El impacto de supuestos no explícitos en las predicciones de las dinámicas de sistemas vivos

Resumen

Este artículo, tiene como objetivo mostrar el modelado predictivo de los sistemas vivos, en particular algunos supuestos del modelo comúnmente estructurado, que simplifican el comportamiento de los sistemas vivos. También, tiene en cuenta el modelado estocástico de los mecanismos básicos de expresión génica, tales como la transcripción y la traducción, reafirmando el efecto que tienen las simplificaciones en el comportamiento predictivo de los sistemas vivos. Estos mecanismos, dependen de la mayoría de las expresiones de genes, las vías de señalización y los modelos de red de interacción proteína-proteína. Este artículo, se señala que el uso de abstracciones de modelados ingenuos, pueden dar lugar a comportamientos predictivos lejanos de la realidad.

Palabras clave

Sistemas Vivos, Modelos de Red, Comportamientos Predictivos.

L'impact des hypothèses de modélisation sur les dynamiques prévisibles des systèmes vivants

Résumé

Cet article traite de la modélisation prédictive des systèmes vivants et en particulier de certaines hypothèses de modélisation exposant les simplifications du comportement des systèmes. Nous examinerons ici la modélisation stochastique de base des mécanismes d'expression tels que la transcription et la traduction, ainsi que l'effet que certaines simplifications peuvent produire sur le comportement prévisible des systèmes vivants. Ces mécanismes sont à l'origine de la plupart des expressions génétiques, des modèles d'interactions en réseau de type protéine-protéine. Nous montrerons que l'utilisation de modélisations simples peut entraîner des comportements prévisibles assez loin de la réalité.

Mots clés

Modèles en réseaux, Systèmes Vivants, Prédiction des Comportements

Impacto de suposições escondidas de modelagem em dinâmica predita de sistemas vivos

Resumo

Este artigo abrange a modelagem preditiva de sistemas vivos, e em particular algumas suposições comuns de modelagem que levam a simplificações no comportamento dos sistemas. Será considerada a modelagem estocástica dos mecanismos básicos de expressão gênica, como transcrição e tradução, para estabelecer o efeito que as simplificações têm no comportamento predito de sistemas vivos. Estes mecanismos estão na base da maioria da expressão gênica, mostrando caminhos e modelos de redes de interação proteína-proteína. Mostra-se que o uso ingênuo de abstrações de modelagem pode levar a comportamentos preditos que estão muito longe da realidade.

Palavras-chave

Sistemas Vivos, Modelos de Redes, Comportamentos Preditos.

1. Introduction

Until 1977, the discrete simulation of biochemical systems at molecular level was considered to be intractable, due to the needs of tracking the speed and position of every molecule in the system. Then, a paper by D.T. Gillespie (Gillespie, 1977) proposed a low-complexity and easy to implement discrete event simulation algorithm for predicting the evolution of the abundances of molecular species in a biochemical system.

This algorithm came with a theoretical proof of exactness grounded on statistical mechanics, which precisely defined the scope of its applicability. Since then, Gillespie's simulation method has been extensively applied, and many computational approaches and tools based on it have appeared (Cao and Samuels, 2009).

The principal reasons for this success are found in the intuitiveness of the algorithm, which establishes a simple link with the classical description language of chemical reactions, and in its inherent simplicity. In fact, Gillespie's algorithm for stochastic simulation proposes a method that follows the intuitive understanding of how a system composed of biochemical species that interact through relations would evolve over time.

However, such simplicity has its own limits of applicability, and is subject to a precise set of assumptions, precisely described by Gillespie itself. These assumptions should always be confirmed or rejected, and in any case questioned and not taken a priori. In this respect, there is a myriad of examples of studies assuming the general validity of such assumptions without discussion.

The objective of this paper is to explore the effects of extending the applicability of Gillespie's result beyond its limits. We shall demonstrate, through some simple examples, that the predictions about system dynamics may be affected by significant errors, which may lead to wrong conclusions about the modeled living systems.

The rest of this document is organized as follows. In Section 2 we outline the main contributions proposed by Gillespie. Section 3 discusses one

crucial condition that Gillespie considered for ensuring the validity of its stochastic modeling approach and presents cases when such hypothesis is not trivially satisfied, to point out where approximations may be introduced. The potential impact of those approximations is considered in Section 4. Finally, conclusions and references are provided in Section 5 and 6, respectively.

2. A short Recap of Gillespie's Results

In a nutshell, Gillespie proved that, under the assumptions described later on, the time to the next reaction event is exactly a sample from a random variable that has a negative exponential distribution. The rate of the exponential is solely determined by the abundance of species before the reaction occurs and by the reaction rate. This astonishing simple result has profound consequences: it implies that the evolution of molecular counts can be modeled over time by a homogeneous continuous-time discrete-space Markov process, a class of stochastic processes for which a wealth of analytical results and solution algorithms exist, see for instance (Norris, 1997).

Gillespie proved that this result is valid under fairly general assumptions. In particular, he demonstrated that there are only three limitations defining its applicability:

- The biochemical system is homogeneous in terms of concentrations, i.e. it is well-stirred and diffusion is very fast, which ensures the velocity of reactions is the same irrespectively of the position of the reactant molecules.

- The system is under thermal equilibrium, which provides for the homogeneity of the Markov process. Note indeed that the reaction speed would change with temperature variations.

- The reactions modeled are elementary, where elementary means that they are not hiding intermediate species. For instance, it would be not elementary to model an abstract reaction $A \rightarrow C$ when the real reaction has indeed an intermediate species B and is in fact $A \rightarrow B \rightarrow C$. The elementariness is required to ensure reaction times are distributed as negative exponential random variables.

All the three conditions above are required for Gillespie's results to be applicable, but very rarely their validity is checked. More often, they are assumed to be true. The only condition that is easy to verify is number 2), because temperature stability is ensured in many biological systems of interest at the molecular level, and furthermore this situation can be easily replicated in experimental conditions.

As for condition 1), diffusion has been proven to be very efficient even in the densely crowded medium of the intracellular spaces. However, homogeneity is obviously violated inside different cellular compartments, which are structural means the cell employs to provide for distinct concentrations of biochemical species in different places. However, this is not posing operative limitations to the applicability of Gillespie's result, because it is possible to model cellular compartments and to add shuttling reactions between them to represent inter-compartment diffusion reactions.

The major problems are in fact related with the verification of condition 3), which is usually not taken into consideration. We shall see in the next section what it means in terms of models to assume this condition is verified when indeed it is not.

3. Elementary and non-elementary Reactions

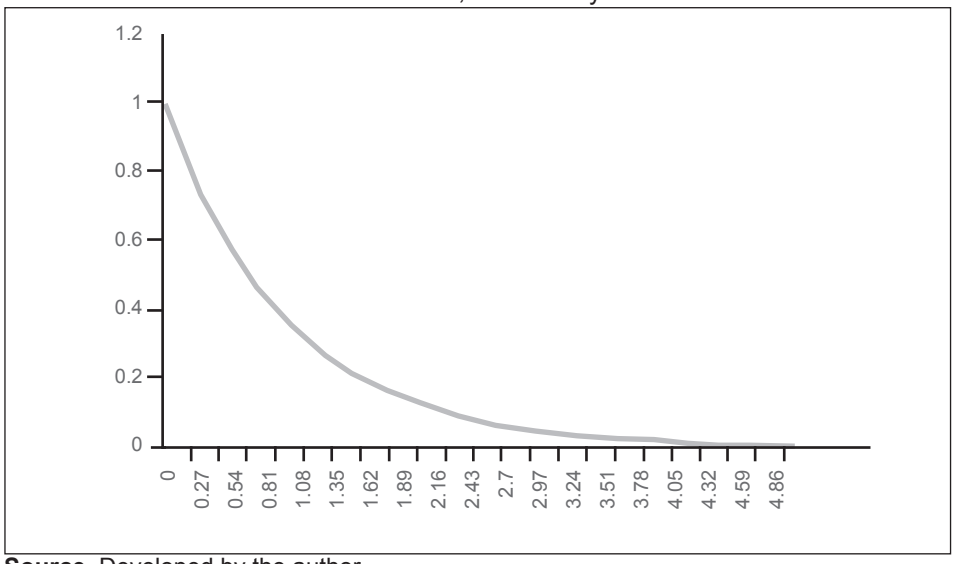
Following the definition stated above, an elementary reaction is one that does not abstract any intermediate species. This apparently simple condition is in fact not obvious to verify. For instance, consider a bimolecular reaction of the form $A+B \rightarrow C$, whereby two reactants are used to produce a third species. There may be multiple reasons for which this reaction may indeed not be elementary. We list here some of them:

- There is an unknown species X that is participating in the reaction, for instance an enzyme that catalyzes it, so that the actual reaction has the form $A+B+X \rightarrow X+C$. Now, this latter reaction is not an elementary one and violates condition 3) for Gillespie's result applicability. It is indeed easy to understand that the probability of 3 species meeting at random in the cellular medium is practically zero. This is a very common modeling risk.
- There is an unknown species Y that forms as an intermediate before C is formed, so the reaction is actually a sequence of reactions, of the form $A+B \rightarrow Y$, $Y \rightarrow C$. Species Y may be a short-lived one and therefore very difficult (or impossible) to detect by experimental means.
- The same reaction may happen with and without a catalyzer enzyme, so we may have, in a concurrent fashion, $A+B+X \rightarrow X+C$, $A+B \rightarrow C$, where the first reaction would be faster than the second one.

In all of the cases above, a naive modeling in the form $A+B \rightarrow C$ turns out in a modeling error. Let us now understand where the error is coming from. If we assume the reaction is elementary, we conclude from Gillespie's result that the time to the next occurrence of the reaction is an exponentially distributed random variable, and therefore we are making a very precise modeling choice about its probability distribution. Suppose we only have one molecule of A and one of B, and that we know, from experimental measurements, that in a given medium or living system, the reaction between A and B happens at a rate $\lambda=1$ (measured in unit of time^{-1}).

If the reaction is elementary, it is correct to assume that λ is the rate of a negatively distributed exponential random variable. Therefore, the expected time to the occurrence of the reaction is equal to λ^{-1}, and the distribution of the time to the occurrence of the reaction has a probability density function $\lambda e^{-\lambda t}$, $t > 0$, which is graphically shown in Figure 1.

Figure 1. The probability density function of the time to the occurrence of the reaction. Time is on the horizontal axis, the density value on the vertical one.



Source. Developed by the author.

However, suppose now that the reaction is not elementary, but it is of the form $A+B \rightarrow Y$, $Y \rightarrow C$. Also, suppose that these two reactions are both elementary, and that they happen with rates λ_1 and λ_2, respectively. We want the average time of occurrence of the overall reaction $A+B \rightarrow Y$, $Y \rightarrow C$ to be the same as the average time of occurrence of the elementary one, therefore the following equality must hold:

$$\frac{1}{\lambda_1} + \frac{1}{\lambda_2} = \frac{1}{\lambda} \qquad (1)$$

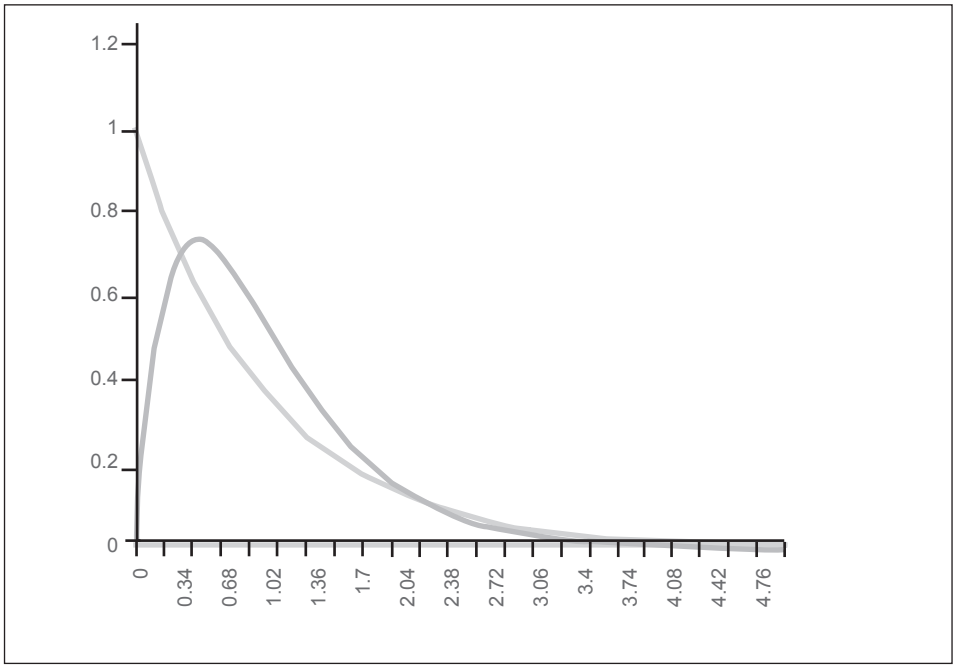
The time to the occurrence of the production of the molecule of C is now a random variable that is the sum of two negatively exponential distributed random variables. The sum of two exponential random variables of rates λ_1 and λ_2 follows a distribution known as hypo-exponential or generalized

Erlang (Amari and Misra, 1997), whose probability density function is as follows:

$$\frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}(e^{-\lambda_2 t} - e^{-\lambda_1 t}),\ t > 0$$

To quantify the modeling error that one may commit when assuming the elementariness of reaction A+B→C whereas it consists of the two reactions A+B→Y, Y→C, we compare in Figure 2 the distributions of the time to the occurrence of the production of C, in the two cases.

Figure 2. Comparison of occurrence time probability distributions: assuming the reaction is elementary (blue-line) and relaxing the assumption of elementariness (red-line).



Source. Developed by the author.

As it can be observed from Figure 2, there are important differences between the two distributions. Even though they have the same expected value $\lambda^{-1}=1$, they have different shapes and different variances. In particular, the variance VAR_{HYPO} in the hypo-exponential case is:

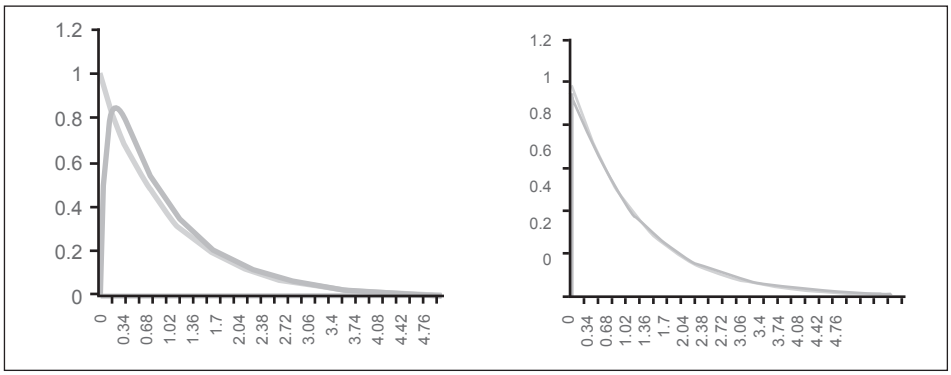
$$VAR_{HYPO} = \frac{1}{\lambda_1^2} + \frac{1}{\lambda_2^2} \qquad (2)$$

Because of relation (1), VAR_{HYPO} *is always lesser than the variance* $VAR_{EXP} = \lambda^{-2}$ *of the exponential case. This means that assuming the elementariness of a reaction when this is not true turns out in an erroneous modeling of the molecular noise in the times of reaction occurrence.*

To speculate on how large the error can be, we make a simple study of the derivative of $VAR_{HYPO} - VAR_{EXP}$ *as a function of* λ_1 *and of the parameter* λ*, eliminating the other variable* λ_2 *by enforcing* (1). *We obtain that the sign of the derivative is that of a second order polynomial with a negative coefficient of the second order term, in other words a concave parabola, with the null value taken for* $\lambda_1 = 2\lambda$. *Therefore, the maximum modeling error occurs when the two elementary reactions that realize* A+B→C *occur at the same speed. This worst case is indeed the one shown in Figure 2.*

We show in Figure 3 the comparison between the distributions when the first reaction A+B→Y *is 10 times and 100 times faster than the second one* Y→C. *As it can be observed, the hypo-exponential distribution converges to the exponential one as the imbalance between the speeds of the two reactions tends to increase. This convergence can be measured by checking the variance of the hypo-exponential distribution. In the worst case scenario shown in Figure 2, the variance is 0.5 (compare with the unitary variance of the exponential, elementary case), in the left case shown in Figure 3 the variance is 0.8347, and in the right case in the same figure is 0.9802.*

Figure 3. Comparison of occurrence time probability distributions. Left side: elementary reaction (blue-line) and two steps reaction (red line) when step 1, A+B→Y is 10 times faster than step 2, Y→C. Right side: elementary reaction (blue-line) and two steps reaction (red line) when step 1, A+B→Y is 100 times faster than step 2, Y→C.



Source. Developed by the author.

Another case we may consider is when we assume elementariness of a reaction but in fact the reaction is the results of multiple paths that combine to generate a product from a substrate. Say that we are modeling A→B whereas what is really happening is that A→X→B and A→Y→B, where X and Y are two unobserved (or not observable) species.

In this case, the product B of the reaction is the result of two different processes, which we assume happen with relative frequencies (i.e. probabilities) p_1 and p_2= 1-p_1 . Also, we suppose that the rate of the first process A→X→B, is λ_1 and the rate of second one, A→Y→B, is λ_2. If we assume that the two processes can be modeled as elementary reactions, then the overall time to the occurrence of the production of molecule B is a random variable defined as the weighted sum of two exponential random variables. This random variable is known as hyper-exponential distribution (Amari and Misra, 1997), and its probability distribution function is simply the weighted sum of the exponential distributions, as follows:

$$p_1\lambda_1 e^{-\lambda_1 t} + p_2\lambda_2 e^{-\lambda_2 t}, t>0$$

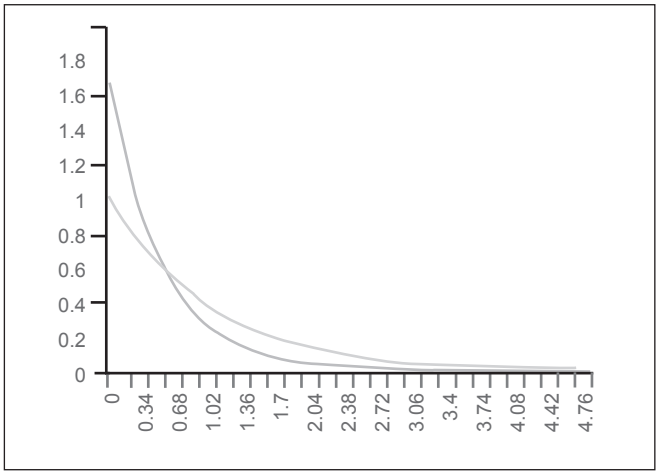
The hyper-exponential distribution above has 3 parameters, which are indeed defining a large family of distributions. The expected value of the

random variable is the weighted sum of the expected times of the two possible reaction paths, which, to simplify the following mathematical treatment, we rewrite in terms of two new variables $\mu_1 = \frac{\lambda_1}{p_1}$ and $\mu_2 = \frac{\lambda_2}{p_2}$ as follows:

$$\frac{p_1}{\lambda_1} + \frac{p_2}{\lambda_2} = \frac{1}{\mu_1} + \frac{1}{\mu_2}$$

To maintain the equivalence of expected times, we need the above expression to be equal to λ^{-1}. We compare in Figure 4 the distributions of the time to the occurrence of the production of B, in the two cases.

Figure 4. Comparison of occurrence time probability distributions: assuming the reaction is elementary (blue-line) and relaxing the assumption of elementariness (red-line) and considering a multi-path reaction.



Source. Developed by the author.

As it can be observed from Figure 4, there are again relevant important differences between the two distributions. In particular, the variance VAR_{HYPER} of the hyper-exponential variable is as follows:

$$VAR_{HYPER} = \frac{2p_1}{\lambda_1^2} + \frac{2p_2}{\lambda_1^2} - \left(\frac{p_1}{\lambda_1} + \frac{p_2}{\lambda_2}\right) \quad (3)$$

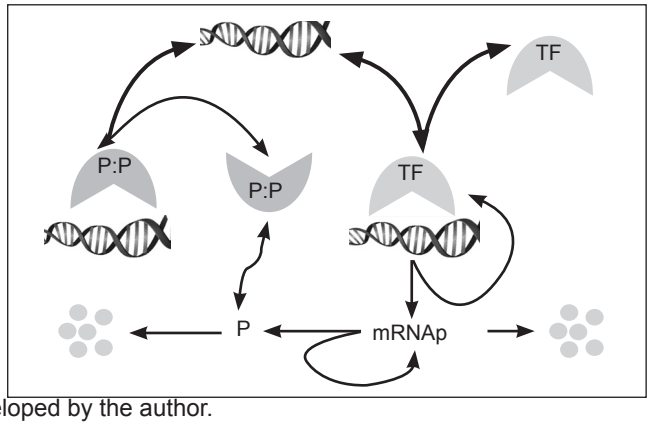
which is always higher than the variance $VAR_{EXP} = \lambda^{-2}$ of the exponential case. There is not a superior limit to the difference $VAR_{HIPER} - VAR_{EXP}$, obviously still considering variables that have the same expected value $\lambda^{-1}=1$. The minimum value of the difference is zero, and it is obtained in the degenerate case, when $\lambda_1=\lambda_2$. This implies that assuming the elementariness of a reaction when this is not true and the biological phenomenon is in fact realized through a multi-path reaction may turn out in very large errors in the modeling of the molecular noise.

4. Do we need to worry about?

To complete our analysis, we still need to understand whether an improper assumption about the elementariness of a reaction may have a consequence on the predictions of a model. In other words, how sensitive is the prediction to an incorrect modeling of the variance reaction times?

To this purpose, we consider a well-known molecular network, that has been studied through modeling and that is a building block of larger models. The network under examination is a gene expression network with a negative self-regulating feedback, and it is a recurrent motif in many organisms, including for instance the well-studied π phage (Arkin et al. 1998). We show (Figure 4) a cartoon representation of the network. The protein TF is a transcription factor for the gene expressing protein P. TF binds to DNA and allows transcribing the mRNA of the protein. The mRNA molecules get translated into protein molecules P. Molecules of P reversibly bind to form the protein homo-dimer species P:P, which reversibly binds to DNA, competing with TF. The complexation of P:P molecules with DNA makes it impossible to further transcribe the mRNA of P, thus establishing a negative self-regulation feedback. The droplet cartoon represents the result of a degradation reaction.

Figure 5. Cartoon of a self-regulating gene expression network



Source. Developed by the author.

The following set of biochemical reactions formally encodes the cartoon network shown in Figure 5. As for the notation, we use the form X: Y to indicate a complex formed by molecules of species X and Y, and the empty set symbol (Ø) to denote the product of a degradation reaction.

TF + DNA ↔ TF: DNA	**(r1)**
TF: DNA →TF: DNA+ $mRNA_p$	**(r2)**
$mRNA_p$ → $mRNA_p$ + P	**(r3)**
P + P ↔ P : P	**(r4)**
P : P + DNA ↔ P: P: DNA	**(r5)**
$mRNA_p$ →Ø	**(r6)**
P → Ø	**(r7)**

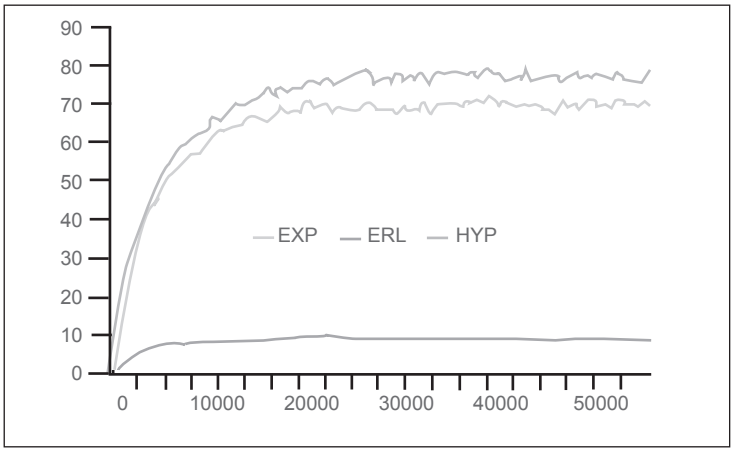
All the reactions in this model are always assumed to be elementary, see for instance (Samoilov and Arkin, 2006). However, let us just consider reaction (r2), which is a gene transcription reaction whereby a molecule of mRNA is produced. Gene transcription is a process that involves many species and thousands of reactions, see for instance Chap. 6 of (Alberts at al. 2002). Therefore, it is a legitimate question to ask whether assuming it is elementary would not have an impact on the predictions of the model.

Suppose the quantity we want to predict is the steady-state amount of protein P molecules. Without entering into the molecular details of transcription, we consider three alternative solutions modeling for reaction (r2). The first option is assuming it is an elementary reaction, the second one is to consider it as having a reduced variance, by using an Erlang distribution (Amari and Misra, 1997) for its occurrence times, and the third one is to model it as a reaction with a variance of occurrence times higher than the exponential case, through the usage of a hyper-exponential distribution (Amari and Misra, 1997). The three different distributions have the same expected value.

We couple each of the 3 modeling choices above with the same modeling of reactions (r1) and (r3)-(r7), for all of which we assume elementariness, and we plot in Figure 5 model simulation results for the three cases. On

the horizontal axis we report simulated time, and along the vertical axis we provide the time-courses of protein P abundance. These results come from stochastic simulations, and the curves have been obtained by the aggregation of 1000 different runs, for each scenario.

Figure 6. Comparison of simulated time-courses for the gene-expression network model, by varying the properties of the transcription reaction noise.



Source. Developed by the author.

We can observe (Figure 6) that the three model versions provide totally different predictions about the steady-state value of protein P and the time by when the steady-state is achieved. Hence, we can conclude that an inadequate modeling choice about the elementariness of a reaction may jeopardize the quality of the predictions obtained through a stochastic model of a living system.

5. Conclusion

We reviewed in this paper the assumptions that are at the basis of the applicability of the stochastic simulation approach a la Gillespie, pointing out that very often those assumptions are not properly considered. In particular, we focused on the assumption of elementariness of reactions, a condition that is normally assumed to be true without much research.

We showed through simple examples that the impact of a wrong modeling choice originated from the assumption of elementariness may profoundly affect the predictions that are obtained via stochastic simulation of models of living systems. This impact is due to the erroneous modeling of the noise in the reaction occurrence time. Therefore, the proper modeling of reaction noise reveals to be essential to gain confidence in the quality of predictions that can be obtained through stochastic simulation.

The conclusion we draw is that modelers should put always additional care in checking the validity of Gillespie's assumptions. Failure to do so may result in predictions that are very far from biological reality. Even more serious is the fact that, when tuning models to match experimental results, modelers may be forced to assign parameter values that do not retain any correspondence with those of the real biological system. Indeed, because of the inherent modeling error, forcing the predicted dynamics to match experimental data requires to assign values to parameters that are de facto meaningless.

An excellent example of this meaningless tuning of parameters can be found in the modeling paper about cell-cycle (Kar et al., 2009), where to match the observed variances of cell cycle time and cell division size distributions the authors were forced to assume very high values of degradation rates for the mRNAs, values that are orders of magnitude higher than the experimentally measured ones. A possible reason that could justify the impossibility of matching the cell-cycle statistics while using the experimental values for mRNA degradation rates is the erroneous modeling of mRNA degradation as an elementary reaction. Indeed, it has been reported that mRNAs and

protein degradations are indeed processes that involve multiple steps, such as deadenylation of mRNA (Decker and Parker, 1993) or poly-ubiquitination of proteins (Ciechanover, 1994). Moreover, in (Pedraza and Paulsson, 2008) it is shown that multi-step mRNA production or removal can significantly decrease protein level fluctuations. A model that considers the non-elementariness of mRNA production and degradation has been presented in (Csikász-Nagy Mura, 2010), which allowed reconciling parameter values and biological predictions.

As a final comment, we would like to stress the fact that quantitative experimental biology has been concerned so far with the measurement of purely average values, without considering noise (for instance variances) and even less distributions of measures. This study not shows that a correct modeling of noise is an important aspect when dealing with model validation, but also that the experimental characterization of noise has indeed to guide the choice between different modeling alternatives.

The emphasis on the quantification of average values is also largely due to the typical experimental approaches, which were normally aggregating measurements conducted at a population level. Nowadays, the technological advances allow conducting single-cell analysis of the dynamics of molecular species, therefore paving the way for new avenues of experimental and modeling research in biology. Hopefully, progresses in the quantitative measurement techniques will also allow getting the much needed information about the noise and other important characteristics of the reaction time occurrence.

6. References

Arkin, A., Ross, J. and McAdams, H.H. (1998). *Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected Escherichia coli cells*, Genetics, 49(4), pp. 1633-48.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P. (2002). *Molecular Biology of the Cell,* 4th edition, Garland Science. ISBN-10: 0-8153-3218.

Amari, S.V. and Misra, R.B. (1997). *Closed-form expressions for distribution of sum of exponential random variables,* IEEE Transactions on Reliability 46, pp. 519–522.

Cao, Y. and Samuels, D.C. (2009). *Discrete stochastic simulation methods for chemically reacting systems,* Methods in Enzymology, 454, pp. 115-140, doi: 10.1016/S0076-6879(08)03805-6.

Ciechanover, A. (1994) The ubiquitin-proteasome proteolytic pathway. Cell 79, 13-21.

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

Decker, C.J. and Parker, R. (1993). *A turnover pathway for both stable and unstable mRNAs in yeast: evidence for a requirement for deadenylation.* Genes Dev 7, 1632-1643.

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

Kar, S., Baumann, W.T., Paul, M.R. and Tyson, J.J. (2009). *Exploring the roles of noise in the eukaryotic cell cycle.* Proc Natl Acad Sci USA 106, 6471-6476.

Norris, J.R. (1997). *Continuous-time Markov chains I,* Cambridge Series in Statistical and Probabilistic Mathematics (No. 2), Cambridge UNIVERSITY Press, pp. 60-107, ISBN: 9780521481816, DOI: 10. 1017/CBO9780511810633.004.

Pedraza, J.M. and Paulsson, J. (2008). *Effects of molecular memory and bursting on fluctuations in gene expression.* Science 319, 339-343.

Samoilov, S.M. and Arkin, A.P. (2006). *Deviant effects in molecular reaction pathways,* Nature Biotechnology 24, pp. 1235-1240, doi:10.1038/nbt1253