A NOTE ON STOCHASTIC MODELING OF BIOLOGICAL SYSTEMS: AUTOMATIC GENERATION OF AN OPTIMIZED GILLEPSIE ALGORITHM

Quentin Vanhaelen (vanhaelen@insilicomedicine.com)

Abstract

Signaling pathways and gene regulatory networks (GRNs) play a central role in the signal transduction and regulation of biochemical processes occurring within the cellular environment. Understanding their mechanisms and dynamics is of major interest in various areas of life sciences and biological sciences. For example controlling stem cell fate decision requires a comprehension of the dynamical behavior of the networks involved in stem cell differentiation and pluripotency maintenance. In addition to analytical mathematical methods which are applicable for small or medium sized systems, there are many computational approaches to model and analyze the behavior of larger systems. However, from a dynamical point of view, modeling a combination of signaling pathways and GRNs present several challenges. Indeed, in addition to being of large dimensionality, these systems have specific dynamical features. Among the most commonly encountered is that the signal transduction controlled by the signaling pathways occurs at a different time scale than the transcription and translation processes. Also, stochasticity is known to strongly impact the regulation of gene expression. In this paper, we describe a simple implementation of an optimized version of the Gillespie algorithm for simulating relatively large biological networks which include delayed reactions. The implementation presented herein comes with a script for automatically generating the different data structures and source files of the algorithm using standardized input files.

Keywords: signaling pathway, chemical reactions, gene regulatory network, kinetic model, binary tree, dependency graph, delay, stochastic simulation, Gillespie algorithm.

Code availability: The Fortran90 implementation of the code and the R script described here as well as the tutorial with practical instructions are stored on the following github repository `qvhaelen/typhon`

1 Introduction

Using the new proteomic [1] and genomic data available, it is possible to study the organization of the cellular environment and its components [2, 3]. A current challenge concerns the elaboration of a comprehensive view of how these different components interact together. This includes the understanding of protein-protein interactions and gene regulation taken separately as well as the interactions between the proteosome and the transcriptome. This is not a trivial task because proteomic and transcriptomic data are obtained by different well defined experimental protocols. Studies are undertaken to establish dynamical links between these two types of data and although many questions remain unanswered, recent encouraging results have been obtained for example about the relationships between mRNA and protein concentrations [4, 5, 6, 7]. It is well-known that considering the complexity of the interactions taking place within the cellular environment, gaining an accurate understanding of the cellular biochemical processes requires a systemic approach where the cell is considered as a complex dynamical system [8]. Within this framework, it is possible to build a systemic description of the biological processes in terms of their functional and dynamical properties [9, 10]. From a practical point of view, a systemic approach is also necessary to understand how a local dysfunction of a small set of molecules affecting a restricted number of defined epigenetic [11] and metabolic processes [12, 13] may propagate to all parts of the cell leading to a progressive disruption of the general homeostasis [14]. The studies performed so far provide an interesting picture of how the cellular components are organized as dynamical motifs

and cycles [15, 16]. These dynamical motifs communicate together to achieve specific tasks. Inside the nucleus for example, the transcription factors interact together forming structured dynamical patterns called gene regulatory networks. These networks control and modulate genes expression via promoter silencing but supplementary mechanisms such as mRNA splicing [17, 18], chromatin remodeling [19] and epigenetic changes also intervene. Epigenetic modifications are a set of elaborate dynamical adaptations of the structure of the chromatin [20, 21] which contribute significantly to the regulation of gene transcription [22, 23, 24, 25].

Understanding and controlling the interaction between external perturbation, signal transduction and biological response is especially important within the field of stem cell research which focuses on understanding the two main mechanisms forming the backbone of stem cell fate decision, i.e., Cellular differentiation and pluripotency maintenance. Observations show that pluripotency state maintenance is a function of the external environment and input received. A dynamical balance between environmental clues and cell signals is required to preserve the self-renewal and tissue regenerative capacity of stem cells. Stem cell commitment and differentiation into a specific cell lineage is induced by modifying this dynamical balance with new external perturbations in order to activate or inhibit specific sets of signaling pathways which transmit external inputs to the core of trasnscription factor (TF) networks. Indeed, it is now experimentally established that pluripotency maintenance is essentially under the control of transcription factors [26, 27, 28] which act as master regulators of highly connected transcription networks [29, 30, 31, 32]. These TFs are continually attempting to specify differentiation to their own lineage of interest. This is the main reason why direct external intervention, through activation or inhibition of one or several signaling pathways is required to reinforce the pluripotent state or to drive differentiation [33, 34, 35]. Thus, the pluripotency state could be considered as a metastable state. The maintenance of the pluripotency state being a function of the external environment, any realistic dynamical description should include the cellular components responsible for processing this external signal. The TF core network being localized inside the nucleus, it receives this external signal through a second network of signaling pathways located inside the cytoplasmic compartment. Thus, it is essential to integrate The TF network and the signaling pathways within a single framework [36]. The behavior of the resulting extended regulatory networks can be better understood when defined as a complex dynamical system [8].

This discussion illustrates the importance of being able to simulate the dynamics of relatively large systems including those eliciting properties such as stochasticity and multiple scale dynamics. Although other computational methods such as constraint-based methods or pathway perturbation analysis methods using pathway maps can handle models of large dimensionality, kinetic-based models are the most appropriate for a detailed study of the dynamical features of the model [37, 38, 39]. Many works have been published regarding stochastic kinetic modeling of biological systems. Reviewing those contributions is outside the scope of this paper. The aim of this work is to shortly describe a stand alone implementation of an optimized version of the exact Gillespie algorithm. This code comes with a practical methodology to describe simple kinetic models which can be used as standardized inputs by a specifically designed R script to automatically generate the source files needed to perform the simulation. The complete description of the R script and a template of the standardized input files are stored with the code. In what follows, the mathematical framework to build the kinetic models is shortly described with a emphasis on the types of reactions allowed. Then, the algorithmic method and the mean features of its implementation are summarized. This is followed by the discussion of a small case study using a single signaling pathway as an example.

2 Definition of the system

2.1 Mathematical formulation

An extended regulatory network is a dynamical system of biochemical components interacting together through a set of biochemical reactions. In order to allow the easy combination of various pathways together in a flexible way, the kinetic is described using a scheme based on the law of mass action. It offers a good mathematical stability and easy systematic interpretation in terms of biological processes. Considering the nature of the dynamics encountered in signaling cascades and gene regulatory networks, all reactions are described in terms of elementary processes and other more specific kinetic schemes and approximations will not be considered.

Formally, when analyzing the dynamics of a biological network, one is concerned with the temporal evolution of the set of m species[1] $X_i \quad (i = 1, ..., m)$ which composes the network. These species react together through a set of r chemical elementary reactions. The derivation of the dynamics of the network is based on the law of mass action which states that, for an elementary reaction, that is, a reaction in which all of the stoichiometric coefficients of the reactants are one, the rate of reaction is proportional to the product of the concentrations of the reactants [40]. Although the mass action kinetics scheme expands the dimension of the model, it provides a regularized and simplified mathematical structure. This feature makes further analysis more tractable. Formally, one defines the stoichiometric matrix $\mathcal{N}_{ij} = p_{ij} - r_{ij}$ with $(i = 1, ..., m)$ and $(j = 1, ..., r)$ and the reaction rate vector as $v_j = k_j \prod_{s=1}^{n} X_s^{r_{sj}}$ with $(j = 1, ..., m)$ where the p_{ij} term is the stoichiometric coefficient for the $i-$th species in the $j-$th reaction if appearing on the right of the reaction arrows, r_{ij} if on the left; k_j is the forward rate constant [41]. Using $\mathcal{N}$ and $\vec{v}$, the function $\vec{F}(\vec{x})$ is built such that $F_i(\vec{X}) = \sum_{j=1}^{r} (p_{ij} - r_{ij})\, v_j \qquad (i = 1, ..., m)$. Each $F_i(\vec{X})$ represents the contribution of the reactions acting on the species X_i.

In our description all the chemical reactions occurring in the extended regulatory network are modeled using the following elementary reactions[2]. Association (between two different molecules or between similar molecules (dimerization)) $X_1 + X_2 \rightarrow X_3$, dissociation (giving two identical molecules or two similar molecules) $X_3 \rightarrow X_1 + X_2$, creation (this reaction mimic the dynamic of molecules for which no explicit creation pathway is included) $\varnothing \rightarrow X$, degradation (last step of any degradation process including lysosomal and proteosomal pathway) $X \rightarrow \varnothing$, direct activation/inhibition (when enzyme can be neglected, like phosphatase, etc.) $X \rightarrow X^*$. Nevertheless, For any chemical process involving enzymes (E) which can not be ignored we use the Michaelis-Menten scheme:

$$S + E \rightleftharpoons (ES) \rightarrow P + E \tag{1}$$

Where S is the substrate and P the product of the reaction. We model these three reactions independently as follows:

$$S + E \xrightarrow{k_1} (ES) \tag{2}$$

$$(ES) \xrightarrow{k_{-1}} S + E \tag{3}$$

$$(ES) \xrightarrow{k_2} P + E \tag{4}$$

[1]The term *species* must be understood as holding for any molecular component present, even for a short time, inside the system: a receptor is a species, a ligand is another type of species and the molecule composed by a receptor bound to its ligand is also another species. A given species and its activated form, i.e. after phosphorylation, are considered as two distinct species.

[2]As usual each reaction is characterized by a kinetic constant giving the rate of the reaction.

The parameters appearing in the Michaelis-Menten formula

$$V_0 = V_{\text{max}} \frac{S}{K_{\text{M}} + S} \tag{5}$$

are connected to the rate constants as follows

$$\begin{cases} V_{\text{max}} = k_2 E_T \\ E_T = E + (ES) \\ K_{\text{M}} = \frac{k_{-1}+k_2}{k_1} \end{cases}$$

In general, the reversible chemical reactions are always modeled as a set of separate irreversible forward reactions. The set of reactions described above is enough to describe any kind of process occurring in most of the signaling pathways.

The dynamics of GRNs is mainly concerned with the regulation of gene expression [42, 43, 44, 45, 46]. In this work, DNA and ribosome are not explicitly taken into account in the kinetic equations. Ribosome could be included with a complete description of ribosomal proteins assembly but, as discussed below, it leads to much more complexity. We propose to model the regulation of gene expression using the following scheme:

$$\text{TF}_{\text{cytosol}} \rightarrow \text{TF}_{\text{nucleus}} \tag{6}$$

$$\text{TF}_{\text{nucleus}} \rightarrow \text{TF}_{\text{cytosol}} \tag{7}$$

$$\text{TF}_{\text{nucleus}} + \text{DNA}_{\text{free}} \rightarrow \text{DNA}_{\text{free}} + \text{mRNA}_{\text{free-nucleus}} + \text{TF}_{\text{nucleus}} \tag{8}$$

$$\text{mRNA}_{\text{free-nucleus}} \rightarrow \text{mRNA}_{\text{free-cytosol}} \tag{9}$$

$$\text{Ribosome}_{\text{free}} + \text{mRNA}_{\text{free-cytosol}} \rightarrow \text{Ribosome}_{\text{free}} + \text{mRNA}_{\text{free-cytosol}} + \text{Protein}_{\text{cytosol}} \tag{10}$$

$$\text{mRNA}_{\text{free-cytosol}} \rightarrow \varnothing \tag{11}$$

$$\text{Protein}_{\text{cytosol}} \rightarrow \varnothing \tag{12}$$

When a gene is repressed by one or several genes binding the promoter site, the effect can be included using a modification of the transcription rate constant as follows:

$$\tilde{k}_l = k_l \frac{1}{1 + \sum_{i=1}^{N} k_i x_i} \tag{13}$$

Where k_l is the non modified transcription rate, N the number of genes that can repress the promoters, x_i the concentration in the nucleus of repressor i and k_i the intensity of the repression.

2.2 GRNs and stochasticity

Experimental evidences show that a lot of mechanisms occurring in a cell and especially within GRN elicit stochastic properties. For example, stochastic expression of key genes (Nanog, Sox2, Oct4,) can lead to heterogeneous populations of stem cells even when starting from a homogeneous culture in a homogeneous environment. In a regulatory network, stochasticity can have different origins including the low number of molecules usually involved in molecular processes occurring in a cell, the effect of external perturbations and response of signaling pathways. Furthermore, the transcription/translation processes are also stochastic with an impact on the release of proteins. It has also been shown that dynamical sub-networks can increase (positive feedback) or reduce (negative feedback) stochastic effects. Taking into account the stochastic nature of the dynamics is thus of primary importance when simulating the dynamics of an extended regulatory network and appropriate methods must be chosen.

2.3 Extended regulatory network and multi-scale processes

Molecular events, such as dimerization and phosphorylation, occurring in signaling pathways and processes such as TF binding, transcription/ translation and mRNA transfer occur on different timescales. Regarding the transcription and translation, one observes a delay between the beginning of these processes and the release of the final product. It is known that systems involving delayed dynamics can elicit specific behavior [47, 48]. From a computational perspective, including dynamical events occurring at various timescale within a single dynamical model can be complex. Two approaches can be used. If one stands from the point of view of the slowest processes, one will assume that the fastest processes have already reached their equilibrium state when the slowest events take place. Otherwise, one takes the point of view of the fastest processes and in that case, one needs to take into account the fact that there is a time delay between the release of the products by the fastest reactions and the release of the product by the slowest reactions. It is this second point of view which is considered in this work. Thus, we divide the set of reactions into two distinct subsets. Firstly, we assume that the fastest chemical reactions occurs immediately (products are released at once). Secondly, the slower chemical reactions are supposed to release their products with a delay which must be specified. Thus we need to know the delay between the beginning and the release of the products for the slower processes. The exact duration of transcription rate and translation rate usually varies from one gene to another and other chemical or physiological parameters can intervene. Nevertheless, these rates are always slower than the dimerization mechanisms. Modeling multi-scale stochastic systems can be done using dSSA (exact formulation of a SSA including delay, see description below) which allows considering the delay when the model includes processes occurring on different time scales.

3 Stochastic simulation: an optimized version of the Gillespie algorithm

There are two formalisms for mathematically describing the time behavior of a spatially homogeneous chemical system. The deterministic approach regards the time evolution as a continuous, wholly predictable process which is governed by a set of coupled ODEs. The stochastic approach regards the time evolution as a kind of random-walk process which is governed by a single differential-difference equation (the master equation). In practice, many modeling studies rely on using different flavors of ODE-based formulations, probably because many scientists are more familiar with purely deterministic approaches and that formulating a model in terms of ODEs and solving this system is most of the time straightforward. Although, deterministic modeling relies on several strong assumptions, it can provide suitable and accurate results in many situations. Furthermore, formulating the dynamics of the system in terms of ODEs makes easier to perform additional analytical or numerical analysis of the underlying dynamical properties of the system.

Moreover, when considering dynamical systems eliciting stochastic behavior,such as GRNs or signaling pathways whose dynamics is directly connected to a GRN, fairly simple kinetic theory arguments show that the stochastic formulation of chemical kinetics has a firmer physical basis than the deterministic formulation. However, the fundamental stochastic master equation becomes quickly mathematically and computationally intractable when the dimensionality of the system becomes large, i.e., for systems with more than 5 or 6 variables. This problem is well known and different approaches have been suggested to overcome this issue [49, 50, 51, 52, 53, 54]. One of the most famous is the Gillespie method, also known as the stochastic simulation algorithm (SSA), which allows making exact numerical calculations within the framework of the stochastic formulation without having to deal with the master equation directly [55]. It uses a rigorously derived Monte Carlo procedure to simulate the time evolution of the

given chemical system. Like the master equation, the SSA correctly accounts for the inherent fluctuations and correlations that are necessarily ignored in the deterministic formulation. In addition, unlike most procedures for numerically solving the ODE, this algorithm never approximates infinitesimal time increments.

The SSA or Gillespie algorithm, were initially described in the fundamental paper [55]. The key steps of this algorithm, called the direct method (DM), can be summarized as follows:

Step 0**:** (Initialization). Input the desired values for the M reaction constants $c_1, ..., c_M$ and the N initial molecular population numbers $X_1, ..., X_N$. Set the time variable t and the reaction counter n both to zero. Initialize the Pseudo Random Number Generator (PRNG).

Step 1**:** Calculate and store the M quantities $a_\mu = c_\mu h_\mu$, (where h_μ is the number of distinct molecular reactant combinations available in the state for reaction μ), for the current molecular population numbers. Also calculate and store as a_0 the sum of the M a_μ values.

Step 2**:** Generate two random numbers r_1 and r_2 using the PRNG, and calculate τ and μ according to

$$\tau = \frac{1}{a_0} \ln \frac{1}{r_1} \tag{14}$$

and

$$\sum_{\nu=1}^{\mu-1} < r_2 a_0 \leq \sum_{\nu=1}^{\mu} a_\nu \tag{15}$$

Step 3**:** Using the τ and μ values obtained in step two, increase t by τ, and adjust the molecular population levels to reflect the occurrence of one R_μ reaction. Then increase the reaction counter n by 1 and return to step one.

The SSA is a very accurate method but when the number of reactions becomes large, the DM becomes too slow to be useful in practice. It is necessary to use several specific algorithmic techniques which can speed up the algorithm while keeping its accuracy intact. Here we provide a general description of the different approaches used in our implementation. The strategy is to focus on the steps of the DM which are the most computationally expensive.

Dependency graph In the DM, after a reaction has been fired, all the a_μ coefficients are updated. It easy to see that only a few of them are actually modified when a reaction occurs. An appropriate way to take this fact into account is to implement a dependency graph. For each reaction, this data structure stores the label of the a_μ which are modified. For each reaction, the set of a_μ to be updated is constructed as follow: we consider the label of the species whose belongs to the set of reactants or products of the given reaction. If any of these species appears in the set of reactants of a reaction i, then the corresponding $a_{(\mu=i)}$ is added in the set of a_μ to be updated. It is also useful to save the smallest label which must be updated [56].

Partial Summation Another expensive step in the DM is the computation of a_0. It is possible to reduce the number of operations by building a set of partial sum over the a_μ. For a system with M chemical reactions, a set of M partial sums is constructed. A partial sum is defined as follows:

$$\tilde{a}_\mu = \sum_{\nu=1}^{\mu} a_\nu \tag{16}$$

Using a recursive formula it is possible to compute each partial sum from the previous one. Taking into account that $a_0 = \tilde{a}_M$, and the fact that we know the smallest label $\tilde{\mu}$ which has been modified it is possible to restrict the computation of a_0 to the subset of partial sums $\nu \in \{\tilde{\mu}, M\}$.

Binary tree data structure and binary tree search Locating the reaction to fire next is the most expensive step of the DM. A way to reduce the cost of this step is to make use of a specific data structure called binary tree and the related binary tree search algorithm [57]. This approach is known as the Logarithmic Direct Method (LDM) [58] and the use of binary tree search reduces the search depth to $O(\log M)$. Furthermore, for LDM, the average search depth is independent of the ordering of the reactions. Hence there is no need for a pre-simulation.

Contracting the stoichiometric matrix Another step of the DM which is computationally expensive is the update of the number of molecules of each species. This update depends on the label of the reactions fired. In the direct method the stoichiometric coefficients of the chemical reactions are stored in a matrix whose dimensions scale with the number of reactions and species of the system. A key property of this matrix is that it is usually sparse, that is, for each reaction there are only a few species which must be updated. It means that in the DM, most of the operations are additions of zero terms. A way to reduce the computational cost of the update is to use sparse matrix techniques. In our case we have implemented a supplementary structure which contains for each reaction the label of the non-zero stoichiometric coefficients. Thus the update is done by taking into account the labels appearing in this structure.

In addition to optimizing the performances of the standard SSA, supplementary adaptations must be done to include the effects of multi scale dynamics. Indeed, in the initial formulation of the Gillespie method, it is assumed that all reactions occur instantly, i.e., products are released instantly. While this is true in many cases, it is also possible that some chemical reactions take certain time to finish after they are initiated. Thus, the product of such reactions will emerge after certain delays. It is important to distinguish two different kinds of delayed reactions. The *non consuming reaction*, where the reactants of an unfinished reaction can participate in a new reaction and the *consuming reaction*, where the reactants of an unfinished reaction cannot participate in a new reaction.

The main difference between consuming and non consuming reactions occurs at the level of the evolution of the reactant: When a non consuming reaction occurs, the population of the reactants does not change and the involved species can participate in a new reaction; however, when a consuming reaction occurs, the population of the reactants changes immediately, and the molecules involved in a consuming reaction can not participate in a new reaction.

There are many algorithms which have been released to integrate delayed reactions into the standard SSA scheme. However, many of them rely on various approximations which although allowing to mimic the effects of delayed reactions, do not constitute a mathematically accurate implementation of delayed reactions with th SSA. In this work, the algorithm chosen [59], is an exact SSA algorithm for chemical reaction systems with delays. The algorithm is exact, since it is rigorously developed based upon the fundamental premise of stochastic chemical kinetics derived by Gillespie. The difference between the complexity of the algorithm [59] and other exact methods like rejection algorithm mainly lies in the number of random variables generated by two algorithms. Specifically, the rejection algorithm needs to generate up to 50% more random variables than the algorithm presented in [59], if the chemical reaction system is dominated by reactions with delays. On the other hand, if the reaction system is dominated by reactions without delays, the rejection algorithm still generates slightly more random variables than the algorithm [59].

4 Examples of simulation

4.1 The model: The LIF pathway

The model is based on a work published by Mahdavi et al. [60] to characterize the dynamics of the LIFR-GP130 signaling pathway. This pathway is well known as a key regulator involved in the pluripotency maintenance of mouse embryonic stem cell [44]. Initially formulated as a set of ODEs, this model includes the formation of the activated receptor complex between LIFR and GP130 as well as its trafficking along the membrane and the effects of the negative feedback loop by SOCS3. The second part of the model is the activation of the STAT3 dimer via endogenous and exogenous phosphorylation. Activation of STAT3 dimer is followed by its nuclear translocation where it acts as a transcription factor for SOCS3 and other key pluripotency factors such as NANOG and OCT4. Moreover, it has been recently shown that NANOG is also involved in a positive feedback loop to increase STAT3 signaling by inhibiting SOCS3 transcription [61]. Although rather simple, this model is useful as it provides the user with a concrete example of how the different types of reactions described above should be implemented to use the code described herein. In what follows, we use the results of simulations performed with the SSA to illustrate the main dynamical features of this model.

4.2 Dynamics without multi-scale dynamics

The simulation starts without ligand and only GP130 receptors, LIF receptors and cytoplasmic STAT3 dimers are present. The model assumes that endogenous phosphorylation of cytoplasmic STAT3 dimers (Tyr705-phosphorylated 2STAT3) takes place. Thus as chemical reactions occur, the concentration of phosphorylated STA3 dimers (p2STAT3) increases inside the cytoplasm and can translocate inside the nucleus. Although present at low concentration, the level of p2STAT3 inside the nucleus is enough to induce the basal transcription of SOCS3 mRNA. SOCS3 mRNA can translocate into the cytoplasm where translation of SOCS3 protein occurs. Once the system has reached its equilibrium, LIF inducer is added. From a dynamical point of view, addition of LIF can be considered as an external perturbation which is applied on the cellular network[3]. LIF first binds to LIFR receptor which then form a receptor complex with GP130 co-receptor. Figure 1 shows the decrease of LIFR and GP130 concentration upon addition of external ligands for various concentrations.

Once the formation and activation of the LIFR-GP130-JAK is achieved, STAT3 dimers bind to the complex and become phosphorylated. As illustrated on Figure 2, the equilibrium switches from a state of a high concentration of cytoplasmic STAT3 dimer combined with a low concentration of nuclear p2STAT3 to a new state characterized by a lower concentration of cytoplasmic 2STAT3 with a higher concentration of nuclear p2STAT3.

Higher concentration of p2STAT3 inside the nucleus leads to a significant increase of SOCS3 mRNA, see Figure 3. As a result, level of cytoplasmic SOCS3 increases. Upon addition of LIF, the negative feedback loop established between SOCS3 proteins and the receptor complex plays an important in the regulation of the signal strength. This is essentially because binding of cytoplasmic SOCS3 to the LIFR-GP130-JAK complex is the main step leading to the final degradation of the receptor complex. The strong decrease in SOCS3 mRNA concentration which is observed right after LIF addition is the direct

[3]For each simulation, it is required to wait until initial equilibrium in absence of LIF is reached. In the corresponding pictures, the timescale has been rescaled accordingly to hide this first part of the dynamic.

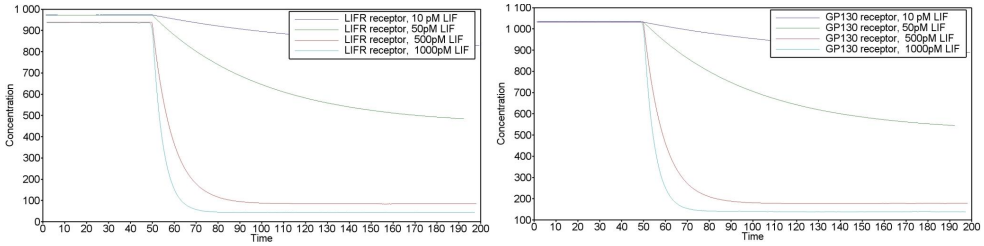

Figure 1: Response of receptors with respect to addition of LIF inducer. Response of LIFR (Left). Response of GP130 (Right).

consequence of the activation of this feedback, Fig. 3.

The purpose of the negative feedback loop is to improve the stability of the system by reinforcing the regulation of the signal. In the situation analyzed here, the feedback loop allows the degradation of the receptor complex which in turn induces the reduction of the exogenous phosphorylation of STAT3 dimers. As a result, nuclear p2STAT3 concentration decreases from its highest level almost directly after LIF addition, Fig. 2, and the system reaches a new equilibrium state characterized by higher level of p2STAT3 and SOCS3 (C). The kinetic parameters controlling the rate of association and dissociation of the receptor complexes as well as the translocation rate of 2STAT3 have a critical influence on the behavior of the cellular network.

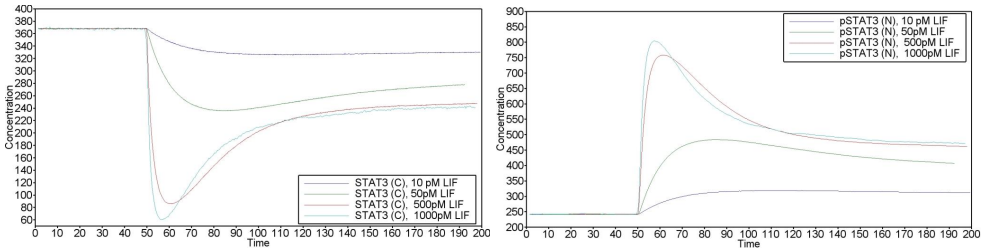

Figure 2: Evolution of the 2STAT3 concentration upon LIF addition in the cytoplasm (Left). LIF addition leads to an increase of the activated form of 2STAT3 dimer concentration inside the nucleus (Right)

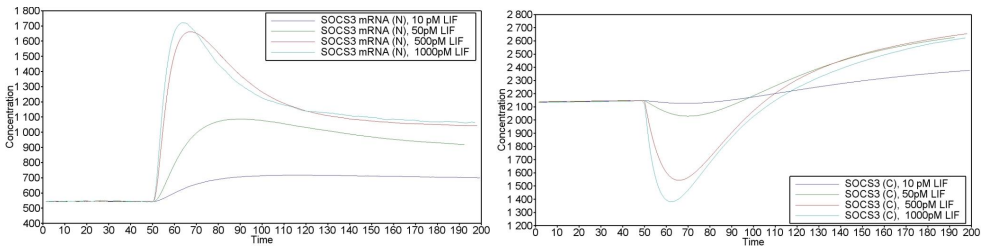

Figure 3: Following LIF addition, transcription of SOCS3 increases. SOCS3 mRNA in the nucleus (Left). SOCS3 protein concentration inside the cytoplasm (Right)

The model presented here also contains the PIAS3 inhibitor. PIAS3 is present inside the cytoplasmic environment as well as inside the nuclear compartment. In the cytoplasmic compartment, PIAS3 acts as an inhibitor by forming a complex with the activated form of 2STAT3 preventing its translocation into the nucleus. In the nucleus, PIAS3 may also binds to the 2STAT3 preventing its export and further activation by phosphorylation. As a consequence, the level of available STAT3 within the phosphorylation loop is

reduced. This second mechanism also contributes to the regulation of the signal strength.

Upon withdrawal of LIF, the ability of the cellular network to come to back to its initial state is a function of the kinetic parameters controlling the deactivation and degradation of the receptor complexes. Thus the frequency of addition and withdrawal of LIF can have an impact on the level of cytoplasmic p2STAT3 and SOCS3 protein. To illustrate this, we show the behavior of the main components of the cellular network for a periodical addition and withdrawal of LIF on Fig. 4.

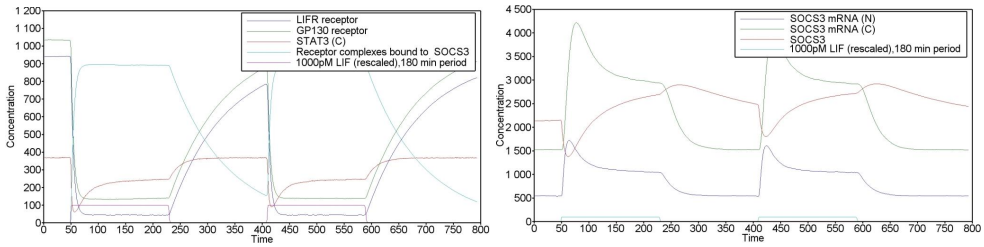

Figure 4: Temporal evolution of the main components of the pathway for a periodical addition of LIF inducer. Data are shown for a periodical addition of 180 minutes.

4.3 Multi-scale dynamics: effects of delayed reactions

To conclude we illustrate the effects of adding delay within the dynamical system for some reactions. We consider a delay for the transcription of SOCS3-mRNA (1.58 min), for the nuclear export of SOCS3-mRNA (1.92 min), for the import and export of the STAT3 dimer (1.2 min and 1.4 min) and for the translation of SOCS3 protein (100 min). The results are shown of Fig. 5 and Fig. 6. In order to understand how the imposed delay can affect the dynamics, the results are shown for different values of the delay for the translation of SOCS3. As expected, the addition of delay impacts how the system react to the LIF addition but also the concentration levels corresponding to the stationary states with and without LIF. The saturation observed for the activated form of the STAT3 dimer (Fig. 5 bottom) and SOCS3-mRNA inside the cytoplasm (Fig. 6 top) and the complete absence of SOCS3 proteins (Fig. 6 bottom) upon LIF addition varies according to the value chosen for the delay of the SOCS3 translation. From a dynamical point of view thecounter-reaction of the system to the absence of SOCS3 protein due to very large delay is a direct effect of the negative feedback loop involving SOCS3 described above. Indeed, upon LIF addition, the complex formed by LIFR and GP130 can be activated and trigger, without delay, the degradation of the amount SOCS3 already present within the system. As before, there is also a translocation of a higher concentration of activated STAT3 dimer which can increase SOCS3-mRNA transcription. However, because the time required to release SOCS3 (up to 100 minutes) is significant, the increase in SOCS3 transcription cannot directly compensate for the degradation immediately taking place. When all SOCS3 is eliminated from the system, this feedback loop ceases to function. This switches the equilibrium for the concentration levels of activated STAT3. When SOCS3 proteins begin to be released, the system undergoes a dynamics similar to the case without delay with a strong increase of SOCS3 proteins and a strong decrease of activated STAT3 dimer, the behavior being more accentuated as the value of the delay increases. The system finally reaches a new steady state which takes longer to be established than in the case without delay.

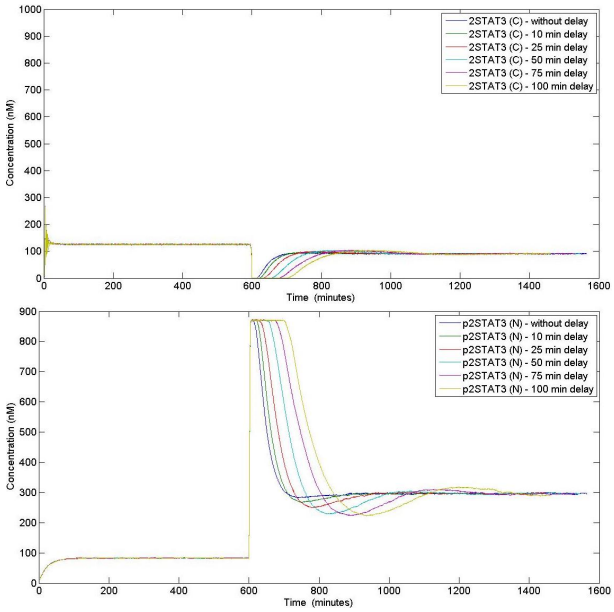

Figure 5: Temporal evolution of the dimer STAT3 inside the cytoplasm (Up) and for the activated dimer STAT3 inside the nucleus (Bottom) for different values of the delay for the translation of SOCS3.

5 Conclusion

In this document, we have described the mathematical and computational characteristics of a simple and ready to use implementation of the Gillespie algorithm suitable for simulating the dynamics of stochastic chemical systems with delayed reactions. The implementation and its associated R script is adapted to dynamical systems (re)written as a set of chemical reactions following the law of mass action and assuming that the system is written in terms of elementary reactions only. The practical advantage of this R script is that the source files and especially heavy data structures, such as binary trees, required for this kind of code can be automatically generated. The only assumption is that the model obeys the restrictions on the type of reactions and the order of these reactions.

Although the present code can handle relatively large systems[4], it remains a serial version only and for simulating very large systems, other implementations (either parallelized or running on GPU for instance) should be used. It is worth emphasizing that what makes stochastic simulations so time consuming is the necessity to simulate hundreds or thousands of times the behavior of the system to obtain correct averaged physical quantities. A straightforward way to speed up the process is to run several batches of runs on different cores rather than using a single core. The performance of this kind of simulator depends on many factors ranging from the kind of machine used to the properties of the system to be simulated such as number of reactions, number of different species and types of reactions. Furthermore, adding delayed reactions can significantly increase the computational time and increasing the total number of molecules within the system decreases the average time step of the algorithm (this is obvious when checking the formula given by eq. 14) which in turn leads to an increase of the computational required for reaching a given physical duration.

Nevertheless, this implementation should be useful for students and researchers familiar with R and Fortran90 and looking for gaining some experience with the use of stochastic simulation algorithms. In all cases, one should keep in mind that when studying the dynamics of biological networks, the simulation

[4]The present implementation was used to simulate a extended network made of 509 species and 892 elementary chemical reactions obtained by combining different signaling pathways using an extended automatized pipeline

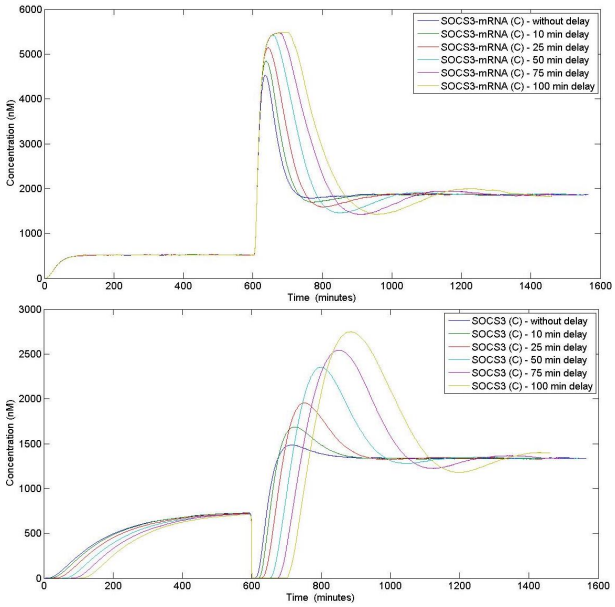

Figure 6: Temporal evolution for SOCS3-mRNA inside the cytoplasm (Up) and for SOCS3 protein inside the cytoplasm (Bottom) for different values of the delay for the translation of SOCS3.

itself is usually the very beginning of a long journey during which various kinds of computational and analytical methods might required depending on the purpose of the study and the questions which must be answered.

References

[1] Altelaar AF, Munoz J, Heck AJ. Next-generation proteomics: towards an integrative view of proteome dynamics. Nat Rev Genet 2012; 14(1): 35-48.

[2] Lowell S. Getting the measure of things: the physical biology of stem cells. Development 2013; 140: 4125-8.

[3] Gruebele M, Thirumalai D. Perspective: Reaches of chemical physics in biology. J Chem Phys 2013; 139(12): 121701.

[4] Jansen R, Greenbaum D, Gerstein M. Relating Whole-Genome Expression Data with Protein-Protein Interactions. Genome Research 2002;12:37-46.

[5] Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J et *al.* Global quantification of mammalian gene expression control. Nature. 19 May 2011;473:337-342.

[6] Li JJ, Bickel PJ, Biggin MD. System wide analyses have underestimated protein abundances and the importance of transcription in mammals. 2014. PeerJ 2:e270; doi:10.7717/peerj.270

[7] Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet; 13(4): 227-232. doi:10.1038/nrg3185.

[8] Liang J, Luo Y, Zhao H. Synthetic biology: putting synthesis into biology. Wiley Interdiscip Rev Syst Biol Med 2011; 3(1): 7-20.

[9] Barabasi AL, Oltvai ZN. Network biology: understanding the cells functional organization. Nat Rev Genet 2004; 5: 101-13.

[10] Bruggeman FJ, Westerhoff HV. The nature of systems biology. Trends Microbiol 2006; 15(1): 45-50.

[11] Gentilini D, Mari D, Castaldi D, Remondini D, Ogliari G, Ostan R et al. Role of epigenetics in human aging and longevity: genome-wide DNA methylation profile in centenarians and centenarians offspring. AGE. 2013; 35: 1961-73.

[12] Tavernarakis N. Ageing and the regulation of protein synthesis: a balancing act? Trends Cell Biol 2008. 18(5): 228-35.

[13] Shimizu I, Yoshida Y, Suda M, Minamino T. DNA Damage Response and Metabolic Disease. Cell Metab 2014; 20: 967-77.

[14] Vanhaelen Q. Aging as an Optimization Between Cellular Maintenance Requirements and Evolutionary Constraints. Current Aging Science, 2015, 8, 110-119

[15] Rajapakse I, Scalzo D, Tapscott SJ, Kosak ST, Groudine M. Networking the nucleus. Mol Syst Biol 2010; 6: 395.

[16] Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet 2007; 8: 450-61.

[17] Matera AG, Wang Z. A day in the life of the spliceosome. Nat Rev Mol Cell Biol 2014; 15: 108-21.

[18] Bentley DL. Coupling mRNA processing with transcription in time and space. Nat Rev Genet 2014; 15: 163-75.

[19] Varga-Weisz PD. Chromatin remodeling: a collaborative effort. Nat Struct Mol Biol 2014; 21(1): 14-6.

[20] Jenuwein T, Allis CD. Translating the Histone Code. Science 2001; 293: 1074-80.

[21] Richards EJ, Elgin SCR. Epigenetic Codes for Heterochromatin Formation and Silencing: Runding up the Usual Suspects. Cell 2002; 108: 489-500.

[22] Marmorstein R. Protein modules that manipulate histone tails for chromatin regulation. Nat Rev Mol Cell Biol 2001; 2: 422-32.

[23] Grummt I, Pikaard CS. Epigenetic Silencing of RNA polymerase I transcription. Nat Rev Mol Cell Biol 2003; 4: 641-9.

[24] Klose RJ, Zhang Y. Regulation of histone methylation by demethylimination and demethylation. Nat Rev Mol Cell Biol 2007; 8: 307-18.

[25] Wagner EJ, Carpenter PB. Understanding the language of Lys36 methylation at histone H3. Nat Rev Mol Cell Biol 2012; 13: 115-26.

[26] Thomson M, Liu SJ, Zou LN, Smith Z, Meissner A, Ramanathan S. Pluripotency Factors in Embryonic Stem Cells Regulate Differentiation into Germ Layers. Cell 2011; 145: 875-89.

[27] Walker E, Ohishi M, Davey RE, Zhang W, Cassar PA, Tanaka TS. et al. Prediction and Testing of Novel Transcriptional Networks Regulating Embryonic Stem Cell Self-Renewal and Commitment. Cell Stem Cell 2007; 1: 71-86.

[28] Tantin D. Oct transcription factors in development and stem cells: insights and mechanisms. Development 2013; 140: 2857-66.

[29] Iglesias-Bartolome R, Gutkind JS. Signaling circuitries controlling stem cell fate: to be or not to be. Curr Opin Cell Biol 2011; 23: 716-23.

[30] Ng HH, Surani MA. The transcriptional and signalling networks of pluripotency. Nat Cell Biol 2011; 13(5): 490-6.

[31] Dalton S. Signaling networks in human pluripotent stem cells. Curr Opin Cell Biol 2013; 25(2): 241-6.

[32] M. Herberg and I. Roeder, Computational modelling of embryonic stem-cell fate control, Development, 2015, 142(13), 2250-2260.

[33] Silva J, Smith A. Capturing Pluripotency. Cell 2008; 132: 532-6.

[34] Nowick K, Stubbs L. Lineage-specific transcription factors and the evolution of gene regulatory networks. Brief Funct Genomic 2010; 9(1): 65-78.

[35] Loh KM, Lim B. A Precarious Balance: Pluripotency Factors as Lineage Specifiers. Cell Stem Cell 2011; 8: 363-9.

[36] Chen X, Xu H, Yuan P, Fang F, Huss M, Vega VB *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. Cell 2008 June 13;133:1106-1117.

[37] Vanhaelen Q. , Aliper A.M., Zhavoronkov A. A comparative review of computational methods for pathway perturbation analysis: dynamical and topological perspectives. Mol. BioSyst., 2017, 13, 1692. doi: 10.1039/c7mb00170c

[38] Li X, Shen L, Shang X, Liu W Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway, PLoS One, 2015, 10(7), e0132813.

[39] Khatri P, Sirota M, Butte AJ, Ten years of pathway analysis: current approaches and outstanding challenges, PLoS Comput Biol., 2012, 8(2), e1002375.

[40] Steinfeld JI, Francisco JS, Hase WL. Chemical Kinetics and Dynamics. Upper Saddle River, NJ: Prentice-Hall,1989.

[41] Beretta E, Vetrano F, Solimano F, Lazzari C. Some results about nonlinear chemical systems represented by trees and cycles. Bulletin of Mathematical Biology. 41. 641-664.

[42] Chickarmane V, Peterson C. A Computational Model for Understanding Stem Cell Trophectoderm and Endoderm Lineage Determination. PLoS ONE 2008;3(10):e3478. doi:10.1371/journal.pone.0003478

[43] Xu H, Ang YS, Sevilla A, Lemischka IR, Maayan A. Construction and Validation of a Regulatory Network for Pluripotency and Self-Renewal of Mouse Embryonic Stem Cells. PLoS Comput. Biol. 2014;10(8):e1003777. doi:10.1371/journal.pcbi.1003777

[44] Martello G, Berton P, Smith A. Identification of the missing pluripotency mediator downstream of leukaemia inhibitory factor. The EMBO Journal 2013;32:2561-2574.

[45] Zhang X *et al.* Esrrb Activates Oct4 Transcription and Sustains Self-renewal and Pluripotency in Embryonic Stem Cells. J. Bio Chem. 2008;283(51):35825-35833.

[46] Ye S, Li P, Chang Tong, Ying QL. Embryonic stem cell self-renewal pathways converge on the transcription factor Tfcp2l1. The EMBO Journal 2013;32:2548-2560.

[47] Batzel JJ., Kappel F. Time delay in physiological systems: Analyzing and modeling its impact. Mathematical Biosciences 234 (2011) 61-74. doi:10.1016/j.mbs.2011.08.006

[48] Glass L, Beuter A, Larocque D. Time Delays, Oscillations, and Chaos in Physiological Control Systems. Mathematical Biosciences. 90:111-125 (1988)

[49] Morton-Firth CJ, Bray D. Predicting Temporal Fluctuations in an Intracellular Signalling Pathway. J. theor. Biol. (1998) 192, 117-128.

[50] Shimizu TS, Aksenov SV, Bray D. A Spatially Extended Stochastic Model of the Bacterial Chemotaxis Signalling Pathway. J. Mol. Biol. (2003) 329, 291-309. doi:10.1016/S0022-2836(03)00437-6

[51] Caia X, Xu Z. K-leap method for accelerating stochastic simulation of coupled chemical reactions. J. Chem. Phys. 126, 074102 (2007). doi: 10.1063/1.2436869

[52] Marquez-Lago et al. Probability distributed time delays: integrating spatial effects into temporal models BMC Systems Biology 2010, 4:19.

[53] Ramaswamy R, Gonzalez-Segredo N., Sbalzarinic F.I. A new class of highly efficient exact stochastic simulation algorithms for chemical reaction networks. J. Chem. Phys. 130, 244104 (2009)

[54] Barik D, Paul MR, Baumann W.T.,Cao Y., Tyson J.J. Stochastic Simulation of Enzyme-Catalyzed Reactions with Disparate Timescales. Biophysical Journal Volume 95 October 2008 3563-3574

[55] Daniel T. Gillespie, Exact Stochastic Simulation of Coupled Chemical Reactions,The Journal of Physical Chemistry, Vol. 8 1, No. 25, 1977

[56] Gibson MA, Bruck J. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. J. Phys. Chem. A 2000, 104, 1876-1889

[57] Knuth D. The art of computer programming vol 1. Fundamental Algorithms, Third Edition. Addison-Wesley, 1997

[58] Li H, Petzold L. Logarithmic Direct Method for Discrete Stochastic Simulation of Chemically Reacting Systems, 2006

[59] Cai X, Exact stochastic simulation of coupled chemical reactions with delays, J. Chem. Phys. 126, 124108 (2007)

[60] Mahdavi A, Davey RE, Bhola P, Yin T, Zandstra PW. Sensitivity analysis of intracellular signaling pathway kinetics predicts targets for stem cell fate control. PLoS Comput Biol. 2007;3(7):e130. doi:10.1371/journal.pcbi.0030130

[61] Stuart HT, van Oosten AL, Radzisheuskaya A, Martello G, Miller A, Dietmann S *et al.*. NANOG Amplifies STAT3 Activation and They Synergistically Induce the Naive Pluripotent Program. Current Biology 2014 Feb 3;24:340-346.