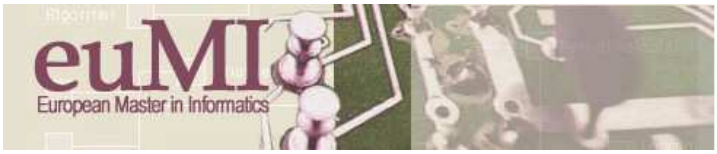

UNIVERSITÀ DEGLI STUDI DI TRENTO

Facoltà di Scienze Matematiche, Fisiche e Naturali

Corso di Laurea Specialistica in Informatica
within European Master in Informatics

Final Thesis

Sensitivity Analysis of Models of Biochemical Reactions

Relatore/1st Reader:

Dr. Stephen Gilmore
University of Edinburgh

Corelatore/ 2nd Reader:

Prof. Paola Quaglia
Università degli Studi di Trento

Laureando/Student:

Andrea Degasperi

Anno Accademico 2006 - 2007

Abstract

In the last decade, models of biochemical interactions, able to describe the time evolution of biomolecular systems, have been the subject of increasing interest. The use of such models has helped to understand the dynamics of biological processes and to accelerate drug discovery, moving part of the experiments from the laboratory to the computer. But models are nothing but a representation, an approximation of the real biological system, therefore they require to be analysed and validated. The analysis of the models *in silico* may lead to the discovery of properties that can be tested and validated in *in vitro* experiments.

When modeling, one of the main issues is the tuning and the correct identification of the parameters of the model. There are factors in a biochemical model, such as kinetic constants and initial concentration of the species involved, that require a specific investigation, in order to assert their role within the system.

Sensitivity Analysis (SA) provides the techniques that can be used to identify those parameters that are the most influential to the outcome of the model. In this dissertation we present a survey about the current SA techniques used to analyse models of biochemical reactions and we introduce new methods that better adapt to the analysis of models that present bifurcation points and bistability. Following the lead of earlier related work, we take in account three different categories of SA (local, global and screening methods) and analyse the aspects and purposes of each of them. The focus is mainly on the differences between the classical approaches, based on time evolutions of the systems computed with *Ordinary Differential Equations*, and the novel techniques, based on time evolutions computed with multiple runs of the *Stochastic Simulation Algorithm*. An example of sensitivity analysis of a real pathway, the MAPK, is then provided.

Finally, we show how these techniques have been implemented in Dizzy, a software that provides tools for the simulation and analysis of models of biochemical reactions.

Acknowledgements

I would like to thank both the Institutions of the University of Trento and of Edinburgh and the people that work for making the double degree of the European Master possible, allowing me and other people like me to live this important and invaluable experience. Then I would like to thank Prof. Stephen Gilmore, for his guide and patience during the writing of this dissertation. Finally, a special thank goes to my family, in particular my parents, that always supported my decisions and that made me feel that I could accomplish everything I wanted.

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Andrea Degasperi*)

Table of Contents

Chapter 1

Introduction

In recent years, Systems Biology has received an increasing amount of attention from different scientific fields, such as Mathematics, Statistics and Computer Science. This has happened and is still happening, because the data collected in the experiments of these last productive years is giving to biologists a new kind of challenge. Discovering new genes or new pathways, correlations between genes, is something that can nowadays be done or suggested by the use of statistical techniques or algorithms that require multidisciplinary competencies. Moreover, experiments *in silico* are becoming more and more popular, particularly to model complex systems, when supported with adequate theoretical foundations and/or mathematical proofs.

As suggested in [23], there are two main disciplines that partially counterpose themselves. *Bioinformatics*, that identifies, catalogues and characterizes the components of a cell, and *Systems Biology*, that studies how the components work and behave together through time.

Moreover, Computational Systems Biology [12] provides tools and techniques that can be used to implement an *in silico* representation of the intracellular dynamics which are the subject of the study. The resulting model will then be used to improve the understanding of the system, simulating possible variants or perturbations and validating them with further experiments. The idea is that the model should suggest new experiments and that the experiments should improve the model.

The most widely used techniques used for this purpose are reaction-based biochemical models. These kinds of models are defined by a set of reactions along with the rate at which these reactions may occur. This information, along with the initial concentration or number of molecules of the species involved, is enough to describe the time evolution of the modelled system. The two methods that often present themselves as

a choice are the *General Mass Action* (GMA) [22], and the *Stochastic Simulation Algorithm* of Gillespie [8, 9, 7]. The former is computationally cheaper while the latter is considered more descriptive and able, in some cases, to capture more complex dynamics. However, it is difficult to identify *a priori* which method would be the most appropriate in a specific case [22].

Once a model is built, the next step is to identify its properties and to verify that these properties belong also to the original real system.

1.1 Why sensitivity analysis

Modeling biochemical reactions requires a large number of parameters that are usually difficult to infer from experimental data. The parameters we are mainly interested in obtaining are initial concentrations and kinetics values. The latter in particular are computed using statistical techniques or found by just randomly trying to reproduce the results obtained in *in vitro* experiments. This leads to uncertainty about the true value of these parameters.

Sensitivity analysis (SA) allows us to see how much a parameter influences the model. If the model is robust with respect to some parameters, we can be more relaxed about their real value, because they will not greatly affect any further analysis. It is also important to quantify this influence, in order to understand whether one parameter is more important than others. If a parameter expresses strong influence on the model results, it is then recommended to identify more precisely or in a more reliable way its real value. SA has shown also to be useful to identify properties and reinforce the understanding of the analysed system [19].

Sensitivity analysis tools for ordinary differential equations models (ODE models) have already been developed and widely used. Usually they provide global SA techniques that can be directly applied to ODE models, considering the steady-state as the output to analyse. An example is the Systems Biology Toolbox [3] (SBToolbox) for MATLAB [1] that includes steady-state parameter SA and period and amplitude parameter SA. Conversely, our focus will be on SA of stochastic models, in particular of the *Chemical Master Equation* (CME) of Gillespie [8, 9, 7] and its Monte Carlo Simulation *Stochastic Simulation Algorithm* (SSA). Some techniques aiming in this direction have been proposed only recently [11], thus this is still a branch of SA that leaves room for new ideas. In particular, new sensitivity measures that better adapt to the highly stochastic nature of certain biological models can be proposed and tested.

In this dissertation we take a measure that has been already presented, the *histogram distance* [5], and we use it as a building block for novel SA methods. The histogram distance is an approximate way to compute the difference between two sets of stochastic simulations. Therefore, the sensitivity measure that makes use of it, can be considered taking into account an approximation of the CME underlying the values of the stochastic simulations.

Finally, the possibility to immediately implement the theory in Dizzy, a software tool for the simulation and analysis of biochemical models [17], makes this work usable by modellers in the short term.

1.2 Content of the dissertation

In the next chapter we begin with the background needed in order to understand the work. This includes some concepts of biology, modeling biochemical models, a short presentation of Dizzy and a survey of the current SA techniques, somehow related or applicable to the analysis of this kind of models. In Chapter 3 we introduce the novel approaches, comparing them to the classical procedures. In Chapter 4 the application of the discussed techniques to a real example is presented. Moreover, in Appendix A we show how this work has been implemented in Dizzy.

Chapter 2

Background

2.1 Basic biological concepts

This dissertation concerns models of biochemical interactions and we will later discuss some real biological examples. Therefore, we wish our reader to be familiar with concepts such as enzyme, signaling pathway or phosphorylation. This first section is indeed a brief guide to those biological concepts that the reader will meet in later chapters.

Moreover, it is important to understand the characteristics, the components and the properties of the systems we wish to model. Our references for this section are [16] and [13].

2.1.1 Forces guiding molecular interactions

In molecular biology there are many kinds of forces that influence the interactions between molecules or atoms, causing, for example, atoms to form solid structures or molecules join with a temporary weak bond. The ones that for us are of main interest are covalent bonds, ionic bonds, hydrogen bonds, hydrophobic interactions and Van der Waals forces.

Covalent bonds: this is a strong bond between atoms, where they share electrons that are in their outermost shell. Atoms tend to establish covalent bonds in order to reach an energetic stability. When they are linked together with covalent bonds they are called molecules. This is the strongest kind of bonds, meaning that it requires a relative large amount of energy to be broken. It can be double or triple if the number of shared electrons is more than one, a situation that makes the bound even more solid.

Moreover, if the shared electrons spend more time in one region of the molecule, the bond is said to be *polar*. An example of a polar covalent bond is the molecule of water, H_2O, that is negatively charged close to the atom of oxygen and positively charged close to the atoms of hydrogen.

Ionic bonds: when an atom, instead of sharing, loses or gains an electron, it becomes an *ion*, i.e. an electron that is positively or negatively charged. Ions of opposite charge may be attracted to one another leading to the formation of electrostatic bonds. Although these bonds are not as strong as covalent bonds, they still require a considerable amount of energy to be broken. An example is the interaction between sodium (Na) and chlorine (Cl): an electron is transferred from the former to the latter generating the ions Na^+ and Cl^- that are holden together by the ionic bond.

Hydrogen bonds: hydrogen atoms with a positive partial charge that are bound to oxygen or nitrogen (as in H_2O or NH_3) are able to interact with free electron pairs of atoms with a negative partial charge. These are charge based interactions that are relatively weak and that usually can be broken by just raising the temperature. Hydrogen bonds are present in many biological structures, such as DNA (deoxyribonucleic acid) or folded proteins, and biological processes, such as enzyme activity.

Hydrophobic interactions: water is electrically polarized and is able to form hydrogen bonds, so hydrophobic molecules are usually non polar molecules that are repelled by H_2O that bonds together. An example of these interactions is the cellular membrane that is formed by molecules that are made of a hydrophobic and a hydrophilic part.

Van der Waals forces: these are forces generated by temporary difference of charge within a molecule, caused by momentary inequalities in the distribution of electrons in a covalent bound. That molecule has a dipole induced. This bipolarity may also induce dipoles in close molecules. These forces can be either attractive or repulsive and they may generate very weak bonds, even weaker than hydrogen bonds. Van der Waals repulsions have an important influence on the possible conformations of a molecule.

2.1.2 Proteins

We cannot talk about proteins without briefly introducing the *central dogma* of molecular biology. The DNA (deoxyribonucleic acid) is that molecule that in every cell contains the information about how to construct and synthesize all those buildings blocks

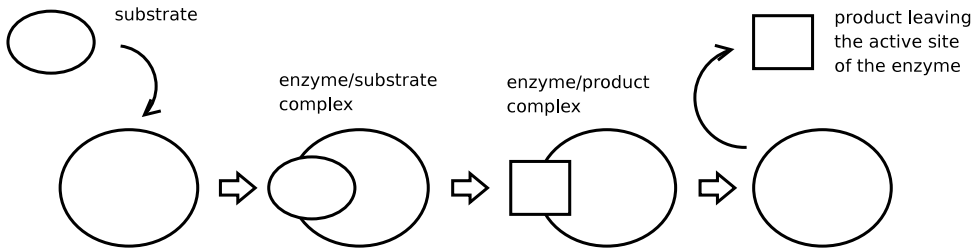

Figure 2.1: Schematic representation of an enzymatic reaction.

that let the cell function properly, the proteins. This information is coded as a sequence of bases: adenine, thymine, guanine and cytosine. When a protein needs to be built, the process of transcription copies the necessary part of DNA into a strand of RNA (ribonucleic acid). Then, in the process of translation, the RNA binds to a molecule called ribosome and the information present in the RNA is used to construct a chain of amino acids, linked together through covalent bonds. After that, this chain folds, thanks to bonds and forces acting on it, leading to the final shape of the protein.

Proteins fulfill numerous important functions in the cell, from being just part of the cellular structure to having roles in the metabolism of the cell or in the delivery of signals. The main characteristic of proteins, that enables them to have so many different functions, is their ability to bind other molecules specifically and tightly. The regions in the protein where other molecules may bind are called *binding sites*. These regions are defined by their shape and by the chemical properties that surround them, allowing only very specific molecules to bind. Proteins can also bind to other proteins or be integrated into membranes. When a protein binds to another molecule, it can also change some of its properties and abilities to bind.

Enzymes. An enzyme is a protein whose role is to catalyze, i.e. to accelerate, a biochemical reaction. Enzymes allow reactions that are normally unfavorable in nature to take place, lowering their activation energy. We will call *substrates* the molecules that take part of catalyzed reactions and *products* the molecules that are generated. Usually enzymes are very specific as to which reactions they catalyze. Complementary shape, charge and hydrophilic/hydrophobic characteristics of enzymes and substrates are responsible for this specificity. Figure 2.1 sketches a possible dynamics of enzymatic activity.

Phosphorylation. Phosphorylation is a biochemical reaction in which a phosphate group links through a covalent bond to a molecule. This usually increases the energy of the resulting molecule, that changes some of its properties and may become active.

Many proteins or enzymes, for example in signaling pathways, have a poor activity until they are phosphorylated and begin to perform important functions. Two categories of enzymes regulate the phosphorylation of a molecule: the *kinase* that adds a phosphate group and the *phosphatase*, that removes it (dephosphorylation). Usually, this kind of reactions depends on the presence of ATP (adenosine triphosphate), that can be considered one of the most diffuse kinds of energy storage in the cell.

2.1.3 Metabolic and signaling pathways

The metabolism of a cell is a highly organized process, that involves thousands of reactions that are catalyzed by enzymes and whose ultimate goal is to provide everything the cell needs to survive and reproduce. Metabolism provides energy and material for building and maintaining the cell. So, metabolic pathways are networks of biochemical interactions that provide mainly mass and energy transfer.

A signaling pathway, instead, is a sequence of biochemical interactions that leads to the transmission of external signals from outside to inside the cell and to the movement of information inside the cell. Examples of signals are hormones, pheromones, heat, cold, light or even the appearance or concentration change of substances such as glucose, K^+ or Ca^+. The interpretation of these external signals triggers the cell response.

Other differences are that metabolic pathways can be determined given the enzymes involved, while it is difficult to have a precise idea of the organization of a signaling pathway, because it is in general complex and it may assemble dynamically depending on the signal. Moreover, in metabolism the enzyme concentration is in general much lower than the substrate concentration, while in signaling processes they are frequently in the same order of magnitude.

2.2 Modelling biochemical interactions

Metabolism or signaling networks can be modelled and studied on different levels of abstraction. It is possible to concentrate on the properties of the individual reactions as well as studying the system as a whole, perturbing it and observing the change in behaviour. In this section we will describe techniques used to model a biological system at a reaction level. This means that, usually, the knowledge of the reactions, the species involved and the rate at which these reactions occur is enough to describe,

and so to model, the time evolution of the system.

2.2.1 The Law of Mass Action

The law of mass action, introduced in the 19th century [10], states that the reaction rate is proportional to the probability of a collision of the reactants. This probability is then proportional to the concentration of the reactants and the molecularity, i.e. how many copies of a reactants are involved in the reaction.

For example, the rate of the reaction

$$S_1 + S_2 \rightleftarrows 2P$$

can be formulated as

$$v = v_+ - v_- = k_+ \cdot [S_1] \cdot [S_2] - k_- \cdot [P]^2,$$

where v is the rate, v_+ is the rate of only the forward reaction, v_- is the rate of the backward reaction and k_+ and k_- are the proportionality factors, called *kinetics* or *rate constant*. The symbol $[\cdot]$ denotes the concentration of the species expressed in moles per liter (mol $\cdot$ L^{-1}). The dynamics of the concentrations of the species can be described by the *Ordinary Differential Equations* (ODEs)

$$\begin{aligned} \frac{d}{dt}[S_1] &= \frac{d}{dt}[S_2] = -v \\ \frac{d}{dt}[P] &= 2v. \end{aligned}$$

The value of the concentrations of S_1, S_2 and P through time are obtained by integration of these ODEs.

2.2.2 Generalized Mass Action

In this section we generalize and formalize the concepts which we introduced in the previous section.

Modelling intracellular dynamics in a quantitative way, is concerned with the estimation through time of the concentration or of the number of molecules of N different species S_i which can interact using M possible biochemical reactions R_μ

$$R_\mu : l_{\mu 1} S_{p(\mu,1)} + l_{\mu 2} S_{p(\mu,2)} + ... + l_{\mu L_\mu} S_{p(\mu,L_\mu)} \xrightarrow{k_\mu} ...$$

where L_μ is the number of reactant species involved in R_μ, $l_{\mu j}$ is the stoichiometric coefficient of the reactant species $S_{p(\mu,j)}$, K_μ=$\sum_{j=1}^{L_\mu} l_{\mu j}$ denotes the molecularity of the reaction R_μ and the index $p(\mu, j)$ selects those S_i participating in R_μ.

Assuming a constant temperature and that diffusion in the cell is fast, such that we can assume a homogeneously distributed mixture in a fixed volume V, the *General Mass Action* (GMA) model of the system can be defined by N ordinary differential equations (ODE) as follows:

$$\frac{\delta}{\delta t}[S_i] = \sum_{\mu=1}^{M} \nu_{\mu i} k_\mu \prod_{j=1}^{L_\mu} [S_{p(\mu,j)}]^{l_{\mu j}} \quad i = 1, 2, ..., N \tag{2.1}$$

where the k_μ's are rate constants and ν_μ denotes the change in molecules of S_i resulting from a single R_μ reaction. $[S_i]$ is the concentration of the species S_i and its unit is usually mol per liter, mol/L. As described in [22], the mathematical representation 2.1 of a biochemical network does not account for noise on the states, neither does it consider measurement noise. For this reason, GMA is often called a deterministic approach. The $[S_i]$ are, however, the most probable value.

There are however some situations in which a GMA model is not able to represent all the characteristics of a system. The differential equation formalism implicitly assumes a continuous variation in concentration of reactants and a deterministic dynamics, assumptions that cannot be valid with very low concentrations (such as regulatory molecules).

2.2.3 Chemical Master Equation

When the number of molecules of a species reduces to few units, another method is often considered more suitable. This is a *stochastic approach* where we wish to determine for each molecular species S_i the probability $P(\#S_i(t) = n_i)$ that at time t there are n_i molecules (with $\#S_i$ denoting the number of molecules of the species S_i). For N molecular species, let $n \in \mathbb{N}^N$ denote the N dimensional state vector. $\nu_\mu \in \mathbb{Z}^N$ are the step changes occurring for elementary reactions indexed by μ. If S is an N dimensional variable, we write $P(\#S = n)$=$P_n(t)$. Describing the changes in random variable S, we consider the following two state transitions:

$$n - \nu_\mu \xrightarrow{a_\mu(n-\nu_\mu)} n$$
$$n \xrightarrow{a_\mu(n)} n + \nu_\mu.$$

The first denotes a transition from another state to the state n; the second denotes moving away from the state n. Most important, $a_\mu(n-\nu_\mu)$ is referred to as the *propensity* function of the reaction R_μ, that is the probability per unit time, of a change ν_μ occurring, given that we are in the state $n-\nu_\mu$.

With these definitions we can define the *Chemical Master Equation* (CME) [8, 9, 7]:

$$\frac{\delta P_n(t)}{\delta t} = \sum_{\mu=1}^{M} [a_\mu(n-\nu_\mu)P_{(n-\nu_\mu)}(t) - a_\mu(n)P_n(t)]. \tag{2.2}$$

This equation describes the probabilities of moving in or out the state n. For each state n we have then a differential-difference equation of this kind. This equation has been derived using physical assumptions about the probability that the single molecules have to collide and therefore react. In particular Gillespie derived the parameter $c_\mu dt$, the average probability that a particular combination of R_μ reactants molecules will react accordingly in the next infinitesimal time interval dt. In fact, the propensity function $a_\mu(n)$ is the product of c_μ and $h_\mu(n)$, the number of distinct combinations of R_μ reactant molecules. $c_\mu dt$ is called the *stochastic rate constant*.

It is interesting to remark that it has been proved that there is a correspondence between c_μ and the GMA rate constant k_μ [22]:

$$c_\mu = \left(\frac{k_\mu}{(N_A V)^{K_\mu - 1}}\right) \cdot \prod_{j=1}^{L_\mu} (l_{\mu j}!) \tag{2.3}$$

where N_A is the Avogadro number and V is the cell volume. This allows to pass from one method to the other as soon as either c_μ or k_μ has been identified from experimental data.

2.2.4 Stochastic Simulation Algorithm

A major difficulty with the CME is that its analytical solution is usually intractable. For this reason, Gillespie developed the *Stochastic Simulation Algorithm* (SSA), a Monte

Carlo simulation of the CME. A single simulation represents one exact possible evolution of the system, while a set of thousands of these simulations can be used to identify an underlying probability function that is an approximation of the CME.

This algorithm proceeds with a loop in which, at every iteration, two parameters are randomly taken from previously defined probability distributions: the time of the next reaction and which reaction will occur next. In order to compute these values, the joint probability that reaction R_μ will be the next reaction and will occur in the infinitesimal time interval $[t, t+\delta t)$, given $(\#S = n)$, is computed:

$$P(\tau, \mu | n, t) = a_\mu(n) e^{-a_0(n)\tau} \tag{2.4}$$

where $a_0(n)$=$\sum_{\mu=1}^{M} a_\mu(n)$.

Starting from 2.4, the probabilities of the next reaction and the time of the next reaction can be obtained:

$$P(\tau | n, t) = a_0(n) e^{-a_0(n)\tau} \quad \tau \geq 0$$
$$P(j | \tau, n, t) = \frac{a_\mu(n)}{a_0(n)} \quad \mu = 1, ..., M$$

From these distributions, random Monte Carlo samples can be taken using two uniform random numbers r_1 and r_2 from $[0,1]$. τ is given by:

$$\tau = \frac{1}{a_0(n)} ln\left(\frac{1}{r_1}\right) \tag{2.5}$$

The index μ of the selected reaction is the smallest integer in $[1, M]$ such that

$$\sum_{\mu'=1}^{\mu} a_{\mu'}(n) > r_2 a_0(n) \tag{2.6}$$

Once these two values are computed, the system is updated adding the selected ν_μ to n and τ summed to t.

2.2.5 Logarithmic Direct Method

During the dissertation we will face the problem of computational complexity, even for very simple models. In order to save computational time, a faster, although still

exact, version of SSA has been adopted. In general, we will use the original SSA and, when specified, the *Logarithmic Direct Method* (LDM) [14].

Locating the next reaction to fire is the computationally most expensive step of the SSA. In LDM, this step can be reduced from $O(M)$ to $O(logM)$. In the original method, a_0 is computed and in a second time the next reaction is determined summing one by one the a_μ, so this sum is computed almost twice. In LDM, while computing a_0, a list is kept, with all the partial sums, from the smallest to the largest, ; then the next reaction is identified performing a binary search on the list. This way, the value in the list such that (2.6) is satisfied can be found in $O(logM)$.

2.2.6 Michaelis-Menten kinetics

We introduce now a model of enzymatic reactions that is well established in the field of systems biology [4, 13]. This applies to the following system:

$$E + S \overset{k_1,k_{-1}}{\longleftrightarrow} ES \xrightarrow{k_2} E + P$$

where E is the enzyme, S is the substrate, ES is the temporary enzyme-substrate complex and P is the product of the reaction. Characteristics of this model are that the process is considered irreversible, i.e. the product cannot become a substrate, and that the enzyme is not affected by the reactions and can be used again after it leaves the substrate or the product. Under determined assumptions, the parameters k_i can be identified with few simple experiments.

These are the ODEs of the model, following the GMA:

$$\begin{aligned}
\frac{d[S]}{dt} &= -k_1 \cdot [E] \cdot [S] + k_{-1} \cdot [ES] \\
\frac{d[ES]}{dt} &= k_1 \cdot [E] \cdot [S] - (k_{-1} + k_2) \cdot [ES] \\
\frac{d[E]}{dt} &= -k_1 \cdot [E] \cdot [S] + (k_{-1} + k_2) \cdot [ES] \\
\frac{d[P]}{dt} &= k_2 \cdot [ES]
\end{aligned}$$

This system of ODEs can be simplified using further assumptions. One of these is that we consider the conversion of E and S into ES and vice versa much faster than the decomposition of ES into E and P ($k_1, k_{-1} \gg k_2$, the *quasi equilibrium* assumption). The other assumption is that during the course of the reactions a state is reached where

the concentration of ES remains constant. This is called the *quasi steady-state* assumption, due to the fact that we consider the concentrations of the intermediates (ES) to reach equilibrium much faster than those of the product and substrate. This means that we assume:

$$\frac{d[ES]}{dt} = 0$$

That leads to the simplification:

$$[ES] = \frac{k_1[E][S]}{k_{-1}+k_2} = \frac{[E][S]}{K_m},$$

where $K_m = (k_{-1}+k_2)/k_1$ is called the *Michaelis constant*. Given the quasi equilibrium assumption we have that $K_m \cong k_{-1}/k_1$. Then, noticing that the total concentration of enzyme $[E_{tot}]$ is equal in every moment to the sum of the concentrations of the free enzyme $[E]$ and of the enzyme-substrate complex $[ES]$,

$$[ES] = \frac{([E_{tot}]-[ES])[S]}{K_m} \Longrightarrow [ES]\frac{K_m}{[S]} = [E_{tot}]-[ES] \Longrightarrow$$
$$[ES]\left(1+\frac{K_m}{[S]}\right) = [E_{tot}] \Longrightarrow [ES] = [E_{tot}]\frac{1}{1+\frac{K_m}{[S]}} = [E_{tot}]\frac{[S]}{[S]+K_m}.$$

At this point we obtain that the velocity of the production of P is given by

$$\frac{d[P]}{dt} = k_2[ES] = k_2[E_{tot}]\frac{[S]}{[S]+K_m} = \frac{V_{max}[S]}{[S]+K_m}$$

where V_{max} (also written k^{cat}) is the maximum velocity of the production of P, given by $k_2[E_{tot}]$. The parameters V_{max} and K_m can be easily estimated with few experiments.

2.3 Sensitivity analysis of biochemical models

Sensitivity Analysis (SA) studies the relationships between the inputs and the outputs of model, often regardless of the particular model which is the subject of the analysis. We explain now what we mean when we talk about SA applied to reaction based biochemical models.

Figure 2.2 shows an example of the time evolution of a biochemical system with four species. These species have the role of the output variables of the model. When

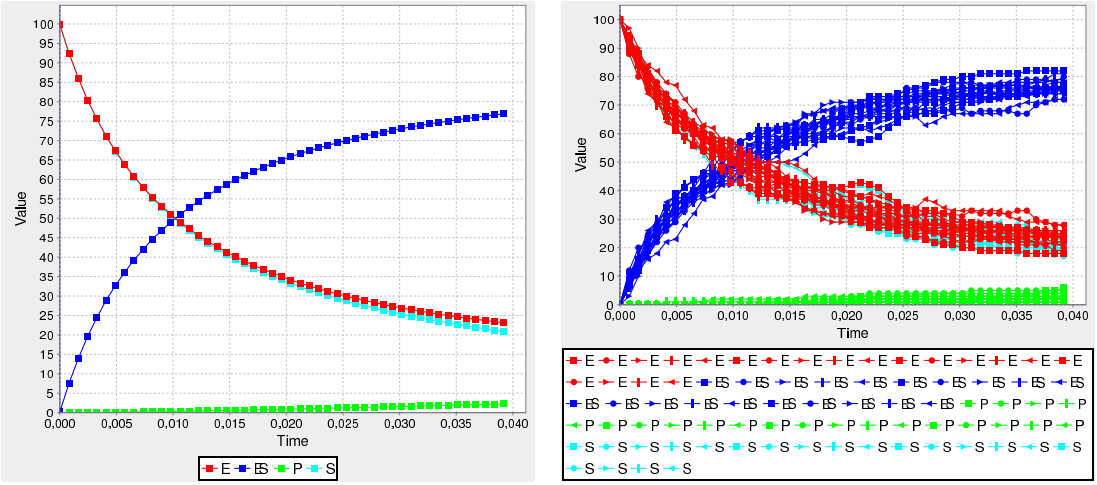

Figure 2.2: Examples of time evolution of a biochemical model computed with ODE (on the left) and with SSA (on the right). This model is the Michaelis/Menten model of the enzymatic reaction which will be presented in Section 3.2.2.

we wish to perform a sensitivity analysis of this model, we need to choose a time point at which to read the output value. In the case of an ODE model, a selected output (species) at a selected time assumes a precise and unique value. Changing one or more parameters of the model may lead to a different output value. On the other hand, in the case of stochastic simulations, the output of a selected species at a selected time can be considered to be the collection of the outputs given by the simulations. This set of values, if it has enough points, will at this point reveal the underlying distribution of the output.

One of the basic operations that can be done during a sensitivity analysis, is to compute the difference between the output of a model and the output of the same model where one or more parameters have been perturbed. While this is possible in ODE models, it is not so clear when facing stochastic simulations. One simple solution is to take as output the mean of the values coming from the simulations. However, this solution can lead to a loss of information: taking the mean we are assuming a normal distribution and we are even neglecting the variance. Another possibility is the use of a distribution distance or histogram distance [5], that, with a sufficient number of simulations, is able to identify more precisely the difference in the outcome of the model (see Section 2.3.1).

2.3.1 Histogram distance

The use of histogram distance in analysing the outcome of reaction based models has been introduced with the purpose of quantifying how well an approximate version of the SSA is able to emulate the original [5]. In the context of sensitivity analysis, it can instead be used to quantify the output difference of a model produced by the perturbation of its parameters.

It is computed as follows:

$$D_k(X,Y) = \sum_{i=1}^{k} \left| \frac{\sum_{j=1}^{N} \chi(x_j, I_i)}{N} - \frac{\sum_{j=1}^{M} \chi(y_j, I_i)}{M} \right|. \tag{2.7}$$

where X and Y are two sets of numbers, k is the number of histogram columns or intervals that divide the range of the output variable, N is the cardinality of the set X and M is the cardinality of the set Y, x_j and y_j are elements of the sets X and Y respectively and the function χ returns 1 if the element x_j belongs to the interval I_i, 0 otherwise. I_i is the i-th interval in the range, that goes from $x_{min} + \frac{(i-1)L}{k}$ to $x_{min} + \frac{iL}{k}$, where $L = x_{max} - x_{min}$.

An interesting measure is then the *self distance*, given by $D_k(X,X')$. This is nothing but repeating the same experiment twice, with the same parameters, and then computing the histogram distance between the two sets of number resulting from the simulations. Perturbations in the parameters that generate values of distances that are less than or very close to the self distance will be considered not to have an influence, or, at least, we can say that we cannot distinguish any effect arising from the perturbation.

2.3.2 Sensitivity analysis classification

According to [20], sensitivity analysis (SA) techniques can be classified as:

- **local methods:** they concentrate the analysis around a particular point in the parameter space. For example, *local one at a time* and *elementary one at a time* approaches belong to this class;

- **global methods:** these techniques try to explore the entire space of the parameters or, at least, explore the subspace that is believed to contain the real value of the parameters and that represents their uncertainty. Examples of members of this class are *variance based* methods, but also OAT methods replicated randomly in the space.

- **screening methods:** these methods are used to select the most important parameters when the complexity of the model is problematic or the number of parameters intractable. The main idea of these methods is that they should be computationally cheap and give the idea of which parameters can be fixed (low importance), even if the information that can be achieved is poor. They are a tradeoff between information and algorithm complexity. Once the most influential parameters have been identified, it is then possible to apply a more informative and computationally expensive technique. An example of these techniques is the *Morris' method* [21, 20].

This classification is not rigid. For example, many screening methods can be considered global or local. Moreover, the last class is characterized by the goal of the analysis (screening) and the other two by how the analysis is performed and the factors treated.

The logic behind this classification follows a wider point of view. With the need to analyse a complex model, with hundreds of reactions and parameters, we would like to have a gradual approach that allows us to achieve rough information with cheap computational power at the beginning and very precise information with high computational effort only at the end. This way we can begin with a local analysis that gives us a preliminary feedback around which are believed to be the nominal values of the parameters. In a second phase we can proceed with a screening method, which allows us to fix some parameters leading to a dimensionality reduction of the model and the selection of a small set of parameters that can be considered dominant. The final step would be the application of a global and most informative technique only on the selected parameters, in order to limit the computational time needed to perform the analysis.

2.3.3 One at a time methods

The classical and most widely used SA is the *One at a time* (OAT) approach: a parameter is perturbed (usually by 1%) and the changes in the output measured. Alternatively it is possible to compute the derivative of the output with respect to each parameter to obtain its sensitivity coefficient:

$$S_{ij} = \frac{\delta y_j(\mathbf{p})}{\delta p_i}$$

where $y_j(\mathbf{p})$ is the j-th output of the model that depends on the parameters and p_i is the i-th parameter.

It is critical to understand exactly what is meant by each kind of sensitivity measure. For example if we chose as sensitivity measure the simple output difference (*Elementary* OAT, EOAT) we can state that varying a variable x_i by 1% of its value influences the output more than varying another variable x_j by 1% of its value. This can be useful when a different order of magnitude of the parameters is involved. The derivative based measure allows a slightly different statement: varying a variable x_i by a fixed value influences the output more than varying another variable x_j by the same fixed value. These are very general techniques and one should always be careful about the answer one is looking for when perturbing a specific model.

In the study of biochemical systems, these methods represent the prevalent practice when analyzing ODE models. Other more complex and informative analysis has been proposed [21]. However, all these analyses are not directly applicable to stochastic models, such as CME [9], whose output is defined as a *probability density function* (pdf) over the number of molecules for each species. The need to consider the entire pdf as the output to analyze is even more evident in the analysis of bistable systems. These present at a certain time a pdf that is not normal, but that presents instead two distinct peaks of likelihood (an example is the Schlögl model [11, 5]). In this particular context an analysis cannot make any assumptions on the pdf resulting from the model. In this regard, SA of stochastic systems has been recently introduced [11]. Here, the change in the output value is quantified in terms of distribution distance, or more precisely, *histogram distance* [5], computed by comparing approximations of the pdfs constructed from several runs of the SSA (see Section 2.3.1).

$$S_i = D(X_n, X_{p_i})$$

where X_n is a *random variable* (r.v.) with nominal pdf = $f(\mathbf{x}, \mathbf{p})$ and X_{p_i} is a r.v. with perturbed pdf = $f(\mathbf{x}, p_1, ..., p_i + \Delta p_i, ..., p_k)$. Also this distance can be divided by Δp_i, leading to a correspondent derivative based approach.

Together, these approaches can be classified as *Local* OAT Sensitivity Analysis and they hold only if the model is linear with respect to the parameters. If we can assume that varying only one parameter at a time affects the output of the model in a proportional way, then these techniques are enough to quantify the input/output influence. However, often, this assumption is not valid for biological systems, a fact that makes

Local OAT not effective enough to give a complete view of the relationships between parameters and output and also between the parameters themselves. They can in any case be considered useful mainly because they are computationally cheap, a key feature when dealing with thousands of stochastic simulations, and because they can give a first idea of the sensitivity indices.

2.3.4 Morris' method

This method [20] can be classified as one at a time (OAT), because it uses as a basic step the local OAT approach, and global, because the experiment covers the entire space over which the factors are believed to vary. Morris estimates the main effect of a factor by computing a number r of local measures, at different random points $\mathbf{x}_1, ..., \mathbf{x}_r$ in the parameter space, and then taking their average. This reduces the dependence on the specific point that a local experiment has.

When applying this method, a computationally expensive model is assumed, or a model with a large number of factors. The goal is to determine which factors have (a) negligible effects, (b) linear and additive effects, or (c) non linear interaction effects. This will help to apply later the most appropriate global sensitivity analysis only on the relevant parameters.

The k-dimensional factor vector $\mathbf{x}$ has components x_i that have p values in the set $\{0,\ 1/(p-1),\ 2/(p-1), ...,\ 1\}$. The region of experimentation Ω is then a k-dimensional p-level *grid* (Figure 2.3). In practice, the values sampled in Ω are then rescaled to generate the actual values of the parameters as sampled from a specific parameter range. Let Δ be a predetermined multiple of $1/(p-1)$. Then Morris defines the *elementary effect* of the ith factor at a given point $\mathbf{x}$ as:

$$d_i(\mathbf{x}) = \frac{y(x_1, ..., x_i + \Delta, ..., x_k) - y(\mathbf{x})}{\Delta}$$

where $\mathbf{x}$ is any value in Ω selected *such that the perturbed point* $\mathbf{x} + \Delta$ *is still in* Ω. After sampling r times, the result will be a distribution F_i of elementary effects. The characterization of this distribution through its mean μ and standard deviation σ gives useful information about the influence of the ith input on the output. A high mean indicates a factor with an important overall influence on the output; a high standard deviation indicates either a factor interacting with other factors or a factor whose effect is nonlinear. As it can be seen, the information acquired is not enough to be more

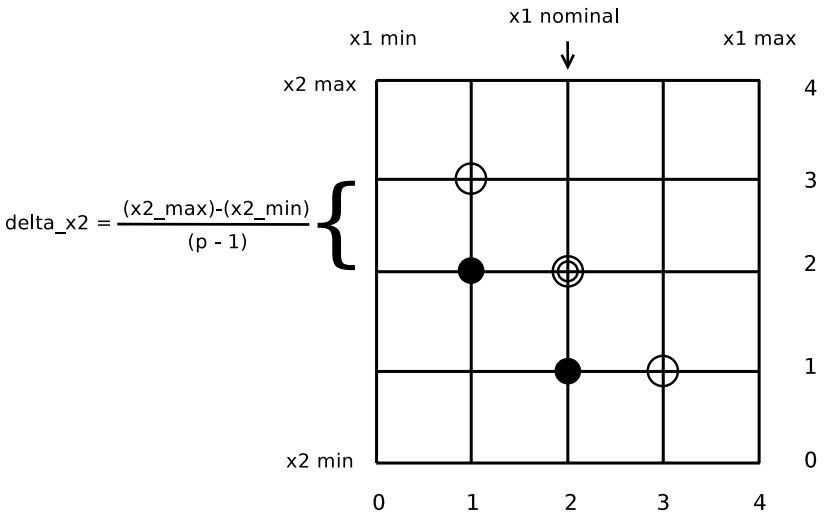

Figure 2.3: Example of a grid in the Morris method. In this case we have two parameters $(k = 2)$ and a grid level of five $(p = 5)$, so the maximum possible combinations are $5^2 = 25$. The black dots are two possible random points, while the circles are other points computed during the algorithm iterations. In this case, an implementation that wants to save computational time would not recompute the point that has been circled twice.

precise in the analysis, but is enough to suggest which parameters have little influence and can be fixed and which technique should be used in a further analysis. If an elementary effect has low mean and low standard deviation, it means that the output is not influenced by the corresponding parameter; it can then be fixed or at least neglected for purposes of dimensionality reduction. If an elementary effect has high mean but low standard deviation, it means that the output is influenced by the corresponding parameter and that the influence is always the same; this means that the model is linear with respect to that parameter and that this parameter is not influenced by others. Finally, if an elementary effect has high mean and high standard deviation, this would require a more complex analysis to understand if the model is simply nonlinear with respect to the corresponding parameter or if the influence of this parameter is controlled itself by the value of other parameters. In this regard it may well be that parameters that seem not influential at all with respect to the output of the model are instead relevant to the effect of other parameters.

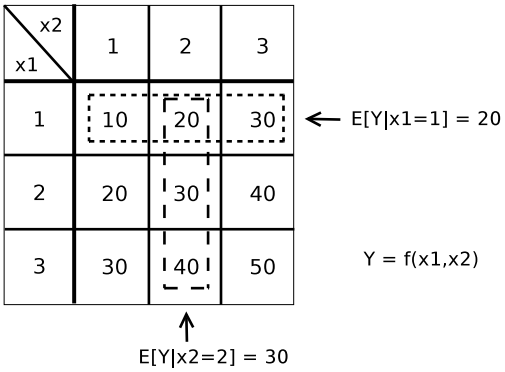

Figure 2.4: An example of a possible deterministic model $f(x_1, x_2) = Y$ that depends on the factors x_1 and x_2 is shown, along with examples of conditional expectations.

2.3.5 Variance-based methods

Variance-based methods are those sensitivity analysis (SA) techniques that use the *variance of the conditional expectation* (VCE) as a measure of importance of the input factors. The goal in these methods is to estimate the VCE by exploring the space made by all the possible values of the parameters. Applied to *Ordinary Differential Equations* (ODE) chemical models, the most famous techniques are *correlation ratio, Sobol', and Fourier amplitude sensitivity test* (FAST) [20, 21, 6].

Probability theory states that:

$$V[Y] = V_{\mathbf{x}}[E[Y|\mathbf{x}]] + E_{\mathbf{x}}[V[Y|\mathbf{x}]] \tag{2.8}$$

where the two components of the variance decomposition are called the variance of conditional expectation (VCE) and the residual part. The term $V_{\mathbf{x}}[E[Y|\mathbf{x}]]$ is the variance of the conditional expectation of Y, conditioned on $\mathbf{x}$, which is a suitable measure of the importance of $\mathbf{x}$, identifying the part of the variance of Y due to $\mathbf{x}$. If we had that all the variance of Y is matched by the VCE of $\mathbf{x}$ we could say that $\mathbf{x}$ is the only parameter (or set of parameters) that influences the outcome Y.

The variance of the conditional expectation is given by:

$$V_{\mathbf{x}}[E[Y|\mathbf{x}]] = \int (E[Y|\mathbf{x}] - E[Y])^2 p_{\mathbf{x}}(\mathbf{x}) d\mathbf{x} \tag{2.9}$$

where $E[Y|\mathbf{x}] = \int y p_{Y|\mathbf{x}}(y) dy$. Here the integral is substituted with the sum over all the possible values of $\mathbf{x}$ sampled from the range of $\mathbf{x}$.

In order to help to visualize this concept, a simple example of a deterministic model is shown in Figure 2.4.

So far, the simplest possible technique to compute that value has been implemented. Similarly to Figure 2.4, the parameter space has been sampled through the use of a grid. After having collected all the results, the conditional expectations are estimated fixing a parameter to its possible values in the grid. As can be expected, increasing the grid level and the number of parameters, the algorithm complexity increases exponentially. On the other hand, a complete analysis of the influence of the parameters on the output and on the other parameters is provided.

Let $S_{\mathbf{x}}$ be the *n*-th order sensitivity index, with $\mathbf{x} \in \mathbb{N}^n$. It corresponds to the VCE fixing the factors in **x** minus the sensitivity indices relative to all the possible combinations of the factors in **x**. For example, S_{12} is given by VCE_{12} - S_1 - S_2 and S_{123} is given by VCE_{123} - S_{12} - S_{13} - S_{23} - S_1 - S_2 - S_3. The VCE relative to **x**, where **x** contains all the factors, is nothing but $V[Y]$.

Following [6] the sensitivity measure that is the most suitable to determine the influence of a parameter on the output of the model is the *Total Sensitivity Index* (TSI) or simply TS_i. This is defined as the sum of all the sensitivity indices that contain *i* in **x**. For example, TS_1 is given by S_1 + S_{12}.

2.3.6 Sensitivity analysis of discrete stochastic systems

In a recent publication [11], new sensitivity measures that better adapt to the analysis of systems described by chemical master equations (CMEs) have been proposed. The authors have been maybe the first ones to highlight the need for a sensitivity measure that is specifically designed to consider not just a single value as an output, but the entire CME. The key idea is that the probability density function of the CME, that describes the probability $P_n(t)$ of the species of a system to be at a certain amount at time *t* (see Section 2.2.3), can be approximated by a *cumulative distribution function* f, obtained using stochastic simulation algorithm (SSA) realizations.

Four measures have been proposed. One is based on simple derivative of the CME:

$$S_j = \frac{\delta f}{\delta x_j} = \frac{f(\mathbf{n}, x_j + \Delta x_j) - f(\mathbf{n}, x_j - \Delta x_j)}{2\Delta x_j} \tag{2.10}$$

that is called centered difference approximation. This is a one-at-a-time approach that represents the influence of a parameter x_j on the probability that, at time t, the number of molecules of the species S_i is equal to n_i for all i.

The other three measures are based on the *Fisher Information Matrix* (FIM). Noticing that sensitivity indices such as:

$$S_j(\mathbf{n},t) = \frac{\delta \log f(\mathbf{n},t)}{\delta x_j}$$

are closely related to the score function in information theory, being the gradient of the log-likelihood function. The FIM

$$J = E[(\bigtriangledown_{\mathbf{n}} \log f)(\bigtriangledown_{\mathbf{n}} \log f)^T]$$

defines therefore the lower bound on the uncertainty in the parameter estimates according to the Cramer-Rao inequality

$$V_{\mathbf{x}} \geq J^{-1}.$$

The gradients are computed approximatively using equation 2.10, but another possibility is to assume that the density function f follows a multivariate gaussian distribution. The three sensitivity measures are the FIM diagonal elements, eigenvalues and the inverse of the diagonal elements of $V_{\mathbf{x}}$.

These methods have shown to have some improvements with respect to the classical methods, though they are of difficult interpretation. It is in particular not clear which is the information that is included in each sensitivity index.

Chapter 3

Sensitivity analysis of stochastic simulations of biochemical reactions

In this chapter we introduce two new sensitivity measures and we present them as variants of the Morris' method and the variance-based approach respectively. When doing this, we compare these new techniques with their original versions.

See Appendix A for the details about the software that has been used to compute the results and that has been extended with tools that implement the techniques introduced in this chapter. Every parameter that refers to a particular technique or algorithm can be used to reproduce the results using this software tool. From now on, when we refer to results obtained with ODE or deterministic methods, we implicitly intend that they are obtained using the 5/4 Dormand-Prince ODE solver with adaptive step-size. Moreover, when we refer to results obtained with stochastic simulations, we implicitly intend that we used the original SSA, if not otherwise stated.

3.1 Local methods

We procede now with the analysis and the comparision of the Local one-at-a-time methods presented in Section 2.3.3.

The analysis of these basic approaches is necessary because they are the building blocks of any more complex technique. For this reason, before introducing the other methods which are the subject of this dissertation, we need to show the kind of results that can be achieved by these basic approaches.

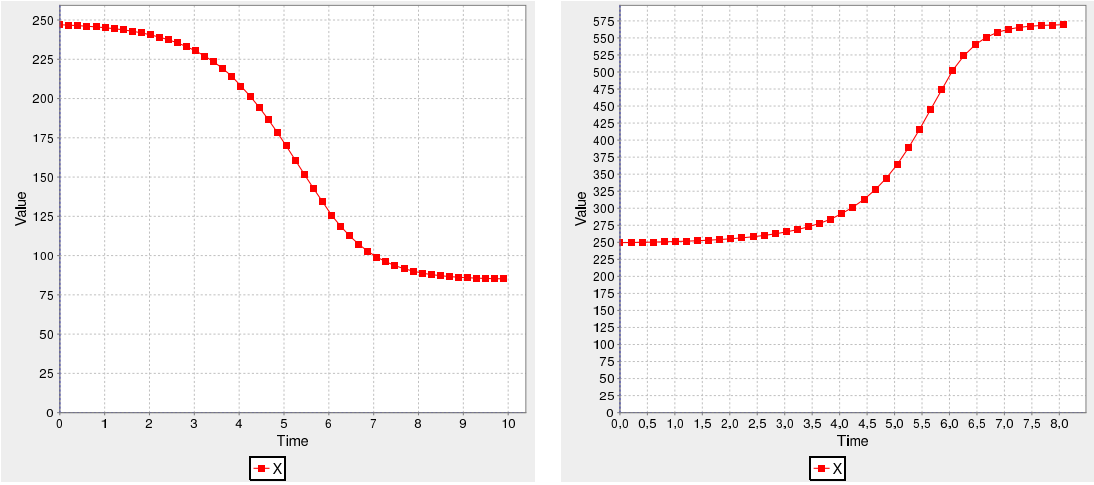

Figure 3.1: In the figure on the left, the time evolution of the output variable X of the Schlögl ODE model is shown. This is obtained with the nominal value of the parameters, as stated in the text. In the figure on the right, the time evolution of X changing only the initial number of molecules of X from 247 (nominal) to 250 is instead shown. The behaviour seems to completely change.

3.1.1 The Schlögl model

The prototype Schlögl model [11, 5] has been considered the most suitable model to show the differences between usual Local OAT approaches and the one based on histogram distance. It is defined as follows:

$$
\begin{aligned}
&A + 2X \xrightarrow{a_1} 3X,\\
&3X \xrightarrow{a_2} A + 2X,\\
&B \xrightarrow{a_3} X,\\
&X \xrightarrow{a_4} B,
\end{aligned}
$$

where A and B are kept constant (buffered). The propensity functions of Gillespie's CME (see Section 2.2.3) are:

$$
\begin{aligned}
a_1 &= k_1 A X (X-1)/2,\\
a_2 &= k_2 X (X-1)(X-2)/6,\\
a_3 &= k_3 B,\\
a_4 &= k_4 X,
\end{aligned}
$$

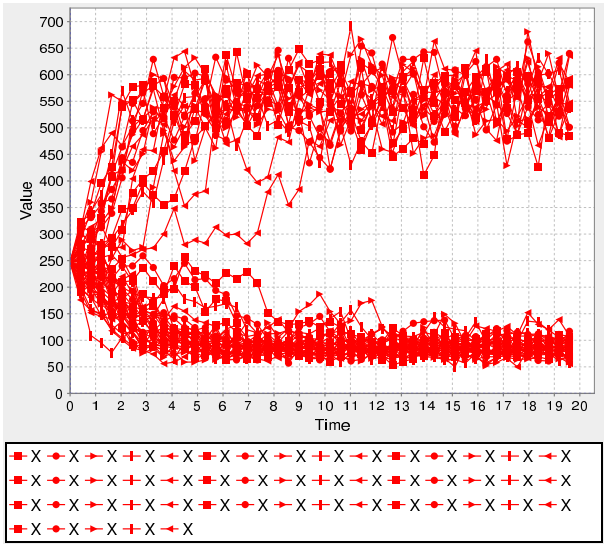

Figure 3.2: Evolution of the output variable X in 50 distinct runs of SSA of Gillespie on the Schlögl model. This shows the real behaviour of the system, in a more informative way with respect to the ODE model.

and the nominal values of the parameters are set as follows:

$$
\begin{aligned}
X_0 &= 247 \\
A &= 1 \cdot 10^5, \\
B &= 2 \cdot 10^5, \\
k_1 &= 3 \cdot 10^{-7}, \\
k_2 &= 1 \cdot 10^{-4}, \\
k_3 &= 1 \cdot 10^{-3}, \\
k_4 &= 3.5.
\end{aligned}
$$

The nominal parameter values are set close to a bifurcation point, where a small perturbation in them can lead to completely different results in the ODE time evolution, as can be seen in Figure 3.1.

With this particular set of parameters, also the time evolution of the stochastic simulations presents singular behaviours: the different runs will follow either one of two possible behaviours, as can be observed in Figure 3.2.

With the goal of describing the behaviour of this system, ODE models, or the simple average of X from different stochastic simulations could be inappropriate if not misleading. The use of estimated distributions can be considered a more suitable choice.

3.1.2 Local Sensitivity Analysis of the Schlögl model

Three LOAT Sensitivity Analysis have been applied to the Schlögl model: *Local OAT (ODE)*, *Local OAT (Gillespie average)* and *Local OAT (Gillespie Density)*. They differ in the way the distance is calculated and the method used to compute the time evolution of the system.

- **Local OAT (ODE):** difference computed between the output resulting from the ODE model. Performing the analysis more than once will lead to the same result, due to the deterministic nature of the ODEs.
- **Local OAT (Gillespie average) :** Many stochastic simulations are computed here, so the result may change from analysis to analysis, reducing its variation if the number of stochastic simulations increases. This analysis is performed exactly in the same way as the previous one. Instead of the ODE output, the average of the simulations output is used.
- **Local OAT (Gillespie Density):** Also in this case, the Gillespie's SSA is used to compute the evolution of the system. In this analysis the histogram distance is used instead of the simple difference.

All the three analysis are performed perturbing each parameter by 1% of its nominal value one at a time, from a nominal set of parameters. In the result, both the Elementary OAT (just distance) and the derivative (distance divided by the perturbation) are computed. In all the cases, the evolution of the system was computed from time zero to 20 seconds, performing the analysis at that final time. This allowed the system to reach a stable point. ODE time evolution was computed with the simulator $ODE - RK5 - fixed$, the 5th order Runge-Kutta algorithm, with fixed time step (see Dizzy manual). Gillespie's direct method was used to perform the stochastic simulations and the number of simulations was 5000. When dealing with histogram distance, two cases were considered, with 50 and 100 histogram columns, showing the same sensitivity indices. Table 3.1 and Figures 3.3 and 3.4 show the outcome of the analysis.

3.1.3 Discussion

Given the difference in the order of magnitude of the parameters of the Schlögl model, we may be more interested in the relative perturbation. For this reason we consider the

	LOAT ODE				LOAT Gillespie Ave				LOAT Gillespie Density			
Params	dist.	rank	deriv.	rank	dist.	rank	deriv.	rank	dist.	rank	deriv.	rank
k_1	503	1	2e11	1	57	3	2e10	1	0.32	2	1e8	1
k_2	0.07	7	7e4	3	69	1	7e7	2	0.16	4	2e5	2
k_3	486	3	5e7	2	8.4	5	8e5	3	0.09	5	9e3	3
k_4	2.1	6	61	5	17	4	485	4	0.24	3	6.9	4
X_0	485	5	196	4	67.5	2	27.3	5	0.08	6	0.033	5
A	503	2	0.5	6	4.5	6	5e-3	6	0.32	1	3e-4	6
B	486	4	0.24	7	2.98	7	1e-3	7	0.08	7	4e-5	7

Table 3.1: Result of the three local one at a time Sensitivity Analysis described in the text. The time of the analysis is 20 seconds with a perturbation of 1%. In the Gillespie density column, the histogram distance is computed with 50 histogram columns and 5000 runs. The histogram self distance for X is 0.068. ODE fractional value 0.001. Results obtained using the simulator Dizzy [17]. A more comprehensible visualization of this Table can be found in Figure 3.3 for the derivative approach and in Figure 3.4 for the distance approach.

simple output difference a more interesting sensitivity index than the derivative and we will discuss that first.

In Table 3.1 the first interesting observation is that ODE and Gillespie Density procedures share common results. They both show that k_1 produces the same variation as A and that k_3 produces the same variation as B. Indeed, we know that k_1 and A are related, because they could have been considered a single parameter (see propensity function definitions) and this fact has been captured by the analysis. The same reasoning holds for k_3 and B. On the other hand, an important and expected difference appears in the X_0 influence: with ODE, the output variation induced by the perturbation of X_0 is comparable with the one of k_1 and k_3, showing high sensitivity. This is due to the crossing of the bifurcation point. The Gillespie Density method shows instead a low value of histogram distance for the same perturbation, revealing it far less influential than k_1. This latter method can easily proved to be the correct one by considering Figure 3.5, where the histograms of the distribution of X at time 20, generated with nominal and perturbed parameter values, nearly coincide. In the figure can also be observed how the perturbation of k_1 influences the outcome of the stochastic simulations. Moreover, k_3, along with B, has been discovered to be not particularly influential, with

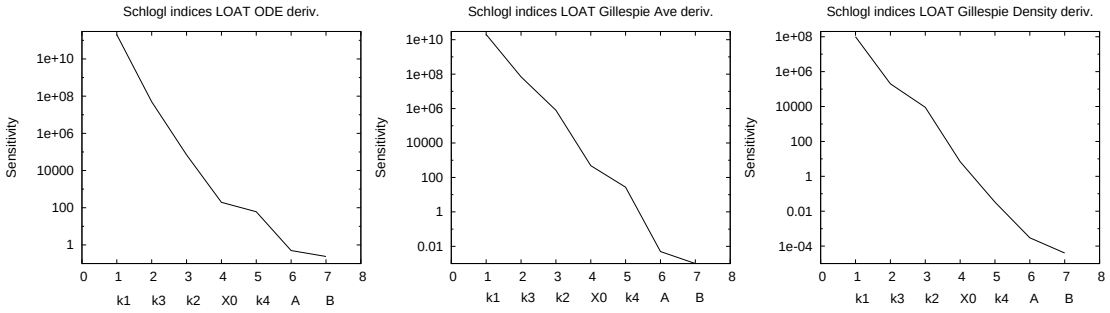

Figure 3.3: Visualization of the results in Table 3.1 obtained with the three methods using the derivative approach. No particular difference is present, the results are not particularly meaningful.

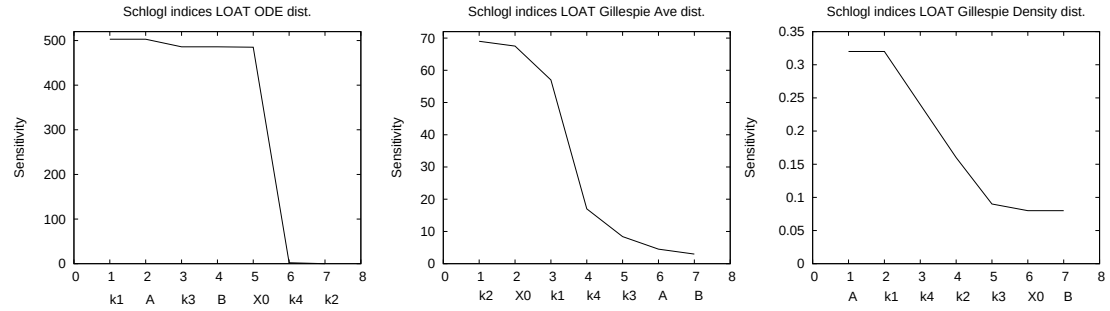

Figure 3.4: Visualization of the results in Table 3.1 obtained with the three methods using the distance approach. These results are discussed in the text.

a histogram distance close to the self distance.

The Gillespie Average approach seems instead inconsistent, without any pattern that could be explained.

To conclude the discussion of the results we can notice how the derivative approach presents more or less the same order of importance of the parameters in all three cases. This is due mainly to their different order of magnitude and it is meaningless in regard to the sensitivity of the system. It is clear that, at least in this contest, a parameter that is estimated to be of the order of 10^{-5} and a parameter that is estimated to be of the order of 10^{3} are not directly comparable.

According to the results of this first part, from now on, we will prefer the simple distance, specifying the relative perturbation in percentage.

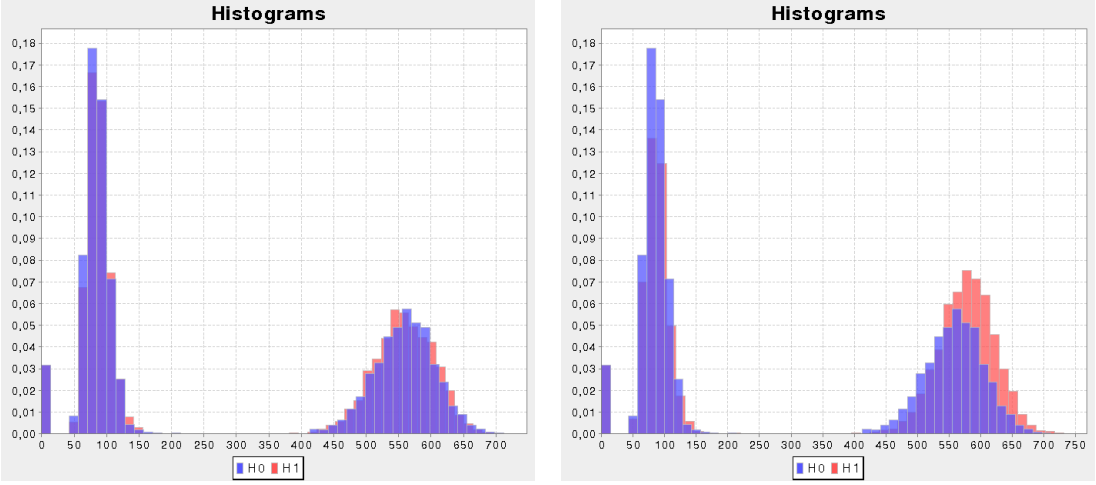

Figure 3.5: Values of output X of the stochastic simulations at time 20. On the left with nominal parameter values (H0) and X_0 perturbed by 1% (H1). On the right with nominal parameter values (H0) and k_1 perturbed by 1% (H1). Each histogram is obtained from 5000 samples grouped in 50 columns.

3.2 Screening methods

In this section we apply and study the Morris'method (see Section 2.3.4) in two different version: an adapted version of the original algorithm that makes use of the output of ODEs and a novel approach, based on the original method, but modified in order to use the information captured by sets of stochastic simulations.

3.2.1 Implementation of Morris' method

Given our previous experience with the local OAT sensitivity analysis, we make use of two different *elementary effects*, one based on simple difference of the output of a ODE model and the other one based on histogram distance of outputs of stochastic simulations. Moreover we consider the possibility of having multiple outputs:

$$d_{ij}(\mathbf{x}) = y_j(x_1,\ldots,x_i+\Delta,\ldots,x_k) - y_j(\mathbf{x})$$

$$d_{ij}(\mathbf{x}) = D(Y_j, Y_j')$$

where in general d_{ij} is the local influence of the ith input on the jth output of the model. Considering a certain fixed time t in which the analysis is performed, y_j is the outcome of the output j at that time and $\mathbf{x}$ is the vector of parameters. Y_j is the *random*

variable (r.v.) of the outcome of the jth output at time t distributed following the pdf $f(y_j, \mathbf{x})$ and Y_j' is the r.v. of the outcome of the jth output at the same time distributed following the pdf $f(y_j, x_1, ..., x_i + \Delta, ..., x_k)$. D is the histogram distance as defined in [5].

In order for these two measures to have meaning, we slightly modified the method to have perturbations that are always comparable. The ranges are chosen as displacement from a nominal value that is proportional ($\pm 10\%$) to that value. In the p-level grid we allow only unitary perturbations (not multiples of $1/(p-1)$ but exactly $1/(p-1)$ every time). This way, every difference corresponds to the same percentage in perturbation with respect to the parameter nominal value, that is central in the grid (Figure 2.3).

A couple of further simple improvements have been implemented. The first one concerns the random points generation: it consists of the insertion of a *taboo list* used to prevent a random point in the grid being selected twice. If a random point is already present in the list, another is immediately chosen. The second one is a *dynamic programming* approach that allows us to save computational time. Noticing that a point in the grid once perturbed is still a point in the grid, it may happen that the model output corresponding to some points risks being computed twice (see Figure 2.3). In our implementation, every result of a simulation of the model is stored and reloaded if a point in the grid is required a second time.

3.2.2 The Michaelis/Menten model

As in the previous Chapter, we pass now to the application of this method, to better show its characteristics. For this reason we now introduce the Michaelis/Menten model of the enzymatic reaction. This is a very simple model and allows us to show that, at least for a small number of simple reactions, the results obtained with the classical method and the version with histogram distance provide analogous results. Once the base cases are coherent, we can analyse the Schlögl model to show the differences with the new approach.

The model is defined as follows:

$$
\begin{aligned}
S + E &\xrightarrow{a_1} ES, \\
ES &\xrightarrow{a_2} E + S, \\
ES &\xrightarrow{a_3} P,
\end{aligned}
$$

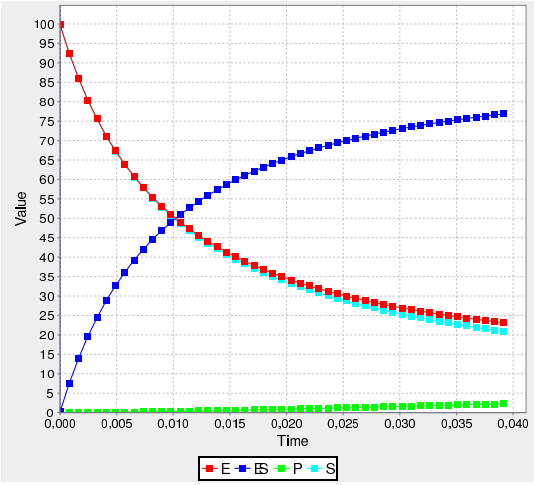

Figure 3.6: Example of time evolution of Michaelis/Menten model computed with ODE.

The propensity functions of Gillespie's CME are:

$$a_1 = enzyme\_substrate\_combine * E * S,$$
$$a_2 = enzyme\_substrate\_separate * ES,$$
$$a_3 = make\_product * ES,$$

and the nominal values of the parameters are set as follows:

$$E_0 = 100$$
$$S_0 = 100,$$
$$ES_0 = 0,$$
$$P_0 = 0,$$
$$enzyme\_substrate\_combine = 1,$$
$$enzyme\_substrate\_separate = 1,$$
$$make\_product = 1.$$

The time evolution of this system is shown in Figure 3.6.

3.2.3 Morris' Methods on Michaelis/Menten model

A new version of the already presented Morris' Method has been implemented to make use of the histogram distance [5]. The method is identical to the analogous one for

	Elementary effects in Michaelis model (ODE)		
Params	mean	standard dev.	rank
S_0	2.27	0.30	1
enzyme_substrate_combine	0.76	0.07	3
make_product	0.08	0.006	4
enzyme_substrate_separate	0.05	0.002	5
E_0	2.16	0.30	2

Table 3.2: Result of the Morris' method on Michaelis model. Time evolution is computed with ODE and the effect is measured as the difference in the output ES. The adopted parameters are 1000 random points (r), grid level 5 (p), time of the analysis 0.04s, $\pm$10% from nominal value, ODE fractional value 0.001.

ODEs: a random repetition of Local Sensitivity Analysis in the parameter space. Using histogram distance this allows us to collect information about how the histograms (stochastic simulations) vary when the parameters explore the space of their possible values. A first application has been performed on the simple Michaelis/Menten model, with results shown in Table 3.2 and 3.3 (visualizations in Figure 3.7). The first important thing we notice is how the two analysis give the exact same result. Both identify the initial concentrations of E and S as the most influential parameters, the kinetic *enzyme_substrate_combine* as relevant, and the other kinetics as not particularly influential. This shows the correctness of the new approach based on histogram distance in analysing a simple and well known model.

Moreover, it can be seen that the precision of the approach based on histogram distance is particularly good in this case. Because the histogram distance is an approximation to the distance between the underlying distributions, it is necessary to use and interpret the results carefully. In this case we observe (still Table 3.3) that the self distance is low and defined with good precision (low standard deviation as well). This is mainly due to the number of stochastic simulations (10000 divided in 50 columns) that assures a good enough approximation. As a consequence, we can state that the initial concentrations of E and S have always the same influence on the distribution of ES and that this influence does not depend on the value of other parameters. The same can be said for *enzyme_substrate_combine*, though its influence is clearly inferior. Not influential are instead *make_product* and *enzyme_substrate_separate*, since the values of their indices are very close to the average histogram self distance.

	Elementary effects in Michaelis model (Gillespie density)		
Params	mean	standard dev.	rank
S_0	0.52	0.06	1
enzyme_substrate_combine	0.18	0.02	3
make_product	0.03	0.008	4
enzyme_substrate_separate	0.03	0.008	5
E_0	0.49	0.07	2

Table 3.3: Result of the Morris' method on Michaelis model. Time evolution is computed with Gillespie's direct method and the effect is measured as the difference in the histograms of ES. The adopted parameters are 100 random points (r), grid level 5 (p), time of the analysis 0.04s, $\pm$10% from nominal value, 10000 simulations and number of histogram columns 50. The average histogram self distance of the random points was 0.026 with std. dev. 0.006.

3.2.4 Morris' Methods on Schlögl model

The two screening methods have been applied also to the prototype Schlögl model (see Section 3.1.1). Table 3.4 shows the outcome of the analysis with ODEs used to determine the time evolution of the system. The average elementary effect has the role of ordering the parameters from the most to the least influential. However, the elevated standard deviation of all the parameters sensitivities makes this classification difficult and reveals that the model is likely to be nonlinear with respect to the parameters and strong dependency between the parameters is also likely to exist. Further analysis is necessary to confirm and quantify dependencies and sensitivities and this analysis would likely involve all the factors.

Table 3.5 shows Morris' method applied using histogram distance. Before discussing the results, we need to put forward the circumstance that, due to computational complexity, we had to limit the precision of the analysis. Instead of using 10000 stochastic simulations for each experiment and 100 random points, as in the analysis of the Michaelis/Menten model, this time we reduced the simulations to 1000 and the random points to 40. The impact of this can be observed immediately on the average histogram self distance, that is at least five times greater than in the case of the analysis of the Michaelis/Menten model. It is then important to bear in mind that all the conclusions are up to the level of precision that is given by the average self distance.

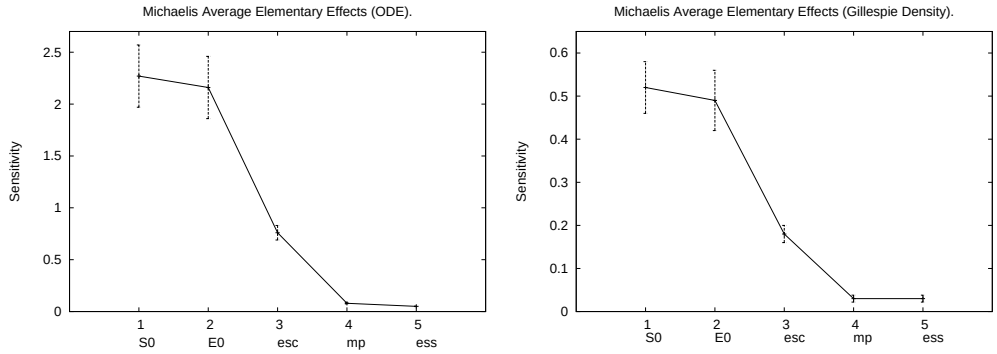

Figure 3.7: Visualization of the average elementary effects of the factors of the Michaelis model, along with their standard deviations. The data is taken from Tables 3.2 and 3.3. Abbreviations of the names of the factors are: **esc** is *enzyme_substrate_combine*, **mp** is *make_product* and **ess** is *enzyme_substrate_separate*.

Observing Table 3.5 we can at this point say that, with the current approximations, the initial number of molecules of the species X is a factor that appears to be not influencing the value of the species X at time 10 seconds. We can also see that the product k_3B has a weak influence and that this influence does not change particularly as other parameters change (relatively low standard deviation). The other three parameters, k_2, k_4 and k_1A show instead that they have a significant influence, specially k_1A, and their relative larger standard deviation implies non linearity and correlations that require further and more specialized analysis.

Figure 3.8 highlights some differences between the results of the two approaches. The most significant one is certainly the reduction of the relative standard deviation that helps us to be more confident when stating which factors are the most important and which require to be further analysed.

Finally, it is interesting to notice how this analysis completes the previous performed local OAT analysis (recall Table 3.1). Thanks to this analysis we are sure that there is more to investigate about the relationships between factors and output.

3.3 Global methods

In this chapter we present and study an alternative method to the one described in Section 2.3.5 for computing the *variance of the conditional expectation*.

	Elementary effects in Schlögl model (ODE)		
Params	mean	standard dev.	rank
k_1A	137.74	210.24	1
k_2	40.99	116.77	3
k_3B	28.13	108.08	5
k_4	107.78	201.60	2
X_0	29.41	116.40	4

Table 3.4: Result of Morris' method on the Schögl model. Time evolution is computed with ODEs and the effect is measured as the difference in the output X. The adopted parameters are 1000 random points (r), grid level 5 (p), time of the analysis 10s, $\pm$5% from nominal value, ODE fractional value 0.001.

	Elementary effects in Schlögl model (Gillespie Density)		
Params	mean	standard dev.	rank
k_1A	0.649	0.199	1
k_2	0.317	0.160	3
k_3B	0.214	0.055	4
k_4	0.451	0.124	2
X_0	0.159	0.037	5

Table 3.5: Result of the Morris' method on the Schögl model. Time evolution is computed with the *logarithmic direct method* and the effect is measured as the difference in the histograms of X. The adopted parameters are 40 random points (r), grid level 5 (p), time of the analysis 10s, $\pm$5% from nominal value, 1000 simulations and number of histogram columns 50. The average histogram self distance of the random points was 0.141 with std. dev. 0.025.

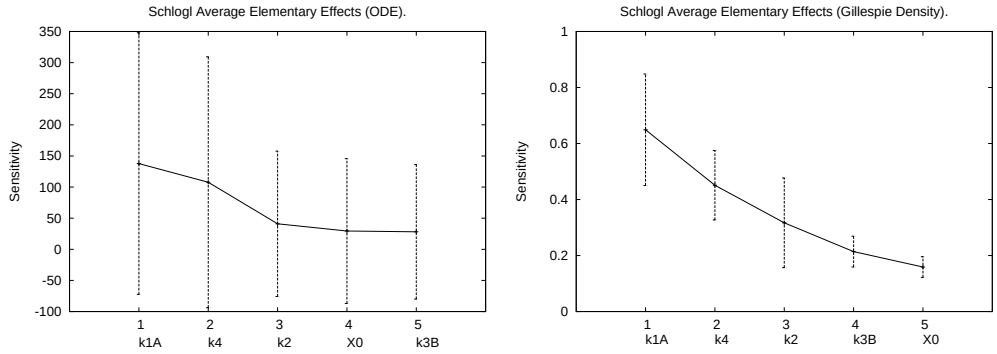

Figure 3.8: Visualization of the average elementary effects of the factors of the Schlögl model, along with their standard deviations. The data is taken from Tables 3.4 and 3.5.

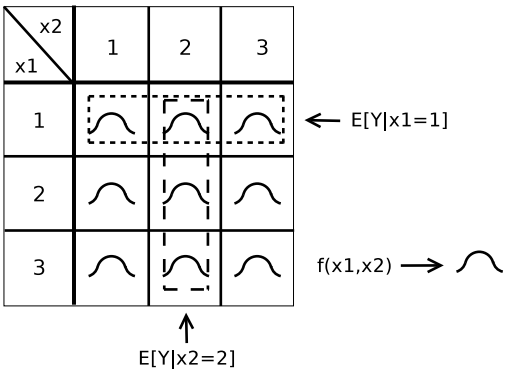

Figure 3.9: A possible model whose output is a probability distribution is shown. A conditional expectation can be obtained using a set of distributions.

3.3.1 Variance-based methods for simulations of biochemical reactions

The possibility of developing analogous techniques that make use of the histogram distance has been investigated. The idea is that we can compute the variance of the conditional expectation introducing the concept of *mean of histograms* and the histogram distance.

Figure 3.9 shows a possible model whose output is a *probability density function* (pdf). Using Monte Carlo simulations, we can obtain a set of values that we can assume to be generated by using that pdf. This pdf can be approximated using a histogram function [5]. The definition of mean is nothing but that value which minimizes the sum of the differences between each value and the mean. The histogram that minimizes the sum of histogram distances between itself and each histogram is nothing but the histogram that collects all the points of all the histograms (as can intuitively be seen by

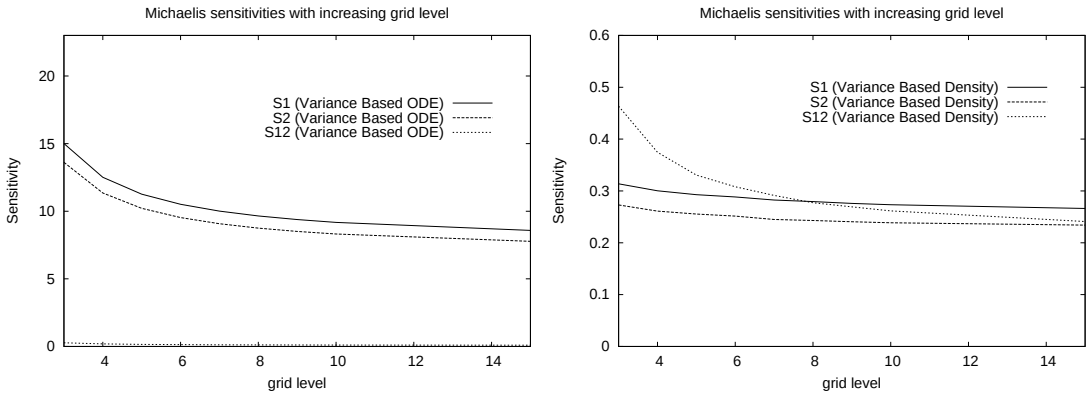

Figure 3.10: First order sensitivity indices as defined in the text for the initial concentrations S_0 (S1), E_0 (S2) and their combined effect (S12). Indices are shown with increasing grid level. Although the index S12 appears to be not coherent, clear analogies are present in the two methods relative to the indices S1 and S2. Time of the analysis 0.04s, fractional step size of ODE method 0.001, number of stochastic simulations 10000, number of histogram columns 50.

looking at how the histogram distance is computed, Section 2.3.1). Using this concept it has been possible to implement a version of the variance based methods also with the histogram distance. The weak point so far is that a mathematical proof of Equation (2.8) has not been provided and the intuition is that it is not valid. However, other results (see next Section) show how this technique is supposed to have some good foundations, providing results that appear to be coherent with the classical approach.

3.3.2 Application to Michaelis/Menten model

Applying the two methods to the Michaelis/Menten model led to the results shown in Figure 3.10. This test has been considered useful as a first glimpse into the correctness of the new approach. In both the approaches, the first order sensitivity indices S1 and S2 (relative to the factors S_0 and E_0 respectively) maintain the same relative importance. The index relative to the combination of the factors S12 shows instead a different behaviour in the two cases. On the other hand, if we look at the results in terms of total sensitivity indices TS_1 and TS_2, we can say that the results are perfectly compatible. In both the cases TS_1 is always greater than TS_2.

These preliminary results have led to further investigations. The variance-based analysis has been extended also to a third parameter (*enzyme_substrate_combine*), that in previous analysis showed to have some influence on the model. The first-order

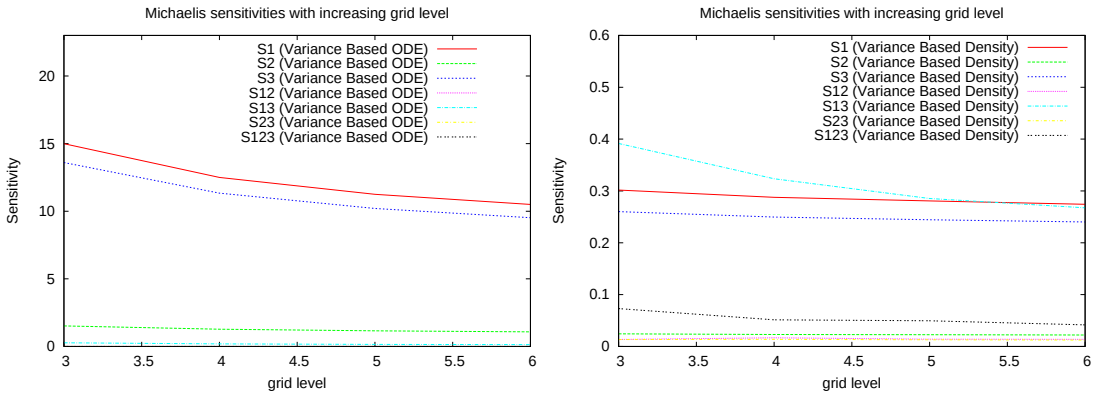

Figure 3.11: First order sensitivity indices as defined in the text for the initial concentrations S_0 (S1), *enzyme_substrate_combine* (S2), E_0 (S3) and various combined effects. Indices are shown with increasing grid level. Same parameters of the results in Figure 3.10.

sensitivity indices, based on the decomposition of variance, are shown in Figure 3.11. The results based on classical analysis still show the indices relative to S_0 and E_0 as the most important. Observing the results of the new approach, we notice that they are coherent with the analysis performed on only two parameters. Even the behaviour of the index corresponding to the combination of the factors E_0 and S_0 is replicated, suggesting this not to be just chance.

Finally, also in this case, from the point of view of the total sensitivity indices, the results of the two techniques are analogous.

3.3.3 Application to Schlögl model

The variance-based analysis has been applied to the Schlögl model, both with the classical and the new approach. The analysis has been performed considering a subset of three parameters, selected as the most important factors arising from a previous analysis with the Morris' method (Table 3.5). The factors are k_4, k_1A and k_2.

The results of the analysis of the two variance based approaches are shown in Table 3.6. We notice that the order of importance of the three parameters is the same, according to the total sensitivity indices. Differences in the first and second order sensitivity indices may be due to the relative weaker importance that k_2 seems to have in the classical analysis. Indeed, sensitivities involving k_2, like S_{13} or S_{23} are weaker in the classical analysis.

	VCE with histogram distance		VCE with ODEs	
index	sensitivity	rank	sensitivity	rank
S_1	0.244	2	21278	2
S_2	0.325	1	30366	1
S_3	0.064	5	4028	5
S_{12}	0.086	4	10033	3
S_{13}	0.008	7	299	7
S_{23}	0.086	3	456	6
S_{123}	0.054	6	8969	4
TS_1	0.392	2	40580	2
TS_2	0.551	1	49826	1
TS_3	0.213	3	13752	3

Table 3.6: Variance-based sensitivity analysis of Schlögl model. First-order sensitivity indices relative to the factors k_4 (1), k_1A (2) and k_2 (3) and other combined effects are shown. The last three rows show the total sensitivity indices. Time of the analysis 10s, grid level 5, fractional step size of ODE method 0.001, number of stochastic simulations 1000, number of histogram columns 50.

Chapter 4

Sensitivity analysis of the mitogen-activated protein kinase (MAPK) cascades

4.1 Introduction to MAPK

When speaking about the mitogen-activated protein kinase (MAPK) cascades [15, 13] we mean one or more signaling pathways that share a particular common structure.

MAPK cascades consist usually of three levels, where the signal is transmitted from one level to another through the phosphorylation of a kinase that, once activated, phosphorylates the kinase at the next level down the cascade (Figure 4.1). The MAPK protein that triggers the cell response usually needs to be activated through a two-site phosphorylation. The catalyst for this reaction is a MAPKK (MAPK kinase) molecule and, at the upper level, the same role belongs to a MAPKKK (MAPKK kinase) molecule. The last molecule in this model is the MKP (MAP kinase phosphatase) which dephosphorylates, and so deactivates, the MAPK molecule.

Usually, the phosphorylation of a MAPK is on two distinct sites and we can distinguish between a form phosporylated on the tyrosine aminoacid alone (MpY) and a form phosphorylated on the theronine aminoacid alone (MpT). These two operations can be performed by one single MAPKK, in one single collision, or, more likely, in two. It is also possible that two MAPKK enzymes are needed, one that catalyzes the phosphorylation on the tyrosine and the other that catalyzes the phosphorylation on the theronine.

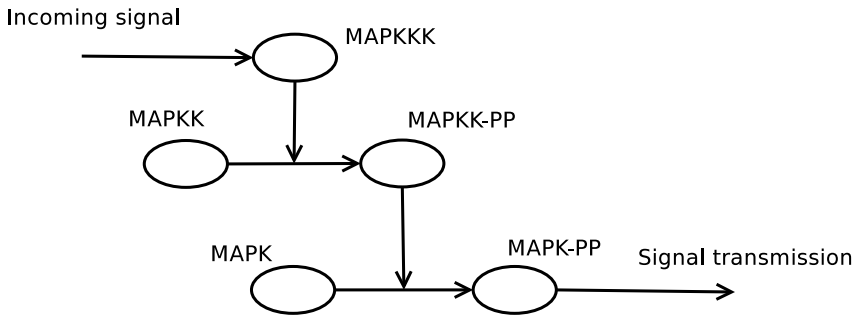

Figure 4.1: Structure of a MAPK cascade. At each level, the enzyme that catalyzes the reaction in the next level is activated by a two-site phosphorylation.

4.2 MAPK model

The model of MAPK that we introduce in this section has been presented in [15]. We consider a single level of the MAPK cascade presented in the previous section, with only one MAPK kinase and without making any distinction between MAPK phosphorylated on tyrosine or theronine. In any case, the model consists of a two step double phosphorylation (Figure 4.2). From now on, talking about this one level of the MAPK cascade, we use M, Mp and Mpp as the unphosphorylated, monophosphorylated and biphosphorylated forms of MAPK.

The original model is written as a system of *Ordinary Differential Equations* (ODEs) that describes the evolution of the concentration of M, Mp and Mpp in time. The rate at which these concentrations change is obtained using assumptions from the Michaelis-Menten kinetics (see Section 2.2.6). We use the same set of equations, but with number of molecules instead of concentrations. This procedure is correct if we assume that the product of the cell volume and the Avogadro number is equal to 1.

The system in Figure 4.2 is defined by the following enzymatic reactions:

$$
\begin{aligned}
&M + MAPKK \overset{k_1,k_{-1}}{\leftrightarrow} \textit{M-MAPKK} \overset{k_2}{\rightarrow} Mp + MAPKK,\\
&Mp + MAPKK \overset{k_3,k_{-3}}{\leftrightarrow} \textit{Mp-MAPKK} \overset{k_4}{\rightarrow} Mpp,\\
&Mpp + MKP3 \overset{h_1,h_{-1}}{\leftrightarrow} \textit{Mpp-MKP3} \overset{h_2}{\rightarrow} \textit{Mp-MKP3} \overset{h_3,h_{-3}}{\leftrightarrow} Mp + MKP3,\\
&Mp + MKP3 \overset{h_4,h_{-4}}{\leftrightarrow} \textit{Mp-MKP3}^{*} \overset{h_5}{\rightarrow} \textit{M-MKP3} \overset{h_6,h_{-6}}{\leftrightarrow} M + MKP3.
\end{aligned}
$$

Notice how, in the first two lines, phosphorylation and product dissociation are considered a single step, while, in the last two lines, dephosphorylation and product release are two distinct steps.

This system can be reduced to only four reactions, under the assumptions of con-

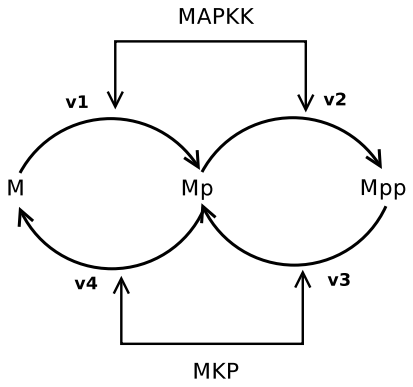

Figure 4.2: Model of a level of the MAPK cascade. M, Mp and Mpp stand for the unphosphorylated, monophosphorylated and biphosphorylated forms of MAPK.

stant number of ATP/ADP molecules and protein-protein complexes at steady-state. These are the resulting reactions:

$$
\begin{aligned}
&M \xrightarrow{v_1} Mp,\\
&Mp \xrightarrow{v_2} Mpp,\\
&Mpp \xrightarrow{v_3} Mp,\\
&Mp \xrightarrow{v_4} M,
\end{aligned}
$$

where the rates v_i are given by the following equations:

$$
\begin{aligned}
v_1 &= \frac{k_1^{cat} \cdot MAPKK \cdot M/K_{m1}}{(1 + M/K_{m1} + Mp/K_{m2})},\\
v_2 &= \frac{k_2^{cat} \cdot MAPKK \cdot Mp/K_{m2}}{(1 + M/K_{m1} + Mp/K_{m2})},\\
v_3 &= \frac{k_3^{cat} \cdot MKP3 \cdot Mpp/K_{m3}}{(1 + Mpp/K_{m3} + Mp/K_{m4} + M/K_{m5})},\\
v_4 &= \frac{k_4^{cat} \cdot MKP3 \cdot Mp/K_{m4}}{(1 + Mpp/K_{m3} + Mp/K_{m4} + M/K_{m5})}.
\end{aligned}
$$

In these expressions MAPKK and MKP3 are the total amount of molecules of the two enzymes and are considered constant through time. The nominal values of the parameters and the relationship with the kinetics of the elementary enzymatic reactions are as follows:

$$
\begin{aligned}
k_1^{cat} &= k_2 = 0.01,\\
k_2^{cat} &= k_4 = 15,
\end{aligned}
$$

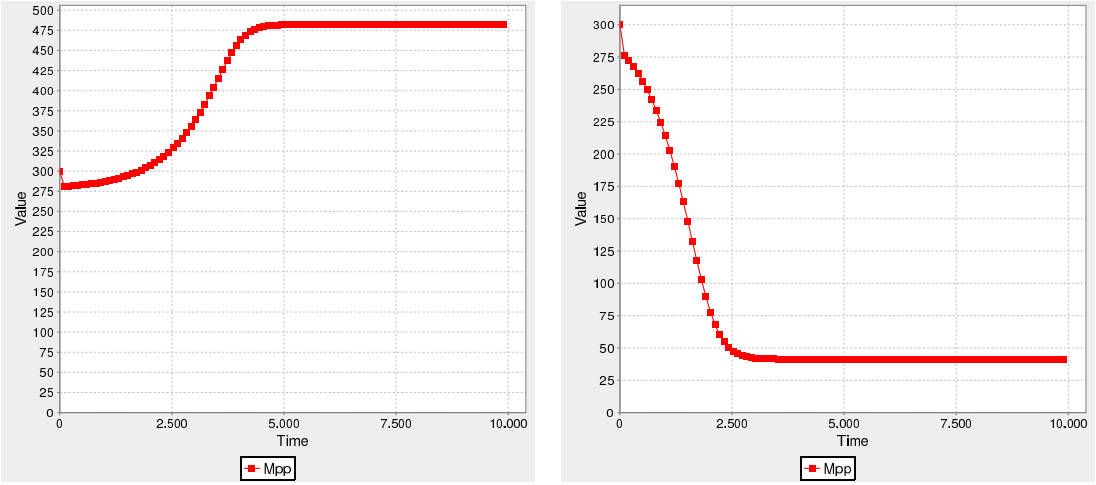

Figure 4.3: Time evolution of the Mpp molecule of the MAPK model computed with ODEs is shown. In the figure on the left, the model has the nominal parameter values described in the text. In the figure on the right, the initial number of molecules of the phosphatase MKP3 is incremented by 5%.

$$
\begin{aligned}
&K_{m1} = (k_{-1} + k_2)/k_1 = 50,\\
&K_{m2} = (k_{-3} + k_4)/k_3 = 500,\\
&k_3^{cat} = h_2/(1 + h_2/h_3) = 0.084,\\
&k_4^{cat} = h_5 \cdot (1 + h_5/h_6 + h_{-3} \cdot (h_{-4} + h_5)/(h_3 \cdot h_4))^{-1} = 0.06,\\
&K_{m3} = (h_{-1} + h_2)/(h_1 + h_1 \cdot h_2/h_3) = 22,\\
&K_{m4} = (h_{-4} + h_5) \cdot (h_4 \cdot (1 + h_5/h_6 + h_{-3} \cdot (h_{-4} + h_5)/(h_3 \cdot h_4)))^{-1} = 18,\\
&K_{m5} = (h_6/h_{-6}) = 78,\\
&M_0 = 200,\\
&Mp_0 = 0,\\
&Mpp_0 = 300,\\
&MAPKK1_0 = 50,\\
&MKP3_0 = 100.
\end{aligned}
$$

The particularity of these parameter values is that they are close to a bifurcation point. As it can be seen in Figure 4.3, the perturbation of a parameter value can lead to a radical change in the behaviour of the ODE time evolution of the double phosphorylated MAPK (Mpp). Again, as it was for the Schlögl model (Section 3.1.1), a set of runs of the *Stochastic Simulation Algorithm* (SSA) (Section 2.2.4) shows that the real behaviour of the system with the nominal parameters is a choice between two stable systems (see Figure 4.4). Moreover, thanks to [15], we know that this system, with the

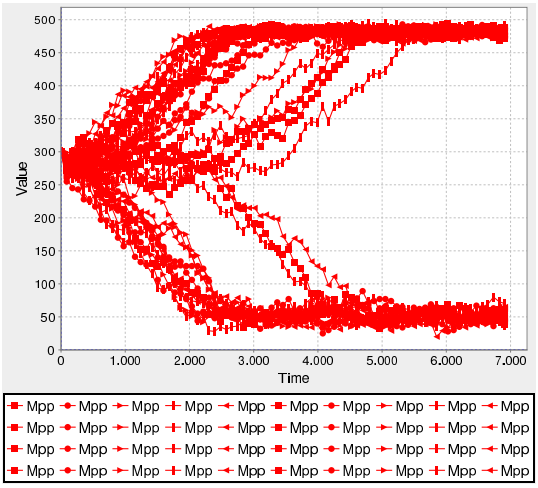

Figure 4.4: Time evolution of the Mpp molecule of the MAPK model computed with the stochastic simulation algorithm (SSA) and the nominal parameters value in the text. These 40 runs of SSA show how the evolution of the system may lead to two different stable systems.

stated parameters, presents three steady-states, that we can consider three attractors for the stochastic simulations. This situation is confirmed by the already mentioned graph of the time evolution of Mpp in Figure 4.4, where, although the choice appears to be between two attractors, this choice is delayed for some runs, that are likely to be influenced by a central attractor.

4.3 Sensitivity analysis

In this section we apply both the classical approach and the techniques we developed in Chapter 3 to the presented MAPK model. We will proceed with a comparison of the methods throughout the analysis. Our choice is to measure the influence of the factors, kinetics and initial number of molecules, on the amount of double phosphorylated MAPK (Mpp). To do so, we choose the time of the analysis to be 2000 seconds. This time, as revealed in Figure 4.4, is at the core of the choice between the two possible behaviours of the system and is within the limits of our possibilities in terms of computational power, when using the SSA.

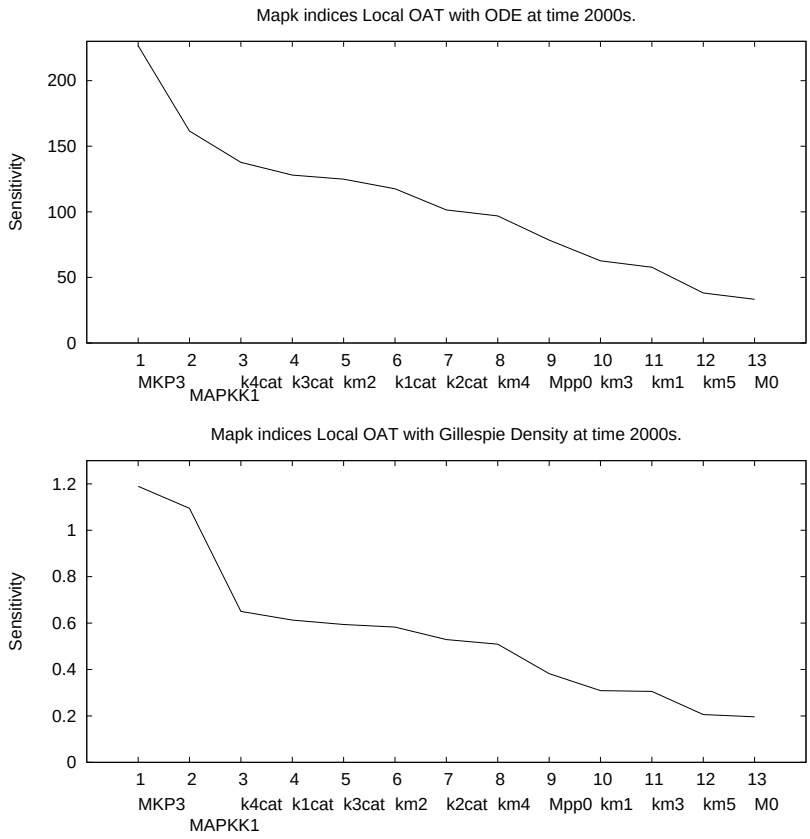

Figure 4.5: Local OAT sensitivity analysis of the MAPK model at time 2000 seconds. The result of classical analysis is shown in the first graph, while the result of the analysis based on histogram distance is shown in the graph below. ODE time evolution is computed with fractional step size of 0.0001, while we used 10000 stochastic simulations and 50 histogram columns in the novel approach. The perturbation of each parameter has been of 5%. The histogram self distance is 0.1.

4.3.1 Local one-at-a-time analysis

As a first step in the sensitivity analysis of the MAPK model, we performed a local one-at-a-time (OAT) analysis. As we have seen in Section 2.3.3, it consists in the perturbation of one of the factors at a time and in the measurement of the corresponding output change with respect to the original model. We used two different measures: the simple difference of the values of Mpp at time 2000 seconds generated using the ODE based time evolution; the histogram distance between the sets of values of Mpp at time 2000 seconds collected using stochastic simulations. With this first and computationally cheap analysis, we can have an idea of the relevance of the factors in the immediate surrounding of the factor nominal values. However, we have to bear in mind that without a global analysis we cannot be certain of the implications that may arise

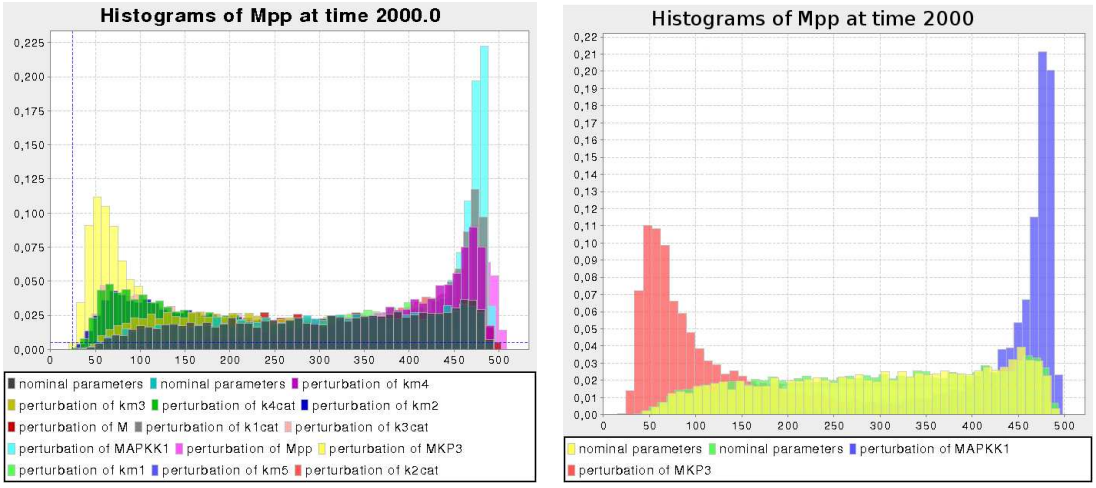

Figure 4.6: Histograms that collect the values of Mpp obtained using 10000 stochastic simulations. Each histogram is divided in 50 columns. On the left, all the histograms resulting from the one-at-a-time (OAT) analysis, one for each factor perturbed, are shown. On the right, a second OAT analysis with only the histograms relative to the perturbations of the initial amount of MAPKK and MKP3 are shown. The histograms labeled with *nominal parameters* are those generated with the values of the parameters stated in the text.

from perturbing more than one factor simultaneously. This last information cannot be neglected when trying to assert the influence of a factor on the model.

The results of the local one-at-a-time analysis are shown in Figure 4.5. The thirteen factors are listed in the graphs from the most relevant to the least. We can notice that the relative order of importance is not particularly affected by the method used for the analysis. However, with the first approach it appears that just the amount of phosphatase MKP3 is the most relevant factor, while with the second approach, the intuition is that both the amount of kinase MAPKK and phosphatase MKP3 are the most relevant factors, above all the others.

This last statement can be proved to be correct, at least in this local analysis, showing the histograms generated using the results of the stochastic simulations of the perturbed models. Figure 4.6 highlights that the initial amount of MAPKK and MKP3 are both the most influential factors. Moreover, they play the strongest role in the choice between the two possible stable systems. They have opposite roles, since increasing the amount of one of the two enzymes leads to opposite choices. It is indeed not surprising that the condition of bistability is guided by the right proportion in the amount of enzymes that catalyze the reactions.

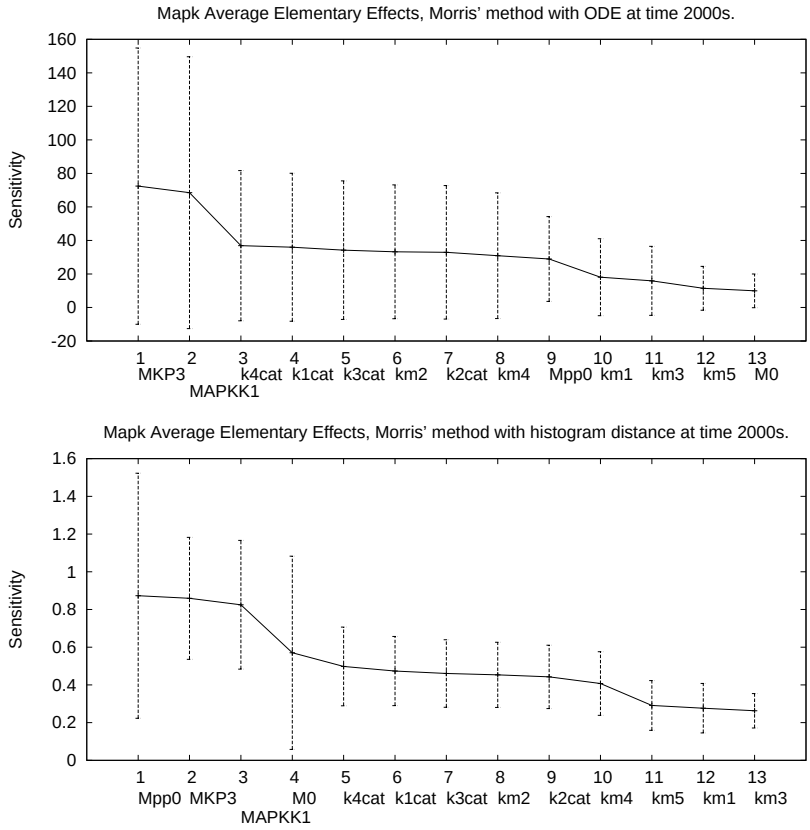

Figure 4.7: The result of the Morris' method applied to the MAPK model. ODE integration uses a fractional step size of 0.0001. Result obtained with a grid level of 5 and an average over 1000 random points. In the approach based on histogram distance, 1000 runs of SSA, 50 columns and 40 random points have been used and the average histogram distance is 0.15 with standard deviation of 0.061. The parameters vary within $\pm$ 10% of their nominal value.

4.3.2 Screening with Morris' methods

Before we proceed to a more detailed analysis, we wish to use a screening method to identify and then exclude those factors that are clearly the least influential. Once we have isolated only a small part of most influential factors, we can proceed with computationally expensive techniques that can provide the most detailed analysis. To do so, we use the techniques we developed in Section 3.2 based on the Morris' method (Section 2.3.4). As we have seen, we consider a range of possible values for each factor and then we sample in the vector space generated by all the possible combinations of values of all the factors. This sampling is done randomly and through the use of a grid. We use here a grid level of five, meaning that each of the thirteen parameters can assume one of five possible values. For each random point selected in the grid of all

the possible combinations of values, a local OAT analysis is performed. The indices resulting from that are the *elementary effects* that are local with respect to that random point. Averaging over all these local analyses reveals whether the degree of importance of a parameter is constant or changes when the other factors assume other values. The results of the Morris' method applied to the MAPK model are shown in Figure 4.7.

Also in this case we compare the results obtained with a method that uses the time evolution computed with ODE and a method that uses time evolution computed with SSA. The ODE based approach highlights that, although the most influential parameters are confirmed to be the initial amount of MAPKK and MKP3, the elementary effects of the factors are extremely variable. In this case it is difficult to say which factors we want to include in the detailed global analysis, if we exclude MAPKK and MKP3. The important standard deviation of the elementary effects is certainly due to a correlation between the factors and the non linearity of the model output with respect to the parameters.

Again, before discussing the results obtained with the Morris' method based on histogram distance, we need to put forward the fact that we had to limit the accuracy of the analysis, due to the high computational complexity of the algorithms and the limited computational power of our means. Each experiment is made of 1000 stochastic simulations, number that leads to a relatively high histogram self distance of 0.150, with also a standard deviation of 0.061. However, we have already seen in the local OAT analysis that the self distance can be considerable even with the more accurate precision of 10000 stochastic simulations (self distance of 0.1, see Figure 4.5). Therefore, it appears that the point in time where we perform our analysis is particularly unstable, with high stochasticity and indecision from the single runs about which stable system to choose. We can then assume that we have two factors that limit the accuracy of our results: a limited number of stochastic simulations and a strong stochasticity already present in the model.

The results of the Morris' method based on histogram distance, shown in Figure 4.7, confirm the high non linearity of the model and the inconstant influence of the parameters on the amount of double phosphorylated MAPK at time 2000 seconds. On the other hand, this method appears to achieve a more precise information with respect to the ODE based analysis. First of all, many parameters have reduced the standard deviation of their elementary effect. We can be more confident when stating that some factors are less influential than others. Moreover, it is more evident the strong influence that is attributed to the initial amount of the enzymes MAPKK and

	Variance-based with ODEs		Variance-based with histogram distance	
index	sensitivity	rank	sensitivity	rank
S_1	15695.65	1	0.350	2
S_2	15308.66	2	0.332	3
S_{12}	5631.88	3	0.811	1

Table 4.1: First and second order sensitivity indices relative to the factors $MAPKK_0$ (1) and $MKP3_0$ (2) of the MAPK model and their combined effect (12), obtained computing the variance of the conditional expectation. The fractional step size used in the ODE integration is 0.0001, the number of stochastic simulations used is 5000 and the number of histogram columns is 50. The parameters vary within $\pm$ 10% of their nominal value.

MKP3. Finally, this second analysis assigns a different role to the factors Mpp_0 and M_0. Here, they appear to have a stronger average sensitivity, though this sensitivity may vary considerably (large standard deviation), showing a strong dependence on the value of the other parameters.

4.3.3 Global analysis with variance decomposition

Thanks to the screening that we practiced in the previous section, we can now apply a global and more informative method to a reduced set of parameters taken from the factors of the MAPK model. The factors that proved to be the most influential are the initial number of molecules of MAPKK and MKP3, so we investigate their influence as single parameters and their combined effect. For this purpose we used the techniques developed in Section 3.3. Again, a method based on differences of outputs of ODEs and one based on histogram distances of executions of SSA are compared. These measures consider the variance of the output: while the former focuses on the variance of the ODE output, the latter estimates the variance in the distribution approximated by histograms. In both cases, the quantity of the variance that is due to each parameter is identified. The results can be found in Table 4.1.

In both the approaches, the initial amount of MAPKK and MKP3 present the same level of importance, with the former that is slightly more influential. The difference lies in the importance that is given to the combined effect of the two factors. While with the first approach the combined effect is considerably less than the single effects, with the second approach it appears that the two parameters are more linked. Changing

Figure 4.8: Distribution of the values of Mpp from 2000 stochastic simulations at time 2000 seconds, with simultaneous perturbation of MAPKK and MKP3. Above, from left to right: number of molecules of the enzymes decreased by 20%, 10% and with their nominal value. Below, from left to right: number of molecules of the enzymes increased by 10%, 20% and 30%.

them together leads to a stronger influence with respect to a one-at-a-time change.

The visualizations in Figure 4.8 help to prove the connection and reciprocal influence of the factors of this model. In this figure one can see that the combined perturbation of MAPKK and MKP3 leads to a variation of the distribution of the set of values obtained with the stochastic simulations. Although the mean of these values appears to be the same, the distributions seem to pass from a compact and largely gaussian shaped (on the left) to a more irregular one, which begins to show the two peaks of the bistability. This observations can be interpreted as the simple fact that increasing the amount of enzymes accelerates the process, allowing the two stable choices to be reached sooner. Other interesting visualizations are those in Figure 4.9, where we can observe that the ODE integration fails to interpret the high stochasticity and indecision present in the system at time 2000 seconds. However, also in this case, incrementing or decreasing the quantity of enzymes accelerates or slows the production of MAPK-PP (Mpp).

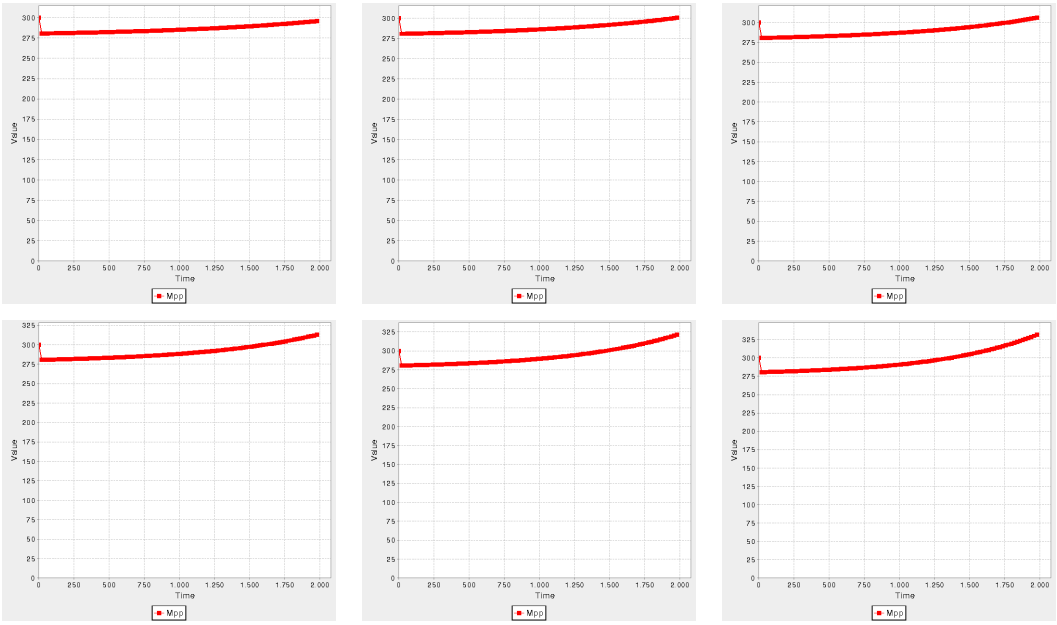

Figure 4.9: Time evolution of the double phosphorylated MAPK (Mpp) with ODE for 2000 seconds, with simultaneous perturbation of MAPKK and MKP3. Above, from left to right: number of molecules of the enzymes decreased by 20%, 10% and with their nominal value. Below, from left to right: number of molecules of the enzymes increased by 10%, 20% and 30%.

4.3.4 Discussion

In this section we showed an example of how a sensitivity analysis of a model of biochemical reactions can be performed using the tools so far developed. As a first result, we have shown how global analysis such as the Morris' method first and the variance decomposition after, are necessary and have to be used to identify the relationship between the factors. If we had to rely only on a local analysis, we would just accept the order of importance given in Figure 4.5. However, thanks to the further application of a global screening method (Figure 4.7), we have been able to state that this order of importance may vary if we change the value of more than a factor at once. This suggested, if not demanded, a further and more informative analysis concerning those factors that seemed the most influential and dependent on the others. In this case, we showed the intuitive relationship between the enzymes MAPKK and MKP3, whose simultaneous increment accelerates the system and whose proportions play the main role in the bistability of the system.

As a second but not less important result, the comparison between classical and novel approaches highlighted how, when dealing with bistable systems near a bifur-

cation point, it becomes necessary to have a sensitivity analysis tool that takes into account the distribution behind a set of stochastic simulations. Although the analytical analysis of the ODEs is fundamental to identify the bifurcation points and the multiple steady-states, ODE integrations cannot model the uncertainty in the time evolution of the system close to those bifurcation points (compare Figure 4.3 with Figure 4.4). In this situation of high stochasticity, a more suitable sensitivity analysis is one that takes into account the variations between sets of stochastic simulations rather than the simple output of a ODE integration. Here, for example, we have seen how a modified version of the Morris' method, identified some properties that the classical version was not able to capture (Figure 4.7).

Chapter 5

Conclusions

In this dissertation we followed a very precise path. First, we described the kinds of model we wanted to analyse and we identified *sensitivity analysis* (SA) as a tool to collect information and understand properties of these models. Then, classical and novel techniques were presented and compared. Our goal was to show how the SA of models of biochemical reactions can be performed with the present means, understand when these give poor or misleading results and investigate new methods that better adapt in such cases.

One of the first conclusions that we verified when we applied SA techniques has been the need for a global analysis that highlights the relationships between the factors of the model. If we assume that all the factors are uncertain within a range of possible values, a local approach (such as the local one-at-a-time) is just able to capture sensitivity indices that are conditioned on the choice of the nominal values. If we want to quantify the influence of the factors on the output of the model and correctly identify the role that these factors play, it is necessary to estimate whether the influence of a parameter changes when the other parameters assume other values, i.e. whether the factors influence the sensitivity of other factors. This is possible only if the analysis takes into account the entire space made by all the possible combinations of values of the parameters. As we have seen, two methods that successfully aim to this are the Morris' method and the variance-based approach.

Since the beginning, during our analyses, we highlighted how *ordinary differential equations* (ODE) integrations cannot model the uncertainty in the time evolution of the system close to bifurcation points. Using the more informative set of stochastic simulations we showed that in this situation the choice is between two possible evolutions, but this cannot be interpreted by ODEs, that for definition can compute only one

of them for a set of possible parameters. For this reason we have considered models with bifurcation points as those models that may be difficult to be analysed using the classical SA approaches, that usually assume the time evolution of the system to be computed using ODEs. Because of the capability of stochastic simulations of capturing these fenomena, we pursued the development of SA techniques that make use of them.

Another and maybe the most important result is that we have demonstrated that the proposed novel techniques lead to more precise results with respect to the classical ones, at least in bistable systems with parameters set close to a bifurcation point. We indeed proved that there is a need for sensitivity analysis measures that take into account the variations between sets of stochastic simulations rather than the simple output of ODE integration. We are confident in stating that the techniques we presented are a good starting point in this direction. Moreover, the application to a real example, the MAPK signaling pathway, validated our hypotheses, giving also a more complete and wider view about the effectiveness of developing sensitivity measures that are specific and that can overcome the limitations of the present ones.

Finally, a fact that shouldn't be underestimated is that the method we analysed in this work, both the classical and the novel ones, are intuitive and the measure that they compute or estimate have a clear and intuitive meaning. This is an advantage with respect to other proposed sensitivity measures like the ones based on *Fisher information matrix* (FIM) that still need a precise interpretation.

5.1 Future Work

We consider our work as a first step, a glimpse into the creation of a branch of sensitivity analysis that is dedicated to analyse models whose time evolution is computed with set of stochastic simulations. The way that has been followed was the adaptation of present techniques and we believe that other techniques can be adapted as well. However, many methods make use of analytical simplifications whose validity may be difficult if not impossible to be proved when passing from ODEs to sets of stochastic simulations.

As we have seen throughout the dissertation, the main issue has been the computational complexity of the methods, specially when using stochastic simulations. For large models, with many species and many factors also using ODEs may lead to a time of the analsis extremely long and unaffordable. Moreover, in order to have accurate

estimations of the density distance between sets of stochastic simulations, often an elevated number of simulations is required (usually 10000). For this reason it is important to work on decreasing the computational complexity of these methods. A first idea is to use approximated and faster versions of the stochastic simulation algorithm (SSA) of Gillespie. Doing this, it would be important to verify whether introducing approximations in the computation of the evolution of the system intoduces errors also in the sensitivity measures and whether these errors are negligible.

Appendix A

Extending Dizzy with a tool for sensitivity analysis

A.1 Dizzy: a tool for modeling biochemical interactions

This section introduces Dizzy [17], a software tool for modeling biochemical interactions, either with general mass action ODEs or stochastic simulations.

High-level description languages. Citing from the Dizzy user manual [18]: "Dizzy is a chemical kinetics simulation software package implemented in Java. It provides a model definition environment and various simulation engines for evolving a dynamical model from specified initial data." In Dizzy a text editor is present, where it is possible to describe a model with an high level language called *Chemical Model Definition Language* (CMDL). This allows us to have a model definition that is independent of the method that will be used in a second stage to model the time evolution. Figure A.1 shows an example of CMDL for the Michaelis model of enzymatic reaction that we introduced in Section 3.2.2. CMDL can also be converted to *Systems Biology Markup Language* [2] (SBML), the most widely used high-level description language for biological models, and a SBML file can be imported in Dizzy as well.

Simulation engines. The simulation engines implemented in Dizzy are either stochastic or deterministic: the SSA in its original version and other versions (like LDM) that improve the computational complexity of the algorithm; approximate versions of the SSA, that improve the complexity but lose the characteristic of being exact Monte Carlo simulations of the CME (we will not discuss these here, because we do not use them in this dissertation); deterministic (ODE-based) algorithms for simulating chemical reaction kinetics, like the 5th-order Runge-Kutta algorithm with fixed or

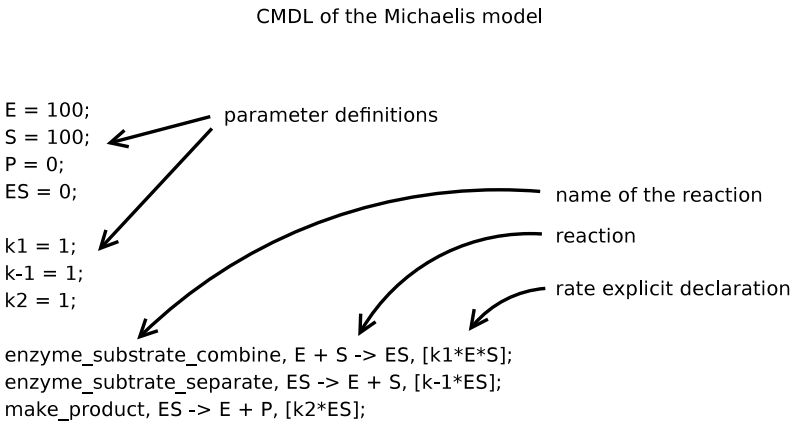

Figure A.1: Example of a CMDL file: the Michaelis/Menten model. Here the reaction rate is computed explicitly, because it is surrounded by [·]. Without the square brackets, Dizzy would have interpreted the value as the kinetic constant or the stochastic rate constant, depending on the selected simulator (deterministic or stochastic). In that case, the reaction rate would have been computed automatically by Dizzy.

adaptive step-size and the 5/4 Dormand-Prince ODE solver with adaptive step-size.

A.2 The extension

In this section we briefly present the part of the Dizzy tool that has been implemented. Here we show and comment some screenshots of the tool.

In Figure A.2, the already existing interface of the simulator is shown. In this panel it is possible to select the algorithm to be used to compute the time evolution of the system. Parameters that are generic (such us the end time of the analysis) and parameters that are specific of one method (such us the number of stochastic simulations for the stochastic simulators) can be tuned by using this interface. On the bottom, a list of outputs of the simulation is available as a set of checkboxes. As a little extension we introduced the checkbox “Histogram” that produces, at the end of a set of stochastic simulations, an histogram like the one shown in Figure A.3. This histogram groups the outputs of the simulations at the last time point in forty columns.

Playing with the simulator is extremely useful before using the sensitivity analysis tools. This helps the modeler to become familiar with the model, understanding which time points are the most interesting and having a first idea of the computational time of the single simulations.

The sensitivity analysis tools are available under the menu “Tools” and then “Anal-

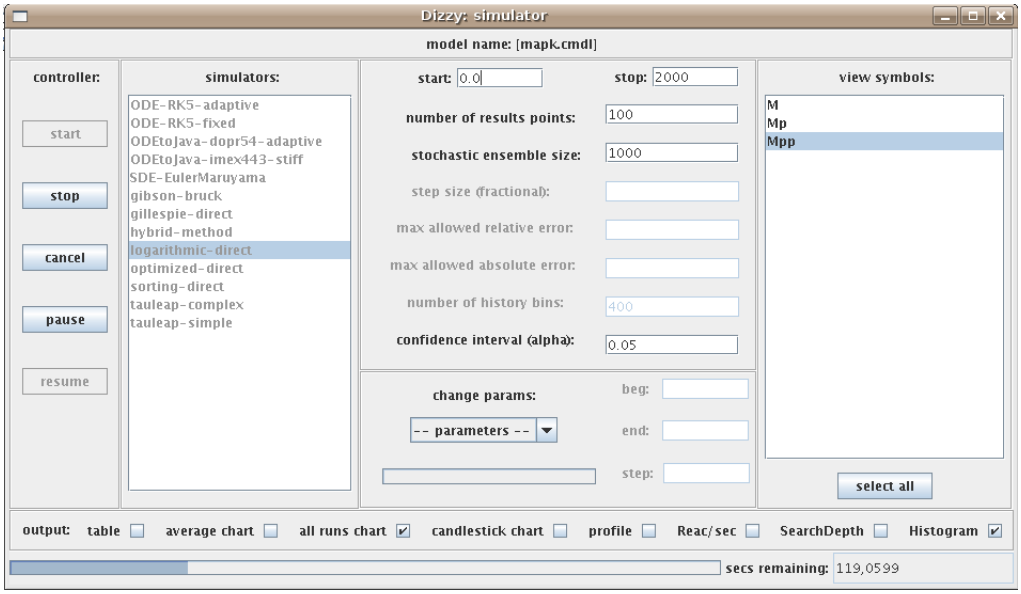

Figure A.2: Simulator interface of the Dizzy tool.

ysis". Three different panels have been created, one for each technique implemented: local one-at-a-time, screening with Morris' method and variance-based. However, the structure of these panels is the basically the same, so we can show just one of them (see Figure A.4). On the left of the panel, inputs and outputs are listed and can be selected. This way, it is possible to limit the analysis only on those factors that one is interested in. The central part of the interface is the one dedicated to the tuning of the parameters, either of the simulators or of the analysis itself. A combobox allows the choice between the different versions of the sensitivity analysis, that can be based on output difference of ODEs or average of stochastic simulations or on histogram distance of sets of stochastic simulations. For the three versions we used the same names introduced in Section 3.1.2, that are ODEs, Gillespie average and Gillespie density. On the right of the panel, informative labels and control buttons are present. Once the analysis starts, the time at which it began is shown, as long as an estimation of the time necessary to conclude, based on the first simulation that is performed.

Once the analysis is concluded, the sensitivity indices and other information specific of the analysis are shown in a table such as the one in Figure A.5. This is a sortable table, i.e. it allows to sort the rows just clicking on the title of the column that one wants to use as the criteria of sorting. Moreover, notes about the analysis, such as histogram self distances or total sensitivity indices, are shown in the space right below the table.

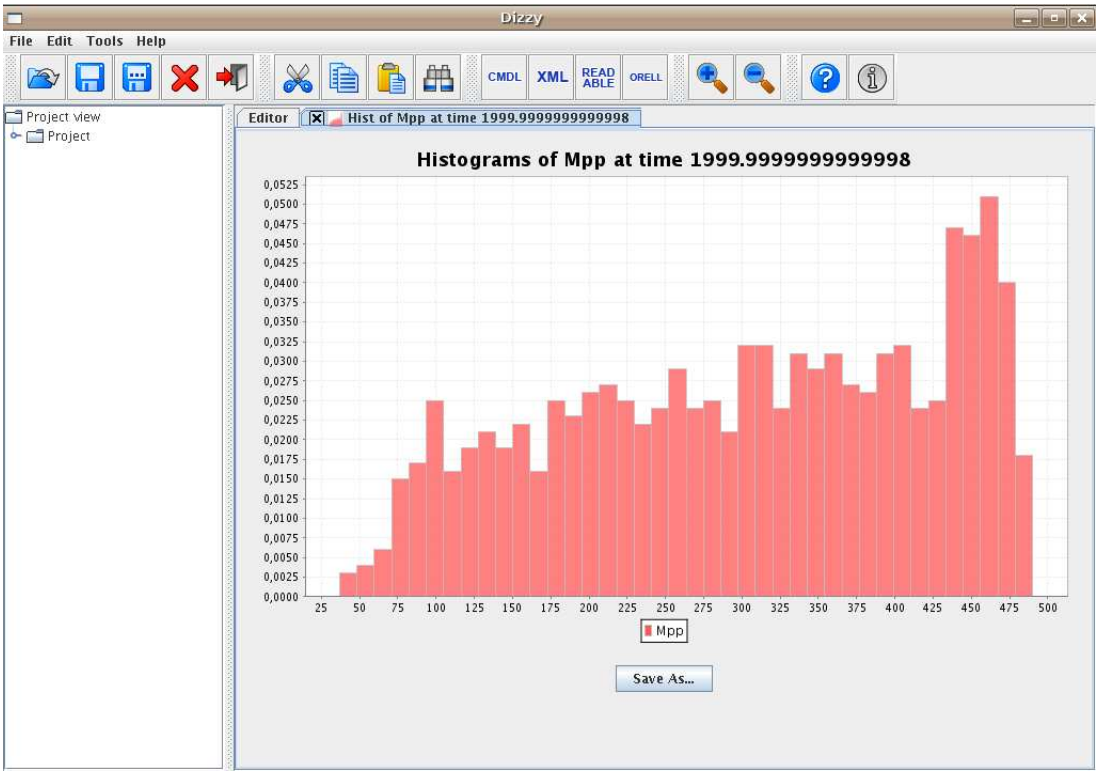

Figure A.3: Histogram generated when using the checkbox “Histogram” in the simulator interface, in Figure A.2.

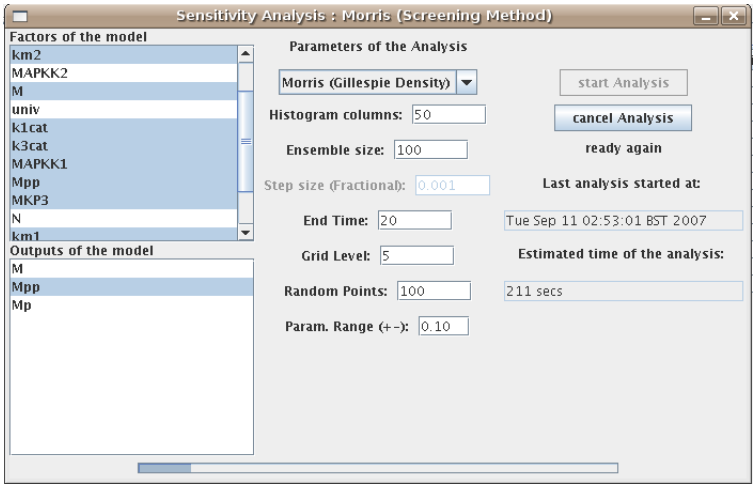

Figure A.4: Example of a panel for a sensitivity analysis.

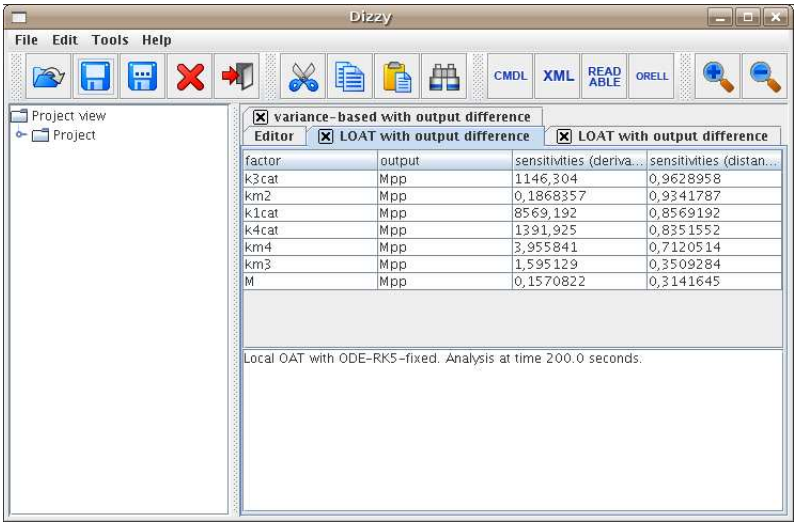

Figure A.5: Table with sensitivity indices, generate by the analysis tool.

Appendix B

Models written in Chemical Model Definition Language (CMDL)

Michaelis model

```
#model "michaelis";

E = 100;
S = 100;
P = 0;
ES = 0;

enzyme_substrate_combine,    E + S -> ES,      1;
enzyme_substrate_separate,   ES    -> E + S,   1;
make_product,                ES    -> E + P,   1;
```

Schlögl model

```
#model myschlogl;

X=247;
k1A=0.03;
k2=0.0001;
k3B=200;
k4=3.5;
```

```
a1, -> X, [k1A*X*(X-1)/2];
a2, X -> , [k2*X*(X-1)*(X-2)/6];
a3, -> X, [k3B];
a4, X -> , [k4*X];
```

MAPK model

```
#model "mapk";

N = 1;
Mp = 0*N;
M = 200*N;
Mpp = 300*N;
MAPKK1 = 50*N;
MAPKK2 = 1*N;
MKP3 = 100*N;

k1cat = 0.01;
km1 = 50;
k2cat = 15;
km2 = 500;
k3cat = 0.084;
km3 = 22;
k4cat = 0.06;
km4 = 18;
km5 = 78;

v1, M -> Mp, [(k1cat*MAPKK1*M/km1)/(1+M/km1+Mp/km2)];
v2, Mp -> Mpp, [(k2cat*MAPKK1*Mp/km2)/(1+M/km1+Mp/km2)];
v3, Mpp -> Mp, [(k3cat*MKP3*Mpp/km3)/(1+Mpp/km3+Mp/km4+M/km5)];
v4, Mp -> M, [(k4cat*MKP3*Mp/km4)/(1+Mpp/km3+Mp/km4+M/km5)];
```

Bibliography

[1] Matlab. *http://www.mathworks.com/products/matlab/*, 2007.

[2] Systems biology markup language (sbml). *www.sbml.org*, 2007.

[3] Systems biology toolbox. *http://www.sbtoolbox.org/*, 2007.

[4] G. E. Briggs and J. B. S. Haldane. A note on the kinetics of enzyme action. *Biochem. J.*, 19:338–339, 1925.

[5] Yang Cao and Linda Petzold. Accuracy limitations and the measurement of errors in the stochastic simulation of chemically reacting systems. *J. Comput. Phys.*, 212(1):6–24, 2006.

[6] Karen Chan, Andrea Saltelli, and Stefano Tarantola. Sensitivity analysis of model output: Variance-based methods make the difference. *Proceedings of the 1997 Winter Simulation Conference*, pages 261–268, 1997.

[7] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annu Rev Phys Chem*, 58:35–55, 2007.

[8] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, December 1976.

[9] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, December 1977.

[10] C. M. Guldberg and P. Waage. Über die chemische affinität. *J. Prakt. Chem.*, 19:69, 1879.

[11] R. Gunawan, Y. Cao, L. Petzold, and F. J. Doyle. Sensitivity analysis of discrete stochastic systems. *Biophys J*, 88(4):2530–2540, April 2005.

[12] Kitano. Computational systems biology. *Nature*, 2002.

[13] E. Klipp, R. Herwig, A. Kowald, C. Wierling, and H. Lehrach, editors. *Systems Biology in practice*. Wiley-Vch, 2005.

[14] Hong Li and Linda Petzold. Logarithmic direct method for discrete stochastic simulation of chemically reacting systems. *Department of Computer Science, University of California Santa Barbara, Santa Barbara*, 2006.

[15] Nick I. Markevich, Jan B. Hoek, and Boris N. Kholodenko. Signaling switches and bistability arising from multisite phosphorylation in protein kinase cascades. *The Journal of Cell Biology*, 164:353–359, 2004.

[16] David L. Nelson and Michael M. Cox. *Lehninger Principles of Biochemistry, Fourth Edition*. W. H. Freeman, April 2004.

[17] S. Ramsey, D. Orrell, and H. Bolouri. Dizzy: stochastic simulation of large-scale genetic regulatory networks. *J Bioinform Comput Biol*, 3(2):415–436, April 2005.

[18] Ramsey S. Dizzy user manual. *CompBio Group, Institute for SystemsBiology*, 2006. http://magnet.systemsbiology.net/software/Dizzy/.

[19] A. Saltelli, S. Tarantola, and F. Campolongo. Sensitivity analysis as an ingredient of modeling. *Stat Sci*, 15:377–395, 2000.

[20] Andrea Saltelli, Karen Chan, and Evelyn Marian Scott, editors. *Sensitivity Analysis*. Wiley, 2000.

[21] Andrea Saltelli, Marco Ratto, Tarantola Stefano, and Camplongo Francesca. Sensitivity analysis for chemical models. *Chem Rev*, 105(7):2811–2828, 2005.

[22] O. Wolkenhauer, M. Ullah, W. Kolch, and K. Cho. Modelling and simulation of intracellular dynamics: Choosing an appropriate framework. *NanoBioscience*, 3:200–207, 2004.

[23] Olaf Wolkenhauer and Mihajlo Mesarovic. Feedback dynamics and cell function: Why systems biology is called systems biology. *Molecular Biosystems*, 1:14–16, 2005.