Statistical Analysis and Comparative Visualization of Cellular Particle-based Simulations

Ana Cristina Pintilie

Course of studies: INFOTECH
Examiner: Prof. Dr. Thomas Ertl
Supervisor: Dipl.-Inf. Martin Falk
Starting date: 2 November 2012
End date: 4 May 2013
CR Classifications: J.2, I.3.4
Abstract

The present work integrates a set of functions, properties and views to the CellVis framework developed by Martin Falk in order to lay the foundations for a statistical analysis and comparative analysis/visualization module.

The CellVis framework employs a stochastic simulation to model and analyze cellular signal transduction. The cell is represented as a sphere consisting of a nucleus, proteins, obstacles and cytoskeleton filaments. The particles can move inside the cell by diffusion or by direct transportation along the cytoskeleton filaments.

The statistical analysis is performed by estimating the first four statistical moments of the sample spatial distribution of the molecules involved in the analyzed simulation. These moments can be visualized in a tree-like view and also with help of a set of time series plots, that illustrate the evolution of these moments through the lifetime of the simulation. The spatial distribution of the concerned particles is separated on the three spatial dimensions. This is possible due to the fact that the motions of one particle on each axis is independent with respect to the other ones. A histogram is also calculated and graphically represented in a 2D plot in order to facilitate understanding the meaning of the estimated sample moments values. To justify the use of moments describing the spatial distribution, an external tool is used to estimate a probability density function that fits the simulated spatial distribution with respect to the estimated sample moments. The results are satisfactory with respect to the matching of the calculated histogram and the histogram of the fitted probability density function. So we could conclude that for typical types of distributions the use of moments can help in finding a suitable probability density function that can be used later on for predicting the spatial distribution for certain types of particles in other simulations.

For a more particular type of analysis of the signal transduction process, the trajectory and reaction history of an individual particle are calculated and determined. To visualize them, 2D graphs are employed and integrated in the framework. The trajectory of a particle is plotted as the distance to the nucleus over time. The reactions and interactions are indicated with markers on the same graph.
The comparative visualization is performed by employing difference plots and difference volume. For this, the framework is extended with the loading of a second data set. The difference graphs are used to show differences in the statistical sample moments, in the concentrations and in the histograms of the two loaded data sets. The difference volume is performed with respect to the differences of concentrations in the considered data sets. The view of the difference volume can be enabled by the user. However the option of viewing both data sets at the same time remains as a possibility of future work.
# Contents

<table>
<thead>
<tr>
<th>Contents</th>
<th>v</th>
</tr>
</thead>
<tbody>
<tr>
<td>List of Figures</td>
<td>vii</td>
</tr>
<tr>
<td>Acronyms</td>
<td>ix</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Motivation and Scope</td>
<td>2</td>
</tr>
<tr>
<td>1.2 Contribution and Aims</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Thesis Outline</td>
<td>4</td>
</tr>
<tr>
<td>2 Biological Fundamentals</td>
<td>7</td>
</tr>
<tr>
<td>2.1 Cellular Architecture</td>
<td>7</td>
</tr>
<tr>
<td>2.2 Signal Transduction</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 MAPK-Cascade</td>
<td>9</td>
</tr>
<tr>
<td>2.3 Transport</td>
<td>10</td>
</tr>
<tr>
<td>2.4 Reactions</td>
<td>10</td>
</tr>
<tr>
<td>2.4.1 First-Order Reactions</td>
<td>11</td>
</tr>
<tr>
<td>2.4.2 Second-Order Reactions</td>
<td>11</td>
</tr>
<tr>
<td>2.4.3 Enzymatic Reactions</td>
<td>12</td>
</tr>
<tr>
<td>3 Moment-Based Statistical Analysis</td>
<td>13</td>
</tr>
<tr>
<td>3.1 First-order moment: Mean</td>
<td>14</td>
</tr>
<tr>
<td>3.2 Second-order moment: Variance</td>
<td>15</td>
</tr>
<tr>
<td>3.3 Third-order moment: Skewness</td>
<td>16</td>
</tr>
<tr>
<td>3.4 Forth order moment: Kurtosis</td>
<td>17</td>
</tr>
<tr>
<td>3.5 Correlation and Covariance</td>
<td>18</td>
</tr>
<tr>
<td>3.6 Results</td>
<td>20</td>
</tr>
<tr>
<td>4 History of Individual Molecules</td>
<td>29</td>
</tr>
<tr>
<td>4.1 Trajectory of Single Molecules</td>
<td>29</td>
</tr>
<tr>
<td>4.2 Identification of Reactions and Reaction Partners</td>
<td>30</td>
</tr>
<tr>
<td>4.3 Results</td>
<td>31</td>
</tr>
</tbody>
</table>
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Eukaryotic cell</td>
<td>7</td>
</tr>
<tr>
<td>2.2</td>
<td>Diagram of the basic structures of an actin filament</td>
<td>8</td>
</tr>
<tr>
<td>2.3</td>
<td>Signal transduction process</td>
<td>9</td>
</tr>
<tr>
<td>3.1</td>
<td>Mean</td>
<td>15</td>
</tr>
<tr>
<td>3.2</td>
<td>Standard deviation</td>
<td>16</td>
</tr>
<tr>
<td>3.3</td>
<td>Skewness</td>
<td>17</td>
</tr>
<tr>
<td>3.4</td>
<td>Kurtosis</td>
<td>18</td>
</tr>
<tr>
<td>3.5</td>
<td>Tree view of statistical parameters</td>
<td>21</td>
</tr>
<tr>
<td>3.6</td>
<td>Time series plot of the first four moments</td>
<td>22</td>
</tr>
<tr>
<td>3.7</td>
<td>Correlation coefficients</td>
<td>22</td>
</tr>
<tr>
<td>3.8</td>
<td>Test configuration</td>
<td>23</td>
</tr>
<tr>
<td>3.9</td>
<td>Histogram first time step</td>
<td>24</td>
</tr>
<tr>
<td>3.10</td>
<td>Histogram last time step</td>
<td>25</td>
</tr>
<tr>
<td>3.11</td>
<td>Distribution fitting for molecule of type mol2</td>
<td>26</td>
</tr>
<tr>
<td>3.12</td>
<td>Distribution fitting for molecule of type Receptor A</td>
<td>27</td>
</tr>
<tr>
<td>3.13</td>
<td>Distribution fitting for molecule of type Mol B Skeleton</td>
<td>27</td>
</tr>
<tr>
<td>4.1</td>
<td>Trajectory view</td>
<td>32</td>
</tr>
<tr>
<td>4.2</td>
<td>Trajectory of type mol3nucleus</td>
<td>32</td>
</tr>
<tr>
<td>4.3</td>
<td>Reaction view of type mol3nucleus</td>
<td>33</td>
</tr>
<tr>
<td>4.4</td>
<td>MAPK simulation configuration</td>
<td>33</td>
</tr>
<tr>
<td>4.5</td>
<td>Particle reaction history-MAPK</td>
<td>34</td>
</tr>
<tr>
<td>4.6</td>
<td>Particle reaction history-MAPKp</td>
<td>35</td>
</tr>
<tr>
<td>4.7</td>
<td>Receptor history</td>
<td>35</td>
</tr>
<tr>
<td>5.1</td>
<td>Radial concentration profile</td>
<td>40</td>
</tr>
<tr>
<td>5.2</td>
<td>Concentration and histogram difference plot-case I</td>
<td>41</td>
</tr>
<tr>
<td>5.3</td>
<td>Concentration and Histogram difference plot-case II</td>
<td>42</td>
</tr>
<tr>
<td>5.4</td>
<td>Concentration and histogram difference plot-case III first time step</td>
<td>43</td>
</tr>
<tr>
<td>5.5</td>
<td>Concentration and histogram difference plot-case III last time step</td>
<td>44</td>
</tr>
<tr>
<td>5.6</td>
<td>Mean and standard deviation difference plot-case I</td>
<td>45</td>
</tr>
<tr>
<td>Section</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.7</td>
<td>Skewness and kurtosis difference plot-case I</td>
<td>46</td>
</tr>
<tr>
<td>5.8</td>
<td>Correlation coefficients difference plot-case I</td>
<td>46</td>
</tr>
<tr>
<td>5.9</td>
<td>Mean and standard deviation difference plot-case II</td>
<td>47</td>
</tr>
<tr>
<td>5.10</td>
<td>Skewness and kurtosis difference plot-case II</td>
<td>48</td>
</tr>
<tr>
<td>5.11</td>
<td>Correlation coefficients difference plot-case II</td>
<td>49</td>
</tr>
<tr>
<td>5.12</td>
<td>(a) Mean and (b) standard deviation difference plot-case III</td>
<td>50</td>
</tr>
<tr>
<td>5.13</td>
<td>(a) Skewness and (b) kurtosis difference plot-case III</td>
<td>50</td>
</tr>
<tr>
<td>5.14</td>
<td>Correlation coefficients difference plot-case III</td>
<td>51</td>
</tr>
<tr>
<td>5.15</td>
<td>Color map</td>
<td>52</td>
</tr>
<tr>
<td>5.16</td>
<td>Concentration difference volume-case I</td>
<td>52</td>
</tr>
<tr>
<td>5.17</td>
<td>Concentration difference volume-case II</td>
<td>53</td>
</tr>
<tr>
<td>5.18</td>
<td>Concentration difference volume-case III</td>
<td>54</td>
</tr>
<tr>
<td>5.19</td>
<td>Concentration difference volume-different density scales</td>
<td>54</td>
</tr>
</tbody>
</table>
Acronyms

**DNA** Deoxyribonucleic Acid  
**DVR** Direct Volume Rendering  
**MAPK** Mitogen-Activated Protein Kinase  
**MAPKp** phosphorylated Mitogen-Activated Protein Kinase  
**MCA** Metabolic Control Analysis  
**SVG** Scalable Vector Graphics
Chapter 1

Introduction

Systems biology is concerned with the study of biological functions and mechanisms. A biological system of interacting entities is analyzed as a whole rather than analyzing each individual entity separately. Systems biology aims to provide means of understanding biological processes on a large scale, like the association of gene with function, the detailed modeling of the interaction among proteins and metabolites and the function of cells. It has a wide-range of applications, as it involves a number of underlying disciplines, including biology, computer science, mathematics, physics, chemistry and the social sciences.

Analyzing the individual functions of genes, proteins or other cellular components is, however, not enough to capture the complexity of a cell, its functions and organisms. The whole system needs to be analyzed as an ensemble. For this we need to determine its organization and its controlling mechanism by investigating the dynamic of genes and proteins, together with the interactions between them. The group of molecules that work together to control the cell functions, such as cell division or death represent a signaling pathway. A signaling pathway consists of a series of chemical reactions that enable passing the external signal from one molecule to another until the signal reaches the target. Signal transduction is the process by which cells communicate with each other and their environment and involves a multitude of proteins that can be in active or inactive states. In their active (phosphorylated) state they act as catalysts for the activation of subsequent steps in the signaling cascade. The result is the activation of a transcription factor which in turn initiates a gene transcription event. The ordered sequence of the reactions in this process is known as the signal transduction pathway. These pathways exhibit a high dynamic complexity and an adaptive behavior.

The general goal of theoretical systems biology is therefore to develop computer models that predict the properties of the large, adaptive, intercon-
1.1. Motivation and Scope

1.1.1. Motivation and Scope

The motivation for this thesis resulted from the need to reduce the complexity of the spatial-temporal data of the particle-based simulations used in systems biology. The intracellular architecture, particle features, the exact molecule count and the molecules precise position can be visualized directly from the simulation, e.g., with glyphs [2]. However, only the spatial distribution at a given time can be easily observed. In order to better describe the spatial distribution at a given time step and to see how this distribution evolves through time, it is necessary to find and estimate a set of parameters that are capable of quantifying the spatial distribution. Once a set of parameters is chosen, their evolution throughout the entire simulation can be easily viewed with a time series plot, thus reducing the spatial-temporal data of the crowded environment of the cell to a simple time series plot where the distribution tendencies of the different types of molecules can be more clearly shown.

Once we have a set of parameters that are able to roughly describe the spatial distribution of the different molecule types involved in one simulation, the idea of a comparative visualization of several simulations comes...
naturally. The simulations might simply differ by the initial random seeds or might employ different parameter sets. For the comparative visualization we can consider either two data sets that are shown side-by-side or several data sets are combined and the differences are highlighted. Though the set of parameters that describe the evolution and the outcome of one simulation can be quite large, the comparison between two simulation runs has to be done by considering those parameters that can actually provide a meaningful information. Since the possibilities for comparison are plenty, this work proposes a comparison based on concentration and statistical sample moments.

The diffusion process of one specific molecule and the reactions in which the molecule is involved are also hard to spot [2]. Therefore the utility of visualizing the path that one molecule follows arises. The trajectory of one molecule can be very important since the main interest is to determine if and how many of the molecules of interest reach their target, in this case the nucleus. Having a view of the several molecule paths provides the possibility to analyze the rate of success in reaching the target and also to understand the role of the motor proteins in propagating the signal. Besides the interest in whether or not one molecule reaches the target, the determination of the other molecules that may interact or react with the selected molecule might also be of interest. Developing an approach that determines when reactions take place and which particles are involved provide the means to estimate a reaction distribution and to estimate the approximate region where different reactions are more likely to take place.

Considering the motivation of developing a module capable of providing a number of statistical parameters in order to simplify the analysis of the simulation data by experts, the scope of this work is to implement and integrate this module in the existing framework CellVis of Martin Falk. The main goals of the thesis are:

- estimation of a set of statistical parameters that describes the spatial distribution of certain molecule types and provide a simple overview of their time evolution
- trajectory visualization of individual molecule, with the possibility to include the potential reaction partners
- developing a visualization for the comparison of two simulation runs

1.2 Contribution and Aims

This thesis lies the basis for a statistical analysis module. It provides the means to analyze the data sets resulting from the simulation with a set of estimated parameters specific to a spatial distribution, in this case: mean,
variance, skewness, kurtosis and correlation coefficients. These parameters represent the basic parameters that come to mind when describing a distribution. But in order to have a deeper understanding of the processes that take place inside the cell other functionalities have to be added to the statistical module, e.g.: determining clusters of molecules and their evolution in time focusing on their movement and size development.

Considering the statistic data involving a single molecule, a trajectory and a reaction history view is integrated. The possibility to track and analyze the activity of a molecule can be extended to track certain types of molecules of interest and extract other statistical information like how many of these molecules reach their target or what is life expectancy of a certain type of molecule. The algorithm for determining the reactions can be used also to provide information about the reaction distribution or to compare the number of reactions at one time point with the theoretical reaction rate for validation purposes.

The CellVis framework was mainly designed to deal just with one data set. In this thesis a functionality that enables loading information about another data set is implemented. At the time being, only a small part of the information that can be extracted from a second data set is used and the visualization is based on differences of the two simulations. For the future work, a complete module of loading a second data set could be included so that two data sets can be viewed at the same time together with a visualization of highlighted differences.

1.3 Thesis Outline

Chapter 2 provides a brief introduction in certain biological aspects and is supposed to give an insight on the biological terms, processes and systems that are taken in consideration in this work.

Chapter 3 describes the statistical methods that were implemented in order to have better understanding not only of the outcome of the simulation but also throughout its entire evolution and how these parameters vary from one time step to another. The concept of a statistical moment is presented together with the problematic of the method of moments used to estimate a probability density function. The estimated sample moments resulting from the simulation are illustrated and discussed in the results section.

The intracellular processes do not involve only the movement of molecules but also different type of reactions and interactions. Thus Chapter 4 presents an approach for visualizing the trajectory of one molecule and also its reaction history. The history does not only include the potential reaction partners of the molecule in question but also the reaction partners of the reaction partners. Thus we have a complete description of the reaction chain involving the molecule of interest.
CHAPTER 1. INTRODUCTION

Though analyzing one data set might be useful, sometimes the information that we obtain from one data set is not enough and it needs to be put in perspective and analyzed together with other data sets. In Chapter 5 the basics of a comparative visualization approach are introduced and the possibility of loading a second data set is added to CellVis in order to view a volume in the difference of molecule concentrations.

The last chapter summarizes the presented work, presenting the resulting conclusions and outlining the possibilities of future development of the statistical analysis module.
Chapter 2

Biological Fundamentals

In order to better understand the role of this thesis, of the implemented statistical methods and of the few simulation aspects that were analyzed, it is important to have a clear description of the biological cell architecture and of the intracellular processes that are considered in this work.

2.1 Cellular Architecture

From the biological point of view cells are of two types: prokaryotic and eukaryotic. Prokaryotes are a type of organisms enclosed by a plasma membrane but lack a membrane-bound nucleus. Eukaryotes, unlike prokaryotic cells, contain a nucleus surrounded by double membrane in which the genetic material is stored. Besides having a nucleus the eukaryotic cells also contain other enclosed compartments called organelles. The organelles of an eukaryotic cell can be seen in Figure 2.1. Since CellVis framework is a tool designed to simulate and visualize signal transduction processes in cells that have a well defined nucleus, the cell architecture of interest in this case will be the one of the eukaryotic cells.

![Figure 2.1: Eukaryotic cell][3]
2.1. Cellular Architecture

The nucleus and mitochondrion are bounded by a double membrane having an intermembrane space as a separator. All other organelles are surrounded by a single membrane. The nucleus can be considered the “brain” of the cell, because it manages and directs the cell activities and it contains the genetic material called chromosomes made out of DNA. Mitochondria are responsible for producing energy from food; they represent also the site of cellular respiration. The lysosome breaks down worn-out or unneeded cellular components and some ingested materials with the help of digestive enzymes. The Golgi apparatus is in charge of modifying, sorting and packaging proteins. The endoplasmic reticulum represents an internal membrane system where certain types of proteins are assembled with the help of ribosomes. The plasma membrane acts as boundary between the cell and its environment and regulates cell’s absorption and diffusion of certain materials.

Between the plasma membrane and the nucleus lies the cytoplasm that comprises all the other organelles and the cytosol consisting of an aqueous medium. The main structure consisting the cytoplasm of the eukaryotic cells is the cytoskeleton. The cytoskeleton is a complex network of protein fibers that define the shape of the cell. The cytoskeleton also serves as a traffic network for the organelles in the cytoplasm and also most of them are attached to the cytoskeleton. This protein fiber network is consisted of three types of filaments that differ by size, type of subunit and the arrangement of these subunits (Figure 2.2) [3].

![Diagram of the basic structures of an actin filament (AF), intermediate filament (IF) and microtubule (MT)](image)

Figure 2.2: Diagram of the basic structures of an actin filament (AF), intermediate filament (IF) and microtubule (MT)[3].

Actin filaments (AF), also called microfilaments (7 - 8nm) have a twisted
two-stranded structure. The microtubules (MT) are hollow structures resembling a tube (24 - 25nm) and the intermediate filaments (IF) have a rope-like structure (10nm) [3, 4]. The cytoskeleton also plays an important role in the transport of signal molecules, as the different filaments that contain the cytoskeleton serve as a “highway” towards the target, e.g. the nucleus, for the proteins involved in the signal transduction [5].

2.2 Signal Transduction

The multitude of processes involved in how a cell converts an external stimulus into a specific response is referred to as signal transduction.

Signal transduction starts with an external signal to a receptor and results in a change of the cell status or of one its functionalities (Figure 2.3). This process is intermediated via a cascade of messenger signal molecules inside the cell. It begins with a signal to a receptor and ends with a change in cellular function which is typically mediated via a cascade of additional messenger signals within the cell.

There are two main ways of adjusting the state of the cell. One way is binding an external signal-molecule as a ligand to a membrane receptor. The receptor protein is activated by this binding and can further activate other signaling molecules that eventually carry the message to the desired target. The second way is when the external signal can itself move through the plasma membrane and to the intracellular receptors [4].

Figure 2.3: Signal transduction process [4]

2.2.1 MAPK-Cascade

Sequential activation of kinases is a common mechanism of signal transduction in many cellular processes. This work deals with the MAPK (mitogen-activated protein kinase) signaling cascades. The MAPKs are a group of
2.3 Transport

Protein serine/threonine kinases that are activated as a result of an external signal and mediate signal transduction process. The MAPK pathway can be generalized by the following sequence:

\[ \text{Stimulus} \rightarrow \text{MAPKKK} \rightarrow \text{MAPKK} \rightarrow \text{MAPK} \rightarrow \text{Response} \]

where MAPKK is the kinase of MAPK and MAPKKK is the kinase of MAPKK. The MAPKKK is activated by G proteins such as Ras, Rac and Rap1 or other enzymes that are activated by external signals [6, 7]. This cascade amplifies the transmitted signals that can eventually activate several regulatory molecules in the cytoplasm and in the nucleus to initiate a number of cellular processes such as proliferation, differentiation, survival, development and death [8]. Deviated or uncontrolled course of the MAPK signaling pathways is involved in the development of many human diseases including Alzheimer’s disease (AD), Parkinson’s disease (PD), amyotrophic lateral sclerosis (ALS) and various types of cancers [9, 10]. Thus these pathways are considered potential therapeutic targets.

2.3 Transport

In the crowded environment of the cell, molecules can move by diffusion which is characterized by a random movement on very short distances. The signaling cascade involves several reactions and dozens of proteins. Diffusion is a very slow transportation method for a reaction product to the next reaction partner along the pathway. To overcome this inefficiency in traveling through large spaces between different compartments of a cell, cells have developed a mechanism that cluster molecules in a common pathway together, assembling on a common “scaffold” [3].

Besides the undirected diffusion, some signaling cascades also involve the directed transportation along the cytoskeleton with the help of motor proteins. Most often the target of the signaling molecules is the nucleus, where these molecules can influence and trigger certain genes. The transport process in this case is a combination of undirected diffusion and linear transport along the cytoskeleton [11].

2.4 Reactions

The partitioning in organelles of the eukaryotic cell permits different kinds of biochemical reactions to take place. Although each organelle performs a specific function in the cell, together they form an ensemble and cooperate in order to meet the overall needs of the cell. For example, biochemical reactions in a cell’s mitochondria transfer energy from fatty
acids and pyruvate molecules into an energy-rich molecule called adenosine triphosphate (ATP) [3].

One or more particles can participate in a reaction, resulting in one or more products. For a reaction to take place it is necessary that the involved molecules are sufficiently close to each other. In addition they have to feature the correct orientation with respect to their position and that the energetic conditions are proper to initiate the reaction.

### 2.4.1 First-Order Reactions

A first-order reaction depends linearly only on the concentration of just one reactant (see eq. 2.4.1), even though other reactants might be present. This translates in a spontaneous decay or in a change in the particle’s state. An example for first-order reactions can be considered the deactivation or unbinding process.

$$\frac{dc_i}{dt} = -k_i c_i \quad (2.4.1)$$

where $k_i$ represents reaction rate constant and $c_i$ represents the concentration of molecule i and the reaction rate is defined by the reaction law as follows:

$$r_i = k_i c_i \quad (2.4.2)$$

The rate of the reactions can be measured in Ms$^{-1}$, where M represents moles per liter and s represents seconds, therefore the rate unit is molar per second. As the reactant is being depleted, the reaction rate decreases proportionally. The first-order reaction rate can be viewed also as a probability per unit of time and this probability is independent of the molecules concentration.

### 2.4.2 Second-Order Reactions

Second-order reaction, also called bimolecular reactions, involve two reactants. These two molecules collide with each other if they are close enough together. The rate of this reaction is determined by the reactants concentrations and by a reaction rate constant.

$$r_{ij} = k_{ij} c_i c_j \quad (2.4.3)$$

The value of the second-order reaction rate constant $k_{ij}$ is determined by the rate of the diffusion of the molecules. The rate of the diffusion depends on the size and shape of the molecules and the temperature and viscosity of the medium. The influence of these factors is summed up in the diffusion coefficient $D$, with unit of $m^2 s^{-1}$. The diffusion coefficient represents a measure that describes how a particle moves in its medium. Therefore, the reaction
rate constant can be defined by the diffusion coefficient and the area of interaction between molecules and is described by the Debye-Smoluchowski equation [12]:

\[ k_{ij} = 4 \pi b (D_i + D_j) N_0 10^3 \]  \hspace{1cm} (2.4.4)

where \( b \) represents the interaction radius between the two molecules, measured in meters, \( D_i \) and \( D_j \) are the diffusion coefficients of the respective particles and \( N_0 \) represents Avogadro’s number. The scaling with \( 10^3 \) is done so that the unit is converted into \( M^{-1}s^{-1} \).

### 2.4.3 Enzymatic Reactions

An enzyme is a protein that speeds up the chemical reaction in a living organism. It acts as a catalyst for specific chemical reactions, converting a certain types of reactants called substrates into specific products.

Enzymatic reactions represents a special case of the second-order reactions, because the enzyme involved in the reaction has the role of both reactant and product. As mentioned before, the enzyme \( E \) converts the substrate \( S \) into the product \( P \) by forming an intermediate complex \( ES \).

\[ E + S \rightleftharpoons ES \rightarrow E + P \]  \hspace{1cm} (2.4.5)

The reaction rate constant can be determined using just two parameters: the unimolecular reaction rate constant \( k_{cat} \) and the Michaelis constant \( K_M \).

\[ k_{ES} = \frac{k_{cat}}{K_M + c_S} \]  \hspace{1cm} (2.4.6)

However, in this case \( k_{ES} \) is not a constant anymore, since it depends on the concentration of the substrate, Michaelis constant \( c_S \).
Chapter 3

Moment-Based Statistical Analysis

Statistical analysis is considered an essential research component that plays a vital role in a considerable number of research areas like: agriculture, biology, education, economics, management, engineering and psychology. Systems biology usually is dealing with large sets of data, whether they are result from simulations or from experiments. The raw numerical data is hard to analyze and the need to extract corresponding features and characteristics in order to put in evidence certain trends emerges. Statistical analysis refers to a set of methods that process large amounts of data with the purpose to reveal these trends, if any exist.

It is a common practice in statistical analysis to describe the shape of the distribution of a population by means of a finite set of quantities summarizing the location, dispersion, skewness, peakedness and tail behavior of the unknown population. Classical measures of distributional shape have been defined by means of algebraic moments of different orders, resulting in the mean for location estimation, the variance to measure the spread and the standardized measures of skewness and kurtosis to further reveal any asymmetries and levels of peakedness.

The Method of Moments [13] is one of the most common used methods of estimating a set of parameters that characterize a distribution. It was introduced by Karl Pearson in the 1880’s. From a known probability distribution it relatively straightforward to extract the respective sequence of moments, however describing a distribution with a known sequence of moments is a complex problem that it is still discussed by mathematicians and finding a corresponding probability density function is not always possible [14, 15].

Given a data set that is sufficiently large and has a tendency of clustering around a particular value, it could be useful to try to characterize its shape and spreadness by a set of parameters called moments of distribution. These
moments represent a quantization of the average of powers of the variables included in the considered data set.

The Kth moment \( m_k \) of a probability distribution \( X \) is defined:

\[
\begin{align*}
m_k &= E(X^k) \text{ for } k = 1, 2, 3, \ldots \tag{3.0.1}
\end{align*}
\]

where \( E() \) is the expectation operator. The Kth sample moments \( \hat{m}_k \) is estimated:

\[
\hat{m}_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \text{ for } k = 1, 2, 3, \ldots \tag{3.0.2}
\]

and the Kth centered sample moment or Kth sample moment \( \hat{m}_k' \) about the mean is:

\[
\hat{m}_k' = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{m}_1)^k \text{ for } k = 1, 2, 3, \ldots \tag{3.0.3}
\]

where \( X_1, X_2, \ldots, X_n \) represent a random sample of size \( n \) from the considered \( X \) distribution.

It is important to differentiate between dealing with a complete population and samples taken from it. When dealing with a complete population, the value of an individual moment is actually a constant and represents a parameter of the population’s distribution, but when dealing with a sample population an individual moment is considered a random variable and its value differs from sample to sample. Therefore the value of one sample moment represents an estimation of the respective moment of the complete population.

In the present case, the considered data set in represented by the positions of each type of particle present in the simulation. Though sometimes it is useful and necessary to analyze just one particle, most times it is better and easier to first look at the all particles of one type as a group.

In this thesis, the first four moments will be computed and analyzed so that the evolution of the spatial distribution of the inner cell particles can be quantized and described with a set of standardized parameters. The first two moments-mean and variance-measure the ability of a distribution to center around certain value. The third moment -skewness- is most useful to investigate when dealing with non-normal distributions, as it measures the asymmetry of the distribution and the forth moment -kurtosis- also used to describe non-normal distributions, as an indicator that helps determining the flatness of the distribution.

### 3.1 First-order moment: Mean

The first-order moment about zero is also called mean or expectation
CHAPTER 3. MOMENT-BASED STATISTICAL ANALYSIS

Figure 3.1: Mean of distribution

and it can be computed using the following formula:

\[ \hat{m}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i \]  

(3.1.1)

The mean is also considered the mass center because it indicates the balance point of the respective distribution. As it can be seen, the mean represents the average of all sample points of the given distribution.

3.2 Second-order moment: Variance

Variance represents the second-order moment about the mean:

\[ \hat{m}_2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{m}_1)^2 \]  

(3.2.1)

and it indicates the spread of the distribution, where \( \hat{m}_1 \) is the first order moment: mean. As it can be observed, the term \( (X_i - \hat{m}_1)^2 \) measures the deviation of \( X_i \) from the estimated sample mean, therefore the variance is also called the mean squared deviation. Another valuable indicator of the spreadness of the data is the standard deviation, which is calculated as the square root of the variance 3.2.2. The standard deviation is a measure of the dispersion about the mean and it is often used instead of the variance as an indicator of how much the data is deviated from the center.

\[ \sigma = \sqrt{\hat{m}_2} \]  

(3.2.2)
3.3 Third-order moment: Skewness

In order to obtain an unbiased estimator of the variance, the sample variance is best to be computed as follows:

\[
\hat{m}_2' = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \hat{m}_1')^2
\]  

(3.2.3)

An estimator is considered unbiased if the mean value of an estimator equals the true value of the quantity it estimates, in this case \(m_2\), the variance of the complete population. Therefore the squared differences are divided by \(n - 1\) and not by \(n\). However, when dealing with a large set of data, dividing by \(n - 1\) or by \(n\) doesn’t produce a significant difference.

3.3 Third-order moment: Skewness

The third sample moment about the mean is called skewness. While the mean, variance and standard deviation are dimensional quantities and share the same unit measure as the individuals from the populations, skewness is defined as adimensional, a pure number that describes the shape of the distribution.

\[
\hat{m}_3' = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i - \hat{m}_1'}{\hat{m}_2'} \right)^3
\]  

(3.3.1)

Skewness is an indicator of symmetry or more correct of asymmetry in the distribution. A symmetric distribution is characterized by a skewness equal to zero. A positive skewness indicates that the tail of the distribution
CHAPTER 3. MOMENT-BASED STATISTICAL ANALYSIS

tends more to the right side, as for a negative skewness which indicates a tailing more to the left side (Figure 3.3). In practice, it is almost impossible to have an exact zero skewness, therefore small deviations from zero of the skewness do not actually reflect any asymmetry in the distribution of the data set.

![Figure 3.3: Skewed distribution (a) to the right (b) to the left](image)

Regarding the sample skewness, the same problem as with the sample variance arises. Considering the fact that the actual mean of the distribution is not known and we operate only with an estimator, the actual division should also be with \( n - 1 \).

3.4 Forth order moment: Kurtosis

The forth moment about the mean, called kurtosis, is a also a nondimensional quantity, the same as the skewness. Its value is a relative measure of flatness or peakedness of the distribution. It considered a relative measure, because the meaning of its value is always taken in consideration with respect to the one belonging to a normal distribution.

The sample kurtosis can be computed as follows:

\[
\hat{m}_4 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i - \hat{m}_1}{\sqrt{\hat{m}_2}} \right)^4
\]

(3.4.1)

In order to facilitate the analysis of the sample kurtosis, a term of \(-3\) is introduced as follows:

\[
\hat{m}_4 = \frac{1}{n} \sum_{i=1}^{n} \left( \frac{X_i - \hat{m}_1}{\sqrt{\hat{m}_2}} \right)^4 - 3
\]

(3.4.2)

so that for a normal distribution the sample kurtosis is approximately zero.
A distribution characterized by a positive sample kurtosis is called leptokurtic, as for a distribution with negative sample kurtosis is called platykurtic. The in-between distributions are referred to as mesokurtic.

From Figure 3.4 we can get an intuitive idea of how the value of kurtosis can be interpreted. On the top right corner of the figure the kurtosis values of several distributions are listed.

Figure 3.4: Distributions with different values for Kurtosis [16]. 1.D: Laplace distribution, also known as the double exponential distribution. 2.S: hyperbolic secant distribution. 3.L: logistic distribution. 4.N: normal distribution. 5.C: raised cosine distribution. 6.W: Wigner semicircle distribution. 7.U: uniform distribution.

It can be observed that a high positive kurtosis value for is an indicator of a steep peaked distribution, as for a negative value of kurtosis points to more flat distribution up to uniform distributions.

### 3.5 Correlation and Covariance

When dealing with a set of random variables, there is an interest in how these variables influence each other. Therefore the need of finding a measure that indicates the strengthness of the relationship between these variables. Covariance and correlation are two measures that suggest if there is a relationship between two or more random variables and the direction of this relationship. However these two indicators are not able to indicate the cause and effect of the relationship, just to reflect whether the random variables are independent or not and if not, how strong is their dependency.

Covariance is an unstandardized measure for the relationship between two random variables: X and Y. Covariance represents a generalization of
the variance, when more than one random variable is considered. Since we are working with a sample from the whole population, the sample covariance instead of the covariance will be taken into consideration.

\[ \text{Cov}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \]  

(3.5.1)

where \( \bar{X} \) and \( \bar{Y} \) represent the estimated sample mean of \( X \) and \( Y \). Please note that the division by \( n - 1 \) can be used for the same reasons explained in the previous sections. The covariance is an unstandardized measure because it depends on the units in which the two random variables are measured and so analyzing the strength of the \( X \) and \( Y \) relationship by the value of the covariance becomes a subjective process and is different from experiment to experiment. Thus the notion of correlation and correlation coefficients emerged. Correlation is a standardized measure and it is actually a form of covariance normalized by the standard deviation of each of the random variables. And so, regardless of the unit type of measurement, the correlation coefficients will have the same type of unit. The definition of the correlation coefficient between \( X \) and \( Y \) is:

\[ r_{XY} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \]  

(3.5.2)

where \( \sigma_X \) and \( \sigma_Y \) are the respective standard deviations of \( X \) and \( Y \). Correlation coefficients are therefore invariant to measure units and are actually bounded:

\[-1 \leq r_{XY} \leq 1 \]  

(3.5.3)

If the correlation coefficient is zero, it means that there is a very weak or nonexistent relationship between \( X \) and \( Y \). Since the correlation coefficients only detect linear dependencies between \( X \) and \( Y \), it is not safe to say that these two variables are independent if their correlation coefficient is null. However in the reverse case, if it is known that the two variables are independent then we can consider the correlation coefficients equal to 0.

If \( |r_{XY}| \) is close or equal to 1, it means that between \( X \) and \( Y \) exists a strong dependency and we can say that \( X \) and \( Y \) a strongly correlated. A negative correlation coefficient indicates a negative linear dependency, for example if \( X \) increases then \( Y \) decreases. This dependency is also called anti-correlation. As for a positive correlation coefficient, \( X \) and \( Y \) tend to increase together.

Covariance and correlation coefficients are symmetric, meaning:
\begin{equation}
\text{Cov}(X,Y) = \text{Cov}(Y,X) \quad (3.5.4)
\end{equation}

\[ r_{XY} = r_{YX} \]

When working with more than two random variables it is useful to determine the variance-covariance matrix, in order to have one data structure that is able to characterize the relationship between the considered variables. If we assume three random variables: X, Y and Z, their variance-covariance matrix $C$ looks as following:

\[
C = \begin{pmatrix}
\text{Cov}(X,X) & \text{Cov}(X,Y) & \text{Cov}(X,Z) \\
\text{Cov}(Y,X) & \text{Cov}(Y,Y) & \text{Cov}(Y,Z) \\
\text{Cov}(Z,X) & \text{Cov}(Z,Y) & \text{Cov}(Z,Z)
\end{pmatrix} \quad (3.5.5)
\]

Since on the diagonal lies the covariance between one random variable and itself, this is actually the variance of the respective random variable, thus the name of variance-covariance matrix. We already stated that the covariance is a symmetric operator. This means that the variance-covariance matrix is a symmetric matrix with respect to the first diagonal.

In the same way, the correlation $c$ matrix is defined:

\[
c = \begin{pmatrix}
r_{XX} & r_{XY} & r_{XZ} \\
r_{YX} & r_{YY} & r_{YZ} \\
r_{ZX} & r_{XY} & r_{ZZ}
\end{pmatrix} \quad (3.5.6)
\]

The correlation matrix maintains the symmetry of the covariance matrix and because it contains on the main diagonal the correlation coefficients of the variables with themselves, the diagonal will always contain just terms of 1's.

## 3.6 Results

The implementation was done in C++ and Qt and the new developed functions were integrated in the CellVis framework.

In the present case, the data set representing the sample population was considered as the spatial distribution of different types of particles inside the cell. The spatial distribution is described by a position vector. This vector is three-dimensional, where the dimensions correspond to the Cartesian coordinates $x$, $y$ and $z$. The estimated statistical parameters were computed with respect to each axis and to each type of molecule. Thus, functions to estimate the first four moments and the correlation coefficients between the three coordinates were implemented and a tree view was integrated in the graphical interface of CellVis in order to inspect the numerical values of these parameters and to browse through all the time steps of the simula-
tions and through molecule types. A print screen of this view can be seen in Figure 3.5:

![Statistical Parameters View](image)

**Figure 3.5: Tree view of statistical parameters**

Having a view that shows the numerical value is of help when the interest is focused on one time step or to compare these parameters between several time steps. However, just from this view it is hard to have an insight of these parameters tendencies and to visualize the general trend of the different types of molecules involved in the simulation. To facilitate this, several time series plots are integrated in CellVis. Thus a graphical view of how all these parameters evolve through time is provided. Time series plots are available for all the parameters included in the tree view and they are grouped 3 by 3 in one window, so that in one window the tendencies on each the three axes are visible. Figure 3.6 the time evolution of the first for moments visualized in CellVis and in Figure 3.7 the evolution of the correlation coefficients is depicted.
3.6. Results

(a) (b) (c) (d)

Figure 3.6: Time series plot of the first four moments (a) mean (b) standard deviation (c) skewness (d) kurtosis

The statistical analysis is very sensitive to the number of individuals in the sample population. In order to obtain reliable information from the estimated statistical parameters, the number of molecules of each type should
be sufficiently large. The initial count of the molecules present at the beginning of the simulation can be set by the user. Since some type of particles are products of a reaction between two or more particles, their total number cannot be controlled by the user and might be significantly low. In this case, the statistical analysis will not provide meaningful and reliable information.

As it can be observed from Figure 3.6 the yellow curve is highly unstable, especially for the higher order moments, such as skewness or kurtosis. The yellow curve corresponds to the type of molecules that are able to actually reach the nucleus, as it can be seen in Figure 3.8 and in the current simulation their number varies from 2 to 8, which is much less information than is necessary for the statistical analysis to extract realistic characteristics about this type’s distribution. Severe fluctuations of these parameters can also be observed when particles of a certain type are formed for the first time and their number continues to increase until it reaches a rather stable count. Thus, it is recommended to look at the values of the statistical parameters from the moment when the curve tends to stabilize and has a smoother trend.

To make it more easier to put these parameters into context, a function that computes the histogram of each time step over the cell’s diameter of each molecule type’s distribution was implemented. The size of the interval set to compute the histogram is left as a parameter and can be further adjusted if the simulation characteristics requires it. To view this histogram, a 2-D plot was added to the main window. The plot represents the histogram of the current time step and during the visualization the view is redrawn so that the new loaded time step is taken in consideration.

Considering the same simulation used in Figure 3.6 we can have a look at the histogram of each type of molecules at the first time step (Figure 3.9). From this figure we can already form an opinion on how the data is distributed. The dark green curve corresponds to the receptor molecules.
3.6. Results

that lie only at the membrane of the cell, therefore the respective distribution is almost uniform. If we consider the orange curve, we can observe normal distribution like shape. But if we look at the corresponding kurtosis value, we see that this value is almost constant and is close to -1 for all three axes, meaning that this distribution resembles more a Wigner semicircle distribution. This makes sense, since none of the particles can occupy the space inside the nucleus, so even though the mean value is almost always the same as the center of the nucleus, the number of particles close to the mean is smaller than without the occupied space of a nucleus.

The red curve, initially has a bounded distribution between 3 and 7, as it can be seen in Figure 3.9 and at the end of the simulation its distribution is bounded only the membrane radius (Figure 3.10). If we take a look at the evolution of the statistical parameters, we can see how this evolution takes place. The estimation of the mean doesn’t change, because the majority of the respective type of particles still tends to group around the center of the cell, but the standard deviation increases from 1 to almost 2.5, indicating a spreading of the molecules, which can also be observed from the histogram at the last time step.

In almost all cases, the skewness is approximately zero. Only the light blue curve has a negative skewness on X and Z axis, meaning that the corresponding distribution has a longer tail to the left side of the mean. This is due to the fact that the current simulation is characterized by an asymmetric cytoskeleton, that is polarized in the same direction. The light blue molecules represent signaling molecules that bounded with motor proteins, thus with the cytoskeleton and can move along the filaments. Therefore an asymmetry in the cytoskeleton also translates into an asymmetry in this type of molecule distribution.
Figure 3.10: Histogram view of last time step

The histogram view, together with the computed statistical parameters provide quantifiable information about the particles distribution, making it easier to understand the dynamics involved in the simulation and how the different type of molecules are involved or influenced in the signal transduction process.

In order to verify and justify the need for estimating these parameters, we used an external tool to try to fit a distribution given just the first four moments and the size of the sample population. The tool that was used is called R and it represents both a free software programming language and a software environment for statistical computing and graphics [17]. Using two special packages like “SuppDists” and “PearsonDS” to include additional distributions such as Pearson or Johnson, we were able to find distributions that were characterized by the same mean, variance, skewness and kurtosis as the ones obtained from the simulation. In this case, the Pearson distribution of type IV was used, because there is a direct connection between the first four moments and the parameters that characterize this distribution. Pearson Type IV is used to fit observed distributions obtained from data or Monte Carlo simulations and its probability density function is given by:

$$f(x)dx = k \left[ 1 + \left( \frac{x - \lambda}{a} \right)^2 \right]^{-m} \exp \left[ -\nu \tan^{-1} \left( \frac{x - \lambda}{a} \right) \right] dx \text{ (} m > \frac{1}{2} \text{)}$$

(3.6.1)

where $m$, $\nu$, $a$ and $\lambda$ are real-valued parameters and $-1 < x < 1$ ($k$ is a normalization constant that depends on $m$, $\nu$ and $a$) [18]. The symbols of $m$, $\nu$, $a$ and $\lambda$ described by Karl Pearson in [15].

To obtain the desired distribution, we can just use the first four estimated sample moments: $\hat{m}_1$, $\hat{m}_2$, $\hat{m}_3$ and $\hat{m}_4$ and the Pearson’s type IV
parameters can be derived from the following equations:

\begin{align*}
    r &= 2(m - 1) = \frac{6(\hat{m}_4 - \hat{m}_3^2 - 1)}{2\hat{m}_4 - 3\hat{m}_3^2 - 6} \quad (3.6.2) \\
    \nu &= -\frac{r(r - 2)\hat{m}_3'}{\sqrt{16(r - 1) - m_3^2(r - 2)^2}} \quad (3.6.3) \\
    a &= \frac{\sqrt{\hat{m}_2'}[16(r - 1) - m_3'^2(r - 2)^2]}{4} \quad (3.6.4) \\
    \lambda &= m_1 (r - 2)\hat{m}_3' \sqrt{\hat{m}_2'} \quad (3.6.5)
\end{align*}

Using the Pearson type IV distribution inside the R package and the statistical parameters corresponding to several types of particles at a given time step we can try to find a probability density function that fits our data distribution. For starting we choose to take a look at the molecule type “mol2” from the simulation configuration shown in Figure 3.8 and we can compare the probability density function obtained with R with the histogram obtained from the simulation e.g. at times step zero for X axis (Figure 3.11).

![Figure 3.11: Distribution fitting for molecule of type mol2: (a) mol2 histogram from CellVis. (b) mol2 histogram from R. (c) mol2 probability density function from R.](image)

It can be observed that the fitting results correspond approximately with the initial histogram data and the resulting probability density function can be saved and used further to compare with other simulation results in order to make a generalization. The considered distribution is symmetrical and we would like to see if it is also possible to fit a probability density function on other types of distributions. For example, lets consider the case of the molecules of type Receptor A from Figure 3.8.
CHAPTER 3. MOMENT-BASED STATISTICAL ANALYSIS

Figure 3.12: Distribution fitting for molecule of type Receptor A: (a) Receptor A histogram from CellVis. (b) Receptor A histogram from R. (c) Receptor A probability density function from R.

From Figure 3.12 we can see that also in the case of an almost uniform distribution the fitting operation provides encouraging results for the use of the method of moments. As a last example, we consider the distribution of molecules of type Mol B Skeleton from Figure 3.8, which is a skewed distribution and we would like to establish if we can find a probability density function that has approximately the same four moments and a resembling histogram with the one computed from CellVis.

Figure 3.13: Distribution fitting for molecule of type Mol B Skeleton: (a) Mol B Skeleton histogram from CellVis. (b) Mol B Skeleton histogram from R. (c) Mol B Skeleton probability density function from R.

For the Mol B Skeleton type of particles, the last time step was considered instead of the first, due to the fact that these type of molecules are a result of other particles interaction with the cytoskeleton and therefore they are not present in the initial state of the simulation. Figure 3.13 shows that also in the case of an asymmetrical distribution, fitting a probability density function using the method of moments can provide satisfactory results.

In order to make it possible to handle the data resulting from the statistical analysis module in CellVis with other tools specially designed for statistical analysis and computation, a “save” function was implemented. With this function it is possible to export the plot as gnuplot or svg.
Chapter 4

History of Individual Molecules

In this work the history of individual molecules is represented by the particle trajectory and its possible reaction partners. Traditionally two or three dimensional plots are employed to visually represent the trajectory and dynamics of a single protein [19, 20]. Instead of graphs, Falk et al. proposed a visualization tool that uses three dimensional glyphs [2].

4.1 Trajectory of Single Molecules

It is of interest to investigate the trajectory of a certain particle and to observe if the respective particle changes its type as a result of a reaction with other molecules or interaction with the cytoskeleton or nucleus. A particle is identified by a unique ID. By tracking a certain molecule it is possible to investigate how this particle propagates through the cell’s environment. It is also possible to determine the type of motion of the particle, if is by diffusion or by motor proteins. The trajectory might indicate if the particle was involved in a reaction, if it was replaced by another protein and if the particle maintained its direction towards the nucleus. Particles can also remain blocked within the cytoskeleton filaments.

The trajectory is determined by computing the distance from the considered particle’s position to the nucleus position at each time step available in the visualization. Thus, another 2D plot view is added to CellVis to show the trajectory as a 2D curve, where the X axis represents the distance to the nucleus and is bounded by the nucleus radius as minimum and the membrane radius as maximum and Y axis represents the number of time steps.

The choice of using a 2D graph to show the particles trajectory resulted from the intention of providing a clear and simple graphical representation of the particles motion. Since mostly the particle’s movement is achieved by
diffusion, characterized by Brownian motion, a 3D view of the particle’s trajectory can become fuzzy and the general trajectory or direction may become hard to observe. By plotting the trajectory with respect to the particle’s distance to the nucleus, it is easy to observe the particle’s motion tendency, whether is directed towards the nucleus, tends to remain in certain regions or even switches its direction.

4.2 Identification of Reactions and Reaction Partners

To fully understand the dynamics of a certain particle, besides analyzing its motion, it is important to identify also the transformations the respective particle suffers along its path. These transformations might be the result of a reaction in which the particle might become involved. There are several types of reaction that are covered and can be simulated in CellVis and they correspond to the types of reactions presented in Section 2. In the context of CellVis, reactions are a result of a collision between two particles or the spontaneous decay of one particle [2]. The newly formed particle is placed at the same position of the involved reactants, since other placements might not be possible due to the presence of other cell components.

Not only reactions can influence the state of a particle, but also interactions. Interactions consist of a binding process between two proteins. As a result of this binding, a protein complex might be formed or a protein might carry another protein through the cytoplasm. In this thesis, two types of interactions are considered, the first represents interaction with the cytoskeleton and the second type is an interaction with the nucleus.

With the available information about the particles trajectory, we can now try to determine when a reaction takes place and what are the particles involved. Determining when a reaction occurs is relatively easy, since information about each particle’s type is available at each time step. So when a change of type is found or when the particle’s ID is not found in the next time step, we can suspect that a reaction or interaction took place.

Finding the other reactions partners, is a more sensitive problem, since it can be a matter of guessing. If the data resolution coincides with the simulation resolution then the identification of reaction partners yields precise results. When a reaction or interaction has taken place, can be easily determined by investigating the type of the chosen particle. To determine the other reaction partners, we need to analyze the information corresponding to the time step where the particle’s type has changed and also the information at the previous time step. For each type of particle information about the reactions or interactions that it can be involved in is available from the simulation configuration. Further, for each type of reaction or interaction the number of molecules involved are stored together with the role of each
type of molecule: reactant, product or enzyme. Depending on what role the particle of interest plays in the reaction, e.g. is a reactant, product or enzyme, the type of the other reaction partners is identified and group of potential reaction partners is chosen with respect to their distance to the particle of interest. From this group of potential partners, depending on how many partners are left to be determined, we choose the ones that correspond to a smaller distance. The search for reaction partners is performed by handling two consecutive time steps, the time step when a change in the state of the selected particle has occurred and its corresponding previous or next time step. The pseudocode for determining reactions and reaction partners is listed in the appendix A.

In the case of an interaction, just two proteins are involved and the signaling molecule receives the type of the carrying protein. Therefore we just signal the interaction by changing the color of the protein to the new type and mark the time step when this change is identified.

The reference distance is chosen also to include larger distances up to 1/5th of the cell radius. The information from consecutive saved time steps are not necessary two consecutive time steps in the actual simulations. As a result it may seem that a reaction just took place and we would expect that the reaction partners in the next time step should be really close to the particle of interest, but in the real case it might not be so and the reaction partners can be actually further apart.

A simulation can contain more than 1 million time steps. It is not possible to store all the time steps from the simulation. By omitting information from the simulation, the process of determining the reaction partners is a probabilistic process, since there is no certain way to determine if the chosen reaction partners are the real reaction partners. However, the incremental changes in the particles positions from one time step to another are insignificant and it is possible to choose a data resolution without loosing valuable information.

4.3 Results

To illustrate the results in determining the trajectory and the reaction partners of a certain molecule, a test configuration was considered as in Figure 3.8. An example of trajectory view is shown in Figure 4.1. The curve color matches the color of the molecule type, as it was set initially in the simulation configuration. The red markers are meant to indicate an event in the history of the particle. This event can be either a reaction either an interaction. The two cases shown in the above figure cover two types of interaction as they were defined in Figure 3.8. In both cases it is noticeable that when an interaction takes place, either with the cytoskeleton filaments or with the nucleus the particle’s trajectory becomes much smoother and is
not characterized by a Brownian motion anymore.

![Figure 4.1: Trajectory view (a) with cytoskeleton interaction (b) with nucleus interaction](image)

The particle can be picked from the 3D view of the CellVis by clicking on it. If the add particle trajectory button is pushed, then, besides the 3D view of its path already implemented, also the plot trajectory is triggered. The trajectory computation takes all time steps from the current loaded simulation into consideration. Since the particle’s trajectory is only computed as a result of a user’s interaction, time steps are loaded separately and all other views are not affected by this. Therefore it is possible to compute a particle’s whole trajectory at any time step in the simulation and once the plotting trajectory action is triggered, all the time steps are loaded sequentially and the particle’s position is searched in every time step.

![Figure 4.2: Trajectory of type mol3nucleus](image)

The trajectory view can be used not just for one particle, but also for several molecules, in case we are interested in several particles. To illustrate
this, a group of molecules that have reached the nucleus by the end of the simulation is chosen (Figure 4.2). This view can be useful, if we want to determine and observe any general tendencies that might be indicated by the graph.

Unfortunately, the graph in this case becomes hard to understand and the user can have difficulties to distinguish between the trajectories of different particles. In this sense, another graph is introduced, a graph where only the particles distance to the nucleus when a reaction or interaction takes place are plotted. Consequently, it would be more easy to observe where some types of reaction are more likely to take place (Figure 4.3) and a reaction distribution can be observed.

![Figure 4.3: Reaction view of type mol3nucleus](image)

The X axis remains as the distance towards the nucleus, but Y axis becomes a count of the number of reactions in which a particles is involved. Now the graph is clearer and an idea of how several types of reactions are distributed can be formed. This type of graph can be useful also to illustrate a chain of reactions between several molecules.

![Figure 4.4: Simplified MAPK cascade simulation configuration](image)

To illustrate an example of reaction history for a chosen particle, a simulation with a simplified MAPK cascade was used (Figure 4.4). The enzymes
from the receptor complex activate the MAPK, resulting the phosphorylized MAPKp. The MAPKp can be deactivated, dephosphorylized by the active Phosphatase. As a result of this type of reaction the Phosphatase also becomes inactive, but can return to its active state by suffering the reactivation reaction. It can be observed from the simulation configuration, that in this case we are dealing with enzymatic reactions, first and second order reaction, cytoskeleton interactions and nucleus interactions.

Figure 4.5 shows the trajectory of the molecules involved in the signal transduction process. The gray curve represents the trajectory of an enzyme from the receptor complex situated at the membrane. This enzyme binds with the MAPK and now MAPK becomes MAPKp, which is then attacked by the Phosphatase. At time step 38 we can notice that a dephosphorylation reaction takes place and interferes in the signal transduction process by phosphorylizing the MAPKp and preventing it from reaching the nucleus and transmitting the signal.

Figure 4.5: Particle history (a) Particle history including trajectories (b) Reaction history

The reaction history view is shown in 4.5(b) for a more clearer look at the reactions placement. However, in this case the reaction history view turns out to be inexact. From Figure 4.5 we can see that reactions do not align, since on Y axis we count the reactions for each particle. In this case, the first reaction for the Phosphatase corresponds to the second reaction of the MAPK and thus results in erroneous information. If we were to properly align the reactions, then the Y-axis does not make sense anymore, since the reaction count will be wrong. Therefore, it is recommended to use this type of plot just for illustrating the reaction of one particle per reaction. In order to maintain the original approach, of using a graph that shows all the partners involved in a reaction and the positioning of the reaction, the Y axis will have to eliminated and the graph represented just as tree-type of view.
An example of activation pathway that ends in the signaling protein reaching the nucleus can be seen in Figure 4.6.

As a study case, we might also be interested to determine how many MAPKs one receptor is able to activate. For this we picked a random enzyme from the receptor complex, computed its trajectory and determined several MAPKs that were activated by this enzyme. In Figure 4.7(a) the trajectories of the implicated MAPKs and the chosen enzyme are drawn. It can be observed that the only protein that finally reaches the nucleus is the one that also benefits by transport with the motor proteins. The role of motor proteins in the signal propagation will be further investigated in Chapter 5. For this case, we can also use the Reaction History Plot, since in this case the MAPKs do not react with each other and the reactions do not have to be aligned (Figure 4.7(b)).
Chapter 5

Comparative Visualization

5.1 Previous work

In scientific data analysis, comparison of different data sets is an important function, whether the data comes from the same source or not. It is of interest to find similarities or differences in several data sets. In order to do so, a comparison can be made between different models, different resolutions, differences in algorithms, different experimental results, etc. One of the most important motivation for comparative data analysis is model validation. In this case, the output of numerical models can be compared with experimental data. A more suggestive method for comparative data analysis is to visually represent similarities or differences between the data sets of interest. This method is called comparative visualization. Simple 2D graphs are often used to show time development of numerical characteristics of data sets and therefore, can be easily used to show the differences between these characteristics, by “A-B” type of comparison. More complex methods for comparative visualization can be represented by three main areas: image-based, data-level and topological comparisons [21, 22].

Image level comparison can be considered one of the simplest and common approach for interpreting the differences between two or more data sets. This method first generates images from the data sets that are about to be compared. The simplest way to perform the comparison is to show these images side-by-side. However, this approach burdens the user with the task of identifying the regions that might represent a difference and to quantify it. Although showing images side-by-side does not provide any reliable and quantifiable comparison data, this approach is well suited for comparing more then two data sets.

Another method of image level comparison is to use difference images. In this case, two registered images are simply subtracted and the comparison is done pixel-by-pixel. This approach provides the user with an actual numeric information about the differences between the two data sets and
with the use of different masks or filters, certain types of differences can be highlighted. The interpretation of the image difference is dependent on the used color map. Since the data color mapping is typically a non-linear transformation, the interpretation should be done accordingly. The drawback of this method is that one-to-one correspondence between pixels must exist. Also, image level comparison is dependent on the transfer function used in the direct volume rendering (DVR) process. For example, some DVR algorithms interpolate color values and other interpolate data values, thus the result can be significantly different. Therefore, another drawback of the image based comparison is that this approach does not take in consideration the methods of rendering. To overcome this, Williams and Uselton [23] provided guidelines to objectively compare rendered images, by specifying certain rendering parameters, methodology for difference classification and a set of metrics to quantify and classify the differences, or similarities between two volume rendered images.

Data level comparison techniques use actual raw data of the compared data sets. Thus the comparison is done first, at the data level and after the result is finally visualized. This approach was used by Trapp and Pagen darm [24] to compare aircraft design and more recently by Childs et. al. using cross-mesh field evaluations and derived quantities [25]. To make use of these method, one must chose a proper metric for performing the comparison. This can translate in a drawback for this approach, since the data comparison looses its generality and becomes application dependent. Several metrics used for data level comparison in volume rendering algorithms are analyzed in detail by Kim and Pang [26].

Feature based comparison does not use the raw data but certain features extracted from the data set. This technique is more flexible than the other two comparison method, because the user has a range of features to chose from to identify differences in the data sets. Also, feature level comparison enables a more complex analysis of the data sets. Feature based comparison can be viewed as an extension of data level comparison. The main difference is that feature based comparison doesn’t compare the data itself but the features extracted from it. An example on how to compare two methods of calculating streamribbons using a combination by side-by-side and overlaying comparison was presented by Pagendarm and Post [27].

5.2 Comparative Visualization in CellVis

In this work, the raw data is represented by the particles position and types at each time step. Since comparing the particles position by directly calculating the difference of positions does not make any sense, other measures of quantifying the particles spatial distribution were chosen to perform the comparison. These measures are represented by the concentration of the
particles inside the cell and the histogram of their spatial distribution. In chapter 3 a set of statistical parameters were estimated and visualized for a single data set. Thus, the idea of comparing two data sets with respect to these parameters comes naturally. The two different simulation data can be obtained by setting differently their initial state. This can be accomplished by choosing different random seeds for the same configuration or by keeping the same configuration and modifying the random seeds. The comparison is however limited to data sets that are characterized by the same number of simulation steps and with the same shape characteristics of the cell and nucleus.

The comparative visualization is achieved by performing the comparison between the first four statistical moments of the two data sets and between the concentrations and histograms. These differences can indicate a change in clustering, spreadness, skewness and uniformity of the spatial distribution described by the two data sets. To visualize these differences a set of 2D graphs were employed. Besides the 2D graph, for the differences in concentrations, also a 3D volume is rendered. To render these differences, another transfer function is chosen to map the differences in the density values to color and opacity.

The protein concentration or density is approximated by defining volume regions and computing the corresponding local densities. These regions are defined by subdividing the cell into volume elements (voxels). Then for each voxel the number of enclosed molecules is determined. The density is then computed by dividing the number of molecules by the voxel’s volume [28].

The histogram is computed in very similar way as the concentration, just that in this case the histogram is computed for each of the three dimensions and the range of the histogram is the cell’s diameter. This range is divided into equal intervals and for each interval the number of contained molecules is determined. The decision to have also a histogram view besides the concentration was made in order to illustrate also possible asymmetries in the spatial distribution.

Since CellVis was originally designed to support just one data set at a time, the task of loading and providing the same functionalities for a second data has to first reconsider the original structure of the framework. Therefore, the actual comparison is performed using external tools, like MATLAB and Gnuplot. Since every plot can be saved and its data exported to external files, the manipulation of the actual data using external tools is relatively easy. The difference in the statistical parameters is carried out in MATLAB, the result is saved as gnuplot files and visualized using Gnuplot.

In CellVis a function that loads the information of a second data set is integrated. This function just handles the simulation information, but does not update any views for the second data set. The function is created so that a rendered volume representing the difference in concentrations of two data sets can visualized if the user desires it.
5.3. Results

For comparison, three types of different data sets representing a simplified MAPK cascade (Figure 4.4) are chosen. In the first two cases, data sets with same starting random seeds are considered, but different simulation configurations. In the first case the transport by motor proteins was deactivated for the second data set and in the second case the inactivation reaction of the Phosphatase was disabled. Last, we consider two data sets characterized by the same simulation configuration but with different starting random seed.

5.3 Results

5.3.1 Concentration and Histogram-based comparison

Before performing the comparison between two radial concentration profiles corresponding to different simulation data sets, we used a rectangular filter to smooth the computed concentration profile. Initially the concentrations were computed by dividing radius in a fixed number of intervals. In this manner the resulting concentrations present abrupt changes because each interval is considered separately. By using a rectangular filter, we consider also proteins from neighboring intervals when computing the final corresponding concentration. The window of this rectangular filter is considered a parameter when computing the radial concentrations, thus it can be easily modified to adjust the filter size.

Figure 5.1: Radial concentration profile

Figure 5.1 shows the two concentration profiles at the same time step.
The first graph illustrates the results of using a rectangular filter, as the second one represents the initial calculated concentrations. It is easy to see that by using this filter we obtain smoother curves that might reflect better the reality. However, when dealing with molecules that are localized only on certain region, it is best to adapt the size of the interval according to the size of the region. This could be the case of the particles that form the receptor complex, that can be found only at the membrane. In this case the use of a rectangular filter is not indicated since it would suggest a wider spreading of these molecules than in the real case.

To present the results of the comparative visualization with respect to radial concentration and histogram, we chose to compare the concentration profile and the spatial distribution histogram at the first and last time step of the simulation, in order to indicate the differences in the starting state of the simulation and in its outcome.

However for the first two cases that consider data sets with the same starting spatial distribution, the difference in concentrations at time step zero would be equal to zero and therefore was not chosen to be viewed. In Figure 5.2 the difference in the concentrations and histogram at the end of the simulation are represented and it can be observed that high differences
are indicated in the case of Target type molecules and MAPKp*. Considering the fact that the second data set corresponds to a configuration that does not enable transport by motor proteins, it is to be expected that the respective data set will not include MAPKp that can interact with the cytoskeleton (MAPKp*). By disabling transport by motor proteins, the number of proteins that reach the nucleus (Target), is significantly smaller or even equal to zero, thus indicating the important role of motor proteins in the signal transduction process.

From the histogram difference plot one can see that the number of phosphorylized MAPK (MAPKp) is lower in the first data set (indicated by a negative difference), due to the presence of motor proteins that transform a part of MAPKp into MAPKp*.

In Figure 5.3 the differences in histograms and concentrations corresponding to the last time step of the simulation for the second type of comparison are shown. In this case all types of molecules are present at the end of the simulation. It can be seen that the difference in concentrations is almost insignificant. Considering that the reactivation reaction of the Phosphatase is disabled for the second data set’s configuration, one can conclude that this type of reaction rarely takes place and therefore its deactivation does not produce significant changes in the outcome of the simulation.

Figure 5.3: (a) Concentration and (b) histogram difference plot for two simulations, with second simulation having the inactivation reaction disabled
For the third type of comparison two data sets that differ by the initial spatial distribution of the molecules are considered. In this case it is necessary to have a look at the differences of concentrations and histograms also at the first time step, to identify significant initial similarities or disparities and whether or not they are constant or evolve during the simulation. The initial difference is illustrated in Figure 5.4. There are no significant differences in any of the present molecule types. This can be justified by the fact that even though the spatial distribution is different the total number of molecules and the placement of the cytoskeleton is kept the same. This type of comparison is useful when we are investigating several samples of the same type of biological system to see if its dynamic presents the same characteristics through the entire simulation or it presents a completely random behavior.

![Figure 5.4](image)

Figure 5.4: (a) Concentration and (b) histogram difference plot for two simulations with the same starting random seeds, at time step zero

The difference in the outcome of the two simulation is shown in Figure 5.5, where it can be observed that in both cases the target (nucleus) is being reached by a similar number of signaling molecules and also cytoskeleton interactions occur in similar quantities. Again, the calculated differences are insignificantly small, suggesting that just by varying the initial positions of the particles involved in the simulation, without considering any change in the reaction configuration or cytoskeleton polarization, does not produce
meaningful differences in the signal transduction process. From this we can conclude that the dynamic of the intracellular processes is not affected by changing of the initial particles positions as long as they present the same distribution.

Figure 5.5: (a) Concentration and (b) histogram difference plot for two simulations with the same starting random seeds, at the last time step

### 5.3.2 Statistical Parameters-based comparison

Performing also a statistical parameters based comparison completes the comparative visualization and enriches the data analysis process. Analyzing similarities or differences in these parameters together with the difference of concentrations and histograms provides more information about the actual similarities or differences in the considered data sets, the outcome of the two simulations and in the signal propagation process.

This comparison is accomplished by computing the differences of the evolution of the first four statistical sample moments and the results can be viewed in similar manner as in the previous section, by the use of 2D graphs. The difference of the mean can indicate a different tendency of clustering around a certain value in the considered data sets, as for the difference of the standard deviation can reflect differences in the data spreadness. Taking in consideration also the difference in skewness and flatness of the data distribution results in a more complex comparative analysis.
CHAPTER 5. COMPARATIVE VISUALIZATION

For comparison, the same three cases as analyzed in the previous section were considered. In Figure 5.6, 5.7, 5.8 the results of the comparison of the two data sets corresponding to the first type of comparison are illustrated. The absence of Target and MAPKp* proteins in the case of the data set with deactivated transport by motor proteins can be observed also here.

![Graphs showing mean and standard deviation difference plots between two simulations with different transport conditions.](image)

Figure 5.6: (a) Mean and (b) standard deviation difference plot between two simulation with the same random seeds but, with the second simulation having transport by motor proteins disabled

In Figure 5.6 the evolution through time of the difference between the mean and standard deviation is shown. While the graph representing the difference in mean values does not indicate any significant difference in the time evolution of the two data sets, the standard deviation comparison indicates a larger spreadness in the case of inactive Phosphotase and the phosphorylized MAPK. This is caused by the lack of transport by motor proteins in the second simulation and thus the transport towards the nucleus is significantly slowed and the phosphorylized MAPK tend to remain more closely to the membrane. Since the motion of the MAPKp is slowed down and not many of them can get closer to the nucleus, it means that less dephosphorylizing reactions take place further from the membrane, therefore the inactive Phosphatase is unlikely to appear closer to the nucleus, thus justifying the larger spatial deviation of the inactive Phosphatase in the first data set.
5.3. Results

Figure 5.7: (a) Skewness and (b) kurtosis difference plot between two simulation with the same random seeds but, with the second simulation having transport by motor proteins disabled

Figure 5.8: Correlation coefficients difference between two simulation with the same random seeds but, with the second simulation having transport by motor proteins disabled
CHAPTER 5. COMPARATIVE VISUALIZATION

Figure 5.7 shows the difference of the evolution of skewness and kurtosis of the considered data sets. The difference in skewness is hard to interpret and reflects more the stochastic nature of the signal transduction process. The difference in kurtosis, however reflects the difference in the simulation configuration. If we consider again at the phosphorylized MAPK case, one can observe positive difference in the evolution of kurtosis. This indicates that in the case with inactive direct transportation the spatial distribution of the MAPKp is more flat and does not present a high clustering around the mean (nucleus). This is again an effect of the slow motion of the particles, that in this case move only by diffusion.

Figure 5.8 shows the difference of the evolution of the correlation coefficients that describe the correlation between the spatial distributions of one type of particle on each dimension. However, in the considered simulations, there are no general tendencies of the correlation coefficients and therefore also their difference presents a noisy like evolution.

Figure 5.9: (a) Mean and (b) standard deviation difference plot between two simulation with the same random seeds but, with the second simulation having the inactivation reaction disabled

The results of the second type of comparison, where the effect of disabling the reactivation reaction of the Phosphatase was taken in consideration, are presented in Figure 5.9, 5.10, 5.11. In Figure 5.9 the difference of the evolution of the first two statistical moments can be visualized. It can

47
be observed that when the curves become stable there are no significant
differences. The high differences in both mean values and standard deviation
values are an indicator of a delay in the appearance of certain types of
molecules in the case of the second simulation. This delay can be observed
in the case of the inactive Phosphatase, the phosphorylated MAPK that
binds to the cytoskeleton (MAPKp*) and the the MAPKp that reach the
nucleus.

Analyzing Figure 5.10 and 5.11, where the difference in skewness, kurtosis
and correlation coefficients is visualized, the same delay can be observed
when higher difference values are reached. Besides the observed delay, the
two simulations seem relatively similar with respect to their outcome, con-
clusion that was also drawn in the previous section, when the concentration
profile and histograms were compared.

Figure 5.10: (a) Skewness and (b) kurtosis difference plot between two sim-
ulation with the same random seeds but, with the second simulation having
the inactivation reaction disabled

In Figure 5.11 we can notice differences in the correlation coefficients
that have an absolute value higher then 1. We already know that correlation
coefficients are bound between -1 and 1, therefore their difference will
be bound by -2 and 2. If of interest is just to see if there is a difference in
Correlation and not also the direction of this correlation, the difference can
be evaluated by using the absolute values of the respective correlation coef-

48
ficients and for further analysis of the direction of the correlation to consult directly the coefficients of the two simulations. Since none of the considered configurations provide any linked restraints in the particles movement on each axis, these correlation coefficients are expected to be close to zero when the molecule count stabilizes and their differences would result just in a noisy signal.

Figure 5.11: Correlation coefficients difference plot between two simulation with the same random seeds but, with the second simulation having the inactivation reaction disabled

The results of the simulation comparison between two data sets that share the same simulation configuration but with different initial random seeds can be viewed in Figure 5.12, 5.13 and 5.14. In the first figure, the difference of the mean and standard deviation values are presented. In this case, again it is observable a delay in the inactivation of the Phosphatase, in the use of motor proteins and in reaching the nucleus. Regarding the mean value, on all the three axes the differences vary slightly around zero, when large enough number of molecules of the respective type is reached. The difference in standard deviation values indicate some difference in spreadness in the case of the inactive Phosphatase and MAPKp transported along the cytoskeleton’s filaments.
5.3. Results

Figure 5.12: Mean and standard deviation difference plot for two simulation with different starting random seeds

Figure 5.13: Skewness and kurtosis difference plot for two simulation with different starting random seeds
In Figure 5.13, we can observe differences in the skewness and kurtosis of the two simulations. Since skewness and kurtosis can both take also negative values, to correctly interpret the result of the differences between different values they take, it is necessary to analyze the subtracted values themselves and their sign. For example on the Z axis, in the inactive Phosphatase case, a negative difference in skewness is present. Without knowing the sign of the original skewness values of the two simulations, we can not state whether this indicates a skewness on different sides or simply a more skewed distribution on the same side. Also, observing that the higher differences appear in the first half of the simulation, we can conclude that these differences can be influenced also by the small number of particles of the respective type and that a more reliable trend of these differences can be viewed after a sufficiently large number of particles has been reached, e.g. in the second half of the simulation.

Since in this case the comparison is made keeping the same simulation configuration, it is to be expected that the results of the two simulations would be similar. Therefore, it is not of so much interest in comparing the spatial distributions, when there is no difference in the configuration that might alter the shape of these distributions. In this case, information from several simulation runs can be extracted to make an analysis on how fast the external signal propagates and how many of the signaling proteins are likely to reach the target, therefore performing another statistical analysis with respect to several data sets that present the same characteristics in their molecules distributions.

Figure 5.14: Correlation coefficients difference plot for two simulations with different starting random seeds
5.3.3 Concentration Difference Volume

For rendering the concentration difference volume the difference in concentrations in computed. The protein density computation was already an integrated function in CellVis and is done automatically when loading the particles positions into the 3D view. To compute the difference of densities, we load the second set of proteins over the first one and when setting the shader parameters we use a negative density scale value, therefore subtracting the concentrations of the second data set from the first one. The resulting scalar field is mapped to a color range. For positive values of concentrations the color map from Figure 5.15, where small differences are mapped to blue and big differences are mapped to red. The negative values are mapped to a grayscale color map, where large negative differences are close to black and small negative differences are close to white.

![Color map](a) for positive differences (b) for negative differences

Figure 5.15: Color map (a) for positive differences (b) for negative differences

Very small differences in concentrations, that are approximately zero are mapped with a total transparency, to illustrate the similarities in the clustering of proteins of the two considered data sets.

To present the results for this last part of the comparative visualization section, the same three cases of comparison as in the previous sections were chosen.

![Concentration difference volume](a) for time step 0 (b) for intermediate time step (c) for last time step

Figure 5.16: Concentration difference volume for two data sets, one with no cytoskeleton interaction (a) for time step 0 (b) for intermediate time step (c) for last time step

Figure 5.16 illustrates the evolution over time of the differences in concentrations when one of the two data sets does not include transport by motor proteins. For the first time step, no volume is visible since the initial positions of all implicated molecules are the same for both data sets.
and therefore the difference in null. When analyzing the resulting rendered volume at an intermediate time step, it can be observed that positive differences are migrating towards the center, as for negative differences remain closer to the membrane. This is justified by the fact that the first data set benefits of the transport by motor proteins that facilitate and speeds the signal transduction process. Since proteins can move faster towards the nucleus in first simulation data, a higher concentration for this data set is expected around the target. For the second data set, transport is only ensured by diffusion and most proteins can not move as fast and remain at a considerable distance from the nucleus. The last time step shows the same effect: the outcome of the first simulation data produces a more crowded environment around the nucleus.

Figure 5.17 illustrates the results from the second comparison case, when the reactivation of Phosphatase reaction is deactivated for the second loaded data set. The first time step reveals the same effect as the first case. The two data sets having the same initial states, the starting concentrations difference volume is also equal to zero. The intermediate time step does not suggest any visible tendencies in the concentration difference. The last time step however, reveals a larger distribution of the negative difference values, especially around the nucleus. Considering that in the second simulation data, the Phosphatase can not be reactivated and therefore the activated MAPKp can not be attacked by it and deactivated, it could make sense that in the case of the second data set more phosphorylized MAPKs are able to reach the target.

![Concentration difference volume](image)

(a) (b) (c)

Figure 5.17: Concentration difference volume for two data sets, one with reactivation reaction of Phosphatase disabled (a) for time step 0 (b) for intermediate time step (c) for last time step

Figure 5.18 illustrates concentration differences between two data sets that have the same simulation configuration but have a different initial state. In this case, the evolution and the outcome of both simulations seem to be very similar, as it was deduced also from comparing the radial concentrations
and the statistical parameters. In this case, as a plus, it can be observed that the original distribution of concentration differences tend to remain the same. Therefore we can conclude then that initial crowded regions do not migrate significantly, they just suffer a relatively small change in shape.

![Figure 5.18](image1.png)

(a) (b) (c)

Figure 5.18: Concentration difference volume for two data sets with the same simulation configuration, but different initial random seeds (a) for time step 0 (b) for intermediate time step (c) for last time step

![Figure 5.19](image2.png)

(a) (b) (c)

Figure 5.19: Concentration difference volume for different density scales (a) same scale with a low value (b) same scale with a higher value (c) first data set with higher scale

Figure 5.19 shows how the concentration difference volume can be manipulated. For Figure 5.19(a) and (b) equal density scales were used for both data sets. In the first case a relative low density scale is chosen and the resulting volume can be seen in Figure 5.16(a). If we want to emphasize larger differences (both positive and negative) we can choose a larger density scale as in Figure 5.19(b). In Figure 5.19(c) different density scales were chosen. This could be justified if we want to visualize a weighted difference, in order to emphasize the concentrations of one data set over the other one.
Chapter 6

Conclusions and Future Work

6.1 Summary

The objective of this work is to provide means for a statistical analysis over the previous obtained simulation data and to expand the CellVis framework to handle two sets of data for performing a comparative visualization.

For the statistical analysis module, the method of moments is chosen to describe the spatial distribution of certain types of particles involved in the simulation. The spatial distribution is considered separately on every of the three dimensions, since the movements on each of the three axes of a particle are independent. Hence, the first four sample moments are estimated for every time step stored from the simulation, for the spatial distribution on each axis. The evolution of these moments throughout the simulation is illustrated with the use of a set of time series plots, that are updated with every new loaded time step.

To facilitate placing these estimated parameters into context, the spatial distribution’s histogram is also calculated and visualized. The histogram for each type of particle is shown and computed for each axis, to provide a schematic view of the shape of the spatial distribution. This shape is further characterized and described by the estimated sample moments.

To justify the relevance of using the methods of moments to describe a distribution, an external tool is used to estimate probability function given the first four moments resulting from the analysis of the simulation data. The results are satisfactory with respect to the matching of the histograms, demonstrating that the method of moments can provide enough information to find a probability density functions that fits the distribution of the sample population. This probability density function is not unique and the fitting operation can provide different results with each try. The estimated probability density function can further be used to predict the behavior of
molecules in other simulations.

The correlation coefficients between the spatial coordinates on the three axes were also estimated. The poor correlation between these coordinates is demonstrated by visualizing the evolution of their coefficients. However, if these coordinates are independent, cannot be decided just by investigating these coefficients.

The statistical analysis section focuses on the general aspects of the simulation data. However, it is important also to provide means of investigating particular aspects of the simulation, like the individual trajectory or the chain reaction of a certain particle. The trajectory is computed as the distance from the particle of interest to the nucleus and is illustrated in a 2D graph, showing the trajectory development over time. Analyzing the particles trajectory we can observe if the chosen signaling protein succeeds in transmitting the signal to the target, in this case the nucleus. An analysis of the reactions in which a particle is involved is performed in order to further investigate the particularities of the signal transduction pathway. Determining the reaction partners of a certain molecule can be a probabilistic process when the data disk resolution does not match the data simulation resolution. The only means of identifying the possible reactants is to investigate their position with respect to the particle of interest. It might be possible to detect a reaction, but not find any reactants, since the present search does not cover the whole cell space, but just a volume with the radius of 1/5th of the cell’s membrane around the particle of interest. However, this is an unlikely event, since the considered volume is sufficiently large with respect to the cell’s size. It might also be possible to detect false reactants if two reactions of the same type occur in the same stored time step. This case can also be considered unlikely.

To illustrate the results of determining the trajectory and the reaction partners of a selected molecule, a simulation of a simplified MAPK cascade is considered and an example of how the external signal propagates through the signaling molecules is shown.

The comparative visualization is mostly carried with the help of other external tools, such as MATLAB and Gnuplot, since the task of providing the same functionalities for a second data set is more complex and CellVis’s structure needs to be reconsidered. However, for the visualization of the differences in molecules concentrations as a rendered volume, a function that completes the loading of a second data set is integrated in the CellVis framework.

For the comparative visualization, a set of 2D plots were used to illustrate the differences in statistical parameters, concentrations and histograms of two second data sets. These plots provide quantifiable information about the similarities or differences between two second data sets. For a more suggestive visualization, the difference in the concentration profile of two simulations were mapped to colors according to a predefined transfer func-
tion and rendered in a form of a volume. This type of visualization also provides the user with the possibility to emphasize the differences by tuning a density scale included in the Property View of CellVis.

For the comparative visualization it is important how we choose the considered data sets. For example, if it is of interest to investigate the general tendencies of a type of system, we can choose data sets that are characterized by the same configuration, but differ in their initial state. If we desire to investigate the influence of one type of molecules or of different types of transport, we can choose second configuration that has enabled or disabled a certain reaction or activates/deactivates transport by motor proteins. According to these cases, a set of different configurations representing a simplified MAPK cascade were chosen and compared.

6.2 Outlook

The integrated methods, views and graphs represent just the basics of a data analysis module. While the method of moments is the oldest method of deriving point estimators and almost always produces asymptotically unbiased estimators, the resulting estimators are not always the best. To find a better probability density function that fits the simulation data, we can use the first four moments as starting values for a maximum likelihood fitter.

The estimation of the first four moments provide useful information in the case of typical distributions, however if some types of particles form multiple clusters and do not have just one central value, the estimated moments are not able to reflect that. In this case, it is useful to determine the forming of clusters and to find means to characterize them. The clusters can be described by their size, e.g. estimated volume or shape, if the shape is conventional or by their motion. It is of interest also to investigate how these clusters evolve trough time, whether they increase or decrease in volume or number of molecules and what is their life expectancy or how fast they form and how fast they spread.

Methods to measure the correlation between the motion of several types of particles or the occurring of several types of reactions or interactions can be further developed to understand the role and the connections between all the consisting entities of the cell.

It is also important to determine the multiplication factor of each molecule type from one time step to another or their expected lifetime an average travel distances, therefore completing the statistical analysis process.

Once we are able to identify when reactions take place, we can perform an analysis of the reaction distribution. It might be of interest to determine regions where certain types of reaction are more likely to take place. For this, we can extract all reactions of a certain type and graphically represent their position. We can also compare the number of reactions that take place
6.2. Outlook

in a single time step with the theoretical expected reaction count presented in Section 2, in order to validate the employed model.

The reaction partners identification algorithm can be further improved by using an intelligent search with respect to distance. This algorithm can start with a minimum distance and further increase it until viable reaction partners are found.

For the comparative visualization, first the framework has to be modified to provide enable all the functionalities also for a second data set. Therefore we can integrate the difference graphs obtained with MATLAB and Gnuplot. It could be useful to give the possibility to visualize two simulations side-by-side, together with their corresponding volume of concentrations difference. Also other features can be extracted from the simulation data to be used in the comparative visualization process. We can consider the number of clusters, their occupied volume, their lifetime. The multiplication factor of each type of proteins can be used to make a comparison of two simulation runs, in order to compare the efficiency of the activation process of MAPK.

To sum up, the statistical analysis together with the visualization process provide a complete tool for interactive data analysis of the cell’s mechanisms and functionalities. The need for comparative analysis and visualization is present in any research area. Future work can consist of several improvements and extend the data analysis process with other functionalities or by employing other approaches. The possibilities mentioned above are just a few improvements that can be further integrated in this framework. However, a collaboration and consultation with biologists can reveal many other needs and other ideas of enriching the data analysis process can emerge.
Bibliography


Appendix A

Identification of reactions and reaction partners

for (time step t = last: first)
{
    load time t
    if particle type has changed
    {
        if the particle of interest is reactant
        {
            identify the types of the other n reactants, and m enzymes
            for each particle of these types
            {
                compute the distance to the particle of interest
                chose the ones with a distance smaller then 1
                place them in an ordered list according to type
            }
            take the first n, and m members from each list
            load the time step t+1
            identify the types of the s corresponding products
            for each particle of these types
            {
                compute the distance to the particle of interest
                chose the ones with a distance smaller then 1
                place them in an ordered list according to type
            }
            take the first s members from each list
        }
        if the particle of interest is enzyme
        {
            identify the types of the other n reactants
            for each particle of these types
            {
                compute the distance to the particle of interest
                chose the ones with a distance smaller then 1
                place them in an ordered list according to type
            }
            take the first n members from each list
            load the time step t+1
            identify the types of the s corresponding products
            for each particle of these types
            {
                compute the distance to the particle of interest
            }
        }
    }
}
chose the ones with a distance smaller then 1
place them in an ordered list according to type
}
take the first \( s \) members from each list
}
if the particle of interest is product
{
identify the types of the other \( s \) products
for each particle of these types
{
compute the distance to the particle of interest
chose the ones with a distance smaller than 1
place them in an ordered list according to type
}
take the first \( s \) members from each list
load the time step \( t-1 \)
identify the types of the the other \( n \) reactants, \( m \) enzymes
for each particle of these types
{
compute the distance to the particle of interest
chose the ones with a distance smaller then 1
place them in an ordered list according to type
}
take the first \( n,m \) members from each list
}
Declaration

I hereby declare that I have independently written the thesis submitted by me. I have indicated all passages whose exact wording or analogy have been derived from published or unpublished works of others. All sources and other means used for this thesis have been indicated by me. Neither the exact contents nor major segments of the thesis have yet been submitted to an examination authority.

Stuttgart, 3 May 2013

Ana Cristina Pintilie