Design and Implementation of a Stand-Alone Tool for Metabolic Simulations

Milad Ghiasi Rad
University of Nebraska-Lincoln, ghiasirad.milad@huskers.unl.edu

Follow this and additional works at: http://digitalcommons.unl.edu/computerscidiss

Part of the Computer Engineering Commons, Databases and Information Systems Commons, Numerical Analysis and Scientific Computing Commons, Other Computer Sciences Commons, Software Engineering Commons, Systems Architecture Commons, and the Theory and Algorithms Commons

http://digitalcommons.unl.edu/computerscidiss/139

This Article is brought to you for free and open access by the Computer Science and Engineering, Department of at DigitalCommons@University of Nebraska - Lincoln. It has been accepted for inclusion in Computer Science and Engineering: Theses, Dissertations, and Student Research by an authorized administrator of DigitalCommons@University of Nebraska - Lincoln.
Design and Implementation of a Stand-Alone Tool for Metabolic Simulations

By

Milad Ghiasi Rad

A THESIS

Presented to Faculty of

The Graduate College at University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Juan Cui

Lincoln, Nebraska

Nov 2017
DESIGN AND DEVELOPMENT OF A STAND-ALONE TOOL FOR METABOLIC SIMULATIONS

Milad Ghiasi Rad, M.S.

University of Nebraska 2017

Advisor: Juan Cui

In this thesis, we present the design and implementation of a stand-alone tool for metabolic simulations. This system is able to integrate custom-built SBML models along with external user’s input information and produces the estimation of any reactants participating in the chain of the reactions in the provided model, e.g., ATP, Glucose, Insulin, for the given duration using numerical analysis and simulations. This tool offers the food intake arguments in the calculations to consider the personalized metabolic characteristics in the simulations. The tool has also been generalized to take into consideration of temporal genomic information and be flexible for simulation of any given biochemical model. After implementation, experimental results have demonstrated the numerical effectiveness of optimization for model selection and the feasibility of the proposed tool for the given metabolic simulation. The proof of concept analysis on the energy metabolism and insulin-glucose metabolism revealed this tool can be promising for a variety of healthcare applications.
ACKNOWLEDGEMENT

I would like to thank the department of Computer Science and Engineering whose funding helped me complete this work.

I also would like to express my sincere appreciation to Dr. Juan Cui, my advisor, who her support, guidance, patience, presence, and dedication helped me to complete this work. I have learned a lot from his expertise, and it is my honor to complete my Master of Science under her supervision.

I would like to thank my committee members, Dr. Cui, Dr. Pierobon, and Dr. Revesz from the Department of Computer Science and Engineering. It is my pleasure to have them in my committee. I also thank my friends Bruno, Jiang, Jayce, and Aditya who their help saved me from a lot of troubles.

Finally, I would like to thank my family, my parents, my wife Marzieh, and my sister Zahra, who their support made me stay strong and determined and be confident and resolute by hearing their voice. I appreciate their patience and positive energy through every step of my life.
# Table of Contents

1  INTRODUCTION ............................................................................................................. 1
   1.1  MOTIVATION ........................................................................................................... 1
   1.2  SUMMARY OF CONTRIBUTION .............................................................................. 2
   1.3  OUTLINE .................................................................................................................. 4

2  BACKGROUND .................................................................................................................. 5
   2.1  METABOLIC NETWORKS ....................................................................................... 5
   2.2  PROPOSED BIOCHEMICAL SIMULATION TOOLS .................................................. 6

3  THE DESIGN OF BIOCHEMICAL SIMULATION TOOL .............................................. 9
   3.1  NUMERICAL ANALYSIS OF BIOCHEMICAL SYSTEMS ......................................... 9
      3.1.1  Numerical Modeling ......................................................................................... 11
      3.1.2  An Example for the Enhanced Biochemical System Simulation Model .......... 15
   3.2  MODELING OF METABOLIC PATHWAYS ............................................................ 16

4  IMPLEMENTATION OF THE TOOL .............................................................................. 22
   4.1  STRUCTURE OF THE PROGRAM .......................................................................... 23
   4.2  DATA EXTRACTION ............................................................................................... 23
      4.2.1  SBML Model Parser ....................................................................................... 24
      4.2.2  External Concentration Import ....................................................................... 24
   4.3  ODE CALCULATION ............................................................................................... 25
   4.4  GENE EXPRESSION DATA IMPLEMENTATION .................................................... 25
   4.5  FOOD INTAKE IMPLEMENTATION ........................................................................ 27
   4.6  THE WORKFLOW .................................................................................................... 28

5  VALIDATION AND RESULTS ......................................................................................... 30
   5.1  MODEL PREPARATION AND SIMULATIONS ......................................................... 30
5.1.1 SBML Model Preparation ................................................................. 30
5.1.2 Case Study Simulations ................................................................. 31
5.1.3 Time Complexity ........................................................................... 36
5.2 Mobile Based Diet Monitoring System ............................................. 37
5.3 Glucose-Insulin Model Simulations .................................................... 40
5.4 OSG Implementation ........................................................................ 42
6 Conclusion and Future Work ............................................................... 45
7 Bibliography ...................................................................................... 48
List of Figures

Figure 3.1 The basics of SBML model and the information that it provides .................. 10

Figure 3.2 An example for the structure of a SBML model for a set of chemical reactions
                                                                                           .......................................................................................................................... 11

Figure 3.3 The order of concentrations updates for the given i and i+1 iterations ........ 14

Figure 3.4 The complete schema of metabolic pathway presented by KEGG ............ 19

Figure 3.5 Major metabolic pathways participating in ATP production ....................... 21

Figure 4.1 Program Structure of the proposed tool.................................................. 23

Figure 4.2 A sample ODE extracted by the proposed stand-alone tool ...................... 25

Figure 4.3 Effect of the gene expression on the reaction rates ....................................... 26

Figure 4.4 The workflow of the proposed stand-alone metabolic simulation tool ....... 29

Figure 5.1 The change in concentration of alpha-Aminoadipoyl-S-acyl as one of the
species participating in the model for 100 minutes ...................................................... 33

Figure 5.2 The reaction associated with degradation of alpha-D-glucose to alpha-D-
glucose 6-phosphate. C00267, C00008, C00020, and C00668 represent alpha-D-
glucose, ADP, AMP, and alpha-D-glucose 6-phosphate................................................. 34

Figure 5.3 The change in concentration of alpha-D-glucose as one of the species
participating in the model for 100 minutes ..................................................................... 34

Figure 5.4 The change in concentration of AMP as one of the species participating in the
model for 100 minutes .................................................................................................... 35

Figure 5.5 The change in concentration of Ammonia as one of the species participating in
the model for 100 minutes .............................................................................................. 36
Figure 5.6 (a) Glucose concentration in starvation state, (b) Glucose concentration for a normal diet of three main meals within a day, (c) ATP concentration of the normal diet, (d) the load applied to mimic the activity (the ATP consumption) captured by exercise tracker, and (e) the ATP concentration for one day after applying the given load.

Figure 5.7 The insulin signaling pathway extracted from KEGG.

Figure 5.8 Glucose-Insulin correlation tested using the proposed model.

Figure 5.9 Two OSG simulation results for high initial glucose concentrations.

Figure 5.10 Two OSG simulation results for high initial glucose concentrations.
List of Tables

Table 3-1 The initial concentrations of the chemicals participating in the example system ................................................................. 15
Table 3-2 The reaction rates of the given example system ........................................................................................................ 15
Table 3-3 The concentrations of the chemicals participating in the example system ...... 16
Table 4-1 The complete list of the arguments .............................................................................................................................. 22
Table 4-2 Some of the most common amino acids and their molecular weight ............ 28
Table 5-1 Processing time in presence and absence of gene expression data .............. 37
Chapter 1

1 Introduction

The biochemical systems contain chains of chemical reactions. Each of these chemical reactions may include numerous chemicals which are called substrates that participate in the reactions. Each reaction has its own rate which may be low or high for its forward or backward direction. These rates show how fast a reaction is likely to move forward or in a reverse direction.

1.1 Motivation

Metabolic networks are the networks of biochemical reactions inside a living organism that are responsible for energy production alongside with some other substrates. Usually, in the metabolic networks, the enzymes are responsible for the rates of the reactions. The higher activity of the enzymes inside a cell can lead to have faster reactions between the species having those enzymes in their way of transforming to another product. This will affect the other reactions indirectly. Therefore, the study of biochemical networks simulation is critical to understand the living organisms.

One way to study the metabolism inside a cell goes through the experimental studies. The experiments take place on the organisms to estimate the concentration of the specific
chemicals. Although this method is reliable, the process is very expensive and is not proper to be used in repetitive researches. Reconstructing the results is very important in bioinformatics studies, as the data generated using a specific method will be used in different researches or be the input of some simulation tools. A simulation tool, gives more freedom to a researcher to analyze a reaction faster and with more reliability, while the cost of experiment remains low. The disadvantage of a simulation tool is in the absence of enough knowledge to make a complete model. The small models do not represent all the characteristics of a big system, and the more complex models are very costly to be used in a computational method. Thus, a flexible tool, that is not dependent on the size of the input model is the key to a reliable analysis.

The metabolism is an area in which, its simulation tools need to have some extra features, such as considering the food intakes and gene expression levels, compared to the other tools [1]. The currently used tools only consider the metabolic networks as a basic chemical reaction chains which do not have the foretold features.

### 1.2 Summary of Contribution

Currently, the tools being used are limited to the small fraction of abilities required to analyze the metabolic networks. The biochemical simulations having the initial concentrations and the *System Biology Markup Language (SBML)* [2] model are the only feature that is accessible in most of the tools like COPASI [3]. The other tools are not that different to be considered designed specifically for the metabolic analysis.

In this research, we aimed to propose a new tool that is specifically designed for the analysis and simulation of the metabolic networks. In this tool, some useful features other
than the SBML import have been added to make it more flexible for the research purposes. This tool is a stand-alone package designed using the libraries in Java and is capable of being used for different purposes.

The features importable by the user are gathered based on the needs of a stand-alone metabolic simulation tool. In every step of the tool preparation, the required features are estimated, and finally the below list has been concluded.

- The simulation duration
- Number of steps
- The specie to be observed
- SBML model import
- External initial concentration import
- Gene expression data import
- Food intake import
- Steps of the food intake
- Weight of the study subject

Based on the design target proposed above, we have designed a tool that benefits a variety of capabilities alongside the Taylor method [4] which is used to simulate the outcome numerically. In our tool, the simulation time is the total number of simulation duration in seconds. The simulation duration is divided by the number of steps which gives the step sizes. A specific specie can be chosen from the list of substrates to be seen as a primary target.
SBML model includes the initial concentrations, but sometimes, there is a need to import the external concentrations to adjust the previously imported initial concentrations. This is another feature that has been added.

The enzymes can also be related to the gene expression levels inside a cell. In many cases, the high expression of a specific gene in long term can lead to increase in activity of the associated enzymes in that cell. We added the ability to import the gene expression levels for the future studies.

The other important feature that was considered is the food intake characteristics. In the food intake arguments, the total carbohydrate, sugar, lipids, and proteins are given to the program. The number of steps in which the food is injected to the cell is also added. The calculations are not completed until the weight of the subject is considered. The weight and type of the subject define the total blood volume available in that test subject, which are in relation with the observable changes in concentrations. These changes are implied by the food intake in a given period.

1.3 Outline

This study follows by Chapter 2 which discusses the background of this work and some literature on biochemical simulations. Then, Chapter 3 presents the theory behind the tool design. Chapter 4 discusses the challenges on the implementation of the tool. Chapter 5 demonstrates the numerical results and their validation. At the end, Chapter 6 illustrates our conclusion and some proposed future work.
Chapter 2

2 Background

The simulation of chemical systems and specifically the biochemical reactions has been studied for several years. In this chapter, the background of study on metabolic networks and some useful tools and methods developed for this aim are discussed.

2.1 Metabolic Networks

Cell metabolism, as the main focus in this thesis, has been the study subject in a lot of experimental and computational researches. Different studies have discussed the role of glucose on the metabolic pathways [5] and its effect on the functionality of the living organisms. [6], [7], [8], and [9] studied the insulin-glucose relationship. Another study [10] showed how oxidative phosphorylation or aerobic glycolysis can affect the proliferation of cancer cells.

Other research focused on protein metabolism. The amino acids are the main components of the proteins. Their role in different metabolic pathways has been studied in several publications. [11]–[14] demonstrated the findings on the catabolism of amino acids and how this affects their related metabolic pathways. On the other hand, [15]–[17] focused on the lipids’ role on the metabolism of living organisms.
Some databases are also implemented, as their organization affects the performance of the metabolic networks analysis due to high volume data. [18], [19] introduced a universal database of the metabolic pathways called MetaCyc to show the chemical reactions supported by the experimental data. Other databases like KEGG [20] and HMDB [21] are used for the extraction of metabolic pathway models. In smaller scale, data is stored using Systems Biology Markup Language (SBML) [2] model due to its ability to store all the information required by the biochemical reactions’ simulation.

### 2.2 Proposed Biochemical Simulation Tools

Since the time researchers developed a method to simulate the behavior of some reactions using other chemicals [22] or mathematical models [23], the chemical system’s simulation has been in the center of attentions. Different tools have been developed to help the simulation of chemical reactions. Different programming languages and techniques have been manipulated to implement these tools.

In [24] a Hill’s formal theory of physical clusters was used to find the equilibrium constants for the formation of physical clusters of molecules. This led to a molecular dynamic calculation that calculates the average potential energy of the molecule cluster. Besides, [25] discussed that the classical reaction kinetics are unsatisfactory when the reactants are spatially bounded by any physical constraint or force field. Other literature such as [26] presented a Monte Carlo method for the simulation of chemical systems undergoing any combination of reaction and phase equilibria. In addition, [27] tried to introduce a new method for modeling of the chemical reactions instead of using the Monte Carlo method to optimize the throughput and parallelism instead of minimizing the
operation count. [28] used FORTRAN to develop a new method for the simulation of kinetic progress curves in the chemical reactions. [29] and [30] again used FORTRAN to develop a prediction tool for gas phase chemical kinetics using sensitivity analysis. Some other literatures discussed on how to increase the accuracy of the computer simulation tools. In [31], a procedure is described to elevate the accuracy of the computer chemical models to study the chemical reaction.

In the matter of chemical reaction pathways, there are some literatures specifically studied this problem. In [32] a theoretical approach was specially designed to determine the energy along the reaction path of a specific enzyme. This paper discusses how the represented simulation predicts the role of the enzyme in the reactions. Other computer simulation techniques for analysis of enzyme involved reactions are discussed in [33], [34]–[36], [37], and [38].

There are also some useful tools developed based on various methods to ease the analysis of the chemical systems. Spyder-Py3 is one of these tools that provides a command line interface using Python 3 to the users to simulate the chemical reactions at different instances. COPASI [3] is the other tool which uses a graphical user interface to import and export data. While this tool benefits a lot of great features such as importing various file types and reaction definitions, it lacks the flexibility needed for metabolic networks analysis such as feeding and user characteristics’ compatibility.

A tool was developed in [1] to display genomic data onto diagrams of metabolic pathways. The combination of this tool with other chemical kinetic simulation tools can benefit both the chemical and genomic data of the metabolic systems to have a more
accurate simulation, as the genomic data can help the researchers to understand the activity of the enzymes participating in the reactions.

For the metabolic simulation purposes only, some tools have been developed by different research groups. MOST [39], which is a software environment for constraint-based metabolic modeling and strain design, has been developed to solve the flux balance analysis problems. Other tools are available if nucleotide or protein sequence data is available. KAAS [40] and PathoLogic [41] are categorized in this group. SEED [42] focuses on the high throughput generation of genome-scale metabolic models.

None of the introduced literature practically have studied the food intake in their calculation. Moreover, the gene expression level data is not considered into account as a factor that may affect the enzyme activities.
Chapter 3

3 The Design of Biochemical Simulation Tool

In this research, a tool was developed and implemented to perform the simulation of metabolic networks and analyze its performance in biochemical analysis. The tool that we developed in this research was assumed to be capable of modeling the metabolic pathways, investigating any chemical system, and predict the concentrations of the species participating in the chemical reactions by previously proven numerical methods. The model needs to be prepared by the user, which in our research SBML model is the best representation that can hold all the necessary data from the chains of reactions and is easy to extract data from it. In this thesis, the theory and modeling of the proposed tool is summarized in three sections. First, an introduction will be given on numerical analysis of the biochemical systems. Then, the modeling of metabolic pathways will be discussed. At last, a model will be discussed for the glucose-insulin interactions as an implementation made using the proposed tool.

3.1 Numerical Analysis of Biochemical Systems

The first step in simulation of the biochemical systems is to prepare a feasible model to represent all the characteristics of chemical reactions participating in it. A feasible model
should contain all the required information for simulation such as, rates of the reactions, the species participating in those reactions, their initial concentrations, the favorability of the reactions, and their reversibility. The scales of the values, the interaction between the species, and other extra parameters also should be contained in such a model.

System Biology Markup Language (SBML) is one of the most powerful and favorable databases in biochemical systems used by the researchers to import and export data into the simulation tools. SBML model uses XML language basics to represent the chemical reactions. A basic of the SBML model and what information it provides is shown in Figure 3.1.

![Figure 3.1 The basics of SBML model and the information that it provides](image)

The information provided by the SBML model indicates that initial concentrations, reaction constants, reaction rates, reversibility of the reactions, the substrates, the products, and so on can be extracted. An example of a SBML model’s structure is shown in Figure 3.2. A normal SBML model consists of annotations, list of unit definitions, list of compartments, list of species, and list of reactions. Each of which follows a data structure
that links all the quantities together in a way that is extractable by different machine language.

![XML code]

**Figure 3.2** An example for the structure of a SBML model for a set of chemical reactions

### 3.1.1 Numerical Modeling

To solve a numerical problem, proper methods are needed to solve the simulation process. One of the most common methods is to use the Taylor series [4] which is represented in Equation 1 as a limit form.

\[
f(t + t_0) = \lim_{h \to 0^+} \sum_{n=0}^{\infty} t_0^n \frac{\Delta_h^n f(t)}{n!} h^n
\]

where \( f(t + t_0) \) represents the function needed to be estimated at time \( t + t_0 \), \( t \) is the current time, \( t_0 \) is the time interval, \( n \) is the harmonic number, and \( \Delta_h^n \) is the \( n^{th} \) finite gradient operator, while \( h \) is the limit variable.

\[
f(t + t_0) = f(t) + f'(t)\times t_0 + e
\]

By simplifications, Equation 1 is transformed to the first-degree Taylor series that simply can be used for the simulation purposes in which, \( f'(t) \) represents the rate of change which can be found using traditional classical chemical kinetics (CCK), and \( e \) shows the
error. Although the first-degree Taylor series does not represent the full spectrum of a function, its error in short enough step sizes is negligible.

Having the SBML model, a general form of a standard chemical reaction with four species is shown in Equation 3. This equation can represent a more general form of a chemical reaction that the number of species participating in it are more than four.

\[ k_A A + k_B B \xleftrightarrow{k_{1},k_{-1}} k_C C + k_D D \]  \hspace{1cm} (3)

where, \( A \) and \( B \) are considered as substrates, \( C \) and \( D \) are the products both in \( \frac{\text{mole}}{\text{liter}} \) or related scales, \( k_1 \) and \( k_{-1} \) are the forward and reverse reaction rates, and \( k_A \) to \( k_D \) are the coefficients of the species \( A \) to \( D \).

For the chemical concentration simulation purposes, \( f'(t) \) can be represented as \( \frac{\text{[S](t)}}{dt} \) for a general specie \( S \). In a chemical reaction, every specie can have a trend for its concentration. This trend is possible to be simulated using Equation 2 that finds the value for the next iteration using the ordinary differential equations (ODE) and the value from the previous iteration.

To learn more about the ordinary differential equations, model in [43] was used. In this model, the chemical reactions are characterized as the more general form of equation 1. The chain of reactions, the variable species and their coefficients work together to estimate the total gradient in the concentration change of a specific specie. Based on this model, the concentration of a specie is not independent from the other species participating in the same list of reactions.

The ODE can be calculated from the SBML file provided by the user of the simulation tool. The SBML contains all the required data necessary to start the calculation of \( [S]'(t) \) which generally can be referred to as the equivalent of the ODE for the specie \( S \) at time \( t \).
Although there are some tools like COPASI in which, the reaction data and ODEs can be extracted, the information gathered with this tool is not easy to use and is not in a string format correctly to be associated with the variables participating in the reactions. COPASI only provides a small portion of the ODEs as the independent ODEs and the rest should be calculated indirectly with the given ODEs. Therefore, another advance in the proposed tool can be in developing a method to find the ODEs in a proper format. The ODEs given with this format are more reliable and easier for presentations and quality check. A general form of an ODE for a specie $S_i$ derived from [43] is presented in below.

$$\frac{d[S_i](t)}{dt} = \sum_{j=1}^{m} v_{ij} V_j$$

(4)

$$V_j = k_1 \prod_{i=1}^{n} [S_i]^{v_{ij}} - k_{-1} \prod_{i=1}^{n} [S_i]^{v_{ij}^{-1}}$$

(5)

Equation 4 follows the Equation 5 in which, $V_j$ is defined. The $V_j$ values are found for each reaction separately, which will be calculated prior to the Equation 4, where $v_{ij} = v_{ij}^1 - v_{ij}^{-1}$ while, $v_{ij}^1$ and $v_{ij}^{-1}$ associate with the forward and reverse coefficients of the $i^{th}$ specie in $j^{th}$ reaction respectively, $m$ is number of the reactions, and $V_j$ is defined in Equation 5 in which, $n$ is the total number of species in all of the reactions [43].

These two formulations are the key to the modeling and simulation of the biochemical systems as they represent the derivation of a specific concentration at any given time. The higher positive of negative ODE, at any given time, represents the higher change at the next iteration in the concentration of a specific specie. At each iteration, the list of all the species which contains the substrates and products in all of the reactions will be assigned with the new concentration values to be used in the future steps. This process is done by chronological orders as there is no preference of doing the analysis at this time. The other
Ordering heuristics can be implemented accordingly. The order of this calculation has been shown in Figure 3.3.

Figure 3.3 The order of concentrations updates for the given i and i+1 iterations

In the figure, \( C_1(i) \), \( C_k(i) \), and \( C_n(i) \) represent the concentrations of the metabolites 1, \( k \), and \( n \) at iteration \( i \) respectively, while the indices show the iteration number. As it is shown in the Figure 3.3, the first specie in the order is updated using the last assignments of the complete set of species from the previous iteration. The next specie is updated using the latest assignment of the first specie of the current iteration alongside with the last assignments of the other species from the last iteration. This process goes on, until the last specie, which only uses the latest assignments of the previous species at the current iteration.
3.1.2 An Example for the Enhanced Biochemical System Simulation Model

As an example, for the given theory, a system of two main reactions has shown below:

\[ 2A \overset{k_1, k_2}{\underset{k_3, k_4}{\rightleftharpoons}} B \]  

\[ B + C + E \overset{k_5}{\rightleftharpoons} D \Rightarrow B + C + E + F \]

where, in the two-way reactions, the first reaction rate represents the forward reaction rate, and the second term shows the backward reaction rate. The initial concentrations for the given chemicals and the reaction rates associated with the reactions have been shown in Table 3-1 and Table 3-2.

**Table 3-1 The initial concentrations of the chemicals participating in the example system**

<table>
<thead>
<tr>
<th>Reaction Rate Names</th>
<th>$k_1$</th>
<th>$k_2$</th>
<th>$k_3$</th>
<th>$k_4$</th>
<th>$k_5$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Value</strong></td>
<td>1</td>
<td>0.1</td>
<td>1</td>
<td>1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

**Table 3-2 The reaction rates of the given example system**

<table>
<thead>
<tr>
<th>Chemical</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Initial Concentration</strong></td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>30</td>
<td>0</td>
</tr>
</tbody>
</table>

Equation 7 shows a one line representation of two chain reactions. For the ease of understanding, one can separate the two reactions. Each of the chemicals participating in this system has its own ODE. These ODEs are calculated using the Equations 4 and 5 and is demonstrated in Equations 8-13.

\[
\frac{d[A(t)]}{dt} = -2(k_1[A]^2 - k_2[B]) \]  

\[
\frac{d[B(t)]}{dt} = (k_1[A]^2 - k_2[B]) - (k_3[B][C][E] - k_4[D]) + k_5[F] \]  

\[
\frac{d[C(t)]}{dt} = k_5[D] - (k_3[B][C][E] - k_4[D]) \]
\[
\frac{d[D](t)}{dt} = k_3 [B][C][E] - k_5 [D] \quad (11)
\]

\[
\frac{d[E](t)}{dt} = k_5 [D] - (k_3 [B][C][E] - k_4 [D]) \quad (12)
\]

\[
\frac{d[F](t)}{dt} = k_5 [D] \quad (13)
\]

After simulating this example system, the concentrations for 10 iterations with the step size of 0.1 second are computed using Equation 2 and the results are illustrated in Table 3-3.

**Table 3-3 The concentrations of the chemicals participating in the example system**

<table>
<thead>
<tr>
<th>Iteration Number</th>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>10.000</td>
<td>0.000</td>
<td>1.000</td>
<td>0.000</td>
<td>30.000</td>
<td>0.000</td>
</tr>
<tr>
<td>1</td>
<td>8.000</td>
<td>0.640</td>
<td>1.000</td>
<td>0.191</td>
<td>29.808</td>
<td>0.000</td>
</tr>
<tr>
<td>2</td>
<td>6.721</td>
<td>0.938</td>
<td>0.811</td>
<td>0.414</td>
<td>29.583</td>
<td>0.001</td>
</tr>
<tr>
<td>3</td>
<td>5.820</td>
<td>1.117</td>
<td>0.591</td>
<td>0.603</td>
<td>29.393</td>
<td>0.001</td>
</tr>
<tr>
<td>4</td>
<td>5.145</td>
<td>1.255</td>
<td>0.403</td>
<td>0.745</td>
<td>29.250</td>
<td>0.002</td>
</tr>
<tr>
<td>5</td>
<td>4.618</td>
<td>1.378</td>
<td>0.264</td>
<td>0.842</td>
<td>29.152</td>
<td>0.003</td>
</tr>
<tr>
<td>6</td>
<td>4.194</td>
<td>1.495</td>
<td>0.167</td>
<td>0.906</td>
<td>29.089</td>
<td>0.004</td>
</tr>
<tr>
<td>7</td>
<td>3.845</td>
<td>1.606</td>
<td>0.104</td>
<td>0.944</td>
<td>29.050</td>
<td>0.005</td>
</tr>
<tr>
<td>8</td>
<td>3.553</td>
<td>1.710</td>
<td>0.066</td>
<td>0.967</td>
<td>29.028</td>
<td>0.006</td>
</tr>
<tr>
<td>9</td>
<td>3.304</td>
<td>1.806</td>
<td>0.044</td>
<td>0.979</td>
<td>29.015</td>
<td>0.007</td>
</tr>
<tr>
<td>10</td>
<td>3.089</td>
<td>1.894</td>
<td>0.032</td>
<td>0.986</td>
<td>29.009</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 3-3 shows the gradual change in the concentrations of the chemicals participating in the example system.

### 3.2 Modeling of Metabolic Pathways

The tool that is represented in this thesis is particularly developed for the analysis of the metabolic networks. The metabolic networks consist of various pathways that are categorized as the metabolic pathways. The knowledge on the reactions inside a cell is
limited to the areas as the metabolic, genetic information processing, environmental information processing, cellular processes, systems or living organisms, human diseases, and drug development [20]. The focus in this thesis is on the metabolic pathways. In these pathways, unlike the signaling pathways, the chemical reactions play more important roles in the intracellular interactions, as they focus on releasing energy from the metabolites using the chemical reactions to help the continuing operation of the vital processes inside the cell. Using the KEGG database [20], the main branches of the metabolic pathways are extracted. The database is available online and although it is still under growth, its influence in different aspects of cell metabolism is not deniable.

In the KEGG database, the metabolism pathways are divided into 12 categories:

1. Carbohydrate Metabolism
2. Energy metabolism
3. Lipid metabolism
4. Nucleotide metabolism
5. Amino acid metabolism
6. Metabolism of other amino acids
7. Glycan biosynthesis and metabolism
8. Metabolism of cofactors and vitamins
9. Metabolism of terpenoids and polyketides
10. Biosynthesis of other secondary metabolites
11. Xenobiotics biodegradation and metabolism
12. Chemical structure transformation maps
In this thesis, we have used the carbohydrate, lipid, and amino acid pathways for most of our research. Although the pathways usually crosstalk with each other, the general schema of the model to be implemented for this research as a case study has been focused on the given three pathways. Figure 3.4 shows the general schema of the current KEGG database for the metabolic pathways. As it is shown, it represents a very complex network, in which the various pathways are mapped over each other.
Figure 3.4 The complete schema of metabolic pathway presented by KEGG
Each node in this map demonstrates a specific specie in the list of reactions of all the known pathways. There are some species which are participating in different pathways. These species are hard to predict, as there is not enough information on their behavior in different pathways.

The protein, lipid, and glucose pathways are the main pathways that will be used in the modeling of a sample metabolic pathway as a test case for the proposed tool. The three pathways crosstalk with each other in some points. The Glucose is the main specie that contributes in the carbohydrate related metabolism and has the most effect in the energy generation inside the cell. The amino acids are found inside the proteins, and the lipids are found in fats. Each of these contribute to the whole model and the change in them may change the outcome of the whole network. As stated earlier, the metabolic networks are very complex and it is not possible to predict the behavior of a specific specie based on an incomplete model. The purpose of this study is to provide a strong tool to analyze the big models with ease and providing enough flexibility to the user. The general model used in this thesis for the required metabolic pathways are shown in Figure 3.5. These pathways are the main pathways that contribute in the generation of ATP and also in glucose regulation. ATP can be a good representation of the energy level inside the cell as ATP regulates the glucose consumption. At first, ATP is consumed to make some products from glucose, then in other steps, ATP is generated at the end. One of the main ATP production pathways are the TCA cycle in which, the generated ATP imposes the ATP consumed in the glycolysis pathway earlier.
This model can be used to analyze different metabolites. We used this model in two different projects of analyzing the user’s dietary intake [44] by developing a mobile application to observe the person’s diet, and also in glucose insulin regulation. In the first project, the user is able to take a picture of the food and get the nutrient information, which includes the main groups of carbohydrates, lipids, and proteins. These values alongside with the model imported with the SBML, make the food intake information to affect the simulations.
Chapter 4

4 Implementation of the Tool

Based on the theory that we illustrated in the previous chapter, the tool was implemented in Java. The implementation of each feature into the program had its own difficulties, which solving any of them required more attention. Some arguments are added into the tool for the user inputs, which their complete list is shown in Table 4-1.

Table 4-1 The complete list of the arguments

<table>
<thead>
<tr>
<th>Argument</th>
<th>Description</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;-d&quot;</td>
<td>The simulation duration</td>
<td>In seconds</td>
</tr>
<tr>
<td>&quot;-s&quot;</td>
<td>The number of steps</td>
<td>The total number of iterations</td>
</tr>
<tr>
<td>&quot;-m&quot;</td>
<td>The simulation method</td>
<td>Taylor as default</td>
</tr>
<tr>
<td>&quot;-w&quot;</td>
<td>Observed specie</td>
<td>One name in SBML model</td>
</tr>
<tr>
<td>&quot;-f&quot;</td>
<td>The SBML model</td>
<td>To be imported to program</td>
</tr>
<tr>
<td>&quot;-e&quot;</td>
<td>Associated genes</td>
<td>The genes expressing each enzyme</td>
</tr>
<tr>
<td>&quot;-g&quot;</td>
<td>Gene expression levels</td>
<td>Gene expression levels for each gene</td>
</tr>
<tr>
<td>&quot;-i&quot;</td>
<td>Food intake information</td>
<td>Total ingredients, effective iterations, and subject's weight</td>
</tr>
<tr>
<td>&quot;-c&quot;</td>
<td>External initial concentrations</td>
<td>External concentrations to replace SBML initial concentrations</td>
</tr>
</tbody>
</table>
4.1 Structure of the Program

The structure of the program follows model-view-controller structure. The user uses the controller packages to model the problem and at the end, a view will be provided to the user in a numerical tab delimited format. This structure is provided in Figure 4.1 that follows the model-view-controller [45] structure. All the packages are connected clearly and the classes are defined accordingly to follow the structure presented in the above figure.

![Figure 4.1 Program Structure of the proposed tool](image)

4.2 Data Extraction

A class and its related methods were defined to extract all the required data from the user input models and values. This class used the Java library called jsbml to read all the data inside the SBML model.
4.2.1 SBML Model Parser

The first feature that is needed to be considered for the stand-alone metabolic simulation tool is the ability to have the SBML model as an input. A SBML model can be generated by any way or tool, and as long as it is in the standard form, it follows the requirements necessary for the data parsing.

As mentioned earlier, the SBML model is in XML format. The data in this format is stored using non-relational databases. In these databases, the information from the model must be imported using the associated methods in a Java library to parse SBML files called jsbml [46]. In this library, all of the functions to extract the reactions, species, rates, initial concentrations, direction of the reactions, their reversibility and so on are implemented.

Once the values were extracted, they will be stored in linked hash tables, as using dictionaries for the non-relational databases increase the speed and efficiency. The extracted values were appointed to the calculations with no order, while the species were ordered chronologically. The keys to each hash table varies by its type. The data associated with the species have the specie names as their keys, while the data associate with the reactions have the reaction names for their keys. For here, the standard KEGG annotations were used for both species and reactions in the test case model.

4.2.2 External Concentration Import

The tool is also flexible to import any external initial concentration if it does not conform with the SBML model initial concentrations. This feature has been added to test any given SBML model for several times with different initial concentrations. The values
imported to the program should follow the tab delimited structure having the specie name and the initial concentration value in mili mole per mili liter.

### 4.3 ODE Calculation

One of the most interesting features that was added to the tool is the ability to extract the ODEs for each specie in string or numerical formats. The ODE is a representation of the gradient change in the concentration of that specific specie at a given time. The program finds all of the related values to the specie and follows the equations 4 and 5 to make the ODEs. One of the sample ODEs for a specie having KEGG ID of C00267 also known as alpha-D-Glucose is shown in Figure 4.2. This ODE is extracted from a very complex model of almost 123 species and 85 reactions.

$$
\text{ODE of C00267 is:} \\
\quad + (-1.0 \times 5.0 \times [C00002]^1.0[C00267]^1.0) + (1.0 \times 0.8 \times [C00008]^1.0[C00668]^1.0) + (1.0 \times 1.0 \times [Exo70]^1.0[CIP4_2]^1.0)
$$

*Figure 4.2 A sample ODE extracted by the proposed stand-alone tool*

In the given ODE, the brackets are the representation of concentrations for each species. This ODE shows how rate of change in C00267 is related to the concentrations of C00002, C00008, C00668, Exo70, CIP4_2, and itself and also the related reaction rates.

Other than the string form, the ODE data is extractable as hash tables. The values, the signs, the directions, and the species involved are stored for every specie to be used in the simulations later.

### 4.4 Gene Expression Data Implementation

One of the very important factors in the biochemical systems’ simulations is the rate of the reactions. Each reaction has a specific rate. These rates are highly related to the
enzyme activity associated with each reaction in any direction. The enzymes are proteins which their structure is forced by some genes. Each gene can be associated with some enzymes and vice versa. This means that the expression level of each gene can affect the activity of some enzymes.

Although the gene expression levels are not always concordant with the enzyme activities, the observation of some persistent gene expression increase or decrease in a long period of time can be a sign of change in the associated enzyme activities. Therefore, it will be smart to consider a logical method to take into account the change of the expressions during those periods. Considering the expression levels at time 0 as the base level, the enzyme activities are assumed to be changing by the expression levels at the future intervals. This process takes place right before the numerical simulations start.

The method proposed in this thesis considers linear change in the gene expression levels when. The intervals between each expression level is given by the user. Each interval is divided into sub-intervals by user’s preference. In this case, the change can be modeled linearly, but to fill the gap of knowledge for the intervals in-between, some randomness is considered into the calculations. For now, maximum of 10% is assumed to be a good randomness in deviation from this linear behavior. Figure 4.3 shows this idea of considering gene expression levels into account.

![Figure 4.3](image)

Figure 4.3 Effect of the gene expression on the reaction rates
4.5 Food Intake Implementation

The developed program is maintained to consider the food intake information in simulations. The food contains four main ingredients of carbohydrates, sugar, lipids and protein. For each of them, different approaches were considered to make the values imported by the user feasible to be used in the metabolic simulations.

The carbohydrates and sugars are mostly considered as alpha-D-Glucose. This metabolite is the main specie in the chain of glucose transformation into energy. The weight of the subject under study is also imported by the user. According to [47], for human subjects 7% and for mouse subjects 5% of the body weight can be considered as the total amount of blood in the body. Using these values, the added carbohydrate concentration can be summed up with the current alpha-D-Glucose in any iteration. Every iteration will be calculated using these values until it reaches to the final iteration of the food intake. The total number of food intake iterations is a user input value. The user should justify this number using the experimental data.

The lipids are normalized in another way. Most lipids are 4 to 28 carbon structures called long chain fatty acids [48]. Most of the large fatty acid molecules do not cross the cell membrane due to their big size. Therefore, the medium level and smaller fatty acid molecules are the ones that pass this gate. This gives a rough estimation on the total number of lipid molecules. By having the subject weight and the lipids molecular weights and the amount of blood in the subject, the added concentration is calculated and will be added to the given iterations.

The proteins are different in some ways. The proteins are made of amino acids. The amino acids separate from the structure of the protein after the digestion. These amino acids
go to the blood stream and add up to the current amino acid deposit inside the cell. Five most common amino acids were extracted using [49] as the most observed amino acids in digestable proteins. Table 4-2 shows the considered amino acids and their molecular weights.

**Table 4-2 Some of the most common amino acids and their molecular weight**

<table>
<thead>
<tr>
<th>Amino Acid</th>
<th>Molar Weight (grams)</th>
<th>KEGG Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Serine</td>
<td>105.09</td>
<td>C00065</td>
</tr>
<tr>
<td>Valine</td>
<td>117.15</td>
<td>C00183</td>
</tr>
<tr>
<td>Isoleucine</td>
<td>131.18</td>
<td>C00407</td>
</tr>
<tr>
<td>Lysine</td>
<td>146.19</td>
<td>C00047</td>
</tr>
<tr>
<td>Tyrosine</td>
<td>181.19</td>
<td>C00082</td>
</tr>
</tbody>
</table>

The total amount of proteins in the food intake is converted to the total amino acids molecule count, and then the total concentration increment in the blood will be calculated.

\[
C_{i,n+1} = C_{i,n} + \frac{C_g}{MW_i \times W \times BP}
\]

Equation 14 shows how the new concentrations after the protein intake is calculated. \(C_{i,n+1}\) is the \(n + 1^{th}\) iteration concentration of the metabolite \(i\), \(C_{i,n}\) is the same parameter for iteration \(n\), \(C_g\) is the gram weight of the injected metabolite, \(MW_i\) is the molecular weight of the metabolite, \(W\) is the body weight \(kg\) and \(BP\) is the blood percentage in the body for that specific specie found in [47].

4.6 The Workflow

By gathering all the information required for the simulation of metabolic networks and developing the methods like ODE extraction, a workflow was developed to be used in the implementation of the tool. Figure 4.4 shows the proposed workflow of the tool.
At first, the arguments are imported in the tool by the user. These arguments express the duration of simulation, step sizes, the input files, and the food intake information. Then, the data from the SBML model will be extracted to find all the required fields necessary for simulation. The ODE will be extracted using this data. At this step, the initial concentrations and the gene expression data are also added to the ODE model. At this state, the tool is ready to start the simulations. For the given food intake information, the tool updates the simulations having the information about the dietary behavior of the subject under study.

![Diagram](image)

**Figure 4.4** The workflow of the proposed stand-alone metabolic simulation tool

The proposed tool is available in the SBBI server at the address https://sbbi.unl.edu. In the page, a download request form gives access to the jar file of the program along with the instructions on how to run different simulations is placed.
Chapter 5

5 Validation and Results

After implementation, using the theory and structures from the last chapters, we tested the tool on different cases. At first, we prepared a model to test our tool. Then, we used the model in two different projects with some changes. The results showed that the tool is able to present the required output given the input data.

5.1 Model Preparation and Simulations

A SBML model was developed using the available databases to be used as a case study for the proposed tool. Besides, the simulations were performed to test the abilities of the tool.

5.1.1 SBML Model Preparation

Using the KEGG database, a list of related metabolic pathways was studied and added to the local database to make a proper test case for the given stand-alone tool. To do so, glycolysis, TCA, Pyruvate to Acetyl-CoA, Gala-Degradation, Oxidative phosphorylation, lipids and amino acid pathways were extracted. All of the reactions were listed and all of the participating species were added to the program’s database. The direction of the
reactions and their reversibility were other parameters that were extracted using the literature and KEGG step by step.

The initial concentrations of the species and rates of the reactions were reached using HMDB and [21], [50] and BRENDA which is an enzyme information system. A web crawling program was developed to help the search in the online databases. The data acquisition process ended up with a very complex network of 123 species and 85 reactions.

When the chemical reactions database was ready, we extracted a SBML model using COPASI as an easy tool which provides a friendly interface to import a system of chemical reactions and export the model. The SBML helped us to compress all of the information into a single database format.

Besides, we used the gene expression level data from mouse to help test the model under the effect of gene expressions alongside the case with no expression available. The tool helps to observe the change in any of the species participating in the model.

5.1.2 Case Study Simulations

The tool was tested with the prepared model. Because of the model’s complexity, a simple prediction for the behavior of the reactions’ substrates is not possible without a simulation tool. Therefore, the program that we developed helps to have a better understanding on the behavior of the biochemical systems with more ease. The running settings of this program are as follows:

- The simulation duration: 6000 seconds
- Number of steps: 600
- The specie to be observed: alpha-Aminoadipoyl-S-acyl
• Gene expression data import: Three external files
• Food intake import: 4.25, 11.55, 74.83, 42.62 (Car., Su., Li., Pr.)
• Steps of the food intake: 260
• Weight of the study subject: 180 lb

Figure 5.1 shows the concentration of alpha-Aminoadipoyl-S-acyl for the given settings. Two different cases were studied in which, for the blue line, the gene expression data was considered into account using the proposed method, and for the red line, no gene expression level change was considered. By the simulation results, we conclude that, this metabolite is active in the Lysine metabolism which its concentration gives information on how Lysine metabolism is active in the given model.

In absence of the gene expression data, this protein is more abundant at the first 3500 seconds, while it drops dramatically after its peak at around 1800 seconds. The correlation of the simulated data can be modeled experimentally to adjust every metabolic model and help the researchers to consider more parameters while studying the metabolic networks.

Although there is not enough experimental data or a complete model in hand to proof the correctness of the simulations, the current simulation result shows a very good response given the prepared model. The gene expression data helped the faster convergence to the final value for the given specie. This result makes more sense as we expect the higher activity at the beginning of the chemical reaction process.
Figure 5.1 The change in concentration of alpha-Aminoadipoyl-S-acyl as one of the species participating in the model for 100 minutes

A similar behavior was observed for Adenosine 5-monophosphate (AMP), which is one of the first products of the glucose synthesis in glycolysis pathway, to help the reaction from alpha-D-glucose to alpha-D-glucose 6-phosphate. The change in the concentration of AMP shows that, although the glucose degradation is fast, the model without the gene expression data does not show the associated reaction as fast as it should be. The reaction of R09085 of AMP from KEGG, and the alpha-D-glucose and AMP responses are shown in Figure 5.2, Figure 5.3, and Figure 5.4 respectively. As glucose level goes higher, the product should go higher, as AMP is produced again in this pathway, but without the gene expression data, this increase is not significant nor does not show the peaks in the alpha-
D-glucose concentration. Thus, the model considering the gene expression levels is more effective and is another proof that the proposed tool can be effective in analysis of the models with the proper inputs.

**Figure 5.2** The reaction associated with degradation of alpha-D-glucose to alpha-D-glucose 6-phosphate. C00267, C00008, C00020, and C00668 represent alpha-D-glucose, ADP, AMP, and alpha-D-glucose 6-phosphate.

**Figure 5.3** The change in concentration of alpha-D-glucose as one of the species participating in the model for 100 minutes.
Figure 5.4 The change in concentration of AMP as one of the species participating in the model for 100 minutes

Another interesting behavior was observed in the concentration of Ammonia in the given model. Ammonia’s trend followed another direction than alpha-Aminoadipoyl-S-acyl in presence of gene expression data. Ammonia showed higher concentrations compared to the presence of gene expression data. Figure 5.5 demonstrates this simulation for 100 minutes, which shows how the proposed tool is effective in analysis of any specie participating in a form of SBML model given the proper inputs and arguments.

In the results, it can be seen that the Ammonia concentration is higher in the presence of gene expression data. This proves that, the tool not always considers higher activity at the beginning of the reaction for the given gene expression level data and shows a better result for the given specie as it is converging to a stable point for the given input compared to the case where it is diverging.
Figure 5.5 The change in concentration of Ammonia as one of the species participating in the model for 100 minutes

5.1.3 Time Complexity

Based on the nature of this problem, the complexity of the algorithm that we followed is polynomial. This program is highly dependent to the size of the input model. With increase in the size of the model, the speed drops, but this is not a significant amount for the smaller models. The tests showed that for the simulations without the gene expression analysis, the simulations take less than two seconds, while the gene expression analysis increases the simulation time to more than 4 seconds for the model proposed in this section. Although this processing time is not high, for the bigger models the process will be much more costly. The processing time for 123 species and 85 reactions are shown in Table 5-1.
Table 5-1 Processing time in presence and absence of gene expression data

<table>
<thead>
<tr>
<th></th>
<th>Number of Cores</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Processor</td>
<td>2.2 GHz</td>
<td></td>
</tr>
<tr>
<td>RAM</td>
<td>16 GB</td>
<td></td>
</tr>
<tr>
<td>Processing Time</td>
<td>No Gene Expression Analysis</td>
<td>1.8 Sec</td>
</tr>
<tr>
<td></td>
<td>With Gene Expression Analysis</td>
<td>4.5 Sec</td>
</tr>
</tbody>
</table>

5.2 Mobile Based Diet Monitoring System

Another implementation of the program was to provide a model to enable the use of this simulation tool for those who are using mobile-based food logging systems. In this implementation, an application was developed [44] to help the user in choosing the best dietary style by first, using the image analysis tools to predict the type and the amount of the food that the user is consuming, and then by simulating the effect of the taken food on the user’s body. In this project, new methods were developed to increase the accuracy of the food detection, and then a library in Java was developed to be used as the core of processing unit to consider the personalized SBML model and simulate the behavior of ATP, Glucose and effect of the loads on the cell. A cell can be a very good representation of the body. Higher or lower energy levels in the cell can be problematic and the maintained interval is preferred. Figure 5.6 Shows the response of the model to a specific feeding behavior.

Figure 5.6 illustrates the body’s response to starvation and to the intake of three meals (breakfast, lunch and dinner) by the user with the moving window of 100 seconds. In this
example, the user has consumed 118 grams of banana (1 medium size) with 244 grams of 2% milk (1 cup) as the breakfast at 8am, 185 grams of white rice with 100 grams of chicken breast for the lunch at 12pm, and 182 grams of apple with 178 grams of pear at 6pm as dinner. Here, we set 100 seconds as interval in a period of 1 hour for each small portion of the meal to be absorbed.

After running the simulations, the ATP and glucose concentrations are shown starting from 8am within a day. In Figure 5.6(a), the glucose concentration is shown for the entire day in the starvation state. Once starvation starts, the free glucose is consumed at first; then, glycogen starts to be consumed for the next 8 to 10 hours to maintain the glucose level in normal condition. If glycogen is completely burned, the third phase will start by consuming the fat resources. Figure 5.6(b) and (c) show the trend of glucose and ATP concentration with three different meals during a day, respectively. In Figure 5.6(d), an arbitrary periodic load was applied to the system to mimic the ATP consumption captured by the exercise tracker. At the end in Figure 5.6(e), the overall ATP balance after applying the load has been demonstrated where the ATP concentration goes to the same level at 8am of the previous day. This information is helpful to alert the user to adjust his/her diet and exercise to manipulate the energy balance.
Figure 5.6 (a) Glucose concentration in starvation state, (b) Glucose concentration for a normal diet of three main meals within a day, (c) ATP concentration of the normal diet, (d) the load applied to mimic the activity (the ATP consumption) captured by exercise tracker, and (e) the ATP concentration for one day after applying the given load.

This response, generated using a 100 second moving window, is a very good example of the proposed tool’s abilities. The model used in this implementation is a flexible SBML file. It can be updated later with more versatile one having more parameters with better accuracy.
5.3 Glucose-Insulin Model Simulations

The integration between glucose and insulin is an important role in integration of metabolic and signaling pathways. To study the model described in the previous section, another layer, to study the glucose-insulin relation, was added to the model to show the abilities of the proposed model other than the concentration estimation.

We added the insulin signaling pathway extracted from KEGG to the test model. In the new model, the insulin concentration in the blood will affect the glucose concentration using some genes such as APS, Cb1, and GRF2 to affect the glutamine in the carbohydrate pathways, after the signaling pathway is triggered by the INSR which is the insulin receptor. The model extracted from KEGG is shown in the figure below.

Figure 5.7 The insulin signaling pathway extracted from KEGG
Because of the complexity of this model, only a few steps were added to the SBML model to add a delay in glucose concentration in the reference glucose. Insulin is an important parameter in the glucose regulation. The high level of insulin in the blood triggers chains of reactions in which, the glucose will be absorbed inside the cell leading to intake of the extra glucose from the blood stream. Without the insulin regulatory procedure, the glucose can increase in the blood stream no regard to its vital limits. We studied this effect in another research and we present it in this thesis as a case study.

The metabolic model and the developed tool was used in the study of glucose and insulin correlation. In the simulations having two food intakes at time 1 and 400 seconds, the study showed that the insulin follows the trend of glucose. Another proof that the provided tool is able to be used in different studies. The delay of insulin in following of glucose is mostly because of some extra reactions defined in the database.

![Blood Glucose and Insulin](image)

**Figure 5.8** Glucose-Insulin correlation tested using the proposed model
5.4 OSG Implementation

An implementation of the given tool was made on Open Science Grid (OSG) to perform parallel analysis on several samples. OSG provides a very good support for parallel computing. We used this ability to test the program for several test cases of the same model and various initial concentrations. This method can also be followed to test the tool with numerous models, but making a new model is extremely costly and is not always dependable.

In the OSG implementation, a stand-alone jar file was generated to be addressed in the simulations. The other models, initial concentrations, and gene expression data, feedings and other required documents imported into the OSG. A script was written to extract the files and run the jar file having the arguments and 100 random initial concentration instances. After the simulations, the results will be compressed and returned to the user. Although the tool expressed in this thesis is fast in biochemical simulations, using OSG increases the ability to test on high numbers of different test cases after the data is gathered from various databases. Some different test cases showed the necessity of the given approach in biochemical simulations. Of course, the improvement in the SBML model can lead to a more accurate simulation result.
Figure 5.9 Two OSG simulation results for high initial glucose concentrations

Figure 5.9 shows that how parallel computing helps to detect the special cases with special behaviors. As it is shown, the simulation is run for 10 minutes on the proposed metabolic SBML model. The result expresses that the high abundance of glucose leads to lower ATP level as at first, it consumes all of the ATP. While the glucose is gradually decreasing, the ATP is compensated and reaches higher values.
Figure 5.10 Two OSG simulation results for high initial glucose concentrations

Figure 5.10 shows the results for some moderate level glucose initial concentrations. It is demonstrated that even though the initial concentrations are in a more moderate level, the systems response follows the same pattern. In this case, the moderate glucose level drops to a certain point. This will consume the ATP which makes the concentration of ATP to drop rapidly to very low values. As ATP is no longer highly available, the reverse glycolysis pathway takes more glucose. This pattern can be seen in some of the other peaks in the demonstrated results.
Chapter 6

6 Conclusion and Future Work

The main focus in this thesis was to present a dependable stand-alone java developed tool to help the simulation of metabolic biochemical networks. This tool was expected to be a stand-alone unit able to import any SBML model, any necessary external source or input like adjusted initial concentrations, gene expression level data, simulation duration, number of step sizes and the number of iterations, and the food intake data to simulate the concentration of every metabolite in the network using Taylor method.

The tool was implemented in Java to be used in the terminal environment other than the IDEs, using the extracted jar file. The arguments were defined to accept any necessary information in the text or xml formats. The program was given the ability to extract all of the information necessary for the simulations from the SBML model. For this aim, the jsbml library was used, which is the most complete library found to read the SBML data in Java.

The ability to read the gene expression levels, in order to be effective in the simulations, is another important feature of the proposed tool. The genes associated to each enzyme were found and their expressions were modeled to represent the change in the activity of the enzymes participating in the reactions in each direction.
To check the tool, a model was developed using the necessary main pathway data extracted from KEGG. All of the reactions were listed; all of the species were marked; and the different literatures were studied to get the model characteristics and a web crawling program was developed in python to extract the initial concentration data from HMDB [50]. Furthermore, the reaction rates were assigned from the ranges found in Brenda, a comprehensive enzyme information system which was designed at the technical university of Braunschweig.

Different simulations were tested using the prepared model and the proposed tool. Simple to more complex models were the test subjects to check the feasibility of the given simulator. Besides, this tool and the model along with it were used in two different projects which their manuscripts were sent to be published accordingly.

For the future, there are some suggestions are needed to be made from the tool design and implementation, to the model design, extraction, and estimation. The next generation of this tool can be more flexible with the numerical simulation methods. At this stage, only Taylor method has been considered to be the main method available to the users. The powerful methods like different orders of Runge-Kutta [51] are very good examples of gradient methods to appear in the arguments.

Another development is to have a better model estimation. Due to high complexity of the metabolic networks, finding the best parameter assignments to the model parameters makes a big difference in the study and analysis of the metabolic networks. Most of the metabolic parameters, including the rates and the initial concentrations, are not a constant number and only conform a particular range for any reaction or specie. Even in a specie,
the ranges vary based on the environment that a subject is tested. Besides, the model that represents a normal metabolism of a cell can contain more pathways which helps it to be more general to be considered in other studies, too. Adding more pathways having dependable parameters which are extracted from experimental researches will have a more comprehensive result.

The tool, that we developed for this thesis, is designed for the analysis of one cell. Certainly, the response for a pack of cells will be different as it depends on the distribution of nutrients inside the organism. The faster metabolite gets distributed inside the body leads to a better similarity to the experimental results. This forces us to design the simulation method with a more efficient method or enhancing the parallel computing to analyze all of the participating cells at the same time.

The last but not least improvement in the tool can be in development of the gene expression level analysis method. For now, linear interpolation of the gene abundances seems to be working, but considering the huge number of genes and enzymes and their correlations, the introduction of a more advanced method seems necessary.
7 Bibliography


256, 1997.


