Genome-scale Integrative Modelling of Gene Expression and Metabolic Networks

A thesis submitted to the University of Manchester for the
degree of Doctor of Philosophy in the Faculty of Life Sciences

2011

Delali Anku Adiamah

List of Contents

Total word count: 46,979

List of Tables

List of Figures

Abstract

The elucidation of molecular function of proteins encoded by genes is a major challenge in biology today. Genes regulate the amount of protein (enzymes) needed to catalyse metabolic reactions. There are several works on either the modelling of gene expression or metabolic network. However, an integrative model of both is not well understood and researched. The integration of both gene expression and metabolic network could increase our understanding of cellular functions and aid in analysing the effects of genes on metabolism.

It is now possible to build genome-scale models of cellular processes due to the availability of high-throughput genomic, metabolic and fluxomic data along with thermodynamic information. Integrating biological information at various layers into metabolic models could also improve the robustness of models for *in silico* analysis.

In this study, we provide a software tool for the *in silico* reconstruction of genome-scale integrative models of gene expression and metabolic network from relevant database(s) and previously existing stoichiometric models with automatic generation of kinetic equations of all reactions involved. To reduce computational complexity, compartmentalisation of the cell as well as enzyme inhibition is assumed to play a negligible role in metabolic function.

Obtaining kinetic parameters needed to fully define and characterise kinetic models still remains a challenge in systems biology. Parameters are either not available in literature or unobtainable in the lab. Consequently, there have been numerous methods developed to predict biological behaviour that do not require the use of detailed kinetic parameters as well as techniques for estimation of parameter values based on experimental data. We present an algorithm for estimating kinetic parameters which uses fluxes and metabolites to constrain values. Our results show that our genetic algorithm is able to find parameters that fit a given data set and predict new biological states without having to re-estimate kinetic parameters.

Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning

Copyright Statement

1. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.

2. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

3. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions ofcopyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Propertyand Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Propertyand/or Reproductions.

4. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Propertyand/or Reproductions described in it may take place is available in the University IP Policy (see http://www.campus.manchester.ac.uk/medialibrary/policies/intellectualproperty.pdf), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's policy on presentation of Theses

List of Abbreviations

Enzyme Names

aceA - isocitrate lyase

aceB - malate synthaseA

ackA - acetate kinase

acn - aconitase

eno - enolase

fba - fructose bisphosphatealdolase

fum - fumarase

g3pdh - glycerol 3-phosphatedehydrogenase

gap - glyceraldehyde 3-phosphatedehydrogenase

glk – glucokinase

glt - citrate synthase

gnd - 6-phosphogluconate dehydrogenase

icd- isocitrate dehydrogenase

mdh - malate dehydrogenase

pdh - pyruvate dehydrogenase

pfk - 6-phosphofructokinase

pfl - pyruvate formate-lyase

pgi - phosphoglucose isomerase

pgk - 3-phosphoglycerate kinase

pgl - 6-phosphogluconolactonase

pgm - phosphoglucomutase

poxB - pyruvate oxidase

ppc - phosphoenolpyruvate carboxylase

pta - phosphate acetyltransferase

pts - phosphotransferase

pyc - pyruvate carboxylase

pyk - pyruvate kinase

rpe - ribulose phosphate3-epimerase

rpi - ribose-5-phosphate isomerise

sdh - succinate dehydrogenase

sucAB - 2-oxoglutarate dehydrogenasecomplex

sucCD- succinyl-CoA synthetase

tal - transaldolase

tkt1 - transketolase 1

tkt2 - transketolase 2

tpi friose phosphateisomerase

zwf - glucose-6-phosphate-1-dehydrogenase

Metabolite name

2PG (2PGA) - 2-phosphoglycerate

3PG (3PGA) - 3-phosphoglycerate

6PG - 6-phosphogluconate

AC – acetate

ACCOA - acetyl-CoA

ADP - Adenosine triphosphate

ATP - Adenosine-5'-triphosphate

BPG - 1,3-diphosphateglycerate

CIT - citrate

DHAP – dihydroxyacetonephosphate

E4P - erythrose-4-phosphate

F16bP - Fructose 1,6-bisphosphate

F6P - fructose-6-phosphate

FBP - fructose-1,6-bisphosphate

FUM - fumarate

G1P - glucose-1-phosphate

G3P - glyceral-3-phosphate

G6P - glucose-6-phosphate

GAP - glyceraldehdye-3-phosphate

GAPDH - glyceraldehyde-3-phosphate dehydrogenase

GLYX - glyoxylate

ICIT - D-isocitrate

MAL – malate

NAD - Nicotinamide adenine dinucleotide

NADH - Nicotinamide adenine dinucleotide, reduced

OAA - oxaloacetate

PEP - phosphoenolpyruvate

PYR – pyruvate

R5P - ribose-5-phosphate

Ru5P - ribulose-5-phosphate

S7P - sedoheptulose-7-phosphate

SUC - succinate

SUCCOA - succinyl-CoA

X5P - xylulose-5-phosphate

Acknowledgement

I would like to take this opportunity to acknowledge and give many thanks to Dr Jean-Marc Schwartz, my supervisor, for his continual help and direction. He is always available when needed. I would also like to thank my advisor, Dr Jim Warwicker, for his contribution to the work I am conducting. His comments and criticisms received are always welcomed. Additional thanks to Julia Handl for her help with the genetic algorithm and input with the first manuscript.

I would like to give thanks to BBSRC who provide me with funding, without which my research would not be possible. Finally, I will like to thank my beautiful wife, Nodumo, for her support and love shown to me during the difficult times while writing up. And a big thank you to my family, without them, I won’t be here.

Short Abstract

This thesis is structured into the following:

Chapter 1: gives a review of works that have already been done relative to this research. The advantages and disadvantages, including limitations of methods, of these works are given.

Chapter 2: We introduce the modelling concepts used in this research and present our software tool, GRaPe, for modelling integrative gene expression and metabolic networks. The results of this chapter appeared in "Streamlining the construction of large-scale dynamic models using generic kinetic equations." Delali A. Adiamah, Julia Handl and Jean-Marc Schwartz. *Bioinformatics*, 26 (10), pp. 1324 – 1331. Julia Handl provided the genetic algorithm. The rest of the work, including development of GRaPe was done by me.

Chapter 3: In this chapter, we validate our methodology by applying it to the yeast glycolysis pathway. The results of this chapter are also published in the above paper.

Chapter 4: In this chapter, we show that our method works even on a very large-scale model by applying it to a genome-scale metabolic model of the *M. tuberculosis* bacterium. We show that we are able to replicate different steady states using this approach. To be submitted.

Chapter 5: Here, we present an integrative model of *E. coli* central metabolism – comprising genes, mRNAs, proteins and reactions. We then show that we are able to represent several steady states using our methodology. We show the predictive prowess of our model by predicting different states in a gene knockout experiment. To be submitted.

Chapters 6, 7 and 8 we discuss our methodology, the overall success and limitations of our findings and methodology; and suggest possible future improvements.

Chapter 1

General Introduction

"I believe there is no philosophical high-road in science, with epistemological signposts. No, we are in a jungle and find our way by trial and error, building our road behind us as we proceed."

Max Born (1882-1970)

In this section, a review of current approaches of modelling gene expression and metabolic is presented. The review also covers the difficulties and challenges of modelling in biological systems and a discussion about computational modelling tools which are of relevance to this research.

1.1 Background Review

Gene expression and metabolic reaction networks are two principal components of living organisms. There are various works on either systems but the integration of both systems is not well researched (Yeang and Vingron 2006). Genes encode proteins that catalyse a metabolic reaction, a process known as gene expression. In gene expression, the gene serves as a template for protein synthesis, with mRNA as an intermediate product. Both mRNA and protein concentrations degrade over time at a constant rate.

In the post-genomic era, there is the need to build integrative models of gene expression and metabolic reaction networks to understand the role of genes and enzymes in cellular functions. Enzymes are expressed differently under different nutrient conditions or enzyme knock-outs. This suggests two logical interpretations 1) that metabolic reactions are controlled by the concentration of enzymes besides the concentration of substrates, 2) enzymes are indirectly regulated by metabolites (Yeang and Vingron 2006). Figure 1 shows the abstract representation of an integrative gene expression and metabolic reaction system.

With available genome and metabolic data, it is now possible to reconstruct metabolic networks integrated with genomic data. Metabolic network reconstruction and simulation allows for an in depth insight into the molecular mechanisms of a particular organism. A metabolic reconstruction involves the breakdown of metabolic pathways into their respective reactions and metabolites. Kinetic modelling provides the best accurate method of analysing the behaviour of systems by modelling the dynamical change of metabolites in a system. It requires kinetic parameters and reaction rate constants to yield the best results. However, the kinetic parameters and rate constants needed to fully define a model are often unavailable. As a result, there is the need to develop methods for either estimating kinetic parameters or find a way of measuring them in a systematic manner. An integral part of this study was to provide an optimisation technique for estimation kinetic parameters from time-series of experimental data.

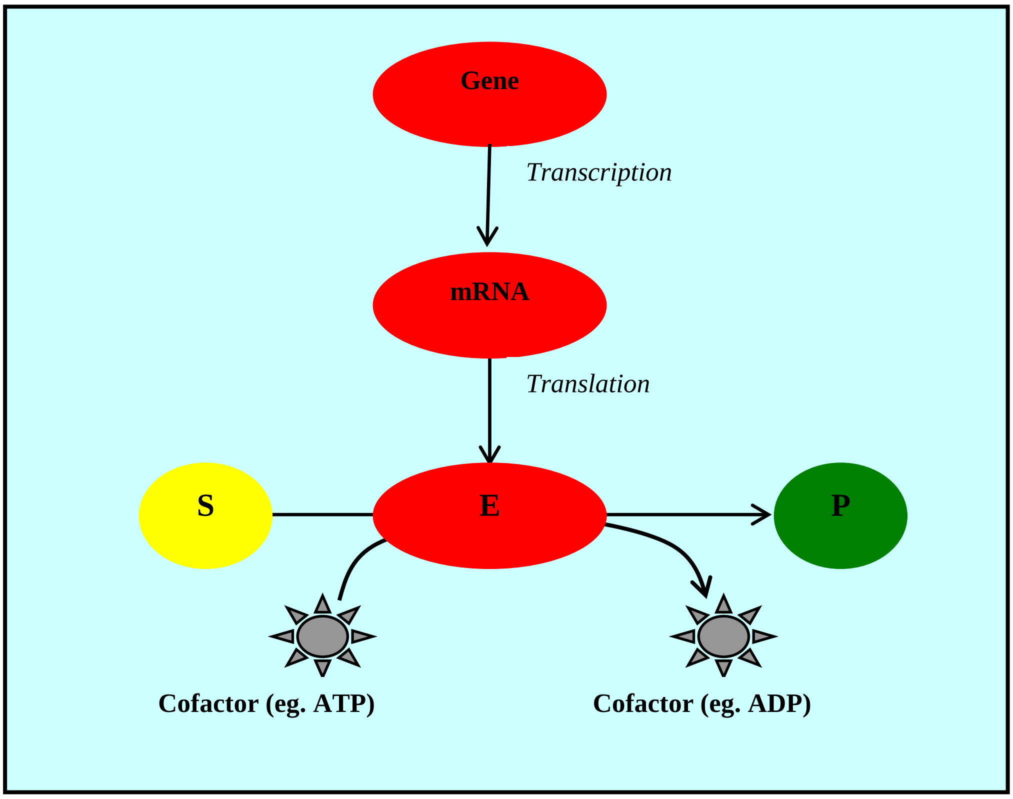


Figure 1: An abstract representation of an integrative model of gene expression and metabolic reaction. An enzyme, E, converts a substrate, **S**, into a product, **P**. The amount of E is not fixed but is a function of the gene and mRNA.

Current modelling tools allow metabolic networks to be reconstructed manually. Reconstruction on a genome-scale metabolic network is very tedious and time-consuming. For example, *Staphylococcus aureus* strain N315 consists of 619 genes that catalyse 640 metabolic reactions (Becker and Palsson, 2005). For dynamic simulation, the user additionally has to express rate equations for every reaction. We focus on rectifying this costly and time-consuming process by generating the reaction rate equation automatically for every reaction in the metabolic network.

One of the aims of this study was to provide a computational tool capable of reconstructing genome-scale integrative models of gene expression and metabolic network from relevant database(s) or some pre-existing stoichiometric models. The software will then generate the reaction rate equations automatically based on the number of substrates and products of each reaction. With SBML files accepted as the main format in systems biology, the software must be capable of reading and writing SBML files.

1.2 Gene Expression

Gene expression is a process that ends with the production of a protein needed to catalyse a particular reaction. The processes occurs in order i.e. the gene is first transcribed (by a process known as *transcription*) into messenger RNA (mRNA), which then gets translated (in a process known as *translation*) into the required amount of proteins.

1.2.1. Transcription

Transcription is the process by which genetic information from DNA is transcribed into messenger RNA (mRNA). mRNA serves as a template for protein synthesis. Transcription is divided into 3 stages: *initiation*, *elongation* and *termination*.

Initiation: RNA polymerase (RNAP, an enzyme that synthesizes RNA) binds to the DNA and unwinds the DNA. This creates an initiation bubble so that the RNAP has access to the single-stranded DNA template, together with other cofactors as shown in Figure 2 (Mathews and Ahern, 2000).

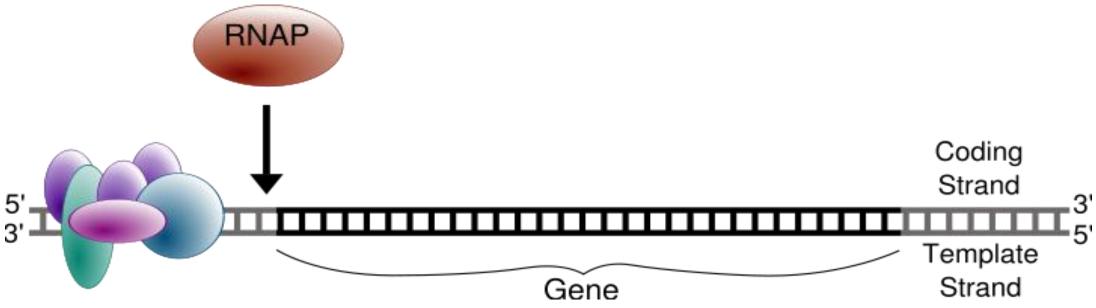


Figure 2: Initiation process of transcription.

Elongation: The template strand, which is one strand of the DNA, is used as a base template for RNA synthesis. As transcription proceeds, RNAP traverses the template strand and uses base pairing complementarily together with the DNA template to create a RNA copy as shown in Figures 3 and 4. Although RNAP traverses the template strand from 3' → 5', the non-template (the coding) strand is often used as the reference point, so transcription is said to go from 5' → 3'. This produces an RNA molecule from 5' → 3', which is an exact copy of the coding strand (with the exception that thymines, T, are exchanged with uracils, U, and the nucleotides are composed of a ribose (5-carbon) sugar where DNA has deoxyribose (one less oxygen atom) in its sugar-phosphate backbone) (Mathews and Ahern, 2000).

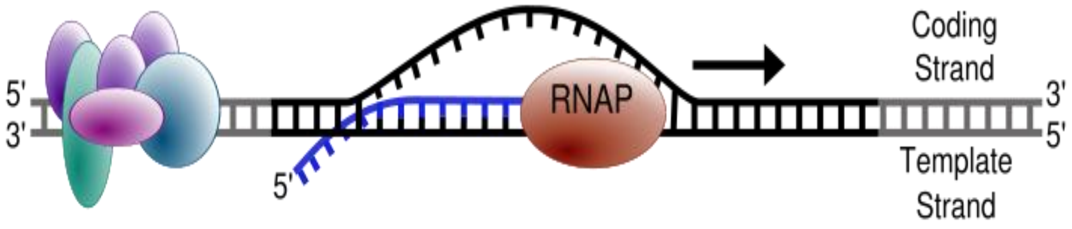


Figure 3: Elongation process of transcription.

Termination: The transcription process terminates when the newly synthesized mRNA forms a *hairpin loop*, followed by a run of *U*s (Figure C) (Mathews and Ahern, 2000).

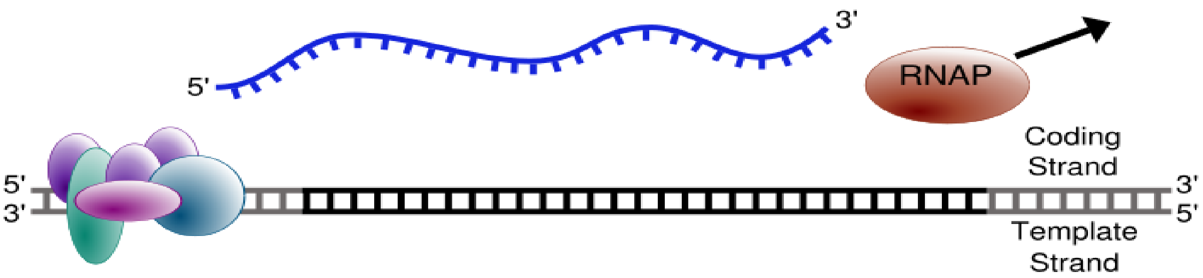


Figure 4: Termination process of transcription

1.2.2. Translation

Translation is the process where the mRNA is translated into amino acid sequence of a polypeptide, namely proteins. This is done by the ribosomes in the cytosol. The genetic information encoded by the mRNA is turned into amino acid sequence to form a polypeptide.

Gene expression has an impact on the ability of the cell to maintain vitality, perform cell division and respond to stimuli in its environment (Garcia-Martinez, Gonzalez-Candelas et al. 2007). During *transcription*, genetic information from DNA is transcribed into messenger RNA (mRNA). The mRNA serves as a template for protein synthesis. In *translation*, the mRNA is translated into amino acid sequence of a polypeptide, namely proteins. The full gene expression process is shown in Figure 5. In this research study, transcription and translation are considered as the main processes in gene expression. The individual stages in transcription, post-translation modification and regulatory elements in translation are considered to be outside the scope of this research study.

Garcia-Martinez *et al*., (2007), studied the relationship among six variables that characterize gene expression at the genome-level in living organism. These variables are *transcription rate* (TR) and *translation rate* (TLR), mRNA concentration (RA) and protein concentration (PA), and mRNA stability (RS) and protein stability (PS). Their studies concluded that the amount of both mRNA and proteins depends on their synthesis and degradation rates (Smolen, Baxter et al. 2003). This suggests that the concentration of mRNA alone is not a good predictor of the amount of proteins (Wolfe 1972).

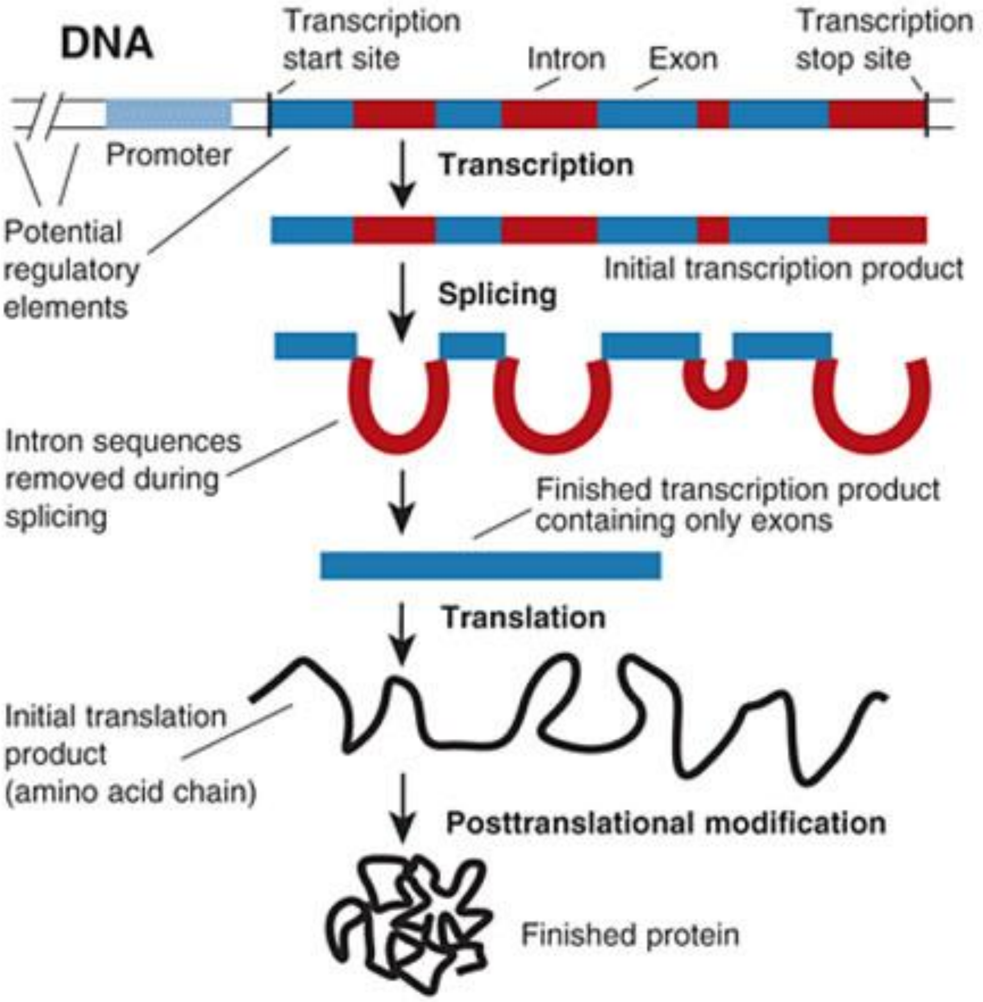

DNA
Transcription start site
Intron
Exon
Transcription stop site
Promoter
Potential regulatory elements
Transcription
Initial transcription product
Splicing
Intron sequences removed during splicing
Finished transcription product containing only exons
Translation
Initial translation product (amino acid chain)
Posttranslational modification
Finished protein

Figure 5. The gene expression process. During transcription, the genetic information coded on DNA is transcribed into mRNA (messenger RNA). The mRNA is then transferred from the cell nucleus into the cytoplasm where it undergoes protein synthesis to specify the amino acids that make up the proteins in a process known as Translation. (Figure taken from NCBI, source: http://www.ncbi.nlm.nih.gov/projects/genome/probe/doc/ApplExpression.shtml).

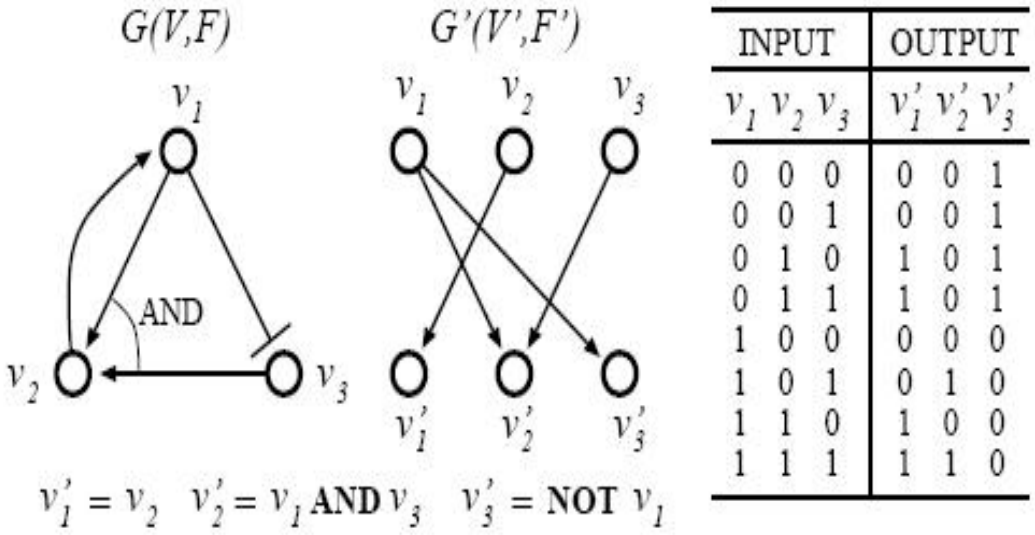
1.3 Methods for Modeling Gene Expression

Mathematical modelling and computer simulations can help us understand the dynamics of biological processes and also provide a platform for a variety of computational analyses (Feist and Palsson, 2008).

There are many computational methods used in modelling gene networks. Among them are Boolean networks (Akutsu et al, 1999; Kim, 2007; Li, 2007), ordinary differential equations (Chen, 1999), Dynamic Bayesian Networks (Murphy and Mian, 1999; Li, 2007), linear difference equations (D'haeseleer *et al*. 1999), state-space equations (Wu *et al*., 2004) and neural networks (Gagneur and Klamt ,2004). Among these methods, neural network is the least frequently considered technique due to its high computational time (Gagneur and Klamt, 2004).

1.3.1.Boolean Networks (BN)

A common approach to modelling gene expression is by Boolean Networks (BNs) where a gene has either one of only two states (binary on/off switch). The state of a gene at any time step is determined by a Boolean function of the state of some genes. The state of the network is defined as the *n*-tuple of 0s and 1s indicating if a gene is expressed or not at the particular moment. The total number of states depends of the number of genes present in the network. In total there are 2^n different possible states. For example a network consisting of 3 genes, the possible states will be (0,0,0), (0,0,1), (0,1,0), (0,1,1), …, (1,1,1) (Murphy and Mian, 1999; Schlitt and Brazma, 2007). An example of a BN is shown in Figure 6.



INPUT			OUTPUT		
v_1	v_2	v_3	v'_1	v'_2	v'_3
0	0	0	0	0	1
0	0	1	0	0	1
0	1	0	1	0	1
0	1	1	1	0	1
1	0	0	0	0	0
1	0	1	0	1	0
1	1	0	1	0	0
1	1	1	1	1	0

Figure 6: Simple Boolean network. A BN consists of a set of nodes (genes) *G (V, F)*, where *V = {v_1…v_n}*, and a set of Boolean functions *F = {f_1…f_n}*. The *f_i (v_{i1}…v_{ik})* is the Boolean function for the nodes *(v_{i1}…v_{ik})* and assigns the value to v_i. The state of the system at any time *t* +1 can be calculated from the state at time *t* with prior knowledge of the *f_i (v_{i1}…v_{ik})*. Yeang, C.-H. and M. Vingron (2006).

BN are able to reproduce features of biological systems, such as global complex behaviour, self-organization, stability, redundancy and periodicity and can explain the dynamic behaviour of living organism (Chen *et al.*, 1999; Kim 2007). Another benefit is that BN require no knowledge of kinetic parameters. This technique only considers the Boolean relationships between genes. For a large number of genes, BNs are very inexpensive in respect to computational complexity (Li *et al.*, 2007).

1.3.2. Bayesian Networks (Dynamic Bayesian Network)

Bayesian networks are a special case of graphical models in which nodes represent random variables, and the arcs represent dependence assumptions (Murphy and Mian, 1999).

Given a set of variables $U = \{X_1, X_2 ... X_n\}$ in a gene network, a Bayesian network, U is a pair $B = (G, \Theta)$ which encodes a joint probability distribution over all states of U. It is composed of a directed acyclic graph G whose nodes correspond to the variables in U and Θ, which defines a set of local conditional probability distributions to qualify the network.

For example, if an arc connects from node *A* to another node *B, A* is called a *parent* of *B*, and *B* is a *child* of *A*. The set of parent nodes, *U,* of a node X_i is denoted by parents(X_i) Given *G* and Θ, a Bayesian network defines a unique joint probability distribution over U of the node values that can be written as the product of the local distributions of each node and its parents as :

(1)

An advantage of Bayesian networks, like Boolean networks, is that they do not require the use of kinetic parameters. Their probabilistic nature makes them represent properties due to the random and unpredicted events that can occur (Murphy and Mian, 1999; Li, 2007). Dynamic Bayesian Networks (DBNs), an extension of Bayesian network analysis, is a general modelling approach that is capable of representing complex temporal stochastic processes (Li, 2007). The two main disadvantages of both BN and Bayesian networks are that 1) by treating genes as either “on” or “off”, some genes with regulatory effects but having an expression level outside the required threshold can be ignored or wrongly classified, and 2) for large gene-scale models, it is very computationally expensive and time-consuming.

1.3.3. Ordinary differential Equations (ODEs)

Bayesian networks like BN are poor in capturing some important aspects of network dynamics (Schlitt and Brazma, 2007). Gene expression, figure 7, can be modelled mathematically using differential equations (Li,*et al*, 2007). Differential equations allow more details of the dynamics of the network by explicitly modelling continuous changes in concentration of molecules over time (Chen *et al.*, 1999; Schlitt and Brazma, 2007).

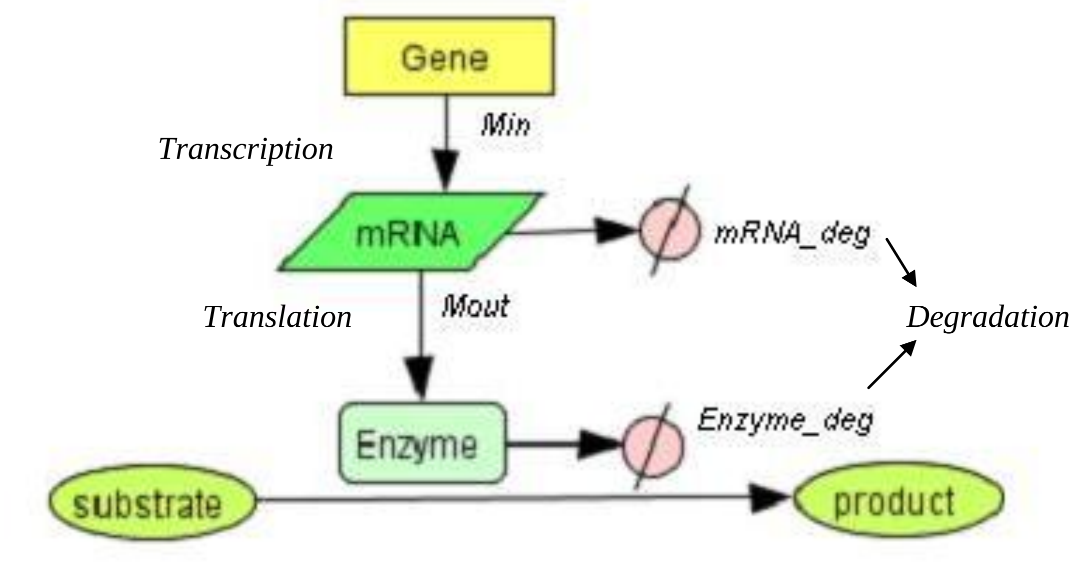


Figure 7. A gene expression dynamic system. The gene is first transcribed into mRNA which is then translated into proteins. The change in mRNA and protein concentrations is modeled as a function of time. Both mRNA and proteins degrade over time by their respective degradation rates. The enzyme (protein) encoded by the gene then catalyses the reaction to convert the substrate into a product.

In the past, degradation of mRNA and proteins have been assumed to occur randomly over time (Chen *et al.*, 1999; Schlitt and Brazma, 2007). Chan *et al.*, (1999) modelled transcription and translation, where the variables are functions of time, as:

a) Transcription

$$\text{———} \qquad (2)$$

Where [*mRNA*] is the concentration of mRNA**,** *f(mRNA)* is the transcription function, *V* is the degradation rate of mRNA. *mRNA degradation* was expressed as being directly proportional to the amount of mRNA concentration.

b) Translation

$$\text{—} \qquad \text{-} \qquad (3)$$

where *L* is the translation constant, *p* is defined as the concentration of protein and *U* is the protein degradation rate.

The change in concentration of proteins (d*p/*d*t)* equals a translation constant multiplied by the concentration of protein minus degradation of proteins over time. The degradation was expressed using first order mass action kinetics.

Klipp et al (2007) modelled transcription and translation, dynamically, using ODEs. The transcription function was modelled as the inflow flux minus the outflow flux. The flux is directly proportional to the concentration of the reactant(s) by laws of mass action i.e ΔmRNA/Δt = *flux in – flux out* (Klipp et al., 2007).

Klipp's method did not account for degradation as compared to Chen's method although degradation of mRNA and proteins are important reactions for regulating gene expression. Chen also assumes translation and degradation rates to be constant (Chen *et al.*, 1999). Whereas Klipp's method considers volume change of proteins as they are transported from the nucleus into the cytoplasm, Chen's method takes no notice of volume change and assumes that the change of volume has minimal effect on the dynamics of the system. However, both models show a significant role of proteins in both transcription and translation.

Schilt et al 2007 showed that gene expression can be modelled using difference and differential equations (Schlitt and Brazma, 2007). The fundamental difference equation model given as:

$$g_1\ (t + \Delta t) - g_1\ (t) = (w_{11}\, g1\ (t) + \ldots\ w_{1n}\, g_n\, (t))\ \Delta t \qquad (4)$$

...

$$g_n\ (t + \Delta t) - g_n\ (t) = (w_{n1}\, g1\ (t) + \ldots\ w_{nn}\, g_n\, (t))\ \Delta t \qquad (5)$$

where $g_1\ (t + \Delta t)$ is the level of gene i expressed at time $t + \Delta t$ and w_{ij} specifies the weight at which gene j influences gene $i\ (i,j = 1....n)$. It is a linear model i.e. gene expression level at time $t + \Delta t$ depends on the expression level at time t. This model is useful if we need to know the subset of genes expressed at a particular time and how they affect the gene network. Differential equations used by Chen *et al* (1999) and Klipp *et al* (2007) in modelling gene expression are similar to difference equation used by Schlitt and Brazma, but differential equations are continuous in nature while difference equations are discrete (Schlitt and Brazma, 2007).

Modelling using differential equation can provide an understanding of different nonlinear behaviours of gene networks and it is preferred over network approaches because of its accuracy and ability to capture the dynamical behaviours. Gene expression rates are continuous in nature, no on-off switches and no discretization of data is required in the methods proposed by Chen and D'haeseleer *et al*. (1999) Also, these models can be integrated with recent large gene expression datasets (Smolen *et al.*, 2003).

However, dynamic models generally require much more computer time than logical network models of comparable size (Smolen *et al.*, 2003). Additionally, differential and difference equations generally rely on numerical parameters, which are difficult to measure experimentally (Schlitt and Brazma, 2007).

1.4 Metabolic Reaction Network

1.4.1. Cellular Metabolism

Cellular metabolism is a network of biochemical fluxes, metabolic compounds and regulatory interactions. The emergent global behaviour of such a network of interactions is impossible to predict, evaluate and understand by intuitive reasoning alone. Therefore, mathematical modelling provides a framework for understanding the organization and dynamic nature of these networks (Steuer, 2007). A metabolic pathway consists of enzyme-catalyzed reactions that convert substrates (reactants) into product(s). Each reaction starts with one or more substrates and terminates with one or more products. A substrate can participate in any number of reactions or pathways. A metabolic network is a linked set of complex interconnected pathways (Walton *et al*., 2006). For a metabolic system, the dynamical properties are essential for the system to ensure and maintain its function and stability (Steuer, 2007). The velocity of a metabolic reaction depends on the enzyme kinetics of the reaction.

1.4.2. Fluxes

Metabolic fluxes, referred to in this thesis as fluxes, relate to the rate of flow of metabolites along a metabolic pathway from reaction to reaction. The flux of a reaction can be expressed as a function of three main components; i) the level of activity of the enzyme catalyzing that particular reaction, ii)the concentration of the metabolites, including reactants and products, that affect the activity of the enzyme and iii) the properties of the enzyme itself which can include activators and inhibition. (Nielsen, 2003). The methods currently used in measuring metabolic fluxes are mainly carbon-13 (^{13}C) tracer experiments - namely ^{13}C –GD-MS, ^{13}C NMR and ^{13}C-LC-MS (Steward *et al*., 2010). All three experimental methods require the labeling of isotopic patterns of metabolic end point but not directly measuring fluxes as the experiments' name might suggest. Computational methods are then used to calculate fluxes based on the isotopic labeling patterns. Fluxes are usually measured in units of mmol/gDW.h. (Steward *et al*., 2010).

1.4.3. Enzyme kinetics

.

Enzyme kinetics is the study of the rate of chemical reactions. A catalyst is a substance that increases the rate of a reaction without modifying the substrate or the energy change in the reaction. This means that the stoichiometric expression of a complete reaction does not include the catalyst. Enzymes are biological catalysts (Moss, 1992).

Enzymes bind temporarily to the substrate to lower the activation energy needed to convert the substrate into a product. Therefore, enzymes are proteins that act as a catalyst.

Kinetic equations are commonly expressed as functions of the *amount-of-concentrations* of the chemical species involved. This *amount-of-concentration* is the amount of substance divided by the volume; and usually abbreviated to *concentration* since it is the only kind of concentration used in biochemistry (Moss, 1992). The rate of a chemical reaction can be influenced by several factors such as temperature, pH, the amount of concentration of substrate and the presence of inhibiting compounds.

Collision theory states that molecules can react only if they come into contact with each other. Therefore, any factor that increases the rate of collision such as increased concentration of the reactant or increased temperature will increase the reaction rate. However, not all molecules that collide will react. An important reason for this is that not all colliding molecules possess sufficient energy to undergo a reaction (Palmer, 1995).

As the concentration of substrate molecules increases, the quicker the enzyme molecules will collide with the substrate molecules. The concentration of substrate molecules is designated [S] and measured in unit of molarity (M).

Increasing temperature increases molecular movements. Hence, probability of collision between substrate and enzyme molecules increases. However, since enzymes are proteins, they have an upper limit bound beyond which the enzyme becomes denatured and inactive. The effects of temperature may be incorporated into the kinetic relation.

There are two types of inhibition that can be present in an enzyme-catalysed reaction, competitive and non-competitive. Competitive inhibitors are molecules that bind to the same region as the substrate, thereby preventing the substrate to bind to that enzyme, but are not changed into product(s). It will take a higher substrate concentration to reach the same velocity as if no inhibitors existed.

Non-competitive inhibitors are molecules that bind to different site or region of the enzyme reducing the catalytic power and efficiency of that enzyme. A substrate bound to an enzyme with a non-competitive inhibitor will take longer to convert the substrate into a product (Matthews, 2000). This work does not investigate the effects of inhibition on reactions. We assume that inhibition can be neglected in this study to reduce modelling complexity.

1.4.4. Methods for Flux Measurement

Systems biology models rely on experiments carried out in the lab to collect data and parameters. As seen in 1.4.2, measuring metabolic fluxes require ^{13}C tracing experiments which like every experiment carried out in the lab is prone to errors. In general, most experiments are targeted and performed under a specific experimental condition which makes it difficult to compare results for the same organism. Current methods in analysing enrichment patterns in metabolites use nuclear gas chromatography-mass spectrometry (GC-MS) or magnetic resonance (NMR) (Nielsen, 2003). An advantage of using ^{13}C sources in measuring fluxes is that the network topology can be inferred based on the direction of fluxes in the network. However, one needs to combine experimental data about carbon transition and mathematical algorithms before being able to calculate fluxes which can increase computational complexity (Wiechert , 2001; Nielsen, 2003).

DNA microarray and proteomic analyses are methods for measuring quantites of cellular molecules. Additionally, they are able to aid in the investigation of the components' composition of cellular molecules (Ishii *et al.*, 2007). However, DNA microarray has been extensively used in determining the quantities of genes and proteins in response to perturbations. Other methods used in the quantification of molecules in the cell such as metabolic flux analysis and quantitative reverse transcription polymerase chain reaction (qRT-PCR) have been found to be targeted.

However, both methods are useful in the detection of small changes in molecules in response to perturbation (Ishii *et al.*, 2007). A study by Teusink *et al* (2000) showed that when data is experimentally determined data for a biological model, the results could still differ.

Furthermore, advances in the proteomics, such as the use of stable isotope labeling techniques, have allowed for the quantification of proteins under different experimental conditions. (Bateman *et al.*, 2007; Picotti *et al.*, 2009). Quantitative proteomic data could be integrated into our kinetic modeling approach, described in Chapter 2, in our attempt to provide a methodology for building integrative kinetic models.

1.4.5.Mass Action

Mass action kinetics describes the behaviour of all chemical compounds (reactants and products) in elementary chemical reaction (E + S → ES → P + E) as an equation where the rate of chemical reaction is directly proportional to the concentration of the reactant(s). An elementary reaction is a chemical reaction where one or more chemical species react directly to form product(s) in a single reaction step and with a single transition state.

In chemistry, there are two aspects of the law of mass action.

1) Equilibrium aspect which concerns the composition of a reaction mixture at equilibrium and 2) kinetic aspect which deals with rate equations for elementary reactions. With the kinetic aspect, there are various rules that determine the reaction rate for a particular reaction based on the number of substrates and products.

For a zero-order reaction, the reaction rate is independent of the concentration of any reactant.

If we consider a reaction [S → P] obeying first-order kinetics, the rate is given by:

$$— \qquad — \qquad\qquad (6)$$

where the formation a molecule of P at a given time is directly proportional to the concentration of S.

With a first-order reaction, the reaction rate is deemed proportional to the concentration of a single reactant. The conversion of an enzyme-substrate complex into product(s) or into another intermediate complex is an example of a first-order reaction is also known as a *unimolecular reaction*.

For a second-order reaction [S + T → P + Q], the reaction rate is given by

$$— \qquad — \qquad — \quad — \qquad\qquad (7)$$

The reaction rate would be proportional to the concentration of each of the reactants S and T. It could also be proportional to the square of the concentration of a single reactant. The binding of an enzyme molecule to a substrate molecule is a typical example of this type of reaction, commonly known as *bimolecular reaction*.

1.4.6. Single-substrate Kinetics

The mechanism of a typical enzyme-catalysed reaction involving a single substrate and a single product may be expressed as

$$E + S \underset{k_{-1}}{\overset{k_{1}}{\rightleftharpoons}} ES \xrightarrow{k_{2}} E + P$$

where E is the enzyme, S is the substrate, ES the intermediate complex, P the product. The terms k_1, k_{-1} and k_2 are rate constants for, respectively, the association of substrate and enzyme, the dissociation of substrate from the enzyme and the dissociation of product from the enzyme.

The arrows represent the direction of the reaction. Double arrows indicate that the reaction occurs in both directions.

1.4.7.Michaelis-Menten Equation

Leonor Michaelis and Maud Menten derived a simple mathematical equation of a single-substrate reaction based on the following assumptions:

- The concentration of the enzyme is very small compared to [S], so that the formation of ES does not significantly diminish [S].
- For irreversible reactions only, the concentration of [P] is effectively zero. This is the 'initial-rate' assumption, and implies not only that P is absent at the outset, but also that the amount of P formed in the time required for a rate measurement is too small to give rise to a significant reverse reaction.
- Although the product-releasing step is fast, however, it is much slower than the reaction in which S is released from ES. E and ES are considered to be at equilibrium.

The *equilibrium assumption* assumes that "the rate of formation of ES equals the rate of dissociation to E + S" expressed mathematically as:

$$\mathbf{k_1[E][S] = k_{-1}\,[ES].} \qquad (8)$$

The *transition-state theory* states that all chemical reactions proceed via the formation of an unstable intermediate between reactants and products (Palmer 1995). ES is an example of such an unstable intermediate complex. The state at which concentrations of the reactants and products exhibits no change over time is known as the *chemical equilibrium.* Usually, this state occurs when the forward reaction rate is the same as the reverse reaction rate with no net change in the concentration of either the reactants or products.

Steady state assumption assumes that "the concentration of E and ES remains constant over a period of time". It requires that the formation of ES should be equal to its breakdown rate in any direction, including product formation, which need not be slow relative to the back-dissociation to E + S. Mathematically expressed as:

$$\mathbf{k_1[E][S] = k_{-1}[ES] + k_2[ES].} \qquad (9)$$

The *initial velocity or rate of a reaction*, *v*, is the reaction rate at time, *t*, equals 0, *t* = 0. *v* depends upon the concentration of a *substrate* S that is present in large excess over the concentration of an enzyme E with the appearance of saturation behaviour following the Michaelis-Menten equation:

$$\frac{}{} \qquad (10)$$

where *v* is the observed initial rate , V_{max} is its limiting value at substrate saturation (i.e. [S] > > K_m), and K_m the substrate concentration when $v = V_{max} / 2$ i.e. the K_m of an enzyme is therefore the substrate concentration for which the reaction occurs at half of the maximum rate. In physical terms, K_m is an indicator of the affinity that an enzyme has for a given substrate, and hence the stability of the enzyme-substrate complex. Figure 8 shows a graph of the effect of substrate concentration on the initial rate of an enzyme-catalysed reaction.

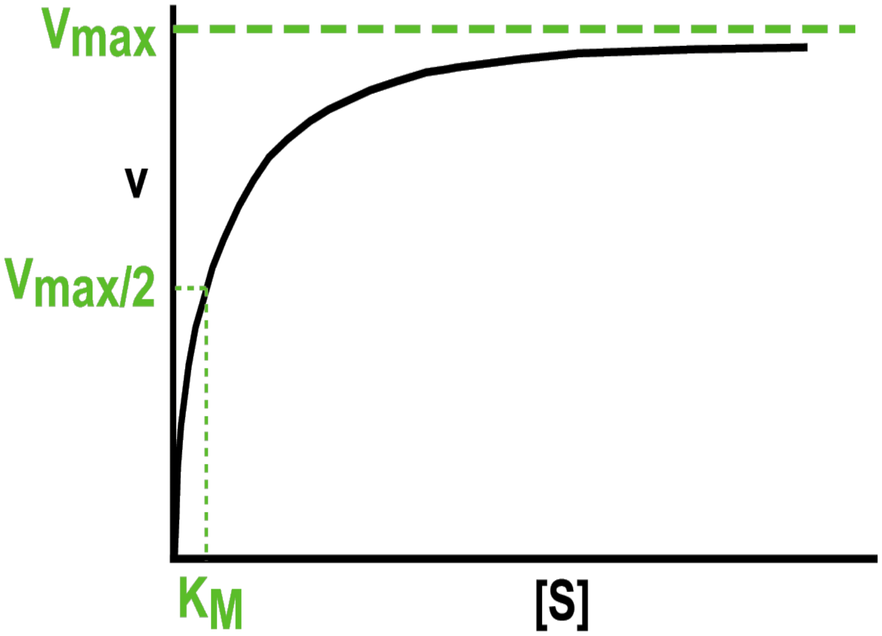


Figure 8. Effect of substrate concentration on the initial rate of an enzyme-catalysed reaction.

1.4.8. The Reaction Mechanism

For reactions involving more than one substrate or product, the kinetics can depend on a particular reaction mechanism. A reaction mechanism may either be sequential, where both substrates bind to the enzyme to form a ternary complex before a product is formed, or non-sequential.

The random-order ternary-complex mechanism is one in which any substrate can bind to the enzyme first and any product can be formed first i.e. a substrate can bind to the enzyme in random order. It is a sequential mechanism which means that a ternary complex is involved for a two-substrate reaction.

The compulsory-order (simple ordered) mechanism is also sequential. This mechanism requires the specification of the precise order of the binding to and leaving from the enzyme. It may be that no binding site is present on the enzyme for one of the two substrates until the other has bound. This makes it a compulsory order of binding. A two-substrate reaction will involve the formation of a ternary complex.

The last possible mechanism for a reaction with two substrates, the Ping-pong bi-bi mechanism, is non-sequential. It allows for a single substrate to be present on an enzyme at any one time. This leads to the assumption that there may only be a single binding site.

1.5 Methods for Modelling Metabolic Networks

The traditional modelling of metabolic processes was done, mathematically, based on explicit enzyme-kinetic rate equations. Metabolic reactions can be modelled qualitatively (e.g. network analysis, stoichiometric analysis) or quantitatively (e.g. structural kinetic models and kinetic models). Alternatively, models are said to be grouped into two classes: kinetic models and stoichiometric models (Patil *et al*., 2003).

1.5.1.Topological approaches and Network analysis

At the basic level, a metabolic network can be considered as a bipartite graph that consists of a set of nodes (metabolic substrates or metabolites) and a set of direct or undirected links (metabolic reactions) between them. For example, Figure 9 shows the metabolic network of the *tricarboxylic* (TCA) reaction cycle showing the level of connectivity that exists among metabolites.

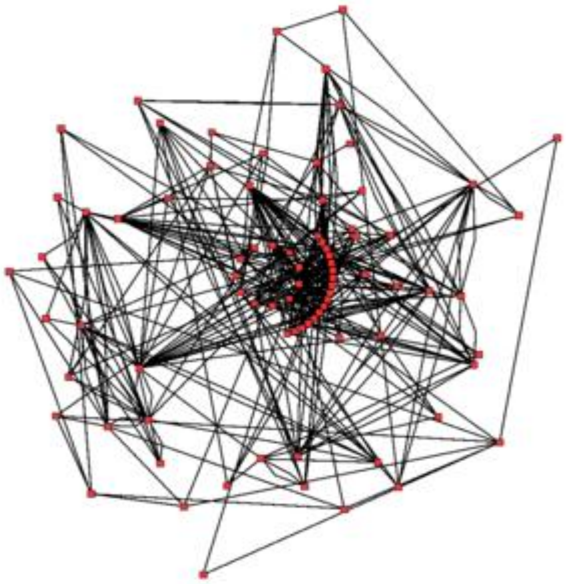


Figure 9 :Metabolic network of the *Arabidopsis thaliana* citric acid cycle. Enzymes and metabolites are shown as red squares and the interactions between them as black lines. [http://en.wikipedia.org/wiki/Metabolic_network_modelling]

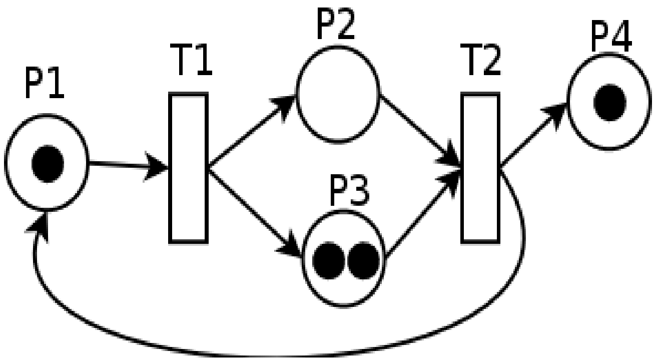
Network-based analysis can facilitate the assessment of the properties that emerge from networks such as reaction correlations, redundancy of pathways and distribution of reaction connectiveness.

Metabolic networks can be represented as hypergraphs, i.e. networks with all the edges (reactions) connecting to several nodes (metabolites) and as such , more advanced method should be used for their analysis rather than simple graph theory (Steuer, 2007).

1.5.2 Directed Graphs

Petri nets, an extension of graph models, have been successfully used to model metabolic networks (Schlitt and Brazma, 2007). Generally, they are directed graphs that consist of two kinds of nodes, place and transition nodes, and arcs that connect the place nodes to transition nodes and vice versa. The dynamic aspect is represented by tokens. Every place node can contain tokens. The number of tokens needed for a transition along an arc is determined by the 'weight' of the arc.

In metabolic networks, the place nodes represent the metabolites; transition nodes represent reactions and arcs represent metabolite concentration as shown in figure 10.



P → Place node

T → Transition node

●→ Token

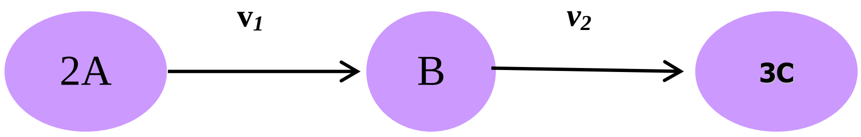
Figure 10: Example of a Petri net representing a metabolic reaction. P1, P2, P3 and P4 represent the place nodes. T1 and T2 represent transition nodes. Black dots represents token.

The main advantage of Petri nets is that there is no need to know the reaction rates equation and kinetic parameters (Schlitt and Brazma, 2007).

1.5.3 Stoichiometric Models and Analysis

Stoichiometric modelling relies on mass balances over intracellular metabolites and the assumption of pseudo-steady-state conditions to determine intracellular metabolic fluxes (Patil *et al.*, 2003). The information contained in a stoichiometric model itself results in a system of linear equations, which is generally under-determined thus not sufficient to calculate a unique flux distribution. The models are therefore combined with additional experimental data or assumptions to yield a well-defined flux map (Kesson, Forster et al. 2004).

Stoichiometric analysis makes use of the structural properties of metabolic systems. It uses the stoichiometric matrix whose element indicates the involvement of each compound consumed and produced in a reaction. The stoichiometric matrix, **S**, serves as the basis for genome-scale metabolic analysis (Jamshidi and Palsson, 2008). In Figure 11, a metabolic network is modelled by an *m* by *n* stoichiometry matrix **S**, which relates the flows ***v*** through the *n* reactions to changes ***c*** in the concentrations of the *m* metabolites by ***c*** = **S** ***v*** (Urbanczik and Wagner, 2005). **S** describes all chemical transformations in a network in an accurate matrix format and is a requisite for dynamic models (Jamshidi and Palsson 2008).



$$S = \begin{array}{cc} V_1 & V_2 & \\ \left(\begin{array}{cc} -2 & 0 \\ +1 & -1 \\ 0 & +3 \end{array}\right) & \begin{array}{c} A \\ B \\ C \end{array} \end{array}$$

Figure 11: A reaction network consisting of three metabolites (A, B & C) and two reactions with reaction rates (v_1 and v_2). The stoichiometric matrix **S** describes relationship between metabolites and reactions. A has a stoichiometric coefficient of 2 and is a substrate in reaction with rate v_1. This is represented as –2 in S (- indicates where a metabolite is substrate and + indicates product).

Knowledge of the stoichiometry puts constraints on the feasibility of flux distributions, hence providing information for the prediction and understanding of the functional capabilities of metabolic network.

1.5.4 Metabolic Flux Analysis (MFA)

In metabolic engineering, the quantification of metabolic fluxes is considered as being essential to the understanding regulation in the cell, identifying problems with product formation and gaining further insight into the biological processes. MFA is concerned with the determination of cellular metabolic fluxes and works by quantifying carbon and nitrogen flow within a metabolic network (Boghigian *et al*., 2010). Several exchanges fluxes are usually measured to produce a determined system of equations (Patil *et al*., 2003). There are optimization-driven studies and data-driven studies in quantifying fluxes used within MFA.

With the optimization-driven approach, the stoichiometric matrix is under-determined (i.e. there are fewer equations than variables) and the flux distribution is determined by the use of optimization method. In the data-driven studies, there is reduction of the stoichiometric matrix to an over-determined form (where the number of unknown variables is less than the independent equations). Then either least squares linear regression (where isotope data is available) or least square nonlinear regression (where data for isotope labeling is unavailable) is used for the determination of flux distribution for a particular organism **(**Boghigian *et al*., 2010)**.**

One disadvantage of MFA, under the data-driven approach, is the considerable amount of experimental measurements needed to obtain a flux distribution and as a result, it is more suited towards smaller metabolic models. However, the optimization-driven approach requires relatively few measurements (usually, biomass composition and the rate of carbon-source uptake) and can solve large models with over 1000 reactions quickly by using modern optimisation techniques (Boghigian *et al*., 2010). A very common example of MFA is flux balance analysis (see below), where an objective function can be used to determine an optimum flux distribution using linear programming. MFA has been used to successfully predict fluxes in the central metabolic network from a genome-scale model of *Arabidopsis thaliana* (Williams et al., 2011). In this example, the direction and magnitude of the changes in fluxes were successfully predicted using MFA.

1.5.5 Flux Balance Analysis (FBA)

Flux Balance Analysis (FBA) is a constraint-based modelling approach that uses physiochemical constraints such as mass balance, energy balance and flux limitation to describe the behaviour of an organism. The model assumes that an organism will attain a steady-state under any given environmental condition which satisfies the physiochemical constraints. Multiple steady-states are generally possible as many constraints on cellular systems are unknown. Therefore, an optimisation process needs to be carried out to find the optimal value for a specified objective. This is done with respect to the constraints identified in order to identify a physiological meaningful steady-state (Kauffman, et al. 2003). The common objective functions include the production of biomass, maximization or reduction of ATP, and maximization of the rate of synthesis of a specific product (Shlomi, et al. 2007).

Although this approach has proven to be very useful - see examples (Almaas, et al. 2004; Serge, et al. 2004; Deutscher, et al. 2006) it requires the knowledge of many variables like the identification of all metabolites, reactions and metabolic enzymes in the pathway. It also requires the definition of an objective function which crucially determines the FBA result (e.g. the production of biomass). The objective function can either be a static (where the function is optimised for a single condition) or dynamic (where the function is optimised for numerous conditions) (Kauffman, Prakash et al. 2003). Further research by Kauffman et al (2003), found that static optimal approach was computationally simple to implement provided all the constraints were linear. Incorporating experimental data was, however, more flexible with the dynamic optimal approach but computationally costly.

Although FBA is a useful method for analysing the behaviour of metabolic fluxes, it requires a model to be optimised towards one objective function. The method in a way assumes that a biological cell is only optimised for a single task (for example, the production of biomass) at a particular time. This assumption makes this approach questionable as in practice; a cell can prefer a suboptimal state that makes it less energy consuming to adjust to multiple tasks (Schuster et al., 2007). Schuster et al (2000) proposed another concept related to constraint-based analysis known as Elementary Modes (EM). This approach decomposes a metabolic network into distinct but overlapping pathways. An EM is a minimal set of reactions capable of working together in a steady state. Every metabolic network has a unique set of EMs and all feasible flux vectors can be expressed as linear combinations of these EMs (Schuster, Fell *et al*. 2000; Poolman *et al*., 2004; Schwartz and Kanehisa, 2005; Steuer 2007).

Elementary modes (EM) and extreme pathways (ExPas) are stoichiometric pathway analysis methods based on convex analysis. However, calculating elementary modes and extreme pathways is computationally challenging, and thus difficult to use on a genome-scale (Papin, Price *et al*. 2003; Gagneur and Klamt 2004; Papin, Stelling *et al.* 2004). The cause of this computational difficulty is that the number of ExPas and EM increases considerably with the size and complexity of the network (Yeung *et al*., 2007).

Using EMs, the full set of all possible flux distributions can be determined (Terzer and Stelling, 2007). However, because of computational difficulties EM analysis is generally limited to small and medium scale models (Klamt *et al*., 2005). Consequently, there have been methods presented for computing elementary modes for large scale biological networks by means of decomposition of the flux distribution (Schwartz and Kanehisa, 2006) and network compression (Klamt *et al*., 2005). The ExPas of a network are the minimal set of linearly independent EMs for a particular metabolic network and can be useful in identifying network redundancy (Yeung *et al*., 2007). Steuer et al (2007) suggests that stoichiometric analysis be considered the most successful computational approach to metabolism, to date, based on the required knowledge and predictive power.

It is not straightforward to incorporate dynamic properties into a metabolic system based on the topological approach. The relationship between the topological structure and dynamic and functional properties of a system still remains unclear.

The incorporation of kinetics information into elementary modes and extreme pathways will present a more complete cellular function where there is a dominating influence of kinetics (Papin, Price *et al*. 2003). An attempt of such integration was made by Schwartz and Kanehisa (2006) when the authors integrated kinetic modelling with EMs to access the range of achievable states in a metabolic network. Their results showed that by combining both modelling approaches, the possible behaviours of metabolic system can be significantly constrained.

1.5.6 Structural Kinetic Modelling

This is a new modelling approach which aims to integrate dynamic modelling with stoichiometric analysis to bridge the gap between topological and dynamic metabolic network modelling. The method augments stoichiometric analysis with kinetic properties without the use of a detailed set of differential equations

This approach was proposed by Steuer *et al.,* (2006; 2007) to offer an alternative to traditional kinetic modelling. Structural kinetic modelling does not require knowledge of the enzyme-kinetic rate equation and parameters. The method describes the dynamics of the systems for small variations around metabolic steady states, and the stability of steady states. Relevant interactions and parameters governing the dynamic properties of the systems can be identified using this method. At each point in parameter space, a local linear model is constructed in a way that the local model has a clear biochemical interpretation (Steuer *et al*., 2006). The linear models, which accounts for all possible explicit kinetic models, are then grouped together. This large ensemble of models then enables the parameter space to be statistically analysed. The method is based on decomposition of the Jacobian matrix for a metabolic system and computation of eigenvalues to determine the stability of metabolic states.

The structural kinetic modelling approach assumes that knowledge of the Jacobian matrix, computed from the stoichiometric matrix, alone is sufficient to determine certain characteristics of metabolic systems (Steuer *et al*., 2006).

1.5.7 Ensemble Modelling of Metabolic Networks

The Ensemble modelling approach uses phenotypic data, such as changes in fluxes as a result of changes in enzyme expression levels, to explore the behaviour of biological systems (Tan *et al*., 2008). The method uses the ensemble of models that are capable of reaching all given steady-states in relation to concentration of metabolites and fluxes distribution. The ensemble models are seen as spanning all kinetic space allowable by thermodynamics (Tan *et al*., 2008). Once the construction of the models is completed, all possible phenotypic states of a metabolic system such as effects of enzyme over-expression can be examined. Furthermore, this approach allows for the integration of flux variability data, data pertaining to the changes in fluxes due to enzyme expression, enzyme regulation and thermodynamic data, to reduce the size of the ensemble. This method does not require detailed knowledge of the kinetics of a metabolic network but can be useful in capturing the phenotypic changes of metabolic networks (Tan *et al*., 2008).

Ensemble modelling represents enzymatic reactions by a collection of elementary reactions as shown in the schema below:

$$\mathrm{A + E} \quad \underset{v_{-1}\quad k_{-1}}{\overset{k_1\quad v_1}{\rightleftarrows}} \quad \mathrm{EA} \quad \underset{v_{-2}\quad k_{-2}}{\overset{k_2\quad v_2}{\rightleftarrows}} \quad \mathrm{P + E}$$

where the reaction rate of individual elementary reaction are modelled using mass action kinetics such as:

$$v_1 = k_1[\mathrm{A}][\mathrm{E}] \qquad (11)$$

where [A] is the concentration of metabolites and [E] is the concentration of free enzyme. Ensemble modelling requires detailed knowledge of the mechanism of enzyme reactions in a model.

1.5.8 Kinetic Modelling

Previous studies have suggested that to improve the effectiveness of target-based drug discovery, efforts must be focused on understanding organism at the system-wide level (Cascante *et al*., 2002; Davidov *et al*., 2003; Klipp et al., 2005; Cho et al., 2006; Hornberg et al., 2006). This requires the detailed modelling and analysis of biological behaviour at the system-wide level.

In contrast to stoichiometric modelling, kinetic modelling aims at characterizing the mechanisms of all enzymatic reaction in relation to how changes in the concentration of metabolites affect reaction fluxes. The initial stage of kinetic modelling requires the definition of the metabolic pathway of interest and its boundaries (Smallbone *et al*., 2007). A kinetic model and its boundary condition can be defined as:

$$\mathbf{x}' = \mathbf{N}\,\mathbf{v}(\mathbf{x}, \mathbf{y}, \mathbf{p});\ \mathbf{x}(0) = \mathbf{x}_0 \qquad (12)$$

where **N** is the stoichiometric matrix (based on the topology of the network), **x** is the metabolites concentration vector, **y** is the boundary metabolites vectors. The concentrations of **y** do not change over time but whose values do affect reaction rates. The initial concentration of both **x** and **y** must be defined; however, only the concentration levels of **x** varies over time. Reaction rates are denoted **v** whose value is dependent on the kinetic parameters, **p**, reaction mechanism and concentrations of metabolites. When the boundary condition is set as true for a particular metabolite (represented as **y** in equation 12), during integration and parameter estimation, these values are treated as independent variables in the system. The use of boundary conditions is similar to mechanical systems where the environmental conditions are usually treated as external parameters which have some control of the physical systems (Nishikawa, 2002). Likewise, in systems biology, boundary conditions exert a degree of influence over the state of the system and can be considered as defining the system from the outside (Nishikawa, 2002).

With kinetic modelling, it is possible to describe the detailed dynamic changes of metabolites and enzymes in a mathematical model by integrating the kinetics of reactions with the stoichiometry of a metabolic pathway (Gombert and Nielsen, 2000). Consequently, kinetic modelling can capture the dynamics of biological systems as the details of metabolite concentrations and enzyme levels are all accounted for in the model. This was exemplified in Teusink et al. (2000) where a detailed kinetic model was constructed for the glycolysis pathway in *Sacchromyces cerevisiae*.

Over the last few years, the amount of kinetic models for biological systems has increased due to the availability of data required to define these models. The challenging aspect of kinetic modelling still lies with the difficulty or unavailability of kinetic parameters and detailed knowledge of kinetic properties of enzymes in a metabolic network (Tan *et al*., 2008). This has lead to the development of optimisation techniques and the use of statistical approaches in estimating kinetic parameters to define kinetic models.

Furthermore, the lack of detailed knowledge of enzyme mechanism has led to the development of generic rate equations which are capable of predicting biological behaviour (Liebermeister and E. Klipp, 2006; Ao *et al*., 2008). For example, linlog kinetics aims at bridging the gap between kinetic and stoichiometric modelling (Smallbone *et al*., 2007; Smallbone et al., 2010). The method relies on the use of FBA in estimating fluxes through the system. These fluxes are then varied dynamically according to linlog kinetics. The results of linlog kinetics, according to the authors were not perfect in its predictions, but it still offers an approximation of a metabolic network when a detailed kinetic model and parameters are not available.

This research study uses kinetic modelling in an attempt to predict the dynamical behaviour of biological networks. Where kinetic parameters are not available for a particular model, there are numerous parameter estimation tools that can be used. Parameter estimation has become an important and integral part of kinetic modelling due to the lack of or unavailability of experimental parameter values needed to fully define a kinetic model.

1.6 Sources of Kinetic Parameters and Enzyme Information

1.6.1 Databases for Biological Models and Information

The Kyoto Encyclopedia of Genes and Genomes (KEGG) database holds information on proteins and small molecules involved in all metabolic reactions. It is the most comprehensive database for metabolic reactions (Gille, Hoffmann et al. 2007). KEGG holds information about metabolic pathways and networks, and a collection of gene catalogues for all complete genomes.

BIOCYC, REACTOME and UM-BBD are databases of biochemical reactions, but for substrate specificity of enzymes BRENDA (BRaunschweig ENzyme DAta base) is more valuable. BRENDA held information of at least 21,000 different enzymes from more than 4330 different organisms in 2001 (Schomburg, Hofmann et al. 2001). In 2006, the database held information on 83,000 different enzymes (Liebermeister, Klipp et al. 2006). Currently it holds information on structure, occurrence, function and application of enzymes of more than 100,000 enzymes (Scheer *et al.*, 2011). BRENDA provides information about kinetic parameters, enzyme-catalysed reactions and pathways, substrates and products, reaction type and cofactors i.e. reaction and specificity and functional parameters. Importantly, the majority of data held in BRENDA are manually extracted from scientific literature (Scheer *et al*., 2011).

The BioModels database allows biologists to store, search and retrieve published mathematical models of biological interest. The models are annotated and linked to relevant data resources such as databases, publication, literature etc; some of these are curated models (Novère, Bornstein et al. 2007). BioModels now incorporates an online simulation tool which embeds the SOSlib dynamic solver to perform basic simulations (Li et al., 2010). The number of curated models as of September 2011 was 366 and 398 non-curated models (BioModels Database, 2011).

The different models, based on their annotations, are categorised in Figure 12, showing that cellular metabolism and signal transduction represent about 65% of models stored in the database.

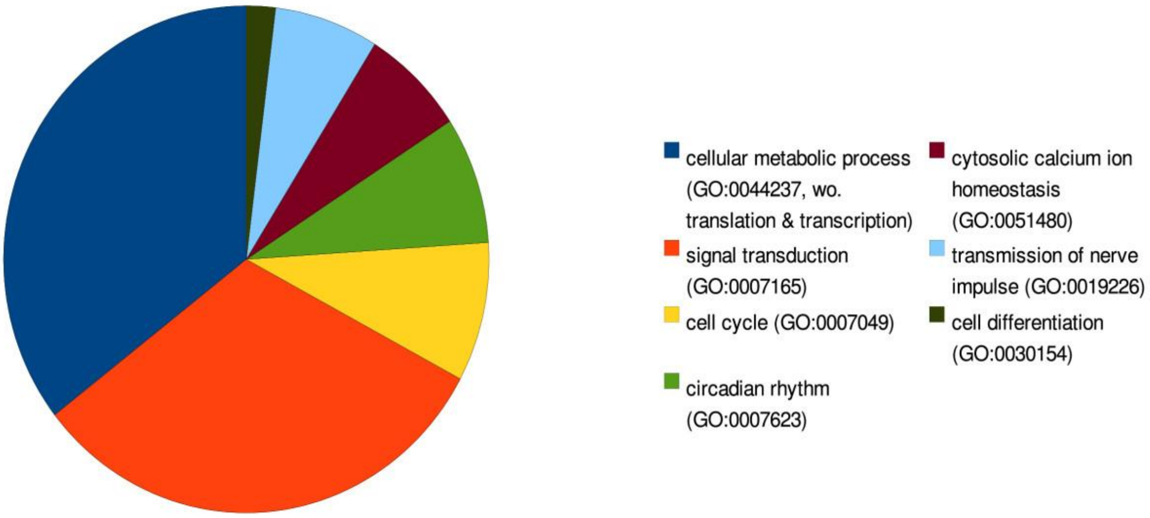


Figure 12: Categories of models in the BioModels database by using Gene Ontology (GO) terms (Li et al , 2010).

JWS Online Cellular Systems Modelling is an online tool for simulation of kinetic models from a curated model database. The database was created to provide a central repository for biologists to store kinetic models. The JWS database stores 85 curated models, which are mirrored in the BioModels database. Models can be downloaded in SBML, COPASI and PuSCes formats (Snoep and Olivier, 2003). Following the success of the BioModels database, JWS has been integrated into BioModels (Li et al, 2010).

The SABIO-Reaction Kinetic database is a database with biochemical reactions and their kinetic properties. There are two main data sources for SABIO-RK; 1) data is manually extracted from literature (publications) and verified by curators 2) data is extracted from the KEGG database before being manually curated. Information about the type of kinetic mechanism and corresponding rate equation of a reaction is also included. Information about reactions and their kinetic data can be exported to an SBML file (Wittig, Golebiewski et al. 2006). The SABIO-RK database focuses primarily on the description of individual reactions (Rojas et al., 2007). Most of the data and models from the last three databases can be used for verification and validation purposes while some of these databases are organism-specific as shown in Table 1.

Organism	Database	Website address
E.Coli-Specific Database	*E. coli* Gene Expression Database EcoCyc EchoBase	http://chase.ou.edu/oubcf/ http://ecocyc.org http://www.ecoli-york.org
S.Cerevisiae-Specific Database	Saccharomyces Genome Database	http://www.yeastgenome.org/
Enzymes (Proteins)	BRENDA KEGG	http://www.brenda.uni-koeln.de http://www.kegg.com
Pathway Databases	KEGG MetaCyc EMP BioSilico	http://kegg.com http://metacyc.org http://www.empproject.com http://biosilico.org
Kinetic Parameters	SABIO-RK	http://sabio.villa-bosch.de
Quantitative Models	BioModels Database	http://www.ebi.ac.uk/biomodels-main

Table 1: Available metabolic pathways and model databases

1.7 Current Modelling and Parameter Estimation Tools

1.7.1 Modelling Tools

Over the last few years, the number of modelling tool for systems biology has risen considerably. In this section, a review of a few of these modelling tools that are relevant to this research is presented. Our attention is focused on modelling tools that are platform-independent and include simulation and parameter estimation functionalities.

CADLIVE is a tool that analyzes and designs large-scale biochemical networks at the molecular interaction level. It has a GUI network constructor, database, a pathway search module, a network layout module, and a dynamic simulator that automatically converts biochemical network maps into mathematical models (Hiroyuki 2006). A drawback of CADLIVE is that it is hard to configure and use.

CellDesigner is a diagram editor tool for drawing biochemical and gene regulatory networks. The networks are drawn based on process diagram using the System Biology Graphical Notation (SBGN) (Hucka, Finney et al.). It has a functional and user-friendly graphical user interface and is easy to use (it comes with online tutorials, good documentation and examples). It supports the import and export of SBML files. Additionally, CellDesigner is now linked with databases such as the MIRIAM database which allows for the referencing of models. Kinetic equation laws can be manually written for each reaction which will make it tedious when modelling and simulating large or genome-wide systems. Alternatively, SBMLSqueezer (Dräger, Hassis et al. 2008), a plugin for generating rate equations, or CellDesigner's own selection of pre-defined equations can be used in describing reactions in a model.

COPASI is an application for simulation and analysis of biochemical networks. It handles large systems better than *CellDesigner* as it can generate the reaction equations automatically from a selection of pre-defined types. It is hard to model integrative systems (e.g. gene expression network with metabolic reaction network) as this will require specifying how the systems are linked and the reaction or kinetic functions associated with their linkage (Mendes, Hoops et al. 2006). This is the same with *CellDesigner* and other tools.

The *E-CELL* software allows users to define functions of proteins, protein–protein interactions, protein–DNA interactions, regulation of gene expression and other features of cellular metabolism, as a set of reaction rules (Tomita, Hashimoto et al. 1999). E-CELL simulates cell behaviour by integrating the differential equations describing the reaction rules. The user can observe dynamic changes in concentrations of proteins, protein complexes and other chemical compounds in the cell through graphic interfaces. With E-Cell, the user has to specify the reaction rules or extract them from the literature (Tomita, Hashimoto et al. 1999; Takahashi, Ishikawa et al. 2002).

Virtual Cell allows users to build complex models with a web-based Java interface, allowing the user to specify all relevant parameters. Virtual Cell simplifies the task of modelling biochemical systems by translating reactions to mathematics (ordinary and/or partial differential equations) (NRCAM 2008).

There are over 100 modelling tools for building either kinetic or structural model of biological systems that support SBML. A full review can be found at (Bergmann et al., 2011). Only a few modelling tool such as ScrumPy (Poolman, 2006) are designed for the construction and analysis of both kinetic and structural modelling with emphasis on large-scale models. Usually, a modelling tool is designed either for kinetic or structuring modelling.

1.7.2 Parameter Estimation Tools

Due to the lack of kinetic parameters required for kinetic biological models, many parameter estimation tools have been developed to estimate kinetic parameters that fit a given experimental dataset.

COPASI, described above, has extensive support for parameter estimation and optimization. COPASI provides thirteen different parameter estimation methods including a genetic algorithm. However, the estimation of kinetic parameters is not straightforward and fluxes are excluded in its estimation procedure.

JigCell is a computational tool for developing and analysing complex biochemical regulatory systems (Vass et al., 2004). Although the modelling tool has been discontinued, the parameter estimation process has been developed into a stand-alone application. The parameter estimation requires the manual initialisation of conditions and parameters which makes it impractical for a large model with thousands of parameters.

The Systems Biology Markup Language-based Parameter Estimation Tool (SBML-PET) is a tool designed for biologists to estimate parameters in systems biology models. SBML-PET supports the importation and exportation of SBML models. Another important feature of SBML-PET is that a variety of experimental data types can be used in estimating kinetic parameters. It also supports the definition of events in SBML models. SBML-PET uses an ordinary differential equation (ODE) solver know as ODEPACK to solve ODEs in a system and utilises a stochastic ranking evolution strategy (SRES) for parameter estimation (Zi and Klipp, 2006).

qPIPSA is a software for estimating missing kinetic parameters. A drawback of this software is that it requires the knowledge of reaction mechanism of similar enzymes. It is a good tool for investigating the enzymatic structural-functional relationship and enzyme mechanisms (Gabdoulline and Matthias 2007), in particular to verify the mechanism of an enzyme-catalysed reaction in order to derive the rate equation of a metabolic reaction.

Other parameter estimation tools such as simBiology (MathWorks, 2011) and SBToolbox2 (Schimdt et al., 2006), are MATLAB products that require a licence for their use.

It will be time-consuming and tedious to reconstruct genome-scale kinetic metabolic networks using any of these software tools. Table 2 shows the number of genes, metabolites and reactions in a genome of *H. influenzae, E. coli, M. genitalium, S. cerevisiae* and *H. sapiens.* To manually reconstruct metabolic models for these organisms at a genome-scale level and adding rate equations for each reaction will be a time-consuming and tedious task. As a result, we need tools to automatically generate rate equations to define a large biological model.

Organism	Number of Genes	Number of Metabolites	Number of Reactions
H. influenzae	296	343	488
E. coli	660	436	627
M. genitalium	482	274	262
S. cerevisiae	708	584	1175
H. sapiens	1496	2766	3311

Table 2. Size of genome-scale metabolic network at different layers. Number of genes, metabolites and reactions in a genome-scale metabolic network of *H. influenzae (*Edwards and Palsson, 1999); *E. coli,* (Reed et al, 2002); *M. genitalium,* (Suthers et al., 2009); *S. cerevisiae* (Förster et al, 2003) and *H. sapiens* (Ma et al, 2007).

1.8 Integration of Gene Expression and Metabolic Network

1.8.1 Integrating Genomic and Metabolic Data

In April 2003 the human genome sequence was completed. A challenge currently facing biologists is the elucidation of the molecular function of the proteins encoded by the genes (Wishart 2007). With the availability of annotated genomes and detailed bibliomic data, it is now possible to reconstruct genome-scale reaction networks that include the identification of all reactions, metabolites and enzymes that participate in the network. However, to reconstruct genome-scale kinetic models, we need to integrate metabolomic, fluxomic data along and thermodynamic information with genomic data. Figure 13 shows this process of integration of various data types (Jamshidi and Palsson 2008).

Gene expression and metabolic reactions are two different functions of a cell. Many works have been directed at either system, but the joint modelling of both systems has not been well researched (Yeang and Vingron 2006). Integrative metabolic and gene expression network can increase our understanding of cell functions. This could help identify genes whose expression levels quantitatively determine a metabolic function, that play a key part in regulating a cellular function and understand their role in the metabolic network (Li 2004). Our ability to identify these genes will increase our understanding of pathways that are active as a function of the environment and in turn help to uncover the interplay between gene and metabolic networks.

Previous studies have suggested that by accomplishing the integration of gene expression and metabolism, it will be possible to perturb cellular functions *in silico* and predict cellular or physiological function across a range of conditions, thus, enable the engineering of biological behaviour (Li 2004).

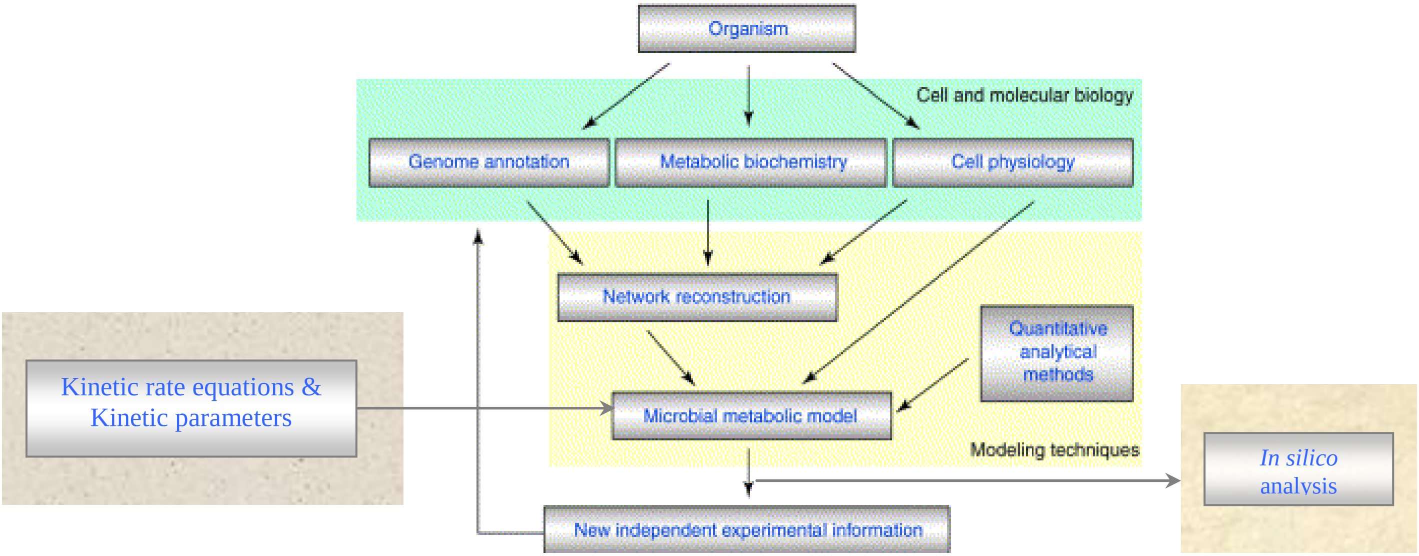

Organism
Cell and molecular biology
Genome annotation
Metabolic biochemistry
Cell physiology
Network reconstruction
Quantitative analytical methods
Kinetic rate equations & Kinetic parameters
Microbial metabolic model
Modeling techniques
In silico analysis
New independent experimental information

Figure 13. Integrated process of microbial metabolic model construction as proposed by Jamshidi and Palsson (2008). Such construction requires a comprehensive knowledge of the metabolism of an organism. From the annotated genome sequence and the experimentally determined biochemical and physiological characteristics of a cell, the metabolic reaction network can be reconstructed. This network is then modified in the context of other physiological constraints to produce a mathematical model, which can be used to generate quantitatively testable hypotheses *in silico*. As the model is used to direct an experimental plan, it can be important in further re-examining the biological properties of the organism.

The number of constraint-based models (CBM) utilising flux balance analysis (FBA) or structural modelling approaches to analyse biological systems has considerably increased. In contrast, the number of large-scale kinetic models still remains relatively low. The reason for the slow growth in detailed kinetic model on a large-scale is due to the unavailability of kinetic parameters for these models. Likewise, CBMs have increased as they only rely on the stoichiometry of the network with no knowledge of kinetic parameters required in the model building process. The success prediction rate of CBM models for *in silico* experiments has been observed to be as high as 86% for *E. coli* gene deletion experiments (Price *et al*., 2003). The current availability and development of high-throughput experiments including genome sequencing and DNA microarray (Pease *et al*., 1994; Schena *et al*., 1995) analysis have made it possible to measure the quantitative levels of gene expression on a genome-scale (Sherlock *et al.*, 2001). In addition to modern high-throughput experiments, the availability of modern 'omics' data about individual organisms signals that it may be time to start building whole cell-wide models. There is now a variety of data for many organisms on different biological layers.

If we are to fully understand the response of an organism to environmental changes, it is essential to include detailed quantitative levels of genes, mRNA transcripts, proteins and metabolites and their subsequent interactions (Zhang *et al*., 2010). It has been previously demonstrated in Ter Kulie and Westerhoff (2001) that the control of glycosis was shared between genomic, proteomic and metabolic levels. This example highlights the significance of building integrative models. Many integrative models of various organisms have been constructed but most are built using constraint-based modelling approach (Fellenberg, 2003; Cavalieri and Filippo, 2005; Çakir *et al*., 2006; Joyce and Palsson, 2006; Herrgård *et al.* 2006; Yizhak *et* al., 2010; Zhang *et al*., 2010).

Previous studies have shown that kinetic modelling is more accurate in capturing the detailed dynamics of biological systems as biological systems are not discrete in nature (Puchalka and Kierzek, 2004; Resat *et al*., 2009). However, only a few detailed integrative kinetic models have been constructed to date. The building of detailed integrative kinetic models is obstructed due to the incompleteness of heterogeneous data needed to fully define such models and the unavailability of kinetic parameters. As a result, parameter estimation has become a very important and central part of computational systems biology (Mendes and Kell, 1998**;** Moles *et al.*, 2003; Goel et al., 2008; Ashyraliyev *et al*., 2009) and efforts are now being focused on predicting the dynamical behaviour of biological systems rather than producing accurate models which require the detailed knowledge of reactions mechanisms, concentration of metabolites and protein-protein interactions (Ao *et al*., 2008; Liebermeister and Klipp, 2006; Adiamah *et al*., 2010; Liebermeister *et al.*, 2010).

Building genome-scale kinetic models has proven difficult as information needed to build detailed genome-scale metabolic for individual organisms tends to be stored in various databases and mostly under different experimental conditions (Radrich *et al*., 2009). This leads to a situation of unemployed enzymes in some models (Fell *et al*, 2010) as genes and enzymes are left unmapped. Additionally, apart from inconsistencies in stoichiometric models (Gevorgyan et al, 2008), there is the issue of missing reactions or inactive reactions in published genome-scale models (Fell *et al*., 2010).

With the current problems in building precise and accurate genome-scale models, our attention on this research was mainly focused with producing a framework for streamlining the large-scale integration of gene expression and metabolic models which are capable of predicting the dynamical behaviour of biological systems.

1.9 References

1. Akutsu, T., S. Miyano, et al. (1999). Identification of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model. *Pacific Symposium on Biocomputing*.

2. Almaas, E., B. Kovacs, et al. (2004). "Global organization of metabolic fluxes in the bacterium *Esherichia coli*." *Nature*. 427: 839-843.

3. Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp J.A., Blom J.G. (2009). Systems biology: parameter estimation for biochemical models. *FEBS J*. 276 (4):886-902.

4. Ao P, et al (2008). Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AM1 growth as validation. *Chin. J. Biotechnol*. 24:980-994

5. Bakker, B. M., P. A. M. Michels, et al. (1997). *Glycolysis* in Bloodstream Form *Trypanosoma brucei.* Can Be Understood in Terms of the Kinetics of the Glycolytic Enzymes. *The Journal of Biological Chemistry*. 272: 3207-3215.

6. Bateman, R.J, Munsell, L.Y., Chen, X., Holtzman, D.M.and Yarasheski, K.E. (2007). Stable isotope labeling tandem mass spectrometry (SILT) to quantify protein production and clearance rates. J AM Soc Mass Spectrom: 16 (6). 997 - 1006.

7. Becker, S. A. and B. Ø. Palsson (2005). Genome-scale reconstruction of the metabolic network in *Staphylococcus aureus* N315: an initial draft to the two-dimensional annotation. *BMC Microbiology* 5(8).

8. Bergmann, F.T, Shapiro, B.E, and Hucka, M. (2011) SBML Software Summary. [Online] Available at <http://sbml.org/SBML_Software_Guide/SBML_Software_Summary> (Accessed 21/09/11).

9. BioModels Database (2011) 20th BioModels Database Release [Online] Available < http://www.ebi.ac.uk/biomodels-main/static-pages.do?page=release_20110901> [Accessed 21/09/2011].

10. Boghigian, B.A., Seth, G., Kiss, R. and Pfeifer, B.A. (2010) Metabolic flux analysis and pharmaceutical production. *Metabolic Engineering*. 12. 81-95.

11. Çakir, T., Kiran Raosaheb Patil, P. K., Önsan, I.Z., Ülgen, Ö. K., Kirdar, B. and Nielsen, J. (2006). Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular Systems Biology*. 2:50.

12. Cascante, M. *et al.* Metabolic control analysis in drug discovery and disease. *Nat. Biotechnol.*, 20 (2002), pp. 243–249.

13. Cavalieri, D. and Filippo, C.D. (2005) Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug Discovery Today*. 10:727-734.

14. Chassagnole, C.,Noisommit-Rizzi,N., Schmid, J.W., Mauch, K., Reuss, M., (2002). Dynamic modeling of the central carbon metabolism of *Escherichia coli*. *Biotechnol. Bioeng*.79,53–73.

15. Chen, T., H. L. He, et al. (1999). Modeling gene expression with differential equations. *Pacific Symposium of Biocomputing*: 29-40.

16. Cho, C.R. *et al.* The application of systems biology to drug discovery. *Curr. Opin. Chem. Biol.*, 10 (2006), pp. 294–302

17. Cornish-Bowden, A. (2004). *Fundamentals of Enzyme Kinetics*, Portland Press.

18. D'haeseleer, P., *et al*. (1999) Linear Modeling of mRNA Expression Levels During CNS Development and Injury. *Pacific Symposium on Biocomputing* 4: 41-52.

19. Davidov, E. *et al.* Advancing drug discovery through systems biology. (2003). *Drug Discov. Today*, 8 , pp. 175–183.

20. Dean, J.T., Rizk, M.L., Tan, Y., Dipple, K.M., Liao, J.C., (2009) .Ensemble modeling of hepatic fatty acid metabolism with a synthetic glyoxylate shunt. *Biophys J*. 98, 1385–1395.

21. Deutscher, D., I. Meilijson, et al. (2006). "Multiple knockout analysis of genetic robustness in the yeast metabolic network." *Nature Genetics Volume* 38: 993-998.

22. Dräger, A., N. Hassis, et al. (2008). "SBMLsqueezer: A CellDesigner plug-in to generate kinetic rate equations for biochemical networks." *BMC Systems Biology* 2(39).

23. Edwards, J, S and Palsson, B.O. (1999) Systems properties of the *Haemophilus influenza* Rd Metabolic Genotype. *The Journal of Biological Chemistry*. 274. 17410-17416.

24. Feist, A. M. and B. O. Palsson (2008). The growing scope of applications of genome-scale metabolic reconstructions using *Eschericihia coli*. *Nature Biotechnology*. 26(6).

25. Fell, D, Poolman, M and Gevorgyan, A (2010) Building and analysing genome-scale metabolic models. *Biochem Soc Trans*, 38 (Part 5). 1197 - 1201.

26. Fellenberg, M. (2003) Developing integrative bioinformatics systems. *BIOSILICO. 1:177-183*.

27. Förster J, Famili I, Fu P, Palsson B.Ø and Nielsen J. (2003) Genome-scale reconstruction of the *Sacchromyces cerevisiae* metabolic network. *Genome Research*. 13. 244 – 253.

28. Gabdoulline, R. R. and S. Matthias (2007). "qPIPSA: Relating enzymatic kinetic parameters and interaction fields." *BMC Bioinformatics* 8(373).

29. Gagneur, J. and S. Klamt (2004). "Computation of elementary modes: a unifying framework and the new binary approach." *BMC Bioinformatics* 5(175).

30. Garcia-Martinez, J., F. Gonzalez-Candelas, et al. (2007). Common gene expression strategies revealed by genome-wide analysis in yeast. *Genome Biology*. 8: R:222.

31. Gille, C., S. Hoffmann, et al. (2007). METANNONGEN: compiling features of biochemical reactions needed for the reconstruction of metabolic networks. *BMC Systems Biology* 1(5).

32. Goel G, Chou IC, Voit EO. (2008). System estimation from metabolic time-series data. *Bioinformatics*.24:2505-2511.

33. Gombert, A. K. and Nielsen, J. (2000) Mathematical modelling of metabolism. *Current Opinion in Biotechnology*. 11. 180 – 186.

34. Haseong Kim, J. K. L., Taesung Park (2007). Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics*. 8(37).

35. Henry,C.S., Jankowski,M.D., Broadbelt,L.J., Hatzimanikatis,V. (2006). Genome-scale thermodynamic analysis of Escherichia coli metabolism. *Biophys. J*. 90. 1453–1461.

36. Herrgård M. J., Fong S. S. and Palsson B. Ø. (2006). Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* 2006, 2:e72.

37. Hiroyuki, K. (2006). "CADLIVE: Computer-Aided Design of biological systems and its application." Bioscience & Industry 64(9): 508-511.

38. Hornberg, J.J. *et al.* (2006) Cancer: a systems biology disease. *Biosystems*, 83, pp. 81–90

39. Hucka, M., A. Finney, et al. (2008) "SBML."   Retrieved 01/07/2008, from http://www.sbml.org/Main_Page.

40. Hynne, F., S. Dano, et al. (2001). Full scale model of *glycolysis* in *Saccharomyces cerevisiae*. *Biophys Chem* 94: 121-163.

41. Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., et al. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science*, 316, 593-597.

42. Jamshidi, N. and B. Ø. Palsson (2008). Formulating genome-scale kinetic models in the post-genome era. *Molecular Systems Biology* 4(171).

43. Joyce, A. R. and Palsson B. Ø. (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* 2006, 7:198-210.

44. Kauffman, K. J., P. Prakash, et al. (2003). Advances in flux balance analysis. *Current Opinion in Biotechnology* 14: 491-496.

45. Kesson, M. A., J. Forster, et al. (2004). Integration of gene expression data into genome-scale metabolic models. *Metabolic Engineering* 6(4): 285-293.

46. Klamt, S., Gagneur, J. And Kamp, A.V. (2005) Algorithmic approaches for computing elementary modes in large biochemical networks. *Systems Biology*. 152. 249-255.

47. Klipp, E. *et al.*(2005) *Systems Biology in Practice: Concepts, Implementation and Clinical Application*, Wiley/VCH

48. Klipp, E., B. Nordlander, et al. (2007). Integrative model of the response to yeast to osmotic shock. *Nature Biotechnology* 23: 975-982.

49. Lambeth, M. J. and M. J. Kushmerick (2002). A Computational Model for *Glycogenolysis* in Skeletal Muscle. *Annals of Biomedical Engineering* 30: 808-827.

50. Li, Z., Chan, C (2004). Integrating Gene Expression and Metabolic Profiles. *Journal Biological Chemistry*. 279: 27124-27137.

51. Li, C., Donizelli, M, Rodriguez, N., Dharuri, H., Endler, L., Chelliah, V., Li, L., He, E., Henry, A., Stefan, M. I., Snoep, J.L., Hucka., M, Novere, N.L. and Liabe, C. (2010). BioModels Database: An enchanced, curated and annotated resource for published quantitative kinetic models. *BMC Systems Biology*. 4. 92.

52. Liebermeister, W., E. Klipp, et al. (2006a). "Prediction of Enzyme Kinetic Parameters Based Statistical Learning." *Genome Informatics Series* 17(1): 80-87.

53. Liebermeister, W. and Klipp, E. (2006b) Bringing metabolic networks to life: convenience rate law and thermodynamic constraint. *Theoretical Biology and Medical Modelling,* 3, 41.

54. Likic, V.A., McConville, M.J., Lithgow, T. and Bacic, A. (2010) Systems Biology: The Next Frontier for Bioinformatics. *Advances in Bioinformatics*. 2010:268925

55. Link, H.,Weuster-Botz,D.,2007.Steady-stateanalysisofmetabolicpathways: comparing the double modulation method and the lin-log approach. *Metab.Eng.* 9, 433–441.

56. Ma, H., Sorokin, A. Mazein, A. Selkov, A. Selkov, E. Denim, O. and Goryanin, I. (2007) The Edinburgh Human metabolic network reconstruction and its functional analysis. *Molecular Systems Biology*. 3: 135.

57. Mathews, C.K. and Ahern, K.G, *Biochemistry*. 3rd ed. 2000.

58. MathWorks (2011) SimBiology. [Online] Available at < http://www.mathworks.co.uk/products/simbiology/index.html> [Accessed 21/09/2011]

59. Mendes P, Kell D. (1998) Non-linear optimization of biochemical pathways: application to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869-883.

60. Mendes, P., S. Hoops, et al. (2006). "COPASI – a COmplex PAthway SImulator." *Epub*. 22(24).

61. Moles C.G, Mendes P, Banga J.R. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res*, 13 (11):2467-2474.

62. Moss, G. P. (1992). *Symbolism and Terminology in Enzyme Kinetics*. 2nd Edition, Portland Press.

63. Murphy, K. and S. Mian (1999). "Modeling Gene Expression Data using Dynamic Bayesian Networks." Technical report, Computer Science Division, University of California, Berkeley. [Online] Available < http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.30.9391&rep=rep1&type=pdf> [Accessed 21/09/2011]

64. Nielsen, J. (2003) It Is All about Metabolic Fluxes. *J Bacteriol*. 185. 7031-7035.

65. Nishikawa, K. (2002) Information concept in Biology. *Bioinformatics*. 18. 649-651

66. Novère, L., B. Bornstein, et al. (2007). BioModels Database: a free, centralized database of curated, published, quantitative kinetic models of biochemical and cellular systems. *Nucleic Acids Research*. 34: 689-691.

67. NRCAM. (2008). "National Resource for Cell Analysis and Modeling." Retrieved 22-March, 2008, from http://www.nrcam.uchc.edu/about_nrcam/aboutNRCAM.html.

68. Palmer, T. (1995). *Understanding Enzymes*. Prentice Hall / Ellis Horwood.

69. Papin, J. A., N. D. Price, et al. (2003). Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences*. 28(5): 250-258.

70. Papin, J. A., J. Stelling, et al. (2004). Comparison of network-based pathway analysis method. Trends in Biotechnology, 22(8): 400-405.

71. Patil, K.P., Åkesson, M. and Nielsen, J. (2003) Use of genome-scale microbial models for metabolic engineering. *Current Opinion in Biotechnology*. 15 64-69.

72. Pease, A. C., et al. (1994) Light-Generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis" *Proc. Natl. Acad. Sci*. USA 91: 5022-5026, (1994).

73. Picotti P, Bodenmiller B, Mueller LN, Domon B, Aebersold R. (2009) Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell*. 138(4):795-806

74. Li, P., C. Z., Edward J Perkins, Ping Gong and Youping Deng (2007). "Comparison of probabilistic Boolean network and dynamic Bayesian network approaches for inferring gene regulatory networks." *BMC Bioinformatics* 8.

75. Poolman MG, Venkatesh KV, Pidcock MK, Fell DA. (2004) A method for the determination of flux in elementary modes, and its application to Lactobacillus rhamnosus. *Biotechnology and Bioengineering* 88:601-612.

76. Poolman, M. (2006) ScrumPy: metabolic modelling with python. *Systems Biology*. 153. 375-378.

77. Price, N.D., Papin, J.A., Schilling, C.H. and Palsson, B.Ø. (2003) Genome-scale microbial *in silico* models: the constraints-based approach. *Trends in Biotechnology*. 21:162-169.

78. Puchalka J and Kierzek AM. (2004) Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophysical Journal*.86:1357–1372.

79. Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A. et Schwartz, J.-M. (2009). Reconstruction of an *in silico* metabolic model of Arabidopsis thaliana through database integration. *Available from Nature Precedings*, DOI: 10101/hpre.2009.3309.1.

80. Resat, H. Petzold, L and Pettigrew, M.F. (2009) Kinetic modelling of biological systems.*Methods Mol Biol.* 541:311-355.

81. Rizzi, M., M. Baltes, et al. (1997). "*In Vivo* Analysis of Metabolic Dynamics in *Saccharomyces cerevisiae*:II. Mathematical Model." Biotechnology and Bioengineering: 592-608.

82. Rojas, I., Golebiewski, M., Kania, R., Krebs, O., Mir, S., Weidemann,S., and Wittig, U. (2007) SABIO-RK: a database for biochemical reactions and their kinetics. *BMC Systems Biology*. 1:S6

83. Sauer. U: High-throughput phenomics: experimental methods for mapping fluxomes. *Curr Opin Biotechnol* 2004, 15:58-63.

84. Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J and Schomburg, D. (2011). BRENDA, the enzyme information system in 2011. *Nucleic Acids Research*. 39. D670 – D676.

85. Schena, M., et al. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467- 470.

86. Schilling, C. H., M. W. Covert, et al. (2002). "Genome-Scale Metabolic Model of *Helicobacter pylori* 26695." *Journal of Bacteriology*, 184(16): 4582-4593.

87. Schlitt, T. and A. Brazma (2007). "Current approaches to gene regulatory network modeling." BMC Bioinformatics 8.

88. Schmidt, Henning and Jirstrand (2006) Systems Biology Toolbox for MATLAB: a computational platform for research in systems biology. *Bioinformatics*. 22. 514-515.

89. Schomburg, I., O. Hofmann, et al. (2001). Enzyme data and metabolic information: BRENDA, a resource for research in biology, biochemistry, and medicine. *Gene Function and Disease* 1: 109-118.

90. Schuster, S., D. A. Fell, et al. (2000). A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature Biotechnology*. 18: 326-332.

91. Schuetz R, Kuepfer L, Sauer U. (2007) Systematic evaluation of objective functions for predicting intracellular fluxes in Escherichia coli. *Molecular Systems Biology*. 3:119.

Schwartz, J.-M., Kanehisa, M. (2005) A quadratic programming approach for decomposing steady-state metabolic flux distributions onto elementary modes. *Bioinformatics*. 21:ii204-205.

92. Schwartz, J-M. and Kanehisa,M. (2006) Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis. *BMC Bioinformatics*. 7:186.

93. Serge, D., A. DeLuna, et al. (2004). Modular epistasis in yeast metabolism. *Nature Genetics Volume*. 37: 77-83.

94. Sherlock, G., et al. (2001) The Stanford Microarray Database. *Nucleic Acids Research.* 29: 152-155.

95. Shlomi, T., Y. Eisenberg, et al. (2007). A genome-scale computational study of the interplay between transcriptional regulation and metabolism. *Molecular Systems Biology*. 3(101).

96. Sible, C. and J. J. Tyson (2006). Mathematical modeling as a tool for investigating cell cycle control networks. *Methods.* 41(2): 238-247.

97. Smallbone, K, Simeonidis, E, Broomhead, D.S and Kell, D.B. (2007) Something from nothing - bridging the gap between constraint-based and kinetic modelling. *FEBS Journal*. 274: 5576-5586.

98. Smallbone K, Simeonidis E, Swainston N, Mendes P. (2010) Towards a genome-scale kinetic model of cellular metabolism. *BMC Systems Biology* 4:6.

99. Smolen, P., D. A. Baxter, et al. (2003). Mathematical modeling of gene networks. *Neuron* 26: 567-580.

100. Snoep, J. L. and B. G. Olivier (2003). JWS Online Cellular Systems Modelling and Microbiology. *Microbiology.* 149: 3045-3047.

101. Steuer, R., Gross, T., Selbig, J and Blasius, B. (2006) Structural kinetic modelling of metabolic networks. *PNAS.* 103. 11868-11873.

102. Steuer, R. (2007). Computational approaches to the topology, stability and dynamics of metabolic networks. *Phytochemistry*. 68: 2139-2151.

103. Suthers, P.F, Dasika, M. S, Kumar, V.S, Denisov, G, Glass, J.I and Maranas, C.D (2009) A Genome-scale metabolic reconstruction of *Mycoplasma genitalium, i*PS189. *PLoS Computational Biology*. 5. e1.0000285.

104. Takahashi, K., N. Ishikawa, et al. (2002). E-CELL2: E-CELL Simulation System for Windows. *Genome Informatics* 13: 240–241.

105. Tan, L.M., Rizk,M.L., Liao,J.C., (2008).Ensemble modeling of metabolic networks. *Biophys. J.*95,5606–5617.

106. Terzer, M. and Stelling, J. (2007) Elementary flux modes – states-of-the-art implementation and scope of application. *BMC Bioinfomatics*. 1: P2.

107. Teusink, B., J. Passarge, et al. (2000). "Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry." *European Journal of Biochemistry* 267: 5313-5329.

108. Tomita, M., K. Hashimoto, et al. (1999). E-CELL: Software environment for whole cell simulation. *Bioinformatics* 15(1): 72-84.

109. Urbanczik, R. and C. Wagner (2005). Functional stoichiometric analysis of metabolic networks. *Bioinformatics* 21(22): 4176-4180.

110. Vass, M., N Allen, C.A. Shaffer, N. Ramakrishnan, L.T. Watson, and J.J. Tyson, The JigCell Model Builder and Run Manager, *Bioinformatics.* 20. 3680-3681.

Visser, D.,Heijnen,J.J.,2003.Dynamicsimulationandmetabolicre-designofa branched pathwayusinglinlogkinetics.Metab.Eng.5,164–176.

111. Walton, S. P., Z. Li, et al. (2006). Biological network analyses: computational genomics and systems approaches. *Molecular Simulation* 32(3-4): 203-209.

112. Wiechert, W.(2001) ^{13}C metabolic flux analysis. *Metab. Eng.* 3:195-206.

113. Williams, T.C.R, Poolman, M.G.,Howden, A.J.M, Schwarzlander, M, Fell, D.A, Ratcliffe, G and Sweetlove, L.J. (2011) Agenome-scale metabolic model accurately predicts fluxes in central metabolism under stress conditions. *Plant Physiology*. 154.311 – 323.

114. Wishart, D. (2007). HMDB: The Human Metabolome Database. *Nucleic Acids Research* 2007 35: 521-526.

115. Wittig, U., M. Golebiewski, et al. (2006). "SABIO-RK: Integration and Curation of Reaction Kinetics Data." Bioinformatics 4075: 94-103.

116. Wolfe, S. L. (1972). *Biology of the Cell.* Wadsworth Pub. Co.

117. Wu, F.X, Zhang, W.J, and Kusalik, W.J. (2004) Modeling Gene Expression from Microarray Expression Data with State-Space Equations. *Pacific Symposium on Biocomputing*. 9:581-592.

118. Yeang, C.-H. and M. Vingron (2006). A joint model of regulatory and metabolic networks. *BMC Bioinformatics*. 7(332).

119. Yeung, M., Thiele, I and Palsson, B. Ø. (2007) Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics.* 8:363

120. Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. and Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. Bioinformatics. 26. i225 – i2560.

121. Zhang, W., Li, F. and Nie, L. (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. Microbiology. 156: 287-301.

122. Zi, Z. and Klipp, E. (2006) SBML-PET: Systems Biology Markup Language-based Parameter Estimation Tool. Bioinformatics. 22. 2704-2705.

Chapter 2

Building Integrative Models of Biological Networks Using GRaPe

“*It is unworthy of excellent men to lose hours like slaves in the labor of calculation which could be relegated to anyone else if machines were used*.”
Gottfried Wilhelm von Leibnitz (1646 – 1716).

This chapter appeared in “Adiamah, DA., Handl, J, Schwartz, J-M. (2010) Streamlining the construction of large-scale dynamic models using generic kinetic equations. *Bioinformatics*, **26**, 1324 – 1331”. Here, we present GRaPe, a platform independent tool for modelling integrative biological systems. The novel features and uses of GRaPe are presented in this chapter. .

2.1 Abstract

Studying biological systems, not just at an individual component level but at a system-wide level, gives us great potential to understand fundamental functions and essential biological properties. Despite considerable advances in the topological analysis of metabolic networks, inadequate knowledge of the enzyme kinetic rate laws and their associated parameter values still hampers large-scale kinetic modelling. Furthermore, the integration of gene expression and protein levels into kinetic models is not straightforward.

The focus of our research is on streamlining the construction of large-scale kinetic models. A novel software tool was developed, which enables the generation of generic rate equations for all reactions in a model. It encompasses an algorithm for estimating the concentration of proteins for a reaction to reach a particular steady state when kinetic parameters are unknown, and two robust methods for parameter estimation. It also allows for the seamless integration of gene expression or protein levels into a reaction and can generate equations for both transcription and translation. We applied this methodology to model the yeast glycolysis pathway; our results show that the behaviour of the system can be accurately described using generic kinetic equations.

Availability and implementation

The software tool, together with its source code in Java, is available from our project web site athttp://www.bioinf.manchester.ac.uk/schwartz/grape

2.2 Introduction

The current availability of annotated genomes and detailed ‘-omics’ data makes it possible to construct stoichiometric genome-scale metabolic networks that include all reactions, metabolites and proteins. Systems biology aims to examine the properties and dynamics of cellular processes as a whole rather than in isolated parts of a cell or an organism (Kitano, 2002). Integrating cellular components is essential for our understanding of how interactions between these components influence cellular functions. One aspect of this integration is the need to integrate proteins and other cellular components into metabolic networks. For example, Förster and Palsson (2003) manually reconstructed a genome-scale integrative model of gene expression and metabolism of *Saccharomyces cerevisiae* (1175 metabolic reactions, 584 metabolites and 708 open reading frames).

Stoichiometric models, which describe the topology of a metabolic network, provide limited insights into the functioning of cellular processes. To understand the detailed dynamics of cellular functions and their regulation, it is necessary to advance toward kinetic models where the behaviour of a system can be perturbed. The construction of a genome-scale kinetic model of a biological cell requires the integration of genomic, proteomic, metabolomic and fluxomic data along with thermodynamic information (Jamshidi and Palsson, 2008). Attempts for building such large-scale kinetic models are now starting to emerge. Ao *et al.* (2008) provided a systematic method for constructing large-scale kinetic metabolic models and addressed the problem of estimating kinetic parameters. Jamshidi and Palsson (2008) described a framework for building and analysing large-scale kinetic models and presented the mathematical challenges associated with the construction of such models. In a cell-scale model, the number of reactions, metabolites and proteins can reach several thousands, making it time-consuming and costly, if not impossible, to accurately measure individual concentrations of metabolites, fluxes and kinetic parameters.

There is often inadequate knowledge of enzymatic–kinetic laws and their associated parameter values, and usually parameters obtained from the literature are dependent on specific *in vitro* or *in vivo* experimental conditions. Yet, there is growing awareness that exact rate equations and parameters are often not crucial in determining the dynamic properties of large systems. This principle has been illustrated by the development of methods for 'bridging the gap' between structural and kinetic modelling. Steuer *et al.* (2006) proposed a method that aimed to give account of the dynamical capabilities of metabolic systems without requiring explicit information about the rate equations, and they showed that it was possible to acquire a detailed quantitative representation of metabolic systems without explicitly referring to a set of differential equations. Smallbone *et al.* (2007) presented a method for building kinetic models solely based on reaction stoichiometries of a network using linlog kinetics. Their results showed good agreement between the real dynamics and their approximation in a yeast glycolysis model. Ao *et al.* (2008) also acknowledged that the scarcity of experimental data for rate equations and parameters is a major difficulty in the construction of large kinetic models, and to alleviate this difficulty, they used a generic form of rate equations with a minimum set of parameters to construct a metabolic model of *Methylobacterium extorquens*AM1. Their results showed that attaining the correct dynamical behaviour of a system is possible without the use of extensive and accurately measured rate equations and kinetic parameters. Furthermore, through an analysis of several systems biology models, Gutenkunst *et al.* (2007) suggested that parameter fitting to experimental data still leaves parameters poorly constrained and that biological systems are often robust to large parameter variations. The success of a model is therefore more dependent on an accurate prediction of the main behaviours of a system rather than on a thorough determination of large numbers of parameter values.

In order to streamline the construction of large-scale dynamic models, the difficulties related to the manual assembly of large networks and the generation of customized rate equations and parameters need to be addressed. For this reason, we developed a software tool named GRaPe (Gene-Reaction-Protein integration). GRaPe uses generic reversible Michaelis–Menten rate equations based on the number of substrates and products for all reactions in the network. The Michaelis–Menten relation offers a reliable approximation of the kinetics obeyed by most enzyme-catalytic reactions. Furthermore, most reactions of importance in biochemistry are reversible in the practical sense (Cornish-Bowden, 2004). We make two distinctive assumptions, namely that compartmentalization of the cell and metabolite–enzyme interactions play a negligible role in determining the behaviour of a system. GRaPe then creates a kinetic model of the metabolic system using ordinary differential equations (ODEs) that are automatically generated based on the stoichiometric matrix of the network.

While many tools exist for the modelling and simulation of complex biological dynamic systems, e.g. CellDesigner (Funahashi *et al.*, 2003), COPASI (Mendes *et al.*, 2006), Biological-Networks (Baitaluk *et al.*, 2006), E-Cell (Tomita *et al.*,1999), CADLIVE (Kurata *et al.*, 2003) and Cellware (Dhar *et al.*, 2004), none of these tools allows for the generation of rate equations from the stoichiometry and for the seamless integration of gene or protein levels into a metabolic network without time-consuming and error-prone manual intervention. Our aim is not to duplicate these tools by creating another simulation software, but to introduce an upstream solution for the rapid generation of large-scale dynamic models, which can be exported for simulation by existing software applications. Being consistent with the standards of systems biology, GRaPe supports the exchange of Systems Biology Markup Language (SBML; Hucka *et al.*, 2003) level 2 version 1 and 2 documents.

We provide an overview of the main features of GRaPe and present a proof-of-principle of the applicability of our approach to the construction of large-scale kinetic models. In particular, we compare the features of a model of the yeast glycolysis pathway based on generic equations automatically generated by GRaPe with a model of the same pathway constructed by Teusink *et al.* (2000) that was based on an experimental determination of rate equations and parameters. Our results show an excellent agreement between both models, supporting the hypothesis that kinetic models using generic equations could successfully reproduce the global behaviour of large metabolic systems without requiring detailed knowledge of the *in vivo* kinetics of each individual reaction.

2.3 Features of GRaPe

2.3.1 General Features

GRaPe provides a user-friendly graphical user interface (GUI) for importing, creating, editing and exporting biological models in SBML. GRaPe automatically integrates every metabolic reaction with either an enzyme species or with a gene expression process (Figure 14). When only proteomic data is defined in the SBML document, GRaPe adopts the Reaction-Protein (RP) representation. When transcriptomic data is given, GRaPe then adopts the Gene Reaction-Protein (GRP) representation. Transcription, translation and degradation of both mRNA and proteins are then expressed mathematically.

GRaPe also provides functionality to manually construct GRP and/or RP network models. In both model-building processes, GRaPe automatically generates a Michaelis–Menten reversible rate equation for each reaction based on its stoichiometry and enzyme mechanism by iterating through the metabolic network. Each reaction in the network can have up to two substrates and products. See the methods section in this chapter for the detailed list of rate equations used by GRaPe.

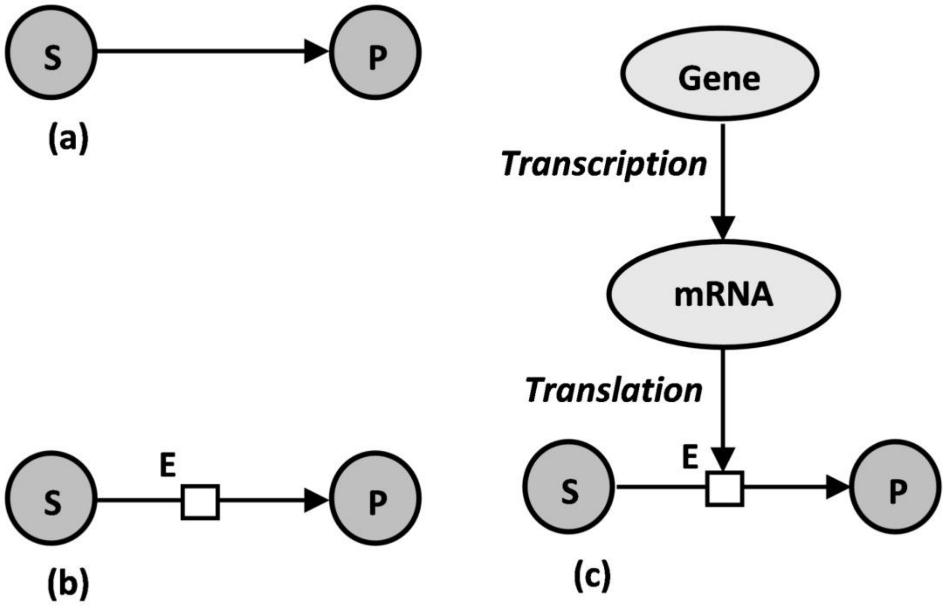


Figure 14: (**a**) A traditional representation of a metabolic reaction where a substrate, S, gets converted into a product, P. The protein concentration is expressed as being fixed, usually in the $V_{\max}$, a constant in the rate equation. (**b**) An RP representation, where reaction (a) is integrated with only its protein concentration, E. In the rate equation, E is expressed as an independent variable, which can be varied. (**c**) A GRP representation, where an RP reaction is fully integrated with its gene expression module. E is now expressed as a function of transcription, translation and degradation of both mRNA and E.

GRaPe implements two robust methods, the Levenberg–Marquardt method (LMA) and a genetic algorithm (GA), for estimating kinetic parameters, in addition to the Steady-State Enzyme Estimator (SSEE) method. See the ‘Methods’ section of this chapter for more details about the estimation methods and their application. The two parameter estimation methods attempt to find the values of missing kinetic parameters, given the experimental time series data. Both methods work interchangeably so that when one method fails to find a suitable solution set (a low objective function) the other is employed. The parameter set from the method that returns the best objective function is then taken.

The time series data used as an input for all estimation methods must be in a tab-delimited plain text file. GRaPe matches the identifiers (ids) in the model to the ids in the data file during parameter estimation and throws an error if any of the ids in the model are not found in the input data. The data file for all parameter estimation methods is the same; which means GRaPe treats the last row of the data file as the steady-state data for the SSEE method. The time series data can correspond to experimental or simulated, continuous dynamic or steady-state data.

The estimation procedure constrains parameters to experimental data and therefore assignment of initial concentrations of species must be the same in the model as in the experimental data. It is also recommended that the precision of experimental values be limited to two decimal places for faster computing of kinetic parameter values. The time taken to estimate parameters for all reactions in a model is dependent on the amount of input data. For dynamic time series data, the total estimation time tends to take longer compared to steady-state time series data. Although there is no limit to the number of data points in the input data, large datasets increase estimation time.

2.3.2 System Architecture

In order to make GRaPe platform independent and easy to use by the biological community, Java was chosen as the programming language. Figure 15 shows the architecture of GRaPe and the interactions between the main components of the system. GRaPe uses the JigCell Parser (Vass *et al.*, 2004) for importing and exporting SBML level 2 version 2 documents. Each reaction in the model consists of its substrates, products and either the enzyme alone or the full gene expression process of that reaction. The reactions are stored in a list that later gets converted by the JigCell parser into a SBML file. The inverse of this process can be achieved, i.e. by parsing a SBML file that is then decomposed into reactions, species/metabolites and enzymes.

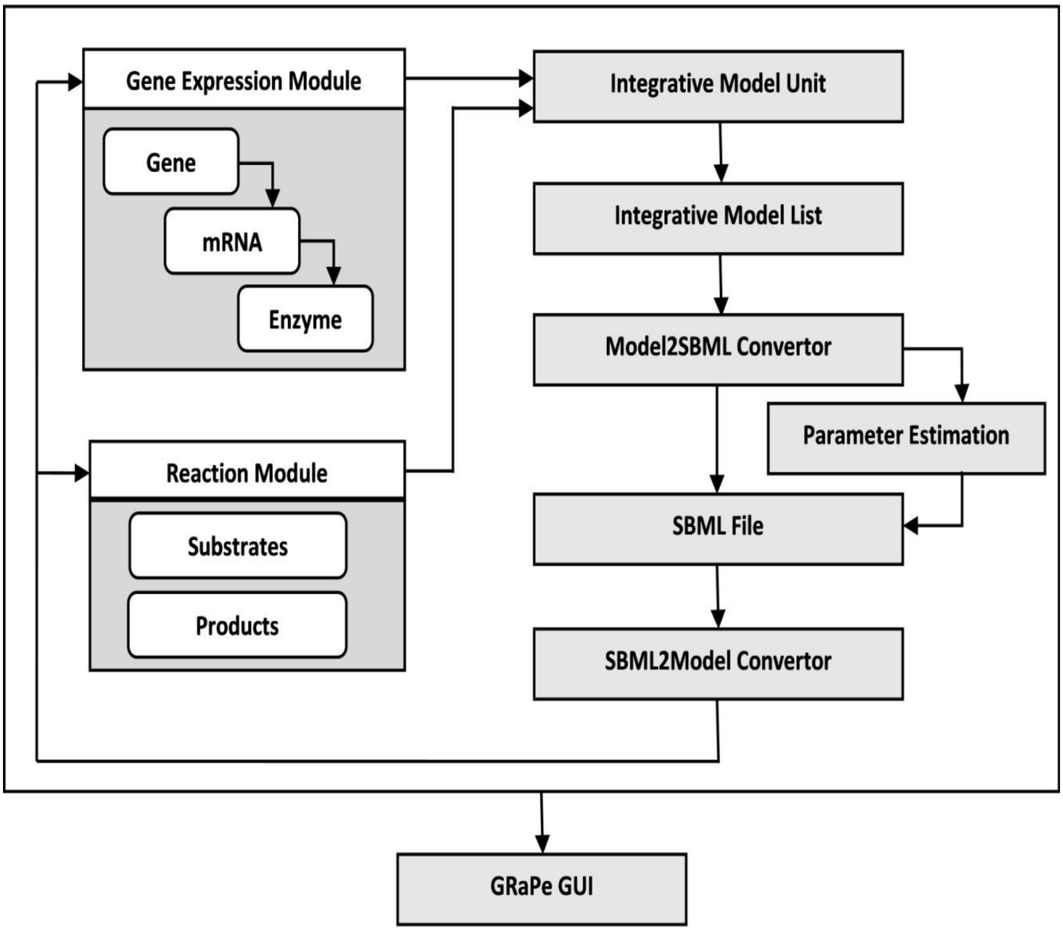


Figure 15: System architecture of GRaPe. The 'Gene Expression' module takes in the gene(s), mRNA and enzyme species of a reaction. Transcription, translation and degradation of mRNA and enzyme are expressed in this module. The 'Reaction' module constructs a reaction based on the number of substrates, products and reaction mechanism. An 'Integrative Model Unit' module is then created from either 'Gene Expression' or just the enzyme species only and 'Reaction' modules. These units are stored in a list ('Integrative Model List') before being converted into a SBML file by the Model2SBMLConvertor. The 'Parameter Estimation' module provides methods for estimating kinetic parameters in a model. SBML files imported by GRaPe are disintegrated into separated units by the SBML2ModelConvertor. The GUI serves as a platform for creating and editing the model.

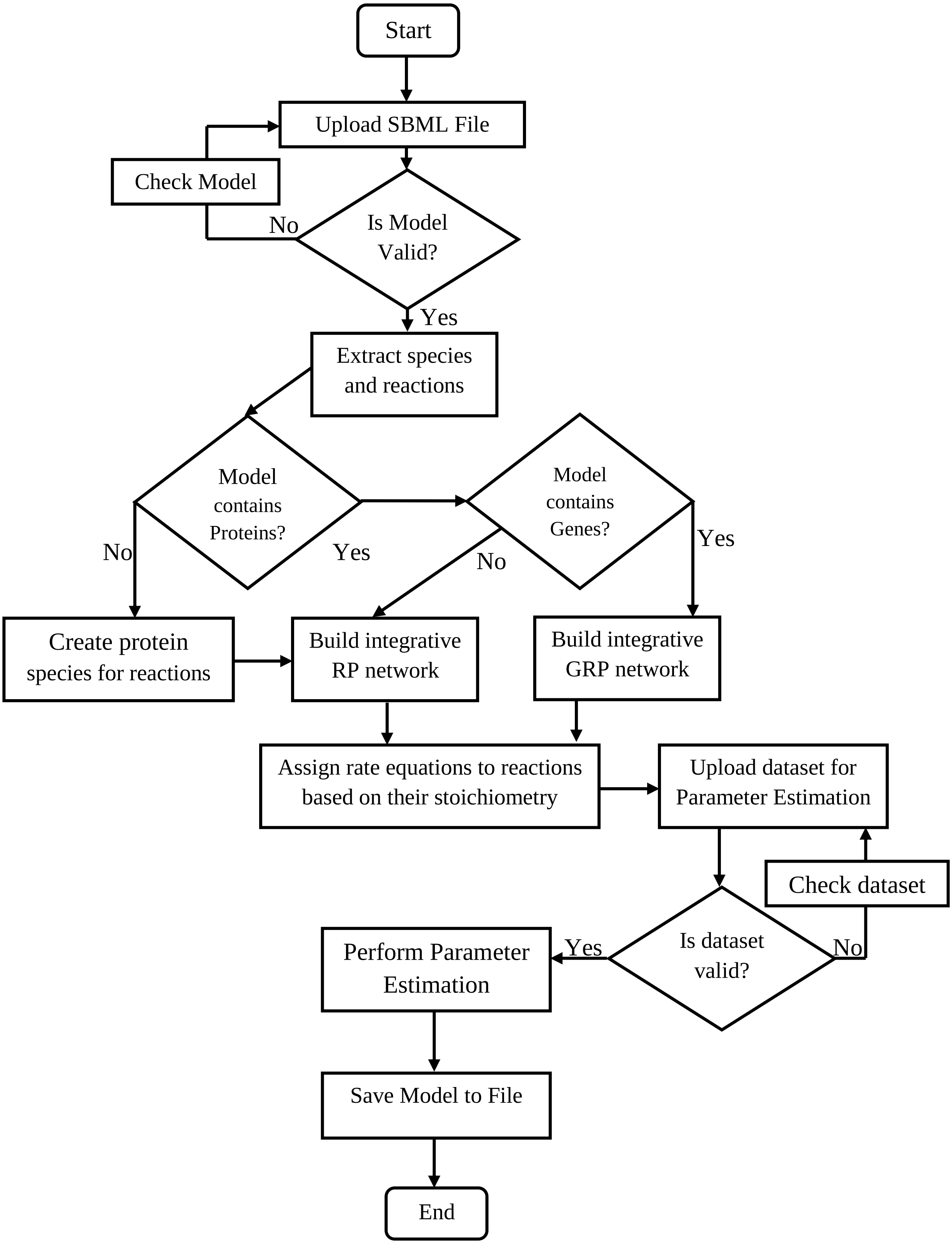

Start
Upload SBML File
Check Model
No
Is Model Valid?
Yes
Extract species and reactions
Model contains Proteins?
No
Yes
Model contains Genes?
No
Yes
Create protein species for reactions
Build integrative RP network
Build integrative GRP network
Assign rate equations to reactions based on their stoichiometry
Upload dataset for Parameter Estimation
Check dataset
Is dataset valid?
Yes
No
Perform Parameter Estimation
Save Model to File
End

Figure 16 depicts a flow chart for the model building process using GRaPe. Diamonds represents critical decision points. Rectangles represent tasks or activities. The success of GRaPe relies on the careful handling of both decisions and tasks.

2.3.2.1 Gene expression module (GEM)

The GEM is a template for creating gene expression process in the system. It takes in a gene, mRNA and an enzyme species. Each species has a unique identifier which is generated by the system, name (which could be a list of names when different genes encode the same enzyme) and an initial concentration which is a real value or the default value, 0.0. GEM generates the reaction rates of transcription, translation, degradation of both mRNA and proteins (enzyme) using the mass action kinetics equation which is then passed to the Integrative Model Unit.

The study initially makes an assumption that genes have one-to-one relationship with their corresponding enzyme and enzymes also have a one-to-one relationship with their reactions. However, this assumption that gene-protein-reaction (GPR) association is one-to-one is not necessarily true. Many genes encode a single protein which catalyses a single reaction. However, there are genes that also encode proteins/enzymes that can catalyse more than one reaction; these enzymes are known as *promiscuous enzymes.* For example, *Succinate Dehydrogenase*, SDH, is encoded by four genes and catalyses two reactions. We are still thinking of way(s) of reverting the one-to-one relationship of GPR to many-to-one or one-to-many relationship. The gene is a template for transcription and does not degrade over time. See future works for more details on how the study aims to model the expression of a gene at time, *t*.

2.3.2.2 Reaction module

The reaction module is responsible for creating individual metabolic reactions. Each reaction has a unique id, a name, a list of reactants and products. The type of reaction can either be reversible or irreversible. The reaction type can be set to 1 for reversible reactions and 0 for irreversible reactions. The reaction can take a maximum of two substrates and products, which are compounds. Each compound has a unique id, name, initial concentration, stiochiometric coefficient and a boundary condition, which can either be set to true or false.

For example, a reaction with A+B+C → P+Q, will be decomposed into two reactions, A+B→P, P+C→Q. This decomposition process, without a doubt, is a major challenge in this study as we will need to know the enzyme mechanism for that reaction.

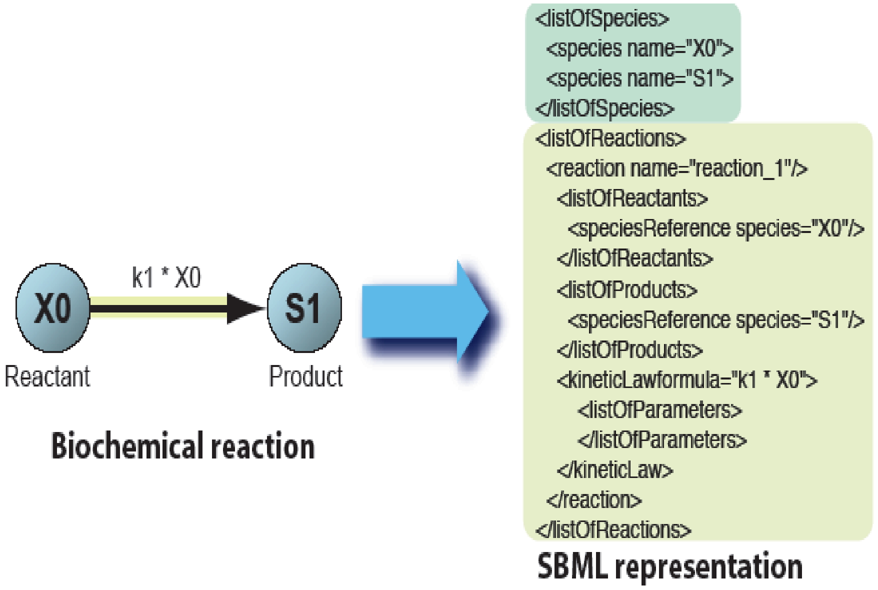
Depending on the number of substrates and products and the reversibility of that reaction, a Michaelis-Menten kinetic rate equation is generated by the system for that reaction. This is one of the main goals of this study. With automatic generation of reaction rate equations, large metabolic networks can be translated into mathematical equations which can then be integrated and analysed.

2.3.2.3 SBML2ModelConvertor and Model2SBMLConvertor

The SBML2ModelConvertor reads in a SBML file and converts it into java objects. The hierarchical structure of SBML file means it is easy to decompose the elements into objects which represent species, which can be metabolites, genes, mRNA or proteins, and reactions. The parser is able to read in SBML level 2 files in either version 1 or 2. With version 1 files, the kinetic equation is converted into MathML format. The gene expression module unit is assigned ‘null’ if the SBML file which is imported contains only reactions. Model2SBMLConvertor class converts a java model into a SBML level 2, version 2 file. The model is a list of individual IMUs which consists of a reaction with its gene expression module. IMU automatically generates the reaction rate for the reaction, together with the transcription, translation and degradation rates of both mRNA and protein.

2.3.3 SBML

Systems Biology Mark-up Language (SBML) is an extension of XML language, for representing models of biochemical reaction networks and it is considered the standard format in systems biology (Hucka, Finney *et al*, 2003). SBML can represent metabolic networks, cell-signaling pathways, regulatory networks, and other kinds of systems studied in systems biology (Hucka, Finney *et al*., 2003). This format provides interoperability between different modeling tools so that a description of a model by one program can be read and processed by other programs. Our system produces SBML level 2, version 2 documents, as it was the current version at the start of this study. SBML level 3 contains new elements and components that does not compromise our use of SBML level 2, version 2 documents. An example of the SBML structure is shown in Figure 17.



```
<listOfSpecies>
  <species name="X0">
  <species name="S1">
</listOfSpecies>
<listOfReactions>
  <reaction name="reaction_1"/>
    <listOfReactants>
      <speciesReference species="X0"/>
    </listOfReactants>
    <listOfProducts>
      <speciesReference species="S1"/>
    </listOfProducts>
    <kineticLawformula="k1 * X0">
       <listOfParameters>
       </listOfParameters>
    </kineticLaw>
  </reaction>
</listOfReactions>
```

Figure 17: XML representation of the structure of SBML.

A reaction, X0 → S1 and the SBML representation of this reaction in SBML level 2 version 1 format.

2.4 Methods

2.4.1 Modelling gene expression

Many methods have been used to model gene expression, which include Boolean networks (Kim *et al.*, 2007; Li, 2007), differential equations (Chen *et al.*, 1999), dynamic Bayesian networks (Kim *et al.*, 2003; Li, 2007) and neural networks (Gagneur and Klamt, 2004). Boolean networks are inexpensive with respect to computational complexity (D'haeseleer, 1999). An advantage of Bayesian networks, like Boolean networks, is that they do not require the explicit determination of kinetic parameters. However, both methods are poor in capturing some important aspects of network dynamics (Schlitt and Brazma, 2007).

We model gene expression using ordinary differential equations; this enables details of the dynamics of the network to be captured by explicitly modelling changes in concentrations of mRNA and proteins over time (Chen *et al.*, 1999). Also, gene expression levels tend to be continuous rather than discrete; discretization can lead to loss of information (D'haeseleer, 1999). Smolen *et al.*(2003) and Garcia-Martinez *et al.* (2007) studied the relationship among variables that characterize gene expression at the genome-level in living organisms. Their studies concluded that the amount of both mRNA and proteins primarily depends on their transcription rate (K_{Tr}), translation rate (K_{Tl}), mRNA concentration ([*mRNA*]) and protein concentration ([*Protein*]).

We model the change in mRNA concentration over time as:

$$\frac{d\left([mRNA]\right)}{dt} = v_{TR}Gene(t) - k_{mRNADeg}\,[mRNA] \qquad (13)$$

where [*mRNA*] is the mRNA concentration, v_{TR} is the transcription or mRNA synthesis rate, and k_{mRNADeg} is the mRNA degradation rate. *Gene*(*t*) is an expression function which may take any real value, enabling external regulators of gene expression to be incorporated into models. However, in the examples presented here,

no such regulation was included and these values were taken to be Boolean, where a value of 1 indicated that the gene was expressed and 0 otherwise.
The change in protein concentration over time is modelled as:

$$\frac{d([P])}{dt} = k_{Tl}\,[mRNA] - k_{\Pr oteinDeg}\,[P] \qquad (14)$$

where $[P]$ is the protein concentration, k_{Tl}, the protein synthesis or translation rate, $k_{\mathrm{ProteinDeg}}$ is the protein degradation rate and $[mRNA]$, the concentration of mRNA. Both the equations (13) and (14) are based on first order mass action kinetics.

Methods have been developed to determine or estimate the synthesis and degradation rates from microarray data (D'haeseleer, 1999; Wu *et al.*, 2004). When these parameters are unknown, we present an alternative modelling approach where only the protein level is integrated into the reaction instead of the full gene expression process (Figure 14b). The concentration of enzyme can be set as fixed over time or varied during simulation. This can be achieved by the use of SBML 'Events' that represent time-dependent changes within the system. An event can be triggered if a certain condition is reached; for example, set the concentration of enzyme 'A' to 0.8 mM if time is >10 min. Furthermore, the modelling framework enables the incorporation of isoenzymes that can be modelled as individual enzyme species with their own gene expression processes.

2.4.2 Enzyme kinetics and rate equations

GRaPe automatically generates a rate equation for a reaction in a network based on the assumed enzyme mechanism governing that reaction, and its number of products and substrates. The enzyme mechanism of a reaction can either be of random order or compulsory order. If the binding order of substrates and releasing order of products are unknown then the random-order mechanism is recommended. The compulsory-order mechanism requires the knowledge of the correct order of binding of substrates to the proteins and releasing of products to be known. This mechanism proceeds in an ordered series of steps, i.e. the substrate must bind in particular order and the product is released in a specified order.

The automatic generation of generic rate equations is a key advantage of GRaPe. This is time-efficient and less error-prone for a relatively large model compared to the manual definition of each rate equation by the user. COPASI provides predefined rate equations for reactions in a system, but these equations must be manually assigned to reactions by the user and protein levels cannot be explicitly assigned in rate equations. SBMLSqueezer is a plug-in for CellDesigner, which generates rate equations for a biochemical reaction (Dräger, 2008); but the user has to manually select the type of rate equation for each reaction in the network, and protein levels are not integrated in rate equations. On a large scale, these solutions are impractical and leave the model with unknown parameters.

GRaPe uses the King & Altman method (King and Altman, 1956) to derive rate equations based on a reaction's stoichiometry and the enzyme mechanism under the steady-state assumptions. The description of generic Michaelis–Menten rate equations used by GRaPe for the different reaction types is given below.

It generates kinetic rate equations for reactions of up to two substrates or products, i.e. reactions can be of type uni–uni, uni–bi, bi–uni or bi–bi. A reaction of more than three substrates or products needs to be decomposed into these reaction types based on its biochemistry.

2.4.2.1 Reversible Michaelis-Menten rate equations

For a Uni-Uni reversible reaction (both random and compulsory order)

A ←→ P:

$$v = \frac{e_0 (\frac{v_+^m a}{K_{mA}} - \frac{v_-^m p}{K_{mP}})}{1 + \frac{a}{K_{mA}} + \frac{p}{K_{mP}}} \tag{15}$$

where *a* is the concentration of substrate A, *p* is the concentration of product P. V^m_+ is the rate of consumption of A (velocity of the forward reaction) and V^m_- is the rate of formation of P (velocity of the backward reaction); K_{mA} and K_{mP} are the Michaelis constant for A and P respectively; e_0 is the total concentration of the enzyme and *v* is the rate (or velocity) of reaction. $V^m_+ / K_{mA} = K_A$ and $V^m_- / K_{mP} = K_p$ in a traditional reversible Michaelis-Menten rate equation.

For a Bi-Uni reversible reaction, A + B←→P

Compulsory order:

$$v = \frac{e_0 \left(\frac{v_+^m ab}{K_{iA} K_{mB}} - \frac{v_-^m p}{K_{mP}} \right)}{1 + \frac{a}{K_{iA}} + \frac{K_{mA} b}{K_{iA} K_{iB}} + \frac{ab}{K_{mB} K_{iA}} + \frac{p}{K_{mP}}} \tag{16}$$

Random order:

$$v = \frac{e_0\left(\dfrac{v_+^m ab}{K_{iA}K_{mB}} - \dfrac{v_-^m p}{K_{mP}}\right)}{1 + \dfrac{a}{K_{iA}} + \dfrac{b}{K_{iB}} + \dfrac{ab}{K_{mB}K_{iA}} + \dfrac{p}{K_{mP}}} \qquad (17)$$

In both equations (16) and (17), V^m_+ is the rate of consumption of A and B, and V^m_- is the rate of formation of P. *a* is the concentration of substrate A, *b* is the concentration of substrate B and *p* is the concentration of product P. e_0 is the total concentration of the enzyme and *v* is the rate (or velocity) of the reaction., K_{iA} and K_{iB} are the substrate dissociation constants of A and *B.* K_{mB} and K_{mP} are the Michaelis constants for B and P respectively, K_{iA}, K_{iB}, K_{mB}, K_{mP} are dissociation constants. Since the binding of A and B are interchangeable, $K_{iA}K_{mB} = K_{iB}K_{mA}$.

For a Uni-Bi reversible reaction, A←→P + Q

Compulsory order:

$$v = \frac{e_0\left(\dfrac{v_+^m a}{K_{mA}} - \dfrac{v_-^m pq}{K_{mP}K_{iQ}}\right)}{1 + \dfrac{a}{K_{mA}} + \dfrac{K_{mQ}p}{K_{mP}K_{iQ}} + \dfrac{pq}{K_{mP}K_{iQ}} + \dfrac{q}{K_{iQ}}} \qquad (18)$$

Random order:

$$v = \frac{e_0\left(\dfrac{v_+^m a}{K_{mA}} - \dfrac{v_-^m pq}{K_{iP}K_{mQ}}\right)}{1 + \dfrac{a}{K_{mA}} + \dfrac{p}{K_{iP}} + \dfrac{pq}{K_{iP}K_{mQ}} + \dfrac{q}{K_{iQ}}} \qquad (19)$$

In (18) and (19), $V^{m}{}_{+}$ is the rate of consumption of A, and $V^{m}{}_{-}$ is the rate of formation of P and Q. a is the concentration of substrate A, p is the concentration of substrate P and q is the concentration of product Q. e_0 is the total concentration of the enzyme and v is the rate (or velocity) of the reaction, K_{iP} and K_{iQ} are the product dissociation constants of P and Q. K_{mA}, K_{mP} and K_{mQ} are the Michaelis constants for A, P and Q respectively, K_{mA}, K_{iP}, K_{mQ}, K_{iQ} are dissociation constants. Since the release of P and Q are interchangeable, $K_{iP}K_{mQ} = K_{iQ}K_{mP}$.

For a Bi-Bi reversible reaction, A + B←→P + Q

Compulsory order:

$$v = \frac{e_0\left(\frac{v_+^m ab}{K_{iA}K_{mB}} - \frac{v_-^m pq}{K_{mP}K_{iQ}}\right)}{1 + \frac{a}{K_{iA}} + \frac{K_{mA}b}{K_{iA}K_{mB}} + \frac{K_{mQ}p}{K_{mP}K_{iQ}} + \frac{q}{K_{iQ}} + \frac{ab}{K_{iA}K_{mB}} + \frac{K_{mQ}ap}{K_{iA}K_{mP}K_{iQ}} + \frac{K_{mA}bq}{K_{iA}K_{mB}K_{iQ}} + \frac{pq}{K_{mP}K_{iQ}} + \frac{abp}{K_{iA}K_{mB}K_{iP}} + \frac{bpq}{K_{iB}K_{mP}K_{iQ}}}$$

(20)

Random order:

$$v = \frac{e_0\left(\frac{v_+^m ab}{K_{iA}K_{mB}} - \frac{v_-^m pq}{K_{mP}K_{iQ}}\right)}{1 + \frac{a}{K_{iA}} + \frac{b}{K_{iB}} + \frac{p}{K_{iP}} + \frac{q}{K_{iQ}} + \frac{ab}{K_{iA}K_{mB}} + \frac{pq}{K_{mP}K_{iQ}}} \qquad (21)$$

where in both (20) and (21), $V^{m}{}_{+}$ is the rate of consumption of A and B, and $V^{m}{}_{-}$ is the rate of formation of P and Q. a is the concentration of substrate A, b is the concentration of substrate B, p is the concentration of substrate P and q is the concentration of product Q. e_0 is the total concentration of the enzyme and v is the rate (or velocity) of reaction.

In (20) K_{iA}, K_{iP} and K_{iQ} are the product dissociation constants of A, P and Q. K_{mA}, K_{mB}, K_{mP} and K_{mQ} are the Michaelis constants for A, B, P and Q respectively. In (21), K_{iA}, K_{iB}, K_{iP}, K_{iQ} and K_{mB} are dissociation constants. Since the release of P and Q are interchangeable, $K_{iP}K_{mQ} = K_{iQ}K_{mP}$, and the binding of A and B are interchangeable, $K_{iA}K_{mB} = K_{iB}K_{mA}$.

2.5 Parameter Estimation

Kinetic models are shown to produce accurate and testable results. However, the number of large-scale kinetic models has been very low due to the enormous number of kinetic parameters needed to define the system. Furthermore, as observed in Teusink *et al.* (2000), *in vitro* measurements of kinetic constants may not necessarily be representative of their numerical values *in vivo* (Jamshidi and Palsson, 2008). Various software tools can now perform parameter estimation: COPASI has a list of methods for estimation including a GA; SBML-PET (Zhike and Klipp, 2006) uses a stochastic ranking evolution strategy method for parameter estimation; however, it excludes constraints on the flux of a reaction implying that a zero flux may be obtained even in non-equilibrium conditions. COPASI requires that columns specified in the experimental data file must be associated with model elements. Having a flux for a reaction in the experimental data file throws an error with COPASI, as fluxes are not explicitly expressed in a model.

While the exact values of kinetic parameters are not necessarily crucial to determine the behaviour of a biochemical system, it is nevertheless necessary to estimate the values of missing parameters in order to run simulations. In GRaPe, we have introduced a simple but effective algorithm, the SSEE that estimates the concentration of enzyme needed for a reaction to reach a steady state, *v*. The SSEE algorithm focuses on solving for e_0, the concentration of the enzyme, by assigning all kinetic parameters in the models to a constant value of 1. For a uni–uni reaction, e_0 is calculated as follows:

$$e_0 = \frac{v(1 + \frac{a}{K_{mA}} + \frac{p}{K_{mP}})}{(K_A a - K_P p)} \qquad (22)$$

where *KmA*, *KmP*, *KA* and *KP* are the kinetic parameter associated to substrate A and product P, respectively, assigned a value of 1; *a* is the concentration of substrate and *p* is the concentration of the product. SSEE allows for the rapid simulation of steady-state behaviour in a system without prior knowledge of kinetic parameters.

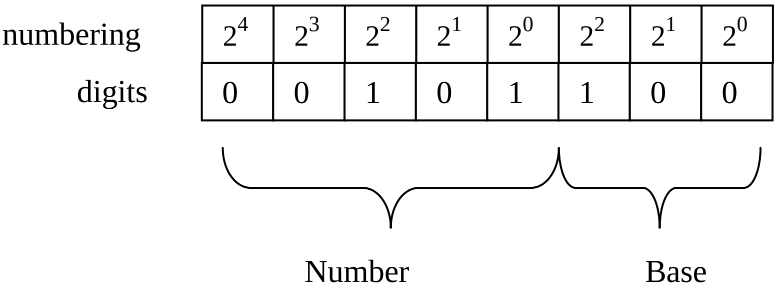
In addition, GRaPe implements two methods for parameter estimation from time series of experimental data: the LMA (Levengberg-Marquardt algorithm) and GA. The LMA is an upgrade from Nocedal and Wright (1999), which was constrained to work with our integrative models. GA is the predominant algorithm for estimating kinetic parameters in GRaPe. However, when GA does not return a good solution based on an objective function, then the LMA is used. GRaPe returns the solution of the algorithm with the better objective function.

2.5.1 Genetic Algorithm

Parameter optimization was performed using a genetic algorithm (GA). Genetic algorithms are heuristic optimization techniques that take their inspiration from concepts of natural evolution. Specifically a basic genetic algorithm works by "evolving" a population of solutions to a problem. Evolution (i.e. improvement) of solutions is achieved through subsequent rounds of (i) reproduction, (ii) variation and (iii) selection of solutions.

For the parameter optimization problem, a solution to the problem was required to provide estimates for all of the parameters within a given reaction. Each parameter was encoded as a bit string of size twelve (eight bits encoding for the number and four bits encoding for the base) and gray coding was used to map the individual bit strings to real numbers within the interval [1.0e-10, 1.5e10]

Here, gray coding was used instead of binary encoding in order to reduce the number of local optima in the fitness landscape. To avoid convergence to local optima, the GA also used a large population size of 1000 individuals in combination with tournament selection of size three, resulting in relatively low selection pressure. The variation operators used were one-point crossover and bit-flip mutation, which were applied with standard probabilities of 0.7 and 1/L, respectively (where L is the length of a complete solution, i.e. the number of parameters times twenty). The GA was run until the summed least squared error (equation 23 below) dropped below 1e-7 or until the maximum number of generations (2000000) had been achieved.



numbering	2^4	2^3	2^2	2^1	2^0	2^2	2^1	2^0
digits	0	0	1	0	1	1	0	0

Fig 18. Shows an example of an eight-bit string. The first 5 bits encodes for the number and the last 3 encodes for the base. The above example translates as 5_4 = 625 (0*16+0*8+1*4+0*2+1*1).

The summed least squared error is expressed as follows:

(23)

where *s* is the sum of the squares of the error or distance between the actual outcome, $x_{i,}$, and the predicted outcome, $y_{i.}$ *n* is the number of data points in the dataset.

2.5.2 Example of how GA works

The input file for the optimization program is a tab-delimited data file (I_data) with a full listing of metabolites' concentrations, fluxes and enzyme concentrations for a metabolic network (See sample dataset below of first two reactions of glycolysis in yeast with proteins *HK* and *PGI*)..

GRaPe extracts the reaction data (R-data) for a reaction based on the reaction id (or flux id), substrate(s) id, product(s) id and enzyme id (and not their ordering in the file).

v1_HK	s1	s4	s2	s3	s6	s5	v2_PGI	s7
47.9937	0.573074	1.5	2.1	4.2	1.0	0.49	47.9938	1
47.9937	0.573074	1.5	2.1	4.2	1.0	0.49	47.9938	1
47.9937	0.573074	1.5	2.1	4.2	1.0	0.49	47.9938	1
47.9937	0.573074	1.5	2.1	4.2	1.0	0.49	47.9938	1

Table 3: A sample input data for GA. The first row represents IDs for fluxes, metabolites and enzymes. The flux data values for reaction, *HK*, are in the first column. s1 and s4 are the substrates and s2 and s6 are the products of reaction *HK;* s6 is the enzyme¸*HK*. GRaPe matches IDs in the SBML model to IDs in the parameter estimation input file.

GRaPe starts by fetching the first reaction from a list (V1_HK) of reactions. For each reaction, the appropriate rate equation is chosen based on the stoichiometry of the reaction. GRaPe then uses the GA to estimate kinetic parameters for the reaction rate equation by minimising the sum of the squared errors (equation 23) between the actual and predicated reaction rate values.

The Levenberg-Marquardt algorithm, LMA or genetic algorithm, GA may be used. If one method produces a solution not "good" enough the other method is employed to find a better solution.

2.6 Discussion

We have introduced GRaPe, a platform independent software tool aimed at streamlining the construction of large-scale dynamic models. GRaPe enables the automated construction of reaction-protein or gene-reaction-protein networks. A novel feature of GRaPe is its ability to generate generic rate equation for models of relatively large sizes. Another important feature is its capability to explicitly integrate gene expression processes or enzyme species into reactions, making it a convenient tool for the construction of integrative protein-reaction networks.

A few manually constructed integrative metabolic models have now been created (Förster and Palsson 2003, Jamshidi and Palsson 2008, Ao *et al*. 2008), however, no computational tool for integrating protein levels into metabolic models exists. The integration of proteomics data into metabolic models could increase our understanding of the role of enzymes on metabolism. Another important feature of GRaPe is its ability to convert existing metabolic models in SBML format into either gene-reaction or gene-protein-reaction networks. This will enable, for example, the import of high-throughput quantitative proteomics data into metabolic models.

Parameter estimation (optimisation) has become an area of significant importance in kinetic modelling due to the fact that it is often prohibitively expensive and time-consuming to measure vast numbers of kinetic parameters experimentally. Some repositories such as Sabio-RK (Rojas *et al*, 2007) and BRENDA (Schomburg *et al.* 2002) store kinetic parameters and enzymatic information for various pathways in different organisms.

However, it is difficult to compare parameters of the same pathway in different models due to different assumptions and experimental conditions. Gutenkunst *et al.* (*2007*) suggested that modellers should focus on predicting the behaviour of the system rather than parameters due to parameter "sloppiness", meaning that parameters are often poorly constrained. GRaPe introduces a simple but effective method for estimating the amount of enzyme concentration required to give a particular steady state. This method enables analysis of the steady state behaviour without detailed knowledge of kinetic parameters. GRaPe also has two methods for parameter estimation: the Levengberg-Marquardt algorithm (LMA) and a genetic algorithm (GA). Both methods are robust and work interchangeably in estimating kinetic parameters.

The capability of GRaPe to convert reactions into ODEs based on their stoichiometric matrix for small or large-scale networks, its main innovation, makes it complementary to other existing software tools. GRaPe is not designed to compete with well-developed simulation software but to complement existing applications by providing an upstream solution for the efficient design of large-scale dynamic models. Models created using GRaPe can be run using existing simulation tools such as CellDesigner, COPASI and any other tools that support SBML. In the future, we aim to interface GRaPe with existing databases of metabolic reactions and kinetic parameters in order to make it capable of rapidly constructing large or genome-scale integrative kinetic models.

2.7 References

1. Ao P, *et al*. (2008).Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AM1 growth as validation. *Chin. J. Biotechnology*. 24:980-994

2. Baitaluk,M. *et al*. (2006) Biological Networks: visualization and analysis tool for systems biology. *Nucleic Acids Research*, 34 (Web Server issue), W466-W471.

3. Chen T, *et al*. (1999). Modeling gene expression with differential equations. *Pac. Symp. Biocomput*. 4. pp. 29-40.

4. Cornish-Bowden,A. (2004). Fundamentals of Enzyme Kinetics, 3rd ed, Portland Press. 30-51.

5. D'haeseleer, P., Wen, X. *et al*. (1999) Linear modeling of mRNA expression levels during CNS development and injury. *Pacific Symposium on Biocomputing*, 4, 41-52.

6. Dhar, P. *et al*. (2004) Grid Cellware: The first Grid-enabled tool for modelling and simulating cellular processes. *Bioinformatics*, 20(8), 1319-1321.

7. Dräger, A., N. Hassis, *et al*. (2008) SBMLsqueezer: A CellDesigner plug-in to generate kinetic rate equations for biochemical networks. *BMC Systems Biology*, 2, 39.

8. Förster, J. Famili, I., Fu, P., Palsson, B. and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*. 13: pp. 244-253.

9. Funahashi, A. *et al*. (2003) CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*, 1, 159-162.

10. Gagneur,J. and Klamt,S. (2004) Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5, 175.

11. Garcia-Martinez, J., Gonzalez-Candelas,F. *et al*. (2007) Common gene expression strategies revealed by genome-wide analysis in yeast. *Genome Biology*, 8, R222.

12. Gutenkunst, R. N., Waterfall, J. J. *et al*. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* 3(10), 189.

13. Haseong,K. *et al*. (2007) Boolean networks using the chi-square test for inferring large-scale gene regulatory networks. *BMC Bioinformatics*, 8, 37.

14. Hucka, M. *et al*. (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19, 524-531.

15. Jamshidi, N. and Palsson, B. Ø. (2008) Formulating genome-scale kinetic models in the post-genome era. *Molecular Systems Biology*, 4, 171.

16. King, E. L. and Altman,C. (1956) A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *Journal of Physical Chemistry*, 60, 1375-8.

17. Kitano, H. (2002) Systems Biology: A Brief Overview . *Science*, 295, 1662-4.

18. Kurata H, Matoba N, Shimizu N. (2003) CADLIVE for constructing a large-scale biochemical network based on a simulation-directed notation and its application to yeast cell cycle. Nucleic Acids Res. 15;31(14):4071-84.

19. Mendes,P., Hoops,S. *et al*. (2006). COPASI: a COmplex PAthway SImulator. *Bioinformatics*, 22, 3067-74.

20. Schlitt, T. and Brazma, A. (2007). Current Approaches to gene regulatory network modelling. *BMC Bioinformatics*. 8(Suppl 6)**:**S9.

21. Smallbone, K. *et al*. (2007) Something from nothing: bridging the gap between constraint-based and kinetic modelling. *FEBS J*. 274(21), 5576-85.

22. Smolen,P., Baxter, D.A. *et al*. (2003) Mathematical modeling of gene networks. *Neuron*, 26(3), 567-580.

23. Steuer, R. *et al.* (2006) Structural Kinetic modelling of metabolic networks, *Proc Natl Acad Sci USA*, 103 (32), 11868-73.

24. Teusink,B., Passarge, J. *et al*. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *European Journal of Biochemistry*, 267(17): 5313-5329.

25. Tomita,M., Hashimoto, K. *et al* (1999) E-CELL: Software environment for whole cell simulation. *Bioinformatics*, 15(1), 72-84.

26. Vass,M. *et al* (2004) The JigCell Model Builder and Run Manager. *Bioinformatics*, 18, 3680-3681.

27. Wu,C.C., Huang H.C. et al (2007) GeneNetwork: an interactive tool for reconstruction of genetic networks using microarray data. *Bioinformatics*, 20, 3691-3.

28. Zhike, Z. and Klipp, E. (2006) SBML-PET: a Systems Biology Markup Language-based parameter estimation tool. *Bioinformatics*, 22(21), 2704-5.

Chapter 3

A Generic Model of Yeast Glycolysis

"*It is a capital mistake to theorise before one has data. Insensibly one begins to twist facts to suit theories instead of theories to suit facts."* Sherlock Holmes, the fictional creation of Arthur Conan Doyle (1859-1930)

The work in this chapter appeared in "Streamlining the construction of large-scale dynamic models using generic kinetic equations." Adiamah, DA., Handl, J, Schwartz, J-M. (2010) *Bioinformatics*, **26**, 1324 – 1331".

The aim of this study was to test our modelling approach and software tool together with its features including parameter estimation technique to determine the success rate of our approach. The yeast glycolysis pathway was reconstructed using GRaPe. We show that using generic rate equations, the dynamical behaviour of system can achieved.

3.1 Abstract

There has been a rapid development in the construction of stoichiometric models of metabolic systems at the scale of entire organisms, for many species from microorganisms to humans. However, these models provide limited insights into the functioning of cellular processes since their use is restricted to steady-state simulations. Efforts are now being directed toward the development of dynamic modelling techniques that are able to cope with the large scale required for cell-wide simulation.

The construction of large scale metabolic models is challenging as it requires the assembly and solving of systems of several hundred of non-linear differential equations. Inadequate knowledge of the enzyme kinetic rate laws and their associated parameter values still hampers large-scale kinetic modelling. Nevertheless, due to the robustness and resilience of biological systems to perturbation, exact rate equations and accurate parameters values are often not crucial in determining the fundamental dynamic properties. Therefore, a degree of generalization and simplification may be applied to reduce model complexity and to streamline model construction.

Using the yeast glycolysis pathway as an example, we showed that a model based on generic kinetic equations and estimated parameters is able to simulate a metabolic system with a comparable accuracy to a detailed model based on experimental determinations. We have implemented this methodology in the GRaPe software. These principles may be used to automate the construction of large models of intracellular metabolism and achieve the scales needed for cell-wide dynamic modeling. Our model also accurately describes dynamic experiments with a changed glucose influx, even though such data were not used for parameter estimation, showing that the generic model has predictive value.

3.2 Introduction

The development and analysis of constraint-based modeling techniques have seen a significant rise. This has subsequently led to the increase in the construction of stoichiometric models of metabolic systems on both small and large scale (Dobson *et al*, 2010). However, constraint-based modeling is limited in capturing the dynamics of cellular behavior (Schlitt and Brazma, 2007). As a result, many studies have geared towards developing generic techniques to building kinetic models which are capable of capturing the dynamical properties of a network. One of the early attempts was by Hynne *et al* (2001) where a general method of fitting experimental data to the mechanism of a model was provided. Here, simple algebra was used in calculating rate constants and maximum velocities for all reactions in the full-scale model of glycolysis in *S. cerevisiae*.

There is now an increasing number of *in silico* metabolic models which can serve as platform for ‘dry lab’ prototyping of experiments with the aim of developing hypothesis which can then be performed in the lab (Jamshidi and Palsson, 2009). Constructing these *in silico* metabolic models on a large scale is challenging as it requires the assembly and solving of systems of several hundred of non-linear differential equations. The main problem of building kinetic models is the lack of or inadequate knowledge of the enzyme kinetic rate laws and their associated parameter values (Jamshidi and Palsson, 2008). Other challenges in building *in silico* model were highlighted by Palsson (2000) to include the difficulty in integrating biological processes and issues with simplifying systems properties.

It is now known that due to the robustness and resilience of biological systems to perturbation, the exact rate equations and accurate knowledge of kinetic parameters values are often not crucial in determining the fundamental dynamic properties (Gutenkunst *et al*, 2007). As a result, there is now the need of generalization and simplification of the model building process in order to reduce model complexity and to also streamline model construction.

Molecular and cell biology of yeast, in particular, have been widely studied in recent times – leading to a large amount of both qualitative and quantitative data (Klipp, 2007). This makes it the ideal model to test our generic approach to modelling metabolic networks. Intuitively, one assumes that once the generic approach works on a smaller test case – the yeast glycolysis pathway with a small number of reactions – then if the same approach is implemented on a large-scale, the results should be coherent. However, what has been acknowledged in the past is that, the behavior of a complex system is often too difficult to understand by intuition alone as feedback loops, cycles or interplay of processes makes a considerable difference (Klipp, 2007).

A comprehensive study by Teusink *et al* (2000) measured *in vitro* kinetic parameters of most of the glycolytic enzymes of *S. cerevisiae* in an attempt to understand whether *in vivo* behavior can be understood in relation to *in vitro* kinetic properties. Interestingly, the results from the Teusink yeast model, built with accurately measured data, showed discrepancies with experimental results. However, to date, the Teusink yeast glycolysis model still remains as one of most accurate models as kinetic parameters and enzyme concentrations were experimentally measured. This process was, undoubtedly, tedious and time-consuming – again prompting the need for powerful but generic approaches to building models with good predictive prowess. Additionally, on a large-scale, experimentally measuring all kinetic parameters and enzymatic properties can be an impossible task.

In our attempt to validate our approach to building models using generic kinetic rate equations, we build a model of yeast glycolysis pathway based on data presented in a previous study.

3.3 *S. cerevisiae* Glycolytic Pathway

We modelled the glycolysis pathway in *S.cerevisiae* using generic rate equations and kinetic parameters estimated by GRaPe and compared it to a detailed model by Teusink *et al.* (2000). Since the main objective of this work is to show that a generic kinetic model can provide results of similar quality to a detailed model, we compare our results with simulations of the Teusink model rather than experimental values.

The yeast glycolysis pathway has been extensively studied (Bakker *et al.*, 1997; Hynne *et al.*, 2001; Lambeth and Kushmerick, 2002; Pritchard and Kell, 2002; Teusink *et al.*, 2000) and a vast amount of genomic and enzymatic data is therefore available. In Teusink *et al.* (2000), the authors examined whether *in vivo* kinetics behaviour can be understood in terms of *in vitro* kinetics of enzymes in yeast glycolysis. They produced two models, one where branched reactions were ignored and a second comprehensive model that included all branched reactions. Their results suggested that half of the enzymes matched their *in vivo* fluxes within a factor of 2, and the calculated deviation between *in vivo* and *in vitro* kinetic characteristics of the other enzymes could explain discrepancy between *in vivo* and *in vitro* kinetics. Fluxes and metabolites concentrations were experimentally determined. Fluxes of trehalose and glycogen were expressed in units of glucose, and kinetic parameters were also determined under the same experimental condition. The unbranched model used experimentally determined metabolite concentrations and calculated conserved moieties but no steady state was reached. The branched model, however, reached a steady state with the original parameter set that had been determined *in vitro*. Both models used a set of ordinary differential equations to describe the time dependence of metabolite concentration.

3.4 Methods

We modelled the glycolysis pathway of *S.cerevisiae* using GRaPe based on the branched topology used by Teusink *et al.* (2000). The initial concentration of metabolites was the same as in the Teusink model. The model includes all enzymes involved in the pathway from glucose uptake to the production of pyruvate and ethanol. All reactions were assumed to be of a random-order mechanism. GRaPe then generated generic reversible rate equations for the reactions in our glycolysis model, which were combined with the stoichiometry of the network to produce ordinary differential equations. We made three distinctive changes in our model. First, in the Teusink model, a metabolic pool represented by an independent variable, *P*, was defined to represent the sum of high-energy phosphate in adenine nucleotides. In our model, an equation is used for the conservation of adenine nucleotides moiety as:

$$\frac{d[ADP]}{dt} = -\frac{d[ATP]}{dt} \qquad (24)$$

where [ADP] and [ATP] are the concentration of adenosine diphosphate (ADP) and adenosine triphosphate (ATP), respectively and *t* is the time.

Secondly, since adenosine monophosphate (AMP) does not partake in any reaction, we excluded the adenylate kinase reaction. Thirdly, in the Teusink model the triosephosphate isomerase (TPI) reaction was modelled using an equilibrium equation such that the ratio of glyceraldehyde 3-phosphate (GAP) to glycerone phosphate (DHAP) was at equilibrium. An independent variable, *Trio2-P*, was introduced, which was the sum of the concentration of GAP and DHAP. In our model, we included the TPI reaction and modelled the change in DHAP and GAP concentrations using uni–uni reversible rate equation. These changes make it possible to study the effects of varying the concentration of ATP, ADP, DHAP and GAP on the system. The initial concentration of all metabolites was then assigned using data given in the Teusink model. The concentrations of ATP and ADP were calculated based on conserved moiety equations given in Teusink *et al.* (2000). As cofactors play an important part in the global regulation of glycolysis, their concentrations were treated as free metabolic variables. The kinetic equation for each reaction was generated automatically by GRaPe based on the number of substrates and products and the enzyme mechanism of the reaction. Our glycolysis model in Figure 18 contains 23 metabolites (22; here, the bracketed data corresponds to respective entity in the Teusink model), 15 enzyme species (0), 116 kinetic parameters (88) and 18 fluxes (17). Below is a list ordinary differential equations used to describe the time-dependence of metabolites in the network:

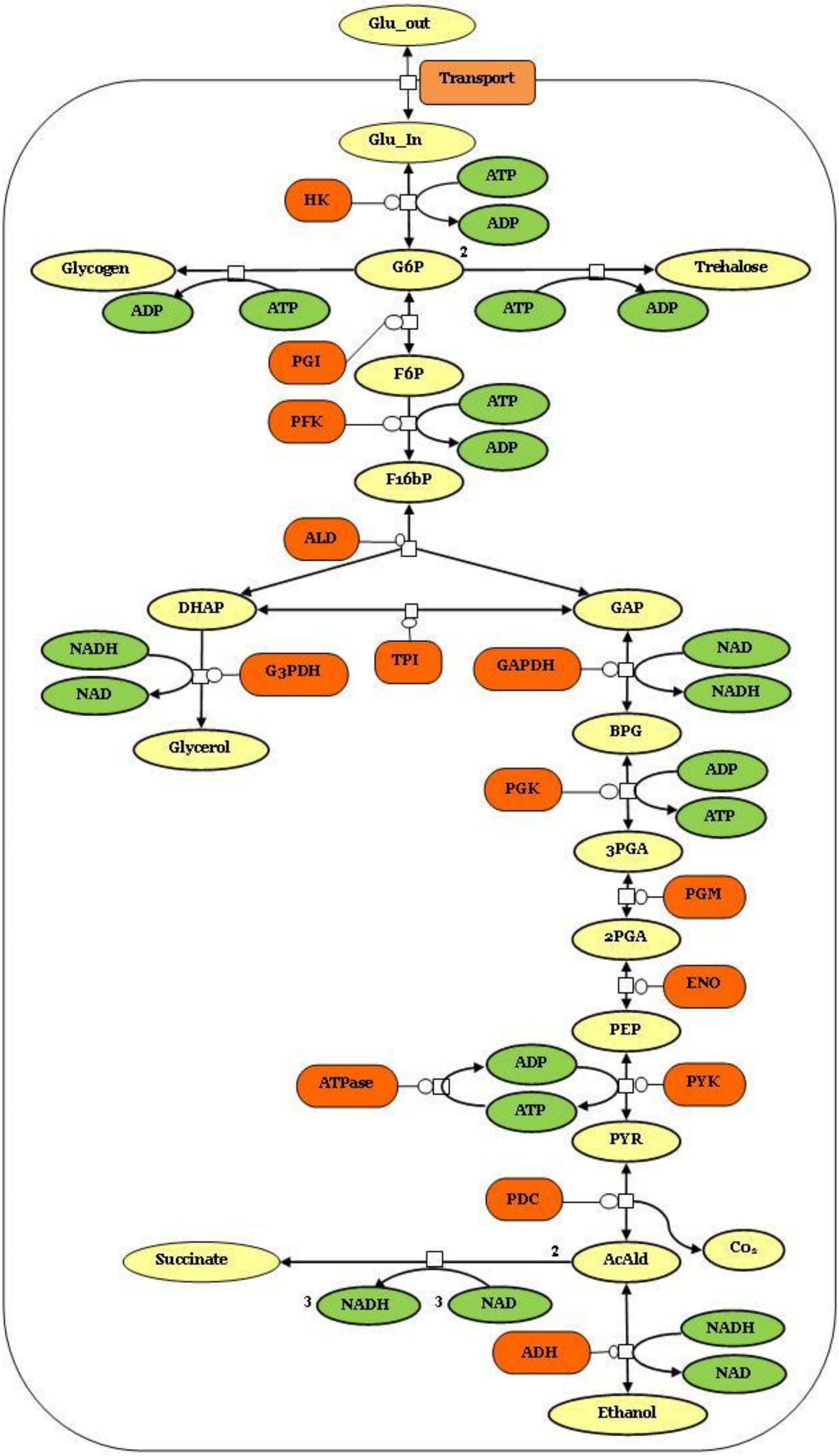


Figure 19: Topology of our yeast glycolysis model.

$$\frac{\quad}{\quad} \qquad - \qquad (25)$$

$$\frac{\quad}{\quad} \qquad - \qquad - \qquad (26)$$

$$\frac{\quad}{\quad} \qquad - \qquad (27)$$

$$\frac{\quad}{\quad} \qquad - \qquad (28)$$

$$\frac{\quad}{\quad} \qquad - \qquad - \qquad (29)$$

$$\frac{\quad}{\quad} \qquad - \qquad (30)$$

$$\frac{\quad}{\quad} \qquad - \qquad (31)$$

$$\frac{\quad}{\quad} \qquad - \qquad (32)$$

$$\frac{\quad}{\quad} \qquad - \qquad (33)$$

$$\frac{\quad}{\quad} \qquad - \qquad (34)$$

$$\frac{\quad}{\quad} \qquad (35)$$

$$\frac{\quad}{\quad} \qquad - \qquad - \qquad (36)$$

$$\frac{\quad}{\quad} \qquad \frac{\quad}{\quad} \qquad (37)$$

Boundary Conditions are set to “True” for the following metabolites Glucose_out, Trehalose, Glycogen, Glycerol, Succinate, CO_2 and Ethanol as we assume that the level of these metabolities are not affected by internal and external reactions or conditions.

3.5 Data acquisition and parameter estimation

Using JWS online (Olivier and Snoep, 2004), a web tool for simulating kinetic models, we collected steady-state data for metabolites and fluxesfrom the Teusink model with glucose uptake concentrations of 10 and 50 mM. These values were then merged to create our input dataset for parameter estimation. The dataset contains values of all metabolite concentrations and reaction fluxes in the glycolysis pathway at every 10 min; the glucose concentration was at 50 mM from time 0 to 30 min and at 10 mM from 30 to 100 min; the precision of values was limited to two decimals points for faster estimation.

Next, GRaPe was used to estimate the kinetic parameters for each reaction in our model so that the distance between the input dataset (Teusink experimental data) and our predicted values calculated by the model was minimized. Due to the absence of gene expression and enzyme amount data in this example, the concentration of enzymes was set to a default value of 1 in both the model and dataset for parameter estimation. GA was used to estimate the kinetic parameters for each reaction after just one run of estimation. The calculated error over our input data ranged from 4.5e-13 to 2.13e-10 for all reaction's kinetic parameter sets.

3.6 Results

3.6.1 Experiment 1: Model validation on training data

After parameter estimation was completed, the model was simulated in CellDesigner using the SBML ODE Solver (SOSlib). The results obtained from our simulations were then compared with results from the Teusink model. The first experiments were to verify whether the model correctly reproduced the behaviour of the system at steady state, without any perturbation, when glucose uptake is at 10 and 50 mM. Our results show a near-perfect agreement between our model and the Teusink model (Tables 4 and 5). These results confirmed that the training and parameter estimation algorithms successfully identified a solution, where the model reproduces the correct concentration and flux values used in the training data. The dataset used for parameter estimation is provided on our lab group page (pg. 148).

Using ‘events’ in SBML, we moreover replicated the effect of a dynamically reduced uptake of glucose: after 30 min, the concentration of glucose was reduced from the original 50 to 10 mM. Results from this experiment (figures. 20a–b) again show an excellent agreement between our model and the Teusink model when the same reduction in glucose uptake is applied. These results confirm that the integrity of our estimated parameters is conserved in a dynamic experiment. Our model of glycolysis, with events for changing the level of glucose uptake, has been provided on our lab group page (pg.148).

3.6.2 Experiment 2: Model validation outside the training range

The second experiment was carried out to verify how well the model would predict a new state of the glycolysis pathway outside the range of training experimental data and without re-estimating the kinetic parameters. We carried out simulations by changing the level of glucose to 1, 100 and 200 mM. The results, also shown in Tables 4 and 5, show an excellent agreement between our model and the Teusink model with glucose increased to 100 and 200 mM.

Our model still produced results with very low concentrations of glucose (1 mM), while the Teusink model reported an error during simulation when the glucose input was <2 mM. The generic model thus appears to be more robust than the detailed model. We repeated a dynamic experiment, changing the glucose concentration from 50 to 100 mM after 30 min (figures 20c–d), which was again successful. The results obtained from these experiments demonstrate the ability of our generic model to predict new steady-state behaviours that were not used for training. Overall, our results demonstrate that it is possible to predict system behaviour using generic reversible rate equations, without addressing detailed mechanisms at the level of each component.

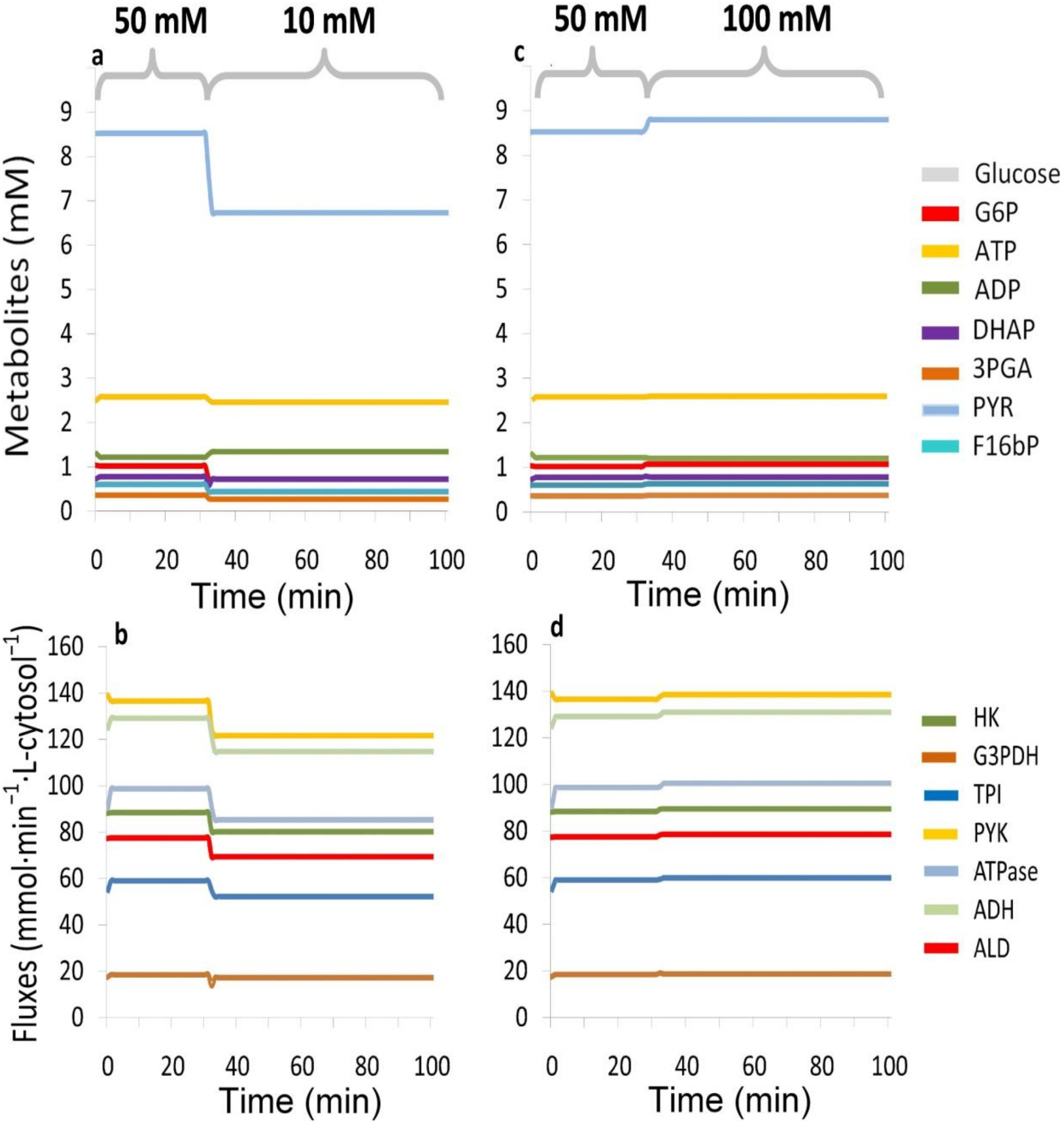


Figure 20: Dynamic experiments using the generic glycolysis model with the level of glucose uptake being changed after 30 min. **(a-b)**. A decreased glucose uptake from 50 mM to 10 mM. **(c-d).** An increased glucose uptake from 50 mM to 100 mM.

	Concentration of glucose (mM)				
Metabolite concentrations (mM)	1	10	50	100	200
Glucose (In)	0.002	0.01	0.1	0.16	0.37
ADP	2.35	1.34	1.29	1.2	1.2
ATP	1.45	2.46	2.51	2.6	2.6
G6P	0.19	0.73	1.03	1.07	1.1
F6P	0.014	0.07	0.11	0.11	0.12
NAD	1.45	1.54	1.55	1.55	1.55
NADH	0.14	0.05	0.04	0.04	0.04
F16bP	0.1	0.44	0.59	0.63	0.64
DHAP	0.032	0.72	0.74	0.78	0.79
GAP	0.026	0.03	0.03	0.03	0.03
BPG	8.01e-06	1.91e-04	3.30e-04	3.88e-04	4.03e-04
3PGA	0.068	0.27	0.36	0.37	0.38
2PGA	0.008	0.03	0.04	0.04	0.04
PEP	0.01	0.05	0.07	0.08	0.08
PYR	1.88	6.73	8.52	8.8	8.92
AcAld	0.067	0.16	0.17	0.18	0.18

	Concentration of glucose (mM)				
Fluxes (mmol·min–^{1}· L-cytosol–^{1})	1	10	50	100	200
Glucose Transport	88.92	88.85	88.15	87.7	89.81
HK	38.75	80.13	88.15	89.48	89.97
Glycogen	4.95	5.96	6.0	6.06	6.06
Trehalose	1.98	2.39	2.4	2.42	2.43
PGI	29.84	69.39	77.35	78.58	79.06
PFK	29.84	69.39	77.35	78.58	79.06
ALD	29.84	69.39	77.35	78.58	79.06
G3PDH	7.6	17.2	18.2	18.67	18.73
TPI	22.24	52.19	59.15	59.91	60.33
GAPDH	52.08	121.59	136.5	138.49	139.39
PGK	52.08	121.59	136.5	138.49	139.39
PGM	52.08	121.59	136.5	138.49	139.39
ENO	52.08	121.59	136.5	138.49	139.39
PYK	52.08	121.59	136.5	138.49	139.39
ATPase	28.64	85.30	99.1	100.44	101.26
PDC	52.08	121.59	136.5	138.49	139.39
ADH	49.04	114.71	129.22	131.03	139.90
Succinate	1.52	3.44	3.64	3.73	3.75

Table 4: Metabolite concentrations (in mM) and fluxes (in mmol·min–^{1}·L-cytosol–^{1}) at steady state for the model generated by GRaPe with glucose uptake levels at 1, 10, 50, 100 and 200 mM. The kinetic parameters were trained on experimental data with glucose levels at 10 and 50 mM. After estimation, the model was simulated in CellDesigner using the SBML ODE Solver.

Metabolite concentrations (mM)	Concentration of glucose (mM)			
	10	50	100	200
Glucose (In)	0.01	0.1	0.1	0.1
ADP	n/a	1.29	n/a	n/a
ATP	n/a	2.51	n/a	n/a
G6P	0.72	1.03	1.09	1.13
F6P	0.07	0.11	0.12	0.13
NAD	1.55	1.55	1.55	1.55
NADH	0.05	0.04	0.04	0.04
F16bP	0.44	0.59	0.63	0.64
DHAP	n/a	n/a	n/a	n/a
GAP	n/a	n/a	n/a	n/a
BPG	2.00e-04	3.30e-04	3.56e-04	3.71e-04
3PGA	0.27	0.36	0.37	0.38
2PGA	0.03	0.04	0.05	0.05
PEP	0.05	0.07	0.08	0.08
PYR	6.72	8.52	8.85	9.03
AcAld	0.16	0.17	0.17	0.17

lluy

Fluxes (mmol·min–1· L-cytosol–1)	Concentration of glucose (mM)			
	10	50	100	200
Glucose Transport	80.16	88.15	88.12	88.1
HK	80.16	88.15	89.25	89.81
Glycogen	6.0	6.0	6.0	6.0
Trehalose	.4	2.4	2.4	2.4
PGI	69.36	77.35	78.45	79.01
PFK	69.36	77.35	78.45	79.01
ALD	69.36	77.35	78.45	79.01
G3PDH	17.24	18.2	18.34	18.41
TPI				
GAPDH	121.48	136.5	138.57	139.62
PGK	121.48	136.5	138.57	139.62
PGM	121.48	136.5	138.57	139.62
ENO	121.48	136.5	138.57	139.62
PYK	121.48	136.5	138.57	139.62
ATPase	85.04	99.1	101.03	102.01
PDC	121.48	136.5	138.57	139.62
ADH	114.59	129.21	131.23	132.25
Succinate	3.45	3.64	3.67	3.68

Table 5: Metabolite concentrations (in mM) and fluxes (in mmol·min^{-1}·L-cytosol^{-1}) at steady state for the Teusink model of glycolysis with glucose uptake levels at 10, 50, 100, and 200 mM. The experimental data for result comparison was obtained using the JWS Online web simulation tool. No simulation with glucose level at 1 mM was obtainable using the Teusink model.

3.7 Discussion and Conclusion

The current availability of high-throughput fluxomic, metabolomic, proteomic and genomic data makes it possible to envisage building integrative genome-scale metabolic models, but convenient tools for assembling such heterogeneous data on a large scale are still lacking. An aim of systems biology is to understand cellular processes as a whole rather than in isolation. Integrating cellular components is essential for our understanding of how interactions between these components influence cellular functions.

Our research ties in with previous investigations indicating that the dynamic behaviour of metabolic systems can be predicted without accurately measuring all rate equations and detailed kinetic parameters. Ao *et al*. (2008) have already used generic rate equations to construct a metabolic model of *Methylobacterium extorquens* AM1. Their results showed that it is possible to attain the dynamical behaviour of a system without the use of extensive and accurately measured rate equations and kinetic parameters. GRaPe follows this principle to enable the building of large models. It generates generic rate equations for all reactions in a metabolic network and thus assumes that the global behaviour of a system should be relatively independent of precise kinetic properties and parameter values. It is worth noting that metabolic systems have long been known to be robust to perturbations and maintain relatively stable intracellular metabolite and flux levels in response to changing external conditions. This property was reflected by the Teusink model, as we have shown in a previous study (Schwartz & Kanehisa, 2006), and it is conserved in our generic model.

What we have demonstrated here re-affirms the hypothesis that generic rate equations can be used to successfully predict the behaviour of biological systems. Due to the unavailability of kinetic parameters – which are time-consuming to accurately measure – it might be worth developing techniques to understand the behaviour of a system in an attempt to understand the interplay of biological systems and also build successful models with very good predictive capabilities.

3.8 References

1. Ao P, *et al*. (2008). Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AM1 growth as validation. *Chin. J. Biotechnology*. 24:980-994

2. Bakker, B. M. *et al.* (1997) Glycolysis in Bloodstream Form *Trypanosoma brucei*. Can Be Understood in Terms of the Kinetics of the Glycolytic Enzymes. *The Journal of Biological Chemistry*, 272, 3207-3215.

3. Dobson, P.D., Smallbone, K., Jameson, D., Simeonidis, E., Lanthaler, K., Pir, P., Lu, C., Swainston, N., Dunn, B.W., Fisher, P., Hull, D., Brown, M., Oshota, O., Stanford, J. N., Kell, B. D., King, D.R., Oliver, G.S., Stevens, D.R. and Mendes, P. (2010). Further developments towards a genome-scale metabolic model of yeast. *BMC Systems Biology,* 4:145.

4. Förster, J. Famili, I., Fu, P., Palsson, B. and Nielsen, J. (2003). Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*. 13: pp. 244-253.

5. Gutenkunst,R.N., Waterfall,J.J. *et al*. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol* 3(10), 189.

6. Hynne, F., Dano,S. *et al*. (2001) Full scale model of glycolysis in Saccharomyces cerevisiae. *Biophysical Chemistry,* 94(1-2), 121-63.

7. Jamshidi, N. and Palsson, B. Ø. (2008) Formulating genome-scale kinetic models in the post-genome era. *Molecular Systems Biology*, 4, 171.

8. Jamshidi,N. and Palsson, B. Ø. (2009) Using *in silico* models to simulate dual perturbation experiments: procedure development and interpretation of outcomes. *BMC Systems Biology*, 3, 44.

9. Klipp, E. (2007). Modelling dynamic processes in Yeast. *Yeast*, 24. 943-959.

10. Lambeth,M.J. and Kushmerick,M.J. (2002) A Computational Model for Glycogenolysis in Skeletal Muscle. Annals of Biomedical Engineering, 30, 808-27.

11. Olivier B.G. and Snoep J. L. (2004). Web-based kinetic modelling using JWS Online. *Bioinformatics*. 20. 2143-2144.

12. Palsson, B. Ø. (2000). The Challenges of *in silico* biology. *Nature Biotechnology*, 18. 1147-1150.

13. Pritchard, L. and Kell, D. (2002). Schemes of Flux Control in a Model of *Saccharomyces cerevisiae* Glycolysis, *European Journal of Biochemistry*, 269, 3894-3904.

14. Schlitt, T. and Brazma, A. (2007). Current Approaches to gene regulatory network modelling. *BMC Bioinformatics*. 8(Suppl 6):S9.

15. Schwartz,J.M., Kanehisa, M. (2006), Quantitative elementary mode analysis of metabolic pathways: the example of yeast glycolysis, *BMC Bioinformatics*, 7, 186.

16. Teusink,B., Passarge, J. *et al*. (2000) Can yeast glycolysis be understood in terms of in vitro kinetics of the constituent enzymes? Testing biochemistry. *European Journal of Biochemistry*, 267(17): 5313-5329.

Chapter 4

Using Generic Equations to Replicate Steady States and Predict New States in a Genome-scale Model

The following work appears in “Construction of a genome-scale kinetic model of *Mycobacterium tuberculosis* using generic rate equations to replicate growth conditions.” Adiamah, DA and Schwartz, J-M. (2011) *To be submitted.*

The aim of this study is to show that our modelling approach is capable of replicating biological behaviour on a large-scale. As flux data was unavailable for the *Mycobacterium tuberculosis* model, we use FBA in computing a flux distribution for parameter estimation. Additionally, we also analysed redundancy in our estimated kinetic parameters and established the computational cost of performing parameter estimation on such a large-scale model.

4.1 Abstract

Genome sequencing and annotation has made it possible to construct genome-scale metabolic networks. These genome-scale models allow for the integration of different data types that can be analysed mathematically. As a result, studying biological systems at the genome-scale level has the potential to increase our knowledge and understanding of fundamental functions and essential biological properties. These models are mostly analysed using constraint-based methods as detailed rate equations and kinetic parameters are unavailable for most genome-scale models. However, constraint-based analysis is limited in capturing the dynamics of cellular processes. This has made it important to build kinetic models to understand the detailed dynamics of cellular functions and their regulation.

In this paper, we present, to our knowledge, the first attempt to build a genome-scale kinetic model of *Mycobacterium tuberculosis* metabolism using generic rate equations and convenience kinetics. *M. tuberculosis* causes tuberculosis which remains one of the largest killer infectious diseases. As such, there is a need to investigate new methods and techniques to identify drug targets and further understand the biology of *M. tuberculosis.* Using a genetic algorithm, we estimated kinetic parameters for a genome-scale model of *M. tuberculosis* based on flux distributions. Our results show a near perfect agreement with flux values obtained under different growth conditions. We also show results of our Parameter Variability Analysis which indicates a degree of redundancy in parameters of our model.

4.2 Introduction

Despite numerous efforts which have led to the production and availability of effective Bacille-Calmette-Guerin (BCG) vaccine and chemotherapy, tuberculosis (TB) still remains one of the largest killer infectious diseases (Chopra *et al.,* 2003*;* Raman *et al.,* 2008). Although significant advances were achieved in understanding the biology of *Mycobacterium tuberculosis,* including functional genomics tools such as proteomics and microarray analyses combined with modern approaches, surprisingly no new drug to treat tuberculosis has been developed in the last 30 years (Chopra, 2003). As a result, there is a need to investigate new methods and techniques to identify drug targets and further understand the biology of *M. tuberculosis.*

Genome-scale metabolic models are essential in bridging the gap between the metabolic phenotypes and genome-derived biochemical information by providing a platform for the interpretation of experimental data related to metabolic states and enabling *in silico* experimentation of cell metabolism (Durot *et al.,* 2009). The annotation and sequencing of genomes has made it possible to reconstruct genome-scale metabolic networks (Price *et al.,* 2003). Using constraint-based models and *in silico* simulation, we can define the phenotypic functions of these genome-scale metabolic networks. Furthermore, current advances in high-throughput experimental technologies and computational systems biology have enabled metabolic models to be reconstructed for an increasing number of species. Using these computational methods to explore bacteria metabolic models has increased our understanding of bacterial evolution and metabolism (Durot, 2009). Additionally, genome-scale models allow for the assembly of various data types which can be analysed mathematically. The integration of regulation with metabolic networks has been used successfully in analysing phenotypes from gene-deletion studies and phenotypic arrays (Borodina and Nielsen, 2005). Other types of data, such as metabolomics and proteomics, have also been integrated with constraint-based methods (Yizhak *et al.,* 2010).

The availability of genome-scale models has accelerated the development of methods to analyse system-wide metabolic behaviours. Systems biology aims at predicting cellular behaviours *in silico* by examining the dynamics and properties of cellular processes (Kitano, 2002). As a result, it is necessary to go beyond static constraint-based models and build kinetic models where perturbation of a system is possible, in order to understand the detailed dynamics of cellular functions and their regulation (Adiamah *et al.,* 2010). However, it is time-consuming and costly to experimentally measure all metabolite concentrations, reaction fluxes and kinetic parameters at the genome scale. This has led to recent efforts in providing methods to build kinetic models using other approaches, such as linlog kinetics (Smallbone *et al.*, 2007), generic equations (Ao *et al.,* 2008; Adiamah *et al.,* 2010), parameter balancing (Lubtitz *et al*., 2010) and convenience kinetics (Liebermeister and Klipp, 2006).

Stoichiometric models only provide limited insight into the functioning of cellular processes as they only describe the topological and steady-state properties of a metabolic network. These models are mostly analysed using constraint-based analysis (Covert *et al.,* 2003). Constraint-based modelling uses energy balance, flux limitations, mass balance and thermodynamics in an attempt to describe the behaviour of an organism (Smallbone *et al.*, 2010). However, constraint-based modelling fails in capturing the dynamics of cellular behaviour and is unable to provide insight into the changes in the concentrations of metabolites and enzymes.

The lack of complete biological knowledge of *M. tuberculosis* makes it difficult to build a detailed kinetic model capable of *in silico* perturbation and analysis. In a previous paper, we presented a method for streamlining the construction of large-scale dynamic models using generic kinetic equations based on the stoichiometry of the reactions (Adiamah *et al.,* 2010). We modelled the yeast glycolysis pathway to test our methodology; our results showed that using generic kinetic equations, the behaviour of the system could accurately be described. However, our approach was limited to reactions with up to two substrates and two products; when the number of substrates and products were more than two for a particular reaction, we required that reaction to be broken down. For a small network such as the yeast glycolysis pathway, reactions of more than two substrates and products are few and breaking down these reactions is possible. However, on the genome-scale where the number of reactions can exceed 1500, manually verifying and breaking down of reactions into their chemical or biological constituents can be time-consuming and tedious. As a result, convenience kinetics (Liebermeister and

Klipp, 2006), which have proved successful in capturing the dynamical behaviour of metabolic reactions without placing a limit on the number of substrates and products, was introduced in our model-building process to streamline the construction of large-scale kinetic models.

Systems biology often uses reverse engineering in an attempt to reconstruct biological interactions from experimental or measured data for a particular organism when parameters are unknown (Banga, 2008). Here, experimentally-measured data are used to constrain kinetic parameter values and other constants required in characterising a metabolic model. It is usually unlikely to have a comprehensive dataset comprising metabolic, genomic and proteomic data needed to constrain kinetic parameter values and as such, simulated or calculated data may be used as a substitute. Flux Balance Analysis (FBA), which enables the calculation of an optimal flux distribution using linear programming, has proved successful in representing different metabolic phenotypes under various experimental conditions with successful prediction rate found to be approximately 60 and 86% for *H. pylori* and *E. coli* respectively in gene deletion studies (Price *et al*., 2003). As kinetic parameters are not required in FBA, calculating fluxes for a model is relatively easy and straightforward when the structure of the metabolic network and flux constraints are known. Our parameter estimation approach uses flux data, together with metabolic concentration data, to constrain kinetic parameter values needed to define our models. When input flux data is omitted from the parameter estimation process a zero flux for a reaction may be obtained even in non-equilibrium conditions. To avoid this caveat, when there is no flux data available for parameter estimation under a particular experiment, flux data calculated using FBA can be used as an input in constraining kinetic parameters for that metabolic model.

Optimisation techniques can be used to estimate kinetic parameters based on simulated or experimental data (Mendes and Kell, 1998; Kell, 2006; Adiamah *et al.*, 2010). However, these estimated parameter values are usually not unique given a set of an input data due to mathematical redundancy (Chou and Voit, 2009). This redundancy means that multiple sets of parameter values can fit to an experimental data series equally well. There have been attempts in the past to reduce redundancy in parameter estimation. One noticeable approach is the use of Dynamic Flux Estimation (DFE) proposed by Goel *et al*, (2008) where there is a verification of mass conservation within metabolic time-series data and fluxes are expressed as functions of the relative variables affecting them. Although results from DFE show that redundancy can be reduced, the approach is computationally very expensive and time-consuming due to the internal verification process.

Another method proposed to constrain parameter estimation and reduce redundancy in systems biology was presented by Lubitz *et al* (2010). Here, the authors used a technique known as 'parameter balancing', which is based on Bayesian parameter estimation, to explore the thermodynamic dependencies that exist between biological quantities in order to estimate kinetic parameters. Although their results, which were validated on the phosphofructokinase reaction, are encouraging, on a large-scale network it might be an impossible task to obtain various experimental data for all reactions. Furthermore, flux data is again omitted from the input data set which means that reaction fluxes may be estimated to zero in a non-equilibrium setting. The model building approach presented in Adiamah *et al*. (2010) showed that estimating kinetic parameters using metabolic and flux data can successfully reproduce experimental conditions under both steady- and dynamical states. However, the level of redundancy in our model remained to be determined. Reducing redundancy in models can result in producing more robust models which are able to reproduce experimental conditions *in silico*. As a result, there is also a need to introduce methods to test the reliability of these estimated values.

Taking on board the current issues in building large-scale integrative models, obtaining kinetic parameters and measuring redundancy, we here present, to our knowledge, a first attempt to build a genome-scale kinetic model of *M. tuberculosis* metabolism based on a stoichiometric model by Beste *et al*. (2007) using generic rate equations and convenience kinetics. We show that kinetic model simulations are in good agreement with flux values predicted by FBA under different growth conditions on a large-scale. We also determine the degree of redundancy in our parameter set estimated by our genetic algorithm. The results from our analysis suggest a high degree of redundancy in parameter values when fluxes are the only constraining input for estimation.

4.3 Methods

4.3.1 Enzyme kinetics and rate equations

We used GRaPe (Adiamah *et al.,* 2010) to build our genome-scale kinetic model of *M. tuberculosis.* Rate equations for all reactions in a model are automatically generated by GRaPe based on the stoichiometry of the reaction. Reactions in the model assume a random-order mechanism as the sequential order of binding and releasing of substrates is unknown. A key advantage of GRaPe over other software tools is its ability to automatically generate rate equations for reactions. This makes it less error-prone and more time-efficient in building large-scale models. The King& Altman method (King and Altman, 1956) is used by GRaPe to derive rate equations based on the stoichiometry of a reaction and the enzyme mechanism. Adiamah *et al.* (2010) provides details of the generic Michaelis–Menten rate equations used by GRaPe for the different reaction types.

Generic rate equations were used for all reactions of up to two substrates or products; these reactions can be of type uni–uni, uni–bi, bi–uni or bi–bi. For reactions of more than two substrates of products, the convenience kinetics was used. Convenience kinetics can be used in translating a metabolic network into a dynamical model capable of predicting biological properties (Liebermeister and Klipp, 2006). The equation, a generalised form of Michaelis-Menten kinetics, follows a random-order mechanism and implements enzyme saturation and regulation. Convenience kinetics is able to cover all possible reaction stoichiometries.

For a reaction of type $A_1 + A_2 + ... \leftarrow\rightarrow B_1 + B_2 + ...$, the concentrations of substrates are represented by a vector $a = (a_1, a_2,...)$ and the concentrations of products are represented by a vector $b = (b_1, b_2, ...)$. The flux, v (a, b), is defined using convenience kinetics as:

(38)

where i is the number of substrates, j is the number of products, K_{Ai} represents the kinetic parameter (substrate constants) of the i^{th} substrate, K_{Bj} is the j^{th} product of the reaction (both K_As and K_Bs are measured in mM), e_0 is the concentration of enzyme, V^+ is the substrate turnover rate, V^- is the product turnover rate and v is the flux of the reaction.

4.3.2 Parameter Estimation

Kinetic models have been shown to produce accurate and testable results (Jamshidi and Palsson, 2008). However, due to the enormous number of kinetic parameters needed to define the system, the number of large-scale kinetic models still remains relatively low. Furthermore, it was observed by Teusink *et al.* (2000) that *in vitro* measurements of kinetic constants may not necessarily be representative of their numerical values *in vivo* .Currently, there are various software tools capable of performing parameter estimation: COPASI (Hoops *et al.*, 2006) provides a list of methods for estimation including a genetic algorithm; SBML-PET (Zi and Klipp, 2006) uses a stochastic ranking evolution strategy method to estimate parameters. However, flux constraints are excluded which can allow for a zero flux solution to be obtained even in non-equilibrium conditions. Fluxes are not explicitly expressed as model elements, as a result constraining parameters using those software is still not straightforward. DFE shows that by verifying mass conservation in metabolic time-series data and integrating fluxes in the estimation of kinetic parameters values, the redundancy in models can be reduced (Goel *et al.*, 2008). GRaPe uses a genetic algorithm to estimate all kinetic parameters using flux values to constrain kinetic parameters in our genome-scale model of *M. tuberculosis*. Figure 1 illustrates the process undertaken to reconstruct our kinetic model of *Mycobacterium tuberculosis*. Thermodynamics can be used to constrain the parameter estimation method in a process known as thermodynamic-kinetic modelling (Ederer and Gilles, 2007). Other data sets can also be introduced into the parameter estimation process for constraining purposes. However, the availability of heterogeneous data for parameter estimation on a large-scale is lacking.

4.3.3 Parameter Variability Analysis (PVA)

One of the issues relating to parameter estimation is that of mathematical redundancy. The redundancy results in multiple sets of parameter values that can fit equally to an experimental data set. A simple example of redundancy is when two parameters, say *a* and *b*, are part of an equation in the form of say, *a*+*b* or *a*b*, but if only their sum or product is known it is impossible to identify the value of *a* and *b* individually; if both the sum and product are known, then the value of *a* and *b* can be calculated. This example illustrates that the level of redundancy is dependent on the amount of experimental data used to constrain the estimation. When there is redundancy, the parameter values found in several runs of the estimation algorithm are likely to be different. In this article, we analyse redundancy or ‘sloppiness’ in parameter estimation using Parameter Variability Analysis (PVA). PVA allows us to measure the degree of change in a set of parameter values when estimation is repeated several times.

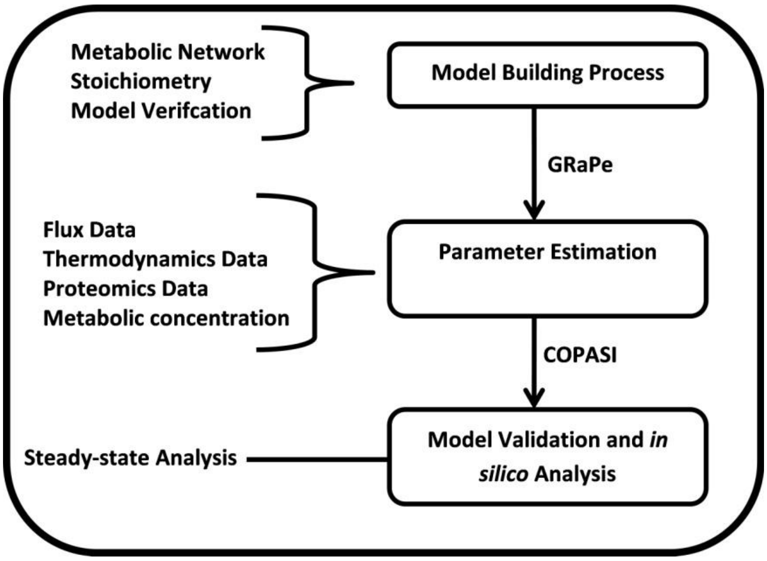


Figure 21: Model construction process and validation. Bracketed text in grey represents data that can be used to constrain kinetic parameters but were not used in this study.

Once a model has been constructed or uploaded in GRaPe, PVA can be performed using the same time-series data required to estimate parameter values for the model. The PVA algorithm works by estimating kinetic parameters for the model using a genetic algorithm (GA) for a number of iterations. GA works by populating a set of random initial parameter values; this is why results may differ after each run of the algorithm when there is redundancy. These values are then optimised in an iterative manner until the maximum number of iterations is reached or a suitable solution is found. In GRaPe, GA uses flux and metabolic data to constrain parameters as illustrated in Figure 21. After each run of estimation, the objective function, which is a measure of the fit of the estimation to the original time-series data, and kinetic parameter values are stored in a data file in a tabbed-delimited format. The results of PVA can then be exported to spreadsheet software and statistically analysed. The PVA function is now fully integrated into our GRaPe software tool.

4.4 Results

4.4.1 The genome-scale kinetic model of *Mycobacterium tuberculosis*

In this section, we present a genome-scale kinetic model of *M. tuberculosis* and provide a comparative analysis of our results with those from Beste *et al.* (2007). The model by Beste was experimentally developed with the accurate measurement of steady-state growth parameters in a continuous culture. The substrate consumption rates were calculated by Beste using Flux Balance Analysis (FBA). Their simulated results showed a close similarity with values determined experimentally. We aim at demonstrating that using generic kinetic equations, we can reproduce different steady states and achieve results of the same quality to those produced by FBA.

We built a genome-scale kinetic model of *M. tuberculosis* using generic rate equations to demonstrate that different experimental conditions could be replicated *in silico* without accurately measuring enzyme concentrations and rate parameters. *M. tuberculosis* is a pathogenic bacterium which causes TB (Bordbar *et al.,* 2010). With TB being one of the major causes of death in the third world, there is still much to be learned about the metabolic and regulatory networks of this bacterium (Chandrasekaran and Price, 2010).

There are now numerous genome-scale reconstruction of *M. tuberculosis* (Beste *et al.,* 2007; Jamshidi and Palsson, 2007; Bordbar *et al,* 2010), which can serve as a basis to construct an integrative genome-scale kinetic model. In Beste *et al* (2007), the authors constructed a genome-scale metabolic network of *M. tuberculosis* using a reconstruction of *Streptomyces coelicolor* as a starting point. Genes were mapped between the two species using gene orthology clusters from the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa *et al.,* 2010). The corresponding metabolic reactions were then transferred to the TB network. Overall, 57% (487) of the unique reactions in the final model were derived using the KEGG orthology gene mapping. Using the KEGG and BioCyc databases, the authors further supplemented the initial model. However, the model could not be constructed using automatic (or semi-automatic) methods alone; the analysis of relevant research articles had to be carried out to identify genes, metabolites and reactions to complete the genome-scale reconstruction. The final metabolic network of *M. tuberculosis* includes reactions needed for the synthesis of the cell membrane, complex lipids and carbohydrates, which are important for both growth and pathogenesis. Additionally, the model includes fatty acid metabolism in *M. tuberculosis* which is thought to be a crucial factor in the pathogenesis of TB. Other metabolic pathways such as respiratory pathways and synthesis of biomolecules, which are specific to mycobacteria, were also modelled manually. Iron metabolism and other transport reactions, including reactions which are responsible for the importing of carbon, nitrogen, minerals and compounds of high molecular weight were also manually added to the model. The final stoichiometric model consisted of 739 metabolites and 849 reactions and included 726 genes. The calibration of their model was done by growing *Mycobacterium bovis* bacilli Calmette Guérin in a continuous culture and parameters for steady-state growth were also measured. FBA was used to calculate substrate consumption rates. Their results showed a close agreement with experimentally determined values. The model was made available as a web-based interactive tool.

Using GRaPe (Adiamah *et al*, 2010), we created a genome-scale kinetic model of *M. tuberculosis* based on the stoichiometric model developed by Beste *et al* (2007). GRaPe assigns an enzyme species to each reaction, which is then mapped to the corresponding gene(s) provided in the model. A major difficulty in building genome-scale kinetic models is the lack of quantitative data available to fully define the model (Jamshidi and Palsson, 2008). The Beste model was a stoichiometric model which did not include any kinetic data. As a result, we set the initial concentration of metabolites and enzyme species to 1 by default. The reactions in our model were assumed to follow a random-order mechanism. We then used GRaPe to generate generic rate equations for all the reactions in the *M. tuberculosis* genome-scale model. The type of rate equation generated for each reaction is based on the stoichiometry of the reaction. The resulting genome-scale model of *M. tuberculosis* contains 739 metabolites, 856 metabolic reactions and 856 enzymes. The model is available in SBML format as Supplementary Online Data.

4.4.2 Parameter Estimation

We obtained flux values for three steady-states with glycerol being the only carbon source using the interactive web-based tool by Beste *et al* (2007). The tool uses FBA in calculating flux distributions for the three steady-states with glycerol consumption at 0, 0.5 and 1 mmol/g dry weight (DW) respectively. Flux distribution data obtained under each experimental condition was then used as an input data source to estimate the parameters of our kinetic model. Since there were no available proteomics and gene expression data for the amount of concentration for both metabolites and enzymes in this study, a default value of 1 was assigned to each metabolite and enzyme species in the model. We also limited the precision of values in each dataset to two decimal places for faster parameter estimation process.

We performed three separate parameter estimations using our genome-scale model of *M. tuberculosis* under the different glycerol consumption rates. The kinetic parameters for each reaction in the model were estimated using GRaPe's genetic algorithm. Model 1, with glycerol consumption rate at 0 mmol/gDW, had 2297 kinetic parameters after parameter estimation. Model 2, with glycerol consumption rate at 0.5 mmol/gDW, had 2537 parameters and Model 3 had 2931 parameters after parameter estimation with glycerol consumption rate at 1 mmol/gDW. The difference in the number of parameters after estimation was due to different numbers of reactions having a zero flux in each case. Furthermore, reactions with negative fluxes had their substrates and products swapped around to prevent having negative kinetic parameter values.

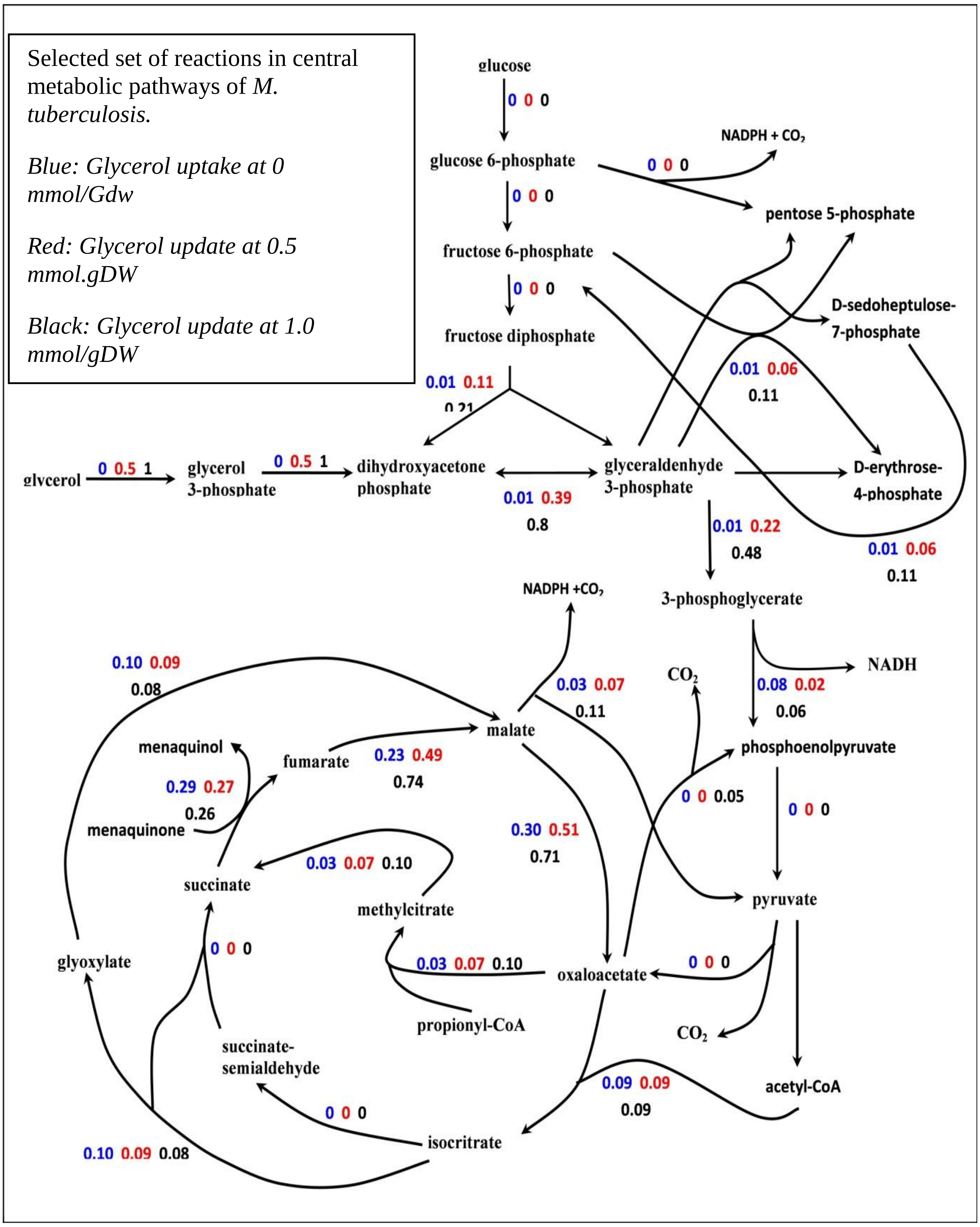


Figure 22: Response of *Mycobacterium tuberculosis* to glycerol uptake rate at 0, 0.5 and 1.0 mmol/gDW and glucose consumption level at 0.003 mmol/gDW. The network shows a selected set of reactions in the central metabolic pathways of *M. tuberculosis.* Reactions in the pathway are represented using arrows and the direction of the flux is indicated by the direction of the arrow (direction of an arrow does not represent reaction reversibility). The flux values are the numbers next to the arrows with blue, red and black colours indicating a glycerol uptake rate of 0, 0.5 and 1.0 mmol/gDW respectively. Flux values were obtained by performing steady-state analysis using COPASI after obtaining kinetic parameters for the M. *tuberculosis* model using GRaPe . $CO_{2,}$ Carbon

dioxide; CoA, coenzyme A; NADH, nicotinamide adenine dinucleotide; NADPH, nicotinamide adenine dinucleotide phosphate..

4.4.3 Model Validation

We performed steady-state analysis for Model 1, 2 and 3 using COPASI. The results were then compared with the FBA flux distribution obtained from the Beste model under the same experimental conditions. Our verification analysis showed a near-perfect agreement between the results from our models and the respective Beste model. Figure 22 shows a part of the central metabolic pathways of Model 2 with glucose uptake level at 0.5 mmol/gDW. The complete flux distribution for Model 1, 2 and 3 are supplied in Supplementary Data 1. Our model demonstrates the ability to accurately reproduce steady-state flux distributions at the genome-scale.

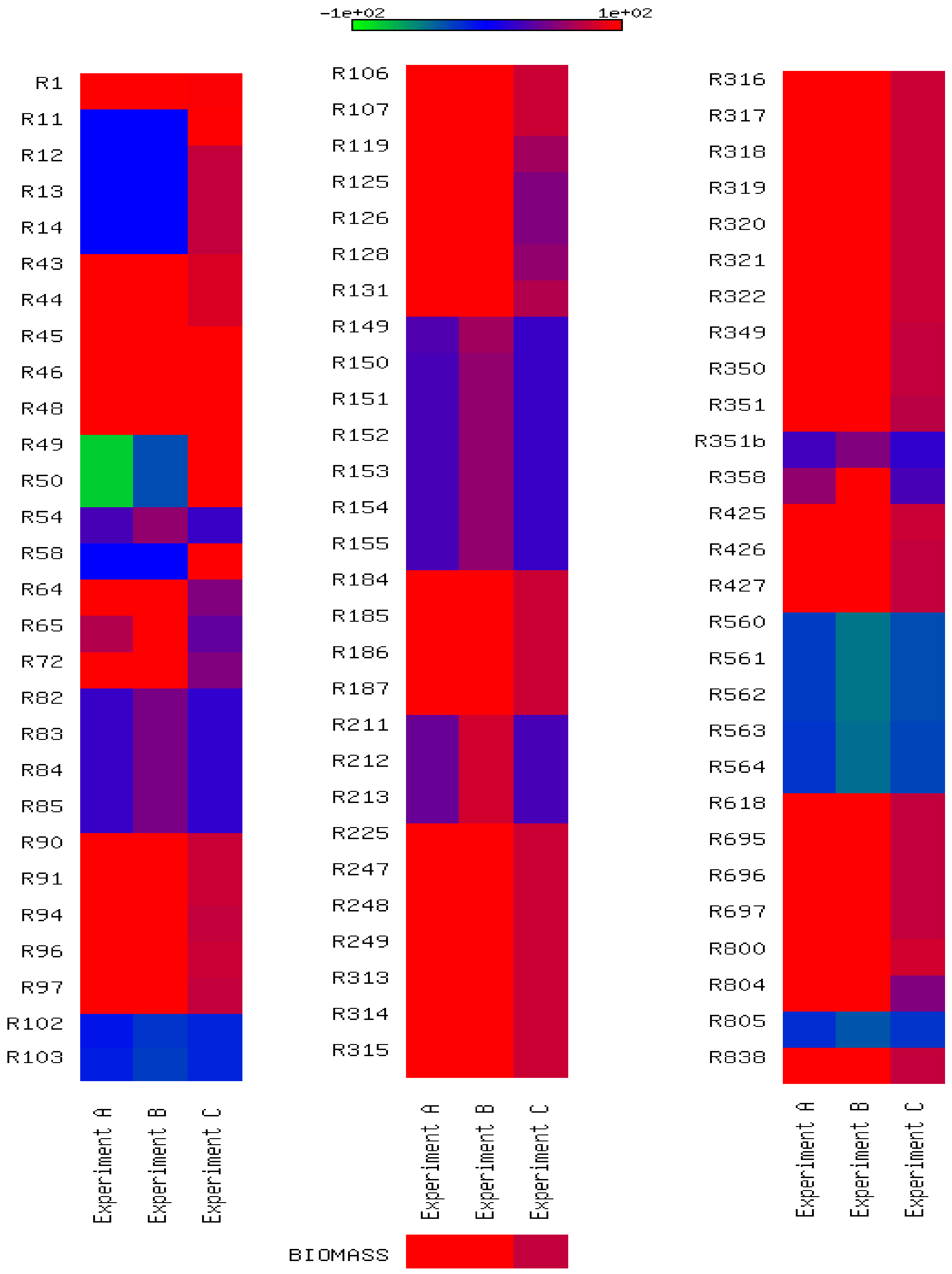


Figure 23: Percentage changes in fluxes with respect to changes in glycerol consumption rate. We compared the response of selected reactions when glycerol uptake rate is at 0 and 0.5 mmol/gDW in *Experiment A*, 0 and 1 mmol/gDW in *Experiment B*, and 0.5 and 1 mmol/gDW in *Experiment C*. Green indicates that the flux of that reaction decreased by 100% or more. Blue indicates that no change was observed. Red indicates an increase in flux of 100% or more. Heat maps were created using matrix2png (Pavlidis and Noble, 2003). Reactions identifiers are the same as in the Beste model.

We also performed an analysis to identify the reactions which showed the greatest change in flux with respect to change in glycerol consumption rate. We determined the relative change in fluxes between glycerol consumption rates at 0 and 0.5 mmol/gDW in Experiment A, 0 and 1 mmol/gDW in Experiment B, and 0.5 and 1 mmol/gDW in Experiment C. Reactions with significant changes in fluxes are represented as a heat map in Figure 23. The most significant changes were observed in the enolase (R49) and pyruvate kinase (R50) reactions where a 100% increase in flux is observed when glycerol consumption is increased from 0 to 0.5 mmol/gDW and from 0.5 to 1 mmol/gDW.

4.4.4 Parameter Variability Analysis (PVA)

To determine the degree of redundancy in the values of our estimated kinetic parameters in the genome-scale metabolic model of *Mycobacterium tuberculosis,* we performed PVA by running our estimation algorithm 100 times. It is well known that different sets of parameters values can fit to an experimental time-series data resulting in mathematical redundancy (Chou and Voit, 2009). This means that running parameter estimation 100 times can produce 100 different sets of parameter values which are able to fit equally well the input data set.

In order to make it easier to interpret the result of PVA, the result of the PVA was split into five different categories based on the stoichiometry of the reaction (uni-uni, uni-bi, bi-uni, bi-bi andConvenience kinetics). The results from our PVA show that many of the parameters are poorly constrained as illustrated in Figure 24. The detailed average value and standard deviation graphs for each parameter under the different reaction types are shown in on our lab group page.

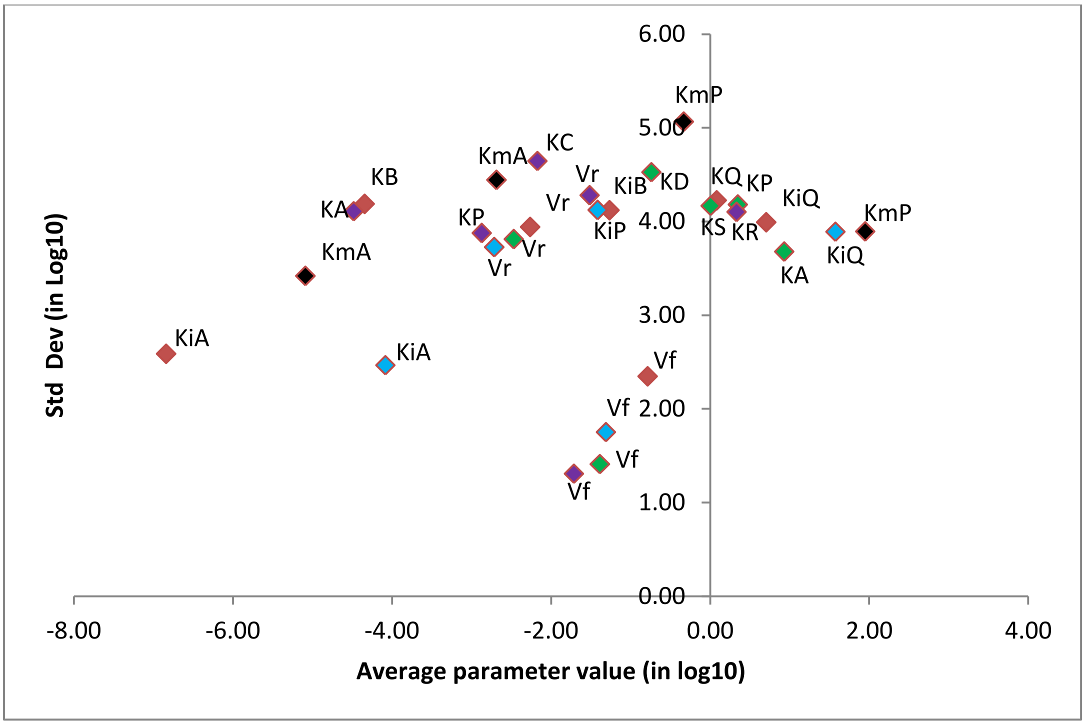


Figure 24: Results of Parameter Variability Analysis (PVA). PVA was performed by running our genetic algorithm 100 times. The results obtained were then subdivided into five reaction types (uni-uni (*in black*), uni-bi (*in red*), bi-uni (*in blue*), bi-bi (*in purple)* and CK (convenience kinetics – *in green*). The graph shows the average parameter value for all 28 parameters under the five different reaction types plotted against their standard deviation (labelled Std Dev above) over the 100 runs. Both axes of the graph are in logarithmic scale.

Our results show that overall *vf*, the velocity of the forward reaction, is the most constrained parameter having the smallest standard deviation (Table 6).

Most Constrained Parameters			
Parameter	**Reaction Type**	**Average**	**Std Dev**
vf	CK	-0.78	2.35
vf	Uni-Bi	-1.39	1.41
vf	Bi-Uni	-1.71	1.31
vf	Bi-Bi	-1.31	1.75

Table 6: Average parameter value and standard deviation (Std Dev) in logarithmic scale over 100 iterations for the most constrained parameters as observed by PVA. Reactions are of type: uni-uni, uni-bi, bi-uni, bi-bi and convenience kinetics (CK) for reactions of more than two substrates or products.

The high degree of redundancy and poor constraining in the parameter values as indicated by the PVA comes in support of our underlying assumption that accurate rate equations and kinetic parameters are not necessarily crucial in constraining the behaviour of biological system. Nevertheless, the integration of genomic and proteomic data, together with metabolic and flux data, is expected to reduce mathematical redundancy as shown in previous studies (Price *et al.*, 2003; Chou and Voit, 2008).

Computation of 100 sets of parameters for each reaction in Model 2 (with 739 metabolites, 856 metabolic reactions, 856 enzymes and 2537 kinetic parameters) for PVA took over 5 hours and 40 minutes. The relatively fast computing time was a result of reducing the objective function to 1.0E-04 and limiting the data points to three decimal places in the input dataset for parameter estimation. The objective function is the summed squared mean distance measured between the simulated data and input data. Reducing the objective function increased computing time but improves the quality of parameter fit to input data. We performed an experiment to determine the relationship between changes in objective function and time taken to compute PVA for one reaction with 2 substrates, 2 products, 1 enzyme and 6 kinetic parameters (Figure 25).

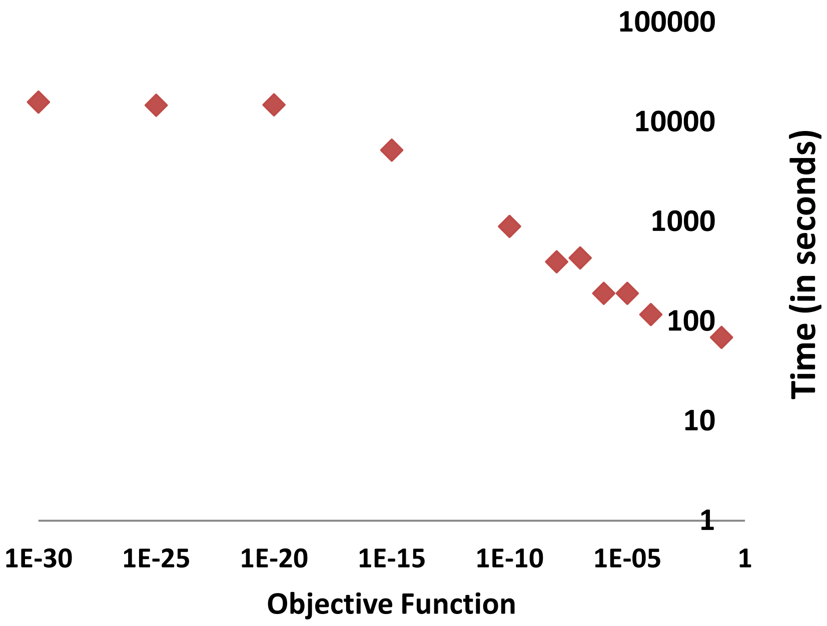


Figure 25: Computing times of PVA against changes in fitness function. PVA is performed for a reaction with 2 substrates, 2 products, 1 enzyme concentration and 6 kinetic parameters. For each PVA run, the expected error distance measured between the input data and simulated data, known as the objective function, is set and the time taken to compute PVA results (running genetic algorithm 100 times) is recorded. The results indicate a linear relation between the objective function and the computing time until the limits of computational precision are reached. Both axes of the graph are in logarithmic scale.

The results of this experiment indicate that the computing time for parameter estimation increases significantly when the objective function is reduced to 1.0e-10 and beyond. The relationship that is observed to exist between the objective function and computing time seems to be linear in nature (PVA was computed on a desktop computer with a quad CPU having 3.00GHz, 2.99GHz processor speed and 4GB of

RAM). The objective function for the PVA for the genome-scale *M. tuberculosis* model was reduced to 1.0E4 for faster computing time.

Another variable that can increase computing time in parameter estimation is the number of data points in the experimental dataset. To examine how the number of data points can influence computing time, an experiment was performed to determine if a relationship did exist. Using PVA, we performed parameter estimation for a single reaction with two substrates, two products and an exzyme and 6 kinetic parameters.

The result of this experiment indicates that the number of data points in the input dataset for parameter estimation increases the computing time in a non-linear manner as shown in figure 26. This explains why a relatively fast time of 5 hours and 40 minutes was recorded when PVA was performed for such a large model with 2537 kinetic parameters as the number of data points was restricted to only three.

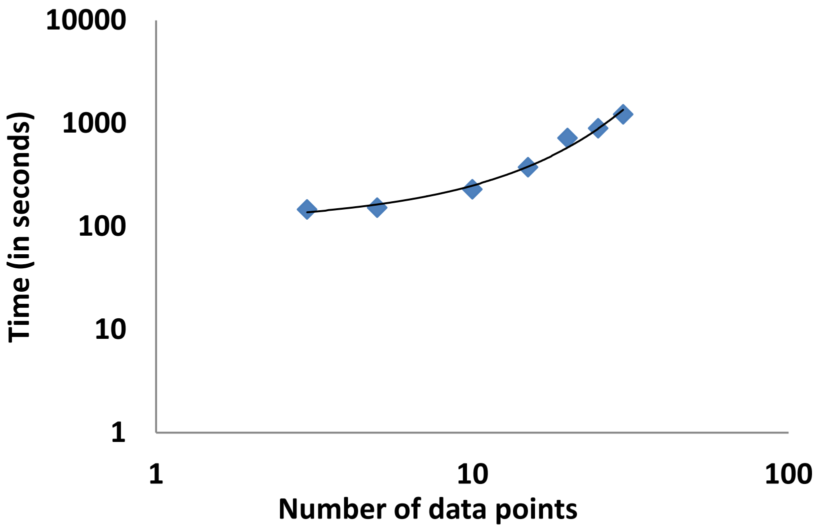


Figure 26. Scatter graph showing the relationship between data points and computing time. PVA was performed for a single reaction of two substrates and two products. There were 6 kinetic parameters needed to define this reaction. PVA was repeated 6 times and for each iteration, the number of data points in the input dataset

for parameter estimation was increased from 3 to 30. The results show a rising curve in a non-linear shape.

4.4.5 Validation on Model Integrity

The key to building predictive models for organisms relies on the construction of sound dynamic biochemical networks (van Riel, 2006). These models can be useful in providing scientific explanation of biological systems in both disease and health and can aide in the discovery of new drug targets. We tested the integrity of our *M. tuberculosis* model and kinetic parameters by trying to predict new states without estimating kinetic parameters again. We took the fluxes at steady-state for three experiments of *M. tuberculosis* with glycerol at 0, 0.5 and 1 mmol/gDW. The flux distribution vector under each experiment was replicated ten times to create a data file with ten data points. , We then merged the three individual files, each with ten data points, into one to create a data set with 30 data points which was then used as an input file for parameter estimation. Performing parameter estimation with this dataset took 125,022 seconds to complete when the objective function was at 1.0E-4. The range of objective function observed for each reaction during parameter estimation was between 1.0e-8 and 20. After parameter estimation, three steady-state analyses of the model were performed with glycerol uptake at 0, 0.5 and 1 mmol/gDW using COPASI.

	Glycerol uptake 0 mmol/gDW		Glycerol uptake 0.5 mmol/gDW		Glycerol uptake 1.0 mmol/gDW	
ID	Exp Data	Predicted Data	Exp Data	Predicted Data	Exp Data	Predicted Data
R1	0.003924	0.503662	0.50366	0.503662	1.003403	0.503662
R7	0.001	1.00E-03	0.001	1.00E-03	0.001	1.00E-03
R8	0.001	0.001	0.001	0.001	0.001	0.001
R9	0.001	0.00100049	0.001	0.00100049	0.001	0.00100049
R13	4.79E-04	0.001962	0.002137	0.001962	0.003749	0.001962
R14	2.97E-04	0.001219	0.001328	0.001219	0.002329	0.001219
R28	0.002901	0.002762	0.002557	0.002762	0.002828	0.002762
R29	0.002901	0.00276215	0.002557	0.00276215	0.002828	0.00276215
R32	0.001267	0.0021828	0.002191	0.0021828	0.00309	0.0021828
R33	0.003	0.003	0.003	0.003	0.003	0.003
R45	-0.007133	0.00713351	0.393944	0.00713351	0.796665	0.00713351
R48	-0.032447	0.032447	0.221728	0.032447	0.479959	0.032447
R49	-0.077595	0.0775879	-0.015883	0.0775879	0.055124	0.0775879
R50	-0.077595	0.0775909	-0.015883	0.0775909	0.055124	0.0775909

R207	0	6.03E-04	0	6.03E-04	0.001809	6.03E-04
R208	0	0.001926	0.001276	0.001926	0.004502	0.001926
R209	0	0.00184166	0.001732	0.00184166	0.003793	0.00184166
R210	0	0.00184168	0.001732	0.00184168	0.003793	0.00184168
R211	0.05515	0.0782287	0.077935	0.0782287	0.101601	0.0782287
R554	0.002845	0.00231102	0.002306	0.00231102	0.001782	0.00231102
R809	0.001	0.00100021	0.001	0.00100021	0.001	0.00100021
R811	0	6.67E-04	0.001	6.67E-04	0.001	6.67E-04
R812	0	0	0.5	0.5	1	1
R814	0	3.33E-04	5.71E-04	3.33E-04	0.001	3.33E-04
R815	0.001	0.00100002	0.001	0.00100002	0.001	0.00100002
R816	1.71E-04	3.33E-04	7.62E-04	3.33E-04	0.001	3.33E-04
R817	0.001	0.00100051	0.001	0.00100051	0.001	0.00100051

Table 7. Selection of reactions from our *M. tuberculosis* metabolic model with kinetic parameters trained on three steady-state fluxes with glycerol uptake (R812). After performing parameter estimation, the model was simulated in COPASI with glycerol uptake at 0, 0.5 and 1 mmol/gDW. Our model is only able to replicate the middle experiment when glycerol uptake is at 0.5 mmol/gDW.

Interestingly, our model was only able to predict the steady-state when glycerol update was at 0.5 mmol/gDW as shown in Table 7. Changing the glycerol level seemed to have no effect on the overall state of the model. A possible explanation for this observation is that the model, which was purposely built for FBA, is unsuitable for dynamic analysis as the input and exchanges fluxes are defined differently in models for FBA and kinetic modelling. Another factor is that training the model with a time series that combines three steady states is not the same experiment as having three separate steady states. A suitable training data set should include intermediate data points covering the dynamics of transition between steady states, which cannot be obtained by FBA but requires extensive experimental measurements.

4.5 Discussion

In this paper, we present the first genome-scale kinetic model of *Mycobacterium tuberculosis* based on generic kinetic equations. In recent years, there has been considerable progress in genome-scale data collection technologies, leading to ever increasing amounts of data in many organisms. However, the exploitation of such large datasets is proving challenging. For example, Ishii *et al.* (2007) measured mRNA, protein and metabolite levels in multiple genetic and environmental perturbations in *E. coli*. Castrillo *et al.* (2007) carried out comprehensive measurements at different growth rates in *S. cerevisiae*. Recently, Yus *et al.* (2009) presented a global and multifaceted analysis of *Mycoplasma pneumoniae*. While each of these studies provided considerable new knowledge about the biology and cellular functions of their respective organism, a comprehensive model that is able to explain, and thus predict, such a large breadth of behaviours is still lacking for each of them. The main reason is that the construction of large kinetic models is arduous and challenging, and there are no established tools and techniques enabling the estimation of numerous kinetic parameters from large sets of heterogeneous data. Our aim is to show that, as an initial step, the construction of such genome-scale models, given a comprehensive set of flux data, can be achieved. The GRaPe tool assigns rate equations to all the reactions in the model based on the stoichiometry of the reaction. We successfully applied our methodology to the *M. tuberculosis* genome-scale metabolic network, resulting in a kinetic model with 739 metabolites, 856 metabolic reactions and 856 enzymes.

Genome-scale metabolic models are essential in bridging the gap between the metabolic phenotypes and genome-derived biochemical information by providing a platform for the interpretation of experimental data related to metabolic states and by facilitating simple *in silico* experimentation of cell metabolism (Durot *et al.,* 2009). Genome annotation and sequencing has made reconstruction of genome-scale metabolic networks possible (Price *et al.,* 2003). By using constraint-based models and *in silico* simulation, phenotypic functions of genome-scale metabolic networks were investigated (Price *et al.,* 2003). Current advances in high-throughput experimental technologies and computational systems biology have enabled genome-scale stoichiometric metabolic models to be reconstructed for an increasing number of organisms (Milne *et al*., 2009).

Predicting cellular behaviours *in silico* by examining the dynamics and properties of cellular processes has the potential to increase our understanding of biological systems. This makes it necessary to advance towards kinetic modelling in our drive to understand the detailed dynamics of cellular functions and their regulation. However, it is time-consuming and costly to experimentally measure all metabolite concentrations, reaction fluxes and kinetic parameters at the genome scale. Additionally, many kinetic equations are unknown and thus, standard rate laws have been used to describe metabolism (Liebermeister *et al.,* 2010). Adiamah *et al*. (2010) and Ao *et al*. (2008) have all shown that using generic rate equations, the dynamical behaviour of systems can be predicted without experimentally measuring all kinetic parameters. Constraint-based modelling fails in capturing the dynamics of cellular behaviour and is unable to provide insights into the changes in the concentration of metabolites and enzymes.

Beste *et al*. (2007) produced a constraint-based simulation of a genome-scale metabolic model of *M. tuberculosis* which was capable of predicting different growth conditions using FBA. The phenotype growth of 78% of mutant strains was correctly predicted by the Beste model. We built a genome-scale kinetic model of *Mycobacterium tuberculosis* based on the stoichiometric model by Beste *et al*. (2007) and showed that our model accurately reproduced genome-scale flux distributions under different growth conditions. The kinetic parameters used in our model were estimated using only flux values, therefore there remains a degree of redundancy in parameter values as illustrated by our PVA (figure 24). The results from our PVA indicate that *vf*, the velocity of the forward reaction, appears to be the most constrained parameter. The rest of the parameters in our model exhibit a high degree of redundancy. Banga (2008) suggests that global optimisation methods are needed in an attempt to avoid finding local solutions which can often be misleading. Additionally, there are suggestions indicating that due to the stochastic nature of biological systems, parameter estimation must account for this degree of stochasticity (Reinker, 2006). Reducing the value of the objective function in parameter estimation improves the quality of the kinetic parameters. However, we observed a significant increase in computing time when the objective function was reduced beyond 1.0E-8. The compromise between computing time and more precise parameter values must always be considered when performing parameter estimation. Furthermore, our results also show that computing time increases non-linearly with the number of data points in the parameter estimation training data. When parameter estimation is being carried out for a system in steady-state, the number of data points can be reduced to lower the computing time.

An attempt was also made to constrain our kinetic parameters by training them with data based on three distinct experimental conditions. However, our model was able to predict only one state revealing the limits of using FBA steady state flux distributions to constrain a dynamic model. In the future, this methodology will be expanded to include metabolite concentrations, proteomics and gene expression data, in order to reduce such redundancy and further advance towards the goal of creating an efficient method to build a fully integrative genome-scale model.

4.6 Conclusion

There has been a recent rise in the number of genome-scale metabolic reconstruction of various organisms (Förster *et al,* 2003*;* Puchałka *et al.,* 2008; Thiele *et al.*, 2009; Smallbone *et al.,* 2010; Park *et al*., 2011). However, many of these models are analysed using constraint-based approaches and techniques due to the difficulty and lack of data to fully characterise kinetic models. Furthermore, rate laws and enzyme mechanisms governing reactions are unknown. This has led to the use of alternative solutions such as linlog (Smallbone *et al.*, 2007), generic equations (Ao *et al.*, 2008; Adiamah *et al.*, 2010) and convenience kinetics (Liebermeister, 2010) when the enzyme mechanism is unknown.

Kinetic models require detailed knowledge of reactions and kinetic parameters. To date, there are still no established software tools and techniques that allow for the estimation of numerous kinetic parameters from large sets of heterogeneous data. Usually, manual sourcing of data from literature and articles is used. This can be extremely challenging and time-consuming when building a large-scale kinetic model. As a result, using generic equations and parameter estimation techniques to build large-scale kinetic models can be a viable option in an attempt to understand the dynamical nature of biological systems.

In this article, we have shown that generic equation and convenience kinetics are capable of reproducing experiments under different growth rates. We developed the first genome-scale kinetic model of *Mycobacterium tuberculosis* based on generic kinetic equations. The model has 739 metabolites, 856 metabolic reactions and 856 enzymes. All kinetic parameters for each reaction were estimated using a genetic algorithm based on stoichiometric and the flux distribution matrices of the network. Our results show a near-perfect agreement with flux distributions under different growth conditions. Nevertheless, the kinetic parameters used in our model were estimated using only flux values, therefore there inevitably remains a degree of redundancy in parameter values. This is evident in our PVA which indicates most of the parameters in our model are not constrained – the most constrained parameter was *vf,* the forward reaction velocity.

Producing quantitative models capable of predicting biological outcomes when perturbed *in silico* is a fundamental aim of systems biology as such models can be used to score biochemical hypotheses. To further improve the predictive power of genome-scale dynamic models, the integration of more experimental data types including gene expression and metabolomics, as well as the use of dynamic training data sets will be needed. Nevertheless, our method for constructing a genome-scale kinetic model of *Mycobacterium tuberculosis* represents a platform for further model development and analysis.

4.7 References

Supplementary Online Data: http://www.bioinf.manchester.ac.uk/schwartz/grape/suppl.html

1. Adiamah, D.A., Handl, J., Schwartz, J.-M. (2010) Streamlining the construction of large-scale dynamic models using generic kinetic equations. *Bioinformatics*, **26**, 1324 – 1331.

2. Banga, R. J. (2008) Optimization in computational systems biology. *BMC Systems Biology*. 2:47.

3. Beste, D.J.V. et al. (2007) GSNM-TB: a web-based genome-scale network model of *Mycobacterium tuberculosis* metabolism. *Genome Biology,* **8**, R80.

4. Bordbar, A., Lewis, N.E., Schellenberger, J., Palsson, B.Ø., Jamshidi, N. (2010) Insight into human alveolar macrophage and *M. tuberculosis* interactions via metabolic reconstructions. *Molecular Systems Biology,* **6**, 422.

5. Borodina, I. and Nielsen, J. (2005) From genomes to *in silico* cells via metabolic networks. *Curr Opin Biotechnol.*, **16**, 350-355.

6. Castrillo, J., Zeef, L., Hoyle, D., Zhang, N., Hayes, A., Gardner, D., et al. (2007) Growth control of the eukaryote cell: a systems biology study in yeast. *Journal of Biology,* **6**, 4.

7. Chandrasekaran, S. and Price, N.D. (2010) Probabilistic integrative modelling of genome-scale metabolic and regulatory networks in Escherichia coli and Mycobacterium tuberculosis. *PNAS*, **107**, 17845-50.

8. Chopra, P., Meena, L.S., Singh, Y. (2003) New drug targets for Mycobacterium tuberculosis. *Indian J Med Res.*, **117**, 1-9.

9. Chou, I-C. and Voit, E.O. (2009) Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences*. **219**, 57-83.

10. Covert, M.W., Famili, I., Palsson, B.O. (2003) Identifying constraints that govern cell behavior: A key to converting conceptual to computational models in biology? *Biotechnology and Bioengineering*, **84**, 763-772.

11. Durot, M., Bourguignon, P.Y., Schachter, V. (2009) Genome-scale models of bacterial metabolism: reconstruction and applications. *FEMS Microbiol Rev.*, **33**, 164-190.

12. Ederer, M. and Gilles, E.D. (2007) Thermodynamically feasible kinetic models of reaction networks. *Biophys. J*. 92, 1846–1857.

13. Förster, J.. Famili, I., Palsson, B. Ø. and Nielsen, J. (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Research*.13. 244 – 253.

14. Goel, G., Chou, I-C. and E.O. Voit (2008) System estimation from metabolic time series data, *Bioinformatics* **24**. 2505-2511.

15. Ishii, N., Nakahigashi, K., Baba, T., Robert, M., Soga, T., Kanai, A., et al. (2007) Multiple high-throughput analyses monitor the response of E. coli to perturbations. *Science,* **316**, 593-597.

16. Jamshidi, N., Palsson, B.O. (2007) Investigating the metabolic capabilities of *Mycobacterium tuberculosis* H37Rv using the *in silico* strain iNJ661 and proposing alternative drug targets. *BMC Systems Biology,* **1**, 26.

17. Jamshidi, N., Palsson, B. Ø. (2008) Formulating genome-scale kinetic models in the post-genome era. *Molecular Systems Biology*, **4**, 171.

18. Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., Hirakawa, M. (2010) KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Research*, **38**, D355-D360.

19. King E.L. and Altman C. (1956) A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *J. Phys. Chem*., **60**, 1375-1378.

20. Kitano, H. (2002) Systems biology: a brief overview. *Science*, **295**, 1662-1664.

21. Liebermeister, W. and Klipp, E. (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraint. *Theoretical Biology and Medical Modelling,* **3**, 41.

22. Liebermeister, W., Uhlendorf, J., Klipp, E. (2010) Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics*, **26**, 1528-1534.

23. Lubitz, T., Schulz, M., Klipp, E. and Liebermeister, W. (2010). Parameter Balancing in Kinetic Models of Cell Metabolism. *The Journal of Physical Chemistry B.* **114**. 16298 – 16303.

24. Kell, D. B. (2006) Metabolomics, modelling and machine learning in systems biology – towards an understanding of the languages of cells. *Febs Journal*. 273**,** 873-894.

25. Mendes, P. and Kell D. B. (1998) Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*. 14 (10):869-883.

26. Mendes, P., Hoops, S., et al. (2006). COPASI: a COmplex PAthway SImulator. *Bioinformatics*, **22**, 3067-74.
27. Milne, C.B., Kim, P.J., Eddy, J.A., Price, N.D. (2009) Accomplishments in genome-scale in silico modeling for industrial and medical biotechnology. *Biotechnology Journal,* **4**, 1653-1670.

28. Park, M. J., Kim, Y. T. and Lee, Y. S. (2011) Genome-scale reconstruction and *in silico* analysis of the *Ralstonia eutropha* HI6 for polyhydroxyalkanoate synthesis, lithoautotrophic growth and 2-methyl citric production. *BMC Systems Biology*. **5**. 101.

29. Pavlidis, P. and Noble W.S. (2003) Matrix2png: A Utility for Visualizing Matrix Data. *Bioinformatics,* **19**, 295-296.

30. Price, N.D., Papin, J.A., Schilling, C.H., Palsson, B.Ø. (2003) Genome-scale microbial *in silico* models: the constraint-based approach. *Trends Biotechnol.,* **21**, 162-169.

31. Puchałka J, Oberhardt MA, Godinho M, Bielecka A, Regenhardt D, et al. (2008) Genome-Scale Reconstruction and Analysis of thePseudomonas putida KT2440 Metabolic Network Facilitates Applications in Biotechnology. *PLoS Comput Biol*. 4(10).

32. Raman, K. Yeturu, K., Chandra, N. (2008) targetTB: A target identification pipeline for *Mycobacterium tuberculosis* through an interactome, reactome and genome-scale structural analysis. *BMC Systems Biology,* **2**, 109.

33. Reinker, S., Altman R. M. and Timmer, J. (2006) Parameter estimation in stochastic biochemical reactions. *Iee Proceedings Systems Biology* 2006, 153(4)**:**168-178.

34. Smallbone, K., Simeonidis, E., Broomhead, D.S., Kell, D.B. (2007) Something from nothing - bridging the gap between constraint-based and kinetic modelling. *FEBS Journal,* **274**, 5576-5585.

35. Smallbone, K. *et al*. (2010) Towards a genome-scale kinetic model of cellular metabolism. *BMC Systems Biology,* **4**, 6.

36. Teusink, B., *et al.* (2000) Can yeast glycolysis be understood in terms of *in vitro* kinetics of the constituent enzymes? Testing biochemistry. *Eur. J. Biochem*., **267**, 5313-5329.

37. Thiele, I., Jamshidi, N., Fleming, M.T. R. and Palsson, B. Ø (2009) Genome-Scale Reconstruction of *Esherichia coli's* Transcriptional and Translational Machinery: A knowledge Base, Its Mathematical Formulation, and It's Functional Charaterization. *PLoS Comput Bio*. **5** (3).

38. van Riel, N. A.W. (2006) Dynamic modelling and analysis of biochemical networks: mechanism-based models and model-based experiments. *Briefings In Bioinformatics*. **7**. 364 – 374.

39. Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E., Shlomi, T. (2010) Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, **26**, i255-i260.

40. Yus, E., Maier, T., Michalodimitrakis, K., van Noort, V., Yamada, T., Chen, W.H., et al. (2009) Impact of genome reduction on bacterial metabolism and its regulation. *Science,* 326, 1263-1268.

41. Zi, Z. and Klipp, E. (2006) SBML-PET: a Systems Biology Markup Language-based parameter estimation tool. *Bioinformatics,* **22**, 2704-2705.

Chapter 5

Integrating Proteins into Metabolic Networks Models to Predict States

The following works appears in “An integrative kinetic model of *Escherichia coli* central metabolism.” Adiamah, DA and Schwartz, J-M. (2011). *To be submitted.*

The aim of the study is to validate our modelling approach by integrating gene expression data with metabolic and proteomic data. We were able to show that integrating biological data at different layers increases robustness of the system and that our model is still able to predict the dynamical behaviour of a biological system. Furthermore, we also analyse the level of redundancy in our estimated kinetic parameters to determine whether adding different data to our estimation technique is able to reduce redundancy.

5.1 Abstract

With the current increase in '*omics'* data, it is now possible to build detailed integrative biological models in an attempt to increase our biological understanding of outcomes when models are perturbed *in silico*. Mathematical models have been shown to describe the complex relationship that exists amongst biological components with models being able to predict new hypotheses and replicate biological states. However, these mathematical models still suffer from the lack or unavailability of parameters needed to describe model behaviour. This has lead to the emergence of different optimisation techniques for estimating parameters models when experimental parameters are unavailable or unobtainable in the lab. However, previous works have shown that estimated parameters do exhibit a degree of redundancy which leads to multiple sets of parameters being able to fit to the same experimental data. Consequently, there is now the need to reduce redundancy in estimated parameter sets by constraining parameters with data at different biological layers.

In this study, we present a model building process integrating genomic and proteomic data with metabolic and flux data. Gene expression is modelled using ordinary differential equations. We apply our approach to building an integrative model of *E. coli* central metabolic network based on experimentally determined datasets. By integrating both genomic and proteomic data into a single model and comparing it with a model integrating only proteomic data, we show that both models are able to replicate biological conditions under steady-state. To further validate our modelling approach, we performed 14 protein knockdown experiments to determine the robustness and predictability of our *E. coli* models. Results from our protein knockdown experiments suggest that integrating genomic and proteomic data in metabolic network increases robustness and stability in biological systems.

We further carried out an experiment to determine the level of constraint in our estimated parameters. Results from our parameter variability analysis indicate that the forward reaction rate is the most-constrained parameter in each reaction while most other parameters are more poorly constrained

5.2 Introduction

Mathematical models are known to be very important in the field of systems biology (Kell and Knowles, 2006) as they describe the complex relationship that exists between components in a biological network (Li *et al.,* 2010). Previous works have shown that mathematical models are able to replicate and predict biological states (Feala *et* al., 2008; Li *et* al., 2010; Adiamah *et al*., 2010; Ruppin *et al*., 2010) and possibly give insights into the emergent properties of biological functions (Feala *et al*., 2008). In the last few years, while the number of constraint-based models utilising flux balance analysis (FBA) in analysing biological systems has grown, detailed kinetic models still remain comparatively low in numbers. Using constraint-based modelling (CBM), which does not require kinetic parameters, the phenotypic functions of an organism can be defined and studied. The successful prediction rate of CBM models *in silico* is observed to be approximately 86% for *E. coli* in gene deletion experiment (Price *et al*., 2003).

Modern high-throughput experiment including genome sequencing and DNA microarray (Pease *et al*., 1994; Schena *et al*., 1995) analysis have made it possible to measure the quantitative levels of gene expression on a genome-scale (Sherlock *et al*., 2001). Furthermore, the availability of modern 'omics' data about individual organisms means it is now possible to collect quantitative data in all layers of cellular organisation (Figure 27). Nevertheless, to transform these datasets into a large-scale quantitative model remaing a highly challenging task.

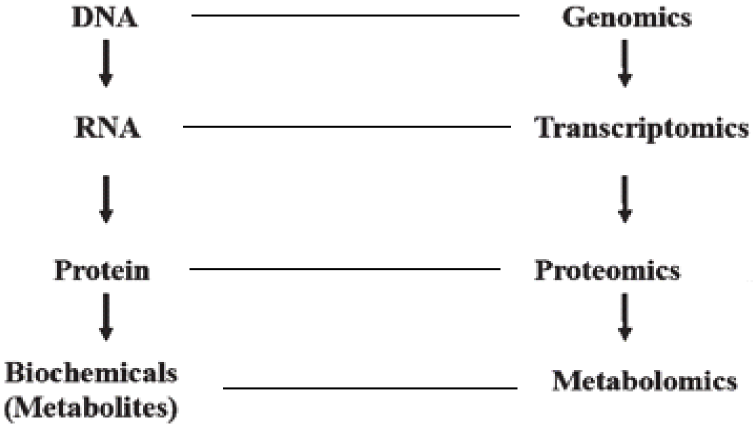


Figure 27. A schematic view of the various 'omics' technologies relating to different biological layers.

The availability of information at different biological layers has spurred the growth of integrative systems biology – where attempts are being made to construct detailed models that include biological data on different layers (Baldazzi *et* al., 2010; Ruppin *et al*., 2010). To fully understand the response of an organism to environmental changes, it is essential to include detailed quantitative levels of genes, mRNA transcripts, proteins and metabolites and their subsequent interactions (Zhang *et al*., 2010). It was shown in Ter Kulie and Westerhoff (2001) that the control of glycosis was shared between genomic, proteomic and metabolic levels. This example highlights the significance of building integrative models. Integrative models of various organisms have been constructed but most are built using constraint-based modelling approach (Fellenberg, 2003; Cavalieri and Filippo, 2005; Çakir *et al*., 2006; Joyce and Palsson, 2006; Herrgård *et al.* 2006; Yizhak *et* al., 2010; Zhang *et al*., 2010). Kinetic modelling has been shown to capture the detailed dynamics of biological systems as biological systems are not discrete in nature (Puchalka and Kierzek, 2004; Resat *et al*., 2009). However, only a few detailed integrative kinetic models have been constructed to date. Detailed integrative kinetic models are still hindered by the availability of kinetic parameters and the incompleteness of heterogeneous data needed to fully define such models. As a result, parameter estimation has become a very important and central part of computational systems biology (Mendes and Kell, 1998**;** Moles *et al*., 2003; Goel et al., 2008; Ashyraliyev *et al*., 2009) and efforts have been made on predicting the correct dynamical behaviour of biological systems rather than measuring highly accurate parameter values (Ao *et al*., 2008; Liebermeister and Klipp, 2006; Adiamah *et al*., 2010; Liebermeister *et al.*, 2010). Additionally, building integrative models has proven difficult as data pertaining to an organism is stored in various databases and mostly under different experimental conditions (Radrich *et al*., 2009).

One of the major problems of parameter estimation is that large amounts of experimental data are usually required to determine the values of all unknown parameters in a network (Heinzle *et al.*, 2007). Gutenkunst *et al*. (2007) and Little *et al*. (2009) have both shown in previous studies that parameterisation of biological model can exhibit a degree of “sloppiness” or redundancy, respectively. These inconsistencies mean that our efforts must be directed towards constraining parameter values in parameter estimation and predicting accurate biological behaviours. If systems biology is to serve as the foundation for genome-scale synthetic biology (Barrett *et al*., 2003**),** then there is the need to build and integrate cellular systems capable of accurate predictions. As a result, parameter variability analysis (PVA) is performed to measure the degree of constraint in parameters estimated using time series experimental data.

Chen *et* al. (1999) provided two methods for modelling gene expression using kinetic equations with feedback loops. However, both methods were limited as the regulation of enzymes on metabolites was not included. In this study, we model gene expression using ordinary differential equations but importantly, we integrate the gene expression process including transcription, translation and degradation with the metabolic network. We show that modelling gene expression using ordinary differential equations is capable of capturing the dynamics of transcription and translation.

By integrating gene expression into metabolic networks, our understanding of how genes and mRNAs affect metabolism under different experimental conditions can be further explored. Additionally, by allowing the inclusion of gene expression data into the parameter estimation process, parameters are better constrained which makes the model more robust. This integration could also serve as a platform for further integration of higher-level data such as signalling or metabolic regulation to extend the coverage of the model.

5.3 Methods

A stoichiometric matrix, N, can be used to present a metabolic network consisting of n metabolites and m reactions where N is an n*m matrix. Under a steady-state condition,

$$\mathbf{N}v = 0 \qquad (39)$$

where v is the flux distribution matrix corresponding to all reaction fluxes in the model. We modelled reaction rate in our model based on the generic rate equations (Adiamah *et al.,* 2010). In our previous study, generic Michaelis-Menten rate equations were used to describe reactions of up to two substrates or products and all reactions were assumed to follow a random-order mechanism as the sequential binding of substrate and release of products were unknown. GRaPe uses the King and Altman (1956) method in deriving rate equations for all reactions in a model based on the stoichiometry and enzyme mechanism of that particular reaction. However, reactions of more than two substrates or products had to be decomposed into reactions of no more than two substrates and products which could be time-consuming and error-prone. Consequently, in our effort to streamline the construction of large-scale dynamical models, we have introduced convenience kinetics to express reactions of more than two substrates or products (Liebermeister and Klipp, 2006). It has been shown in previous studies that convenience kinetics, which describes reactions containing any number of reactants and products, is able to model the behaviour of biological systems with good precision when exact rate equations are unknown (Liebermeister and Klipp, 2006; Liebermeister and Klipp, 2010).

In our model, reactions of type $A_1 + A_2 + ... + A_k \leftrightarrow B_1 + B_2 + ... + B_k$, for $k > 2$, the flux v (a, b) is defined using convenience kinetics as:

(40)

where a_1, a_2, ... represent the concentrations of substrates and b_1, b_2, ... the concentrations of products, i is the number of substrates, j is the number of products, $K_{A\mathrm{i}}$ represents the kinetic parameter (substrate constants) of the i^{th} substrate, K_{Bj} is the j^{th} product of the reaction (both K_As and K_Bs are measured in mM), the concentration of enzyme is represented as e_0, v^+ is the substrate turnover rate or the forward reaction rate, v^- is the product turnover rate or reverse reaction rate.

By incorporating the concentration of enzyme, e_0, as a variable into the equation, rather than a fixed constant, we are able to integrate quantitative levels of proteins and metabolites with metabolic fluxes using equation (40). GRaPe now uses convenience kinetics in the definition of reactions of more than two substrates and products.

When there is no availability of genomic and quantitative proteomic data, a Boolean function can be used to represent the expression of an enzyme in a model as demonstrated in Adiamah *et al.* (2010). In this case, enzymes are represented by 1 when expressed and 0 when not expressed. However, as biological processes are not discrete in nature, using a Boolean style representation to understand gene expression can result in the loss of information. The correlation between the concentration of mRNA and protein levels is nevertheless not straightforward. While some studies have suggested a non-linear relationship between them (Vogel *et al.*, 2010), other studies have suggested that there is no clear correlation (Smolen *et al.*, 2003; Garcia-Martinez *et al.* 2007). In this study, we expand on our previous study by introducing genomic, transcriptome and proteomic data into a model of central metabolism of *E. coli*. The quantitative changes of mRNA and protein concentrations are modelled using ordinary differential equation. The change in concentration of mRNA relative to time is modelled as:

$$\overline{} \qquad\qquad - \qquad\qquad\qquad (41)$$

where Gene(*t*) indicates whether or not the gene is expressed at time, *t*. [mRNA] is the concentration of mRNA. k_{Tr} is the transcription rate and k_{deg} is the mRNA degradation rate.

In a steady-state analysis, equation (41) is defined as

$$\frac{}{} \tag{42}$$

where the concentration of mRNA remains unchanged over time. In a gene knockout analysis, the gene and mRNA concentration are set to zero.

Likewise, the concentration of proteins is expressed as a function of the concentration of mRNA. We model translation – which results in the production of proteins - in our study as

$$\frac{}{} \qquad - \tag{43}$$

where k_{Tls} is the translation rate, *mRNA* is the concentration of mRNA, k_{deg} is the protein degradation rate and *Prot* is the concentration of proteins. By expressing the level of proteins as a function of mRNA, we can analyse the dynamical changes in proteins relative to the concentration of mRNA and whether the gene(s) pertaining to that enzyme is expressed or not. The level of mRNA can also be changed to simulate a gene knockdown analysis which usually results in a change in protein level. In simulating a gene knockdown experiment, the amount of mRNA in equation (43) can be altered appropriately.

Isoenzymes are enzymes with different amino acid sequence but catalysing the same reaction. In this study, we model isoenzymes by summing up the total concentration of individual enzymes as expressed by their respective gene or genes. In this situation, an isoenzyme is expressed mathematically as:

$$E = \qquad (44)$$

where E is the total concentration of the enzyme, GE_i equates to the gene expression of the *ith* gene and *n* is the number of genes. The transcription and translation processes for GE_i are expressed as (41) and (43) respectively.

All these equations are automatically generated by the GRaPe software given the stoichiometric and genetic structure of a reaction network (Adiamah *et al*., 2010).

The complex relationship that exists between components at different biological layers is seen as producing cellular functions and as such to fully understand biological systems, our efforts must be directed at building integrative models which combine global biological information at different levels (Ideker, T. *et al.* 2001; Ma and Zeng, 2003; Reed *et al.,* 2006; Joyce and Palsson, 2006; Herrgård *et al.* 2006). In Ishii *et al* (2010), the authors collected data for wild-type *E. coli* k-12 and 24 single–gene knockouts using multiple high-throughput analyses with all cells grown at a fixed dilution rate of 0.2 hours^{-1} in glucose-limited chemostat cultures. The numerous studies of *E. Coli* metabolism at different biological layers means that information about individual pathways and components can be obtained (Reed *et al.,* 2003; Fujisaki *et al.,* 2005; Imielinski *et al.,* 2006; Baba *et al.,* 2006; Feist *et al.*, 2007, Ishii *et al.,* 2010). We used GRaPe to reconstruct the *E. coli* central metabolic model based on the experimental model by Ishii *et al*. (2007).

5.3.1 Parameter estimation and Parameter Variability Analysis

The relatively low amount of detailed and accurate kinetic models being produced to date can be attributed to the lack of kinetic parameters needed to fully define the metabolic model. As a result, there are numerous optimisation techniques for estimating kinetic parameters and their relative challenges (Goss and Peccoud, 1998; Moles *et al.*, 2003; Bruggeman and Westerhoff, 2007). As there were no kinetic parameters available to fully define our model, a genetic algorithm (GA) was used to estimate kinetic parameters for our *E. coli* central metabolic network. The GA was specifically designed to incorporate flux data in constraining parameter values. The GA, described in Adiamah *et al*. (2010), is also implemented in GRaPe. Our genetic algorithm uses proteins, metabolites and flux data to constrain kinetic parameters.

Previous studies have shown that the effects of imprecise parameter values and missing data points on the predictability of the model can be highly heterogeneous (Bruggeman and Westerhoff, 2007; Gutenkunst *et al.*, 2007). We therefore performed parameter variability analysis (PVA) to establish the degree of redundancy in our estimated kinetic parameters. In PVA, kinetic parameters and rate constants for each reaction are estimated *n* times. The distribution of estimated parameter values over the *n* runs is then statistically analysed. With the aim of producing models capable of producing effective experiments *in silico*, it is important to verify the level of redundancy in the parameterisation of biological models.

5.4 Results

As proof of concept, we demonstrate that by using ordinary differential equations, we can replicate the dynamical changes in the level of protein as determined by the expression of its genes and mRNA transcripts. Firstly, in an attempt to validate our model, we show that our model of the central metabolic network of *E. coli* is capable of reproducing an experimental condition based on trained data without the integration of gene expression. Following that, we show that our model is able to reproduce experimental condition with the integration of differential equations modelling transcription and translation (gene expression). To show that generic ordinary differential equations have the possibility of predicting experimental conditions *in silico*, we perform gene knockdown simulations to validate our methodology. Finally, we show the results of our parameter variability analysis from which we can deduce the level of redundancy in our estimated parameters and estimate the robustness of biological systems.

5.4.1 Proof of Concept

Our first approach was to show that using ordinary differential equations, the quantitative changes in protein level can be modelled. As a proof of concept, we model the gene expression process of the PFK protein in the glycolysis pathway in *E. coli*. The PFK protein is expressed by two genes, *pfkA* and *pfkB*, as shown in Figure 28. The genes are transcribed into mRNA transcripts before being translated into the PFK protein. Using GRaPe, we modelled the PFK reaction in the *E. coli* glycolysis pathway using generic kinetic equations. The gene expression process of PFK protein is modelled using ordinary differential equations. The total concentration of PFK is determined by equation (44). We initialised our model with experimental data from Ishii *et al.*, (2007) and estimated kinetic parameters for the model using the genetic algorithm implemented in GRaPe.

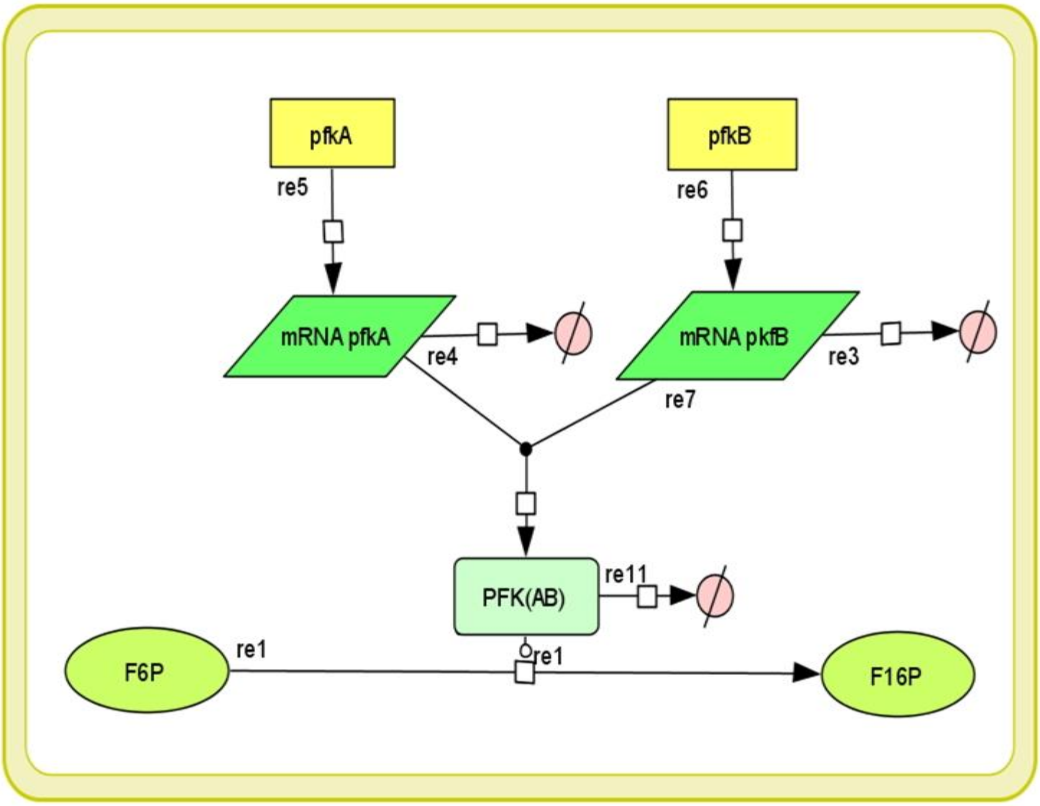


Figure 28: Gene expression processes of the PFK protein. Genes are transcribed into mRNA which then gets translated into proteins. In the PFK protein, the *pfkA* and *pfkB* genes are responsible for the expression of the PFK protein. Transcription of both *pfkA* and *pfkB* is represented by reactions re5 and re6 respectively; which is then followed by the translation of both *pfkA* and *pfkB* mRNA transcripts into PFK protein by reaction re7. The degradation of both mRNA and PFK over time is represented by re3 and re11 respectively. Transcription, translation and degradation are modelled using ordinary differential equations. In glycolysis, F6P (fructose 6-phosphate) is transformed into F16P (fructose 6-phosphate) by the PFK protein, re1. We model this reaction using a generic kinetic rate equation. (Figure drawn using Systems Biology Graphical Notation (SBGN) (Le Novère *et al*., 2009)).

After parameter estimation, we used SBML ‘events’ to replicate the effects of a gene knockout experiment. The effects were then simulated in CellDesigner using the SBML ODE Solver (SOSlib) (Funahashi *et al*., 2003). Our first experiment was to replicate the wild-type effect when both *pfkA* and *pfkB* genes are expressed. We then performed three knockout experiments by firstly, knocking-out only pfkA gene followed by the knockout of only pfkB gene and finally, a knockout both pfkA and pfkB genes using equation (41) with the gene and mRNA concentrations set to zero. Lastly, both mRNA transcripts were reduced by 50%.

Our results from this proof of concept show that using ordinary differential equations, the dynamical behaviour of gene expression can be modelled as shown in Table 8. When there is a knockout of one gene, we observed a 50% reduction in PFK protein level and a knockout of both genes results in the protein level reducing to 0. When mRNA transcripts were reduced by 50% the total concentration of the PFK protein is only reduced by 50% as expected.

pfkA	**pfkB**	**mRNA pfkA**	**mRNA pfkB**	**PFK(AB)**	**Flux**
1	1	8.91E+07	6.27E+07	0.06	84.98
0	1	0	6.27E+07	0.03	42.16
1	0	8.91E+07	0	0.03	42.14
1	1	4.46E+07	3.14E+07	0.28	42.19
0	0	0	0	0	0

Table 8: Quantitative changes in protein level of PFK in *E. coli* glycolysis pathway. Genes *pfkA* and *pfkB* are represented as Boolean values where 1 indicates that the gene is expressed and 0 when not expressed. mRNA transcript of *pfkA* and *pfkB* were measured in mg-protein/g-dry cell weight; PFK(AB), representing the total concentration the PFK protein, is measured in protein in mg-protein/g-dry cell weight; flux expressed as a % of substrate uptake.

5.4.2 Model Building

We reconstructed a kinetic model of the central metabolic network of *E. coli* based on the dataset provided by Ishii *et al.* (2007). Our aim is to show that using generic rate equations to model metabolic reactions coupled with ordinary differential equations to model gene expression, we can predict the behaviour of biological metabolic systems at a steady-state and possibly predict states under different experimental conditions.

In Ishii *et al.* (2007), the global response of *E. coli* K-12 cells to both genetic and environmental perturbations at the gene expression and protein levels were measured and compared to specific metabolic pathways. The central carbon metabolism of *E. coli* K-12 consisted of the glycolysis, pentose phosphate pathway, and the tricarboxylic acid cycle (TCA) as these three pathways play a vital role in the generation of energy and the production of important macromolecular precursors. The quantities of gene and protein products, and the concentrations of metabolites were experimentally determined. The aim of the study by Ishii *et al.* (2007) was to highlight the complex relationship that exists between the different biological layers. The availability of protein, metabolic and flux data pertaining to *E. coli* makes this study a suitable reference sample for our integrative modelling approach.

Using GRaPe (Adiamah *et al.*, 2010), we manually constructed two models of the carbon central metabolic of E. coli K-12. The first model, EC Model 1, had only quantitative protein levels integrated with metabolites while the second model, EC Model 2, had the complete gene expression process integrated into the metabolic network. EC Model 1 allows for an easy analysis of the effects of varying the level of proteins in a model when the knowledge of transcription, translation and degradation rates is missing or limited. The initial concentrations of metabolites and expression levels of genes and proteins were the same as those of the *E. coli* wild-type in the Ishii dataset except for 16 metabolites which were initialised to 0.01 and 3 proteins initialised to 1 as no data was present for them (see Table 9 for a list of metabolites with missing data). The number of genes, proteins and mRNA transcripts differed between our models and that observed in the Ishii dataset due to the availability of experimental data present in the experiments conducted by Ishii.

In both EC Model 1 and 2, the number of proteins was 62 compared with 67 in the Ishii dataset. Proteins shown in Table 9 were removed from our models as they were undetected in the wild-type and gene-knockout experiments in the Ishii dataset. However, three proteins without protein data, Edd, SfcA and PoxB, were included in our models and their concentrations initialised to 1 as the reactions they catalyse had fluxes and gene expression data in the experimental data. The mRNA transcripts and genes for the omitted proteins were also excluded in our EC Model 2. As a result, the number of mRNA transcripts and genes in our model was 69 compared with the 85 seen in the experimental dataset provided by Ishii. Additionally, 12 reactions which are included in cell synthesis and evolution were modelled as exchange reactions with an arbitrary “transport” protein.

Internal Metabolites	External Metabolites	Proteins
DHAP	Lactate	Plk
G3P	Formate	PfkA
2PG	Acetadehyde	GpmG
AcCoA	Ethanol	RpiB
CO_2	Cell	PflC
Gluconolactone-6P	Synthesis	
6PG		
2-KDPG		
X5P		
E4P		
0AA		
CIT		
ICT		
Suc-CoA,		
Glyoxylate		
Ac-P		

Table 9: A list of metabolites and proteins without an experimental initial concentration for the wild-type. As result, the initial concentration for these internal metabolites was set to 0.1 mM and concentration of external metabolites was fixed at 1mM (boundary condition is set to “True” for all external metabolites).. These proteins were removed from our model as they were not detected in the wild-type.

We assume that all metabolic reactions are of a random-order mechanism. All rate equations, including gene expression equations, were then automatically assigned by GRaPe once the models were reconstructed. The generic rate equation for each metabolic reaction in the network is determined based on the number of products and substrates of that reaction. Since cofactors were not included in the Ishii experimental data, likewise they were ignored in our models. There were 9 more reactions in our models (shown in Table 10) than in the Ishii dataset as coupled reactions were broken down into individual reactions as their proteins and mRNA levels were given in the experiment data.

Coupled Reactions	Individual Reactions
G3P → 3PG	G3P → PGP PGP→ 3PG
3PG → PEP	3PG →2PG 2PG → PEP
G6P → 6PG	G6P → Gluconolactone-6P Gluconolactone-6P → 6PG
6PG → G3P + PYR	6PG → 2-KDPG 2-KDPG → G3P + PYR
2-KG → SUC + CO_2	2-KG → SUC-CoA SUC-CoA → SUC + CO_2
AcCoA → Acetate	AcCoA → Ac-P Ac-P → Acetate
Not given	PYY → Acetate (External)
AcCoA → Ethanol	AcCoA → Acetadehyde Acetadehyde → Ethanol

Table 10. A list of coupled reaction in the Ishii dataset (left) and their corresponding separated reactions in our E. coli central metabolic models (right).

Only metabolites included in the central metabolic network were included in our model. This meant that most of the cations and anions in the Ishii dataset were ignored. Overall, there were 32 metabolites and 53 metabolic reactions in both our models compared with 130 metabolites and 43 reactions in the Ishii dataset. EC Model 2 had 257 reactions in all – including transcription, translation and degradation rate reactions. A list of genes, mRNA transcripts, proteins and reactions used in our models is given in Table 11.

Genes	mRNA	Proteins	Reactions	Reaction ID
galM	galM	GalM	**Glu_ex <--> Glucose**	R1
glk	glk	Glk	**Glucose + PEP <-> G6P + PYR**	R2
pgi	pgi	Pgi	**G6P <-> F6P**	R3
pfkA, pfkB	pfkA, pfkB	PfkA, PfkB	**F6P <-> F16P**	R4
fbaA	fbaA	FbaA	**F16P <-> DHAP + G3P**	R5
tpiA	tpiA	TpiA	**DHAP <-> G3P**	R6
gapA	gapA	GapA	**G3P <-> PGP**	R7
pgk	pgk	Pgk	**PGP <-> 3PG**	R8
gpmA, gpmB	gpmA, gpmB	GpmA, GpmB	**3PG <-> 2PG**	R9
eno	eno	Eno	**2PG <-> PEP**	R10
pykA, pykF	pykA, pykF	PykA, PykF	**PEP <-> PYR**	R11
aceE, aceF	aceE, aceF	aceE, aceF	**PYR <-> AcCoA + CO2**	R12
zwf	zwf	Zwf	**G6P <-> Gluconolactone-6P**	R13
pgl	pgl	Pgl	**Gluconolactone-6P <-> 6PG**	R14
gnd	gnd	Gnd	**6PG <-> Ru5P + CO2**	R15
edd	edd	Edd	**6PG <-> 2-KDPG**	R16
rpe	rpe	Rpe	**Ru5P <-> X5P**	R17
rpiA, rpiB	rpiA, rpiB	RpiA, RpiB	**Ru5P <-> R5P**	R18
tktA, tktB	tktA, tktB	tktA, tktB	**X5P + R5P <-> S7P + G3P**	R19
talA, talB	talA, talB	talA, talB	**S7P + G3P <-> E4P + F6P**	R20
tktA, tktB	tktA, tktB	tktA, tktB	**X5P + E4P <-> F6P + G3P**	R21
eda	eda	Eda	**2-KDPG <-> G3P + PYR**	R22
gltA, prpC	gltA, prpC	gltA, prpC	**AcCoA + OAA <-> CIT**	R23
acnA, acnB	acnA, acnB	AcnB	**CIT <-> ICT**	R24
icdA	icdA	IcdA	**ICT <-> 2-KG + CO2**	R25
sucA, sucB, lpdA	sucA, sucB, lpdA	sucA, sucB, lpdA	**2-KG <-> Suc-COA**	R26
sucC, sucD	sucC, sucD	sucC, sucD	**Suc-COA <-> SUC + CO2**	R27
sdhA, sdhB, frdA	sdhA, sdhB, frdA	sdhA, sdhB, FrdA	**SUC <-> FUM**	R28
fumA, fumB, fumC	fumA, fumB, fumC	fumA, fumB, fumC	**FUM <-> MAL**	R29
mdh	mdh	Mdh	**MAL <-> OAA**	R30
pckA, ppc	pckA, ppc	pckA, ppc	**PEP + CO2 <-> OAA**	R31
sfcA	sfcA	SfcA	**MAL <-> PYR + CO2**	R32
aceA	aceA	AceA	**ICT <-> Glyoxylate + SUC**	R33
aceB, glcB	aceB, glcB	aceB, glcB	**Glyoxylate + AcCoA <-> MAL**	R34
pta	pta	Pta	**AcCoA <-> Ac-P**	R35
ackA	ackA	AckA	**Ac-P <-> Acetate_ex**	R36
ldhA	ldhA	LdhA	**PYR <-> Lactate_ex**	R37
poxB	poxB	PoxB	**PYR <-> Acetate_ex**	R38
pflB	pflB	pflB	**PYR <-> AcCoA + Formate_ex**	R39
adhE	adhE	AdhE	**AcCoA <-> Acetadehyde**	R40
adhE	adhE	AdhE	**Acetadehyde <-> Ethanol_ex**	R41

Table 11. List of genes, mRNAs, proteins and reactions used in our model and their respective reaction IDs.

5.4.2.1 Consistency of flux distributions

We carried out a verification process to ensure that fluxes taken from experimental datasets were consistent throughout the network and that mass conservation was preserved in our model. Using the Simple Steady-State Estimator (SSSE) described in Adiamah *et al*. (2010), we calculated only the amount of proteins needed by each reaction in the *E. coli* model to achieve the required experimental flux. SSSE works by assigning all kinetic parameter values to 1 and determining the enzyme level e_0 needed to achieve a given flux, *v*. The rationale behind this is that at any given steady-state, if fluxes are conserved and no changes in fluxes are observed then the concentration of metabolites should remain unchanged in a dynamic simulation. An increase or decrease in metabolite levels indicates an inconsistency in the flux distribution. The model provided by the SSSE was then simulated in CellDesigner using the SOSLib solver. We observed several inconsistencies in the flux distribution through the network as shown in Table 12. Fluxes in the experimental data (second from the right in Table 12) were inconsistent throughout the system. The results showed almost all metabolites either increased or decreased in concentration as the Pgi reaction, R2, consumed more glucose and PEP than required and resulted in both metabolites decreasing to zero over time. No steady-state was found with the raw experimental data. As a result, the fluxes were then adjusted manually by tracing all fluxes through the network to ensure flux consistency and mass conservation.

	Reactions	Experimental Data	Fluxes Adjusted for flux consistency
R1	**Glu_ex <-> Glucose**		100
R2	**Glucose + PEP <-> G6P + PYR**	100	100
R3	**G6P <-> F6P**	80	78
R4	**F6P <-> F16P**	86	85
R5	**F16P <-> DHAP + G3P**	86	85
R6	**DHAP <-> G3P**	86	85
R7	**G3P <-> PGP**	173	172
R8	**PGP <-> 3PG**		172
R9	**3PG <-> 2PG**	162	162
R10	**2PG <-> PEP**		162
R11	**PEP <-> PYR**	46	47
R12	**PYR <-> AcCoA + CO2**	125	129
R13	**G6P <-> Gluconolactone-6P**	18	20
R14	**Gluconolactone-6P <-> 6PG**		20
R15	**6PG <-> Ru5P + CO2**	18	20
R16	**6PG <-> 2-KDPG**		0
R17	**Ru5P <-> X5P**	7	8
R18	**Ru5P <-> R5P**	12	12
R19	**X5P + R5P <-> S7P + G3P**	5	6
R20	**S7P + G3P <-> E4P + F6P**	5	6
R21	**X5P + E4P <-> F6P + G3P**	2	2
R22	**2-KDPG <-> G3P + PYR**		0
R23	**AcCoA + OAA <-> CIT**	86	86
R24	**CIT <-> ICT**		86
R25	**ICT <-> 2-KG + CO2**	86	71
R26	**2-KG <-> Suc-COA**	76	62
R27	**Suc-COA <-> SUC + CO2**	68	62
R28	**SUC <-> FUM**	78	77
R29	**FUM <-> MAL**	78	77
R30	**MAL <-> OAA**	87	89
R31	**PEP + CO2 <-> OAA**	13	11
R32	**MAL <-> PYR + CO2**	0	3

R33	ICT <-> Glyoxylate + SUC	10	15
R34	Glyoxylate + AcCoA <->MAL	10	15
R35	AcCoA <-> Ac-P	0	0
R36	Ac-P <-> Acetate_ex	0	0
R37	PYR <-> Lactate_ex		0
R38	PYR <-> Acetate_ex	0	0
R39	PYR <-> AcCoA + Formate_ex		0
R40	AcCoA <-> Acetadehyde	0	0
R41	Acetadehyde <-> Ethanol_ex		0
R42	G6P -> Cell Synthesis	2	2
R43	F6P -> (Cell synthesis)	1	1
R44	R5P -> (Cell synthesis)	7	6
R45	E4P -> (Cell synthesis)	3	4
R46	G3P -> (Cell synthesis)	1	0
R47	3PG -> (Cell synthesis)	10	10
R48	PEP -> (Cell synthesis)	4	4
R49	PYR -> (Cell synthesis)	21	21
R50	AcCoA -> (Cell synthesis)	29	28
R51	OAA -> (Cell synthesis)	14	14
R52	2KG -> (Cell synthesis)	8	9
R53	CO2 -> (Evolution)	275	274

Table 12. Experimental values of fluxes before adjustments and corrected flux values.

5.4.3 Parameter Estimation

To have a consistent dataset for parameter estimation, data for mRNA, proteins, metabolites and fluxes under the same growth rate were taken from the Ishii experimental data. For each reaction, the flux, protein level and concentration of metabolites for *E. coli* wild-type at 0.2 hours^{-1} in glucose-limited chemostat cultures were merged to create a dataset for parameter estimation of the EC Model 1. The same dataset, but adding the level of mRNA transcripts at the same growth rate, was then used to estimate parameter values for EC Model 2. Using GRaPe, we estimated the kinetic parameters for all reactions in both models such that the summed mean error distance between the experimental input dataset and the values simulated by GRaPe is minimized. The summed error between the input dataset and simulated values was found to be lower than 1.5E-13 after just one run of estimation.

5.4.4 Experiment 1: Model Validation

Once parameter estimation had been completed for EC Model 1, CellDesigner was used to perform a simulation using the SOSlib solver. The results from our simulation were then compared with experimental results from Ishii *et al.* (2010). Previously, we showed that using Boolean values instead of quantitative protein levels, the dynamic behaviour of a system can be reproduced (Adiamah *et al*, 2010). Similarly, an excellent agreement is observed between EC Model 1, without gene expression, and the experimental dataset presented by Ishii for wild-type experiments as shown in Table 13 and 13. Here, we observe that our method is still able to achieve near-perfect results when the levels of proteins are integrated in a model. These results also further validate our methodology and the parameter estimation algorithm in finding a suitable solution given a set of experimental data.

	Concentration (mM)				Concentration (mM)		
Species	**ID**	**Experimental Data**	**Simulated Data**	**Species**	**ID**	**Experimental Data**	**Simulated Data**
Glucose	s1	0.01	0.01	**2-KDPG**	s17	0.10	0.10
PEP	s2	0.01	0.01	**X5P**	s18	0.01	0.01
G6P	s3	0.17	0.17	**R5P**	s19	0.01	0.01
PYR	s4	0.19	0.19	**S7P**	s20	0.01	0.01
F6P	s5	0.01	0.01	**E4P**	s21	0.18	0.18
F16P	s6	0.05	0.05	**OAA**	s22	0.01	0.01
DHAP	s7	0.03	0.03	**CIT**	s23	0.01	0.01
G3P	s8	0.01	0.01	**ICT**	s24	0.01	0.01
PGP	s9	0.01	0.01	**2-KG**	s25	0.01	0.01
3PG	s10	0.01	0.01	**Suc-COA**	s26	0.03	0.03
2PG	s11	0.75	0.75	**SUC**	s27	0.01	0.01
AcCoA	s12	0.01	0.01	**FUM**	s28	0.07	0.07
CO2	s13	0.01	0.01	**MAL**	s29	0.07	0.07
Gluconolactone-6P	s14	0.01	0.01	**Glyoxylate**	s30	0.08	0.08
6PG	s15	0.01	0.01	**Ac-P**	s31	0.01	0.01
Ru5P	s16	0.01	0.01	**Acetadehyde**	s32	0.01	0.01

Table 13: The concentration of metabolites, measured in mM, at simulated steady-state from EC Model 1 (GRaPe model) compared with experimental data from Ishii *et al*. (2007). Results show an excellent agreement between simulated and experimental data without perturbations.

		Fluxes (% of substrate Uptake)				Fluxes (% of substrate Uptake)	
Reaction	**ID**	**Experimental Data**	**Predicted Data**	**Reaction**	**ID**	**Experimental Data**	**Predicted Data**
Glu_ex <--> Glucose	R1	100	99.53	**Suc-COA <-> SUC + CO2**	R27	62	62.01
Glucose + PEP <-> G6P + PYR	R2	100	99.98	**SUC <-> FUM**	R28	77	77.00
G6P <-> F6P	R3	78	78.00	**FUM <-> MAL**	R29	77	77.00
F6P <-> F16P	R4	85	85.00	**MAL <-> OAA**	R30	89	89.00
F16P <-> DHAP + G3P	R5	85	85.00	**PEP + CO2 <-> OAA**	R31	11	11.00
DHAP <-> G3P	R6	85	84.99	**MAL <-> PYR + CO2**	R32	3	3.00
G3P <-> PGP	R7	172	171.99	**ICT <-> Glyoxylate + SUC**	R33	15	15.00
PGP <-> 3PG	R8	172	172.00	**Glyoxylate + AcCoA <->MAL**	R34	15	15.00
3PG <-> 2PG	R9	162	162.00	**AcCoA <-> Ac-P**	R35	0	0.00
2PG <-> PEP	R10	162	162.01	**Ac-P <-> Acetate_ex**	R36	0	0.00
PEP <-> PYR	R11	47	47.00	**PYR <-> Lactate_ex**	R37	0	0.00
PYR <-> AcCoA + CO2	R12	129	129.00	**PYR <-> Acetate_ex**	R38	0	0.00
G6P <-> Gluconolactone-6P	R13	20	20.00	**PYR <-> AcCoA + Formate_ex**	R39	0	0.00
Gluconolactone-6P <-> 6PG	R14	20	20.00	**AcCoA <-> Acetadehyde**	R40	0	0.00
6PG <-> Ru5P + CO2	R15	20	19.91	**Acetadehyde <-> Ethanol_ex**	R41	0	0.00
6PG <-> 2-KDPG	R16	0	0.00	**G6P -> Cell Synthesis**	R42	2	2.00
Ru5P <-> X5P	R17	8	7.98	**F6P -> (Cell synthesis)**	R43	1	1.00
Ru5P <-> R5P	R18	12	11.96	**R5P -> (Cell synthesis)**	R44	6	6.00
X5P + R5P <-> S7P + G3P	R19	6	6.00	**E4P -> (Cell synthesis)**	R45	4	4.00
S7P + G3P <-> E4P + F6P	R20	6	6.01	**G3P -> (Cell synthesis)**	R46	0	0.00
X5P + E4P <-> F6P + G3P	R21	2	2.00	**3PG -> (Cell synthesis)**	R47	10	10.00
2-KDPG <-> G3P + PYR	R22	0	0.00	**PEP -> (Cell synthesis)**	R48	4	4.00
AcCoA + OAA <-> CIT	R23	86	86.00	**PYR -> (Cell synthesis)**	R49	21	21.00
CIT <-> ICT	R24	86	86.00	**AcCoA -> (Cell synthesis)**	R50	28	28.00
ICT <-> 2-KG + CO2	R25	71	71.00	**OAA -> (Cell synthesis)**	R51	14	14.00
2-KG <-> Suc-COA	R26	62	62.00	**2KG -> (Cell synthesis)**	R52	9	9.00
				CO2 -> (Evolution)	R53	274	274.00

Table 14. Fluxes, measured as % of glucose uptake, at simulated steady-state from EC Model 1 compared with experimental fluxes from Ishii *et al*. (2007). Results show a near-perfect agreement between simulated and experimental data at steady-state.

5.4.5 EC Model 2

In EC Model 1, the gene expression process was excluded from the model. Transcription and translation rates, together with degradation rates of mRNA and proteins, have showed to exhibit a degree of control over the concentration of protein (Garcia-Martinez *et* al., 2007). To increase our understanding of the regulation of genes and mRNA on metabolism, the integration of both genes and mRNA data needs to be carried out.

In EC Model 2, after parameter estimation was completed, CellDesigner was used to perform a simulation using the SOSlib solver. Again, the concentration of metabolites, protein expression levels and fluxes show a near-perfect agreement between our model data and the experimental data from Ishii *et* al. (2007) at steady-state as shown in tables 15, 16 and 17 respectively. These results validate our model on training data and as such these results were expected.

Concentration (in mM)		
Species	**Experimental Data**	**Predicted Data**
Glucose	6.33E-03	0.01
PEP	0.01	0.00
G6P	0.17	0.17
PYR	0.1852	0.18
F6P	0.01	0.01
F16P	0.045	0.04
DHAP	0.0312	0.02
G3P	0.01	0.01
PGP	0.01	0.01
3PG	0.0083	0.08
2PG	0.7463	0.64
AcCoA	0.01	0.01
CO2	0.01	0.01
Gluconolactone-6P	0.01	0.01
6PG	0.01	0.01
Ru5P	0.01	0.01
2-KDPG	0.0983	0.10
X5P	0.01	0.01
R5P	0.01	0.02
S7P	0.0127	0.01
E4P	0.178	1.96
OAA	0.01	0.01
CIT	0.01	0.01
ICT	0.01	0.01
2-KG	0.01	0.00
Suc-COA	0.0301	0.03
SUC	0.01	0.01
FUM	0.0686	0.07
MAL	0.0702	0.07
Glyoxylate	0.0835	0.08
Ac-P	0.01	0.01
Acetadehyde	0.01	0.01

Table 15. Concentration of metabolites (in Mm) for EC Model compared with experimental data from Ishii *et al*. (2007)

Protein level		
Proteins	Experimental Data	predicted Data
GalM	0.04	0.04
Glk	1	1.00
Pgi	0.22	0.22
Pfk(AB)	0.0582	0.06
FbaA	1.39	1.39
TpiA	0.4	0.40
GapA	2.97	2.97
Pgk	0.53	0.53
Gpm(AB)	0.97	0.97
Eno	1.24	1.24
Pyk(AF)	0.5125	0.51
Ace(EF)	2.1767	2.18
Zwf	0.04	0.04
Pgl	0.09	0.09
Gnd	0.06	0.06
Edd	1	1.00
Rpe	0.05	0.05
RpiA	0.03	0.03
Tkt(AB)	0.6604	0.66
Tal(AB)	0.2976	0.30
Eda	0.13	0.13
GltA PrpC	2.1189	2.11
AcnB	0.92	0.92
IcdA	3.31	3.33
Suc(AB) LpdA	3.1785	3.18
Suc(CD)	0.61	0.61
Sdh(AB) FrdA	1.1632	1.16
Fum(ABC)	0.5763	0.58
Mdh	0.4	0.40
PckA Ppc	1.6369	1.64
SfcA	1	1.00
AceA	23.22	23.16
AceB GlcB	6.6989	6.68
Pta	0.06	0.06
AckA	0.88	0.88
LdhA	0.24	0.24
PoxB	1	1.00
PflB	0.87	0.87
AdhE	0.18	0.18

Table 16. The expression level of proteins in EC Model 2 compared with experimental data at steady-state. Protein level is measured in mg-protein/g-dry cell weight.

Reactions	Experimental	predicted	Reactions	Experimental	predicted
	Data	Data		Data	Data
Glu_ex <--> Glucose	100	99.53	SUC <-> FUM	77	77.00
Glucose + PEP <-> G6P + PYR	100	99.98	FUM <-> MAL	77	77.00
G6P <-> F6P	78	78.00	MAL <-> OAA	89	89.00
F6P <-> F16P	85	85.00	PEP + CO2 <-> OAA	11	11.00
F16P <-> DHAP + G3P	85	85.00	MAL <-> PYR + CO2	3	3.00
DHAP <-> G3P	85	84.99	ICT <-> Glyoxylate + SUC	15	15.00
G3P <-> PGP	172	171.99	Glyoxylate + AcCoA <->MAL	15	15.00
PGP <-> 3PG	172	172.00	AcCoA <-> Ac-P	0	0.00
3PG <-> 2PG	162	162.00	Ac-P <-> Acetate_ex	0	0.00
2PG <-> PEP	162	162.01	PYR <-> Lactate_ex	0	0.00
PEP <-> PYR	47	47.00	PYR <-> Acetate_ex	0	0.00
PYR <-> AcCoA + CO2	129	129.00	PYR <-> AcCoA + Formate_ex	0	0.00
G6P <-> Gluconolactone-6P	20	20.00	AcCoA <-> Acetadehyde	0	0.00
Gluconolactone-6P <-> 6PG	20	20.00	Acetadehyde <-> Ethanol_ex	0	0.00
6PG <-> Ru5P + CO2	20	19.91	G6P -> Cell Synthesis	2	2.00
6PG <-> 2-KDPG	0	0.00	F6P -> (Cell synthesis)	1	1.00
Ru5P <-> X5P	8	7.98	R5P -> (Cell synthesis)	6	6.00
Ru5P <-> R5P	12	11.96	E4P -> (Cell synthesis)	4	4.00
X5P + R5P <-> S7P + G3P	6	6.00	G3P -> (Cell synthesis)	0	0.00
S7P + G3P <-> E4P + F6P	6	6.01	3PG -> (Cell synthesis)	10	10.00
X5P + E4P <-> F6P + G3P	2	2.00	PEP -> (Cell synthesis)	4	4.00
2-KDPG <-> G3P + PYR	0	0.00	PYR -> (Cell synthesis)	21	21.00
AcCoA + OAA <-> CIT	86	86.00	AcCoA -> (Cell synthesis)	28	28.00
CIT <-> ICT	86	86.00	OAA -> (Cell synthesis)	14	14.00
ICT <-> 2-KG + CO2	71	71.00	2KG -> (Cell synthesis)	9	9.00
2-KG <-> Suc-COA	62	62.00	CO2 -> (Evolution)	274	274.00
Suc-COA <-> SUC + CO2	62	62.01			

Table 17. Fluxes from EC Model 2 compared with experimental data from Ishii *et al*. (2007). Fluxes are measured as % of glucose uptake.

5.4.6 Genetic Perturbations – Validation on untrained Data

In Ishii *et al.*(2007), the authors also carried out 24 single-gene disruptions in an attempt to analyse the effects of genetic perturbation. The genes selected cover most of the glycolysis and pentose phosphate pathways. The disrupted cells were grown at the same fixed dilution rates as the wild-type at 0.2 hours $^{-1}$ in glucose-limited chemostat cultures. This allowed for comparisons between the wild-type and the disrupted cells. To further validate our models of E. coli central metabolism, 14 single-gene disruptionsor protein knock-downs were simulated and the results were compared with their respective experiment in the Ishii dataset. For EC Model 1, the concentration of proteins was changed to that observed in the experimental data as no gene expression process is available for this model. However, in EC Model 2, the amount of protein under a gene disruption experiment is calculated using equation (6) where the expression level of mRNA is obtained from the same experimental data under the same condition. Here, it is worth mentioning that kinetic parameters in the both models were not re-trained using any disrupted experimental dataset.

Figure 28 shows the changes in metabolites concentration under the different protein knock-down experiments from both EC Model 1 and 2. When compared with the same experiment from the Ishii dataset a few discrepancies are observed. In EC Model 1, there are more metabolites (G6P, F6P, E4P, Mal, Glyoxylate) which end up being over-produced in many of the protein knockdown experiments. This number is considerably reduced in EC Model 2 which might suggest that the model as a whole is more robust to perturbation than EC Model 1 due to the integration mRNA and gene regulation. In Ishii *et al.*, it was found that metabolites concentrations did not change significantly which could be a result of the regulation of enzyme concentration. The over-production of E4P and DHAP could be due to some reactions which allow for re-routing of metabolites in a biological system being omitted. In our model, proteins are not linked with each other which means that a decrease in the amount of one reaction has no effect on the concentration level of other proteins. For example, if an in-flux of a reaction is 100 mmol/gDW and out-flux is 100 mmol/gDW, then the systems is seen as being in a steady state. However, if the same in-flux was to be reduced to 60 mmol/gDW as a result of protein knockdown while the out-flux remains the same, then an imbalance is created in the system. In Figure 30 it can be observed that changes in protein levels in our models do not have any significant impact on the protein levels of other proteins. Further, effects of the lack of protein-protein association can explain the abnormal effect of R7 and R9 in fluxes in Figure 31. In the experimental data (Figure 30), it is evident that a single-gene disruption can cause the concentration of other proteins to change appropriately to maintain stability. It is known that the presence of alternative routes and isoenzymes enable enzyme capacity to be expressed through the cell (Ishii *et al*., 2007). This indicates a form of indirect regulation amongst proteins in a cell.

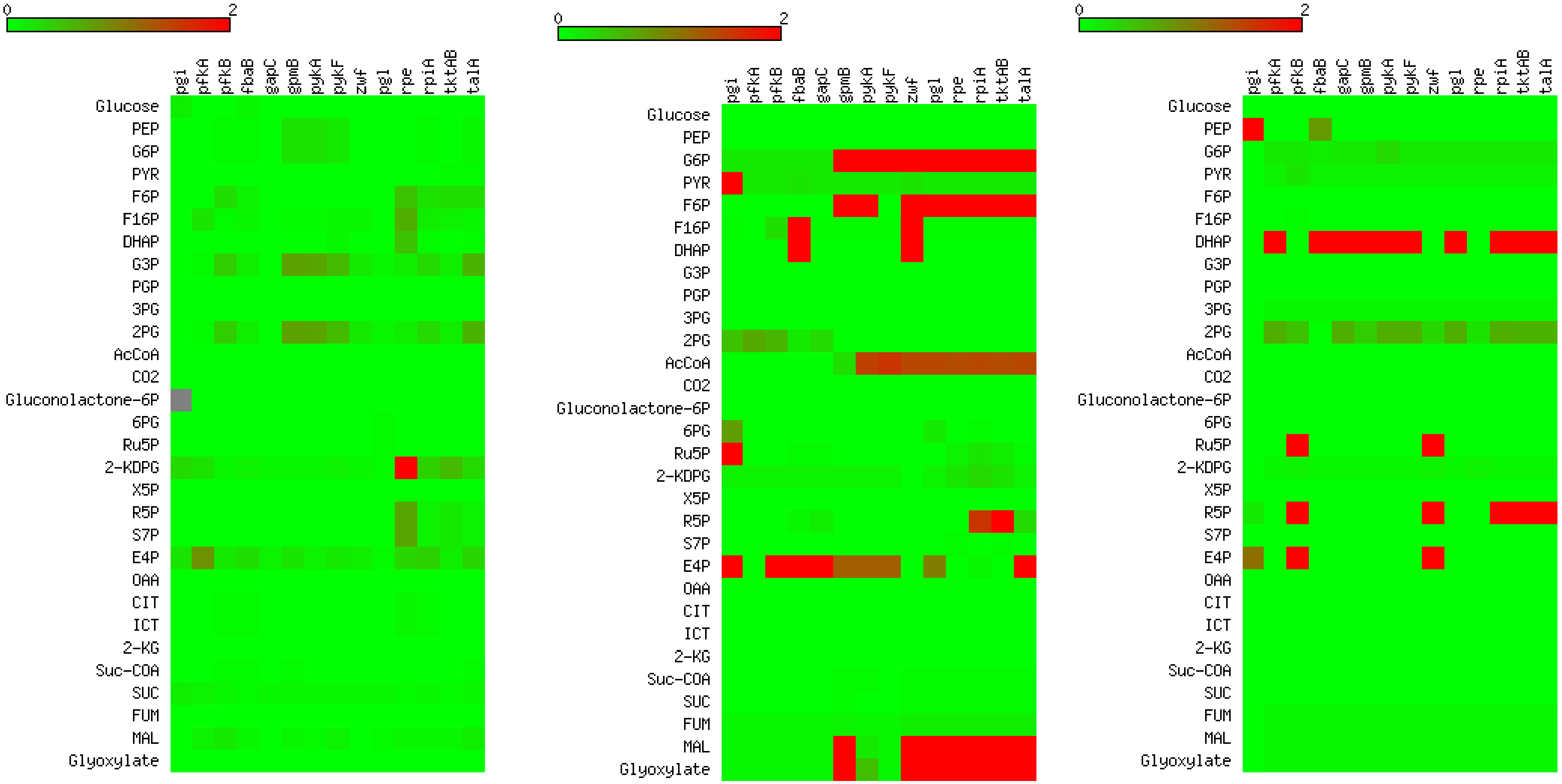


Figure 29: Heat map showing the concentration of metabolites (in rows) under each protein knock-down experiement (in columns). The heap map on the far left is the experimental data from Ishii *et al.* (2010), heat map in the middle and right show changes in the concentration of metabolites in EC Model 1 and EC Model 2 respectively. Concentration of metabolites is measured in mM. Heat maps were created using matrix2png (Pavlidis and Noble, 2003).

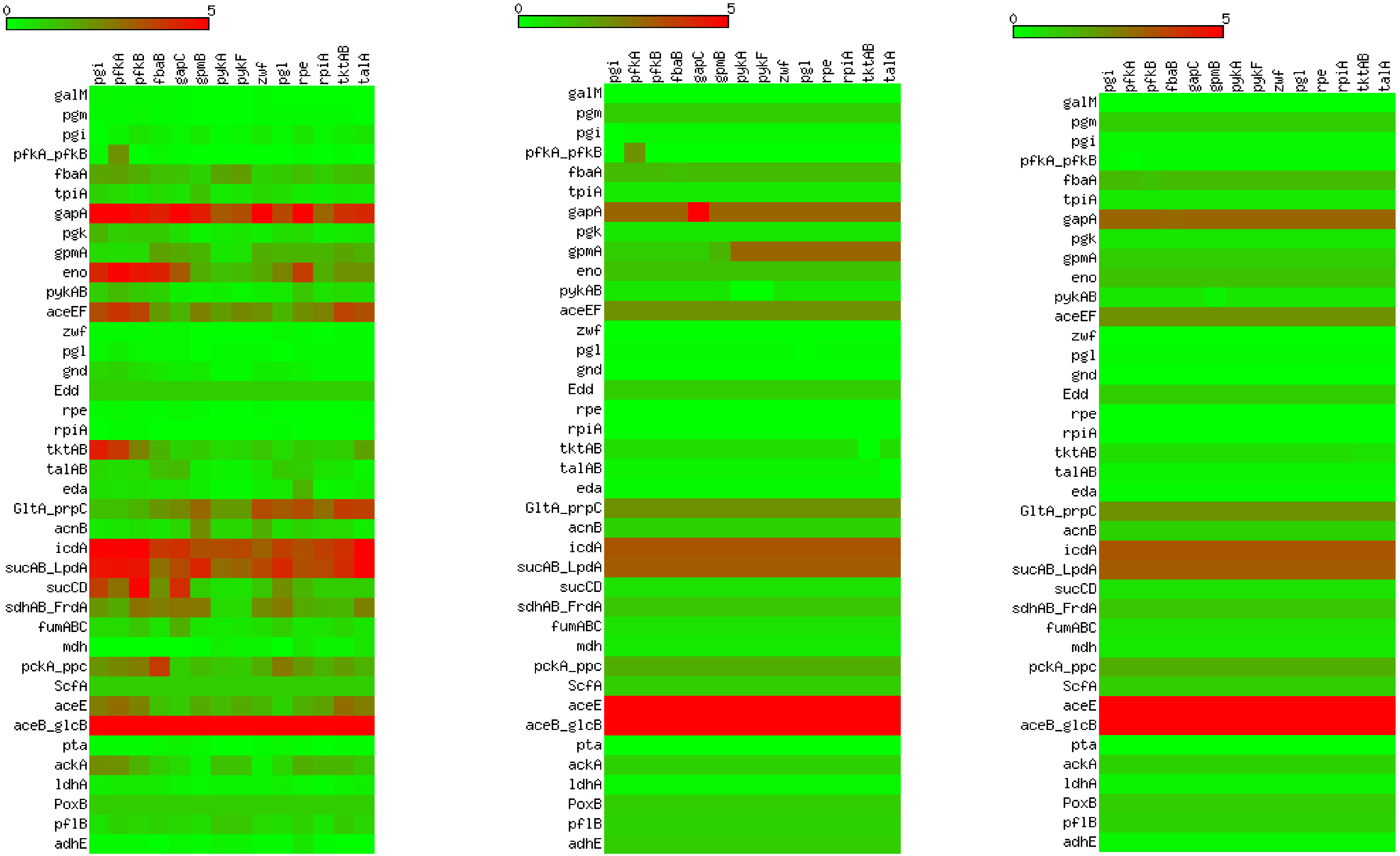
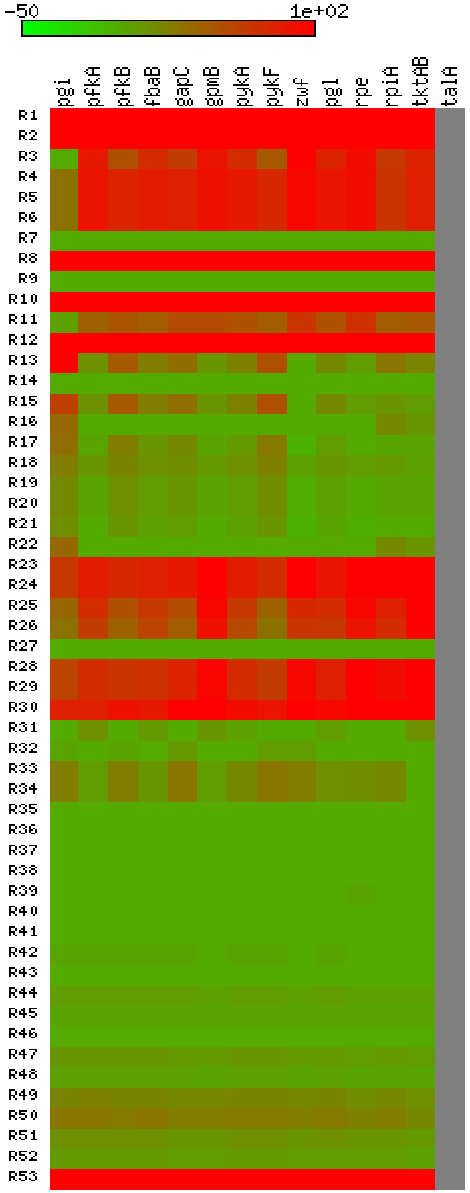
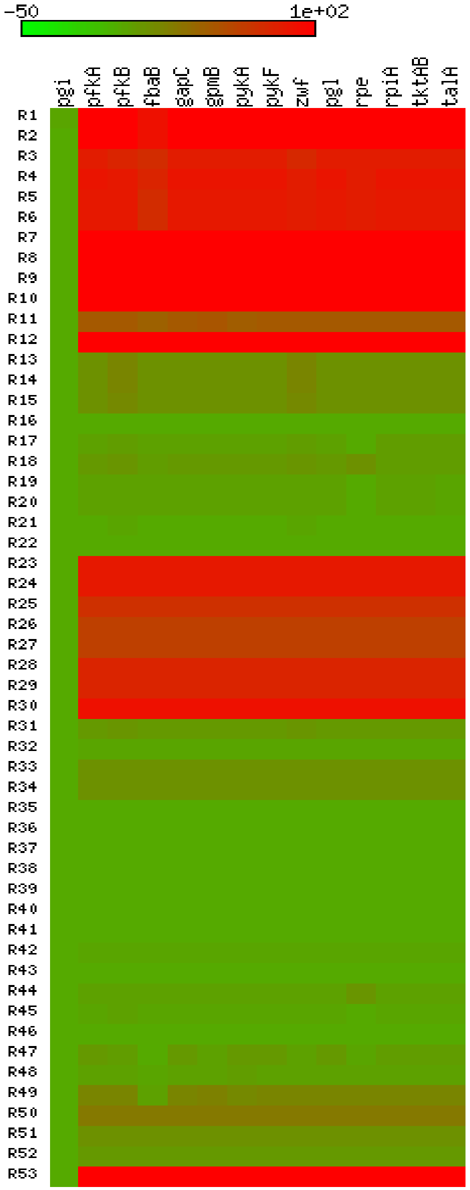


Figure 30: Heat map showing the expression levels of proteins (in rows) under protein knock-down experiement (in columns). The heap map on the far left is the experimental data from Ishii *et al.* (2010), heat map in the middle and right show changes in the amount of proteins in EC Model 1 and EC Model 2 respectively. Protein level is measured in mg-protein/g-dry. Heat maps were created using matrix2png (Pavlidis and Noble, 2003).


-50
1e+02
pgi
pfkA
pfkB
fbaB
gapC
gpmB
pykA
pykF
zwf
pgl
rpe
rpiA
tktAB
talA
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39
R40
R41
R42
R43
R44
R45
R46
R47
R48
R49
R50
R51
R52
R53


-50
1e+02
pgi
pfkA
pfkB
fbaB
gapC
gpmB
pykA
pykF
zwf
pgl
rpe
rpiA
tktAB
talA
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39
R40
R41
R42
R43
R44
R45
R46
R47
R48
R49
R50
R51
R52
R53

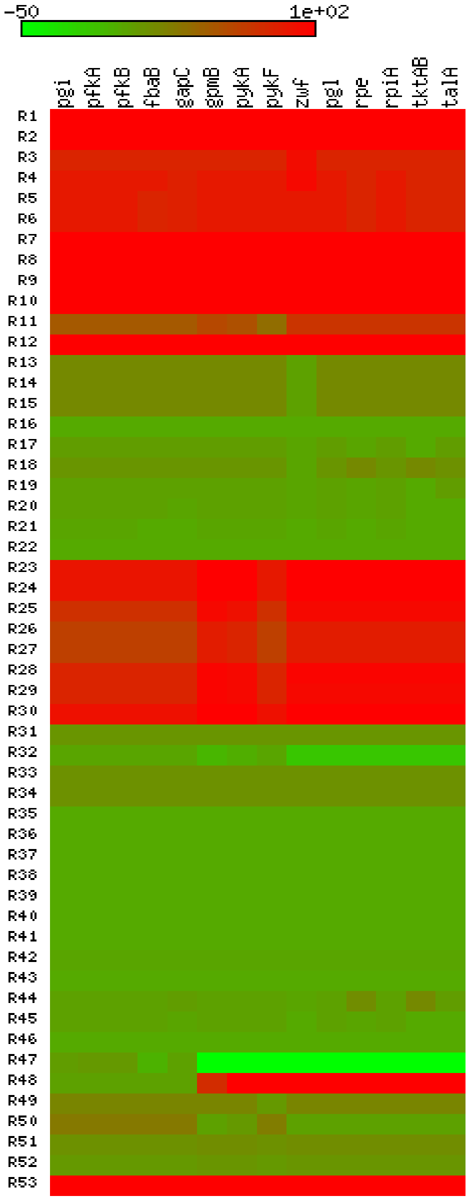

-50
1e+02
pgi
pfkA
pfkB
fbaB
gapC
gpmB
pykA
pykF
zwf
pgl
rpe
rpiA
tktAB
talA
R1
R2
R3
R4
R5
R6
R7
R8
R9
R10
R11
R12
R13
R14
R15
R16
R17
R18
R19
R20
R21
R22
R23
R24
R25
R26
R27
R28
R29
R30
R31
R32
R33
R34
R35
R36
R37
R38
R39
R40
R41
R42
R43
R44
R45
R46
R47
R48
R49
R50
R51
R52
R53

Figure 31: Fluxes measured in mmol/gDW in rows through the *E. coli* central metabolic system under protein knock-down experiement (in columns). The heap map on the left shows the fluxes observed under gene distruptant experiements in Ishii *et al.* (2010), heat map in the middle and right show changes in the fluxes in EC Model 1 and EC Model 2 respectively. EC Model 2 is observed as being more robust as flux for R47, an exchange reaction, remains positive while a negive flux is observed in EC Model 1 under different protein knockdown experiements. Heat maps were created using matrix2png (Pavlidis and Noble, 2003).

In our analysis of fluxes under each knockdown experiment, again it is more evident that EC Model 2 which includes the integration of gene expression performs better than our model without gene expression. In EC Model 2, the fluxes for all exchange reactions that exported metabolites outside the system boundary were positive compared with negative fluxes observed for reactions R47 and R48 (G3P and PEP respective export reactions) in EC Model 1 (Figure 31)

With the exception of the Pgi knockdown in EC Model 2, which reduced fluxes through the glycolysis pathway and subsequently caused most metabolites and fluxes in the system to go down to zero, there was a positive agreement between fluxes from our models and the Ishii datasets. This shows that our modelling framework can be used in a predictive manner to estimate the effects of perturbations in an integrated system of genes, proteins and metabolic reactions

5.4.7 Parameter Variability Analysis (PVA)

We performed parameter variability analysis (PVA) to establish the degree of redundancy in our estimated kinetic parameters. The results of the PVA indicate that the forward reaction rate parameters (K_A or V^+) are the most constrained in our models (Figure 32). The backward reaction rate parameters (K_P or V^-) are observed as being the least constrained. Michaelis constants and dissociation constants tend to be scattered but overall, a high degree of variability was observed among those parameters. The observed summed squared mean error between the experimental data and simulated data was in the range of 1.0E-12 to 3.4E-5 for all repeats of the PVA.

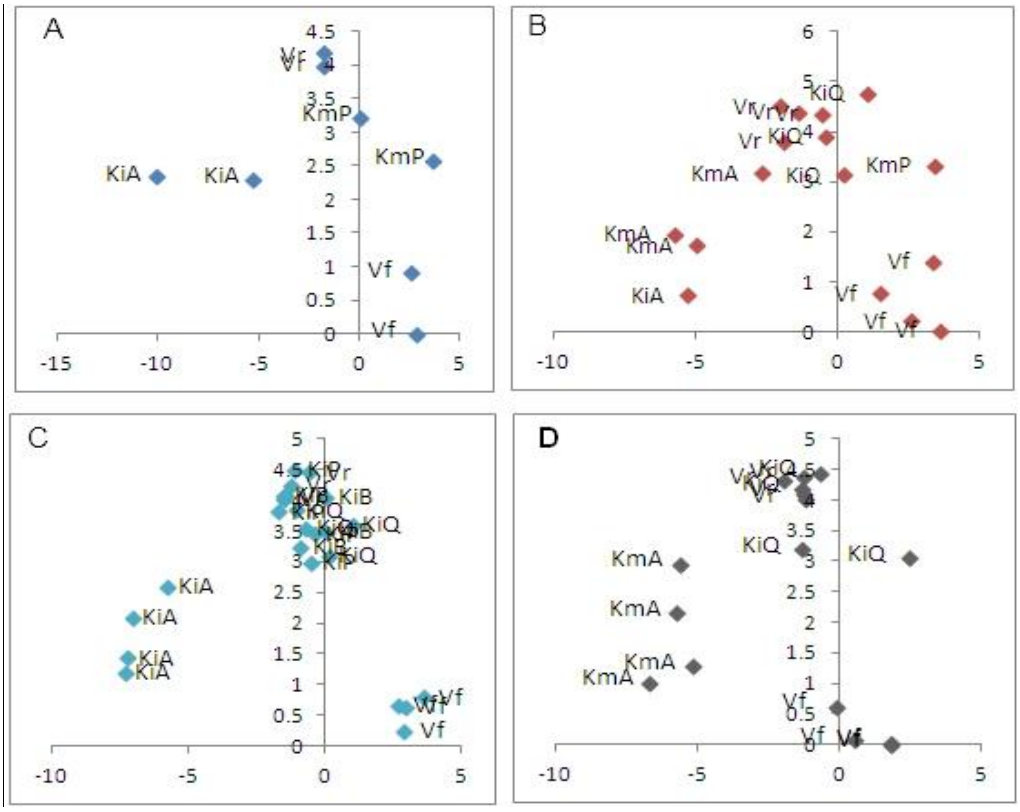


Figure 32. Scatter graph showing the distribution of parameter values obtained from parameter variability analysis (PVA). For PVA, parameter estimation was repeated 100 times for each reaction of the *E. coli* central carbon metabolism. The value of each parameter value was converted to $\log_{10}$. The average for each parameter value in log10 over the 100 runs was then plotted against its standard deviation (STD dev). The most constrained parameters are substrate turnover rate or the forward reaction rate (K_A or V^+) and the least constrained parameters are product turnover rate or the backward reaction rate (K_P or V^-). (For details of each parameter value, see on our lab group page (pg. 148). (Figure **A** shows PVA for uni-uni; **B** shows PVA for uni-bi; **C** shows PVA for bi-bi and **D** shows PVA for bi-uni reactions).

The high remaining degree of redundancy observed by PVA confirms that even a rich dataset containing proteomic, metabolomic and flux data is not sufficient to eliminate redundancy in the parameterisation of a model. This property was already observed by Gutenkunst *et al*. (2007) and Little *et al*. (2009). A possible solution to reduce redundancy in parameterisation will be to incorporate transcriptomic data in the parameter estimation process.

5.5 Discussion and Conclusion

The use of mathematical models to predict biological states is a major component of systems biology (Kitano, 2002). These models enable *in silico* replication and perturbation of biological systems as demonstrated by numerous examples (Förster and Palsson 2003, Jamshidi and Palsson 2008, Ao *et al*. 2008; Feala *et* al., 2008; Li *et* al., 2010; Adiamah *et al*., 2010; Ruppin *et al*., 2010). In the last few years, the constraint-based modelling (CBM) approach has been considerably used for the prediction and analysis of biological systems (Edwards and Palsson, 2000; Covert *et al*., 2001) as it does not require the knowledge of detailed rate equations and kinetic parameters in defining a biological model. Using CBM, the phenotypic properties of an organism can be analysed and successful prediction rates between 70-90% can be achieved in some experiments (Price *et al*., 2003). However, CBM models cannot capture the dynamic properties of biological systems. Kinetic modelling was long hampered by insufficient knowledge of kinetic parameter values and detailed knowledge of enzymatic-kinetic rate laws. As a result, efforts are being made to predict the behaviour of biological systems using estimated parameter values rather than detailed kinetic models (Ao *et al*., 2008; Liebermeister and Klipp, 2006; Adiamah *et al*., 2010; Liebermeister *et al.,* 2010). This approach requires the utilisation of parameter estimation techniques to determine kinetic parameters based on time series of experimental data, as it is usually too expensive and time-consuming to measure every parameter individually in *in-vitro* experiments. Furthermore, there is no guarantee that parameter values measured *in vitro* will still be relevant to physiological *in vivo* conditions.

With the availability of modern '*omics*' data for many organisms, it is time to start building integrative large-scale biological models which are able to integrate fluxomic, metabolomic, proteomic and genomic data. In a previous study, we presented GRaPe (Adiamah *et al*., 2010), a computational tool as a platform for building such integrative models, reducing the time and effort needed to build such models manually (Förster and Palsson 2003, Jamshidi and Palsson 2008, Ao *et al*. 2008). Additionally, a method for modelling an integrative gene expression and metabolism process using ordinary differential and generic rate equations was presented. We applied our methodology to the yeast glycolysis pathway and showed that the dynamical behaviour of biological systems was reproducible without detailed knowledge of enzymatic-kinetic rate laws and accurately measured parameters. However, as no genomic and proteomic data was available for the yeast glycolysis example, genes were excluded from our model and Boolean values of 1 or 0 were used to represent the expression level of an enzyme.

In this study, we present an integrative model of *E. coli* K-12 central metabolism by integrating genomic, proteomic, metabolomic and fluxomic data based on a dataset by Ishii *et al*. (2007). Two models were constructed, one with the integration of proteomic, metabolomic and fluxomic data and the other with genomic, proteomic, metabolomic and fluxomic. The two models were built in an attempt to determine whether the *E. coli* metabolic network became more robust to perturbations with the integration of genomic data and also ascertain the degree of control exerted by genes and proteins in metabolism. Our results showed that the behaviour of a system can be achieved using generic rate and ordinary differential equations with or without genomic data. The models were perturbed *in silico* to determine the predictive value of our model building approach. Our results showed that in most cases our models are able to predict the major features of gene knockout experiments. Overall, the model incorporating gene expression showed better predictive value and was able to avoid major discrepancies such as negative fluxes in the system. Nevertheless changes in protein levels were observed to have no effect on other proteins in the system, which points to the need of further integrating feedback loops or protein-protein interaction networks to model the effects of protein regulation on other proteins.

Parameter estimation has become increasingly important in systems biology as parameters needed to define complex biological models are still lacking, and it is generally impossible to measure all these parameters *in vivo* or *in vitro.* One of the major issues of parameter estimation is that more experimental data is usually required to constrain the values of unknown parameters in a network (Heinzle *et al.,* 2007). It has also been shown in previous studies that even fully parameterised models of biological systems exhibit a degree of “sloppiness” or redundancy (Gutenkunst *et al*., 2007; Little *et al.,* 2009). It is therefore important to estimate the degree of redundancy in parameter estimation, which helps to define the parameters that are most important to constrain the system. Our results from parameter variability analysis showed that for each reaction in our model, the forward reaction rate parameter was the most constrained parameter. Interestingly, the backward reaction rate parameter was seen as being the least constrained parameter in our models.

We have presented an approach for building integrative biological dynamic models. As more high-throughput data becomes available, efforts are needed towards integrating these datasets into comprehensive models of growing precision that are capable of predicting biological outcomes when perturbed *in silico*. In an iterative systems biology approach, the results of each modelling step should be compared to experimental data and are expected to be in accordance with part of the data. But at the same time, it is the discrepancies that generally point us towards the next step needed to improve the precision of the model. In this study, this process has revealed the importance of integrating protein-protein interactions into an integrative modelling building framework to further improve the precision and predictive power of such models in the future.

5.6 References

1. Adiamah, D.A., Handl, J. and Schwartz, J.-M. (2010). Streamlining the construction of large-scale dynamic models using generic equations. *Bioinformatics*. **26**. 1324-1331.

2. Ao, P, et al (2008). Towards kinetic modeling of global metabolic networks: Methylobacterium extorquens AM1 growth as validation. *Chin. J. Biotechnol*. **24**. 980-994.

3. Ashyraliyev, M., Fomekong-Nanfack, Y., Kaandorp J.A., Blom J.G. (2009). Systems biology: parameter estimation for biochemical models. *FEBS J* 2009. **276**(4). 886-902.

4. Baba T, Ara T, Hasegawa M, Takai Y, Okumura Y, Baba M, Datsenko KA, Tomita M, Wanner BL, Mori H. (2006). Construction of Escherichia coli K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol*. **2**. 2006.0008.

5. Baldazzi V, Ropers D, Markowicz Y, Kahn D, Geiselmann J, et al. (2010) The Carbon Assimilation Network in *Escherichia coli* Is Densely Connected and Largely Sign-Determined by Directions of Metabolic Fluxes. *PLoS Comput Biol.,* **6**(6): e1000812.

6. Barrett, C.L., C., Kim, Y. T., Kim, H.U., Bernhard Palsson, B.Ø. and Lee, S.Y. (2003). Systems biology as a foundation for genome-scale synthetic biology. *Current Opinion in Biotechnology*. **17**. 488-492.

7. Bruggeman, F. J. and Westerhoff, V. H. (2007). The nature of systems biology. *Trends in Microbiology*. **15**. 45 – 50.

8. Çakir, T., Kiran Raosaheb Patil, P. K., Önsan, I.Z., Ülgen, Ö. K., Kirdar, B. and Nielsen, J. (2006). Integration of metabolome data with metabolic networks reveals reporter reactions. *Molecular Systems Biology*. **2**. 50.

9. Cavalieri, D. and Filippo, C.D. (2005) Bioinformatic methods for integrating whole-genome expression results into cellular networks. *Drug Discovery Today*. **10**. 727-734.

10. Chen, T., He, H.L. and Church, G.M. (1999) Modeling Gene Expression with Differential Equations. *Pacific Symposium on Biocomputing*. **4**. 29-40.

11. Chou, I-C. and Voit, O. E. (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. *Mathematical Biosciences*. **219**. 57-83.

12. Covert, M., Schilling, C. H., Palsson, B.Ø. (2001) Regulation of Gene Expression in Flux Balance Models of Metabolism. *Journal of Theoretical Biology*. **213**(1)*: 73-88.*

13. Edwards, J.S. and Palsson, B.Ø. (2000) Metabolic flux balance analysis and the in silico analysis of *Escherichia coli* K-12 gene deletions. *BMC Bioinformatics*. **1**.1.

14. Feala, J. D., Coquin, L., Paternostro, G. And McCulloch, D.A. (2007) Integrating metabolomics and phenomics with systems models of cardiac hypoxia. *Progress in Biophysics and Molecular Biology*. **96**. 209-225.

15. Feist A.M, Henry C.S, Reed J.L, Krummenacker M, Joyce A.R, Karp P.D, Broadbelt LJ, Hatzimanikatis V, Palsson BØ. (2007. A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol*, **3**:121.

16. Fellenberg, M. (2003) Developing integrative bioinformatics systems. *BIOSILICO*. **1**. 177-183.

17. Förster, J., Famili, I., Palsson, B.Ø., and Nielsen, J. (2003). Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae. *OMICS*. **7**:193-202.

18. Fujisaki S, Takahashi I, Hara H, Horiuchi K, Nishino T, Nishimura Y (2005). Disruption of the structural gene for farnesyl diphosphate synthase in Escherichia coli. *J Biochem (Tokyo)*, **137:**395-400.

19. Funahashi A, et al (2003). CellDesigner: a process diagram editor for gene-regulatory and biochemical networks. *BIOSILICO*. **1.** 159-162.

20. Garcia-Martinez J, et al., (2007). Common gene expression strategies revealed by genome-wide analysis in yeast. *Genome Biology*. **8.** R222.

21. Gevorgyan, A., Poolman, M.G. and Fell, D.A. (2008) Detection of stoichiometric inconsistencies in biomolecular models. *Bioinformatics*. **24**. 2245 – 2251.

22. Goel G, Chou IC, Voit EO. (2008). System estimation from metabolic time-series data. *Bioinformatics.* **24**(21). 2505-2511.

23. Goss, J. E. P. and Peccoud, J. (1998). Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets. *PNAS*. **12.** 6750 – 6755.

24. Gutenkunst R.N. et al. (2007) Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput. Biol.* **3**:189.

25. Heinzle, E., Yang, H.T. and Deshpande, R. (2007). Analysis and design of metabolic networks - experiments and computer simulation. *Computer Aided Chemical Engineering*. **24**. 925-926.

26. Herrgård M. J., Fong S. S. and Palsson B. Ø. (2006). Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol* , **2**:e72.

27. Ideker, T. *et al.* (2001). Integrative Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. Science. **292,** 929 – 934.

28. Imielinski M, Belta C, Halasz A, Rubin H. (2005). Investigating metabolite essentiality through genome-scale analysis of *Escherichia coli* production capabilities. *Bioinformatics*, **21**:2008-2016.

29. Ishii, N. et al. (2007). Multiple High-Throughput Analyses Monitor the Response of *E. coli*. to Pertubations. *Science*. **316**, 593 – 597.

30. Joyce, A. R. and Palsson B. Ø. (2006). The model organism as a system: integrating 'omics' data sets. *Nat Rev Mol Cell Biol* , **7:**198-210.

31. Kell D. B and Knowles, J. (2006) *The role of modeling in systems biology*. *In System modeling in cellular biology: from concepts to nuts and bolts*. Edited by Szallasi Z, Stelling J, Periwal V. Cambridge: MIT Press. 3-18.

32. King E.L. and Altman C. (1956) A schematic method of deriving the rate laws for enzyme-catalyzed reactions. *J. Phys. Chem*., **60,** 1375-1378.

33. Kitano, H. (2002) Systems Biology: A Brief Overview. *Science*. **295**, 1662-4.

34. Le Novère N, Hucka M, Mi H, Moodie S, Schreiber F, Sorokin A, Demir E, Wegner K, Aladjem MI, Wimalaratne SM, Bergman FT, Gauges R, Ghazal P, Kawaji H, Li L, Matsuoka Y, Villéger A, Boyd SE, Calzone L, Courtot M, Dogrusoz U, Freeman TC, Funahashi A, Ghosh S, Jouraku A, Kim S, Kolpakov F, Luna A, Sahle S, Schmidt E, Watterson S, Wu G, Goryanin I, Kell DB, Sander C, Sauro H, Snoep JL, Kohn K, Kitano H. (2009).The Systems Biology Graphical Notation. *Nat Biotechnol.* **27**(8). 735-41.

35. Li *et al*., (2010). Systematic integration of experimental data and models in systems biology. *BMC Bioinformatics.* **11.** 582.

36. Liebermeister, W. and Klipp, E. (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraint. *Theoretical Biology and Medical Modelling,* **3,** 41.

37. Liebermeister, W., Uhlendorf, J., Klipp, E. (2010) Modular rate laws for enzymatic reactions: thermodynamics, elasticities and implementation. *Bioinformatics***, 26,** 1528-1534.

38. Little MP, Heidenreich WF, Li G (2009) Parameter identifiability and redundancy in a general class of stochastic carcinogenesis models. *PLoS ONE*. **4.** e8520.

39. Ma H, Zeng A-P (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*. **19.** 270-277.

40. Mendes P, Kell D. (1998) Non-linear optimization of biochemical pathways: application to metabolic engineering and parameter estimation. *Bioinformatics*, **14**(10). 869-883.

41. Moles C.G, Mendes P, Banga J.R. (2003) Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Res*. **13**(11). 2467-2474.

42. Palsson, B.Ø. (2006). *Systems Biology: Properties of Reconstructed Networks*. Cambridge University Press, Cambridge.

43. Pavlidis, P. and Noble W.S. (2003) Matrix2png: A Utility for Visualizing Matrix Data. *Bioinformatics,* **19**. 295-296.

44. Pease, A. C., et al. (1994) Light-Generated Oligonucleotide Arrays for Rapid DNA Sequence Analysis" *Proc. Natl. Acad. Sci. USA* **91**: 5022-5026, (1994).

45. Price, N.D., Papin, J.A., Schilling, C.H. and Palsson, B.Ø. (2003) Genome-scale microbial *in silico* models: the constraints-based approach. Trends in Biotechnology. **21**. 162-169.

46. Puchalka J and Kierzek AM. (2004) Bridging the gap between stochastic and deterministic regimes in the kinetic simulations of the biochemical reaction networks. *Biophysical Journal.* **86**. 1357–1372.

47. Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A. and Schwartz, J.-M. (2009). Reconstruction of an *in silico* metabolic model of Arabidopsis thaliana through database integration. *Available from Nature Precedings*, DOI: 10101/hpre.2009.3309.1.

48. Reed JL, Vo TD, Schilling CH, Palsson B Ø. (2003) An expanded genome-scale model of *Escherichia coli* K-12 (iJR904 GSM/GPR). *Genome Bioli*ogy, **4.** R54.

49. Reed JL, Famili I, Thiele I, Palsson B Ø. (2006) Towards multidimensional genome annotation. *Nat Rev Genet*, **7**:130-141.

50. Resat, H., Petzold, L. and Pettigrew, M.F. (2009) Kinetic Modelling of Biological Systems. *Methods Mol Biol.* **541**. 311-335.

51. Ruppin, E., Papin, A.J., Figueiredo, d. F. L. And Schuster, S. (2010). Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Current Opinion in Biotechnology*. **21**. 502-510.

52. Schena, M., et al. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. **270**. 467-470.

53. Sherlock, G., et al. (2001) The Stanford Microarray Database. Nucleic Acids Research. **29**. 152-155.

54. Smallbone, K., Simeonidis, E., Swainston, N. and Mendes, P. (2010) Towards a genome-scale kinetic model of cellular metabolism. *BMC Systems Biology*. **4**. 1.

55. Smolen P, *et al.,* (2003). Mathematical modeling of gene networks. *Neuron*. **26**. 567-580.

56. Vogel,C.,de Sousa Abreu, R., Ko, D., Le, S-Y., Shapiro, B.A., Burns, S.C., Sandhu, D., Boutz, D.R., Marcotte, E.M., and Luiz O Penalva, L.O. (2010) Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line. *Molecular Systems Biology,* 6. 400.

57. Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. and Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*. **26**. i225– i2560.

58. Zhang, W., Li, F. and Nie, L. (2010) Integrating multiple ‘omics’ analysis for microbial biology: application and methodologies. *Microbiology*. **156**. 287-301.

Chapter 6

General Discussion

It is now possible to construct genome-scale metabolic models due to the availability of high-throughput data. Integrating different biological data types could increase our understanding of complex biological systems. Consequently, efforts must be made in building such integrative models as we move towards whole cell models. However, one of the main challenges of building such integrative models is that there are no software tools that are purposefully built for the construction of such models. Usually, software tools require the manual construction of biological components and the description of reaction rate equations for each reaction in the network. This makes it time-consuming and tedious when building large-scale biological models and usually, detailed knowledge of reactions mechanisms is unknown.

We provide a solution to this shortfall in software tools for building integrative genome-scale models. We presented a software tool, GRaPe for constructing an integrative model of gene expression and metabolic reaction. The software generates the reaction rates for each reaction based on its number of substrates and products, thus, reducing the time and effort in writing out these equations manually. Since there is no limit to the amount of reactions that can be instantiated, we can use this software to reconstruct genome-scale metabolic networks. As the detailed knowledge of enzyme kinetics is unknown for many pathways, we use reversible generic rate equations based on random-order mechanism in describing the reaction rates in our models. Although, kinetic models are known to be able to capture the dynamical properties of biological systems, building these models is still hampered by the lack or unavailability of kinetic parameters needed to fully define them. As a result of this, we developed a genetic algorithm which uses metabolites and fluxes data in estimating kinetic parameters for our models. Our first experiment was to validate our modelling approach of using generic rate equations and Boolean values in predicting the dynamical state in yeast glycolysis pathway.

In the results section, we presented models of the yeast glycolysis pathway built using the GRaPe. The SBML file produced by the software was then simulated using CellDesigner and analysed. We showed that these models have the capability of successfully predicting distributions of fluxes and concentrations in several experiments when glucose is either increased or decreased. .

Secondly, we moved towards predicting steady-states for a large-scale model. In this experiment, a kinetic model of the genome-scale model of *Mycobacterium tuberculosis* was reconstructed. Again, we were successfully able to predict steady-states using our modelling approach on trained experimental data. However, we failed in finding a set of parameters capable of predicting various states without re-estimating kinetic parameters as done in our previous experiment. Admittedly systems biology is an iterative process where each round of models is expecting to reveal some shortcomings, enabling us to direct the next research steps towards new objectives. Flux predictions were sometimes inaccurate because of the incomplete specification of the boundary conditions of metabolites, fluxes and gene expression . A possible reason for the deviation in trends of metabolites and fluxes when predicting new states in the *Mycobacterium tuberculosis* model is that,the model which was built for FBA was inaccurate for kinetic modelling as exchange fluxes are defined differently. Another possible reason for this deviation is the lack of reliable kinetic parameters found by our genetic algorithm because of a insufficient training set; dynamic time series of experimental values may be needed to properly constrain the parameters, instead of independent steady states.

Consequently, we performed parameter variability analysis (PVA) to determine the level of redundancy in our model and also to determine the relationship between computing times when the number of points in the dataset is increased The result of our PVA suggest a high degree of redundancy in our estimated parameters. It is known that there is a degree of redundancy in biological systems that fundamentally limit the role of exact parameter values. Our results also indicate that as we increase the number of data points in the dataset for parameter estimation, the time taken to compute estimated parameters rises exponentially. This suggests that trying to improve the quality of the estimated parameters by increasing the number of data points in the dataset is costly and as such a different method of constraining parameter values is required.

Finally, we integrated gene expression and protein levels into our reconstruction of the *E. coli* central metabolic network. Our aim from the start was to develop an approach to building integrative gene expression and metabolic models. This model was to be our main validation of our modelling approach. Using GRaPe, we reconstructed the central metabolic network of *E. coli* using generic rate equations in describing the behaviour of reactions and ordinary differential equation in modelling gene expression. Our results showed an excellent agreement with experimental data after performing parameter estimation and *in silico* simulations. Our methodology was capable of predicting several distributions of metabolic concentrations and fluxes in gene deletion experiments. The results of PVA in this experiment corresponded with what was observed in the PVA from the *Mycobacterium tuberculosis* experiment. We observed that the forward reaction rate parameters are the most constrained while the backward reaction rate parameters tend to be the least constrained in our models. We have showed in three experiments of various metabolic sizes that our methodology is capable of predicting states on trained data but importantly also on unseen data. Nevertheless, there is still more work to be done to improve the predictive power of methodology. Furthermore, we hope to expand the software to be able to reconstruct metabolic pathways from databases and include other data for constraining our parameters.

Chapter 7

Conclusion

We have developed a new tool and presented a series of methods to meet the challenge of building large-scale integrative kinetic models. GRaPe, the new software tool, is a platform independent tool that allows for the streamlined construction of large-scale dynamic models. GRaPe also provides a platform for the construction of reaction-protein or gene-reaction-protein networks as we move towards building integrative biological models. A novel feature of GRaPe is its ability to generate generic rate equation for all reactions in a model automatically, irrespectively of its size. Another important feature is its capability to explicitly integrate gene expression processes or enzyme species into reactions, making it a convenient tool for the construction of integrative protein-reaction networks.

Our fundamental hypothesis was that exact rate equations and parameters are not always crucial to determine the main properties of cellular systems. We showed that by using generic rate equations and estimated parameters, the behaviour of a system can be predicted as shown in our case study of the yeast glycolysis pathway and *E. coli* central metabolic network. Furthermore, we demonstrated that by using Boolean values to indicate the expression level of proteins, it not only possible to reproduce the dynamical behaviour of a system on training data but also predict new biological outcomes when the system is perturbed *in silico* without re-estimating kinetic parameters.

Based on these assumptions, we created this new methodology which is able to address several limitations of previous tools, by:

1. Assigning generic rate equations to all reactions in a network based on the stoichiometry of the reactions without placing much emphasis on the accurate enzyme mechanism of reactions.
2. Using ordinary differential equations in the definition of gene expression processes, which means that the concentration of the enzyme can be explicitly included in rate equations, thereby making it easy to analyse the effects of changes in enzyme expression level on a metabolic reaction.
3. Integrating proteins, mRNA and gene expression levels into metabolic models without extensive knowledge of their enzyme kinetics.
4. Using a genetic algorithm which uses metabolites and flux data to estimate kinetic parameters for our models.

5. Using computationally estimated flux data to constrain kinetic parameter values when experimental data is unavailable

We showed that this approach not only successfully integrates different sets of high-throughput datasets into kinetic models, but can also make predictions (gene deletion and protein knockdown experiments). Finally, we showed that a biological model becomes more robust with the integration of different layers of biological data.

In conclusion, this method of modelling an integrative gene-protein-reactions using ordinary differential and generic rate equations allows for the prediction and analysis of biological outcomes *in silico* and represents a positive step towards the goal of building cell-wide integrative biological models.

Chapter 8

Future Work

In this study, we have demonstrated that dynamical cellular behaviours can be modelled by using generic rate and ordinary differential equations without detailed knowledge of rate equations and kinetic parameters in defining a biological model. There is often inadequate knowledge of enzyme-kinetic laws and their associated parameter values, and usually parameters obtained from literature are dependent on specific *in vitro* or *in vivo* experimental conditions. As such, one distinct advantage of using generic rate equations over constraint-based and detailed kinetic modelling approaches is that extensive knowledge of enzyme kinetic rate laws governing individual reactions in a pathway is not required in predicting or replicating the dynamical behaviour of a metabolic system. Another, advantage is that, constraint-based modelling usually requires an optimisation against an objective function which is not a requirement when using generic rate equations or kinetic modelling.

Nevertheless our method is not perfect and we identified a few limitations that will require further work.

8.1 Inclusion of Haldane Relationships in Parameter Estimation.

We showed that estimated kinetic parameters tend to have a high degree of redundancy in our parameter variability analyses. As only flux and metabolites data is used in constraining our parameters, this high degree of redundancy was to be expected. This means that methods and other techniques are required to reduce this level of redundancy in parameter estimation. One possible way of achieving this will be to relate enzyme kinetic parameters to the equilibrium constant of the reaction, thus creating an additional constraint between parameter values.

In a reversible reaction, both forward and backward reactions are expressed mathematically as (at steady-state):

(45)

Where *Vf* is the forward reaction rate, *Vr* is the backward reaction rate, *S* is the concentration of substrate, *P* is the product concentration, *KS* and *KP* are kinetic constants. ($Vf = V^{+}e_0$ *and* $Vr = V^{-}e_{0;}$ where e_0 is the concentration of the enzyme). At equilibrium, when the reaction is zero equates to:

(46)

where *Seq* and *Peq* are the equilibrium constants for the both the substrate and product. By rearranging equation (46) gives:

(47)

where *Keq* is the equilibrium constant of the reaction. This relationship is known as the Haldane relationship and describes the dependency between kinetic parameters at equilibrium state (Bisswanger, 2008). The inclusion of the Haldane relationship into equation (45), we get:

(48)

By incorporating the Haldane relationship into parameter estimation methods, we could reduce the redundancy in kinetic parameters. Equation (48) can be substituted with equation (45) used in our approach to account for the Haldane relationship. However, this method requires the knowledge of *Keq* constants or relationships that exist amongst enzymatic parameters from literature or the lab. On a genome-scale, such data will be difficult to obtain.

8.2 Thermodynamics Inclusion in Parameter Estimation Process

Another possible way of constraining kinetic parameters will be to use the laws of thermodynamics as biochemical reactions are bound by these laws. In our model building methodology, thermodynamic laws are ignored in an attempt to demonstrate that detailed knowledge of enzymatic parameters and mechanisms are not always necessary in predicting the behaviour of a biological system.

Here, the equilibrium constant, K_{eq}, can be calculated using the following equation.

$$\frac{}{} \tag{48}$$

where ΔG is the standard Gibbs free energy of the metabolic reaction, R is the gas constant and T is the absolute temperature. The relationship between K_{eq} and ΔG can be defined as follows:

When a reaction is at equilibrium, the rate of converting a substrate, A, into product, P, is zero implying that the ratio of A to P is fixed. K_{eq} = B/A = constant and ΔG = 0 at equilibrium.

Past and recent attempts have shown that kinetic parameters can be constrained using the laws of thermodynamics (Henry et al., 2006; Liebermeister and Klipp, 2006; Tan et al., 2010; Jenkinson and Goutsias, 2011). In Jenkinson and Goutsias, the authors introduced a novel method for estimating thermodynamically feasible kinetic parameters for biochemical systems and suggested that their approach of estimating kinetic parameters based on thermodynamic laws reduces computational complexity, dimensionality and data overfitting. In the examples given above, the authors have had to rely extensively on literature and databases in obtaining Gibbs free energies for enzymes for their approach to work.

8.3 Connecting GRaPe to Online Databases

Currently, GRaPe is not connected to any online database or repository to download biological models. As a consequence, users have to obtain the models before uploading it into GRaPe.

CellDesigner allows for the downloading of models from the BioModel database and means that models can be analysed in a short space of time. COPASI now allows for MIRIAM compliant annotation when building biological models making it easy in maintain consistency of naming biological components such as compartment, reactions, and species across models

To reduce time taken in building models and maintain consistency of models, GRaPe will have to be able to connect to various databases to download models. As GRaPe is mainly concerned with building kinetic models, databases such as KEGG, BioModels and SABIO-RK should be the main focus as they store relevant models, and kinetic parameters can be obtained from the SABIO-RK database.

8.4 Protein-Protein Interactions (PPIs)

We showed that our approach is capable of predicting biological outcomes when the system is perturbed *in silico*. However, one of the main limitations of our approach was that protein levels remained fixed in our gene deletion and protein knockdown experiments. This was not observed in the experimental dataset used in estimating our kinetic parameters.

In the future, a possible way around this limitation would be to include protein-protein interactions into our kinetic modelling approach as shown in Figure 32.

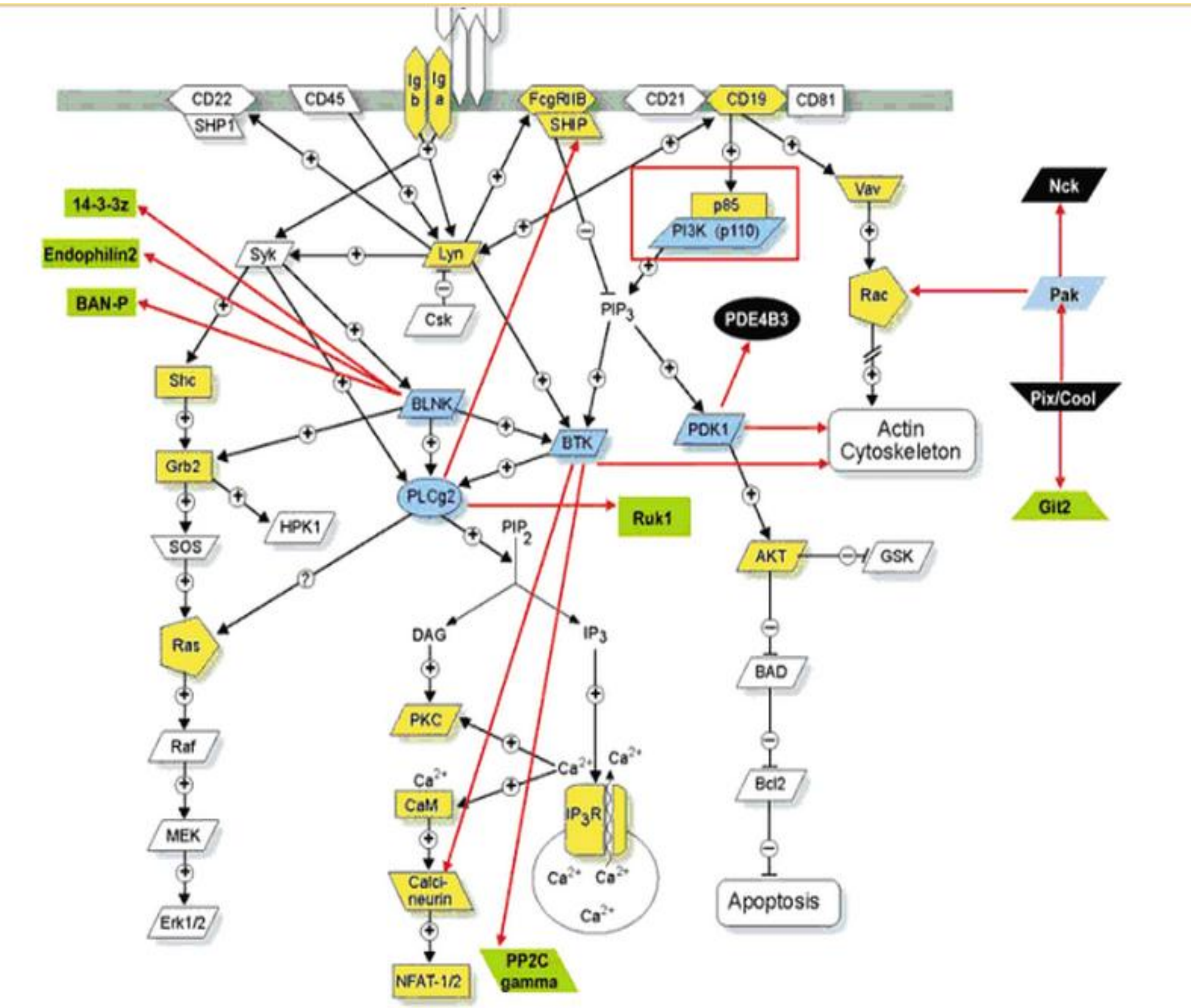


Figure 33: Protein-protein interactions in the B-cell receptor pathway shows that Pak binds to Nck and Rac.

Our results demonstrate a lack of dependency between proteins in our models which makes it impossible for any protein to affect another. To overcome this issue, another possible solution will be to include feedback loops into our models. This will allow us to analyse the effects of changes in protein levels affecting other proteins and importantly, performing gene knockout and protein knockdown experiments could yield better predictive results and extent our progress towards building genome-scale models that integrate biological data at different layers.

8.5 References

1. Bisswanger, H. (2008) Enzyme kinetic: Principles and Methods. 2nd revised ed. Wiley-VCH.

2. Henry, C.S. Jankowski, M.D., Broadbelt, L.J. and Hatzimanikatis, V (2006) Genome-scale thermodynamic analysis of *Escherichia coli* metabolism. *Biophys. J.*. **90.** 1453–1461

3. Liebermeister, W. and Klipp, E. (2006) Bringing metabolic networks to life: convenience rate law and thermodynamic constraint. *Theoretical Biology and Medical Modelling,* **3**, 41.

4. Tan, L.M., Rizk,M.L., Liao,J.C., (2008).Ensemble modeling of metabolic networks. *Biophys. J.* **95**, 5606–5617.