# A DISCRETE EVENT BASED STOCHASTIC SIMULATION APPROACH FOR STUDYING THE DYNAMICS OF BIOLOGICAL NETWORKS

by

SAMIK GHOSH

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

December 2007

## ACKNOWLEDGEMENTS

I take this opportunity to express my gratitude to my advisors Prof. Sajal K. Das and Kalyan Basu, for their moral support and technical insights throughout the course of my research. They introduced me to the invigorating field of scientific research, providing motivation and guidance whenever I needed them. Particularly, I am indebted to Prof. Basu for initiating me into the fascinating field of computational systems biology - providing never-ending motivation and guidance as I traversed through the challenges of this cross-disciplinary area of research. Prof. Kalyan Basu, whom we lovingly call "Kalyan da", has been and will always remain, the unflickering guiding torch in my life, in matters both personal and professional.

I extend my sincere regards to Dr. Mandal, Dr. Waasbergen and Dr. Stojanovic for their comments and suggestions regarding my work in biological modeling and simulation.

I cannot express in words, my appreciation of the valuable discussions and brain-storming sessions put in by all my colleagues at CReWMaN laboratory and for the late nights spent together at work.

I would also like to acknowledge the Computer Science and Engineering Department of UT Arlington and the Deans Graduate Scholarship for providing me financial support. A token of gratitude goes to my *Alma Mater*, Haldia Institute Of Technology, India for shining the first beacon of Computer Science on my life.

Finally, I thank my parents and relatives, who transcended the barriers of geographical distance in providing me encouragement and support without which this work would have never reached its completion.

November 19, 2007

## ABSTRACT

A DISCRETE EVENT BASED STOCHASTIC SIMULATION APPROACH FOR STUDYING THE DYNAMICS OF BIOLOGICAL NETWORKS

SAMIK GHOSH, Ph.D.

The University of Texas at Arlington, 2007

Supervising Professors: Sajal K. Das, and Kalyan Basu

With increasing availability of data resources on the molecular parts of a living cell, biologists are focussing on holistic understanding of cellular mechanisms and the emergent dynamics arising out of their complex interactions. Comprehending the fine-grained signal specificity, gene regulation and feedback mechanisms of molecular interactions at a network level forms a central theme of systems biology.

With the speed and sophistication of computational methods, *in silico* modeling and simulation techniques have become a powerful tool for biologists challenged with understanding the system complexity of biological networks. Numerical simulation of classical chemical kinetics (CCK), agent-based simulations of biological processes, and linear optimization models of metabolic networks, have been applied to the study of cellular behaviors with varying degrees of success. The spatio-temporal scales of cellular processes, coupled with the knowledge gap and complexity of biological networks limit the application of existing computational techniques.

In this dissertation, we present a network-centric modeling and simulation approach to systematically study the stochastic dynamics of cellular processes at a molecular level. The central theme of our approach revolves around abstracting a complex biological process as a collection of discrete, interacting molecular entities driven in time by a set of *discrete biological events* (*bioEvents*). We develop the discrete-event based simulation engine, called *iSimBioSys*, together with an integrated database of biological pathways, which captures the temporal dynamics of the molecules through stochastic interactions of different *bioEvents*.

With an illustrative case study of signal transduction networks in bacterial cells, we highlight the efficiency of a discrete event based approach in capturing high-level system dynamics of a biological process, particularly in reproducing the *switching effect* of the PhoPQ pathway in *Salmonella* cells as reported in experimental work. Next, we build a detailed stochastic model for the fundamental process of gene expression in prokaryotic cells and study the biological events of transcription and translation using the proposed simulation framework. Our results identify the role of transcriptional and translation machinery in controlling the burstiness of protein generation. We extend our simulator to incorporate a hybrid algorithm which combines stochastic models of signalling and regulatory events with a flow-based model for metabolic networks. In order to validate the efficacy of the hybrid simulation approach, we develop an integrated database of signaling and metabolic networks in the bacterial cell *Escherechia Coli*. The hybrid simulation recreates experimentally observed regulation of metabolic flux distributions in the network while providing new insights into the mechanism of regulation.

## TABLE OF CONTENTS

## LIST OF FIGURES

## LIST OF TABLES

# CHAPTER 1

# INTRODUCTION

Traditionally, the key focus of biology has been on detailed understanding of single genes, molecules or processes involved in particular phenotypic manifestations. With the discovery of the double-stranded structure of the di-oxy ribonucleic acid (DNA) molecule which forms the basic building block of a living cell, the field of molecular biology has increasingly generated detailed mechanistic maps of complex molecular machineries working inside a cell. This powerful "reductionist approach" [184] has resulted in a significant understanding of the structure and function of individual genes, proteins as well specific cellular processes.

With the development of efficient and cost effective sequencing techniques and the complete sequencing of the human genome, biologists have obtained huge amounts of data on the molecular "parts list" comprising a cell. The availability of high-throughput micro array experiments and bio chips have further augmented the process of accumulating experimental data on specific disease pathophysiologies at a molecular level. Complete genomic sequencing of new organisms has been completed and advanced databases like Genome Bank (GenBank) [176], Protein Database (PDB) [179], which store comprehensive annotations of genomic and protein structures, are being developed at previously unimaginable rates.

Concomitant with this development, a large body of knowledge is being derived from biological pathways activated by different regulatory genes, hormones and metabolic reactions through fluorescence tagging and other types of advanced *in-vitro* experi-

ments [11]. These results are captured in large volume of scientific papers and public pathway databases like PubMed [145], Ecocyc [52], KEGG [96] etc.

The availability of data in the post-genomic era has opened up further opportunities for researchers. The ability to systematically store and retrieve biological data, and more importantly, to be able to characterize the phenotypic behaviors of a biological system emerging as a whole from the "part-lists" has become the foremost question in biology [67].

As more and more data become available, biologists are now looking beyond assigning functions to individual genes. The focus is shifting from understanding complex biological systems as static models of loosely connected molecular devices to an 'integrated' or 'collective' mode of behavior, encompassing interdependent regulatory controls and multiple interacting components [87].

While system level understanding of biological processes has been a recurrent theme in biology from the days of Norbert Weiner [126], it has found increasing attention in the last decade, particularly due to the vastly improve techniques in molecular biology. The availability of high-throughput experimental data on the molecular entities controlling cellular function has opened up tremendous opportunities for the marriage of system sciences with the biological sciences to provide a complete spectrum of knowledge [67], [46].

Computational systems biology [67] aims to develop a class of integrated mathematical, computational and experimental techniques with the goal of linking the knowledge of different molecular parts of a living cell in comprehending the structure, dynamics, control and design of biological systems. As elucidated by Denis Noble in [46], "Systems biology...is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different....It means changing our philosophy, in the full sense of the term".

## 1.1 Challenges in comprehending molecular complexity

However, development of such integrated techniques for studying biological networks pose several challenges to biologists. While the ability to study biological systems at an integrated level remains the key goal of molecular biology, it is pertinent to identify the salient features of living systems which need to be tackled in building "systems level" views of molecular interactions:

- *Complexity* : Biological systems are characterized by their inherent *complexity*, arising from the non-linear interplay between different entities within a cell, as well as the myriad environmental signals affecting the physiology of tissues and organs. Comprehending the fine-grained interplay between the environment (intracellular as well as extra-cellular) and the cells in a living organism as it maintains its homeostasis [6] remains a major challenge for systems biology.
- *Knowledge gap*: The complexity of natural systems makes it further difficult to obtain detailed information on the molecular mechanisms associated with a particular process. Even with existing techniques, major molecular level knowledge gap exists in the understanding of the biology of various fundamental processes. Even for well-characterized cellular functions, the large number of parameters controlling their interaction dynamics make the study of large scale biological networks a particularly challenging problem. The ability to easily abstract the available level of knowledge on a biological process and bridge the "gap" in a computational framework remain a major hurdle of systems biology.
- *Time and space heterogeneity*: Another challenge for developing integrated models of biological processes is the large order of spatial and temporal dimensions in which cellular systems operate. Signaling networks operate on the order of seconds and minutes while metabolic reactions in a cell typically take microseconds. Also, molecular interactions are constrained to cellular compartments (cytoplasm or mi-

tochondria) while various viruses and bacteria affect tissues and organs in a body. Development of multi-scale models [38] of biological processes is a major goal of systems biology.

- *Computational challenges - need for speed*: Computational techniques and mathematical models provide a promising and powerful tool for capturing the complex diversity of biological systems. Particularly, in the post-genomic era, the use of computers and databases in systematic storage and retrieval of vast amounts of molecular data has become successful [145]. Even with the ever increasing power of computational tools, the need for fast and efficient biosimulation techniques [99] remains an open challenge for both biologists and computer scientists.

Thus, the fundamental challenge [87]in a "wholistic" understanding of biological processes is the complexity involved in the interaction of different components, coupled with the knowledge gap which exists in a complete characterization of their molecular mechanics. The complexity and knowledge gap increases manifold as we move into higher scales such as interaction of large ensemble of cells in a tissue, or interaction of tissues in continuum for rhythmic pumping of the heart [45]. The goal of "systems biology" is to build tools and techniques, both computational and experimental, which allow systematic characterization of biological systems, from cells to tissues and organs, and finally to the physiology of human beings, as outlined in Fig. 1.1.

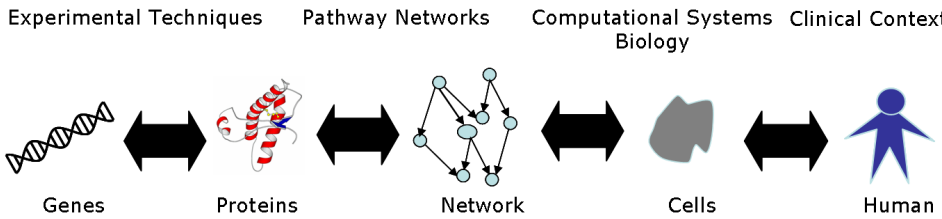



Figure 1.1. Closing the "gap" in system wide study of human physiology.

## 1.2 Role of *in silico* modeling and simulation

The complexity and magnitude of the problem in developing system-level understanding of cellular dynamics have opened up opportunities for the application of computational and network modeling techniques in this domain [6]. In recent years, researchers have recognized the necessity of developing systematic mathematical principles, borrowed from engineering and computer sciences, in an effort to comprehend the complex molecular choreography within living organisms. Computational models, or *in silico* models, supplement *in vitro* and *in vivo* experimental techniques, providing formal methods for storing and querying biological data. Moreover, leveraging the ever-growing power of the silicon chips, such techniques hold the promise of significantly reducing the cost of biological experiments and drug design. Well calibrated *in silico* models will allow biologists to perform systematic *what if* experiments on particular pathways, check the dynamics of different alternative hypotheses and proceed to further wet-lab experiments on only viable pathways as predicted by the computer models. These approaches have huge implications for the pharmaceutical industry, where the lack of computational techniques for reliable hypothesis-testing lead to high costs of drug development time.

Fig. 1.2 shows the typical drug development pipeline. As seen from the figure, the process starts with identification of potential trug target molecules and their subsequent validation and selection. Once the target molecules are selected, new chemical entities (NCE) or leads are developed and optimized for the selected targets. While in this stage, around 300 leads are indentified, severe attrition of the NCEs occur downstream, particularly in preclinical and earlu clinical phases ( phase I and II) leaving around 10-20 NCEs which enter advanced clincal studies. The optimized NCEs finally leads to the development of one new drug application at the end of phase III clinical trials completing the process which takes around 10-15 yrs and nearly a billion dollar in expenditure. Computational systems biology has the potential of significantly augmenting the different

stages of the drug discovery pipeline, providing *in silico* target pathway identification and lead optimization to the development of virtual trials on a suite of virtual patient population.

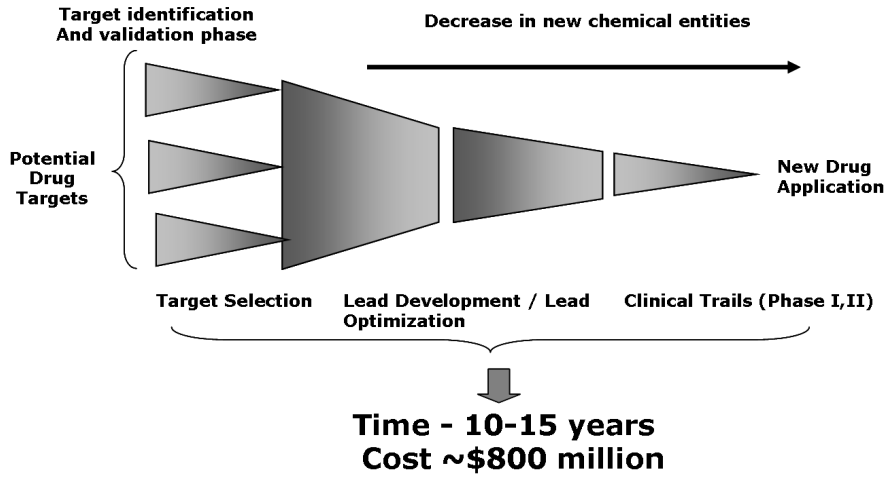



Figure 1.2. The role of computational systems biology.

With the promise of significantly reducing the time and cost of drug discovery, sophisticated *in silico* modeling and simulation techniques have become a powerful tool for biologists who are challenged with understanding system complexity of biological processes. With the unprecedented growth of genomic and experimental databases, bioinformatic tools have been developed to mine information from these raw data sets [145]. Sophisticated physico-chemical and mechanistic models [21] of molecular dynamics have been built and validated based on experimental data. In recent times, graph theoretic and network theory concepts [191] have been applied to extract information from complex pathways of biological entities, which can be combined with mechanistic models to develop cell simulation platforms.

Mathematical models have been particularly successful for complex biochemical reaction networks, using deterministic rate laws [66], which gives a relationship between reaction rates and molecular concentrations. Many simulation tools, based on classical chemical kinetics (CCK) have been developed [115], [141], [182], [70] which represent cellular dynamics in terms of ordinary differential equations (ODE) and employ numerical methods to solve them.

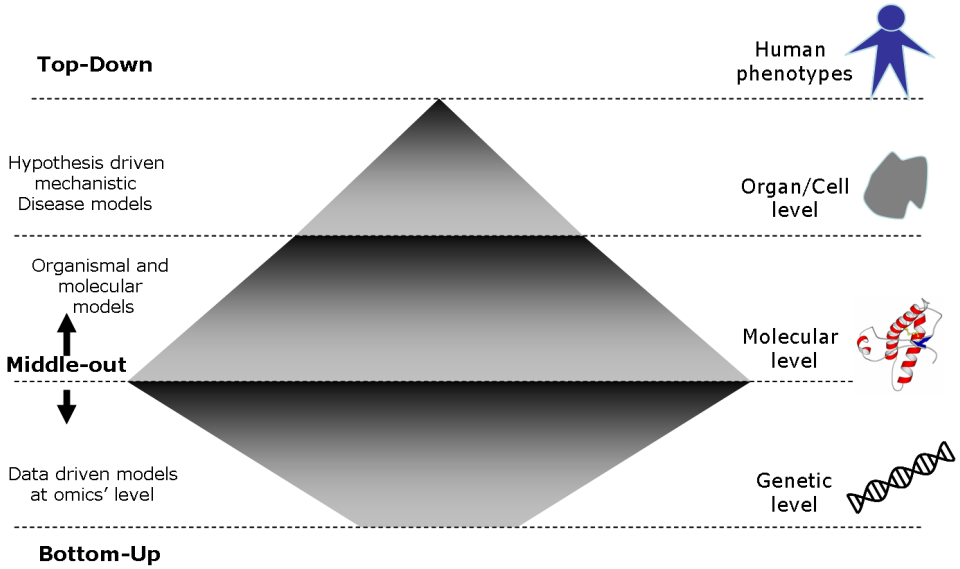



Figure 1.3. *In Silico* modeling and simulation philosophies.

In recent years, to encompass the inherent stochasticity of cellular systems [112], stochastic simulation algorithms have emerged as alternative modeling approaches for biochemical networks. Particularly, in the light of the fact that biochemical reactions involve discrete, random collisions between molecules, the assumptions of chemical reactions to be macroscopic under convective or diffusive stirring, continuous and deterministic, are simplifications of the physical reality [21]. Numerous stochastic algorithms which approximate the chemical master equation (CME) [41], have been developed and

successfully applied by Gillespie et.al [41, 102], Burrage and Tian [180], Novre et.al [127], in environments with low copy number of molecules and small system volume.

Fig. 1.3([47]) shows the broad classification of the different modeling techniques currently prevalent in computational systems biology. As seen from the figure, most modeling and simulation efforts employ either a "top-down" or a "bottom-up" view of the biological space. In a "top-down" approach, mechanistic models of the interaction between different tissues and organs are developed and represented through system of ordinary differential equations. The numerical solution of these equations, which are based on mass action kinetic parameters, capture the temporal behavior of the species in a continuous and deterministic domain. Such physiology-driven, top level models are effective in representing coarse-grained dynamics but do not capture the network level information stored in the more detailed molecular pathway maps.

On the other hand, "bottom-up" approaches start at the level of single molecules and genes, and integrates data from various genomic and proteomic databases to build increasing complex bio-molecular models. While such an approach provides a fine granularity of details, spatio-temporal scalability becomes a major challenge. Particularly, such data-driven approach face significant computational hurdles in encompassing large pathway networks at the high level of granularity. As shown in Fig. 1.3, most of the work on computational bio-modeling has focussed on the two extremes of the inverted pyramids. However, the base of the pyramids provides a wide spectrum of flexibility in terms of capturing physiological as well as "omics" level information of molecular pathways. A "middle-out" approach, wherein data from the bottom layers is integrated in a computational platform to study of dynamics of biochemical pathways and physiological interactions, can provide an efficient and flexible modeling paradigm.

## 1.3 Our Contribution

In this dissertation, we focus on the application of network based modeling and simulation methodologies for the study of biological systems. In this approach, the cell is viewed as a complex network of interacting molecular pathways and its phenotypic manifestations evolve through the dynamic interaction of the molecular entities under different intra-cellular and extra-cellular environmental signals.

Based on a middle-out design philosophy, the central theme of our approach revolves around abstracting a complex biological process as a collection of interacting functions driven in time by a set of *discrete biological events*. Analyzing the system at a molecular level, the temporal dynamics of the system are revealed by the interaction of these events. The stochastic behavior of the interactions is captured through the mathematical formalism characterizing the time associated with each of the biological events, i.e. the event holding time modeling. The discrete event models create the biological process description in time, while the event-based stochastic simulation captures the interaction of these processes through the events to create the dynamics of the biological system.

In this work, we have laid down the framework of a discrete event based stochastic simulation framework for studying the molecular dynamics of cellular processes. The key contributions of this work, which are depicted pictorially in Fig. 1.4, are outlined below:

- *iSimBioSys – discrete event biosimulation engine*: A novel discrete event based stochastic simulation framework is proposed for studying the temporal dynamics of cellular processes at a molecular level. The event level abstraction provides a middle-out philosophy, allowing the definition of biological processes at different levels of granularity (defined through the events) depending on available biological knowledge and focus of study. Based on the discrete event paradigm, a biosimulation platform, called *iSimBioSys* is developed which is able to incorporate biological knowledge and simulate the temporal dynamics of different molecular entities. With

the focus on flexibility of model abstraction at various scales, we study the dynamics of signal transduction and gene regulatory pathways, capturing the various biological functions through discrete events. We simulate the effect of virulence gene regulation in single cell bacteria, *Salmonella typhimurium*, reproducing the *switching effect* of external signals in controlling virulence as reported in experimental work.

- *Stochastic modeling of biological events and logic modules*: The proposed modeling and simulation schema allows the integration of stochastic models of various biological events at different levels of granularity. Each event model is represented by a parameterized mathematical expression for the probability distribution of event time.

  Delving into details of a particular biological event, we employ the proposed simulation methodology in developing stochastic models of prokaryotic gene expression. A stochastic birth-death model is developed for transcription and translation events in bacterial cells. The proposed analytical model is capable of providing a quantitative framework for systematically studying the effect of different molecular actors on gene expression dynamics.
- *Integrated database of biological networks and pathways*: An integrated database schema is developed, which stores information on the various biological pathways in a cell. The pathways database stores the information available in various disparate sources, public databases and biological literature, providing a common interface for querying and simulating interactions between them at different scales of space and time. An illustrative database schema has been developed for regulation of central metabolism reactions in the bacteria *Escherichia Coli* (*E.Coli*), integrating data on 7 global transcription factors and their signal transduction, which regulate

802 genes and their protein products which control the metabolism of around 1000 reactions involving 2000 metabolites.

- *HimSim flow based simulation of metabolic network engine*: A novel hybrid simulation technique called *HimSim* is proposed, which combines discrete event driven models of slow time-scale events (like signaling and gene regulation) with an algebraic data-flow based model for capturing the fast-time scale metabolic reactions. The hybrid scheme alleviates the problem of "stiffness" i.e. inability to simulate the effects of fast time-scale reactions in conjunction with slow reaction models. The proposed methodology, together with the integrated database, provides a generic platform for genome-wide, multi-scale modeling of cellular networks.

  The hybrid approach captures the interplay between signal transduction, gene regulatory and metabolic networks for the bacterial cell *Escherechia Coli*(*E.Coli*) built in the integrated database. We validate experimental reported results on the transcriptional regulation of metabolic reactions in *E.Coli* under different environmental and growth conditions. While simulation studies on this hybrid platform focused on the core metabolism part of the *E.Coli* metabolic network (with around 50 genes and 100 metabolic reactions, the existing system together with the database is capable of simulating the current genome-scale networks for *E.Coli* as available from the EcoCyc database [52].

## 1.4 Organization of the Dissertation

In Chapter 2, we present a detailed taxonomy of existing computational approaches for modeling and simulation of complex biological processes, identifying the key features together with the promises and pitfalls of the existing approaches employed in the study of cellular dynamics.

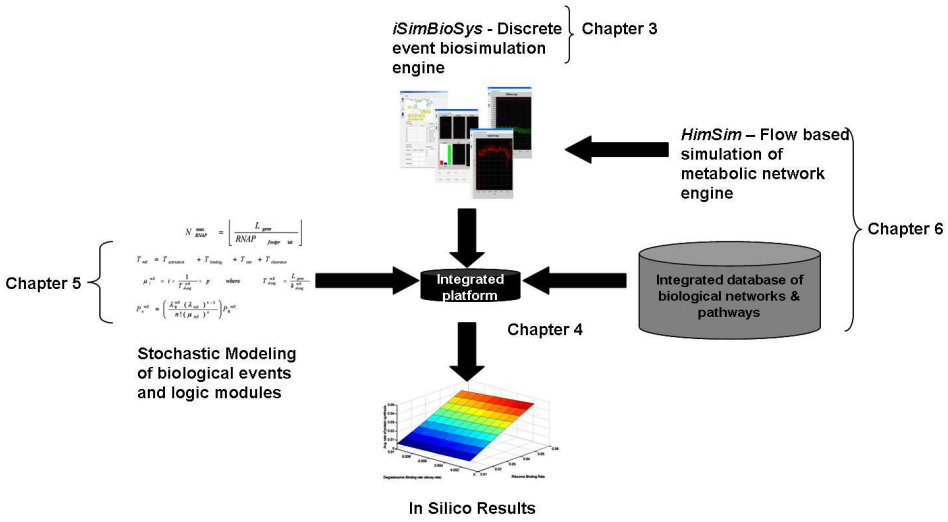



Figure 1.4. Contributions of this dissertation.

In Chapter 3, we outline the discrete event simulation paradigm, which provides an event based abstraction of biological networks and their interactions. Further, we build the architecture of our biosimulation tool, called *iSimBioSys*, elucidating the different components of the framework and outline its salient properties vis-a-vis existing biosimulation platforms.

We validate the efficacy of the discrete-event based approach by simulating the dynamics of signal transduction and gene regulatory networks in the bacterial cell *Salmonella Typhimurium* in Chapter 4. *In silico* models are developed for the key functions involved in the virulence pathogenesis pathway in *Salmonella* and simulation results obtained to validate existing wet-lab data for the signaling network system.

With a goal to model the biological events at different scales of granularity in our discrete event based approach, Chapter 5 develops a detailed stochastic model for the fundamental process of gene expression in prokaryotic cells and studies the biological events of transcription and translation in a simulation framework. Our results identify

the role of transcriptional and translation machinery in controlling the "burstiness" of protein generation.

Chapter 6 elucidates on the hybrid simulation framework proposed for studying biological networks at different time-scales and develops the integrated database schema for storing pathway data. We conclude with the key observations of this work in terms of its efficacy in providing a fast and scalable biosimulation platform for developing large-scale models of various human disease pathophysiologies.

# CHAPTER 2

## *IN SILICO* MODELING AND SIMULATION LANDSCAPE

The wealth of data available from high-throughput experimental systems that populate ever-increasing molecular pathway databases has brought about a paradigm shift in the study of molecular biology [67]. Sophisticated experimental methodologies, like micro-array based genome-wide study of gene expression, are being developed for system wide analysis of cells. With increasing availability of data resources on the molecular parts of a living cell, biologists are focussing on holistic understanding of cellular machinery and the emergent dynamics arising out of their complex interactions.

As elucidated in the previous chapter, the inherent complexity of the different biological networks, coupled with cross-talk and non-linear, spatio-temporal interactions between the molecular components of the pathways, render *in vitro/ in vivo* experimental approaches insufficient in capturing the dynamics of the system. In recent years, computational techniques, derived from different branches of engineering and sciences, have become popular in providing a systematic mathematical formalism in the study of biological processes [115], [141], [182]. Mathematical modeling provides a systematic formalism for capturing molecular details in a physiological context which can be stored in dynamic repositories and subjected to computational studies for uncovering biological insights.

In this chapter, we provide a taxonomical overview of different modeling and biosimulation techniques which have been applied in the study of complex biological systems with varying degrees of success . In section 2.1, we identify the principle parameters governing the different modeling techniques. Next, in sections 2.2 to 2.5, we delve into

specific *in silico* modeling methods and identify their strengths and weaknesses putting in perspective the discrete event based stochastic simulation approach proposed in this dissertation in section 2.6.

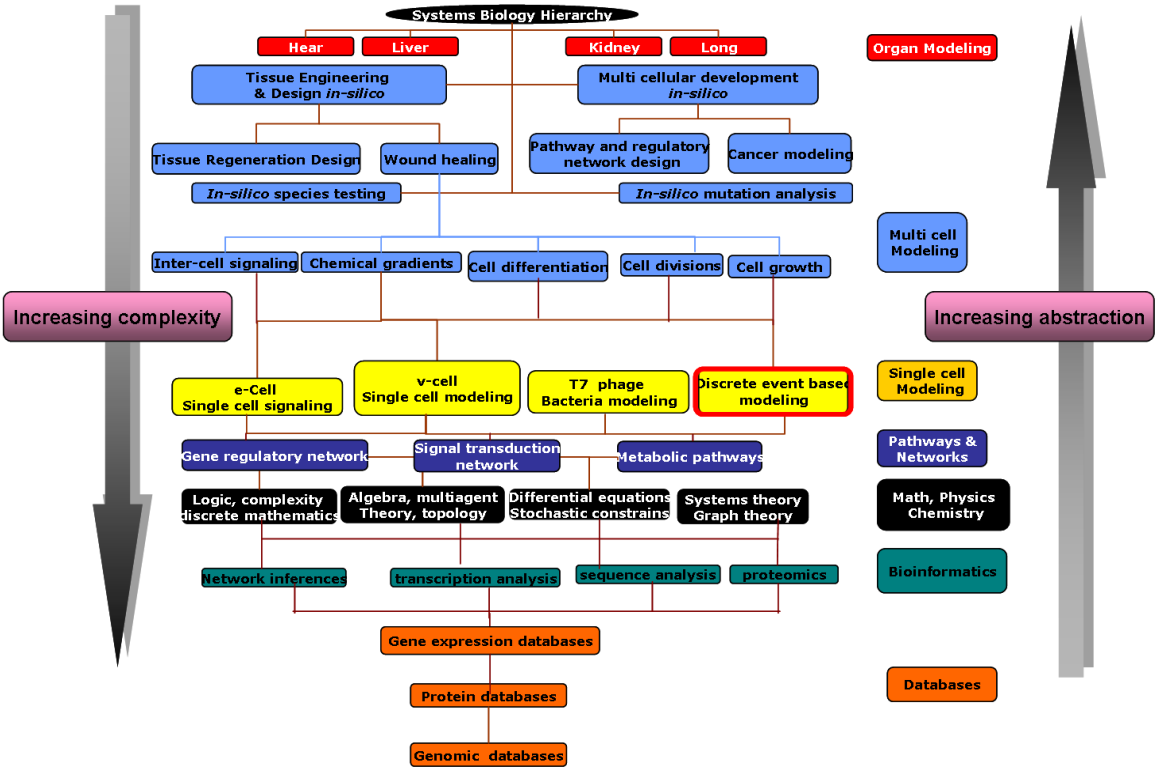



Figure 2.1. Modeling & simulation tools and methods.

## 2.1 Taxonomy of bio-modeling and simulation approaches

In developing modeling approaches for biological systems, it is pertinent to observe the unique features of complex biological functions which can affect mathematical models. As identified in the Chapter 1, biological systems are typically characterized by complex, non-linear interactions of a large number of molecular entities, knowledge

gaps in the understanding of their mechanistic behaviors, together with wide variations in spatial and temporal patterns. With this large spectrum of factors that control biological functions, physicochemical biomodels can be classified on the basis of following salient parameters [21]:

- *Time*: The propagation of time in the system can be continuous (C) or discrete (D)
- *Space*: The state space of the system can be continuous or discrete
- *System Evolution*: The evolution of the system can be considered in terms of being deterministic (D) or stochastic (S)
- *Physical Scale*: The model can consider the system at a microscopic scale (as in molecular dynamic simulations), macroscopic scale (chemical kinetic systems) or mesoscopic (where individual molecules are represented).

Based on the above characteristics considered in a biological model, *in silico* techniques can be broadly classified into four categories:

- Classical chemical kinetic (CCK) approach
- Stochastic simulation algorithm (SSA) approach
- Computational approach
- Agent-oriented approach

We elucidate on these modeling and simulation techniques, together with their promises and pitfalls, next.

## 2.2 Classical chemical kinetic (CCK) approach

The most extensively used modeling approach, which has been applied across different time and scales of biological processes, is based on classical chemical kinetics (CCK) approach. Most common models of molecular networks like kinase cascades and metabolic pathways, gene regulatory networks and protein interaction networks consider

the system as a set of coupled ODEs (ODE network) and use numerical methods to capture the system dynamics deterministically in continuous time and space.

A large number of computational tools, which provide a software platform for building, storing and parameterizing a set of biochemical reactions and solving those using numerical techniques, are available, like Gepasi [141], Jarnac [70], CyberCell [170], [110], Stode [29]. These rate-based models have been successfully applied to study gene expression and other molecular reaction systems [66] as well as build physiological disease models [159].

While continuous-deterministic reaction models are capable of capturing behavioral dynamics for spatially homogeneous systems with large number of molecular species, the inherent stochasticity observed in many biological processes (gene expression and protein synthesis) have proven the limitation of CCK in accurately representing biological processes. As mentioned earlier, due to the stochastic nature of molecular interactions, the assumptions made in deterministic models break under various biologically relevant scenarios:

1. *Volume*: Deterministic systems assume infinite volume to convert a spatial distribution of discrete molecules into single, continuous variable of concentration. As many intracellular reactions occur in small volumes, the assumption affects the accuracy of the model.
2. *Stochastic Fluctuations*: In many cases, fluctuations of the bimolecular systems are amplified (stochastic resonance) and cause observable behavioral changes at the macroscopic level, particularly for low copy number of molecules.
3. *Spatial heterogeneity*: Continuous, deterministic systems assume system homogeneity, which is frequently violated in biochemical systems due to compartmentalization of processes within the cell.

4. *Deviant behaviors*: In a recent work, Arkin et.al [112] systematically identified classical scenarios where the CCK model fails to capture system dynamics even in cases of high molecule counts, and proved that deterministic CCKs are closer to the 'mode' rather than the 'average' behavior of stochastic reaction dynamics.

## 2.3 Stochastic simulation algorithm (SSA) approach

Stochastic models, which present an accurate approximation for the chemical master equation (CME) [41], have been developed, largely based on Gillespie's algorithm [41, 102]. In this method, the next reaction event and the time associated with it are computed based on a probability distribution (Monte Carlo Step). The original Gillespie's algorithm is inherently very slow as it has to generate a large number of random numbers on each Monte Carlo step. Stochastic tools, like StochSim [127], have been developed based on Gillespie's technique and its computationally efficient variants like Gibson-Bruck [102] and tau-leaping [180, 5] which incorporate approximations to speed up the simulation.

Other modeling tools, which provide an integrative environment to build and study biochemical reaction systems in an exchangeable format (like Systems Biology Markup Language (SBML)) using deterministic as well as stochastic techniques are available, like E-Cell [115], Virtual Cell [186], Dizzy [161], and M-Cell [65]. These techniques are based on treating a biological process as a system of equations, represented by their rate constants and other parameters (like volume, cell density etc.) and simulating their interactions through numerical techniques or Monte Carlo based stochastic simulations.

While stochastic techniques provide a more closer mechanistic model of molecular interactions, the computational cost of the models pose a serious problem for large scale biological networks [99]. Moreover, as both CCK and SSA are based on representing the molecular interactions in the form of reaction equations, capturing the biology of

different pathways require building reaction models with thousands of equations and this relevant kinetic parameters. The cost of numerical simulation techniques, together with the difficulty in obtaining kinetic constant values, make these approaches challenging for genome-scale system study.

## 2.4 Computational approach

Computational approaches, based on optimization techniques and graph-theoretic formalisms, are particularly promising in studying biological systems with incomplete mechanistic knowledge where the system behavior evolves as an emergent property through non-linear interaction of molecular entities.

Computational models for biosimulation has been developed based on Petri nets [130, 131, 73] and stochastic process algebra [103, 28, 12]. These methods present a mathematical formalism for representing biochemical pathways within and between cells. In [131], the authors present a stochastic Petri net (SPN) model for studying simple chemical reactions (SPN model of ColE1 plasmid replication) and show how existing softwares can be used to perform structural analysis based on numerical techniques.

Discrete event system specifications based on Devs & StateCharts [154, 155], developed by Harel et.al and Stochastic $\pi$ calculus [12] have been successfully demonstrated to provide a computational platform for temporal simulation of complex biological systems. Hillston et. al have developed a mathematical technique, Performance Evaluation Process Algebra (PEPA) [103, 108], wherein functionality is captured at the level of pathways rather than molecules and the system is represented as a continuous time Markov chain.

Another approach, based on steady-state, constraint-based optimization of cell properties, have been particularly successful in developing metabolic reaction models and flux computation. Flux balance analysis (FBA) [2, 117] abstracts metabolic flux distribution in a cellular network as an optimization problem driven by thermodynamic

and stoichiometric constraints. We provide a more detailed discourse on flux balance methods later in Chapter 6.

While such computational techniques provide a efficient algorithmic platform for the analysis of specific biological systems, the lack of a common interface and data integration techniques render them unsuitable for systematic study of biological pathways, operating on varying time and space scales.

## 2.5 Agent-oriented approach

Simulation methodologies, based on software engineering concepts of object abstraction and modularity have been applied to the development of computational models of biological processes with emergent behaviors.Agent based modeling (ABM) paradigms have been applied in the studied for *in silico* complex bio-processes by Uhramacher et.al [15, 17, 86]. In [171], the authors have developed AgentCell, an ABM based digital assay for the study of bacterial chemotaxis. Another modeling technique, Functional Unit Representation Model (FURM) [160, 166] has been proposed for large scale modeling of in vitro drug metabolism. Simulation platforms, based on discrete events, where the events are modeled on rate constants and measured experimental data, have been demonstrated in [188] and [183].

Agent-oriented approach employ object abstraction and modularity concepts from software engineering to provide a software platform for biological simulations. One of the limitations of such an approach is the requirement of explicit agent definition and their specific functional behavior in a biological process. In many biological pathways, identification of well-define modules as well as functional behaviors becomes difficult due to lack of sufficient data on the particular entities.

## 2.6 Summary

The overarching theme guiding the development of *in silico* modeling and simulation tools, is developing models based on continuous-deterministic ODEs or using stochastic simulation algorithms (SSA) for approximating the chemical master equation, which capture the temporal evolution of the reaction event probabilities. Most of these techniques focus on specific parts of molecular pathways, which are represented in graphical and mathematical formalisms, treat spatial dynamics in terms of well-defined cellular compartments, and abstract the complexity in terms of estimated parameters and rate constants. The different approaches and specific implementations outlined in this chapter, are represented graphically in Fig. 2.1 [47] depicting the trade-off between system scale and complexity in the modeling space.. As seen from the figure, top-down approaches based on mechanistic models of cellular physiology capture the system behavior at a higher scale while compromising on molecular level complexity available from "omics" databases. Bottom-up, data-driven approaches integrate proteomic and genomic level data but suffer from scalability problems at the tissue and organ level physiology.

In this chapter, we systematically built the taxonomy of these different techniques and identified the key design parameters of the bio-modeling and simulation landscape. We have provided an overview of the different modeling and simulation philosophies for studying biological processes at various levels of granularity. Table 2.1 outlines the different approaches and compares their characteristics as identified here.

In the next chapter, we outline our modeling and simulation technique, based on a discrete event system specification, where the biological events (representing reactions, ionic diffusions) are mechanistically modeled depending on their biophysical characteristics to compute the probability distribution of their execution times. A discrete event simulation system then links the biological processes to simulate the behavior emerging from the interaction of the events in time.

Table 2.1. Comparative list of biological modeling and simulation softwares

| ***Modeling technique*** | ***Tool*** | ***Spatial representation*** | ***Temporal evolution*** | ***Reaction model*** |
|---|---|---|---|---|
| Classical chemical kinetics (CCK) | Jarnac [5] | Not explicitly defined | Continuous time | mass action kinetics |
| | Gepasi [141] | Compartments, sub volumes | Continuous time | Mass action |
| | E-Cell [115] | Compartmental | Supports CCK as well as SSA | mass action, chemical master equation (CME) |
| | SimBiology [182] | Not explicitly defined | Supports CCK as well as SSA | mass action, chemical master equation |
| | CyberCell [170] | Off lattice | Inter-particle collisions | MD based |
| Stochastic simulation approach (SSA) | MesoRD [85] | Compartments, sub volumes | Event-driven | CME |
| | M-Cell [65] | Off lattice | Time-step driven | At surfaces, CME |
| | Smoldyn [169] | Off lattice | Interparticle collisions | MD based |
| | Dizzy [161] | Supports CCK as well as SSA | continuous time | CCK, CME |
| | Promot/ DIVA [110] | Not explicitly defined | Continuous time | CME |
| Computational approach | Stochastic Petrinets [130, 131, 73] | Compartments | Continuous/discrete time steps | Graphical model |
| | Flux balance analysis (FBA) [2] | Not explicitly defined | Constraint driven, steady-state flux optimization | linear optimization technique |
| Agent-oriented approach | AgentCell [171] | Not explicitly defined | Time-step driven | Agents model molecular behavior |
| | FURM(Functional unit representation of biological processes [160, 166] ) | Not explicitly defined | Continuous time | Functional modeling |
| Stochastic discrete event based | *iSimBioSys* [157] | Compartmental | Event driven discrete time steps | Based on CME, explicit models of reaction time |

# CHAPTER 3

# A DISCRETE EVENT BASED SIMULATION PARADIGM

A fundamental challenge in computational systems biology [67] is the "..judicious simplification" of the biological system complexity without loosing the ensemble dynamic behavior in an incomplete biological knowledge space. As elucidated in the previous chapter, various computational modeling/simulation tools have been developed to represent biological processes using formal mathematical constructs, either in the form of ordinary differential equations or agents. While these techniques are capable of customizing specific representation of pathways and molecular interactions, a generic framework capturing the different biochemical networks over a wide range of time, space and scale dimensions is needed for developing large *in silico* models of cellular physiologies.

In this chapter, we propose a network centric approach for modeling complex biological pathways through the stochastic interaction of *discrete biological events* (*bioEvents*). In section 3.1, we outline the details of the modeling technique, identifying the biological events and stochastic models associated with them and comparing them with existing stochastic approaches. Section 3.2 builds the biosimulation framework, called *iSimBioSys*, defining the simulation algorithm and software architecture, which provides the computational platform for studying the interaction dynamics of the events. We summarize the contributions of this chapter in section 3.3.

## 3.1 Stochastic discrete event based approach

In the system engineering view of complex biological processes [23], the key notion is to abstract the complexity of the system as a set of *discrete* time and space variables

(random variables), which capture its behavior in time. The entire system is a collection of functional blocks or modules, which are driven by a set of "events". An "event" defines a large number of micro level state transitions between a set of state variables accomplished within its execution time, also termed "holding time". The segregation of the complete state space into disjoint sets of independent events which can be executed simultaneously without any mutual interaction forms the basis of abstraction for the particular system under investigation.

The application of this technique in large complex communication networks [88] has demonstrated the accuracy of the approach for the first and higher order dynamics of the system within the limits of input data and state partitioning algorithms [178]. For example, discrete event based system modeling has been effectively applied for designing routers, the key components responsible for routing traffic through the Internet. Discrete event based simulation techniques have also been used a wide variety of manufacturing processes and system dynamics of complex industrial processes [168].

Motivated by the success of discrete event driven stochastic simulation techniques in large scale complex networks, our approach is based on identifying and modeling key biological functions at a cell, tissue or organ level and mapping those to a set of *discrete biological events*, called *bioEvents*, associated with the model processes. The model captures the behavior of the pathways in time, through the stochastic interactions of the different *bioEvents*, as shown in Fig. 3.1.

### 3.1.1 Biological event identification and definition

Each event represents a biological interaction (chemical reaction, ionic diffusion etc.) and is associated with two characteristics:

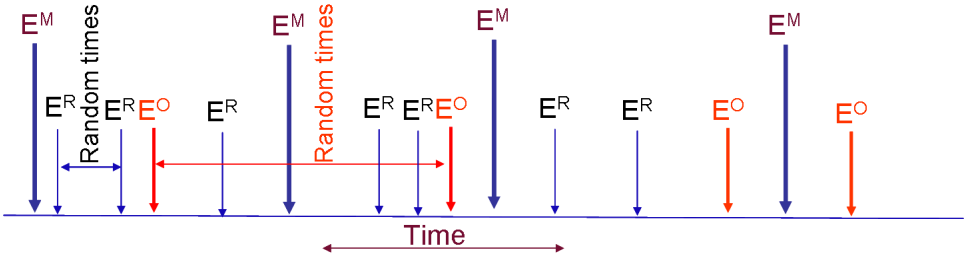



Figure 3.1. Event interactions in time.

- The parametric stochastic model of the underlying physico-chemical process associated with the event. The model, elucidated further in the next section, characterizes the holding time distribution associated with the event.
- The molecular resources associated with the event (e.g the molecular species involved in a reaction event) and their utilization algorithm based on reaction stoichiometry or pathway knowledge.

Thus, to define the discrete events, we first identify a biological process as a system of resources (which can typically be the various molecules, ions, ribosome, chromosome, operons, etc. involved in the system) that are periodically changing their state between "busy" (e.g., a molecule is busy in a reaction), "free"e.g., a molecule is free to enter a new reaction),"created" (e.g., a molecule is created by a reaction) and "killed" (e.g., a molecule is used up by a reaction) based on the underlying resource usage algorithms. The events are marked by the time instant the selected resources change their state in the system. The state transitions from one state to another are governed by transition rates of the functions involved in the process.

The estimation of the transition rates is derived by mathematical model or by experimental observation of the physical processes involved in the functions. As an example, we consider the fundamental function of phosphorylation, which involves the transfer of a phosphate ion from an Adenosine triphosphate (ATP) molecule to another molecule/ion

resulting in the phosphorylated molecule/ion and a molecule of Adenosine diphosphate (ADP). In particular, we consider the phosphorylation of a PhoP molecule (which as we will see later is intra membrane protein signaling molecule involved in the regulation of the PhoPQ pathway in *Salmonella*) to phosphorylated PhoP or PhoPp. In order to capture the dynamics of this basic biological function, we need to account for the state of the resources involved (in this case ATP, count of PhoP molecule and phosphorylated PhoP molecule, and ADP). Further, the time required to perform this function, which is termed as the *holding time*, is estimated on models based on fundamental physical processes like molecular kinetic, diffusion physics and molecule binding mechanism that will be in place at that particular system state. Thus, this holding time will be randomly changing as the system states change and will accurately reflect the actual working of the cellular system. At the end of the "holding time", the phosphorylation molecule can trigger an "event" to drive another functional process.

As the simulation proceeds at a event level, the resource states are determined in terms of the "molecular count" of the individual resources that are affected by the event. For example, after the successful completion of the PhoP-phosphorylation event, the count of ATP in the system is decreased by one while that of ADP is increased by one. The PhoP molecule is "killed" and phosphorylated PhoP molecule is "created". In this way, basic biological molecules and their events are identified, modeled and linked together in a discrete event simulation framework to capture the dynamic interactions of a cellular process in time.

### 3.1.2 Modularity and Module Reuse

As is evident from the above discussion, one of the key challenges of this discrete event modeling of biological processes is the identification of basic functional modules, the resources involved in them and the key events driving the interaction between the dif-

ferent modules. The wide variability and complexity of modules, resources and possible set of events in natural sciences further complicate the problem. However, there exists a core set of basic functional modules which play fundamental roles in a wide variety of biological processes. Identification and modeling of theses functions can greatly facilitate the study of complex processes of life. Some of the basic biological events (and their associated time), which are associated with key biological functions, include:

(1) Reaction Time,

(2) Diffusion Time,

(3) DNA Protein binding time

(4) Transcription Time

(5) Translation Time

(6) Transport Time

(7) Protein Life Time

(8) Protein Folding Time

Identifying 'modularity' of biological processes forms a key step in employing system engineering principles to the study of these complex processes. A discrete event framework allows the identification of such modules, few of which are outlined above, and their characterization in terms of their input and output events, event holding time distributions and resource utilization stoichiometry. Such formal characterization lends reuse of the modules across various biochemical pathways and networks. In the next section, we layout the mathematical underpinning for the modules based on their biophysical and chemical characteristics.

### 3.1.3 Capturing the system behavior in the temporal domain

In discrete event simulation, "simulation time" is the representation of the "physical time" of the system being modeled. Each event time computation is associated with a time-stamp indicating when that event occurs in the physical system being simulated. The event time is computed from the knowledge of the previous event that triggers the current event, together with the event execution time.

This execution time is often called "holding time" of the event function and is generally a random number. The dynamics of resource utilizations with progression in time unveil the complete internal picture of a complex biological process, capturing the evolution of the system in time. The exercise of characterizing the system parameters is performed as follows:

- Identify the list of discrete events that can be included in the model based on the available knowledge of the system. As mentioned previously, due to lack of complete understanding of biological process, at this stage the modeler can abstract their system at different level of event definition, like reactions events, event of molecular assembly like ribosome, or even events at higher structural levels.
- Identify the resources of interest for the execution of the event function which are being used by the biological process for each discrete event. In other words, we need to identify the various types of molecules, cells, tissues etc which are involved in the event function. In addition, we include the biologic understanding of the event execution to define the resource usage algorithm for an event (either in reactions, or as catalysts or end products). For example, the ribosome binding gap on the mRNA for protein synthesis.
- Compute the time taken to complete this biological discrete event, i.e. the holding time of the discrete event. For this purpose, it is important to mathematically relate all the event parameters which affect the interaction of the resources in a particular

biological function. In reality many of these parameters are random variables which are linked through complex algebra of random numbers. The event execution time is a random number drawn from a probability distribution characterized by its two significant moments. Because of the details of the biological function included in this mathematical model, the moments are parametric and change with the change in system state.

Unlike in rate based simulation models, where it is assumed that the system state remains the same during the complete reaction of multiple molecules, the time required for completion of a biological event processing is computed as a function of these parameters. The resource utilization algorithms which determine the holding time of the functional blocks, together with the resources involved and their count in the system, all play a key role in the dynamic behavior of the biological process being simulated.

- Identify the next set of biological discrete events initiated on the completion of an event. If multiple discrete events are possible after completion of an event, it is necessary to find out the probability of the individual next event. This modeling of the probability depends on the biological knowledge captured through micro array or other experimental data that are reported in pathways and other research databases. For example, if one regulatory protein can possibly activate multiple proteins due to the similarity of binding motif, the probabilities are modeled by considering the individual binding site locations and the chromatin configuration of the binding site. Though this knowledge is available, this level of biological detail is not currently used in rate constant based ODE models.

Thus, extraction of the system information from experimental data captured in literature to generate the pathway logic is an important component of any biological system modeling. In current rate-based systems, this complex pathway knowledge (with

positive and negative feedback loops) is converted into a system of kinetic equations with rate constants (ODE network). This transformation potentially looses the temporal behavior of the pathway, as it treats the ODEs as a memory less system of reaction equations. In a discrete event based simulation, this behavior of the system is captured through the sequence of biological events. Once the components are defined and linked in the simulation framework, the dynamics unfold by the interacting of these events in time. The overall functional modules of the simulation framework are outlined in Fig. 3.2.

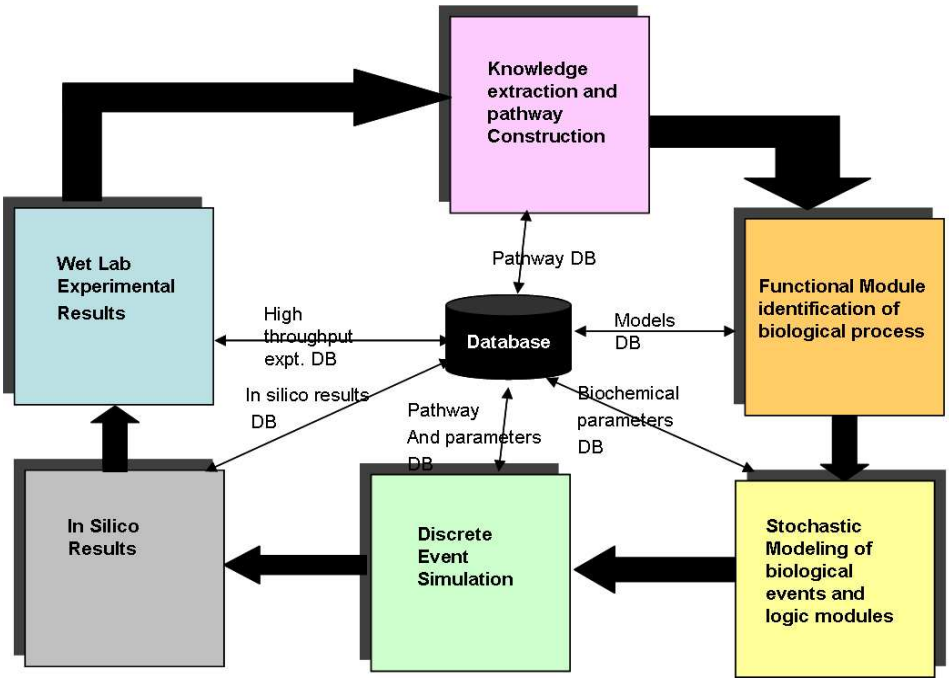



Figure 3.2. The functional modules of the simulation.

As shown from the figure, the simulation methodology starts from extraction of pathway information from different sources (databases as well as existing literature) which store molecular level data from high throughput experiments. Based on the network information captured from the pathways, the biological functions and modules are

abstracted and maps to specific *bioEvents*. The *bioEvents* are characterized by their associated resources and the stochastic models of the event holding times. The event network is fed into the simulation engine block which drives the time-series result on the change in concentration of the molecular species and generates *in silico* results. The *in silico* experiments provide new insights into the biology of the system and contribute to the development of *de novo* wet lab experiments, thereby completing the circular process.

### 3.1.4 Comparison with existing stochastic simulation algorithms

The efficacy of discrete event stochastic simulation techniques in the analysis of system dynamics for complex biological processes has been successfully demonstrated in the literature [88, 178, 168]. A large volume of work in stochastic *in silico* analysis of biological systems is centered on Stochastic Simulation Algorithms (SSA) using Gillespie's technique [42, 43] and its variants [102]. Thus, we highlight the key characteristics of our event based stochastic simulation approach, the basic steps of which are outlined in Table 3.1, vis--vis the general SSA technique employed in Gillespie's algorithm:

- *Event Modeling and Resource Update*: The development of stochastic event models is closely linked to the success of the simulation and forms a central part of the modeling and simulation approach. While the fundamental notion of approximating the Chemical Master Equation (CME) [38] forms the driving principle in any stochastic modeling framework, the event modeling and execution phase (Step 2 in Table 3.1) and the resource update phase (Step. 3) differentiates the two techniques.

  While Gillespie and other SSA algorithms employ a Monte Carlo step to determine the next reaction event and the time-step update from a memory less list of events (event holding time exponential), individual event holding time probability distributions characterize the discrete event approach. Moreover, the time-step is

updated according to the particular temporal sequence of events associated with the biology of a process.

- *Capturing knowledge gap*: Current simulation systems use pseudo-equations to capture knowledge-gap in system behaviors wherein acquiring the rate constants for those equations becomes a challenging problem. In our approach, knowledge gaps are captured by defining biological events at a level of abstraction where knowledge is available, allowing events models to be at different levels of granularity for each functional block.
- *Temporal Evolution*: The Gillespie algorithm makes time steps of variable length, choosing one random variable to determine the next reaction and another to determine the time of the reaction, based on the rate constants and population-size of each species in the system. In another variant, StochSim [127] uses time steps of fixed lengths in the simulation. The time step has considerable impact on the computation speed and accuracy, and has to be tuned for the specific problem under investigation. In our discrete event based algorithm, the system evolves in time through the events and their holding time distributions. The system will adopt the time step depending on the events without any special or prior analysis.
- *Individual-based vs. Population-based Representation*: The discrete-event based simulation traces the system dynamics at the level of individual molecules, i.e it is possible to trace the state of a tagged molecule as it changes its state from "busy" to "free" or "killed" through the progression of events. On the other hand, the Gillespie algorithm treats molecules at a population level and the identity of single molecules are lost in the process.
- *Multi-state molecules and reaction stiffness*: The Gillespie algorithm has been shown to run into combinatorial state space explosion for multi-state molecules because each state has to be treated as a being part of a separate chemical equa-

tion [25]. In StochSim [127], the authors introduce binary flags and associated probabilities to consider multi-state molecules (e.g multi-protein complexes) as part of a single reaction. In our discrete-event framework, the multi-state events are incorporated as sub-class of a single event which are modeled (models consisting of input events, associated event holding distribution, resource utilization stoichiometry and output events) as part of the general framework.

- *Capturing the transient system behavior*: In a system-wide study of biological pathways, involving large number of molecular entities, the biological knowledge of the system is captured in the sequence of events driving the pathway. Maintaining the *sequence* of events, as employed by our discrete event based simulation algorithm, is essential for understanding the system behavior, especially in the transient phase when the number of molecular species is low (for example, in a signal transduction and downstream gene expression pathway, a transcription event (reaction) cannot be executed before signaling and kinase/phosphastase activity events are executed). Stochastic simulation algorithms employ Monte Carlo techniques where the order of the reactions is not considered because the system is assumed to be memory less.
- *Computational Efficiency:* The performance of the discrete event simulation technique is based on the number of events ($e$) generated by the system. Because of longer execution time of regulatory events, the event rate significantly drops in our system and significant speed-up is possible. As analyzed in [127], the computational complexity of the Gillespie technique is on the order of number of reactions, while that of *StochSim* is of the order of $n$, where $n$ is the number of molecules in the system. Also, it may be noted here that unlike the fixed size time-step in *StochSim* [127] wherein reactions may not be executed in a time step, the discrete event simulation, being event driven, moves forward in time at every biological event, adjusting the count and states of molecular resources.

- *Parameter Estimation*: Current simulation models are based on rate constants predominantly estimated from experimental data. This becomes particularly challenging, as for many reactions, the rates are not experimentally available, and considerable assumptions have to be made to complete the model. On the other hand, the stochastic models are parameterized on physicochemical molecular properties, like temperature and reactions energy barriers, and pathway information, which are available in different databases. The parameterized models can be tuned to validate different wet-lab experimental conditions (like modeling pathway behavior at higher temperature) and for testing *de novo* hypotheses on an existing system.

Table 3.1. Stepwise comparison between Gillespie and Discrete event appraoch

| | ***Stochastic Simulation (Gillespie Algorithm)*** | ***Stochastic modeling based Discrete event simulation (iSimBioSys)*** | ***Comments*** |
|---|---|---|---|
| **1** | **Initialization**: Initialize the number of molecules in the system, reactions constants, and random number generators | **Initialization**: Initialize the number of molecules in the system for each species, model parameters and resources and random number generators | The initialization steps are similar in both the algorithms |
| **2** | **Monte Carlo Step**: Generate random numbers to determine the next reaction to occur as well as the time interval. | **Event modeling and execution**: The next reaction or event is selected based on the functional logic hardwired in the simulator.<br>For each process and its associated event, a random number is generated for the event execution time based on the first and second moment of the event holding time distribution computed by the stochastic model. | In this step, Gillespie and other stochastic simulation algorithms employ a Monte Carlo step to determine next reaction event and time while in our approach, the next event selection and random execution time generation are computed differently. |
| **3** | **Update**: Increase the time step by the randomly generated time in Step 1.<br>Update the molecule count based on the reaction that occurred. | **Update**: The global simulation clock is increased by the time-step computed in the previous step as the event holding time.<br>The resource count of molecules are updated based on the last event stoichiometry | The temporal progression takes place in discrete time-steps based on the random event holding times computed in the previous step in our approach. |
| **4** | **Iterate**: Go back to Step 1 unless the number of reactants is zero or the simulation time has been exceeded. | **Iterate**: Go back to Step 1 and repeat the process. In case a particular event cannot be executed because of resource conflicts, it is ignored and simulation proceeds without the update step | The handling of reactions/events with resource conflicts/shortage is different in our approach |

## 3.2 *iSimBioSys* simulation framework

In this section, we develop the software implementation of our discrete event simulation platform, *iSimBioSys*, based on the methodology explained in the previous section. The modular nature of the functional blocks involved in our event based approach lends itself to an object-oriented computing paradigm [61]. Specifically, the Java based [84] implementation encapsulates the stochastic models for the different biological events and links them together in the discrete event simulator.

### 3.2.1 Event Objects

Each functional module is represented as an object, having its own state (the resources involved in the module) and its associated behavior, which is modeled on the functionality of the module. Another characteristic of a module are its associated input events, which drive the functionality of that module and its corresponding output events which are inputs to other modules. The central theme of a discrete event simulator revolves around the event queue, which is the global data structure responsible for storing time-stamped events for the simulation. The event queue maintains an event list containing the events to be executed. Instead of having each event store its corresponding execution time, each event is associated with the corresponding model object (an instance of a model class) which stores the first and second moment of the probability distribution associated with the event, e.g. the diffusion event is associated with the mean and variance of the probability distribution as computed in the model formalism outlined in the previous section.

A central scheduler is in control of the queue, popping events in a time-ordered manner to avoid "causal errors" [82] and sending it to the corresponding modules. At each event triggered, an instance of a random variable following the corresponding probability distribution is computed to calculate the event execution time for the particular event.

Based on the event execution logic, new events are created and pushed into the event list, updating the global simulation clock in the process. The scheduler is also responsible for maintaining the event list as events are generated by a module following its biological process logic. As is evident from the discussion, the scheduler together with the event queue drives the simulation environment while the module objects and their behaviors define the event handlers of the framework. Fig. 3.3 shows the flowchart for the discrete event simulation algorithm.

### 3.2.2 Software Components

Our current framework supports a multi-threaded architecture with the main simulation engine running in one thread while the visualization plane running on another. The basic architecture and framework of the simulation involves four logical packages, identified in the block diagram presented in Fig. 3.4:

- *In Silico Experimental Setup*: These set of classes are responsible for setting up the modeling and system parameters used in the particular simulation block and are generally provided through user interface or plain text files. While certain parameters are based on available biological literature such as cell volume, macromolecule diameters etc., event execution time parameters are computed by the engine internally based on the logic defined in the corresponding model class for each event.
- *Discrete Event Process Modules*: These set of classes, derived from a common base class, essentially the resource utilization algorithms for the biological process being simulated and provide methods to compute event holding times. It may be noted here that the discrete event process modules are a one-to-one mapped implementation of the functional modules. These event modules act on the system resources constructed in the knowledge extraction phase. In our current implementation,

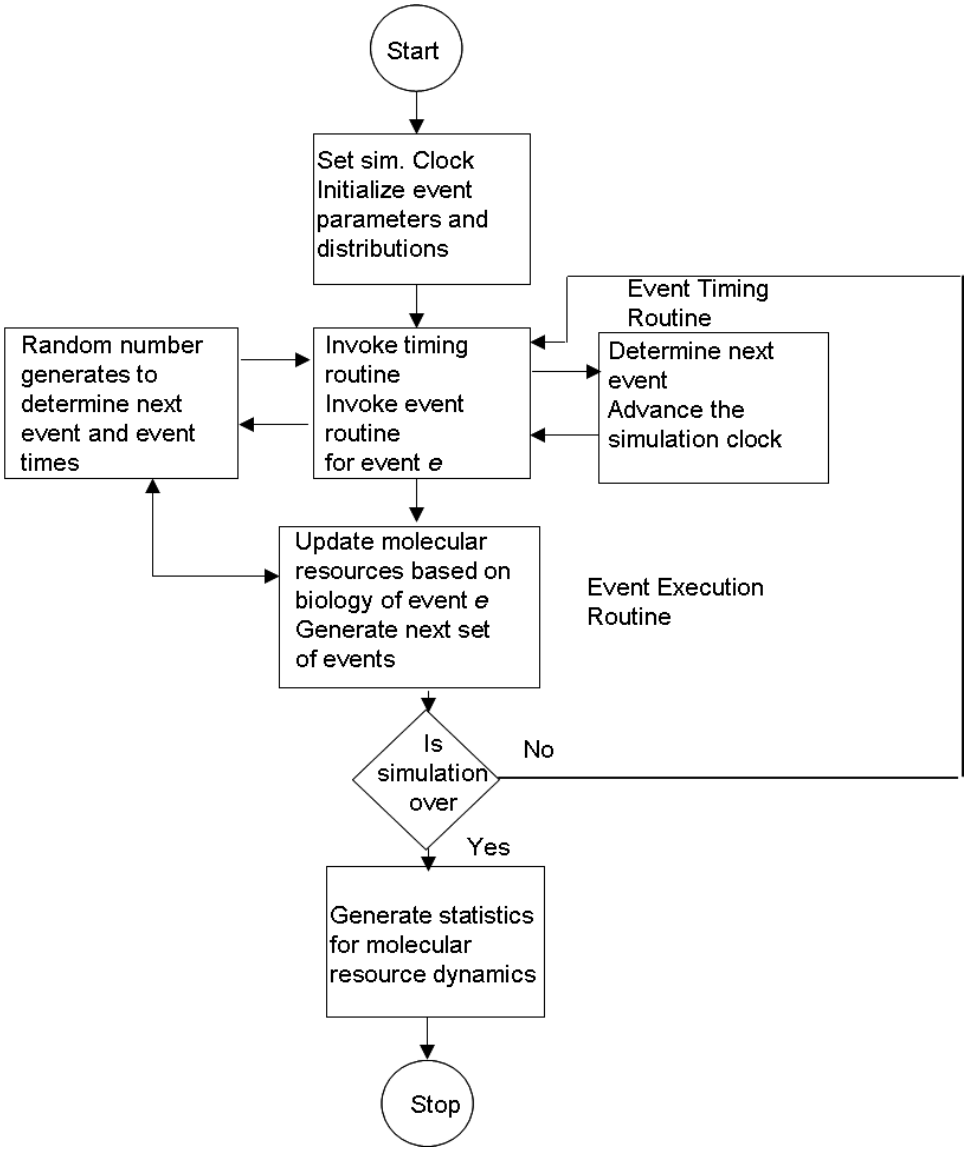



Figure 3.3. The discrete event based simulation algorithm.

the resources are modeled in a two dimensional data structure consisting of the resource state and its regulation logic (up or down regulation) based on the constructed pathway. As the event modules run in time, the resource states change and capture the dynamics of the system. These set of classes form the heart of the

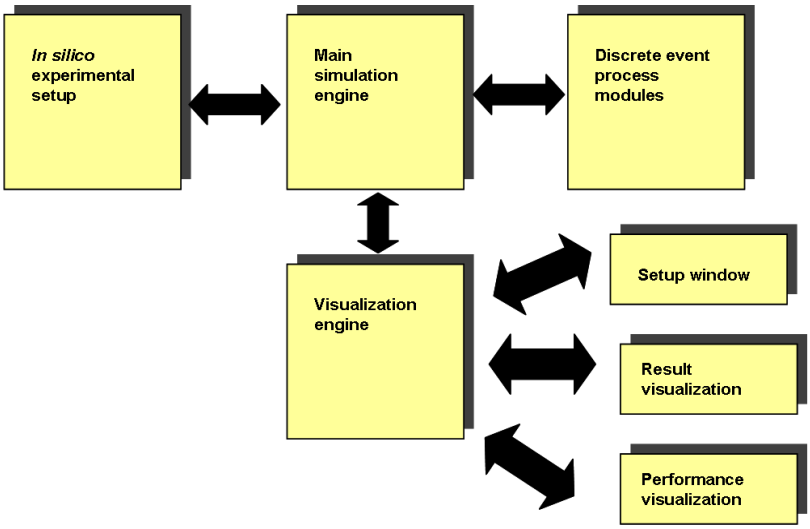



Figure 3.4. The *iSimBioSys* software architecture.

modeling formalism as they realize the stochastic behavior associated with each event in terms of its probability distribution.

- *Main Simulation Engine*: This class is responsible for handling the main thread of the discrete event simulator and implements the global event queue used. This class is responsible for communicating with the global event queue through the scheduler, executing the event process logic, updating the global simulation clock and exchanging resource state information with the visualization unit which updates the system behavior in real-time.
- *Visualization and Data Generation*: These set of classes are responsible for data generation of the simulation and tracing the simulation in terms of change in resource states in the temporal axis.The user interface of the current implementation involved three parts:
    1. *User Interface for experiment setup*: The user interface is presented before the start of the simulation for the user to set up system parameters, simulation runtime environments and visualization data.

2. *Runtime visualization of simulation*: The simulation can be traced in run-time in the visualization plane which runs on a separate thread as discussed earlier. Depending on user inputs, it traces the change in resource concentration of the system and also system signal states. As the dynamics of the system are traced in time, it provides a window for viewing the system behavior while the simulation runs in the background.
3. *Performance visualization*: These screens trace the various performance metrics of the simulation platform as it is executed. In the current implementation, it is trace of the memory and CPU usage of the system.

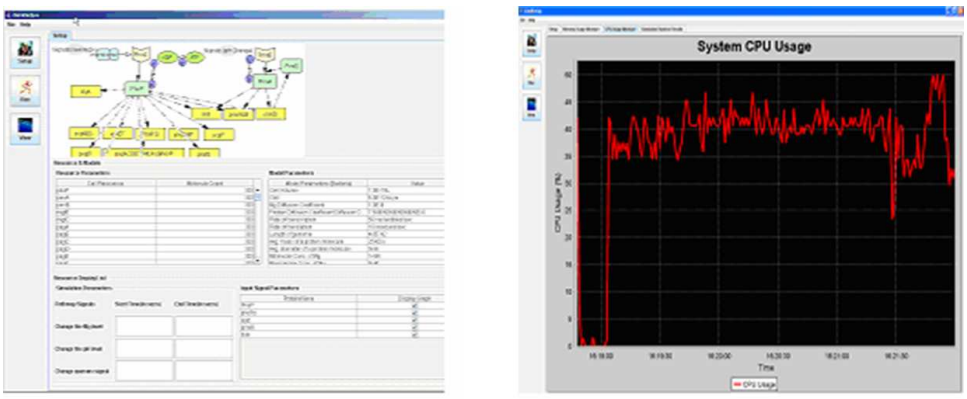



Figure 3.5. *iSimBioSys* software interface.

Fig. 3.5 shows a snapshot of the experimental setup and visualization screens part of the *iSimBioSys* simulation engine. It may be mentioned here that the current implementation of iSimBioSys is based on Java 1.5 SDK and runs on a windows XP service pack 2 (enterprise edition) based Dell XPS Dimension system (Intel Pentium 4 processor with HT technology running at 3.4 GHZ), 2GB DDR2 SDRAM at 533 MHz and 250MB ATI Radeon X850 XT PE video card.

## 3.3 Summary

In summary, our modeling and simulation technique presents a stochastic, event-driven framework which approximates the stochastic dynamics of the chemical master equation by parametric models of biological event time distributions. In this chapter, we provided the basic building blocks of the modeling and simulation paradigm and built the software framework. In the next chapter, we illustrate the simulation methodology, building its different components based on the case study of virulence gene regulatory and signal transduction pathways in the bacterial cell *Salmonella*.

# CHAPTER 4

# SIMULATING THE DYNAMICS OF SIGNAL TRANSDUCTION

One of the most important functions in a cell is the transduction of extra-cellular and intra-cellular signals. A complex set of molecular machinery working in close co-operation is responsible for sensing and transducing changes in environmental conditions of a cell. The cell reacts to these signals by employing different gene regulation and protein assembly mechanisms to maintain cellular homeostasis. Thus, studying the complex dynamics involved in signalling pathways and their downstream regulatory networks forms a fundamental step in the understanding of cellular behavior.

In this chapter, we employ the discrete event based modeling methodology to simulate the regulation of virulence gene in the bacterial cell *Salmonella typhimurium* when it invades a host cell, specifically the effect of external magnesium concentration on the two component PhoPQ virulence gene regulatory pathway. In section 4.1, we start with a brief description of the signal transduction and gene regulation process for this particular two-component system based on available biological literature [185]. Section 4.2 develops the mathematical abstraction of the key biological events and the discrete event simulation implementation of the abstraction, validating the *switching effect* of Magnesium signal on the signaling pathway as reported in experimental work while section 4.3 concludes the chapter.

## 4.1 Virulence gene regulation in *salmonella typhimurium*

Bacterial pathogenesis in *Salmonella Typhimurium* involves the complex interaction of regulatory pathways which play different roles in various stages of infection [121,

62]. While various signaling networks are involved in orchestrating pathogenesis in a host system, we focus on the two component PhoPQ regulatory pathway and its role in accomplishing parasitism of the host. [185] elucidates the role of extra cellular Magnesium ($Mg^{+2}$) concentration as a primary signal of this pathway which acts as a global regulator of *Salmonella* virulence and helps in the survival and replication of the bacteria in the macrophages (host cell), shown in Fig. 4.1. Low extra cellular $Mg^{+2}$ (micromolar concentrations) was shown to cause an increase in the expression of certain PhoPQ activated genes, while high $Mg^{+2}$ concentrations (millimolar) caused an immediate "switch off" of these genes. The knowledge available from the biological studies, together with the qualitative diagram of the system in Fig. 4.1 represents the biological process under study.

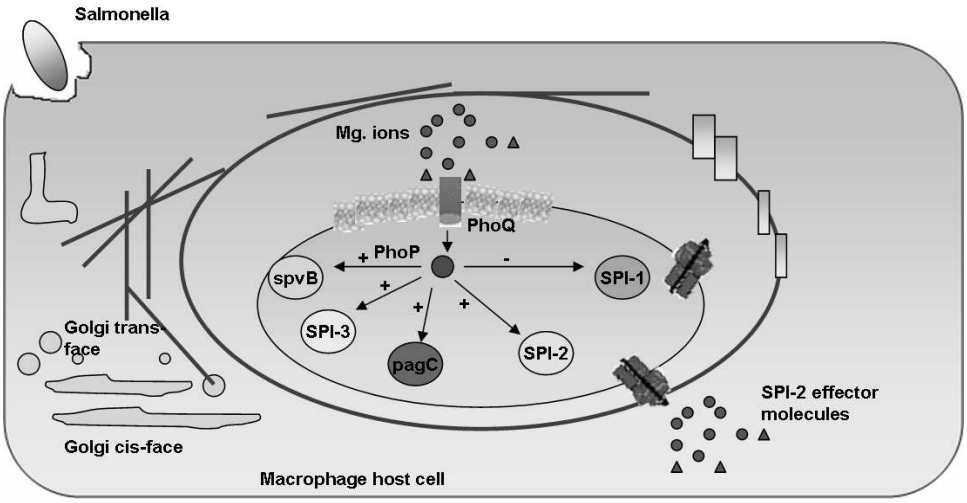



Figure 4.1. Virulence gene regulation in *salmonella*.

### 4.1.1 Modeling the two component pathway

Once the biological system has been defined, the modeling methodology outlined in the previous chapter is employed to translate the qualitative knowledge into a quantitative formalism, characterized by the events and the stochastic models of their execution time.

#### 4.1.1.1 Knowledge extraction and pathway construction

The first step in building a discrete event based model of a biological process is the extraction of molecular pathway information with subsequent construction of their interaction network. We used comprehensive knowledge extraction from PubMed [145] database, to construct the gene regulatory pathways for the PhoPQ network, identifying the common intersection of the pathways i.e. the genes and gene products which are regulated by this system at various stages.

For the PhoPQ two component pathway, the magnesium driven signal transduction is involved in transcriptional regulation of 44 genes, 5 of which are involved in another cascading two component system. A positive feedback loop exists in this pathway, in the form of up regulation of phoP gene by the phosphorylated PhoP protein. Fig. 4.2 shows the complete pathway, with the positive feedback loop marked in deep color. The pathways have been constructed using the Cell Designer 3.0 software which presents a structured (Extensible Markup Language (XML)) format data which can be easily rendered into the discrete event simulation framework. The gene regulatory pathway provides information on the molecular resources involved in the network as well as their biological interactions whose temporal dynamics drive the phenotypic response of the cell to the magnesium signal. Fig. 4.2 marks the first step in the transition of the qualitative knowledge of the biological system into a computational format captured in the pathway network structure.

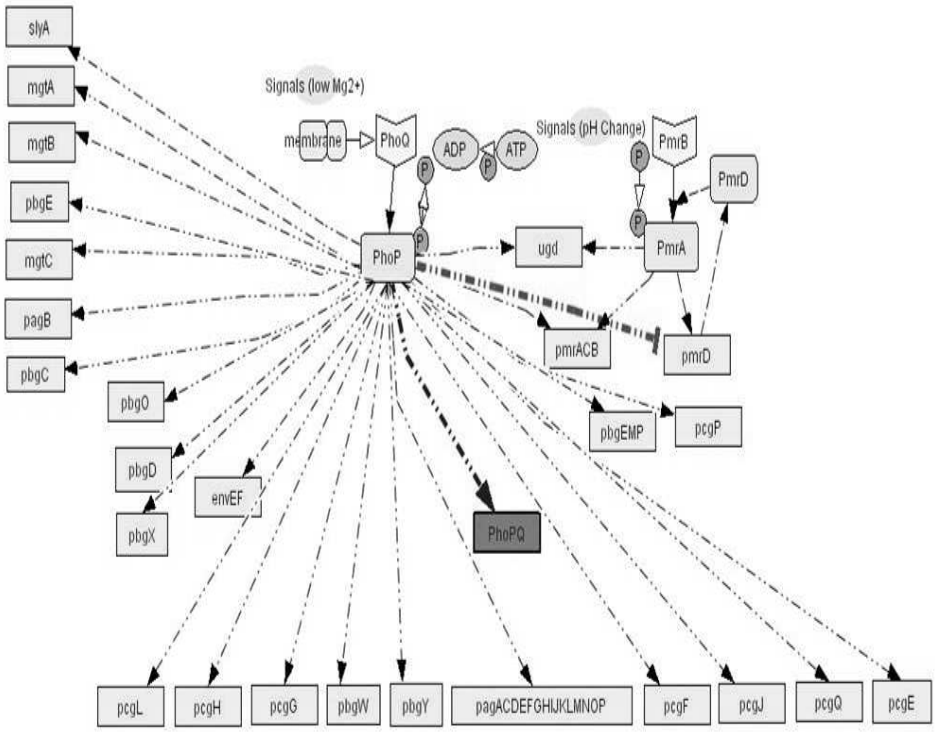



Figure 4.2. The PhoPQ pathway.

#### 4.1.1.2 Functional module identification of biological processes

The next step in model building is the abstraction of the biological functions associated with the signaling pathway. In this case, it translates into the basic processes which are involved in the activation of the PhoPQ system under external magnesium, follows by expression (up regulation) or repression (down regulation) of genes in the pathway. Based on a available literature [185, 121, 62], the main functional modules have been described in Fig. 4.3.

The biological process modules identified here are at different levels of granularity. For example, the autokinase activity [185] of PhoQ receptor molecules involves phosphorylation of a single PhoQ molecule. However, gene expression is a complex process, involving a large number of complex sub processes, all of which are not fully under-

stood currently. Thus, the functional modules need to incorporate these varying levels of granularity in their event models, which we illustrate next.

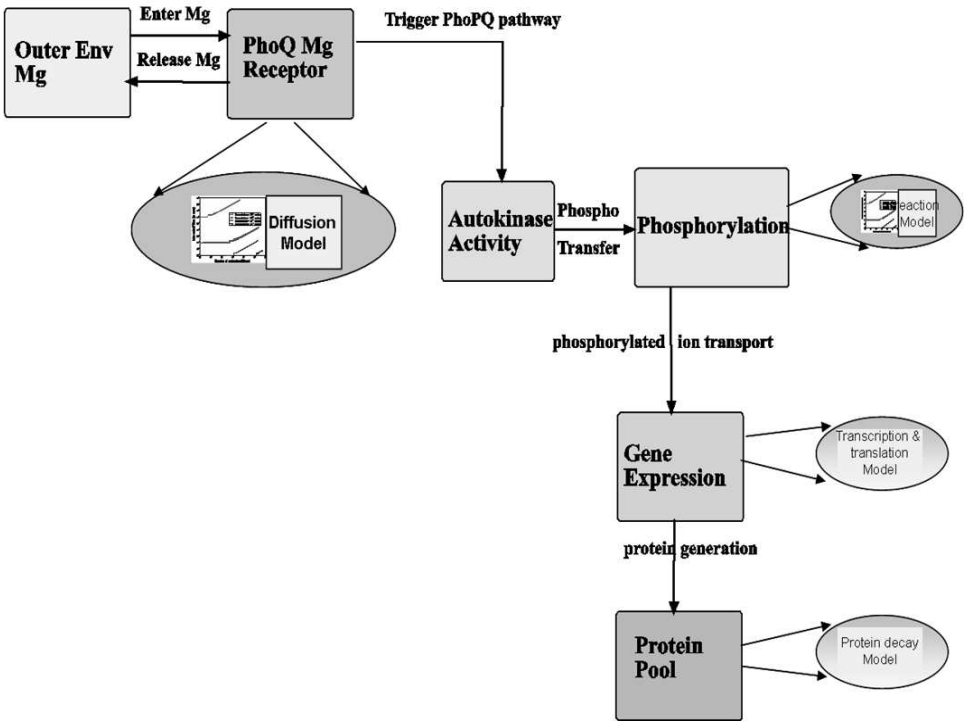



Figure 4.3. Event interaction network for the two component system.

#### 4.1.1.3 Stochastic event modeling

In a discrete event based approach, the mathematical formalism underpinning the simulation is the stochastic modeling of the events associated with the biological processes. As mentioned in the previous chapter, the modeling of the event holding time of the functional modules (the arrows between the modules in Fig. 4.3 denote the events), is a key distinguishing step in our methodology. The random holding time is generated as an instance of the probability distribution associated with a particular event. This parametric distribution, defined in terms of its first and second moments, is computed

from the stochastic modeling of the biophysical and biochemical properties of the process (elliptical modules in Fig. 4.3) and forms the heart of the stochastic event modeling step.

Stochastic models for the different events involved in the signaling and gene regulatory pathway of the PhoPQ system have been developed as part of the simulation framework. The simulation is capable of incorporating models in varying degrees of granularity and abstraction. Below, we summarily present the stochastic model formalism for two key processes used in this study at different levels of abstraction. More details on the stochastic models are available in [136, 137, 138, 139].

- *Transfer of magnesium ions through the cell membrane*: As the PhoPQ pathway is controlled by extra-cellular magnesium ion concentration, the movement of $Mg^{2+}$ through the cell membrane needs to be modeled. This event is modeled [135] as diffusion of charged ions through the cell membrane. Specifically, the event time for a molecule of $Mg^{2+}$ entering or leaving the membrane needs to be computed. This deals with the inter-arrival (departure) time between two molecules or ions, where their movement to/from a cell is controlled by concentration gradient and ion charge potential gradient across the membrane. The inter-arrival (departure) time is controlled by the ion flux in this case. To derive the inter-arrival (departure time between the $i^th$ and $(i+1)^th$ molecules i.e. $t_{i+1} - t_i$ , we determine $t_{N-i+1}$ and $t_{N-i}$ that denote respectively the times to transfer $N-i-1$ and $N-i$ molecules/ions through the channel, where $N$ is total number of molecules/ions. Now,$t_{N-i}$ can be obtained by solving the following equation as reported in [132, 135]:

$$N - i = 2 \times C_0 \times A\Sigma_{m=0}^{\infty}(m^2\pi^2 \frac{1-(-1)^m e^{\frac{-zFV}{2RT}}}{\frac{z^2F^2V^2}{4R^2T^2l^2} + m^2\pi^2}^2 e^{-\beta t_{N-i}}) \tag{4.1}$$

where $A$ is the area of the membrane sheet; $F$ is Faraday's constant; $T$ is the absolute temperature; $R$ is gas constant; $z$ is the total number of positive charges; $l$ is the length of the ion channel; and $\beta$ is,

$$\beta = (\frac{z^2F^2V^2}{4R^2T^2l^2} + \frac{m^2\pi^2}{l^2})D \tag{4.2}$$

The parameterized equations in Eqn.4.1 and Eqn.4.2 capture the different physico-chemical factors affecting the diffusion of charged ions. Based on the parameter values for diffusion of $Mg^{2+}$ ions through the bacterial cell membrane, the inter-arrival time between $Mg^{2+}$ ions can be computed for different concentrations of extra-cellular magnesium.

- *Gene expression modules*: Next, we focus on the complex module of gene expression and protein synthesis which orchestrate the expression dynamics of the different genes involved in a the regulatory pathway. The stochastic nature of gene expression and the multitude of factors both at transcription (RNA polymerase copy number), translation (competition between ribosome and RNaseE molecule for translation initiation or decay respectively) as well as post-translational stages pose modeling challenges in this complex molecular assembly phase. In this case study, the holding time in these blocks have been modeled based on existing experimental data, collected for average bacterial transcription time and translation times. The complex process of protein formation and decay have been modeled as an exponential distribution with the exponent computed based as function of the number of proteins in the system and its half life values, which depends on the conformation and residue length of a particular protein.

  While such modules at varying degree of model granularity can co-exist in our framework, in the next chapter we provide a detailed stochastic model for prokary-

otic gene expression, accounting for the different molecular machinery controlling the process.

- *Discrete event simulation*: This is the heart of the framework, comprising of the core simulation engine responsible for driving the system *in silico*. Based on the functional modules, the key events driving the interaction of these modules are identified. The event times associated with each of these biological events are developed based on the stochastic modeling techniques. The discrete event platform, *iSimBioSys*, elucidated in the previous chapter, incorporates these information in its framework.
- *In silico result and wet lab verification*: The success of the simulation methodology depends on the validation of results with wet lab experiments. This provides a tool for verification of biological processes and for subsequent hypothesis testing of biological functions prior to more advanced and costly in vitro and in vivo experiments.

The components elucidated in this section, powered by large databases of molecular knowledge, iteratively interact to form an *in silico* modeling and simulation platform.

## 4.2 Experimental validation and hypothesis testing

The efficacy of the modeling and simulation approach is governed by (a) validation of the model against existing wet-lab experimental results, (b) effective calibration and sensitivity analysis of the key parameters governing the biological model and (c) hypothesis testing of different conditions on the biological system which can give further insights for novel experiments in the future.

In this section, we employ the discrete event based stochastic simulation framework to model the effect of the PhoPQ two-component signal transduction pathway on the expression of virulence genes involved in bacterial pathogenesis of the gram-negative

bacteria *Salmonella Typhimurium*. While the simulation system can be used to model the temporal dynamics of different regulatory pathways in a bacterial cell, we focus on the particular system in this work as it provides:

- Existing wet-lab experimental setup and results [185] which allow the validation of the simulation results.
- The system involves the interaction of signal transduction with subsequent expression of genes governed by the upstream signals.
- The gene regulation pathway as built based on existing literature on the two-component system provides various regulatory mechanisms including up and down regulation of genes, and positive feedback effects which can serve to test different hypothesis.
- As the system involves complex biological functions like gene regulation and protein expression, whose exact molecular mechanisms are not always well known, it provides a platform to test the efficacy of granular model abstraction based on available knowledge, on the behavior at a systems level.

In the rest of the section, we start with a brief description of the wet lab experimental system, moving on to present the detailed results of simulation. We show how the discrete event simulation framework can be used for hypothesis-driven analysis of different conditions for the PhoPQ system.

### 4.2.1 The wet lab experimental system

The experimental setup, explained in details in [185], consists of reporting the phenotypic output of the virulence gene expression pathway (measured in terms of expression level of the *phoP* gene). Fluorescence measure of expression of destabilized green fluorescence protein (dEGFP) under the control of a PhoPp (phosphorylated PhoP) responsive promoter was used as the reporter system. Thus, the system measure of the dEGFP was

in essence an indication of the PhoPp concentration in the system. In the experimental system, low $Mg^{2+}$ was maintained for a period of 60 mins, during which the system output increased, after which the signal was toggled to high $Mg^{2+}$. The measurements of the fluorometer were taken every 15 mins for the positive activation state.

Fig. 4.4 shows the system output of the cell culture in time, both for high-magnesium as well as low-magnesium conditions, representing the "switching effect" of the magnesium signal. Fig. 4.5 shows the system behavior as observed when the cells were in a culture of low (8 microgram) magnesium medium, highlighting the activation of the PhoPQ pathway (as shown by increase in concentration of PhoPp protein). Similarly, Fig. 4.6 shows the toggling effect of the 'on-off' switch mechanism when the system state was changed from high to low magnesium medium.

Based on these experiments, we run the discrete event simulation to generate simulation results which capture the system output in time. The simulation initialization with different resource and system parameters are key to the success of the model. Also, the platform provides flexibility in changing these conditions and resources to generate synthetic, hypothetical results for a better understanding of the test system. In the next subsection, we outline the system and simulation parameters and present the results of the *in silico* experiment.

#### 4.2.1.1 *In Silico* validation

In this section, we setup the 'dry-lab' experimental system for the signal transduction and subsequent gene regulation pathway involved in the test-bed. The *in silico* experiment is initialized with the system molecular resources and biological parameters associated with the probability distribution functions of the different event holding time modules. In this experiment, we focused on parameters associated with the *Salmonella* bacterial cell based on the CCDB database [170] which are summarized in Table 4.1.

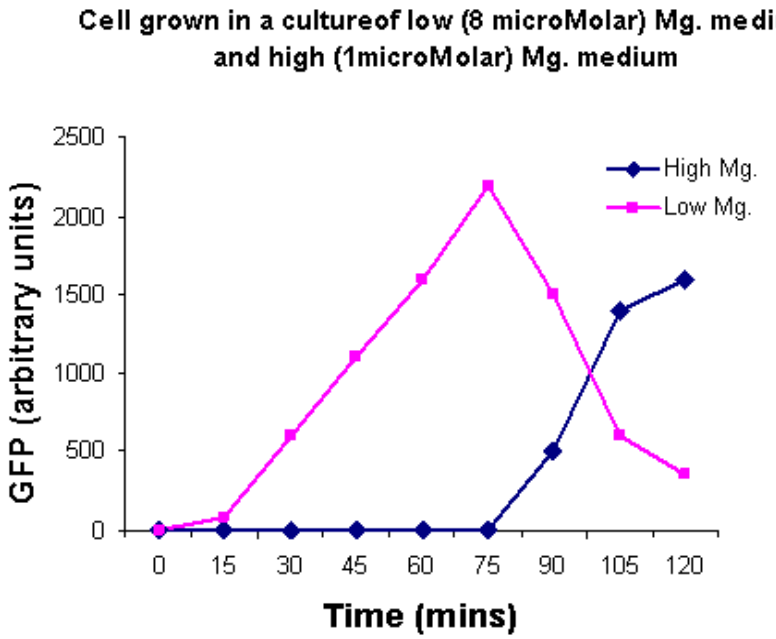



Figure 4.4. Effect of Mg2+ on the system output.

The simulation also initializes other resource parameters like the number of molecules (in terms of concentration) for the different species involved in the system (e.g. ATP, ADP, PhoP, PhoQ, extracellular $Mg^{2+}$ ions) and the gene regulatory pathway information extracted during the PhoPQ pathway creation phase. Once the system is initialized, the event queue is populated with the initial event list which determines the snapshot of the biological environment at simulation start time and the simulation engine is triggered.

For the current system, the simulation focused on tracing the effects of signaling events ($Mg^{2+}$ ion arrival and departures) on the expression dynamics of the PhoPQ pathway. Also, as a reporter protein (GFP) has been used in the wet-lab scenario to trace the system behavior, our results are focused primarily on PhoPp as the main resource whose dynamic temporal behavior was observed in the simulation. Although, the simulation can be configured to monitor and generate results for a wide range of

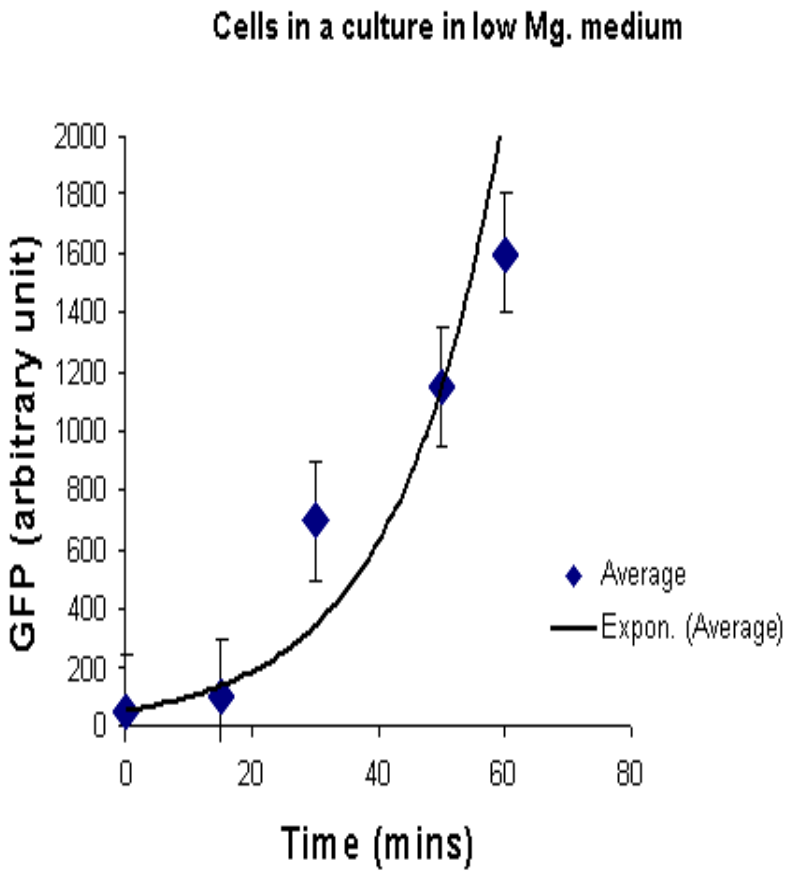



Figure 4.5. Effect of low Mg2+ (8microM) on the system output).

system resources, PhoPp was chosen primarily to verify the wet-lab tests. The simulation experimental results denote resource states averaged over 100 runs of the simulation under the same initial conditions.

In order to simulate similar conditions, the simulation was configured to run with low Mg2+ for 60mins, during which no resource conflicts or starvation were assumed (i.e the simulation would not stop due to lack of any resource). As seen in Fig. 4.7, the simulation responds with continuous growth in PhoPp concentration, implying increasing dEGFP fluorescence.

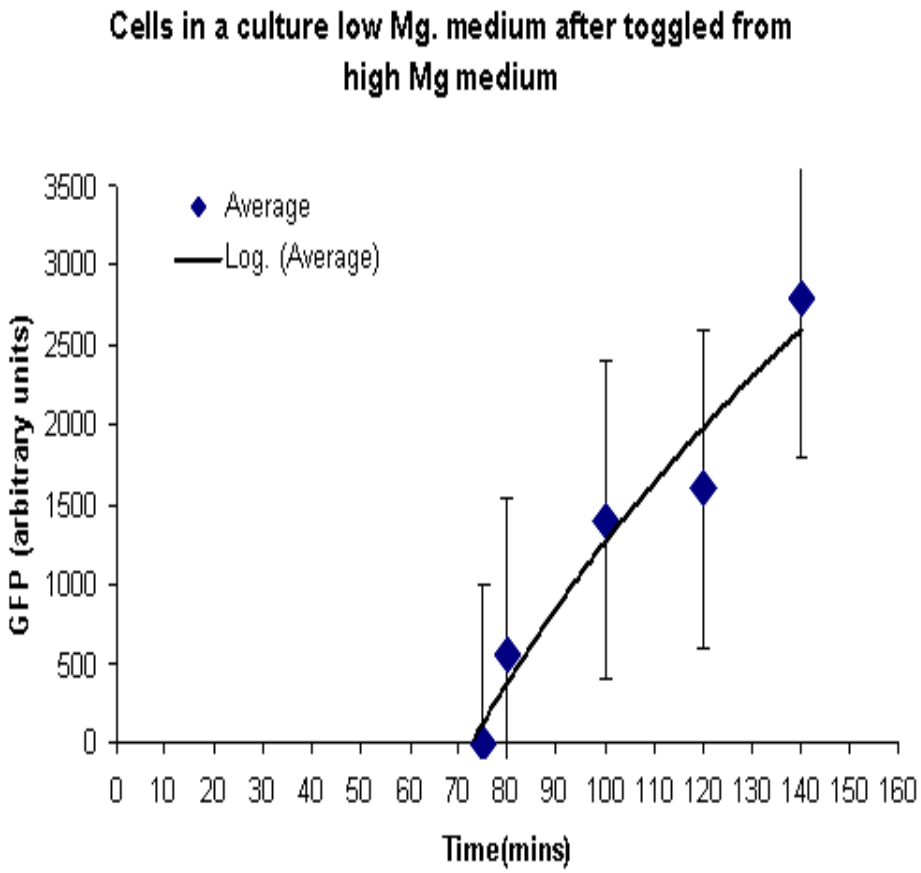



Figure 4.6. Effect on the system output of Mg. switch.

In another simulation experiment, the system was started with high Mg2+ which was switched to low Mg2+ at 20mins which was kept low for 30 mins. and toggled back to high. Fig. 4.8 captures the system response under this scenario, recreating the "toggling effect" of the $Mg^{2+}$ signal on the pathway. The condition of no resource starvation shows relative smoothness in output as obtained from continuous system models since the effect of low copy number of molecules on stochasticity [112] is not displayed.

The simulation platform allows the analysis of the effects of stochasticity on the model by varying the resource states of the molecules involved in the simulation and

Table 4.1. Experimental parameters for the *Salmonella* bacterial cell

| | **Biological Parameter (Bacteria)** | **Value** |
|---|---|---|
| 1 | Length Of Genome | 4857432 |
| 2 | Number Of Genes | 4451 |
| 3 | Rate Of Transcription | 50 Nucleotides/Sec |
| 4 | Rate Of Translation | 18 Residues/Sec |
| 5 | Area Of Cell | 6*10-12 Sq.M |
| 6 | Volume Of Cell | 1*10-15 L |
| 7 | Diffusion Co-Efficient Of Magnesium Ion | 1*10-9 |
| 8 | Diffusion Co-Efficient Of A Protein Molecule | 7.7*10-6 |
| 9 | Avg. Mass Of A Protein Molecule | 25kda |
| 10 | Avg. Diameter Of A Avg. Protein Molecule | 5 Nm |
| 11 | Millimolar Conc. Of Mg | 1.0*10-3 |
| 12 | Micromolar Conc. Of Mg | 8.0*10-6 |
| 13 | Phosphorylation Reaction Time | 5.6*10-9/ (No. Of Reactant Molecules) Secs |
| 14 | Avg. Delay Between Diffusion Of Two Mg. Molecules | 8.5*10-10 Secs |

also the sensitivity of the system outputs to the different parameters governing the event holding time distributions. In the next sub-section, we present a systematic analysis of the different hypothesis tests.

### 4.2.2 *In silico* hypothesis testing

The *in silico* simulation model allows the modeler to test the system under various synthetic conditions, in terms of system resource states, initial conditions and different combinations of environmental cues driving the systems (for example, the diffusion of magnesium ions through the cell membrane in our case study).

In order to capture the effects of varying the rate of diffusion of magnesium on the system output, we ran the simulation with increasing magnesium ion diffusion times (100ms, 1ms,10ms) and reported the results for two key system resources, the proteins

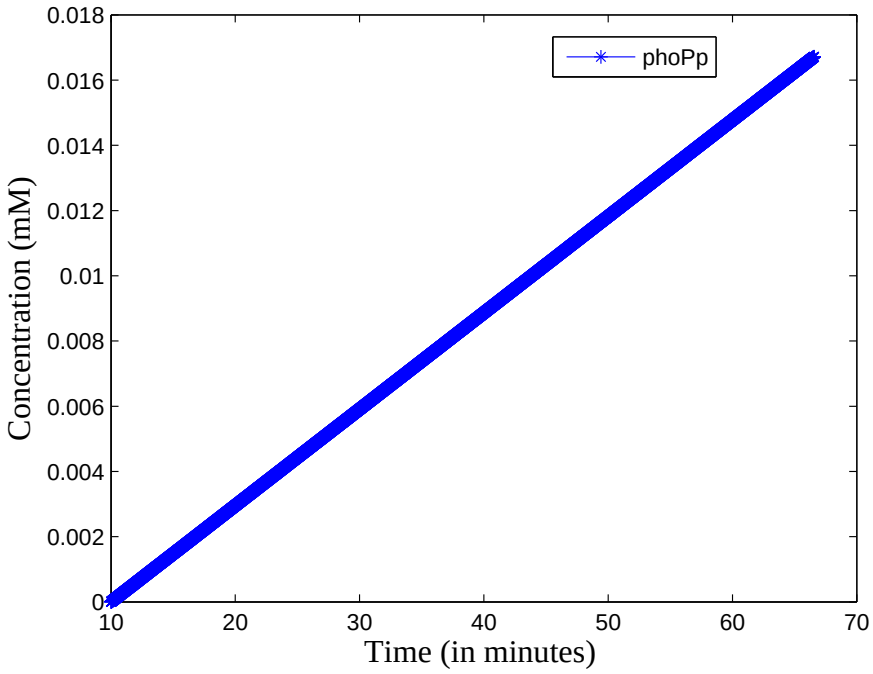



Figure 4.7. Effect of low $Mg^{2+}$ on the *in silico* system.

PhoQ, which is the sensory protein responsible for binding to magnesium ions, and the PhoP protein, which controls the dynamics of the gene expression. Fig. 4.9 shows how the rate of decrease in the concentration of inactive PhoQ (phoQ molecule bound to a magnesium ion) is damped with increasing delay in the diffusion of magnesium ions out of the membrane. Also, capture in this graph is the effect of resource starvation on the biological system. As the Mg. ion initiated signal activates the PhoPQ pathway, the sensory PhoQ proteins are consumed by the system, thereby shutting down the pathway when all phoQ molecules available to the system have been used. Similarly, Fig. 4.10 captures the effect of the same conditions on PhoP.

An interesting observation, not capture in the wet-test lab results, is the orchestration of the positive feedback loop of PhoP, as identified in the knowledge extraction phase. As seen in Fig. 4.10, the concentration of PhoP in the system decreases initially;

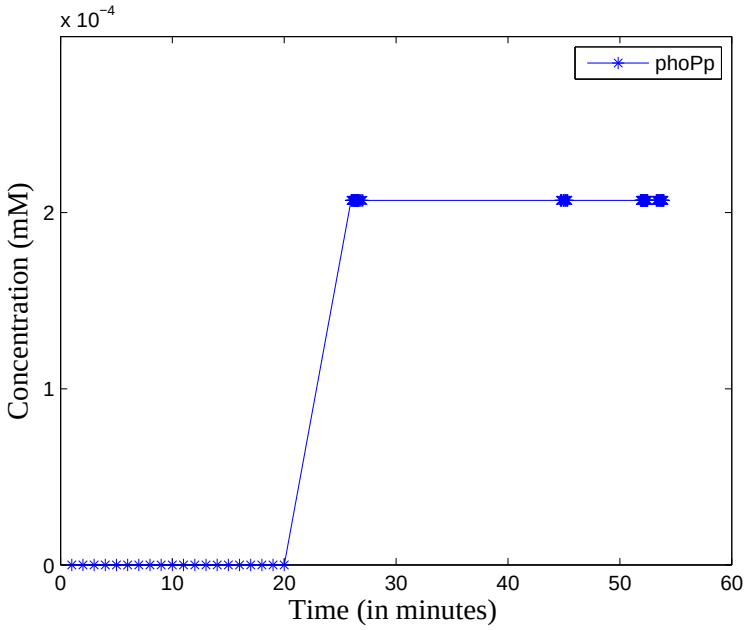



Figure 4.8. Simulation results on the 'switching effect" of $Mg^{2+}$ signal.

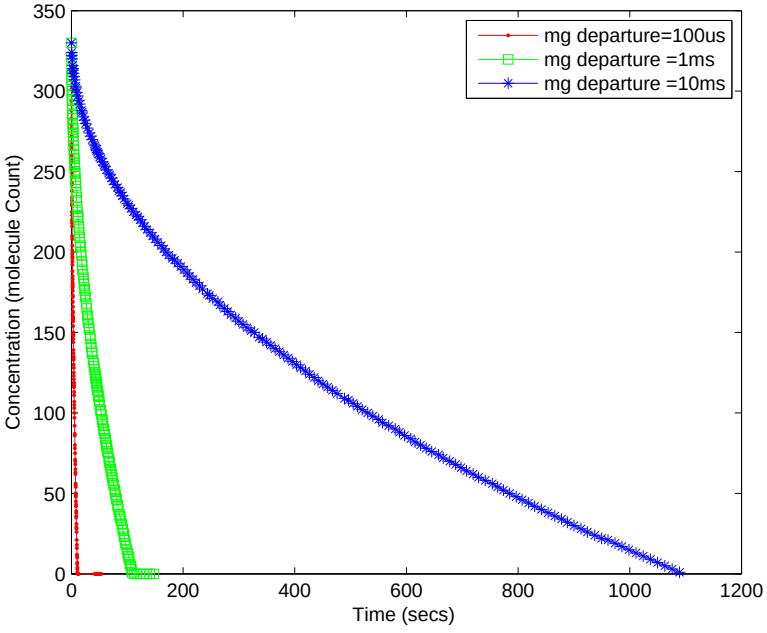



Figure 4.9. Change in conc. of membrane PhoQ.

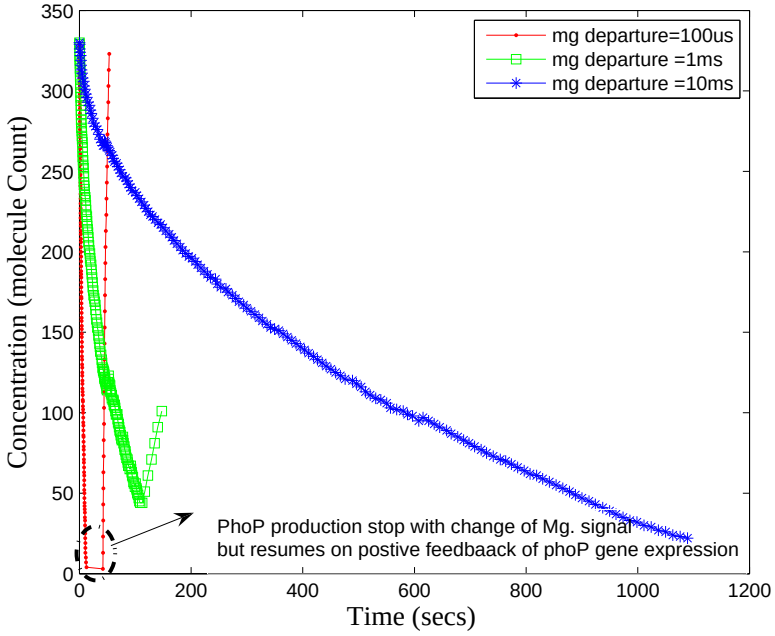



Figure 4.10. Change in conc. of membrane PhoP.

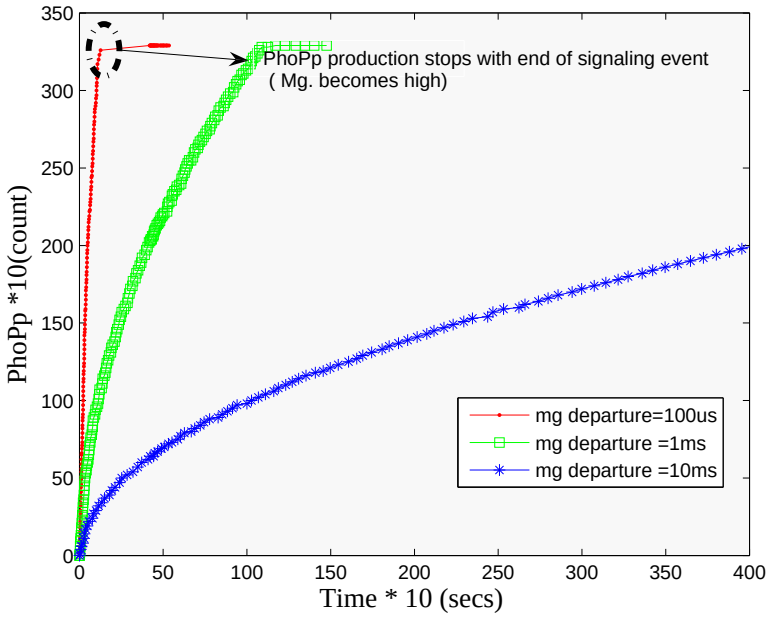



Figure 4.11. Change in conc. of PhoPp.

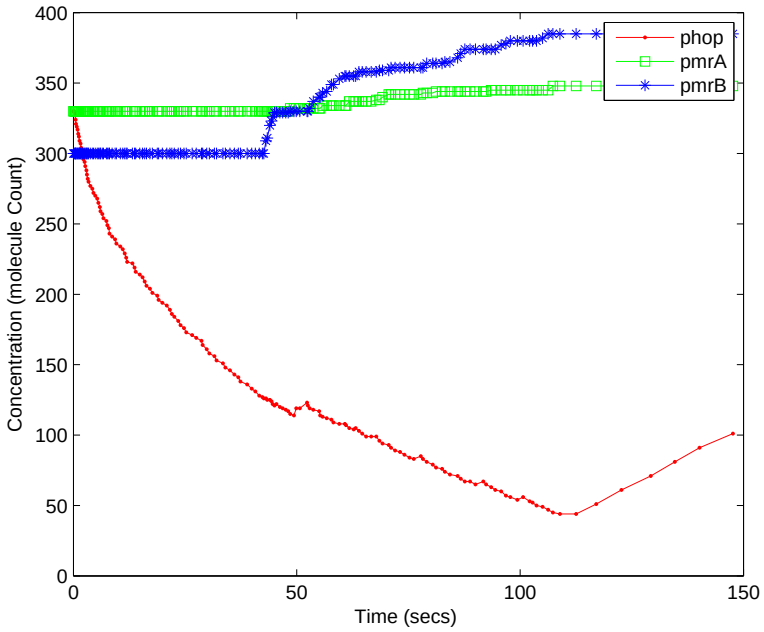



Figure 4.12. *In silico* gene expression profile.

but once, the expression of genes is triggered by phoPp (phosphorylated PhoP), PhoP starts appearing in the system. The corresponding effect on PhoPp, which increases in concentration when magnesium ions depart from the membrane (activating the pathway) is captured in Fig. 4.11. In both the graphs, the slowest rate diffusion does not bring the system into resource shortage phases while the other diffusion rates locks the system (plateau on Fig. 4.11) due to non-availability of phosphorylated PhoP molecules. These results show how the tuning of different parameters (in this diffusion rates) can be synthetically manipulated to study different behaviors of the systems.

Another *in silico* result, which is possible in our simulation framework is the profiling of different resources, which though key to the system as a whole, but may not be the focus of a current experiment. For example, it is possible to profile metabolites and energy molecules like ATP and ADP, to name a few. Also, the expression profile

of a whole range of gene products, like proteins and enzymes can be traced in the simulation, providing a gene profiling microarray. In Fig. 4.12, we show the protein profile of 3 proteins in our test bed pathway, as they unfold in time. The protein expression profile captures the stochastic fluctuations of the PhoP molecule as the system progresses in time, triggering the positive feedback effect of the phoP gene on the two-component pathway.

## 4.3 Summary

The *in silico* results on the test-bed pathway demonstrate the efficacy of the modeling and simulation approach for study single cell dynamics. Particularly, the flexibility in event scheduling and resource state specifications allows a modeler to validate the effects of high and low copy number of molecules on different parts of the biological system. This flexibility allows the simulation to be computationally efficient depending on the required granularity of the biological model and the resource state space considered. In the next chapter, we show how the simulation framework is capable of capturing high level details of granularity in terms of modeling biological events using the fundamental process of gene expression in prokaryotes as the model system.

# CHAPTER 5

# MODELING PROKARYOTIC GENE EXPRESSION

One of the salient features of an event based modeling paradigm is the ability to abstract biological complexity of a particular function at different levels of granularity depending on available knowledge. In the previous chapter, the events of the two component system included gene expression and protein generation as two functional blocks which were modeled with constant holding time based on experimental rate parameters for transcription and translation. However, the process of gene expression, which forms the central dogma of molecular biology, involves complex interactions between a large number of molecular actors. In order to capture these details in a parametric form, it is pertinent to model transcription and translation processes at a higher level of granularity while incorporating the available biological knowledge. In this chapter, we focus on developing a stochastic model for prokaryotic gene expression and study its dynamics using the discrete event simulation technique elucidated in the previous chapters.

Fluctuations in protein number (noise) caused by the stochasticity in gene expression plays a central role in the dynamic behavior of cellular pathways. Deterministic models capture average cell population behavior and are limited in their relevance in modeling stochastic deviations of gene expression in single cells. In this chapter, we develop a birth and death Markov chain model to capture the discrete biological events of transcription and translation in prokaryotic cells. Section 5.1 gives a brief overview of the prokaryotic gene expression process while section 5.2 outlines existing modeling approaches. We derive mathematical models for the expression 'burst frequency' distribution as well as the number of protein molecules per burst in section 5.3. We validate

our stochastic models with recent single cell experiments on the *lacZ* gene in *Escherichia Coli*(*E.Coli*) in section 5.4 and characterize expression noise sensitivity to biological parameters like gene activation ratio in section 5.5. Further, we build a discrete-event stochastic simulation system to study the transient dynamics of *lacZ* gene expression in section 5.6, quantifying the role of promoters in controlling the 'burstiness' of protein synthesis in section 5.7.

## 5.1 Dynamics of gene expression

One of the key goals in studying complex cellular pathways is the understanding of the dynamic interaction between the cell's gene regulatory and metabolic networks. In particular, the dynamics of protein-coding genes play a vital role in controlling the expression patterns of other genes encoding regulatory proteins and metabolic enzymes. Continuous and deterministic differential equation-based models representing the discrete events of transcription and translation have been traditionally used to study gene expression in cell populations [80]. However, the inherent stochasticity in molecular interactions limits the applicability of these models in studying single cell expression deviations and observed cell-to-cell variability [80, 106, 79].

Stochastic fluctuations in the expression pattern of genes have been mathematically studied by Arkin et. al [1, 72] and experimentally observed in [98, 93]. In particular, the 'burstiness' in protein production i.e. proteins are produced in random bursts of short duration rather than in a continuous manner, have been studied by various researchers [80, 106, 109, 4]. The random fluctuations in the number of proteins, termed 'noise', stems from the interplay of a large number of factors: discrete, random nature of molecular interactions like promoter binding and transcription open-complex formation, low copy number of key transcriptional and translational machineries like RNA polymerase, transcription factors, ribosomal units etc., and the random nature of signals

triggering gene expression. We provide an overview of stochastic models, primarily based on Monte Carlo simulation of the biochemical reactions involved in gene expression [7, 4], in the next section.

We model the events of transcription and translation for prokaryotic cells as discrete space, continuous time Markov processes, computing the probability distribution of the two key parameters characterizing the expression of proteins from a gene [93]: (a) frequency of messenger RNA (mRNA) bursts per cell cycle, and (b) the number of proteins molecules per burst. To understand the transient dynamics, we further integrate our mathematical equations into a stochastic discrete event simulation of coupled gene transcription and translation events.

## 5.2 Stochastic models of gene expression

As mentioned in the previous section, gene expression is an inherently stochastic process, governed by random association and dissociation of transcription factors, RNAP polymerases, ribosomes and degradosomes (RNase E). etc. caused by the low copy number of molecular entities. In this section, we briefly overview Monte Carlo simulation and other stochastic models for prokaryotic gene expression, focusing on the molecular actors involved in transcription and translation events as depicted in Fig. 5.1.

Specifically, the events of gene activation followed by transcription initiation, elongation and termination leading to the production of an mRNA molecule are considered. With respect to the translational machinery, the molecular events involved in translation initiation by ribosome binding/unbinding, mRNA decay by degradosome binding, translation elongation and termination are accounted for.

A common theme driving mathematical models studying the stochastic dynamics of gene expression is the elucidation of the different molecular actors affecting transcription and translation through a set of bio-chemical equations. In Fig. 5.2, we provide a listing

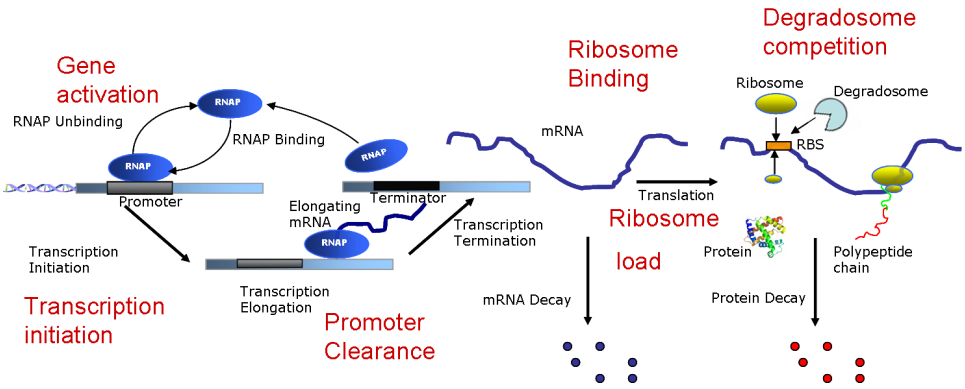



Figure 5.1. Molecular events involved in bacterial gene expression.

of the set of reactions involved in gene expression, including mRNA and protein decay, which have used (in different partial sets) in various reaction models [4, 109, 92, 27]. The reaction model has been generated using the Chemical Model Definition Language (CMDL) format provided by Dizzy [161]. Once the reaction set is defined together with rate constants estimated from experimental and available data, stochastic Monte Carlo simulations (Gillespie algorithm [40]) are used to study their behavioral dynamics. McAdams and Arkin [72] combined a continuous model of transcription initiation with a stochastic model for subsequent processing and protein synthesis capturing the race between translation initiation and mRNA degradation. Kierzek et. al [7] systematically studied the effects of transcription and translation efficiencies for bacterial gene expression through Monte Carlo simulations. In [27], the authors have developed stochastic pi-calculus based techniques for studying gene expression dynamics. Paulsson [92] provides a comprehensive review of different models of gene expression and protein noise analysis, giving a common Fluctuation-Dissipation Theorem (FDT) based framework to encompass various models. Noise characterizations (intrinsic and extrinsic) in single cell gene expressions have also been in studied in [80, 109].

```
// transcription reactions

TF_binding, PROM_IA -> PROM_A,  activationRate;
TF_unbinding, PROM_A -> PROM_IA, deactivationRate;
complex_formation, RNAP + PROM_A -> RNAP_PROM_A_CLOSED , rnapComplexFormationRate;
complex_unbinding, RNAP_PROM_A_CLOSED -> RNAP + PROM_A,  rnapComplexUnbindingRate;
transcription_initiation, RNAP_PROM_A_CLOSED -> RNAP_PROM_A_OPEN, transcriptionInitiationRate;

loop (i, 1, basePair)
{
        "RNAP_DNA_[i]"   = 0;
}

elongation_start, RNAP_PROM_A_OPEN -> "RNAP_DNA_[1]" , transcriptElongationRate;

// mrna chain elongation

loop (i, 1, basePair-1)
{
"RNAP_moves_DNA_[i]", "RNAP_DNA_[i]" -> "RNAP_DNA_[i+1]",  transcriptElongationRate;
}

RNAP_DNA_terminate,  "RNAP_DNA_[basePair]" -> RNAP + mrna, transcriptElongationRate;

// translation reactions

loop (i, 1, AAlength)
{
        "Rib_MRNA_[i]"   = 0;
}


ribosome_binding, ribosome + mrna -> rib_mRNA, ribosomeBindingRate;
ribosome_unbinding, rib_mRNA -> ribosome + mrna, ribosomeunbindingRate;
translation_initiation, rib_mRNA -> "Rib_MRNA_[1]", translationInitiationRate;

loop (i, 1, AAlength-1)
{
  "RIB_moves_MRNA[i]",  "Rib_MRNA_[i]" -> "Rib_MRNA_[i+1]", translationElongationRate;
}

Rib_terminate, "Rib_MRNA_[AAlength]" -> ribosome + protein, translationElongationRate;

// mrna decay
//degradosome_binding, degradosome + mrna -> degradosome_mrna, degradasomeBindingRate;
//mrna_decay, degradosome_mrna -> dead_mrna, mRNADecayRate;

//protein decay

protein_decay, protein -> dead_protein, proteindecayRate;
```

Figure 5.2. Reaction model implementation.

We consider the molecular process of prokaryotic gene expression as a complex stochastic system and abstract the process through probability distributions for the rate of mRNA and protein synthesis. While the estimation of the various kinetic parameters involved in reaction models pose a challenge under different conditions, our stochastic model captures the available information to quantitatively characterize the gene expression process in terms of the average rates of mRNA synthesis and protein generation.

## 5.3 Birth-death markov chain model of gene expression

In this section, we build a stochastic model for gene expression, considering the molecular actors affecting expression dynamics. Stochastic behavior arises in cellular reactions from the individual randomness of molecules in motion and the activation energy required to form the complex when two molecules come close to each other. Our modeling approach is motivated by the fact that the system dynamics of such processes can be captured by algebra of random numbers.

To make the process mathematical tractable, we use the following set of assumptions:

- The effect of promoter arrangement, as in an operon, or regulon or tandem promoters has not been considered.
- The transcription machinery is not limited by the number of RNA polymerase (RNAP), transcription factors etc. available in the system.
- The process of ribosome assembly is not considered and translation is assumed to be initiated by an active ribosome. The translation process is not rate limited by the number of ribosome or amino acids present in the system.

Table. 5.1 outlines the notations used in our models together with their definitions.

Table 5.1. Mathematical notation table

| Notation | Definition |
|---|---|
| $X_{mR}(t)$ | denote the number of RNAP molecules attached to the gene-coding region of the DNA at time $t$ |
| $S_{mR}$ | set of possible states of RNAP molecules |
| $N_{RNAP}$ | max. no. of RNAP molecules attached |
| $L_{gene}$ | length of genes (bp) |
| $RNAP_{footprint}$ | spacing between RNAP molecules |
| $\lambda_i^{mR}$ | transcription birth rate |
| $\mu_i^{mR}$ | transcription death rate |
| $T_{binding}$ | RNAP molecule binding time |
| $T_{clearance}$ | RNAP molecule clearance time |
| $k_{elong}^{mR}$ | transcript elongation rate constant |
| $k_{on}^{mR}$ | transcription activation rate constant |
| $k_{decay}^{mRNA}$ | transcript decay rate constant |
| $P_n^{mR}$ | probability of $n$ RNAP molecules attached to a transcript |
| $\bar{R}_{mR}$ | Avg. rate of transcript synthesis |
| $\underset{mR}{\sigma}$ | Variance of transcript synthesis rate |
| $X_p(t)$ | denote the number of ribosome molecules attached to a transcript at time $t$ |
| $S_p$ | set of possible states of ribosome molecules |
| $N_{ribosome}$ | max. no. of ribosome molecules attached |
| $L_{ribosome}$ | length of protein (residue) |
| $\lambda_i^p$ | translation birth rate |
| $\mu_i^p$ | translation death rate |
| $k_{elong}^p$ | translation elongation rate constant |
| $k_{on}^{mR}$ | ribosome activation rate constant |
| $P_n^p$ | probability of $n$ ribosomes molecules attached to a transcript |
| $\bar{R}_p$ | Avg. rate of protein synthesis |
| $\underset{p}{\sigma}$ | Variance of protein synthesis rate |
| $F_{inter}(\tau_p)$ | Inter-arrival time between protein molecules |
| $< P(t) >$ | Ensemble average of number of protein molecules produced at time $t$ |

### 5.3.1 Modeling transcriptional dynamics

In modeling the dynamics of prokaryotic transcription, we consider the key processes involved: gene activation and deactivation, transcription initiation and RNAP recruitment, mRNA chain elongation, and finally transcription termination. We observe the 'concurrency' involved in the transcriptional process [20], i.e. each gene is simultaneously transcribed by several RNAP molecules, which can be visualized as a combed structure, with the gene-encoding region of the DNA forming the backbone, and the transcripts of increasing length (left-to-right) forming the teeth [20]. Thus, in order to calculate the average number of mRNA transcripts, i.e. the burst frequency distribution, we model the system in terms of the number of RNAP molecules attached to the gene at any instant of time.

Let $X_{mR}(t)$denote the number of RNAP molecules attached to the gene-coding region of the DNA for a particular gene at time $t$. $X_{mR}(t)$ can take values in the discrete state space

$S_{mR} = \{0, 1, ..... \ N^{\max}_{RNAP}\}$ where,

$$N^{\max}_{RNAP} = \left\lfloor \frac{L_{gene}}{RNAP_{footprint}} \right\rfloor \tag{5.1}$$

and, $L_{gene}$ = length of the gene ( in base-pair length)

$RNAP_{footprint}$ = promoter clearance, or distance between two RNAP molecules.

At any time $t$, the system starts at state 0, i.e., no RNAP is attached to the gene. Once the promoter is activated and an RNAP molecule is recruited, the system moves to state 1. Now, assuming at some time $t^{'}$, the system is in state $k$, i.e. $k$ RNAP molecules are attached to the gene, the following events can occur, ( transition steps are captured in Fig. 5.3:

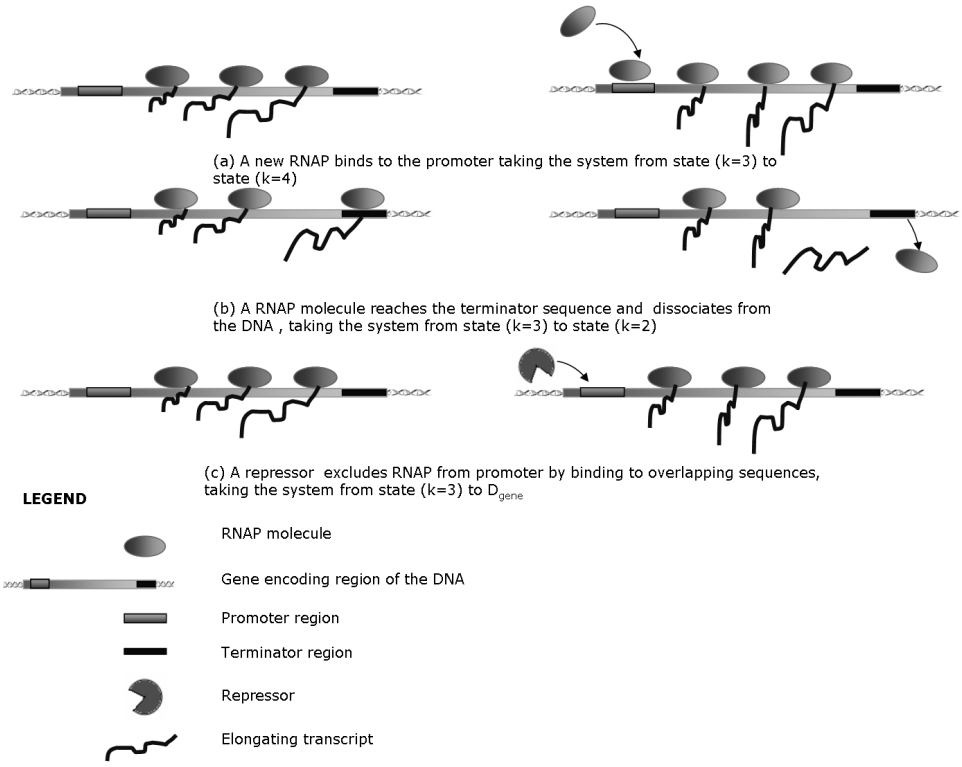



Figure 5.3. Different events in the transcription process.

(i) The gene may be activated again and another RNAP molecule is recruited, taking the system to state $k+1$.

(ii) An RNAP molecule reaches the terminator sequence releasing an mRNA transcript for translation, taking the system to state $k-1$.

(iii) A repressor may exclude further RNAP molecules from binding to the promoter region, deactivating the gene and taking the system to an inactive state ($D_{gene}$).

We now consider a continuous parameter, discrete space Markov chain [63, 123], $\chi_{mR} = \{X_{mR}(t), \quad t \in T\} \qquad and \quad T = \{t : 0 \leq t < \infty\}$ where $X_{mR}(t)$is the state of the system at time $t$, with a finite state space $S_{mR}$.

For initial solution, let us neglect the dead state$D_{gene}$, and use standard birth and death modeling technique to derive the probability of states,

$$\chi_{mR} = \{X_{mR}(t)\}, t \in T = \{t : 0 \leq t < \infty\} \tag{5.2}$$

where $X_{mR}(t)$ is a birth-death Markov chain where the intensity matrix [123] $\boldsymbol{Q}$ is defined as,

$$Q = \begin{Bmatrix} -\lambda_0^{mR} & \lambda_0^{mR} & 0 & 0 & ... \\ \mu_1^{mR} & -(\lambda_1^{mR} + \mu_1^{mR}) & \lambda_1^{mR} & 0 & ... \\ 0 & \mu_2^{mR} & -(\lambda_2^{mR} + \mu_2^{mR}) & \lambda_2^{mR} & ... \\ . & . & . & . & . \\ . & . & . & . & . \end{Bmatrix} \tag{5.3}$$

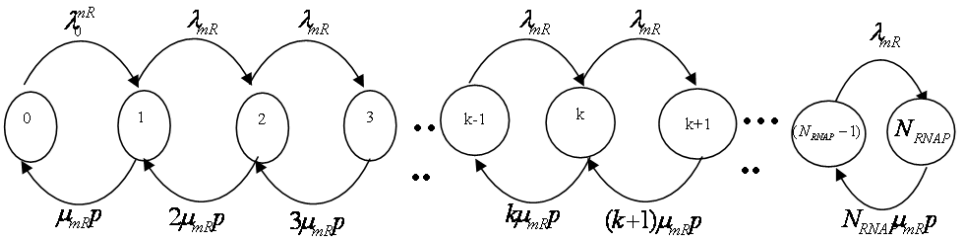



Figure 5.4. Birth-death Markov chain for transcription.

$\lambda_i^{mR}$ is defined as the birth-rate while $\mu_i^{mR}$ is defined as the death rate of the process, which is depicted in Fig. 5.4 along with the possible states. The birth and death rates encompass the dynamics of the different biological actors affecting transcription and are defined as follows:

(i) *Birth rate* ($\lambda_i^{mR}$): defines the rate at which new RNAP molecules arrive in the system, which is given by,$\lambda_i^{mR} = \frac{1}{T_{mR}}$ where, $T_{mR}$ is the sum of the average time taken for

gene activation, RNAP-promoter binding, open-complex formation and the time taken for the previous RNAP molecule to clear the promoter for the next RNAP molecule (promoter clearance time), i.e.,

$$T_{mR} = T_{activation} + T_{binding} + T_{init} + T_{clearance} \tag{5.4}$$

$$T_{mR} = \frac{1}{\lambda^{+}} + \ \frac{1}{k_{on}^{mR}} + \frac{1}{k_{init}^{mR}} + \frac{RNAP_{footprint}}{k_{elong}^{mR}} \tag{5.5}$$

It may be noted here that the arrival rate of the first RNAP molecule does not include the promoter clearance time and is given by,

$$\lambda_0^{mR} = \frac{1}{T_{mR}} \quad where \quad T_{mR} = \ T_{activation} + T_{binding} + T_{init} \tag{5.6}$$

(ii) *Death Rate* ($\mu_i^{mR}$) : defines the rate at which RNAP molecules leave the system, i.e. the rate at which mRNA transcripts are released and is governed by the time taken by the RNAP molecule to elongate the chain and reach the terminator sequence. Now, at the $i$th. state, any one of the $i$ RNAP molecules can release a transcript with probability $p$ and $i$ RNAP molecules are working in parallel to produce mRNA molecules. Thus,

$$\mu_i^{mR} = i \times \frac{1}{T_{elong}^{mR}} \times p \quad where \quad T_{elong}^{mR} = \ \frac{L_{gene}}{k_{elong}^{mR}} \tag{5.7}$$

With the birth and death rates computed as above, the birth-death process can be characterized by the set of differential-difference equations for the state probabilities (obtained from the Chapman-Kolmogorov forward equations [51]),

$$\frac{dp_j^{mR}(t)}{dt} = -(\lambda_j^{mR} + \mu_j^{mR})p_j^{mR}(t) + \lambda_{j-1}^{mR}p_{j-1}^{mR}(t) + \mu_{j+1}^{mR}p_{j+1}^{mR}(t), (j \geq 1) \tag{5.8}$$

$$\frac{dp_0^{mR}(t)}{dt} = -\lambda_0^{mR} p_0^{mR}(t) + \mu_1^{mR} p_1^{mR}(t) \tag{5.9}$$

where $p_j^{mR}(t)$ = probability of being in state $j$ at time $t$. Applying the stochastic balance procedure [123], the stationary state probabilities can be obtained from Eqn. 5.8 and 5.9 for the chain in Fig. 5.4 as,

$$\begin{aligned} P_n^{mR} &= \left(\frac{\lambda_0^{mR}(\lambda_{mR})^{n-1}}{n!(\mu_{mR})^n}\right) P_0^{mR} \quad where \\ \lambda_n^{mR} &= \lambda_{mR} \ , \mu_n^{mR} = \mu_{mR}, n = 1, 2..N_{RNAP} \end{aligned} \tag{5.10}$$

and

$$P_0^{mR} = \frac{1}{\sum_{n=1}^{N_{RNAP}} \frac{\lambda_0^{mR}(\lambda_{mR})^{n-1}}{n!(\mu_{mR})^n}} \tag{5.11}$$

In the above formulation, the cessation of the transcription process by a repressor molecule has not been considered. As mentioned earlier, the repressor binding can prevent access of the RNAP molecule to the promoter region, thereby taking the system to an inactive state, $D_{gene}$. As this can happen from any state in the system, the modified Markov chain is represented by Fig. 5.5.

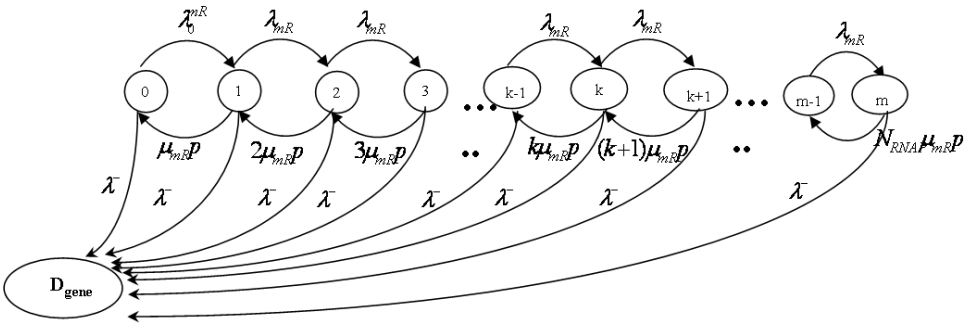



Figure 5.5. Birth-death Markov chain with killing state for transcription.

The Markov chain in Fig. 5.5 represents a special birth-death process with killing [50, 156]. Based on Karlin and McGregor's integral representation for the transition proba-

bilities ($Q$ matrix), it has been shown in [156] that the representation can be extended in case where transition to the death state occurs from any other state, as in Fig. 5.4. Also, [51] shows that at as long as killing is possible from a finite number of states and the criteria of certain absorption and positive decay parameter are maintained, existence of quasi-stationary distributions for a birth-death process with killing are maintained. This particular chain can now be analyzed in a quasi-stationary manner as shown in [51] and reduced to a simple birth-death process with the following mapping for the birth-death process representing the chain,

$$\widetilde{\mu_0^{mR}} = 0, \quad \widetilde{\lambda_0^{mR}} = \lambda_0^{mR} + \lambda^- \widetilde{\mu_i^{mR}} = (\lambda_{i-1}^{mR} / \widetilde{\lambda_{i-1}^{mR}})\mu_i^{mR}, \quad \widetilde{\lambda_i^{mR}} = \lambda_i^{mR} + \lambda^- + \mu_i^{mR} - \widetilde{\mu_i^{mR}}$$

Using the above transformations for the modified birth and death rates, the individual state probabilities,$P_n^{mR}$ can be computed. Now, $P_n^{mR}$ gives the probability of $n$ RNAP molecules in the system producing transcripts at the rate $\mu_n^{mR} = n \times \frac{1}{T_{elong}^{mR}} \times p$.

Thus, the average rate of transcript synthesis, $\bar{R}_{mR}$ is given by,

$$\bar{R}_{mR} = \sum_{n=0}^{N_{RNAP}} P_n^{mR} \mu_n^{mR} \tag{5.12}$$

and the variance is given by,

$$\underset{mR}{\sigma} = \sum_{n=0}^{N_{RNAP}} (\mu_n^{mR} - \bar{R}_{mR})^2 \; P_n^{mR} \tag{5.13}$$

The first and second moments [50] i.e. mean and variance respectively, specified by Eqn. 5.12 and 5.13 characterize the probability distribution of mRNA synthesis rate or frequency of transcript generation. In the next sub-section, we develop the model for the distribution of the number of proteins synthesized from an mRNA molecule.

### 5.3.2 Modeling translation dynamics

Once an mRNA molecule is transcribed, the translation machinery is recruited to initiate the synthesis of proteins. Although translation can be initiated even before the transcription machinery has released the mRNA in prokaryotic cell, we consider the recruitment of the translation machinery after the complete transcription of an mRNA molecule in this model.

We observe a commonality in the role of the ribosomal unit in translation as played by the RNAP molecule in transcription. Translation is initiated by a ribosome macromolecule binding to the ribosome binding site (RBS) on the mRNA. The ribosome then reads out the genetic code from mRNA in three-letter codons corresponding to the amino acids, assembling them into a growing chain of amino acids which subsequently fold into a functional protein.

Concurrency is also observed [14] in the translation process, with multiple ribosomes (forming polysomes) reading out simultaneously from a transcript. Another characteristic in the translation machinery is the competition between a ribosome and a degradosome (RNaseE) molecule to bind to the RBS and initiate translation or decay of mRNA respectively, as shown in Fig. 5.6.

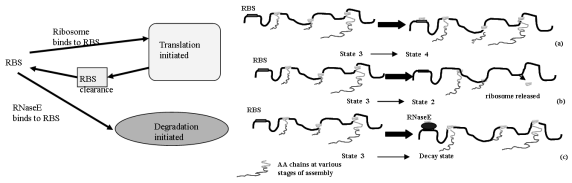



Figure 5.6. Competition and state-space of translation.

Based on the above observations and the fact that the number of ribosomes attached to an mRNA governs the count of protein molecules synthesized before decay of the

mRNA, we consider$X_p(t)$, the number of ribosomes attached to the mRNA at time $t$. The maximum number of ribosomes (ribosome load) simultaneously processing an mRNA depends on the length of the gene and the inter-ribosome spacing ($L_{ribosome}$) and is given by, $N_{ribosome} = \left\lfloor \frac{L_{gene}}{L_{ribosome}} \right\rfloor$ i.e. $X_p(t)$ takes values in the discrete state space, $S_{protein} = \{0, 1, ..... N_{ribosome}\}$. Analyzing the events which can occur at a state $X_p(t) = k$ as in the case of transcription (as shown in Fig. 5.6), we have,

(i) After the time taken for the ribosome to move $L_{ribosome}$ base pairs, another ribosome can get attached to the mRNA depending on the association rate of the ribosome and takes the system to state $k + 1$.

(ii) A previously attached ribosome can complete the process of protein generation and gets released from the system bringing it to state $k - 1$.

(iii) An RNaseE molecule can get attached to the RBS and start the decay process taking the system to a 'death' or 'killing' state.

We now define a Markov chain, $\chi_P = \{X_P(t)\}$ and $\{t \in T\}$ where $X_p(t)$ is the state of the system at time $t$. As in the case of transcription, the Markov chain can be analyzed as a birth-death process with a killing or decay state, for translation. The main differences are the computation of birth and death rate of the translation process and the computation of the maximum possible states of the system. *(i) Computation of birth rate* $\lambda_i^p$: The birth-rate captures the rate at which the system moves forward in state, i.e. another ribosome unit binds to the RBS successfully. This event occurs after the time taken for the ribosome unit bound to the RBS in the previous (*k-1*) th. state to move $L_{ribosome}$ units along the mRNA chain to clear the next ribosome for association at the RBS, together with the time taken for the ribosome to bind to the RBS and initiate translation. Thus,

$$\lambda_i^p = \frac{1}{T_{binding} + T_{clearance}} = \frac{1}{\frac{1}{k_{on}^p} + \frac{L_{ribosome}}{k_{elong}^p}} \tag{5.14}$$

It may be noted here that,$\lambda_0^p = \frac{1}{T_{binding}} = k_{on}^p$, since ribosome clearance time is not required for the first ribosome binding to the RBS.

*(ii) Computation of death rate* $\mu_i^p$: The death event signifies a ribosomal unit successfully completing the synthesis of a protein molecule and getting released from the system. This can be computed on the same lines at the birth-rate with the observation that the ribosome here has to traverse 3 more codons to generate the last amino acid in the chain and stop the synthesis of the protein. At the $i$th. state, any one of the $i$ proteins can successfully complete protein chain synthesis with probability $p_{ribosome}$ . The assembly of the protein in the ribosomal complex is progressing in parallel, so the rate of protein production will be multiplication of the number of parallel production stations (ribosome units) to the rate of an individual unit to complete production of a protein. Thus, death rate is given by,

$$\mu_i^p = i \times \frac{1}{T_{elong}^p} \times p_{ribosome} \quad where \quad T_{elong}^p = \quad \frac{L_{gene}+3}{k_{elong}^p} \tag{5.15}$$

(ii) Decay Rate ($k_{decay}^{mRNA}$) : The decay rate is defined by the binding rate of the RNase E molecule to the RBS which triggers the mRNA decay process.

Based on the above rates, the stationary state probability of the states, $P_n^p$ as,

$$P_n^p = \left( \frac{\lambda_0^p (\lambda_p)^{n-1}}{n!(\mu_p)^n} \right) P_0^p \tag{5.16}$$

$$where \quad \lambda_n^p = \lambda_p \ , \mu_n^p = \mu_p, n = 1, 2..N_{ribosome} \tag{5.17}$$

and

$$P_0^p = \frac{1}{\sum_{n=1}^{N_{ribosome}} \frac{\lambda_0^p (\lambda_p)^{n-1}}{n!(\mu_p)^n}} \tag{5.18}$$

Once the state probabilities are computed, we observe that at the $n$th. state, the rate of protein synthesis is given by $\mu_n^p$ and the probability of this synthesis rate at that state is$P_n^p$. The average rate of protein synthesis, $\bar{R}_p$, can thus be computed as

$$\bar{R}_p = \sum_{n=0}^{N_{ribosome}} P_n^p \mu_n^p \tag{5.19}$$

and the variance is given by,

$$\underset{p}{\sigma} = \sum_{n=0}^{N_{ribosome}} (\mu_n^p - \bar{R}_p)^2 \; P_n^p \tag{5.20}$$

Based on the above equations, we characterize the probability distribution of the number of proteins available from a mRNA or the burst size distribution for protein synthesis.

### 5.3.3 Combined model of protein synthesis

As mentioned earlier, the process of gene expression is marked by *coupling* between the processes of transcription and translation. Thus, the probability distribution for protein synthesis is a combined process arising out of the two stochastic processes of transcription and translation elucidated in the previous sub-sections. Now, the probability distribution of protein arrival, specifically the time between two protein molecules being synthesized will depend on the number of mRNA molecules present in the system. In order to compute the stationary probability of $k$ mRNA molecules being present in the system, we consider the transcript arrival and decay process as a birth-death Markov chain with infinite state space, where the arrival rate is given by $\bar{R}_{mR}$ computed earlier, and death rate is the mRNA decay rate or $k_{decay}^{mRNA}$.

Solving the Markov chain, the probability of k mRNA/transcript in the system is given by

$$P_k^{transcript} = \left( \frac{(\bar{R}_{mR})^n}{n!(\mu_{decay}^{mRNA})^n} \right) P_0^{transcript} \tag{5.21}$$

$$P_0^{trancript} = \frac{1}{\sum_{n=1}^{\infty} \frac{(\bar{R}_{mR})^{n-1}}{n!(\mu_{decay}^{mRNA})^n}} \tag{5.22}$$

Now, each of these $k$ transcripts has a corresponding protein synthesis distribution with average rate $\bar{R}_p$. Thus, we can define the distribution for the time-interval between two protein molecules (inter-arrival time) released in the combined system by an Erlang distribution [51] with shape parameter $k$ and rate parameter$\bar{R}_p$, i.e. $F_{erlang}(x, k, \bar{R}_p)$. Specifically, the cumulative density function (CDF) of protein inter-arrival time is given by,

$$F_{inter}(\tau_p) = \sum_{k=1}^{\infty} P_k^{transcript}.F_{erlang}(\tau_p, k, \bar{R}_p) \tag{5.23}$$

### 5.3.4 Modeling noise dynamics

The quantification of the fluctuation in the number of proteins produced is a key component in understanding the stochasticity of gene expression. Generically, this fluctuation can be defined as the ratio of the variance over mean squared which allows separation of noise sources [13]. If P(t) is the protein concentration at time t, then the protein noise $\eta(t)$is given as,

$$\eta(t) = \frac{\langle P(t)^2 \rangle - \langle P(t) \rangle^2}{\langle P(t) \rangle^2} \tag{5.24}$$

where $\langle P(t) \rangle$ is the ensemble average.

In [92], the author has elucidated that the classification of noise components in terms of 'extrinsic' or 'intrinsic' depends on the definition of the system versus the environment. Thus, the noise contribution of the transcription machinery is intrinsic to transcriptional noise while extrinsic to translation noise. Based on our probability distribution for burst frequency and size, we can characterize the total protein noise ($\eta_{protein}$) as sum of transcriptional noise ($\eta_{transcription}$) and translational noise ($\eta_{translation}$), given as,

$$\eta_{protein} = \eta_{transcription} + \eta_{translation} \tag{5.25}$$

where,

$$\eta_{transcription} = \frac{\sigma_{mR}}{(\bar{R}_{mR})^2} \ , \ \eta_{translation} = \frac{\sigma_p}{(\bar{R}_p)^2} \tag{5.26}$$

In the next section, we estimate the burst frequency and size distributions based on our model and validate it with experimental data obtained from single cell experiments in *E.Coli*.

## 5.4 Model validation

In this section, we validate our stochastic models for gene expression with experimental data obtained from recent single cell experiments. Specifically, we estimate the average rate of mRNA synthesis and protein generation from our model and validate the distributions with actual experiments data. In [98, 93] measurements at the single cell level were reported on the average transcript synthesis rate and protein number distributions for the *lacZ* gene expression in *E.Coli* cells. We briefly overview the *lac* operon system in *E.Coli*, before presenting the model parameters and validation results.

### 5.4.1 The *lac* operon experimental system

The *lac* operon, which encodes a set of genes for the lactose permease, has been extensively used in experimental and analytical systems. The native *lacZ* gene is the first in the operon and is translated into monomers composing of the catalytic protein, $\beta$- galactosidase and galactose. The native operon is positively regulated in the presence of lactose and negatively regulated by glucose. The *lac* operon also encodes LacY (lactose permease) and LacA (galactosidase acetyltranderfase) apart from LacI which encodes a regulatory protein. The *lac* operon, together with the gene products is depicted in Fig. 5.7. We focus on the first gene of the operon, *lacZ*, to validate our model against experimental data obtained on single cell experiments for protein synthesis for the *lacZ* system [98, 93].

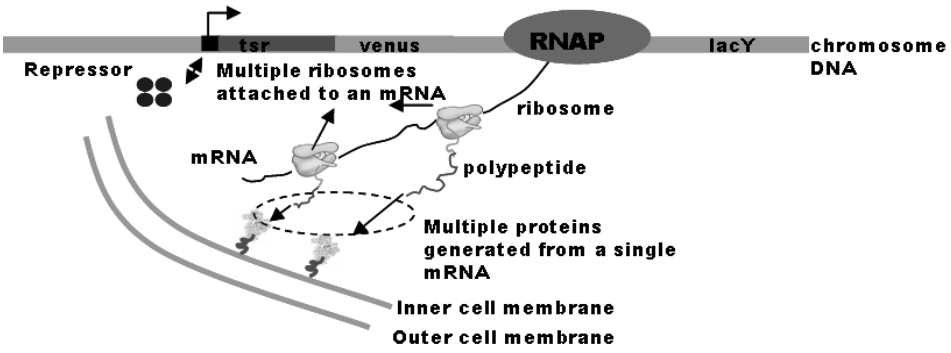



Figure 5.7. The *lac* operon system.

### 5.4.2 Model parameter estimation and validation

The stochastic model for gene expression elucidated in the previous section computes the distributions for mRNA synthesis (burst frequency) and protein synthesis (burst size) characterized in terms of their mean and variance. The mathematical expression

for the average rates involve different biological parameters which are obtained from available literature and databases to estimate the model parameters. Based on single cell experiments of gene expression in *E.Coli*, the average number of mRNA bursts was estimated as 1.2 bursts/cell cycle in [98, 93]. The model estimate for $\bar{R}_{mR}$, using the computed parameter values was calculated as 1.03 $\pm$0.85 burst/cell cycle time.

Based on the fluorescent $\beta$-galactosidase reporter molecules, details of which are provide in [93], the authors obtain an average burst size of 4.2 $\pm$ 1.8 protein molecules. Now, the authors also showed that a burst of proteins occurred from a single mRNA (0.037 $\pm$ 0.013) where the average life-time of the mRNA is 1.5 $\pm$ 0.2mins. Thus, assuming that 4.2 protein molecules are produced from a single mRNA during its life time, we obtain an average protein generation experimental rate of 0.046/s for $]barR_p$. Based on $\bar{R}_p$ and the associated probability, $P_n^{mR}$, Fig. 5.8 shows the probability distribution for burst size of proteins computed from the model (Fig. 5.8(b)) and compared with experimental data (Fig. 5.8(a)) [93]. Thus, the validated protein burst size distribution and mean rate estimated from the model provides a mathematical tool for systematically studying the effect of different parameters, which we elucidate next.

## 5.5 Sensitivity analysis of model parameters

The parameterized models of gene transcription and translation provide a mathematical foundation for systematically studying the sensitivity of the average rate of these events on the different biological parameters. Based on the biological process underlying gene expression encompassed in the scope of our model, we analyze the parameters potentially controlling the burst frequency and size distributions for prokaryotic cells. For completeness, we elucidate sensitivity analysis of all parameters in the following sub-sections.

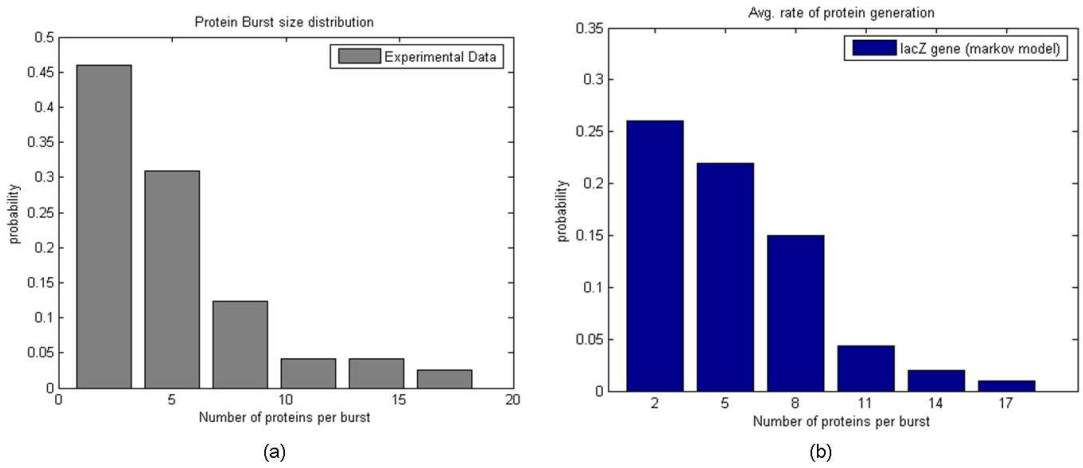



Figure 5.8. Burst size distribution (a) experimental system, (b) Markov model.

### 5.5.1 Effect of activation ratio on transcription rate

The activation ratio, defined as the ratio of the gene activation rate to the deactivation rate, quantifies the strength of the transcription factor. Fig. 5.9 shows the effect of increasing activation ratio on the transcript production rate and the corresponding transcriptional noise. As seen from the plot, increasing activation ratio, increases the efficiency of the transcriptional machinery thereby increasing the average rate of transcript production and decreasing the transcriptional noise.

### 5.5.2 Effect of transcription initiation ratio on transcription rate

The rate of binding of the RNAP molecule to the promoter region exerts control on the rate of transcription by increasing the efficiency of the transcription process. Thus, as seen from Fig. 5.10, increase in the rate of transcription initiation increases the average rate of transcript synthesis and decreases . It may be noted here that the activation ratio has a greater effect on the efficiency of transcription as seen from the slope of the curves in Fig. 5.9 and Fig. 5.10 and noted in earlier work [7].

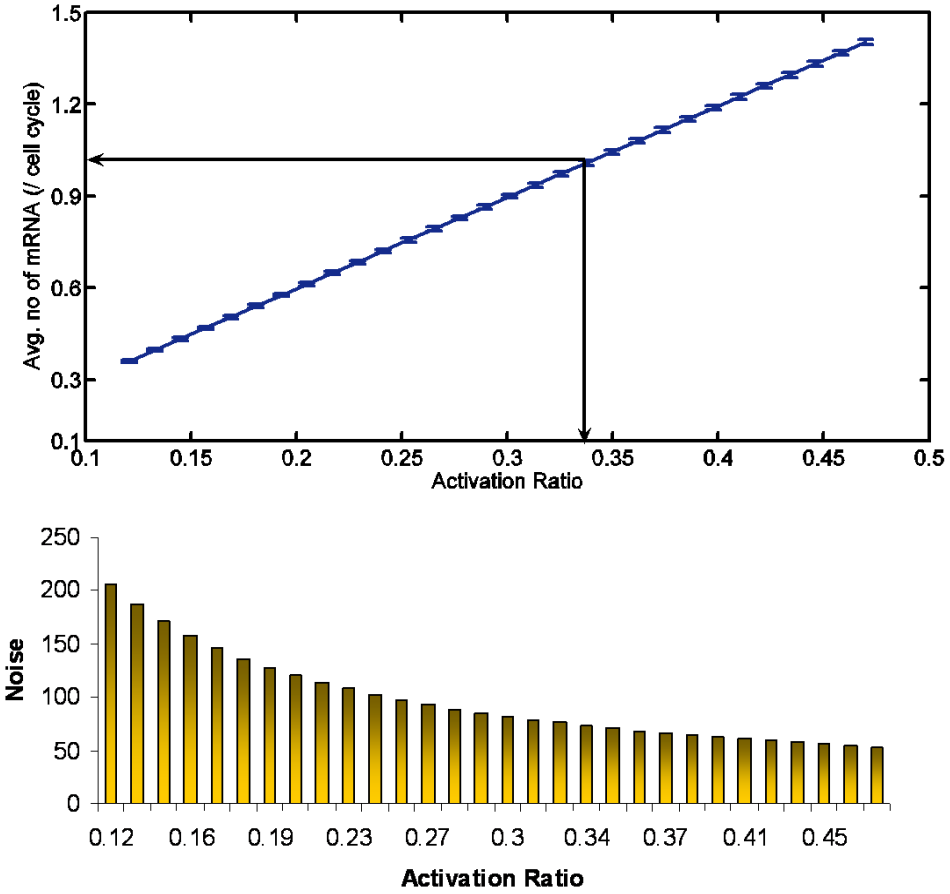



Figure 5.9. Sensitivity of gene transcription to activation ratio.

### 5.5.3 Effect of promoter on transcription rate

The effect of the promoter clearance region, on the transcriptional machinery efficiency is closely coupled with the activity of the promoter for a gene. A strong promoter will activate gene at a faster rate, recruiting RNAP molecules which can get 'queued' up due to a large clearance footprint of RNAP molecules already bound to the DNA, Fig. 5.11. The effect would not be significant for weak promoters which recruit RNAP molecules at a much slower rate providing sufficient time gap for clearing the promoter region, as illustrated in Fig. 5.12.

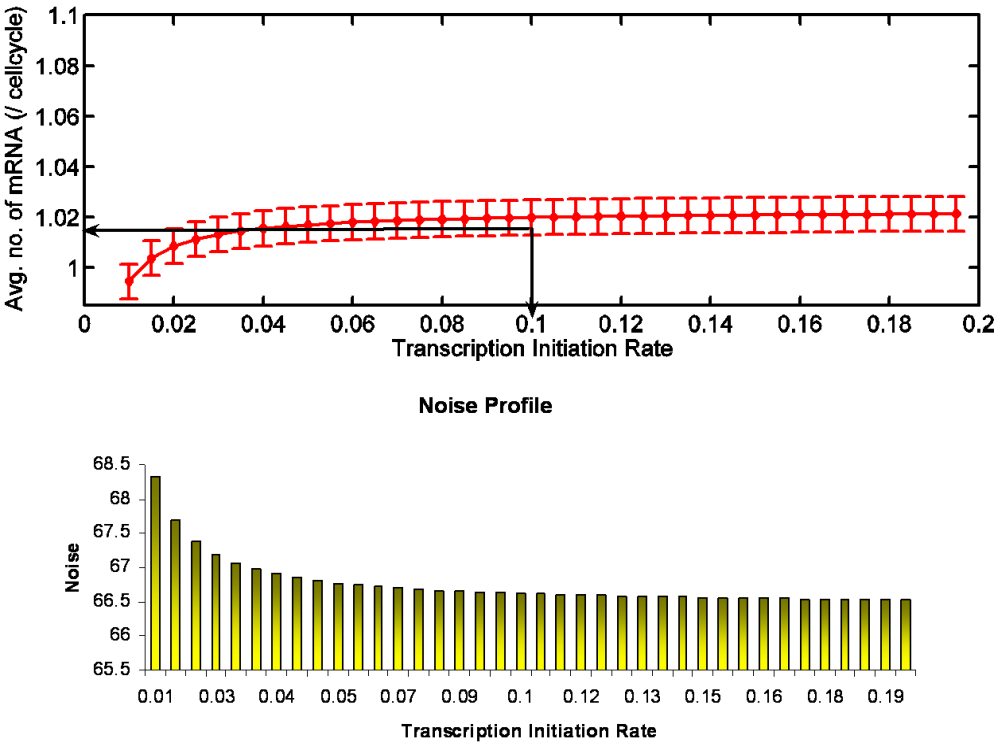



Figure 5.10. Sensitivity of gene transcription to transcription initiation efficiency.

### 5.5.4 Effect of ribosome binding on translation rate

The rate of ribosome binding to the RBS region of the mRNA molecule controls the efficiency of translation. With an increase in the value of ribosome binding rate, the average rate of protein synthesis (burst size) increases and the noise in translation decreases, as shown in Fig. 5.13.

### 5.5.5 Effect of ribosome spacing on translation rate

The clearance region of the ribosome on the mRNA controls the number of ribosomal units (ribosome load) that can concurrently translate an active mRNA molecule. By increasing the spacing between two ribosome molecules, the load decreases, thereby

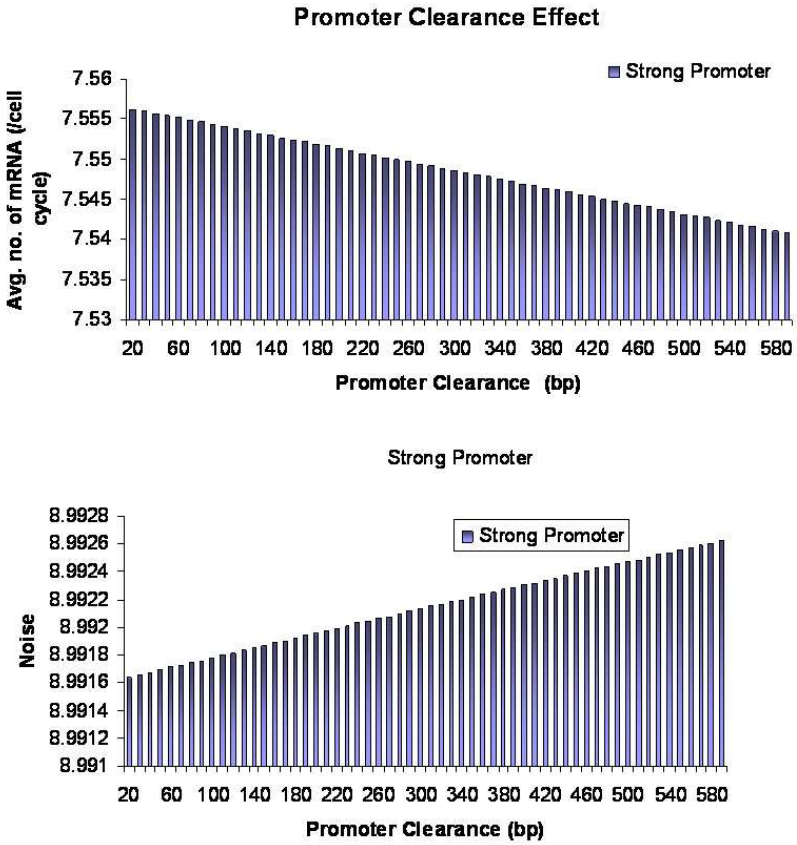



Figure 5.11. Promoter clearance effect (strong promoter).

decreasing the protein synthesis rate and increasing translational noise, depicted in Fig. 5.14.

### 5.5.6 Effect of competition on translation rate

As mentioned earlier, the degradosome competes with the ribosome for binding to the RBS on the mRNA molecule. Thus, increasing the degradosome binding rate will increase the mRNA degradation event rate thereby decreasing the protein synthesis rate, as shown in Fig. 5.15. However, the dynamics of the overall competition are controlled by the ribosome binding rate and the degradosome binding rate. In Fig. 5.16, we show the

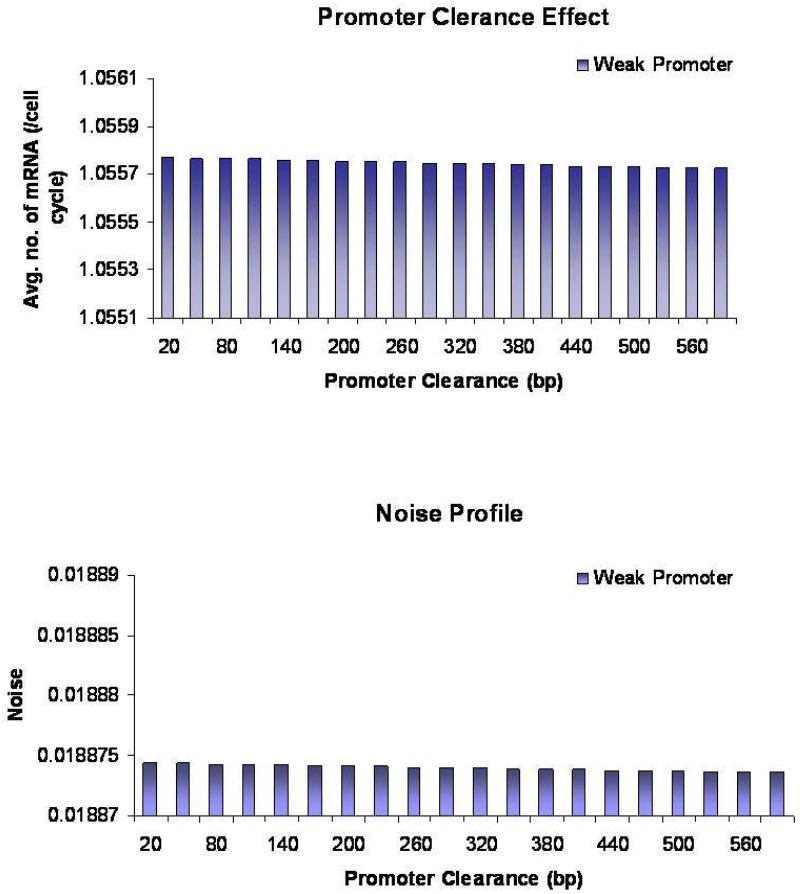



Figure 5.12. Promoter clearance effect (weak promoter).

interplay of these two effects as a surface plot. As seen from the graph, for the *lacZ* gene parameters, the ribosome exerts greater control on the protein synthesis rate compared to the degradosome.

In this section, we have systematically studied the effect of the various biological parameters on the gene transcription and translation processes. One observation of particular interest in this study is the emergence of "queueing effect" of RNA polymerase molecules recruited by strong promoters due to high ribosome clearance. Also, the inter-

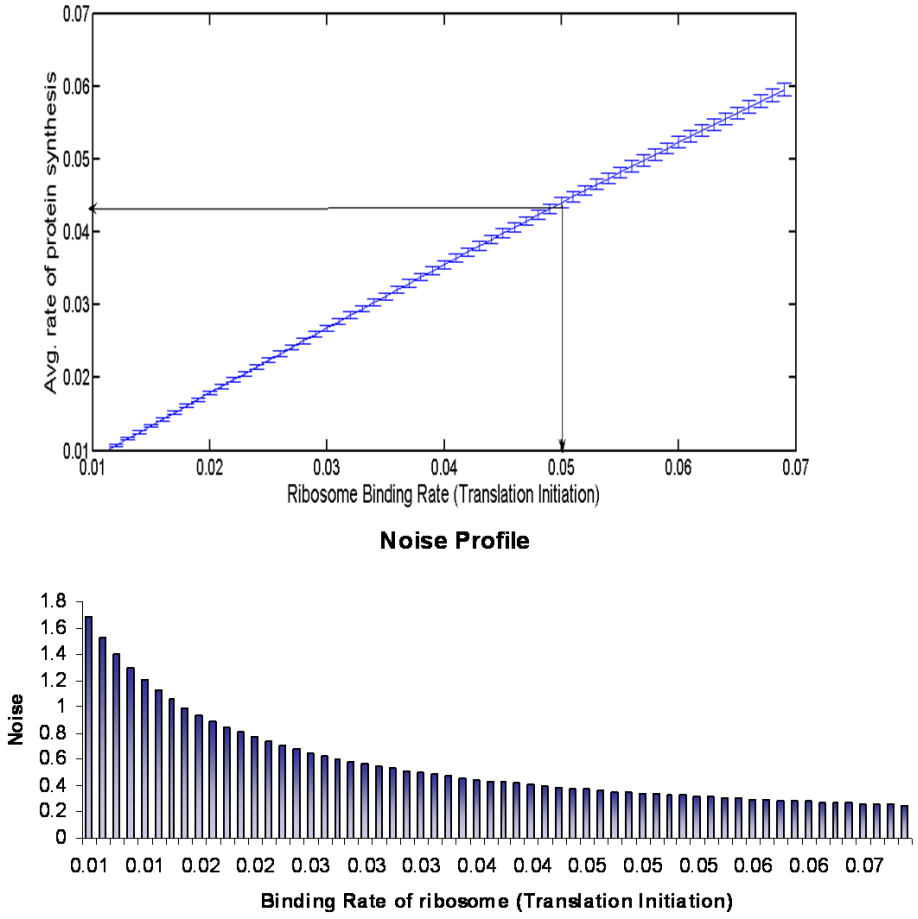



Figure 5.13. Sensitivity of translational machinery to ribosome binding rate.

play of the competition between RNA polymerase and degradosome shows the greater control of the ribosome over the degradosome for binding to the RBS of a transcript.

## 5.6 Simulation framework

In this section, we outline the framework of a *in silico* discrete event based computational framework to study the dynamic interactions of the events involved in prokaryotic gene expression. The stochastic models of gene transcription and translation provide parameterized mathematical expressions for the probability distribution of the number

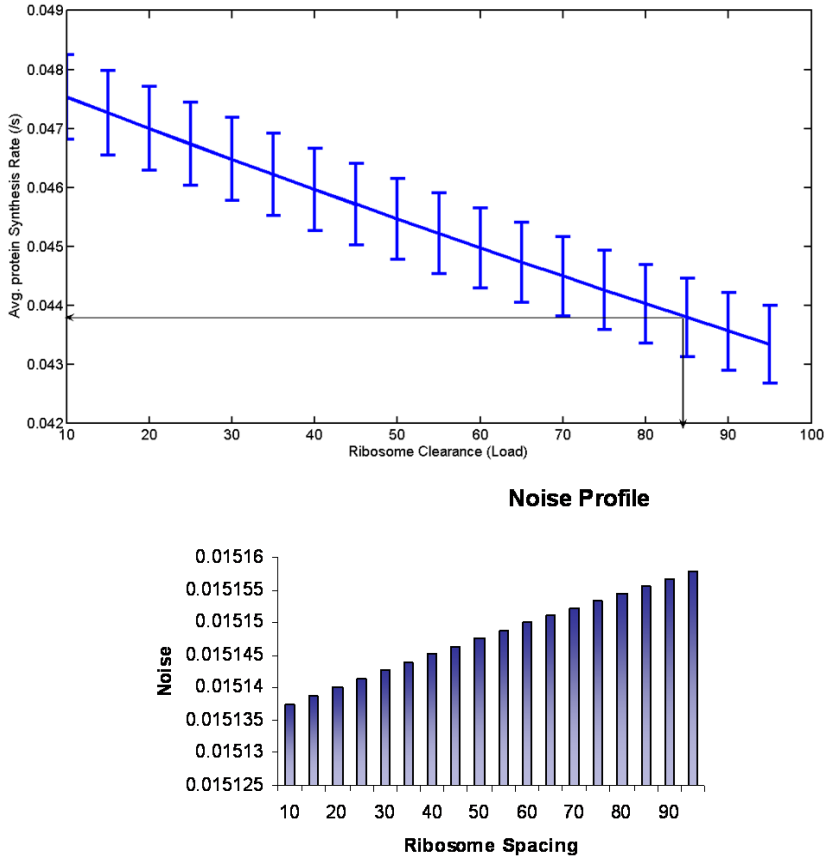



Figure 5.14. Sensitivity of translational machinery to ribosome load.

of mRNA transcripts obtained from a single gene activation event (burst frequency distribution) and the number of proteins obtained from a single transcript (burst size distribution). The distributions are characterized by their mean and variance measures as outlined in the first part of this chapter. In a bacterial cell, the process of gene expression involves a dynamic interaction of the transcription and translation events and their complex "coupling" [20] governs the cellular behavior. Moreover, other cellular events, specifically signalling events and decay events (transcript and protein decay) exert further control on the over-all system dynamics. In order to quantitatively study the fine granular interactions of these events, it is pertinent to build a computational simulation

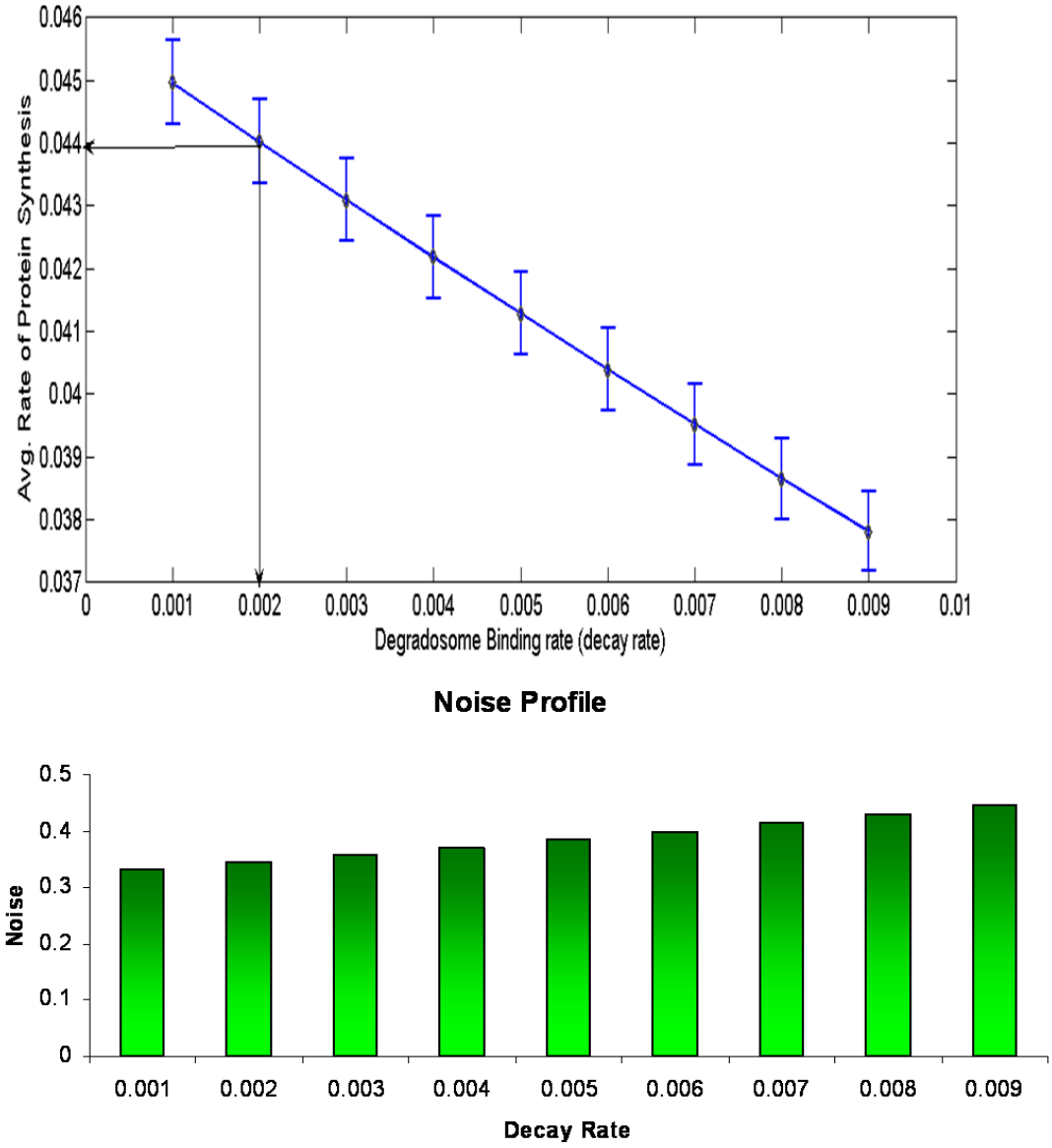



Figure 5.15. Sensitivity of translational machinery to degradosome binding rate.

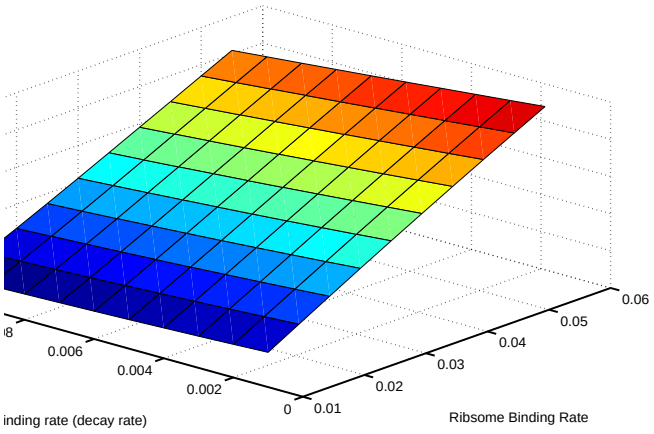



Figure 5.16. Dynamics of competition on translation machinery.

platform which is amenable to perturbations of the different system parameters while capturing the stochastic behavior of the biological process.

The stochastic modeling of cellular events, each characterized by the mean and variance of their probability distributions, lends itself to a computational treatment at a system level. We employ the discrete event based simulation framework elucidated in Chapter 3 to build an *in silico* environment for studying the gene expression process of the *lacZ* system in *E.Coli*.

### 5.6.1 Event implementation

A discrete event simulation provides flexibility in modeling biological processes at different granularities depending on available knowledge. In Chapter 4, we developed a simulation model for studying the dynamics of signal transduction in bacterial cells. In this section, we focus on the key events associated with the process of bacterial gene expression (*lacZ* system in *E.Coli*) as part of the simulation study. The key events associated with the gene expression process, as depicted in Fig. 5.17, are:

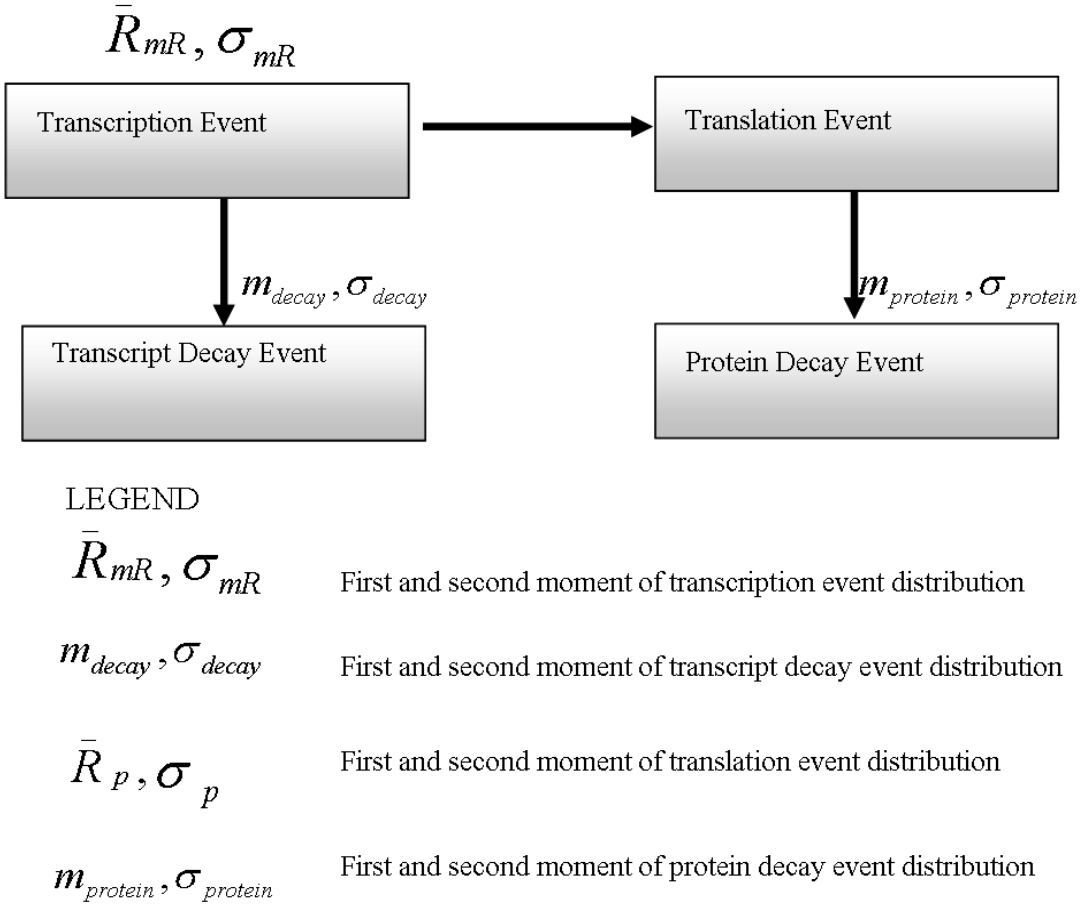



Figure 5.17. Event interaction graph for gene expression.

- *Transcription event*: This event represents the triggering of transcription by the activation of a gene and the eventual release of a mRNA molecule in the system. The probability distribution characterizing the time taken for the event (holding time) is defined by the first and second moments, $\bar{R}_{mR}$ and $\bar{\sigma}_{mR}$ respectively with time between two transcription events is represented by the random variable $\tau_{mR}$. This can be obtained from existing reaction models [106] or from rate constants for reaction models.

- *Transcript decay event*: This event represents the decay of a transcript and is characterized by an exponential distribution with half-life $m_{decay}$ obtained from experimental data. The random variable $\tau_{mR}^{D}$ represents the time between two decay events.
- *Translation event*: This event captures the process of protein synthesis from a single mRNA molecule characterized by the probability distribution of its time, with mean $\bar{R}_P$ and variance $\bar{\sigma}_P$. $\tau_p$ represents the random variable for time between two translation events.
- *Protein decay event*: This event represents the decay of a protein characterized by an exponential distribution with half life of $m_{protein}$.

### 5.6.2 Simulation process implementation

Once the events involved in gene expression have been characterized, the main simulation engine is implemented to capture the temporal interaction of these events *in silico*. As outlined in Chapter 3, the simulation engine consists of the event model library storing the holding time distribution of the events, the molecular resources database which captures the change in molecular count of the different biological entities in time through the simulation run and the event scheduler which controls the engine.

### 5.6.3 Simulation runs

As the discrete event simulator captures the behavior of the system (in this case, prokaryotic protein synthesis) in the probabilistic domain, the simulation results are reported over multiple runs. Each simulation run characterizes the system (captured in the change in molecular resource counts with time) for a specified simulation run-time. The simulation results, presented in the next section are the average values over 50-100 runs for the *lacZ* system.

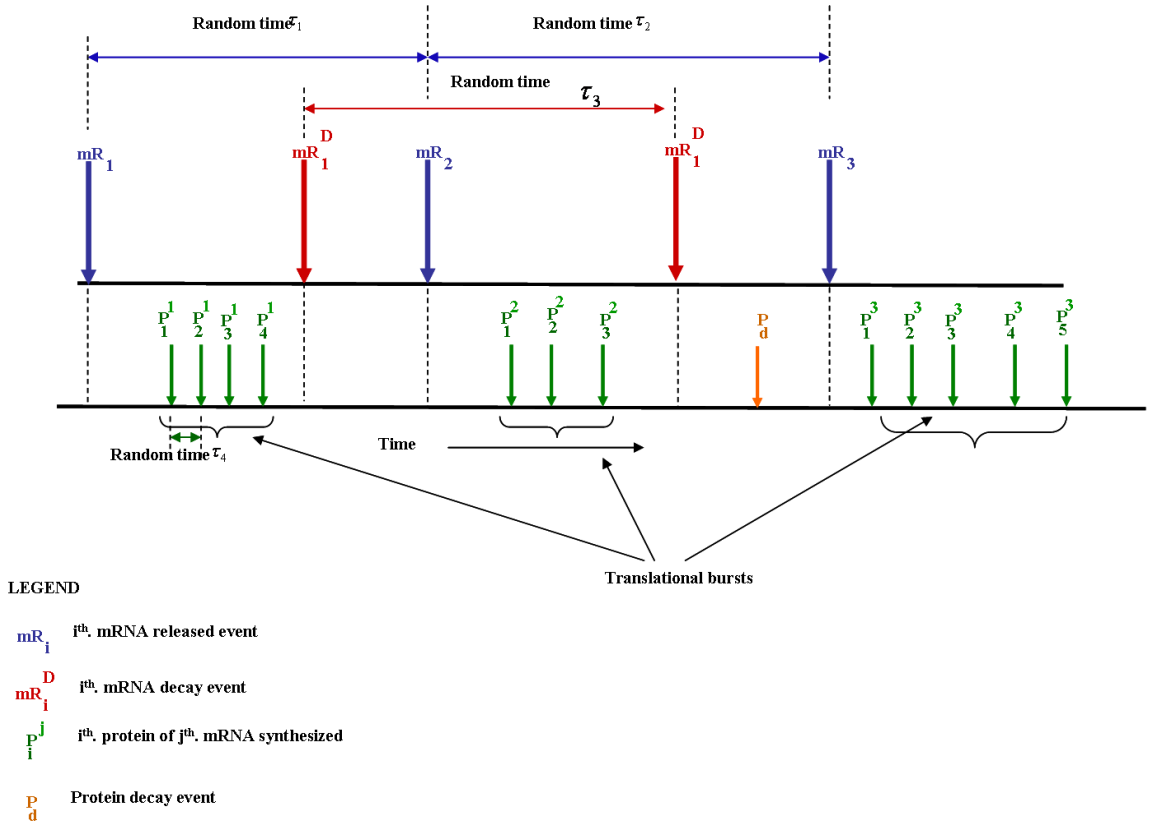



Figure 5.18. Event dynamics (simulation with experimental parameters).

## 5.7 Simulation study of gene expression dynamics

Once the simulation platform is built, we conducted several *in silico* studies to quantify the temporal interaction of the different events outlined in Fig. 5.17, specifically focussing on the contribution of the sensitive biological parameters identified earlier on the overall dynamics of protein generation.

### 5.7.1 Burstiness of protein generation

In the first simulation case study, we observe the system dynamics in time for experimentally validated transcriptional burst frequency and protein burst size distributions. In particular, we show the burstiness in protein production reported in experimental

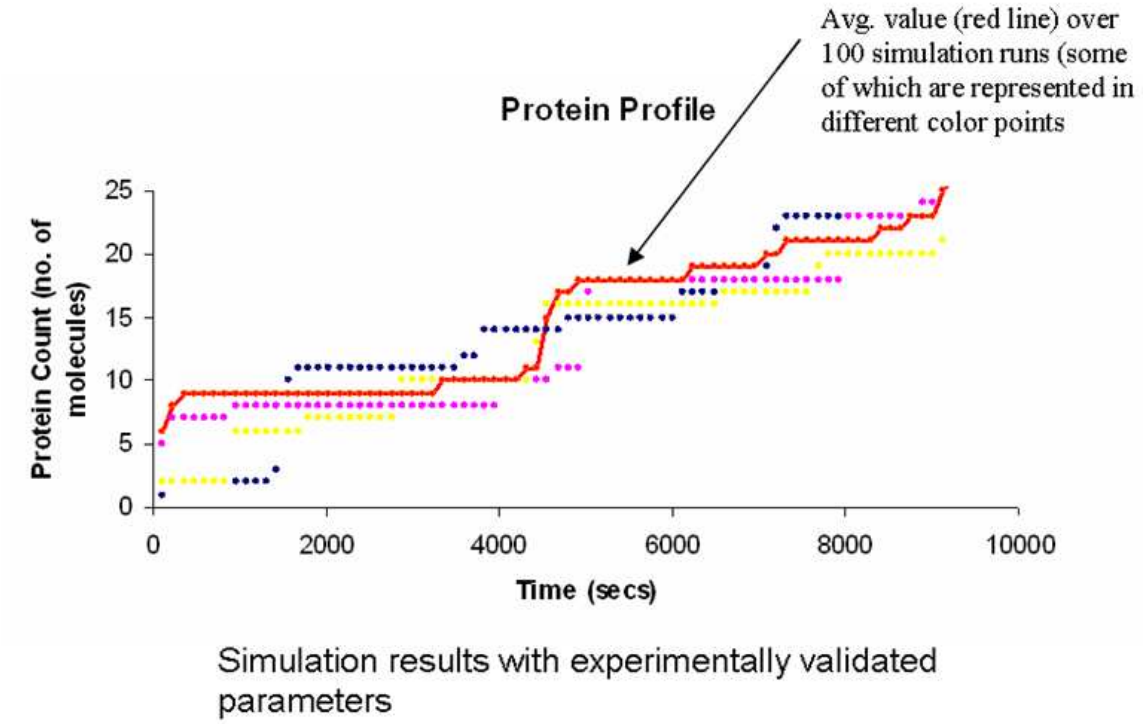



Figure 5.19. Protein profile (simulation with experimental parameters).

work conducted on the lacZ gene in *E.Coli*. Fig. 5.18 shows the event dynamics for the gene expression process, focusing on the rare transcription events which drive the burstiness in protein generation. In Fig. 5.19, we show the time-course of protein generation from the lacZ gene for the simulation results (averaged over 100 simulation runs) which validate the wet-lab data obtained from [51, 123] over a 3 cell cycle period. The results indicate that the *in silico* simulations validate the experimental observations of protein burstiness. Fig. 5.20 shows the low number of mRNA molecules produced during the simulated time, indicating the rarity of transcription events while the noise profile reflects the fluctuations in protein count with the bursts. The protein and noise profiles together characterize the dynamics of gene expression for the *lacZ* system. Next, we

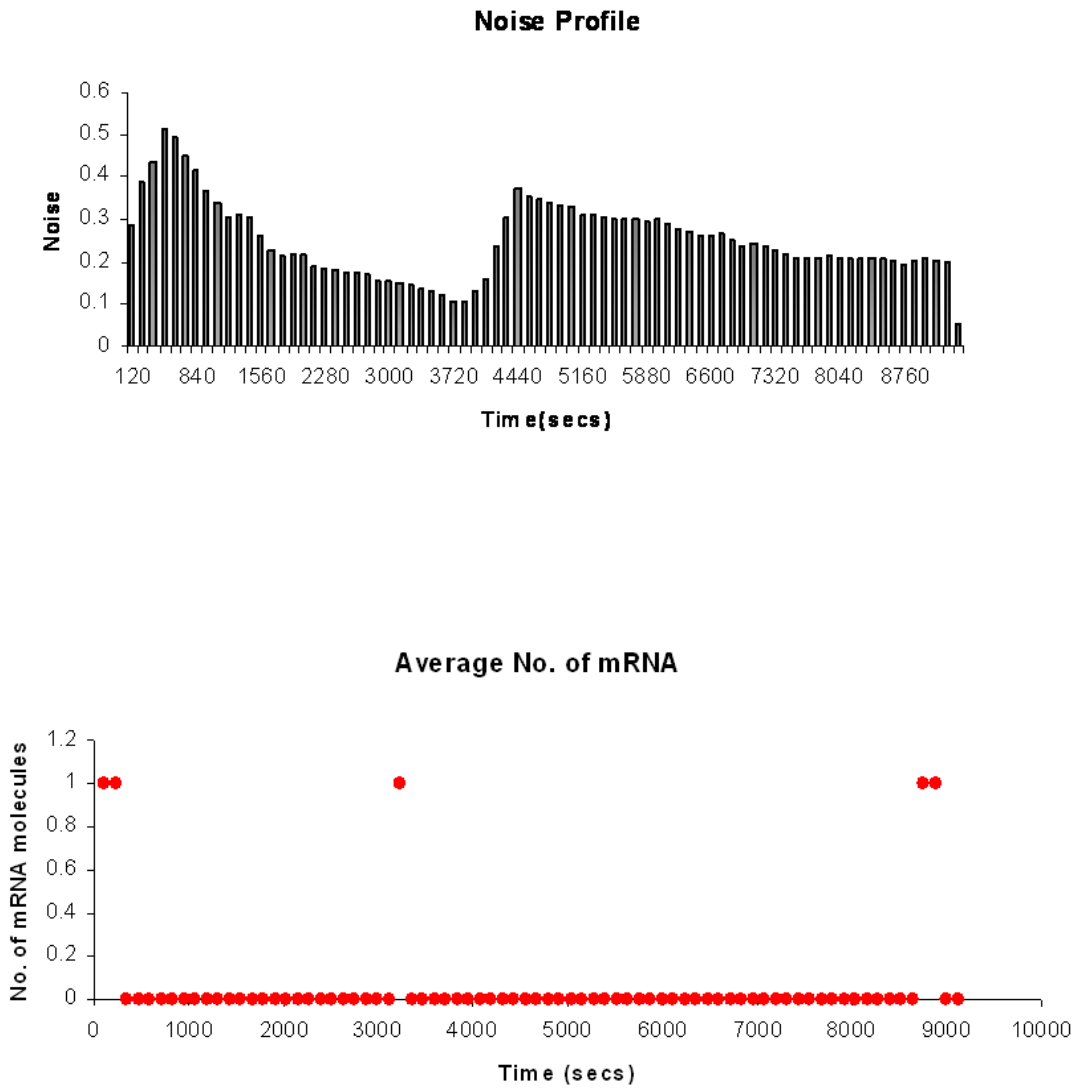



Figure 5.20. Noise and transcript profile.

conduct a suite of simulation studies to analyze the nature of "burstiness" and identify the contribution of the different molecular factors in controlling it.

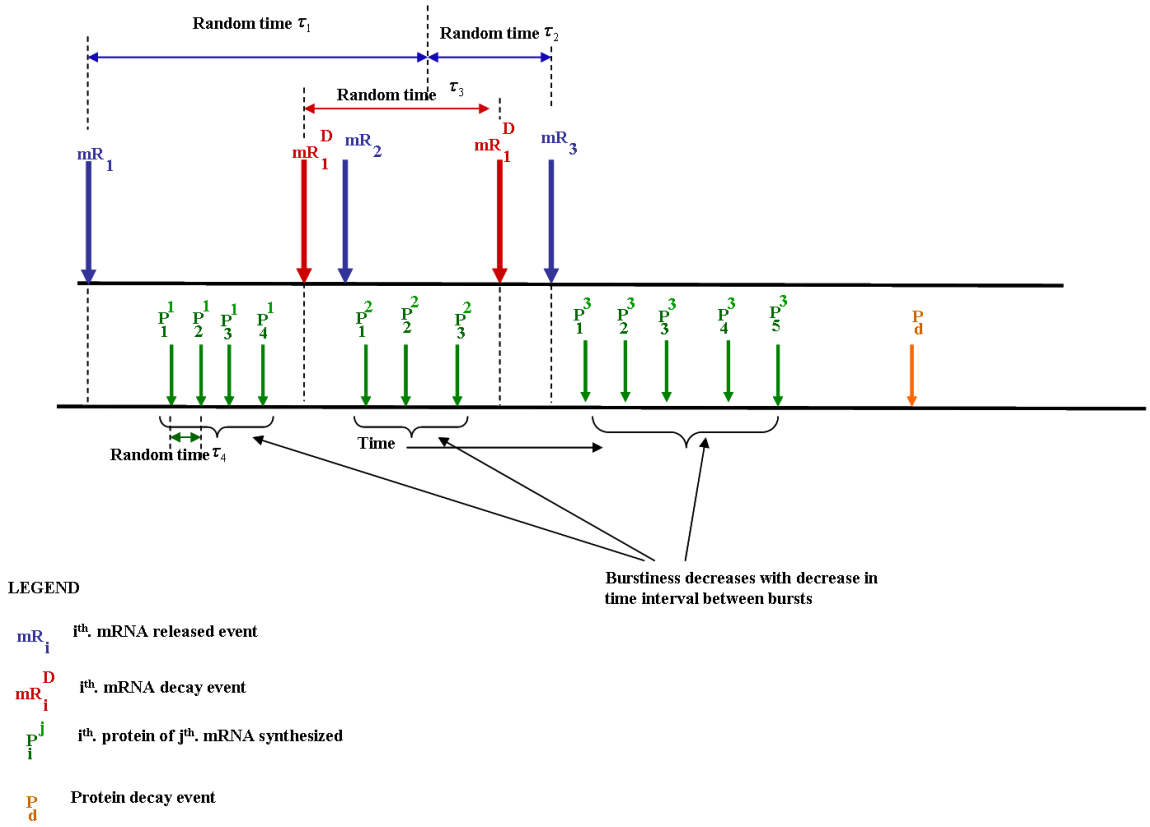



Figure 5.21. Event dynamics (increased transcription rate).

### 5.7.2 Effect of promoter strength on protein burstiness

Previously, we showed the strength of the promoter exerts control on the rate of transcript synthesis. In order to study its effect on the system level generation of proteins, we conducted simulation study with increased rate of mRNA synthesis, i.e. the arrival rate of transcription events is increased. Fig. 5.21 shows the changing event dynamics for this case. Due to the decreased inter-arrival times between transcripts, the protein generation machinery becomes more efficient, thereby decreasing the bursty nature of protein production (shown in Fig. 5.22 with corresponding mRNA and noise profiles in Fig. 5.23) while increasing the number of mRNA molecules produced for the same simulated time of 3 cell cycles.

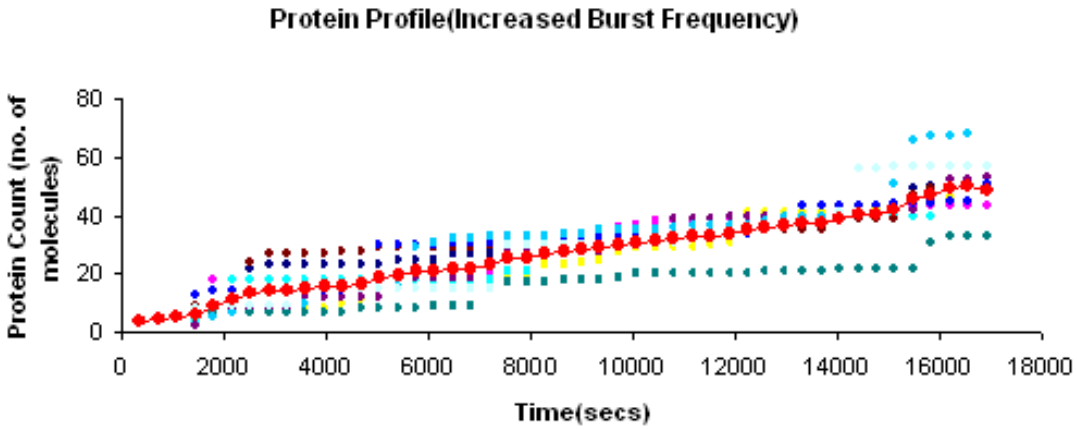



Figure 5.22. Protein profile (increased transcription rate).

### 5.7.3 Effect of mRNA decay on protein burstiness

The simulation environment provides a platform to conduct further experiments on the effect of other parameters on the number of proteins generated. Specifically, we focus on the role of the decay rates of mRNA and proteins on the gene expression phenomena. Fig. 5.24 shows the protein profile for the case study with increased mRNA life, where it can be observed that proteins are produced in a continuous manner with rare occurrences of "bursts". This observation can be explained in terms of the longer life of a single mRNA molecule. Although the transcription events occur at large intervals due to their rarity, the longer life-span of a single mRNA results in more number of proteins being synthesized from it by the translational machinery. This effective improvement of the translational machinery also results in decreased bursts (and therefore noise in protein production Fig. 5.25) although the number of mRNA molecules generated is low as shown in Fig. 5.25.

The previous two simulation studies highlight the fact that control of protein generation bursts can be executed through increased efficiency of the transcription machinery

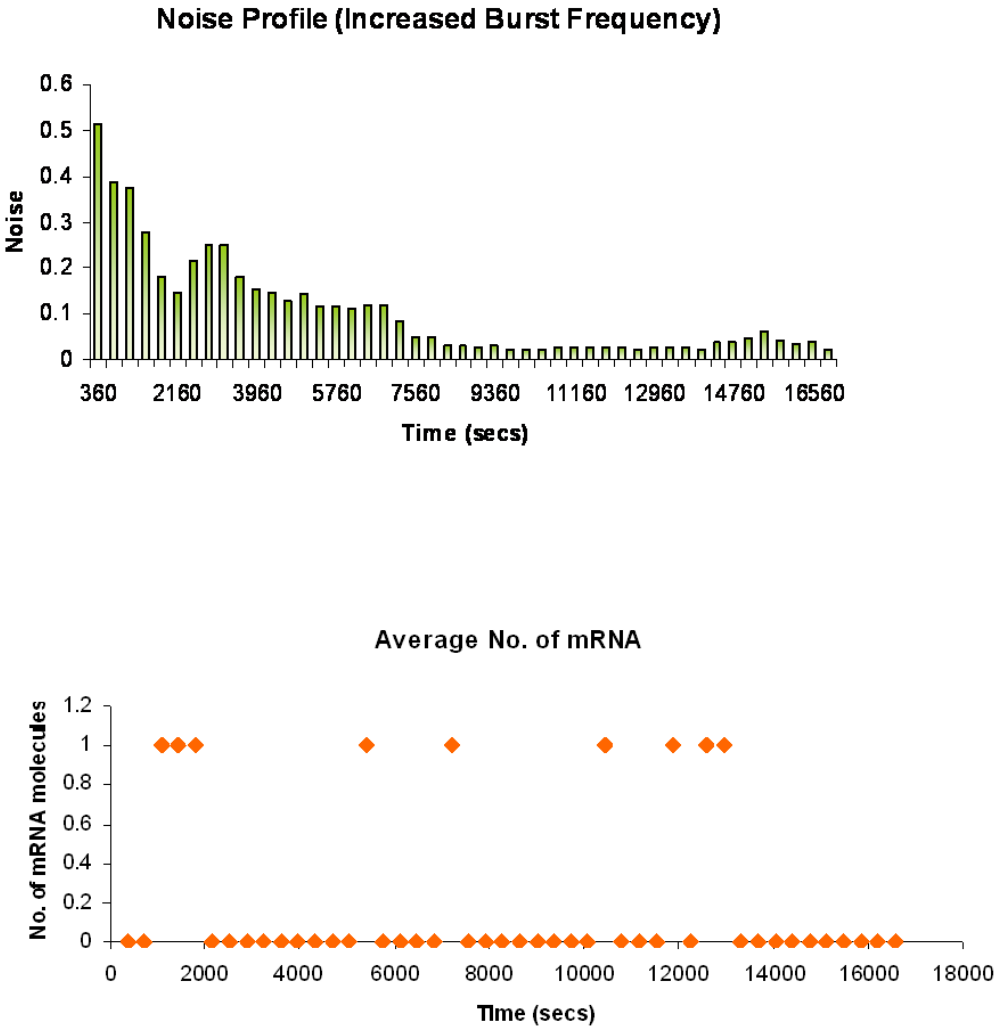



Figure 5.23. Noise and transcript profile (increased transcription rate).

(increased transcript synthesis rate) or indirect increase of translation machinery efficiency. These simulation studies provide insight into various alternate pathways for controlling protein generation rates which can serve as potential therapeutic alternatives for controlling the expression level of various disease genes.

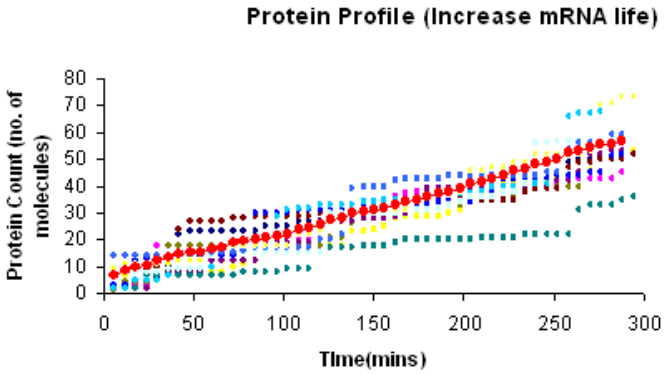



Figure 5.24. Protein profile (increased transcript lifetime).

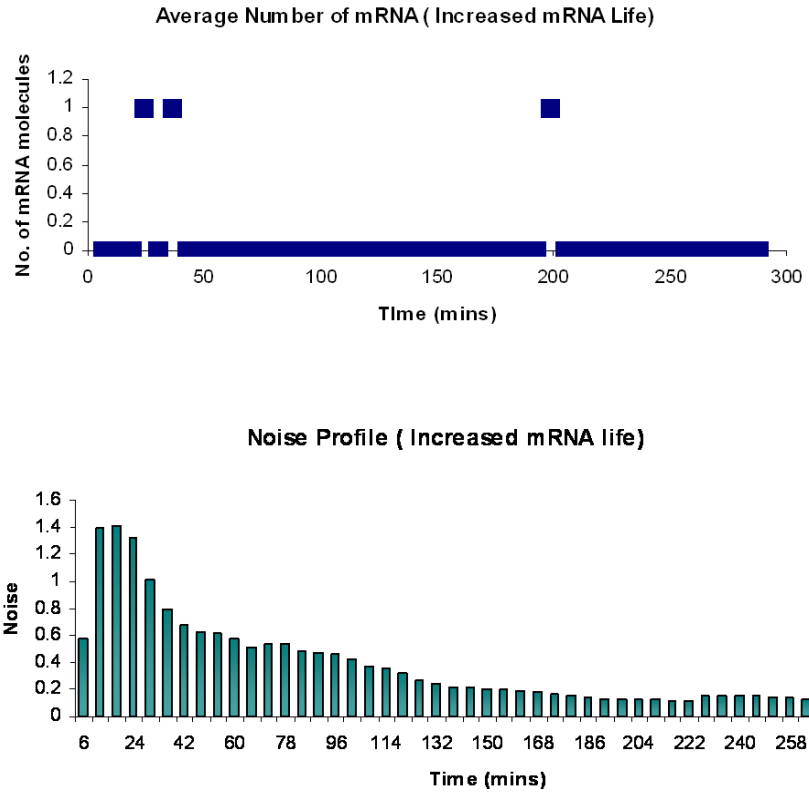



Figure 5.25. Noise and transcript profile (increased transcript lifetime).

## 5.8 Summary

Bacterial gene expression involves a complex process of interaction between multiple molecular actors acting individually or in a concurrent fashion. These actors contribute to the temporal fluctuations in the number of proteins produced in a cell. In this chapter, we have modeled the stochastic process of gene expression, incorporating the effect of various actors in parameterized probability distributions for mRNA and protein synthesis. The parameterized distributions help in systematically analyzing the sensitivity of the noise in protein production to the different molecular actors.

Till this point of time, we have focused on modeling biological events associated with signal transduction and gene regulation. However, for a genome scale study of cell behavior, it is mandatory to include the dynamics of the metabolic network into the overall picture. In the next chapter, we delve into the unique features of metabolic network modeling and outline a hybrid simulation paradigm, which extends the existing discrete event framework to provide a holistic view of cell dynamics.

# CHAPTER 6

# A HYBRID SIMULATION APPROACH

Comprehending the fine-grained signal specificity, gene regulation and feedback mechanisms which control the complex molecular choreography within the cell remain a fundamental theme in systems biology. The complexity of regulatory and metabolic networks coupled with the cross-talk, noise and spatio-temporal variations make genome-scale study of their interaction dynamics a particularly challenging computational problem.

The central issue in understanding the system dynamics of a living cell is to capture the interaction of gene regulatory, signal transduction and metabolic pathways in an integrated *in silico* platform. Such a platform requires systematic integration of different databases and the ability to capture the complex characteristics in a computational framework. Specifically, with the difference in time-scale of regulatory and metabolic events, the problem of "stiffness", i.e. inability to simulate the effects of fast time-scale reactions in conjunction with slow reaction models, affect the efficiency and performance of different *in silico* approaches.

In this chapter, we propose a novel hybrid simulation approach to tackle the interaction dynamics of biological networks. In section 6.1, we identify the major computational challenges in developing integrated network models. We delve into the root causes of the stiffness problem, outlining existing works for the study of metabolic networks and their characteristics in section 6.2. Based on an extension of the discrete event simulation approach outlined in the first part of the dissertation, we present the hybrid simulation architecture, called *HimSim* in section 6.3. The hybrid algorithm incorporates

the stochastic model based discrete event simulation with a flow-based computation of metabolic event dynamics to simulate the interplay between these networks. Section 6.4 outlines the implementation details of the hybrid extension to the existing software platform together with the database model. Section 6.5 and 6.6 show experimental validation and *in silico* results for the bacterial cell *Escherichia Coli*, particularly the interplay of signal transduction, gene regulatory and metabolic networks involved in the central metabolism components of this bacterial cell under different growth conditions. We conclude this chapter in Section 6.7.

## 6.1 Interplay of regulatory and metabolic networks

From a biological perspective, the ability to trace the behavior of cellular pathways at a molecular level opens the window towards holistic understanding of living systems. While reductionist approaches provides detailed molecular mechanisms of specific parts of a cell, e.g. signalling molecules and their interactions, or protein-protein interaction leading to gene expression dynamics, a complete picture of cellar mechanisms arise from collating these disparate components in a continuous spectrum [57], [129].

Recent experimental work on studying cellular networks at a systems level [163] have shown that phenotypic behavior in a cell emerges from complex, non-linear interactions between various molecular entities located in different parts of the cell. Microarray experiments [120] on global gene expression analysis and metabolic engineering have shown that the metabolic flux (defined as the change in the number of metabolites for a metabolic reaction) is controlled by the regulation of metabolic genes which influences the number of active enzymes available in the cell. The genes are further controlled by the complex interaction of other genes in a gene regulatory network structure. In [68], the authors reconstructed the transcriptional regulatory network for the bacterial cell *Es-*

*cherichia Coli*, identifying the key global transcription factors which employ fine-grained control of genomic and metabolic phenotypes.

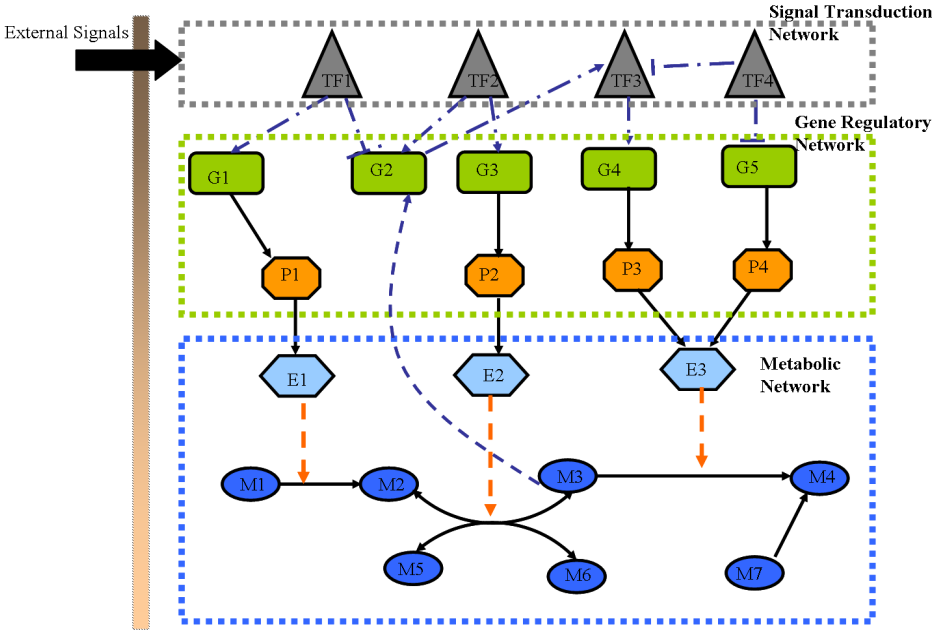



Figure 6.1. Interplay between signaling, gene regulatory and metabolic networks.

While the closely coupled dependance between gene regulatory and metabolic networks in a cell has been identified, one of the other major players in global cellular control is the signaling network. The signaling network, or signal transduction network, governs the behavior of the cell in response to various internal and external environmental conditions. The cross-talk and transduction of signals through a variety of membrane-bound and cytoplasmic protein signaling molecules add further complexity to the dynamics controlling cell behavior. In Fig. 6.1 [101], we show a schematic view of the interplay between the various networks in a cell which work in tandem for a cell's function and growth, in response to external signals (environmental change in ion or nutrient concentrations, stress etc.). In building a computational framework which allows the study of

cellular dynamics on a genomic-scale, it is pertinent to develop models and algorithms which systematically capture the interaction between the molecular entities outlined in the schematic.

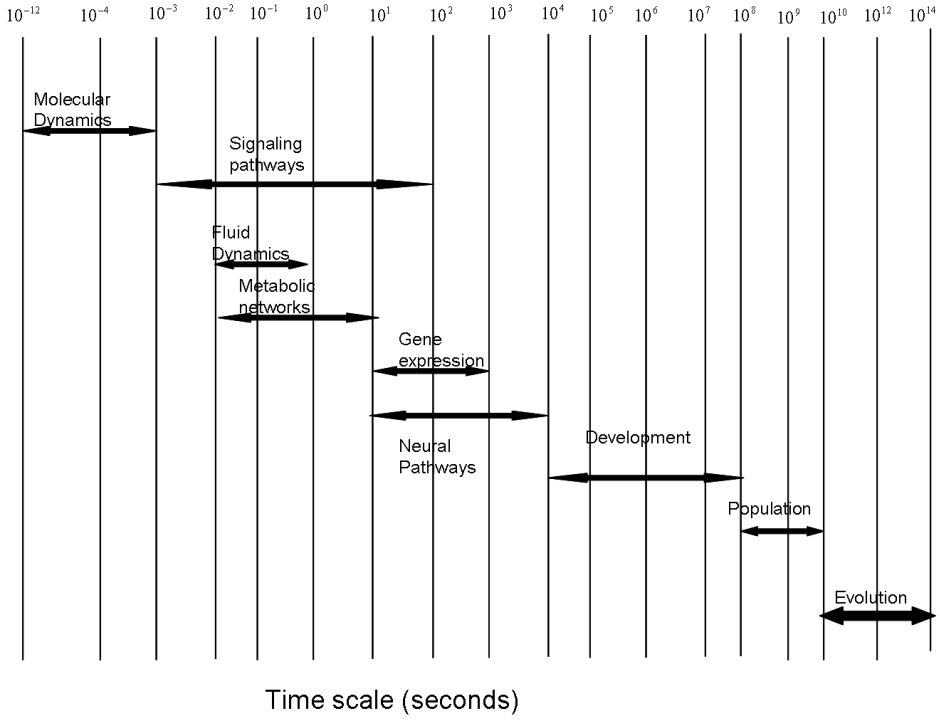



Figure 6.2. Temporal variation in biological phenomena.

## 6.2 Computational approaches to the study of cellular networks

While the need for an integrated platform has been realized, particularly in the light of ever-increasing proteomic and genomic data from high throughput experiments and pathway databases, the constructing of such a framework pose several challenges - both from a computational perspective as well from a biological viewpoint. In this section, we briefly outline the key issues before delving into the advantages and limitations of existing techniques.

### 6.2.1 Challenges in integrated modeling of cellular networks

Some of the major issues in building an integrated platform for *in silico* modeling can be classified as follows:

- *Different temporal scales of biological phenomena* : In order to study biological systems at a holistic level, it is important to realize that different biological processes operate on time-scales which are 10 or more order of magnitude different. As shown in Fig. 6.2 [127], there exists orders of magnitude difference between signaling pathways, gene regulations (which are typically slower) and metabolic reactions which operate of milliseconds and less scale. This difference in time presents a major computational challenge in simulating a system which involves thousands of reactions of these various networks.

  Specifically, as identified in [55], [113], the difference in the rate constants for signaling and metabolic reactions causes "stiffness", i.e. the simulation spends more time in the fast reaction space without simulating the dynamics of the slow time-scale reactions, in classical ordinary differential equation (ODE) based techniques. The problem compounds manifold when the simulation has to scale further orders of magnitude to capture the dynamics of inter-cellular, inter-tissue and organ level interactions.

- *Knowledge gap in the biology of different pathways*: Another problem, from a biological perspective, is the existence of knowledge gap in understanding the molecular mechanisms governing various biological processes. For example, while a particular gene regulatory mechanism may be well understood biologically, the upstream signal triggering the gene regulation may have significant gaps, rendering the development of integrated models challenging.

- *Disparate sources of pathway information*: While the recent surge of genome scale experimental techniques and bioinformatic tools have opened a huge collection of

databases storing data on different molecular entities and their interactions, lack of common interface poses severe challenges in communication between the disparate resources. Each experimental or computational tool employs its own database schema and structure which caters to the specific needs of the biological system, for example signaling network data in KEGG [96], or metabolic reaction networks in EcoCyc [52]. In the absence of common schema or interface between the different databases, integrating the information across various platforms is a major computational issue. The challenge here is to create a balance between the different database structures and the global usability of the information contained them, which were not incorporated in the initial design of the databases.

- *Lack of common computational modeling and simulation tools*: Closely linked with the disparity in data storage and retrieval technique, is the existence of a wide variety of computational modeling and simulation tools. As outlined in Chapter 2, different modeling techniques cater to specific biological systems. While classical ODE models capture chemical reactions at a molecular level, it is computationally infeasible to scale such system of equations for an integrated model. Most of the database information is structured in a network graph form for pathways, signal transaction networks and metabolic reactions. On the other hand, for differential equation based models, the system is represented by a set of linear differential equation of molecular reactions. The mapping of these two structures is a difficult task requiring human intervention. It is necessary to develop tools which allow the dynamism hidden in these networks to be captured automatically in the simulation framework.

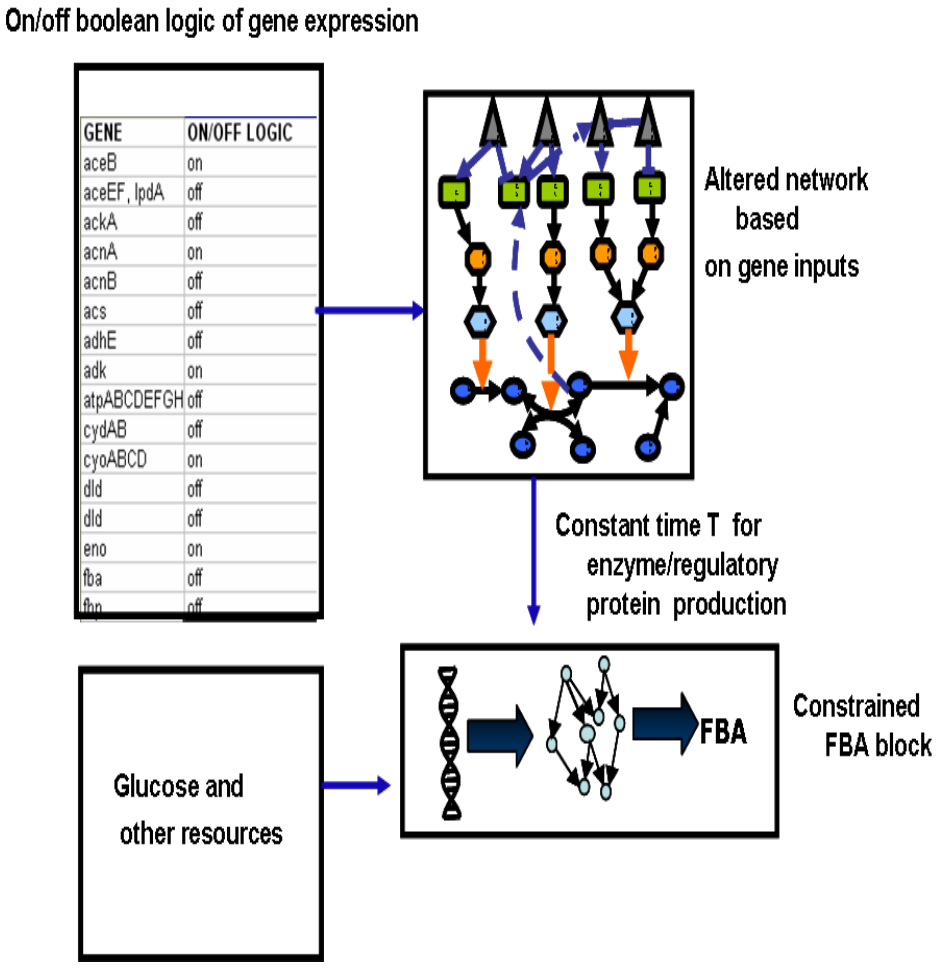



Figure 6.3. Regulatory flux balance analysis approach.

### 6.2.2 Existing computational approaches

Significant efforts have been undertaken in reconstructing genome-scale metabolic and regulatory networks together with computational approaches for the systematic study of their behavioral dynamics [57, 60, 150, 122]. Specifically, constraint-based metabolic models, which employ stoichiometry, thermodynamics and flux capacity to compute the possible distribution of flux across a given network of reactions have been successful in predicting metabolic phenotypes of model organisms like *E.Coli* and yeast [150].

Flux balance analysis (FBA) [150, 117] employs a linear optimization technique to compute the optimal flux distribution across a system of reactions. Based on the assumption that metabolic reactions occur under steady-state conditions and that the cell works towards optimization of a particular cellular entities (typically maximizing biomass yield), FBA formulates the problem of flux computation as a optimization problem where the thermodynamic and stoichiometric properties of the system serve as constraints.

While this particular technique has the advantage of being computationally fast on account of not employing dynamic simulations, it does not take into account the regulatory constraints governing the metabolic reactions. In recent years, several studies have developed integrated models incorporating regulatory constraints on the FBA models [119, 118, 181]. Two basic approaches are used for the hybrid study of such networks: (i) In a regulatory FBA or rFBA approach, outlined by Palsson et.al [120], the flux optimization problem is augmented with a dynamically changing constraint profile based on the regulation of metabolic genes. Thus, the optimization search space changes in every predefined time-step, depending on the gene expression profiles. In this approach, the gene expression dynamics are captured through a boolean matrix formulation (representing with 1 or 0 depending on whether the gene is active or inactive respectively), as outlined in Fig. 6.3. (ii) Extreme Pathway Analysis (EPA) [150] based approach for the identification of consistent, steady-state metabolic and regulatory flux for a given constant environmental signal.

While these techniques have been successful in predicting observed metabolic fluxes for specific systems, the assumption of boolean expression levels for the genes is not reflective of actual cellular conditions where gene expressions changes continuously based on upstream signals. In the FBA approach, this continuous change in gene expression levels is abstracted by a boolean variable, which is set to an arbitrary "on-off" state based on the biological knowledge to constraint the metabolic reaction fluxes. Thus,

the transient dynamics of the change in enzyme concentration are not captured in the binary representation. As elucidated in the next section, the hybrid approach allows the incorporation of gene regulatory effects on metabolic flux distributions on a continuous time-scale.

Also, many enzymes are formed from multiple protein complexes under transcriptional regulation of different genes whose relative abundance governs the number of available enzymes for a metabolic reaction. In a recent work, Shlomi et. al [90] extends the rFBA model to incorporate signaling events as upstream triggers for determining the state of a gene. However, the signaling logic is also expressed in terms of boolean expressions and does not consider their transient molecular dynamics.

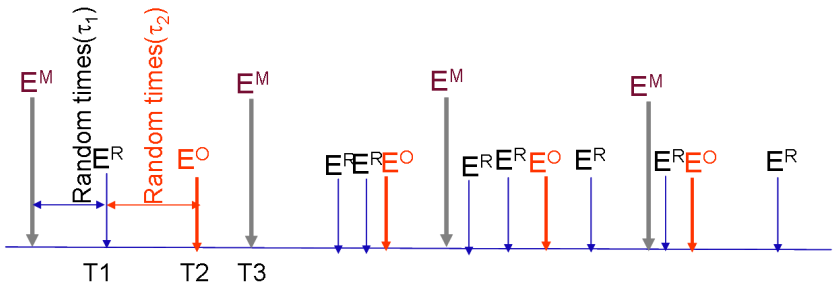
In [188], Wise et.al developed a discrete event based metabolic simulator based on an algebraic data flow model wherein a metabolic reaction is modeled as a discrete event and the flux of the metabolites emerges as the sum of molecules produced or consumed by reactions involving the metabolites. The reaction velocities and enzyme activity coefficients determine the number of reactions executed in each step. However, the data-flow model is only applicable for metabolic reactions and the does not provide an integrated computational framework.

In the next section, we outline the detailed of a hybrid simulation technique, which incorporates a discrete metabolic model based on algebraic data-flow together with the discrete event based stochastic simulation of signaling networks to provide a common platform for integrated study of cell behavior.

## 6.3 *HimSim*: A hybrid simulation approach

In this section, we elucidate the details of the hybrid simulation algorithm which allows the study of the interaction between slow time-scale regulatory reactions and fast

time-scale metabolic reactions. Before outlining the algorithm, we make some observations which motivate the hybrid approach:




T1 DES computes the metabolite change based on Flux provided by DMA due to time $\tau_1$ before processing the event $E_R$ so that the metabolite conditions are correctly reflected in the system

T2 DES computes the metabolite change based on Flux provided by FBA due to time $\tau_2$ before processing the event $E_O$ so that the metabolite conditions are correctly reflected in the system

T3 New values of Flux is given by FBA based on its calculations.

Figure 6.4. Interaction of events in an integrated model.

- In a discrete event based approach, the entire system is viewed as the interaction of events of different types, signaling, protein-protein interactions etc. Thus, the metabolic reactions can be viewed in this domain as metabolic events with the reactants and products being input and output resources respectively. This event interaction view of the system is depicted in Fig. 6.4. The key issue is to capture the time taken for a metabolic reaction which causes change in metabolite flux depending on reaction stoichiometry.
- Since the metabolic events are executed in a couple of orders of magnitude less time-scale compared to the other regulatory events, it is possible to view their behavior as being a change in the total number of molecule count of metabolites

at a given instance of time. The change, however, is governed by the dynamics of the gene expression profiles and enzyme concentrations.

- In order to capture the effects of the cross-talk between the different pathways in an integrated model, it is pertinent to develop a common database schema which stores the information of pathway interactions and provides a consistent interface to query their interaction networks.

Based on the observations outlined above, we develop a hybrid simulation approach, called *HimSim*. The driving intuition behind the approach is the fact that the interplay between signaling and metabolic networks can be captured in time by abstracting the system as the interaction of the respective events.

### 6.3.1 Stochastic simulation of signaling and regulatory events

Since the dynamics of signaling and downstream gene regulatory events evolve through stochastic interaction of the molecular entities, their behavior is captured through the discrete event based simulation approach outlined in the previous chapters. The interplay between these events show the temporal change in the system state, in terms of number of genes and gene products which are expressed or repressed as a result of external signaling events. As the molecular concentration of these gene products and enzymes change, they effect the flux of the metabolic reactions controlled by these enzymes, which need to be computed next.

### 6.3.2 Freezing the system time to capture metabolic events

As mentioned earlier, the metabolic reaction events typically occur on much faster time scales compared to the regulatory events. Thus, at a given instance in time, the metabolic reactions which are "fireable" [188], i.e. whose reaction stoichiometries and enzymatic count allow the reactions to be executed, can be assumed to occur during a

given fixed time-interval. This time interval, called the *metabolic event interval* ($\tau_{metab}$), determines the inter-arrival time between two metabolic events. It is important to note here that the number of metabolic reactions and their types, are different for every instance of a metabolic event. This is governed by the dynamics of the gene regulatory networks (which govern the enzyme count) together with the metabolite counts and enzyme activity at the particular event execution time. In other words, the simulation is "frozen" in time at a metabolic event, wherein all the fireable reactions and executed and molecular resources updated. At the end of metabolic event execution, the control is passed back to the discrete event scheduler to execute the next event (which can be other regulatory or signaling events).

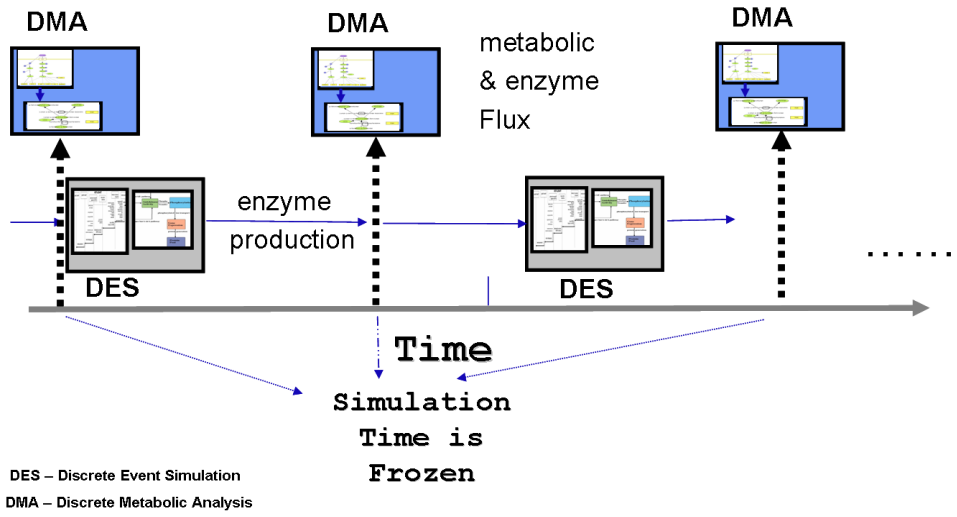



Figure 6.5. Interaction of simulation modules.

In this manner, by defining the metabolic event interval and freezing the simulation during its execution, the hybrid simulation overcomes the problem of "stiffness" associated with ODE based models of integrated reaction networks. Fig. 6.5 pictorially depicts the interaction of the DES and discrete metabolic analysis (DMA) module.

### 6.3.3 The discrete metabolic analysis (DMA) algorithm

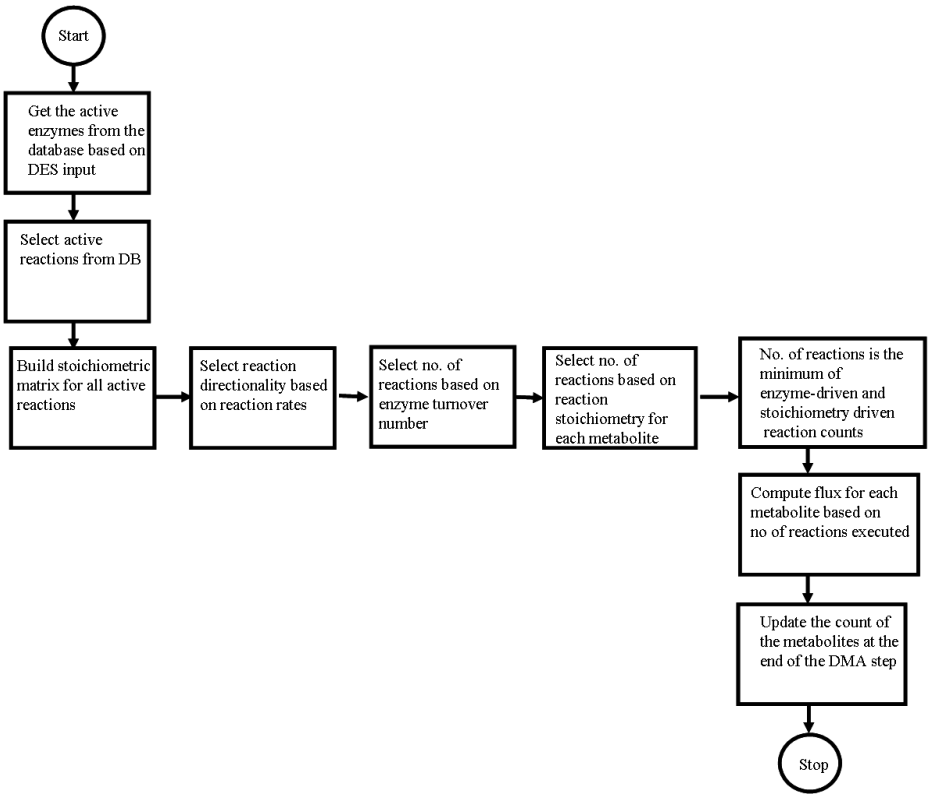



Figure 6.6. Flowchart of DMA algorithm.

The discrete metabolic analysis (DMA) module computes the change in the number of metabolites (i.e. the flux across each reaction) taking into account reaction stoichiometry, metabolite count and the number of available enzymes depending on the state of the signaling and regulatory system. Our algorithm for computation of metabolic flux is based on the data-flow model [188], wherein the flux is not a numerical value arising out of an optimization step as in FBA approach, but reflects the change in actual molecular count of metabolites depending on their role (reactants or products) in the set of

metabolic reactions being executed. The steps of the DMA algorithm, the flowchart of which is outlined in Fig. 6.6, are:

- **Active Enzyme Computation**: In the first step, the total number of active enzymes is computed depending on the state of the corresponding proteins which form the enzyme. The list of all active enzymes is computed. It may be noted here that the list of active enzymes (based on the count of the corresponding proteins) gives the state of the signaling and regulatory part of the system at the initiation of the metabolic event. Since the number of enzymes depends on the active protein component count, the transcriptional regulatory control on the metabolic flux is captured in this step. Thus, instead of using boolean variables to represent the state of the regulatory network as in the FBA approach, the number of active enzyme count provides the dynamic link between these networks.
- **Fireable reactions computation**: Once the active enzyme list is computed, the next step is to determine the metabolic reactions which are fireable, based on the active enzymes which catalyze these reactions. It may be noted here that at any metabolic event execution, the metabolic reactions which are catalyzed by the active enzymes are fireable.

  Since the number of fireable reactions depends on the previous step, the enzyme count constraints the part of the metabolic network which would be executed in a particular instance of a metabolic event. For every instance of the metabolic event, this step captures the dynamically changing nature of the metabolic network.
- **Determination of Reaction directionality**: From the list of fireable metabolic reactions, the computation of the directionality of the reaction is performed for all reversible reactions to determine whether the forward or backward reactions will be executed. This will govern the stoichiometry of the reactants and products involved in the fireable reaction list.

- **Building stoichiometric matrix**: Once the directions have been identified, the stoichiometric matrix of the active enzymes can be built based on the $K$ reactions and corresponding $M$ metabolites in the system. The stoichiometric matrix can be viewed as a bipartite graph with the reactions in one set and metabolites in another with directed edges determining the state of a metabolite as a reactant or product.
- **Enzyme driven reaction count**: The number of reactions fired depends on the catalytic activity of the enzyme together with the stoichiometry and count of the metabolites. Each enzyme is characterized by the turnover number $T_E$, which is defined as the number of substrate molecules catalyzed per second by each enzyme molecule [188].
- **Stoichiometry driven reaction count**: While the previous step gives the maximum reaction count from the catalytic activity of the enzyme, the stoichiometry of the metabolites determines the other bound on the reactions and gives the maximum number of reactions possible from the the stoichiometry point of view of the system.
- **Actual number of reactions fired**: From the above two steps, the actual number of metabolic reactions fired is determined.
- **Computation of metabolic flux**: Now, based on the stoichiometric matrix and the actual reaction count, the metabolic flux for each metabolite is computed and updated in this last step of the DMA.

The hybrid algorithm, together with the discrete event simulation of slow timescale events builds the integrated simulation environment. Next, we outline the implementation architecture of the hybrid algorithm based on the extension of the *iSimBioSys* framework.

## 6.4 Hybrid simulation architecture

The *HimSim* framework has been built on the discrete event based *iSimBioSys* platform outlined in Chapter 3. Specifically, the hybrid simulation algorithm has been developed as a pluggable modular object which interacts with the discrete event (DES) engine under the control of the central event schedular. Moreover, the hybrid module interfaces with an integrated database to incorporate the pathway knowledge of the different networks into the combined simulation platform.

As outlined previously, one of the major hurdles in building an integrated modeling framework is the lack of coherence in storage and retrieval of pathway information stored across disparate databases. As part of the *HimSim* framework, we have developed a custom database schema for storing data on signaling, gene and metabolic networks curated from different databases, like KEGG [96], EcoCyc [52] and CellSignaling [30]. The major schemas of the database are outlined below:

- *Signaling event database*: The signaling events database stores the events associated with a particular signaling pathway. Each entry in the signaling database consists of a list of events associated with the pathway (each event being characterized by the input and output molecular resources and the biological model). The events and the signaling pathway are curated from existing databases and literature search.
- *Gene regulatory network database*: The gene regulatory network (GRN) database stores the transcription factors for a model organism (in this case bacterial cell *E.Coli*) together with a known list of genes which are upregulated and down-regulated by the transcription factor. As elucidated later, the GRN database can be automatically populated from flat files and gene regulatory information stored in System Biology Markup language (SBML) schema formats.
- *Metabolic network database*: Once the signaling molecules and genes have been identified, the database schema for the metabolic reaction network (MTN) is de-

fined, consisting of the reactants, products, stoichiometry and the list of enzymes associated with a particular reaction. Since each enzyme is formed of multiple proteins (regulated by genes in the GRN), a protein-protein interaction (PPI) table is defined which stores the list of enzymes together with the gene-products associated with the particular enzyme.

It may be mentioned here that the database has been currently implemented on an object-oriented (OO) database management schema (DBMS) which provides a middle-layer for storing and querying entries through objects defined in the database. As mentioned earlier, such an integrated database schema provides a single interface for simulating cellular pathways and in identifying possible knowledge gaps.

The overall architecture of the hybrid simulation framework is outlined in Fig. 6.7. As shown in the figure, the core simulation engine consists of the DES module and the DMA module which interact with each other to capture the dynamics of the system through the different signaling, regulatory and metabolic events. Specifically, the events database interacts with these modules which in term communicate with the different pathway databases to simulate the system in time.

### 6.4.1 Discrete metabolic analysis (DMA) simulation engine

The DMA engine forms the heart of the hybrid simulator and implements the hybrid simulation algorithm elucidated in the previous section. When the central event scheduler schedules a metabolic event object, an instance of the DMA engine is invoked. The DMA engine then queries the database to obtain the dynamics of the various gene products and enzymes and computes the fireable metabolic reactions and their molecular concentrations based enzymatic activities and reaction stoichiometries. Once the metabolic reactions have been determined and the flux across each metabolite computed, the DMA engine updates the molecular resources and hands back control to the DES

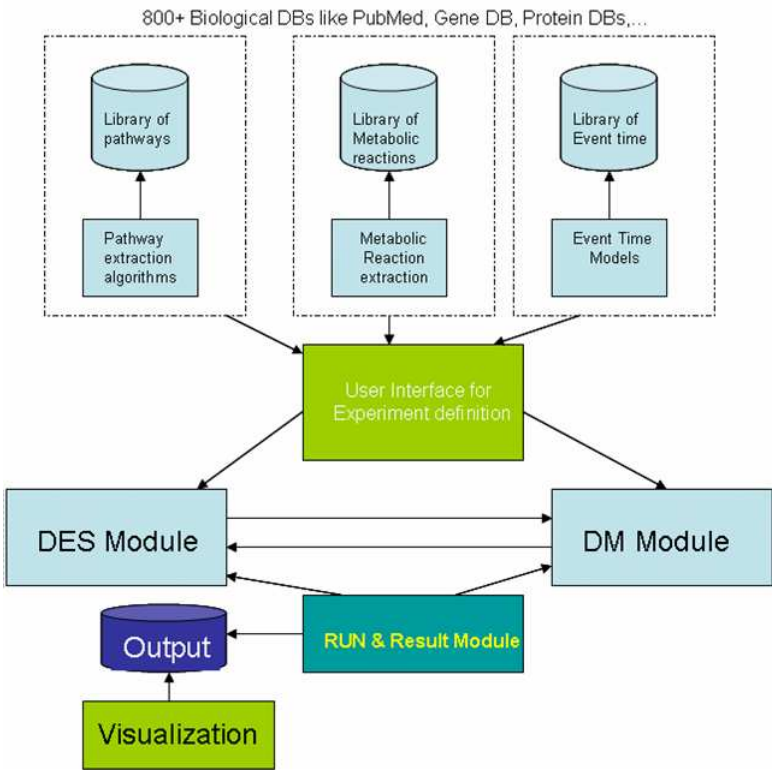



Figure 6.7. The hybrid simulation architecture.

module to continue the simulation of the signal transduction and regulatory pathways. In this way, through the interactions of the two engines and the corresponding database, the simulation traces the temporal evolution of the biological network (in terms of change in the molecular concentration of the different gene products and metabolites).

## 6.5 Experimental validation

In the previous sections, we have systematically built the basic building blocks of the hybrid simulation framework. The *HimSim* architecture provides a generic platform for developing database and simulation models for studying cellular pathways of model organisms. In this section, we develop a detailed genome-scale pathway database for the bacterial cell *Escherichia Coli*, outlining the specific signaling, gene regulatory and

metabolic reaction networks which have been studied as part of this dissertation, both for validation with experimental results as well generating *in silico* predictions.

### 6.5.1 Regulation of central metabolism in *E.Coli*

The model organism chosen as part of the study in this work is the single cell bacteria, *Escherichia Coli*(*E.Coli*), particularly the K-12 MG1655 strain with a circular chromosome of length 4639675 base pairs [97], 4243 protein genes and 157 RNA genes. Because of the long history of research on *E. Coli*, both in the biological and computational communities, a wealth of information on the gene regulatory and signaling networks of the cell are available [117, 90]. Moreover, biological evidence on the metabolic reaction network of the bacteria, particularly metabolite flux data as well as recent microarray data on global gene expression profiles [120, 11, 117] are readily available.

In particular, we focus on the key metabolic reaction pathways involved in central metabolism of *E. Coli* (glycolysis, tricarboxylic acid (TCA) cycle, pentose phosphate pathway, serine biosynthesis, pyruvate oxidation). Together the central metabolic network consists of approximately 500 metabolic reactions, 450 enzymes which are regulated by 800 gene under the control of 7 global transcription factors. The major signal transduction pathways, gene regulatory networks and the metabolic reactions together with the relevant resources are outlined next.

#### 6.5.1.1 Signal transduction pathways

Based on existing literature, [117, 120, 74, 190] on the signal transduction pathways inducing transcriptional regulation on central metabolism, four key signaling networks were identified, which regulate downstream metabolic genes under different concentration of carbon source(mainly glucose medium growth condition) and oxygen source (aerobic and anaerobic growth media).

*Escherichia coli* has several elaborate sensing mechanisms for response to the availability of oxygen and the presence of other electron acceptors. The adaptive responses are coordinated by a group of global regulators, which includes the one component FNR (fumarate, nitrate reduction) protein, and the two-component Arc (aerobic respiration control) system. With the initial onset of anaerobiosis ArcA is activated, and if these conditions persist, FNR is activated leading in turn to the upregulation of ArcA and the amplification of its effect.

The Arc system is a two-component regulatory system composed of ArcA, the cytosolic response regulator, and ArcB, the transmembrane histidine kinase sensor. ArcB is activated during the transition from aerobic to microanaerobic growth, and remains in the activated state during anaerobic growth. The increased level of phosphorylated ArcA represses the synthesis of some enzymes, such as the citric acid cycle enzymes, succinate dehydrogenases, lactate dehydrogenase, fumarase, pyruvate dehydrogenase, and the low oxygen affinity cytochrome oxidase, while it activates the expression of other enzymes such as cytochrome deoxidase and enzymes involved in fermentative metabolism [163].

The FNR protein contains an Fe-S cluster that serves as a redox sensor. The FNR system is active under microaerobic to anaerobic conditions and induces the expression of genes that permit anaerobically growing *E. coli* to transfer electrons to alternative terminal acceptors. It also represses the aerobic genes, cytochrome deoxidase, and NADH dehydrogenase II. It acts as a positive regulator of genes expressed under anaerobic fermentative conditions such as aspartase, formate hydrogenase, fumarate reductase, and pyruvate formate lyase [163]. Fig. 6.8 shows the effect of oxygen on the regulation of metabolic genes involved in central metabolism in *E.Coli*.

Another important pathway which controls glucose uptake and the regulation of genes involved in the phosphotransferase system (PTS) is the glucose mediated Mlc system. As shown in Fig. 6.9, glucose transporter IICB-Glc stimulates the transcription

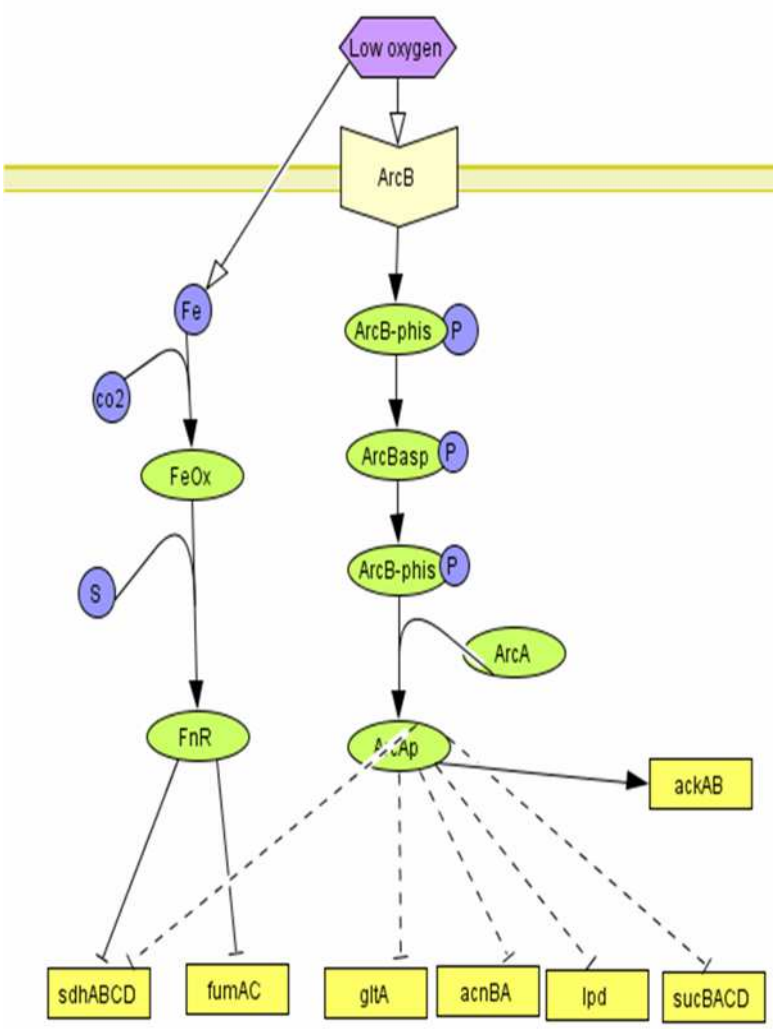



Figure 6.8. The ArcAB two component system.

of several genes involved in PTS by acting as membrane sequester for the Mlc, thereby relieving the negative effects of the global repressor factor. In high glucose, Mlc is sequestered and is not available for regulation of genes like *crr* and *ptsGHI*.

Glucose also causes catabolic repression by lowering the levels of intracellular cAMP and CRP proteins. Thus, under high glucose conditions, CRP proteins are repressed thereby lowering the expression level of genes involved in pyruvate oxidation. However, under low glucose, the pyruvate pathway is activated leading to utilization of acetate as

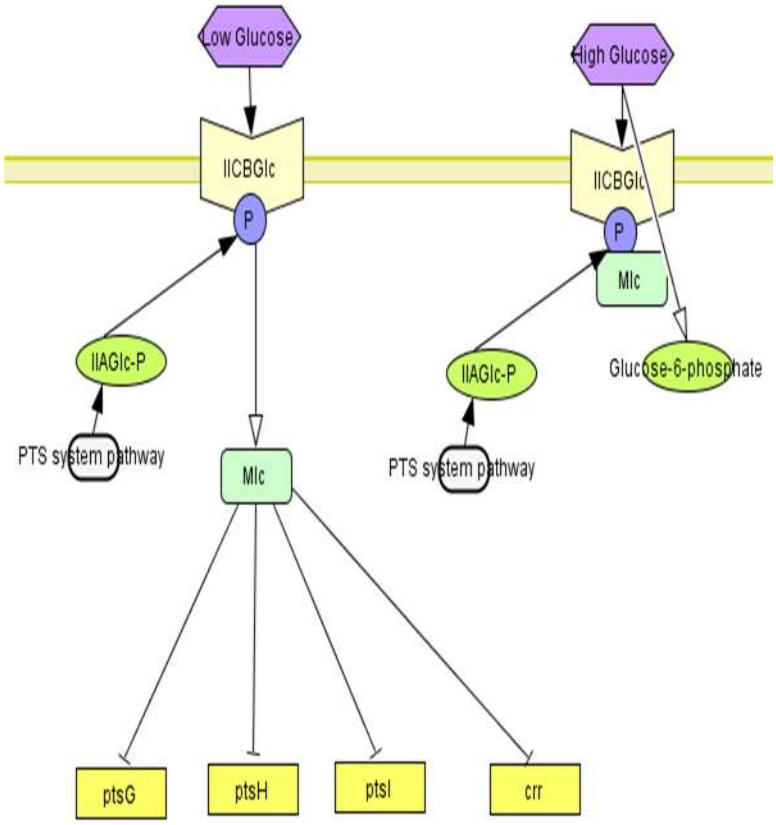



Figure 6.9. Mlc sequestration in glucose media.

the growth medium due to increase in CRP levels which cause down-regulation of *cra* genes. Fig. 6.10 captures the effect of the CRP pathway under glucose medium.

#### 6.5.1.2 Gene regulatory pathways

In order to build the gene regulatory network involving the key transcription factors controlling regulation in *E.Coli*, we reconstruct the network based on data provided by Ma. et. al in [68]. Genome-wide study by Shen. et. al [162] together with data from public databases, reveal a multi-layer hierarchical structure for the entire gene regulatory network which are under the control of around 7 key global regulators including Mlc, ArcAB, CRP, FoxS etc. Recent experiments [11] elucidated the role of these global

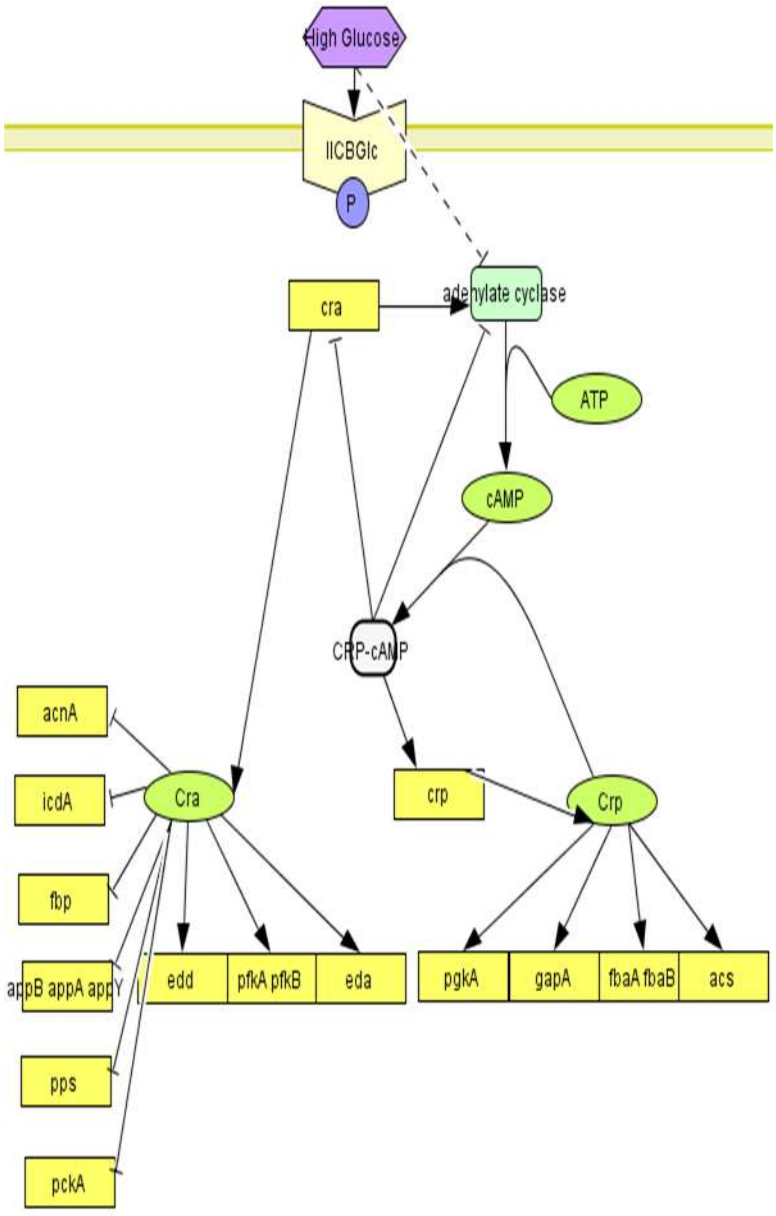



Figure 6.10. Regulation of cAMP and CRP proteins.

regulators in controlling the distribution of metabolic fluxes across the central metabolism network in *E.Coli*. The gene regulatory network used in this study consists of 1024 genes and 2065 interactions under the control of the global transcription factors (data obtained

from EcocCyc and flat files provided by Ma et.al in [68]). An important observation in this regard, is the fact the existence of a multi-tier hierarchy facilitates the building of the signaling networks. Constructing the signal transduction events for the global factors outlined above and establishing the association with the gene regulatory pathways in an integrated database helps to capture the interplay between these pathways.

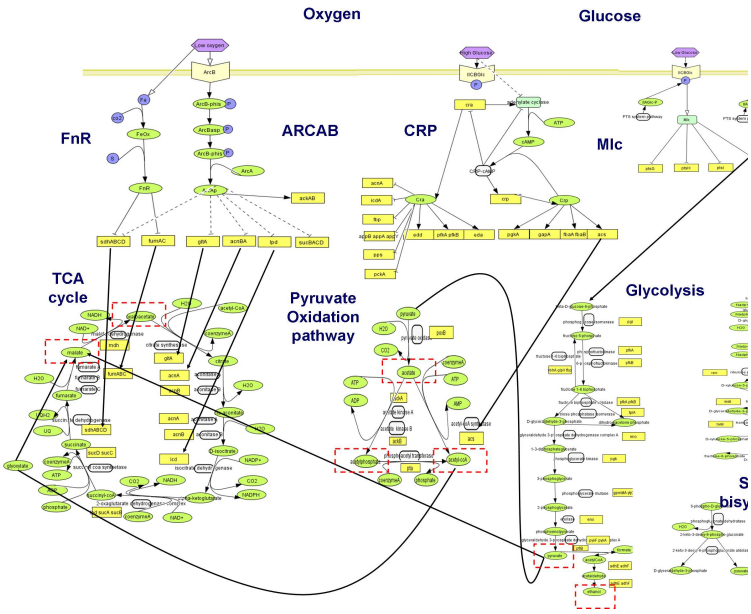



Figure 6.11. Global map of central metabolism and its regulation in *E.Coli*.

### 6.5.1.3 Metabolic reaction pathways

The Ecocyc [52] database provides a comprehensive map of the core reactions involved in central metabolism of *E.Coli*. Reactions, together with their enzymes and protein complexes were curated from the Ecocyc database (available in flat files format from the Ecocyc website [52]) were generated and parsed into the metabolic network database outlined previously. The comprehensive map of the signaling networks, together

with associated gene and the corresponding metabolic network for *E.Coli* (under glucose and oxygen media) is depicted in Fig. 6.11.

### 6.5.2 Dynamics of aerobic growth on glucose

We study the regulation of central metabolism in *E.Coli* under glucose media and oxygen conditions to capture the effects of the metabolic genes controlling the flux across the glycolysis and pyruvate oxidation pathways. The specific network consists of a subset of the global system, with 80 metabolic reactions and 100 genes regulated by the ArcAB and FnR signals under oxygen medium and CRP/Mlc system under glucose growth media. The reduced network is depicted in Fig. 6.12. In [120], Covert. et.al applied the rFBA approach to validate experimentally observed dynamics of the flux across glucose, acetate and ethanol together with the expression patterns for the metabolic genes (expressed in boolean form).

In the first set of simulations, the dynamics of central metabolism were observed across the glycolysis pathway (in terms of change in glucose concentration) along with acetate flux under aerobic growth on glucose media. Fig. 6.13 shows the uptake of extracellular glucose by the cell (both experimental observations as well as hybrid simulation results [1].

As shown from the figure, glucose concentration decreases over time as the cell uses glucose as the primary source of carbon. This causes the flux across the glycolysis pathway to increase as glucose is converted to pyruvate through the glycolytic path. The increase in pyruvate causes the activation of the pyruvate oxidation chain thereby leading to an increase in acetate concentration (shown in Fig. 6.14). As observed in experiments and reproduced in the hybrid simulation, depletion of glucose leads to the reutilization

[1]The simulation results report the observations over 100 runs with the error bar depicted the average value with the 95% error margins

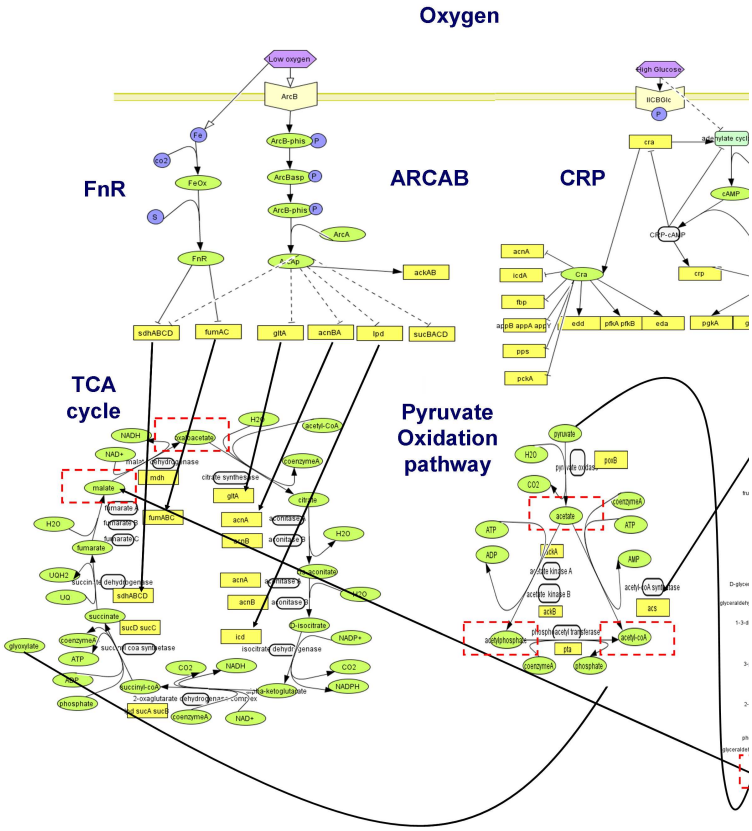



Figure 6.12. Network of central metabolism in *E.Coli*.

of acetate as the growth substrate, causing a decrease in the acetate concentration. The simulation results on glucose uptake and acetate reutilization reproduce the experimentally observed results within the bounds of simulation error (Fig. 6.13 and Fig. 6.14 for the flux distribution across these metabolites).

### 6.5.3 Dynamics of anaerobic growth on glucose

In another experimental scenario, the same network of signaling pathways and metabolic reactions were subjected to growth on glucose medium, but under anaerobic conditions. The glucose uptake and acetate flux profiles observed under these low oxygen concentrations were experimentally observed and reported in [120]. Fig. 6.15 and

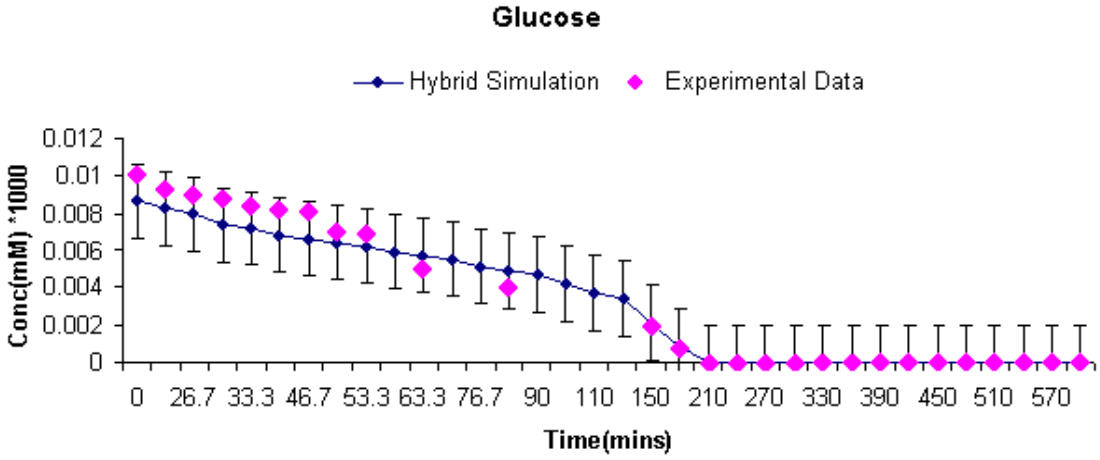



Figure 6.13. Glucose uptake under aerobic conditions.

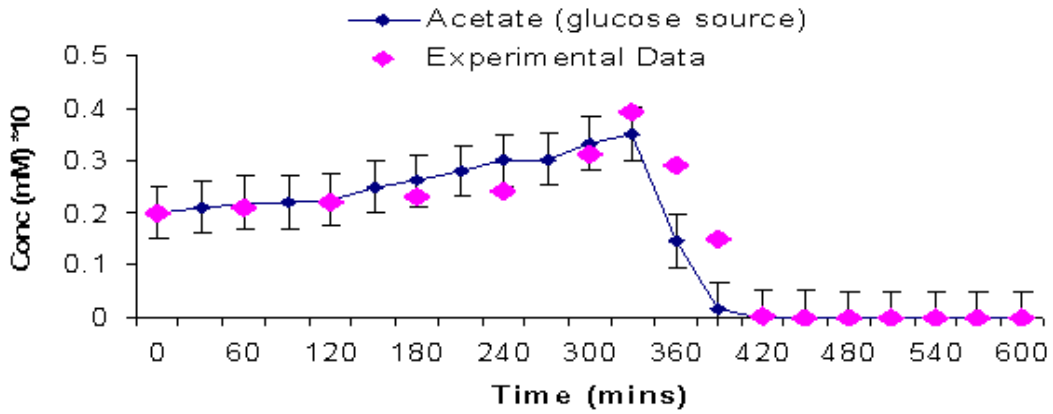



Figure 6.14. Acetate growth and reutilization.

Fig. 6.16 show the simulation results vis-a-vis the experimental data. The simulation results captures the observed effect of glucose uptake and corresponding increase in flux across acetate within the error bounds of the simulation. It may be noted that under the conditions of anaerobic growth, the acetate reutilization is not invoked on glucose

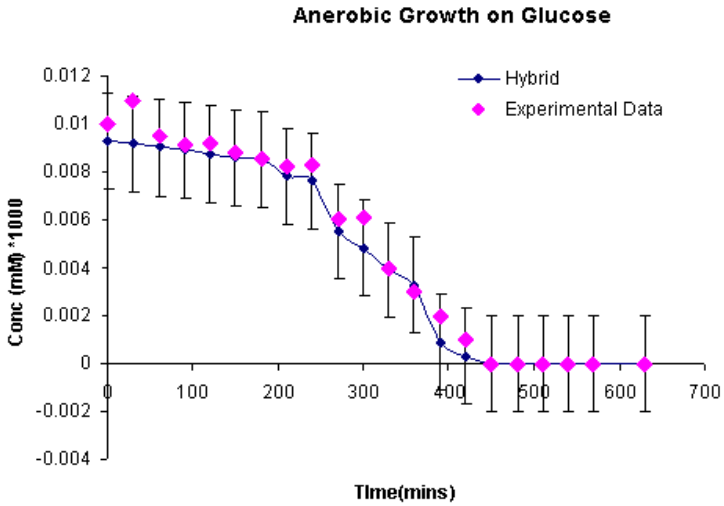



Figure 6.15. Glucose uptake under anaerobic growth.

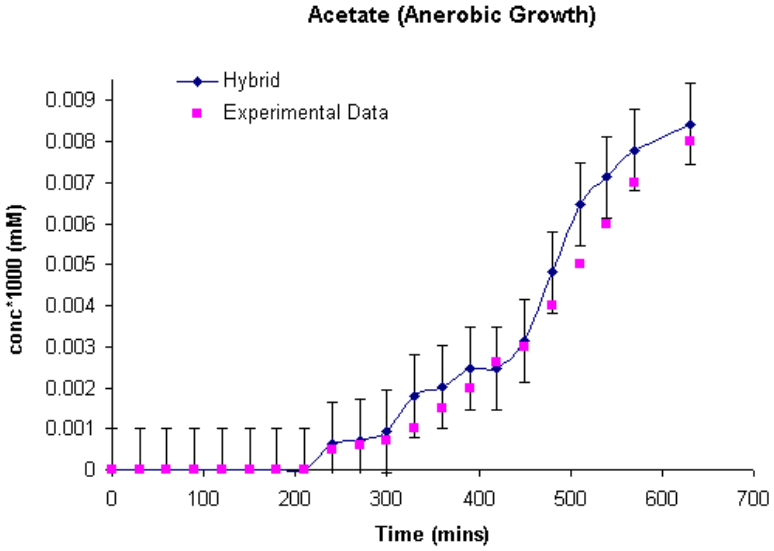



Figure 6.16. Acetate flux under anaerobic growth.

depletion. However, anaerobic conditions trigger the activation of the ArcBA and FnR pathway as the cell senses the lack of oxygen in its environment. The FnR pathway becomes active under complete anaerobic conditions triggering the expression of genes

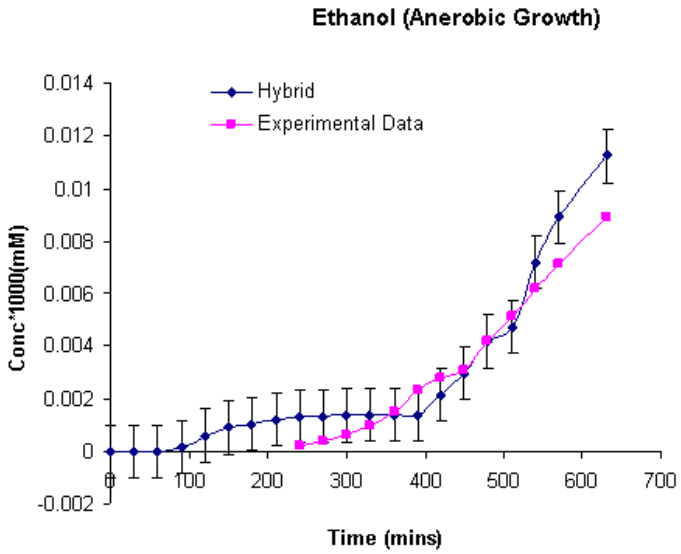



Figure 6.17. Ethanol flux under anaerobic growth.

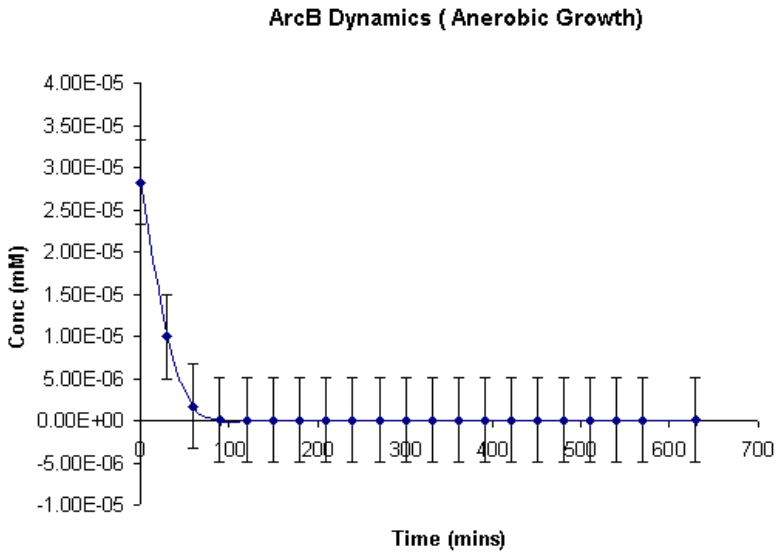



Figure 6.18. ArcB concentration change under anaerobic conditions.

regulating the enzymes catalyzing the conversion of pyruvate to acetaldehyde and ethanol

(specifically adhE and adhF genes). Thus, as shown in Fig. 6.17, the increase in ethanol flux is marked by an initial delay owing to the regulatory effect of FnR.

On the other hand, the ArcAB system is a two-component signaling pathway activated by the low oxygen concentration. This activation leads to a decrease in the number of membrane bound (unphosphorylated) sensory ArcB molecules in the system with time (shown in Fig 6.18). The activation of the ArcAB system causes expression of genes involved in the TCA cycle leading to an increase in flux across it, as reported in the increase in malate concentration in Fig. 6.19.

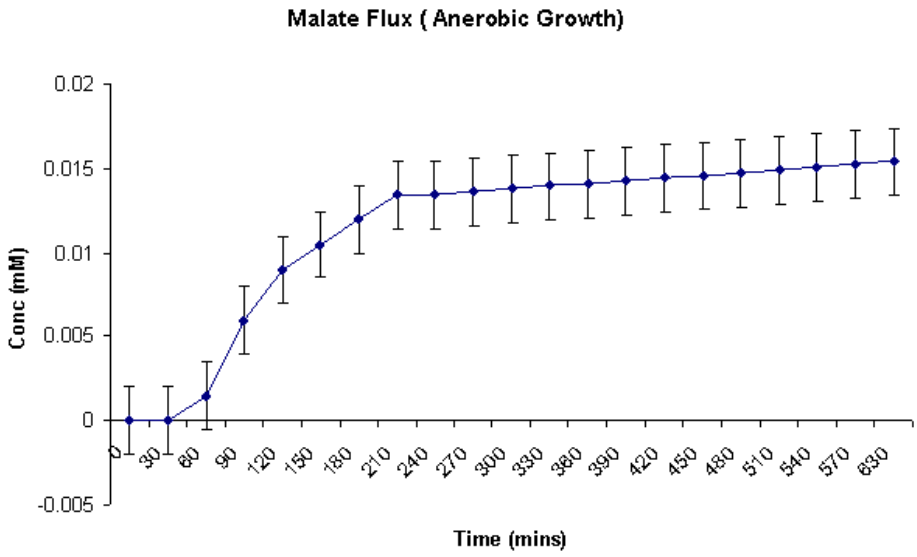



Figure 6.19. Flux across malate (TCA cycle).

## 6.6 *In Silico* results

The hybrid simulation approach provides a global holistic picture of the dynamical interplay of the different signals and molecules. Such a picture provides detailed insights

into the inner working of pathways and signals and facilitates design of further *in silico* experimental protocols.

### 6.6.1 Gene expression profiles for growth on glucose media

In the previous section, we validated experimentally observed flux distributions for the different metabolites involved in central metabolism of *E.Coli* under growth and oxygen conditions. While the hybrid simulation was successful in reproducing the dynamics of flux change for the metabolites, it also allows the study of time-course evolution of the different metabolic genes and thus identify the effect of signal transduction and gene regulation on the metabolic network.

In the experiments on flux change for anaerobic growth on glucose, it was observed that depletion of glucose leads to the reutilization of acetate as the growth substrate, causing a decrease in the acetate concentration. From a system level pathway perspective, this phenomena can be traced to the activation of the CRP signaling pathway under low glucose which causes an upregulation of the *acs* gene responsible for the production of the acetylCoA synthetase protein which increases the flux across acetyl-CoA (refer to Fig. 6.12). The hybrid simulation quantitatively captures the effect of the genes and their expression levels through time-course evolution of their molecular concentrations. Fig. 6.20 shows the gene expression profile for Acs protein under anaerobic growth on glucose media, indicating how the protein profile increases due to the CRP signal.

It is important to note here that the hybrid simulation allows tracing the temporal dynamics of the changing concentration of the genes instead of providing a boolean high-low value for the expression. This allows the study of *in silico gene expression array* which can profile different genes as concentration changes instead of boolean values. Fig. 6.21 shows the profile of the gene AceA under glucose growth conditions. The *aceA* gene is upregulated by the transcription factor *Cra* which is inhibited by CRP. Thus, under

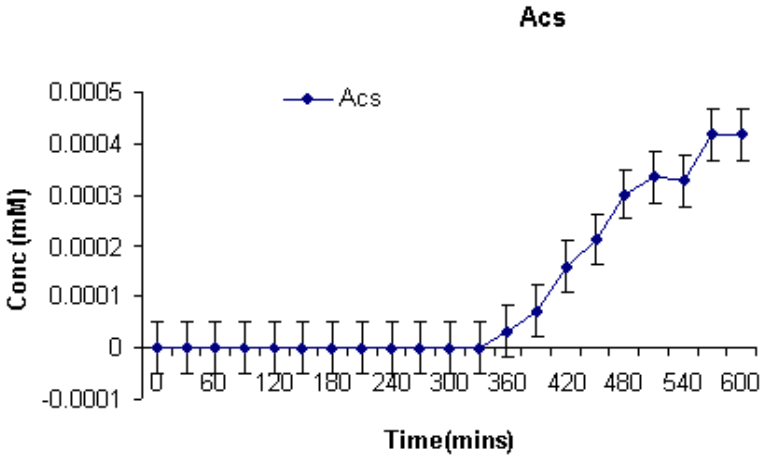



Figure 6.20. Acs expression dynamics under aerobic growth on glucose.

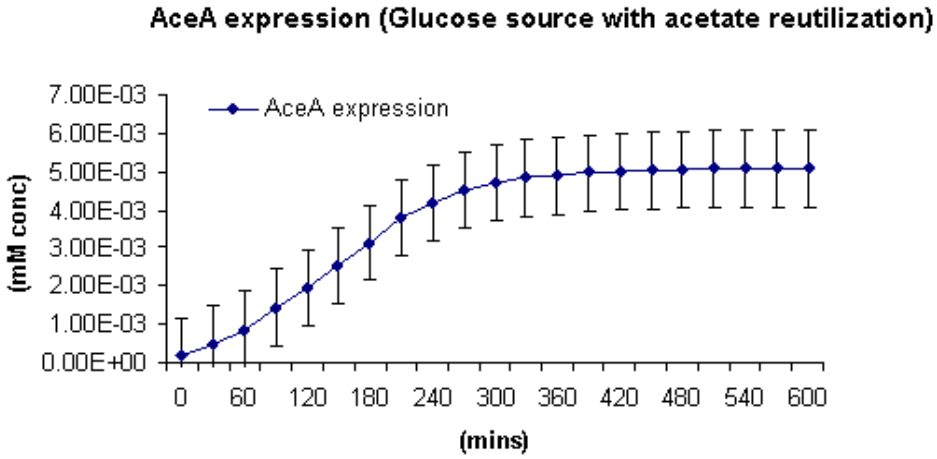



Figure 6.21. AceA gene expression dynamics under CRP signal.

high glucose, CRP is inactive leading to upregulation of AceA by Cra. When glucose becomes depleted, CRP is activated down regulating Cra and its positive regulatory effect on AceA. In Fig. 6.22 and Fig. 6.23, the effect of the Mlc sequestration pathway is outlined together with its effect on the gene *crr*. Under glucose rich media, active Mlc

decreases due to sequestration effect of the glucose transporter IICB-Glc molecules as reported experimentally in [190] causing inhibition of Mlc positive-regulatory effect on Crr protein leading to a decrease in its concentration.

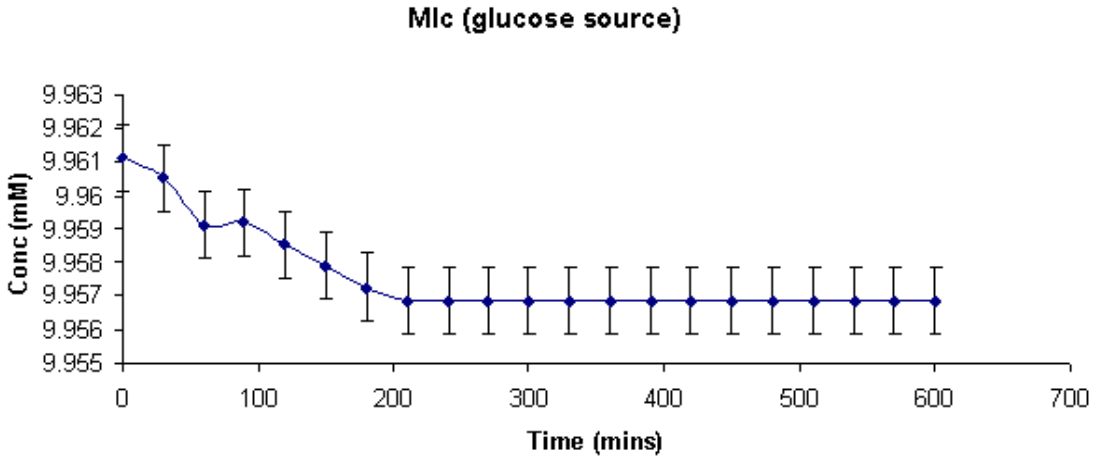



Figure 6.22. Mlc sequestration effects.

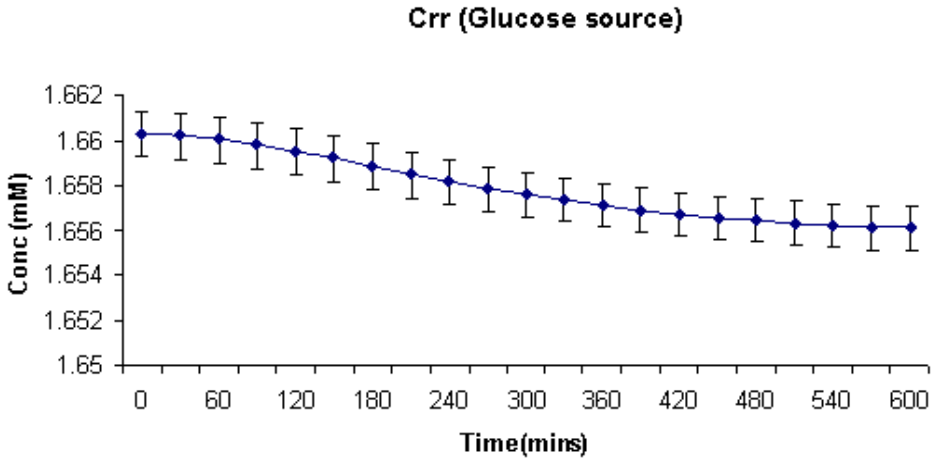



Figure 6.23. Effect of Mlc signal on crr expression.

These gene expression profiles illustrate the efficacy of the hybrid simulation technique in capturing the simultaneous effect of multiple signaling cascades on downstream genes and the corresponding regulation of metabolic phenotypes.

#### 6.6.1.1 *In silico* analysis of gene deletion effects on aerobic growth

As the effect of CRP on Acs was identified as the key driver of acetate reutilization in the previous experiments, we conducted *in silico* simulation under CRP gene deletion (null mutant) conditions. Knockout of the CRP strain causes the acetate reutilization phenomena to disappear as shown in Fig. 6.24. An interesting observation in this gene deletion simulation was the decreased growth rate of acetate under the same glucose rich media. This can be explained by noting the fact that two of the enzymes controlling glycolytic flux, *gapA* and *pgk* are positively regulated by CRP. Thus, knockout of CRP causes these proteins and their corresponding enzymes to operate at basal levels only instead of higher levels under CRP non-mutant conditions causing the flux across acetate to decrease.

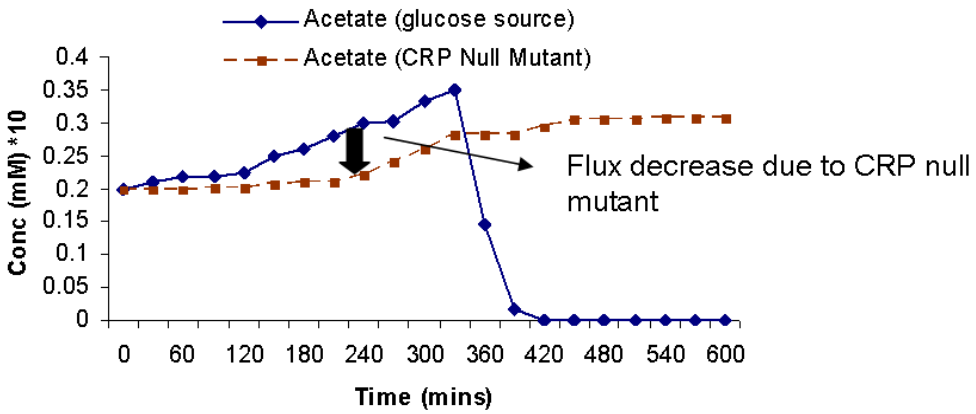



Figure 6.24. Effect of CRP gene knockout on acetate flux.

#### 6.6.1.2 *In silico* analysis of pyruvate oxidation pathway on anaerobic growth

With the dynamics of metabolic flux distribution across central metabolism validated against experimental data under conditions of anaerobic growth on glucose, we focused on specific control of other pathways by the ArcAB system, which is activated under such conditions and is a critical regulator of several metabolic genes. ArcAB positively regulates the acetate kinase genes (*ackA* and *ackB*) which control acetylphosphate formation from acetate in the pyruvate oxidation pathway. In order to quantify the effect of this pathway on the metabolic flux across acetylphosphate, simulations were conducted with anaerobic conditions followed by aerobic conditions leading to shutdown of the pathway. As seen in Fig. 6.25, the fluctuation in oxygen signal causes the AckB gene to be transiently upregulated followed by its decrease when the ArcAB signal switches off. This causes the corresponding flux across acetylphosphate to increase briefly but maintain its basal value once the gene has been turned off, Fig. 6.26.

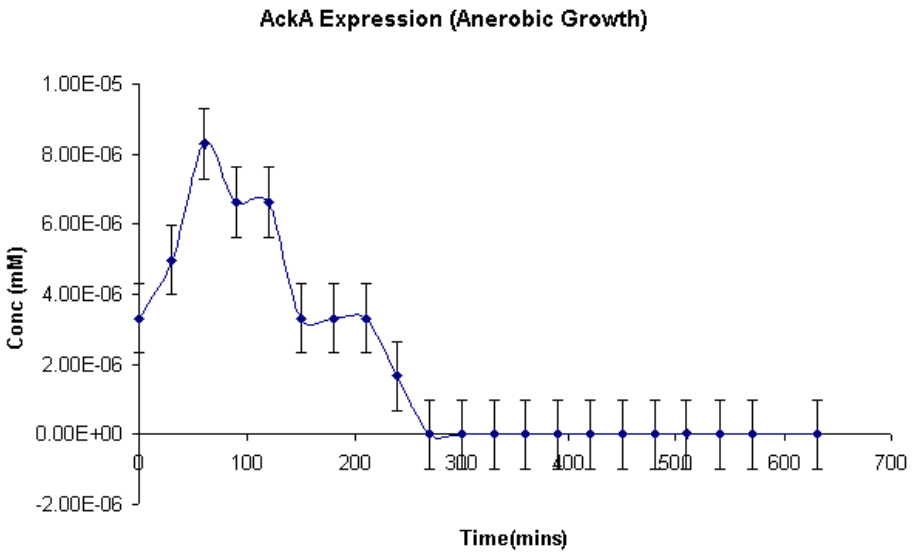



Figure 6.25. AckA gene regulation by ArcAB system.

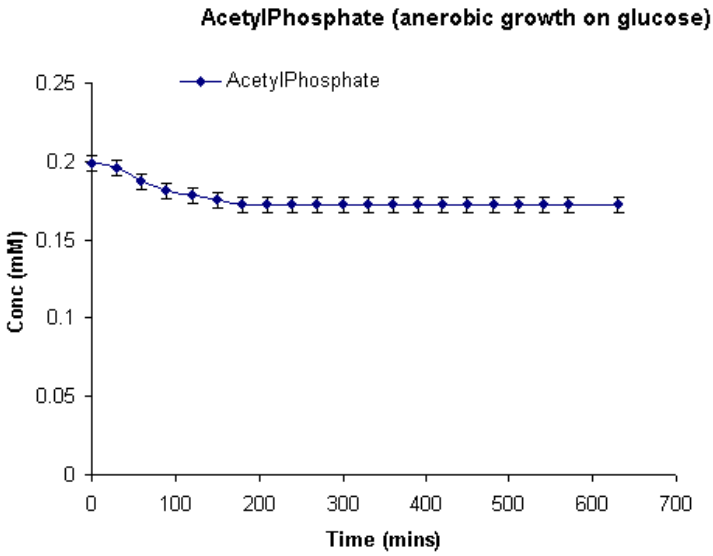



Figure 6.26. Effect of ArcAB signal on acetylphosphate flux.

Thus, the hybrid simulation platform allows the study of gene deletions effects as well obtaining granular quantification of the role of specific signals on their downstream genes and metabolites.

## 6.7 Summary

In this chapter, we outlined the fundamental challenges in building genome-scale simulation models of the dynamical interaction of different molecules pathways operating in different time-scales. With the limitations of current approaches and techniques in perspective, we developed a hybrid computational framework, called *HimSim*, which incorporates stochastic DES with flow-based algebraic model of metabolic analysis (DMA). The interaction of stochastic signaling and regulatory events with fixed time interval metabolic events overcome the "stiffness" of building multi-time scale simulations. We also elucidated on the detailed of an integrated database schema which serves as a com-

mon interface for different pathway databases and interacts with *HimSim* to provide a genome-scale cell simulation platform. The efficacy of the hybrid technique is illustrated through a detailed model of central metabolism in *E.Coli* cells, incorporating comprehensive data on signals, genes and metabolites curated from disparate sources. The simulation experiments on glucose and oxygen growth media illustrate the feasibility of the hybrid simulation technique in both validating experimental data as well as providing biologically relevant insights into the fine-grained control of signaling and regulatory systems on metabolic networks and their phenotypes.

# CHAPTER 7

# CONCLUSION

A discrete event based stochastic biosimulation platform provides a generic computational framework to study temporal variations in cellular processes at single molecule level. It allows systematic analysis of different bio-molecular events and their interactions in an effort to unravel biological intelligence *in silico*. The event paradigm provides flexibility in abstracting system at a micro, meso or macro scale within a common computational model.

In this thesis, we have outlined the framework of discrete event based simulation and modeling, building the computational artefact and implementing the simulation engine. Using the platform, we developed *in silico* models of biological processes, spanning from a top-down signalling network system to a bottom-up mechanistic modeling of prokaryotic gene regulation. Moreover, the hybrid simulation framework, allows the genome-scale study of the interplay between gene regulation, signal transduction and metabolic reaction networks as outlined for the case studies involving the regulation of central metabolism in carbon-rich and oxygen-limited growth environments for the bacteria cell *Escherichia Coli*.

Discrete event based modeling techniques can give computational advantages for molecular level study of biologic pathways and the impact of stochasticity in them. In applying discrete event simulation techniques in the study of biological systems , it is of foremost importance to map available knowledge into parameterized model of events. Also, with different order of time-scales for biological events (typically, between gene reg-

ulatory and metabolic reaction events), a hybrid discrete event based approach provides the flexibility of capturing their interactions in time.

It is pertinent to keep in perspective that for building comprehensive, system-wide computational models of complex disease pathophysiologies, different modeling techniques, from top-down physiological models to bottom-up atomic and molecular interaction techniques, have to integrated in a common platform. In this light, a discrete event paradigm, provides a "middle-out" approach by allowing the characterizing of biological functions through events at different levels of granularity.

This approach is particularly promising for the pharmaceutical industry, as one of the key challenges in the drug discovery process is the prohibitively expensive process of target attrition. With the simulation and modeling framework outlined in this work, it would be possible to study various drug targets at the molecular level while incorporating physiologically relevant information at the level of organs and tissues. The flexibility in developing models at different levels of granularity facilitates capturing network pathway information in a common computational platform.

## 7.1 Future research directions

A fundamental requirement for application of a discrete event based stochastic computational framework for large scale disease modeling is the ability to scale the simulation across a distributed computing architecture. While the focus of this work has been primarily on the development of the simulation and modeling framework for a single processor architecture, future work would involve extending the architecture over a distributed platform.

It is also important to note that the various techniques of computational biomodeling and simulation cater to specific biological systems and processes. While each of the techniques make their own assumptions of the system view, linking them together

through well-defined, inter-operable interfaces to render a coherent global view remains the holy grail for developing the next generation of personalized medicine. Rapidly changing interpretations of existing knowledge gaps, lack of common standards for collating information from a wide gamut of databases are a few of the computational challenges engineers and scientists working in this field have to overcome in their strive towards comprehending the complexity of cellular processes.

## REFERENCES


[1] A. Arkin, J. Ross, H.H McAdams, Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage lambda-Infected Escherichia coli Cells, Genetics 149, 1633-1648, 1998.

[2] A. A. Aristidou, K.Y San, G.N Bennett, Metabolic flux analysis of Escherichia coli expressing the Bacillus subtilis acetolactate synthase in batch and continuous cultures. Biotechnol Bioeng 63: 737-749, 1999.

[3] A. Barnard, A. Wolfe, S. Busby, Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes, Current Opinion in Microbiology 7, 102-108, 2004.

[4] A. Becskei, B.B Kaufmann, A. van Oudenaarden, Contributions of low molecule number and chromosomal positioning to stochastic gene expression, Nature Genetics. 37(9): 937-944, 2005.

[5] A. Chatterjee, K. Mayawala, J.S. Edwards, D.G. Vlachos, Time accelerated Monte Carlo simulations of biological networks using the binomial tau-leap method, Bioinformatics.21(9):2136-7, 2005.

[6] A. D. McCulloch and G. Huber, Integrative biological modeling *in silico*, "In Silico" Simulation of Biological Processes, Novartis Foundation Symposium 247, 2002.

[7] A.M Kierzek, J. Zaim, P. Zielenkiewicz, The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression, J Biol Chem. Mar 16;276(11):8165-72, 2001.

[8] A.L Barabasi, & R. Albert, Emergence of scaling in random networks. Science 286, 509512 1999.

[9] A.J Carpousis, The Escherichia coli RNAdegradosome: structure, function and relationship to other ribonucleolytic multienzyme complexes, Biochemical Society Transactions 30 (2), 150-154, 2002.

[10] A.J Carpousis, The Escherichia coli RNAdegradosome: structure, function and relationship to other ribonucleolytic multienzyme complexes, Biochemical Society Transactions 30 (2), 150-154, 2002.

[11] A. Perrenoud and U. Sauer, Impact of Global Transcriptional Regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc on Glucose Catabolism in Escherichia coli, Journal of Bacteriology, 3171-3179, Vol. 187, No. 9, May 2005.

[12] A. Regev, Representation and simulation of molecular pathways in the stochastic pi-calculus, In Proceedings of the 2nd workshop on Computation of Biochemical Pathways and Genetic Networks, 2001.

[13] A. Regev, E. M. Panina, W. Silverman, L. Cardelli and E. Shapiro, Bioambients: an abstraction for biological compartments, Theoretical Computer Science, 325: 141-167, 2004.

[14] A.M Kierzek, STOCKS: STOChastic Kinetic Simulations of biochemical systems with Gillespie algorithm, Bioinformatics 18 (3), 470-481, 2002.

[15] A.M Uhrmacher, Concepts of Object and Agent Oriented Simulation, Transactions on SCS, 14(2), pp.59-67, 1997.

[16] A.M. Uhrmacher, C. Priami, Discrete event systems specification in systems biology - a discussion of stochastic pi calculus and DEVS, Proceedings of the Winter Simulation Conference, pp. 317-326, 2005.

[17] A.M Uhrmacher, P. Tyschler, D. Tyschler, Modeling and Simulation of Mobile Agents, Future Generation Computer Systems, 17, pp. 107-118, 2000.

[18] A.S Lynch, E.C.C. Lin, Regulation of aerobic and anaerobic metabolism by the Arc system. In: Lin ECC , Lynch AS , editors. Regulation of gene expression in Escherichia coli. New York: Chapman & Hall. p 361-381, 1996.

[19] B. Alberts, D. Bray, and J. Lewis, Molecular Biology of the Cell, Garland Science, 2002.

[20] B. Alberts, et. al, Molecular Biology of the Cell, Fourth Edition, Garland Science, 2002.

[21] B. B. Aldridge, J. M. Burke, D. A. Lauffenburger, P. K. Sorger, Physicochemical modelling of cell signalling pathways, Nat Cell. Biol., pp. 1195 - 1203, 2006.

[22] B.G. Cox, Modern Liquid Phase Kinetics, Oxford University Press, Oxford, 1994.

[23] B.P Zeigler, Theory of Modeling and Simulation, Academic Press, 2000.

[24] C.C Guet, M. B. Elowitz, W. Hsing, S. Leibler, Combinatorial synthesis of genetic networks, Science 296, 1466-1470, 2002.

[25] C.J. Morton-Firth, Stochastic Simulation of Cell Signalling Pathways, PhD thesis, University of Cambridge, 1998.

[26] C. H Yuh, H. Bolouri & E.H Davidson, Genomic cis-regulatory logic: experimental and computational analysis of a sea urchin gene. Science 279, 18961902, 1998.

[27] C. Kuttler, Modeling bacterial gene expression in a stochastic pi-calculus with concurrent objects, Phd Thesis, 2007, http://www2.lifl.fr/ kuttler/

[28] C. Kuttler, Simulating Bacterial Transcription and Translation in a Stochastic pi Calculus , T. Comp. Sys. Biology: 113-149, 2006.

[29] C. van Gend, U.Kummer, STODE - automatic stochastic simulation of systems described by differential equations, ICSB 2002.

[30] Cell Signaling Database, http://stke.sciencemag.org/cm/

[31] Cornish-Bowden, A. Fundamentals of Enzyme Kinetics; Portland Press: 1995.

[32] D. Adalsteinsson, D. McMillen, and T. Elston, Biochemical Network Stochastic Simulator (BioNetS): software for stochastic modeling of biochemical networks, BMC Bioinformatics 5 (1), 24, 2004.

[33] D. Browning, S. Busby, The regulation of bacterial transcription initiation. Nat Rev Microbiol 2, 57-65, 2004.

[34] D. Bray, Protein molecules as computational elements in living cells. Nature 376, 307312, 1995.

[35] D. Degenring, M. Rhl, A.M. Uhrmacher, Discrete Event Simulation for a Better Understanding of Metabolite Channeling, In Proc. Of the International Workshop on Computational Methods in Systems Biology, Rovereto, Italy, Springer, LNCS Series, 2002.

[36] D. Fell, Understanding the Control of Metabolism; Portland Press, 1997.

[37] D.J Shaw, J.R Guest, Molecular cloning of the fnr gene of Escherichia coli K12. Mol Gen Genet 181: 95-100, 1981.

[38] D. G. Vlachos, The emerging field of multiscale analysis: a review with examples from systems biology, materials engineering, and fluid-surface interacting systems, Adv. Chem. Eng., 2005.

[39] D. Orrell, S. Ramsey, P. de Atauri, and H. Bolouri, A method for estimating stochastic noise in large genetic regulatory networks, Bioinformatics 21 (2), 208-217, 2005.

[40] D. T. Gillespie, Exact Stochastic Simulation of Coupled Chemical Reactions, The Journal of Physical Chemistry, Vol. 81, No. 25, pp. 2340-2361, 1977.

[41] D.T. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems, J. Chem. Phys., 115, 17161733, 2001.

[42] D.T. Gillespie, A general method for numerically simulating the stochastic time evolution of coupled chemical reactions, J. Comput. Phys., 22, 403-434, 1976.

[43] D.T. Gillespie, Concerning the validity of the stochastic approach of chemical kinetics, J. Stat. Phys., 16, 311-319, 1977.

[44] D. Kuo, J. D. Keasling, A Monte Carlo simulation of plasmid replication during the bacterial division cycle, Biotechnology and Bioengineering 52 (6): 633647, 1996.

[45] D. Noble, Modeling the Heart, Physiology 19: 191-197, 2004.

[46] D. Noble, The Music of Life, Oxford University Press, 2006.

[47] Drug Discovery Today Magazine, Vol (8) 24, December 2003.

[48] D. Ridgway, G. Broderick, M.J. Ellison, Accommodating space, time and randomness in network simulation, Curr. Opin. Biotechnol. Oct;17(5):493-8, 2005.

[49] E.A Groisman, E. Chiao, C.J Lipps, F. Heffron, Salmonella typhimurium phoP virulence gene is a transcriptional regulator, Proc Natl Acad Sci U S A, 86(18):70777081, 1989.

[50] E.A van Doorn, A.I Zeifman, Birth-death processes with killing, Statistics & Probability Letters, 72 (1). pp. 33-42. ISSN 0167-7152, 2005.

[51] E.A van Doorn, A.I Zeifman, Extinction probability in a birth-death process with killing, Journal of Applied Probability, 42 (1). pp. 185-198. ISSN 0021-9002, 2005.

[52] EcoCyc Database, http://www.ecocyc.org.

[53] E.A Abbondanzieri, W. J. Greenleaf, J. W. Shaevitz, R. Landick, and S. M. Block, Direct observation of base-pair stepping by RNA polymerase. Nature 438, 460-465, 2005.

[54] E.H Davidson, The Regulatory Genome: Gene Regulatory Networks In Development And Evolution, Academic Press, 2006.

[55] E. Selkov, M. Galimova,et.al, The metabolic pathway collection: an update. Nucleic Acids Res. 25, 37-38, 1997.

[56] E. M Sozbudak, M. Thattai, I. Kurtser, A. Grossman, and A. van Oudenaarden, Regulation of noise in the expression of a single gene, Nature Genetics 31, 69-73, 2002.

[57] E.P Gianchandani, J.A Papin, N.D Price, A.R Joyce, B.O Palsson, Matrix formalism to describe functional states of transcriptional regulatory systems, PLoS Comput Biol 2, 2006.

[58] F. Cellier, Continuous System Modeling, Springer Verlag, 1991.

[59] F. Gros, H. Hiatt, W. Gilbert, C. Kurland, R. Risebrough, and J. Watson, Unstable ribonucleic acid revealed by pulse labeling of Escherichia coli, Nature 190, 581-585, 1961.

[60] J. Forster , I. Famili, B.O Palsson, J. Nielsen, Large-scale evaluation of in silico gene deletions in Saccharomyces cerevisiae, Omics 7:193202, 2003.

[61] G. Booch, Object-Oriented Analysis and Design with Applications, Addison-Wesley, 1993.

[62] G. Vscovi, F.C Soncini, E.A Groisman, The role of the PhoP/PhoQ regulon in Salmonella virulence, Res Microbiol. 145(5-6):473480, 1994.

[63] Gross and Harris, Fundamentals of Queueing Theory, 3rd Ed, John Wiley & Sons, Inc., ISBN 0-471-17083-6, 1998.

[64] G.K Ackers, A. D. Johnson, and M. A. Shea, Quantitative model for gene regulation by lambda phage repressor. Proceedings of the National Academy of Sciences USA 79 (4), 1129-1133, 2002.

[65] H. Casanova, F. Berman, T. Bartol, E. Gokcay, T. Sejnowski, A. Birnbaum, J. Dongarra, M. Miller, M. Ellisman, M. Faerman, G. Obertelli, R. Wolski, S. Pomerantz, J. Stiles, The Virtual Instrument: Support for Grid-Enabled MCell Simulations, Intl. J. of High Perf. Comp. App. 18:3-17, 2004.

[66] H. de Jong, Modeling and Simulation of genetic regulatory systems: A literature review, J Comput Biol. 9(1):67-103, 2002.

[67] H. Kitano, Systems Biology: A Brief Overview, vol. 295. no. 5560, pp. 1662 - 1664, Science 2002.

[68] H. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, A. Zeng, An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs, Nucleic Acids Research 2004 32(22), 2002.

[69] H. Ishizuka, A. Hanamura, T. Kunimura, H. Aiba, A lowered concentration of cAMP receptor protein caused by glucose is an important determinant for catabolite repression in Escherichia coli. Mol Microbiol, 10, 341350, 1993.

[70] H.M. Sauro, Jarnac: a system for interactive metabolic analysis. Animating the Cellular Map, 9th International BioThermoKinetics Meeting (eds: Hofmeyr, JH. S, Rohwer, J. M, Snoep J. L) Stellenbosch University Press, Ch. 33, pp.221-228, 2000.

[71] H.M Sauro, Scamp: A general-purpose simulator and metabolic control analysis program, Comput. Applications Biosci., 9 441-450, 1993.

[72] H. McAdams and A. Arkins, Stochastic mechanisms in gene expression, PNAS, 1997.

[73] H. Matsuno, A. Doi, M. Nagasaki, S. Miyano, Hybrid Petri net representation of gene regulatory networks, Pacific Symposium on Biocomputing 5, 341-352, 2000. http://www.genomicobject.net/.

[74] H. Ishizuka, A. Hanamura, T. Inada, and H. Aiba, Mechanism of the down-regulation of cAMP receptor protein by glucose in Escherichia coli: role of autoregulation of the crp gene., EMBO J. 1994 July 1; 13(13): 30773082, 2000.

[75] I. Moll, S. Grill, C. O. Gualerzi, and U. Blasi, Leaderless mRNA in bacteria: urprises in ribosomal recruitment and translational control, Molecular Microbiology 43 (1), 239-246, 2002.

[76] I.R. Epstein, J.A. Pojman, An introduction to nonlinear chemical dynamics: oscillations, waves, patterns, and chaos, Oxford University Press, Oxford, 1998.

[77] I. Golding, J. Paulsson, S. Zawilski, E. Cox, Real-time kinetics of gene activity in individual bacteria, Cell 123 (6), 1025-1036, 2005

[78] J. Gowrishankar, R. Harinarayanan, Why is transcription coupled to translation in bacteria?, Molecular Microbiology 54 (3), 598-603, 2004.

[79] J.M Raser, J. M. & E.K O'Shea, Noise in Gene Expression: Origins, Consequences, and Control, Science 309, 2010-2013, 2005.

[80] J. T. Mettetal, D. Muzzey, J. M. Pedraza, E. M. Ozbudak, and A. van Oudenaarden, Predicting stochastic gene expression dynamics in single cells, PNAS 103, 7304, 2006.

[81] Java 1.5 Platform, www.java.sun.com.

[82] J. Banks, J. Carson, B. Nelson and D. Nicol, Discrete-event system simulation, fourth edition, Pearson, 2005.

[83] J. Green, J. R. Guest, Activation of FNR-dependent transcription by iron: an in vitro switch for FNR. FEMS Microbiol Lett 113: 219-222, 1992.

[84] J. Gosling, B. Joy, G. Steele, and G. Bracha, The Java language specification, third edition. Addison-Wesley, 2005.

[85] J. Hattne, D. Fange and J. Elf, Stochastic reaction-diffusion simulation with MesoRD, Bioinfomatics, 2005.

[86] J. Himmelspach, A. Uhrmacher, A component based simulation layer for JAMES, In Proc. of the 18th Workshop on Parallel and Distributed Simulation (PADS), Kufstein, Austria, 115122, 2002.

[87] J. L. Snoep and H.V Westerhoff, From isolation to integration, a systems biology approach for building the Silicon Cell, Systems Biology: Definitions and Perspectives, Springer-Verlag, 2005.

[88] J. M. Martin, Parallel discrete event simulation of large scale wireless ad-hoc networks, University of California, Los Angeles, 2002.

[89] J.M. Bower, H. Bolouri (Eds.), Computational Modeling of Genetic and Biochemical Networks, MIT Press, 2001.

[90] J.S Edwards, B.O Palsson, The Escherichia coli MG1655 in silico metabolic genotype: its definition, characteristics, and capabilities, Proc Natl Acad Sci USA 97: 55285533, 2002.

[91] J. Plumbridge, Control of the expression of the manXYZ operon in Escherichia coli: Mlc is a negative regulator of the mannose PTS, Mol Microbiol, 27,369380, 1998.

[92] J. Paulsson, Models of Stochastic Gene Expression, Phys. Life Rev. 2, 157-75, 2005.

[93] J. Yu, et.al, Probing Gene Expression in Live Cells, One Protein Molecule at a Time, Science, March 2006.

[94] K. Burrage, T.Tian, P. Burrage, A multi-scaled approach for simulating chemical reaction systems, Prog. Biophys. Mol. Biol. 85,217234, 2004.

[95] K. Kimata, H. Takahashi, T. Inada, P. Postma, H. Aiba, cAMP receptor protein-cAMP plays a crucial role in glucoselactose diauxie by activating the major glucose transporter gene in Escherichia coli. Proc Natl Acad Sci USA, 94, 1291412919, 1997.

[96] KEGG Encyclopedia Database, http://www.genome.jp/kegg/

[97] K.R Heidtke, S. Schulze-Kremer, Biosim - a new qualitative simulation environment for molecular biology. Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB98); Montreal, Canada, 1998; pp 85-94.

[98] L. Cai, et.al, Stochastic protein expression in individual cells at the single molecule level, Nature Letters, 2006.

[99] L. Lok, The need for speed in stochastic simulation, Nat. Biotechnol., 22, 964965, 2000.

[100] L.M Hsu, Promoter clearance and escape in prokaryotes, Biochimica et Biophysica Acta 1577, 191-207. 2002.

[101] L. Shlomi, Yariv Eisenberg, Roded Sharan & Eytan Ruppin, A genome-scale computational study of the interplay between transcriptional regulation and metabolism, Molecular Systems Biology 3 Article number: 101, 2007.

[102] M.A. Gibson, J. Bruck, Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels, Journal of Physical Chemistry A 104 (9):18761889, 2000.

[103] M. Calder, S. Gilmore and J. Hillston, Modelling the influence of RKIP on the ERK signalling pathway using the stochastic process algebra PEPA, Transactions on Computational Systems Biology VII, vol. 4230, pp. 1-23, Springer, 2006.

[104] M. Ehlde, G. Zacchi, G. Mist, A user-friendly metabolic simulator, Comput. Applications Biosci. 1995, 11, 201-207.

[105] M. S. Samoilov & A. P. Arkin, Deviant effects in molecular reaction pathways, Nature Biotechnology - 24, 1235 - 1240 (2006).

[106] M. Kaern, T.C Elston, J.W. Blake, J.J Collins, Stochasticity in gene expression: from theories to phenotypes, Nat. Rev. Genet. 6, 451-464, 2005.

[107] M. Grunberg-Manago, Messenger RNA stability and its role in control of gene expression in bacteria and phages, Annual Reviews Genetics 33, 193-227, 1999.

[108] M. Calder, S. Gilmore and J. Hillston, Automatically deriving ODEs from process algebra models of signalling pathways, Proceedings of CMSB (Computational Methods in Systems Biology) 2005.

[109] M. Elowitz, A. Levine, E. Siggia & P. Swain, Stochastic gene expression in a single cell, Science 297, 1183-1186 (2002).

[110] M. Ginkel, A. Kremling, T.Nutsch, R. Rehner, E.D. Gilles, Modular modeling of cellular systems with ProMoT/Diva, Bioinformatics, 19, pp. 1169-1176, 2003.

[111] M.M Babu, N.M Luscombe, L. Aravind, M. Gerstein, S.A Teichmann, Structure and evolution of transcriptional regulatory networks, Curr. Opin. Struct. Biol., 14, 283291, 2004.

[112] M. S. Samoilov & A. P. Arkin, Deviant effects in molecular reaction pathways, Nature Biotechnology - 24, 1235 - 1240 (2006).

[113] M. S. Dasika, A. Gupta and Costas D. Maranas, DEMSIM: a discrete event based mechanistic simulation platform for gene expression and regulation dynamics, Journal of Theoretical Biology, Volume 232, Issue 1, 7 January 2005, Pages 55-69.

[114] M.S Dasika, A. Gupta, C.D Maranas, A mixed integer linear programming (MILP) framework for inferring time delay in gene regulatory networks. Pac. Symp. Biocomput. 9, 474485, 2004.

[115] M. Tomita et.al, E-Cell: Software environment for whole-cell simulation . Bioinformatics, 15(1), pp. 72-84, 1999.

[116] M. Savageau and F.C Neidhart, Regulation beyond the operon. in Escherichia coli and Salmonella: Cellular and Molecular Biology (ed. Neidhart, F.C.) 13101324, American Society for Microbiology, Washington D.C., 1996.

[117] M. W. Covert, C.H Schilling, B. Palsson, Regulation of gene expression in flux balance models of metabolism, J. Theor. Biol 213: 7388, 2001.

[118] M.W. Covert, B.O. Palsson, Constraints-based models: regulation of gene expression reduces the steady-state solution space, J. Theor. Biol 221: 309325, 2003.

[119] M.W Covert, E.M Knight, J. L Reed, M.J Herrgard, B.O Palsson, Integrating high-throughput and computational data elucidates bacterial networks, Nature 429: 9296, 2004.

[120] M. W. Covert and B O. Palsson, Transcriptional regulation in constraints-based metabolic models of Escherichia coli, J. Biol. Chem, 10.1074 2002.

[121] N.A. Buchmeier, F. Heffron, Induction of Salmonella stress proteins upon infection of macrophages, Science 11;248(4956):730732, 1990.

[122] N.D. Price, J.L. Reed, B.O. Palsson, Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol 2: 886897, 2004.

[123] N. G. van Kampen, Stochastic Processes in Physics and Chemistry, 1992.

[124] N.J Guido, X. Wang, D. Adalsteinsson, D. McGillen, J. Hasty, C. R. Cantor, T. C. Elston, and J. Collins, A bottom-up approach to gene regulation, Nature 439 (16), 856-860, 2006.

[125] N.V Fedoroff, W. Fontana, Small numbers of big molecules, Science 297, 1129-1131, 2002.

[126] N. Weiner, Cybernetics and Communication in the Animal and the Machine, MIT Press, Cambride, MA, 1948.

[127] N. Le Novre & T.S Shimizu, StochSim: modeling of stochastic biomolecular processes , Bioinformatics 17, pp.575-576, 2002.

[128] N. Kam, N.A. Katz, A. Korman, A., D.Peleg, The Immune System as a Reactive System: Modeling T-Cell Activation with StateCharts, Technical Report MCS01-09, Mathematics & Computer Science, Weizmann Institute of Science, 2001.

[129] P. A. Cotter, R.P Gunsalus, Contribution of fnr and arcA gene products in coordinate regulation of cytochrome o and d oxidase (cyoABCDE and cydAB) genes in Escherichia coli. FEMS Microbiol Lett 91: 31-36, 1992.

[130] P. Haas, Stochastic petri nets: Modelling, stability, simulation, Springer Verlag, 2002.

[131] P.J Goss, J. Peccoud, Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets, Proc Natl Acad Sci U S A. 95(12):6750-5, 2000.

[132] P. Ghosh, S. Ghosh, K. Basu and S.K. Das, A Diffusion Model to Estimate the Interarrival Time of Charged Molecules in Stochastic Event based Modeling of Complex Biological Networks, CSB Workshops, pp.19-22, 2005.

[133] P. Ghosh, S. Ghosh, K. Basu and S.K. Das, S. Daefler, An Analytical Model to Estimate the time taken for Cytoplasmic Reactions for Stochastic Simulation of Complex Biological Systems, IEEE Granular Computing Conference, 2006.

[134] P. Ghosh, S. Ghosh, K. Basu, S. Das, S. Daefler, Estimation of the holding Time for Cytoplasmic Reactions in Stochastic Event Based Modeling of Complex Biological Networks, TR CSE-2005-09, 2005.

[135] P. Ghosh, S. Ghosh, K. Basu, S. Das, S. Daefler, Transient analysis of Diffusion for charged molecules to model the input process in a stochastic event based simulation framework for the PhoPQ signal transduction system, TR CSE-2005-07, 2005.

[136] P. Ghosh, S. Ghosh, K. Basu, S.K Das , A Computationally Fast and Parametric Model to estimate Protein-Ligand Docking time for Stochastic Event based Simulation, The Transactions on Computational and Systems Biology, 2007.

[137] P. Ghosh, S. Ghosh, K. Basu, S.K Das, Parametric modeling of protein-DNA binding kinetics: A Discrete Event based Simulation approach, second round of revision at the Elsevier Journal on Discrete Applied Mathematics (DAM), 2007.

[138] P. Ghosh, S. Ghosh, K. Basu, S.K Das, A Markov Model-based Analysis of Stochastic Biochemical Systems, Computational Systems Bioinformatics Conference, 2007.

[139] P. Ghosh, S. Ghosh, K. Basu, S.K Das, Modeling protein-DNA binding time in Stochastic Discrete Event Simulation of Biological Processes, IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology (CIBCB '07), pp. 439-446, 2007.

[140] P.P Dennis, M. Ehrenberg, H. Bremer, Control of rRNA synthesis in Escherichia coli: a systems biology approach. Microbiology and Molecular Biology Reviews 68 (4), 639-668, 2004.

[141] P.Mendes, GEPASI: A software package for modeling the dynamics, steady states and control of biochemical and other systems , Comput. Applic. Biosci. 9, pp.563-57, 1993.

[142] P.P Dennis, M. Ehrenberg, H. Bremer, Control of rRNA synthesis in Escherichia coli: a systems biology approach. Microbiology and Molecular Biology Reviews 68 (4), 639-668, 2004.

[143] P.J Goss, J. Peccoud, Quantitative modeling of stochastic systems in molecular biology by using stochastic Petri nets, Proceedings of the National Academy of Sciences USA 95 (12), 6750-6755, 1998.

[144] P.S. Swain, M.B. Elowitz,E.D. Siggia, Intrinsic and extrinsic contributions to stochasticity in gene expression, PNAS 99 2002.

[145] PubMed Cetral, http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed.

[146] R. Bundschuh, F. Hayot, and C. Jayaprakash, The role of dimerization in noise reduction of simple genetic networks, Journal of Theoretical Biology 220, 261-269, 2002.

[147] R. Heinrich, S. Schuster, The Regulation of Cellular Systems, Chapman & Hall, New York, 1996.

[148] R. Blossey, L. Cardelli, A. Phillips, A compositional approach to the stochastic dynamics of gene networks, Transactions on Computational Systems Biology IV, 99-122. LNCS vol 3939, 2006.

[149] R. Kannan, P. Tetali & S. Vempala, Simple Markov-chain algorithms for generating bipartite graphs and tournaments. Random Structures and Algorithms 14, 293308, 1999.

[150] R. Mahadevan, C.H Schilling, The effects of alternate optimal solutions in constraint-based genome-scale metabolic models, Metab. Eng. 5: 264276, 2003.

[151] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Network motifs: simple building blocks of complex networks. Science, 298, 824827, 2002.

[152] S. Becker, G. Holighaus, T. Gabrielczyk, and G. Unden, O2 as the regulatory signal for FNR-dependent gene regulation in Escherichia coli, J Bacteriol. August; 178(15): 45154521, 1996.

[153] S. Brenner, F. Jacob, and M. Meselson, An unstable intermediate carrying information from genes to ribosomes for protein synthesis, Nature 190, 576-581, 1961.

[154] S. Efroni, D. Harel, and I. Cohen, Towards rigorous comprehension of biological complexity: Modeling, execution and visualization of thymic t cell maturation, Genome Research (13): 24852497, 2000.

[155] S. Efroni, D. Harel, and I. Cohen, Reactive animation: Realistic modeling of complex dynamic systems, IEEE Computer magazine, vol. 38, no. 1, pp. 38-47, 2005.

[156] S. Karlin, J.L McGregor, A characterization of birth and death processes, Proc. Natl. Acad. Sci. USA 45, 375-379, 1959.

[157] S. Ghosh, P. Ghosh, K. Basu, S. Das and S. Daefler, iSimBioSys: A Discrete Event Simulation Platform for 'in silico' Study of Biological Systems , 39th Annual Simulation Symposium, pp.204-213, 2006.

[158] S. Raczynski, When system dynamics ode models fail. Simulation 1996, 67, 343-349.

[159] S. Michelson, M. Cole, The Future of Predictive Biosimulation in Drug Discovery, Expert Opinion on Drug Discovery, 2007.

[160] S. Park, C. A. Hunt, G. E.P Ropella, PISL: A Large-Scale In Silico Experimental Framework for Agent-Directed Physiological Models, Spring Simulation Multiconference (SpringSim'05), The Society for Modeling and Simulation International, San Diego, CA, April 2-8, 2005.

[161] S. Ramsey, D. Orell, and H. Bolouri, Dizzy: Stochastic simulation of large scale genetic regulatory networks, J. Bioinform. Comput. Biol. Apr 3(2):415-36, 2005.

[162] S.S Shen-Orr, R. Milo, S. Mangan, U. Alon, Network motifs in the transcriptional regulation network of Escherichia coli.,Nat Genet. May;31(1):64-8., 2002.

[163] S.S Levanon, K.Y San, G.N Bennet, Effect of oxygen on the Escherichia coli ArcA and FNR regulation systems and metabolic responses, Biotechnol Bioeng. Mar 5;89(5):556-64, 2005.

[164] S. Greive, P. H. von Hippel, Thinking quantitatively about transcriptional regulation, Nature Reviews Molecular Cell Biology 6, 221, 2005.

[165] S. Becker, G. Holighaus, T. Gabrielczyk, G. Unden, O2 as the regulatory signal for FNR-dependent gene regulation in Escherichia coli. J Bacteriol 178:4515-4521, 1996.

[166] S. Sheikh-Bahaei, G. E. P. Ropella,C. A. Hunt, Agent-Based Simulation of In Vitro Hepatic Drug Metabolism: In Silico Hepatic Intrinsic Clearance, Spring Simulation Multiconference (SpringSim'05), The Society for Modeling and Simulation International, San Diego, CA, April 2-8, 2005.

[167] S. J Park, R.P Gunsalus, Oxygen, iron, carbon, and superoxide control of the fumarase fumA and fum C genes of Escherichia coli: role of arcA, fnr and soxR gene products. J Bacteriol 177: 6255-6262, 1995.

[168] S. Van Volsem, W. Dullaert, H. Van Landeghem, An evolutionary algorithm and discrete event simulation for optimizing inspection strategies for multi-stage processes, Eur J Oper Res, 2006.

[169] SmolDyn, http://genomics.lbl.gov/ sandrews/index.html.

[170] The CCDB Database, http://redpoll.pharmacy.ualberta.ca/CCDB/.

[171] T. Emonet, C.M. Macal, M.J. North, C.E. Wickersham, and P. Cluzel, AgentCell: A Digital Single-Cell Assay for Bacterial Chemotaxis, Bioinformatics, Vol. 21, No. 11, pp. 2714-2721, Oxford University Press, Oxford, UK, 2005.

[172] T.E Ideker, V. Thorsson, R.M Karp, Discovery of regulatory interactions through pertubations:inference and experimental design. Pac. Symp. Biocomput. 5, 302313, 2000.

[173] T. Inada, K. Kimata, H. Aiba, Mechanism responsible for glucoselactose diauxie in *Escherichia coli*: challenge to the cAMP model. Genes Cells, 1, 293301, 1997.

[174] T.C Meng, et al., Modeling and simulation of biological systems with stochasticity, In Silico Biol., 4, 0024, 2004.

[175] T.E Turner, et al., Stochastic approaches for modelling in vivo reactions, Comput. Biol. Chem., 28, 165178, 2004.

[176] The GenBank Datatbase, http://www.psc.edu/general/software/packages/genbank/genbank.

[177] T.M Henkin, Transcription termination control in bacteria, Current Opinion in Microbiology 3 (2), 149-153, 2000.

[178] The Network Simulator (NS-2), http://www.isi.edu/nsnam/ns/.

[179] The Protein DataBank, http://www.rcsb.org/pdb/home/home.do.

[180] T.Tian, K. Burrage, Binomial leap methods for simulating stochastic chemical kinetics, J. Chem. Phys., 121, 1035610364, 2004.

[181] T. Shlomi, O. Berkman, E. Ruppin, Regulatory on/off minimization of metabolic flux changes after genetic perturbations, Proc Natl Acad Sci USA 102: 76957700, 2005.

[182] The Simbology Toolbox, The Mathworks, http://www.mathworks.com.

[183] T. Zhang, R. Rohlfs, and R. Schwartz., Implementation of a discrete event simulator for biological self-assembly systems, Proc. 2005 Winter Simulation Conf. pp.2223-2231, 2005.

[184] U. Sauer, et al., Getting Closer to the Whole Picture, Science, vol. 316. no. 5824, pp. 550 - 551, 2007.

[185] V. K. Rangavajhala, S. Daefler, Modeling the Salmonella PHOPQ Two Component Regualtory System, MS Thesis, The University Of Texas At Arlington, Arlington, TX, 2003.

[186] Virtual Cell Project, http://www.nrcam.uchc.edu/.

[187] W.R McClure, Mechanism and control of transcription initiation in prokaryotes, Annual Review Biochemistry 54, 171-204, 1985.

[188] X.Q. Xia, M.H. Wise, DimSim: A Discrete Event Simulator of Metabolic Networks, Journal of Chemical Information and Computer Sciences, 43(3), 2003.

[189] X. Zeng, R. Bagrodia, M. Gerla, GloMoSim: A Library for Parallel Simulation of Large-Scale Wireless Networks, p. 154, 12th Workshop on Parallel and Distributed Simulation, 1998.

[190] Y. Tanaka, K. Kimata and H. Aiba, A novel regulatory role of glucose transporter of *Escherichia coli*: membrane sequestration of a global repressor Mlc, The EMBO Journal 19, 53445352, 2002.

[191] Z. N. Oltvai and A. L. Barabasi, Life's complexity pyramid, Science 298, 763-764, 2002.

## BIOGRAPHICAL STATEMENT

Samik Ghosh received the Bachelor of Technology (Honors) degree in Computer Science and Engineering from Haldia Institute of Technology, Haldia, India, in 2001. After having worked in telecommunication software industry for a year, he began his graduate studies at the Center For Research In Wireless Mobility and Networking (CreW-MaN) Laboratory, Department of Computer Science and Engineering at The University of Texas at Arlington in Fall 2002. He received his Master of Science degree in Computer Science and Engineering from The University of Texas at Arlington in May 2004 before embarking on his doctoral studies. His research interest is primarily focused on modeling and simulation of complex biological networks and his doctoral dissertation involved developing a discrete event based hybrid simulation framework for modeling the dynamics of gene regulatory and metabolic networks in bacterial cells. His secondary research interests include wireless communications, particularly community wireless mesh networks and mobile software development and analysis.