STOCHASTIC MODELS FOR IN-SILICO EVENT-BASED BIOLOGICAL NETWORK SIMULATION

by

PREETAM 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 would like to thank my supervising professors Sajal K. Das and Kalyan Basu for constantly motivating and encouraging me, and also for their invaluable advice during the course of my doctoral studies. I also wish to thank my committee members Dr. Nikola Stojanovic (CSE), Dr. Lorraine van Waasbergen (Biology) and Dr. Doyle Hawkins (Mathematics) for their interest in my research and for taking time to serve in my dissertation committee.

I would also like to extend my appreciation to Nokia, Nortel Networks, the Dean's office, TxTec funding office and the CSE department at UTA for providing financial support for my graduate studies. I am especially grateful to Dr. Simon Daeffler from Mount Sinai school of medicine for the helpful discussions and invaluable comments. I am grateful to all the teachers who taught me during the years I spent in school, both in India and in the Unites States.

Finally, I would like to express my deep gratitude to my parents and sister who have encouraged and inspired me throughout my life. I am extremely fortunate to be so blessed and thank them for their sacrifice, encouragement and patience. I also thank several of my friends and fellow members of the CReWMaN lab at UTA who have helped me throughout my career.

November 26, 2007

ABSTRACT

STOCHASTIC MODELS FOR IN-SILICO EVENT-BASED BIOLOGICAL NETWORK SIMULATION

PREETAM GHOSH, PhD,

The University of Texas at Arlington, 2007

Supervising Professors: Sajal K. Das, and Kalyan Basu

The multi-scale biological system model is a new research direction to capture the dynamic measurements of complex biological systems. The current statistical thermodynamic models can not scale to this challenge due to the explosion of state-spaces of the system, where a biological organ may have billions of cells, each with millions of molecule types and each type may have a few million molecules. We seek to propose a phenomenological theory that will require a smaller number of state variables to address this multi-scaling problem. Discrete Markov statistical process is used to understand the system dynamics in the networking community for a long time. In this dissertation, we focus more specifically on a composite system by combining the state variables in the time-space domain as events, and determine the immediate dynamics between the events by using statistical analysis or simulation methods. In our approach the space-time behavior of the cell dynamics is captured by discrete state variables, where an *event* is a combined process of a large number of state transitions between a set of state variables. The execution time of these state transitions to manifest the event outcome is a random variable called *event-holding time*. The underlying assumption is that it will be possible

to segregate the complete system state-space into a disjoint set of independent events and events can be executed simultaneously without any interaction once the execution conditions are satisfied (removal of resource bottleneck, collision).

In this dissertation, we present the event-time models for some biological functions that will be incorporated in the discrete-event based stochastic simulator. In particular, we present analytical models for the molecular transport event in cells considering charged/non-charged macromolecules. We show, that molecular transport event completion time can be approximated by an exponential distribution. Next we present stochastic models for biochemical reactions in the cell (that can be extended to reactions occurring in the cell cytoplasm, membrane or nucleus). We show that the reaction completion time follows an exponential distribution when one of the reactant molecules enter the cell one at a time, whereas, it follows a gamma distribution when a batch of the reactant molecules enter the cell. We also present stochastic models for the protein-DNA binding and protein-ligand docking events and show that both these events have an exponentially distributed event completion time. We also validate each of the models presented in the dissertation with experimental findings reported in the literature. Finally, we present a markov chain based stochastic biochemical system simulator which can give us the dynamics of more complex events and can be used to improve the scalability of the discrete-event based stochastic simulator. We propose to successfully demonstrate this technique by modeling the complete dynamics of one Salmonella cell.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

During the last decade, the advancement in high-throughput biological experiments has generated a large amount of empirical data on the molecular foundations of biological structures and functions. Complete genomic sequencing of new organisms are being completed and advanced databases storing comprehensive annotations of genomic and protein structures are being developed rapidly. As more and more data become available, biologists are now looking beyond assigning functions to individual genes. Although the functional and structural properties of individual genes and proteins have been studied and characterized, the understanding of their complex interactions in a cell through a set of pathways that create the intelligence of the organisms is still very limited. The complexity of this exercise increases manifold as we move into higher scales: interaction of large ensemble of cells in a tissue or interaction of tissues in continuum for rhythmic pumping of the heart, for example. The new research challenge [1] is to develop a comprehensive modeling framework that integrates molecular and genetic data for a quantitative understanding of physiology and behavior of biological processes at multiple scales – starting from the cell, to the tissues and finally to the whole organism.

Currently, there exist comprehensive models in mathematical physiology [48] and computational cell biology [21] that provide limited understanding on multi scale biological processes. Such models (e.g., the Hodgkins-Huxley equation) work very well in specific problem domains like cell membrane current and conductance. Alongside, researchers from diverse disciplines have focused on developing models to capture the dynamics of biological processes [97, 28]. These spatio-temporal models can be classified

into five categories: (a) quantum mechanics, (b) molecular dynamics, (c) mesoscale dynamics, (d) cellular/organ-level stochastic simulation and (e) rule based model. The first two models are limited in scope, as they cannot handle the complexity of an entire cell. For example, the quantum mechanics based model captures the random environment of the cell at electron level and is very useful to understand the structure of the macromolecules. But because of computational overhead, it can only handle about 1000 atoms. Similarly, the molecular dynamics model uses force field methods based on Newtonian mechanics and is the right tool to understand the function of the macromolecules. This model, for example, is used to study the binding site configurations for protein-protein or protein-DNA interactions and protein folding. Currently, it can handle about 1 million molecules and is not sufficient to model a cell or complex pathways.

The next two models have focused on a narrow range of biological components such as the wave model [97] for ventricular fibrillation in human heart, neural network signaling model [28] to control the onset of sleep in human, or simulation frameworks like E-Cell [61] and Virtual Cell [54] for biological processes. Mesoscale models deal with rate equation based kinetic models and uses continuous time techniques. The rate constants derived from measurements are the most important biological parameters used in this model to represent the biological functions. Experimental measurement of rate constants, in reality, hides all the structural and functional complexities of the corresponding biological function. In addition, biologists often have difficulty to get all the rate constants from valid experiments. The model solves a set of differential equations corresponding to chemical reactions of the pathways with the help of numerical integration. Since a biological system involves a very large number of such differential equations, the model is computationally limited.

Recent experimental measurements at molecular level [40] identified the stochastic nature of the reaction, specially for protein synthesis. This stochastic behavior is further

modulated by the positive and negative feedback loops that exist in biological pathways. The complex interaction of these factors create the stochastic resonance [47, 85] in a biological system. To accommodate this, Gillespie [29] extended the rate based model to a stochastic simulation framework that led to a few other variations such as Cell Designer [39], BioSpice [25], Cell Illustrator [57] etc. The computational overhead of this simulation forced the use of approximation techniques by sacrificing accuracy e.g. the Tau Leap algorithm [30, 58]. Gillespie's technique considers the biochemical system as a discrete Markov process. The limitations of this technique are:

- It assumes that a biological system only consists of different biochemical reactions. Hence, each reaction event is abstracted by the experimentally determined rate constant. the model cannot capture the pertinent details of that biological event. For example, ideally a bimolecular reaction event should incorporate some details of the reactant molecules (e.g., kinetic parameters and size of the molecules).
- The Gillespie technique considers each reaction event completion time to be exponentially distributed with means determined from the kinetic parameters which is not always the case. Depending on the concentration of the reactants, the reaction event completion times might follow different distributions which play an important role in studying the system dynamics (specially for low number of reactant molecules in the system where the stochastic effects are more pronounced).

The system has to be broken down to the reaction level to present the dynamics. This will result in an increase in the complexity and explosion in the number of equations. Due to the large number of protein complexes in a cell, the existing stochastic simulation models lead to a combinatorial explosion in the number of reactions, thus making them unmanageable for complex metabolic and signaling pathway problems.

Finally, the rule based simulation [59] is a new technique to model the complex multi cell interaction at a molecular level and addresses the more complex host-pathogen

interactions. It ignores the stochastic nature of biological functions and considers a set of rules derived from pathways. After reviewing the current status of the modeling methods and their challenge, and also to accommodate the new found stochastic behavior of biological processes, we are motivated to look to an alternative approach to address the problem. We plan to convert the biological process as a stochastic network and solve it as a stochastic network simulation or analysis problem.

1.1 Some preliminary concepts

Here we define some preliminary concepts that will aid in a better understanding of this thesis.

- Cell: The cell is the structural and functional unit of all known living organisms. It is the smallest unit of an organism that is classified as living, and is sometimes called the building block of life. Some organisms, such as bacteria, are unicellular (consist of a single cell). Other organisms, such as humans, are multicellular. (Humans have an estimated 10^{14} cells; a typical cell size is 10 m; a typical cell mass is 1 nanogram.)
- Prokaryotes: Prokaryotes (illustrated in Fig 1.1) are a group of organisms that lack a cell nucleus, or any other membrane-bound organelles. Most are unicellular, but some prokaryotes are multicellular organisms. The prokaryotes are divided into two domains: the Bacteria (e.g., *Salmonella*, *E. coli*) and the Archaea.
- Eukaryotes (illustrated in Fig 1.2): Animals, plants, fungi, and protists are eukaryotic organisms whose cells are organized into complex structures by internal membranes and a cytoskeleton. The most characteristic membrane-bound structure is the nucleus. In the nucleus, the genetic material, DNA, is arranged in chromosomes. Many eukaryotic cells also contain membrane-bound organelles such as mitochondria, chloroplasts and Golgi bodies. The subcellular components labelled

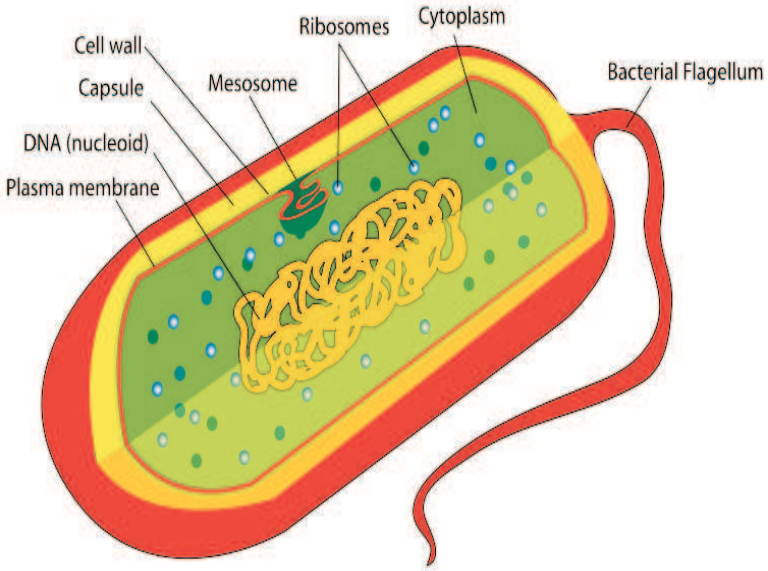

Figure 1.1. Diagram of a typical prokaryotic cell (from Wikipedia).

in Fig 1.2 are as follows: (1) nucleolus (2) nucleus (3) ribosome (4) vesicle (5) rough endoplasmic reticulum (ER) (6) Golgi apparatus (7) Cytoskeleton (8) smooth ER (9) mitochondria (10) vacuole (11) cytoplasm (12) lysosome (13) centrioles within centrosome.

- Discrete-event simulation: In discrete event simulation, the operation of a system is represented as a chronological sequence of events. Each event occurs at an instant in time and marks a change of state in the system. In addition to the representation of system state variables and the logic of what happens when system events occur, discrete event simulations include the following: (a) Clock: The simulation must keep track of the current simulation time, in whatever measurement units are suitable for the system being modeled. In discrete-event simulations, as opposed to real time simulations, time hops because events are instantaneous the clock skips to the next event start time as the simulation proceeds. (b) Events List: The simulation maintains at least one list of simulation events. An event must have a start time,

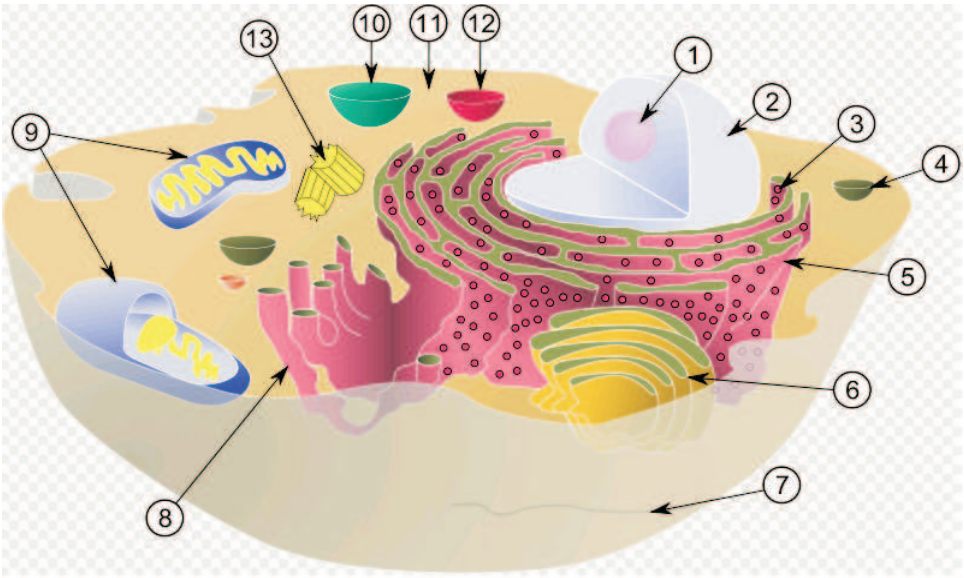

Figure 1.2. Diagram of a typical eukaryotic cell (from Wikipedia).

some kind of code that constitutes the performance of the event itself, and possibly an end time. In some approaches, there are separate lists for current and future events. Events in their lists are sorted by event start time. Typically, events are bootstrapped that is, they are scheduled dynamically as the simulation proceeds. (c) Random Number Generators: The simulation needs to generate random variables of various kinds, depending on the system model. This is accomplished by one or more pseudorandom number generators. (d) Statistics: The simulation typically keeps track of the system's statistics, which quantify the aspects of interest. (e) Ending Condition: Because events are bootstrapped, theoretically a discrete-event simulation could run forever. So the simulation designer must decide when the simulation will end. Typical choices are "at time t" or "after processing n number of events" or, more generally, "when statistical measure X reaches the value x".

- Collision theory: Collision theory qualitatively explains how chemical reactions occur and why reaction rates differ for different reactions. It assumes that for a

reaction to occur the reactant particles must collide, but only a certain fraction of the total collisions, the effective collisions, cause the transformation of reactant molecules into products. This is due to the fact that only a fraction of the molecules have sufficient energy and the right orientation at the moment of impact to break the existing bonds and form new bonds. The minimal amount of energy needed so that the molecule is transformed is called activation energy. Collision theory is closely related to chemical kinetics.

- Markov chain: Markov chain is a discrete-time stochastic process with the Markov property. Having the Markov property means the next state solely depends on the present state and doesn't directly depend on the previous states. At each point in time, the system may have changed states from the state the system was in the moment before, or the system may have stayed in the same state. The changes of state are called transitions. If a sequence of states has the Markov property, then every future state is conditionally independent of every prior state.
- Bernoulli trial process: In probability and statistics, a Bernoulli process is a discrete-time stochastic process consisting of a sequence of independent random variables taking values over two symbols. Prosaically, a Bernoulli process is coin flipping, possibly with an unfair coin. A variable in such a sequence may be called a Bernoulli variable. In this thesis, we use this concept for computing the time taken for some biological events.
- First and second moments of event holding time: In the stochastic models for the different biological events, we estimate the time taken to complete these events. We consider the time as a random variable and determine its distribution by providing the expressions for the first and second moments of this random variable. In the reaction model (shown later) we also compute the *adjusted time of reaction* which considers the changes in probability of the reaction event during a single reaction

process. As our stochastic reaction model is discretized in time, the first and second moments for the reaction completion time does not incorporate these changes in probability due to competing reactions, and we show that this approximation does not significantly affect the accuracy of the reaction models.

- Chemical master equation (CME): A master equation is a set of first-order differential equations describing the time evolution of the probability of a system to occupy each one of a discrete set of states. Many physical problems in classical, quantum mechanics and problems in other sciences, can be reduced to the form of a master equation, thereby performing a great simplification of the problem. The CME refers to the master equation governing the time evolution of the probability of a system of biochemical reactions.
- Gillespie simulation: The Gillespie algorithm generates a statistically correct trajectory (possible solution) of a stochastic equation. It was developed to simulate chemical or biochemical systems of reactions efficiently and accurately using limited computational power. As computers have become faster, the algorithm has been used to simulate increasingly complex systems. The algorithm is particularly useful for simulating reactions within cells where the number of reagents typically number in the tens of molecules (or less). Mathematically, it is a variety of a dynamic Monte Carlo method and similar to the kinetic Monte Carlo methods. It is used heavily in computational systems biology.
- Molecular dynamic simulation: Molecular dynamics (MD) is a form of computer simulation wherein atoms and molecules are allowed to interact for a period of time under known laws of physics, giving a view of the motion of the atoms. Because molecular systems generally consist of a vast number of particles, it is impossible to find the properties of such complex systems analytically; MD simulation circumvents this problem by using numerical methods. It represents an interface

between laboratory experiments and theory, and can be understood as a "virtual experiment".

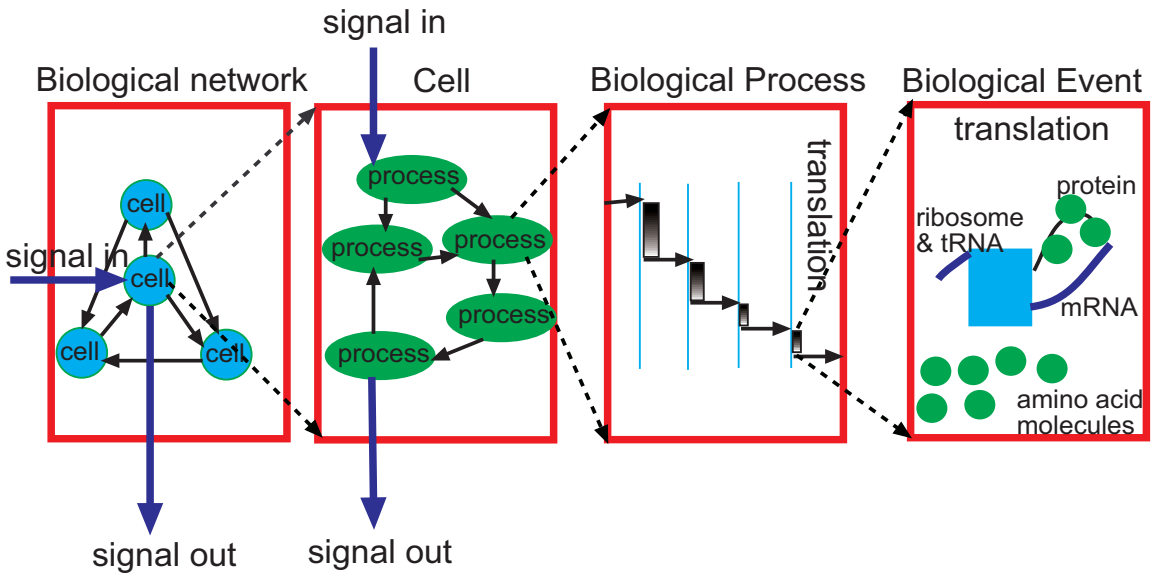

Figure 1.3. Overview of proposed Modeling Concept.

1.2 Stochastic event-based simulation approach

The concept of "in silico" [87] or discrete event based modeling has been successfully applied to study many complex networks and systems. Our main objective is to use this technique to (a) explicitly study the stochastic nature of the biological system, (b) reduce computational complexity of the system so that we can model complete cell dynamics, and (c) use as much as possible the biological knowledge in the modeling, so that all biological complexities are not hidden behind the value of the rate constant measurements. To achieve that, we need mathematical models that are computationally fast and generic in nature, so that by changing the species specific information, the models can be used for other similar species. This stochastic discrete event based framework for complex

biological systems [88, 89] can be easily used for modeling the dynamics of pathways in a single cell, and multiple cell interactions.

In our terminology, a biological network is a collection of biological processes, each comprising a number of functions, where a function will be modeled as an event. A unique pathway will be defined as a biological process consisting of a number of biological functions cascaded through the signal trafficking mechanism. The fundamental entity in our proposed simulation model is an "event" which represents a biological function with relevant boundary conditions. These event models provide the parameters for the temporal displacement of the system states and drive the stochastic discrete-event simulation of the biological system under study. Fig 1.3 illustrates our modeling concept for a biological system. The interactions between cells are captured in the Biological network view. Then for every cell the biological pathways are identified and their relationship is defined. Each pathway is described by the event diagram of the biological process. All these events are modeled by transforming the biological functionality of the event into a stochastic parameter. The main research direction to model the events is to decompose it into a number of biological microevents. Thus we transform the biological function from the thermodynamic and diffusion plane to information plane through a coarse-grained measure of probability. We then use the methods of applied probability to model the temporal and spatial dynamics of the event as a stochastic process of these microevents. Two types of models are required for this method: (1) event execution time, and (2) probability of next event type. Fig 1.4 shows a hypothetical example of the reaction pathway of a biological process with these two types of models (model type-1 and model type-2). A salient feature of this approach is the balance between computational complexity and accuracy of the estimate by including sufficient biological function details. This model allows us to track the important resource counts (typically the various molecules, ions, ribosome-chromosome operon etc involved in the system) in time and space. If the path-

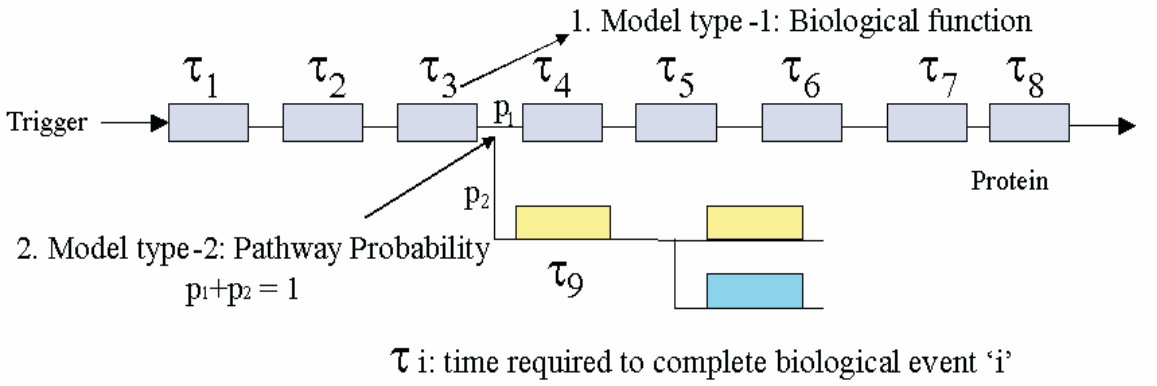

Figure 1.4. Modeling Scheme for Pathway Abstraction.

ways are changed, the logic of the resource usage will also change and the simulation can show the corresponding effect on the system states.

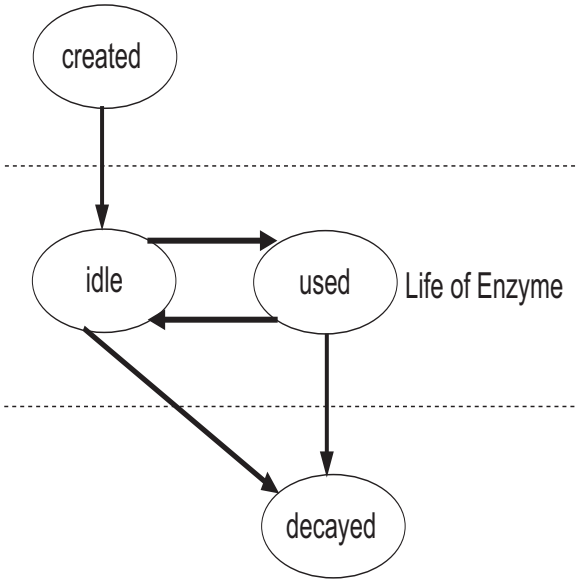

Figure 1.5. State Transition Diagram of an Enzyme during its life cycle.

We identify a biological process as a system of resources (e.g., molecules, ions, ribosome-chromosome operon, tissue, organ, enzyme, etc) that periodically change between one of the following four states based on the resource usage algorithms: (i) 'used'

(e.g, an enzyme is busy in a reaction), (ii) 'idle'(e.g, an enzyme is free to enter a new reaction), (iii) 'created' (e.g, a molecule is created by a reaction) and (iv) 'decayed' (e.g, a molecule is in the process of disintegration at the end of its life-cycle) as shown in Fig 1.5. The state transitions from one state to another are governed by transition flow rates of the process in a cell. The process is initiated by an input signal(s) from the external world to the cell. These input signals initiate a set of events which drive the simulation in time domain resulting in changes in the cell resources with time. For example, the activation of a receptor on a cell starts the pathway to create new proteins or store energy in the ATP or release energy from ATP or breakdown complex sugar molecules to release glucose from the cell to the other tissues. These dynamics can be captured by this process as a time-ordered set of distinct events within a cell. On completion, these events will generate a signal (in this case, may be specific protein or sugar molecules, or ions) that can trigger another pathway in a similar fashion.

The challenges in this approach are:

1. obtain the complete pathway maps that are part of the simulation and identifying the biological discrete events based on system knowledge,
2. identify the set of resources involved in the event and collection of their structural and functional characteristics from different databases,
3. model the event details mathematically to estimate the time taken to complete an event (which is termed in system modeling as the *holding time* of the discrete event),
4. identify the pathway forks and the underlying biological conditions (e.g. location of motif on DNA, hidden motif by chromosome etc) and use this knowledge to model the probabilities for the different fork branches, and
5. create a large "in silico" discrete event simulation framework.

Table 1.1. Comparison of Gillespie Algorithm and Our modeling framework

Stochastic Simulation (Gillespie Algorithm)	Our Discrete event based Simulation	Comments
Initialization: Initialize the number of molecules in the system, reaction 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.
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 molecular event is selected based on the functional logic hardwired in the simulator. 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 we compute them differently.
Update: Increase the time step by the randomly generated time in the Monte Carlo step. 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. The resource count of molecules are updated based on the last event stoichiometry.	We make the temporal progression in discrete time-steps based on the event holding times computed in the previous step.
Iterate: Go back to the Monte Carlo step unless the number of reactants is zero or the simulation time has been exceeded.	Iterate: Go back to the Event modeling step and repeat the process. In case a particular event cannot be executed because of resource conflicts, it is ignored and the simulation proceeds without the update step.	We handle reactions/ events with resource conflicts /shortage differently.

At present, this type of problems are solved by using the Gillespie simulation. It is appropriate at this stage to compare the proposed Discrete event simulation framework with the Gillespie method to explain our contribution. The comparison is schematically presented in Table 1.1.

To illustrate the concept, we present the discrete event modeling of the PhoP/PhoQ two component regulatory system which controls the expression of essential virulence traits in Salmonella Typhimurium depending on the concentration of extra-cellular magnesium [33],[83]. Based on available information, we have developed a functional event diagram (Fig 1.6) of the process that includes both the model types. We identify the list of discrete events that are required for the model based on the available knowledge of the system. In other words, we identify the various types of molecules, cells, tissues etc which are involved in the resource usage algorithm for an event (either in reactions, or as catalysts or as end products). To find the time taken for an event, it is important to identify the parameters which affect the interaction of the resources in a particular

biological discrete event process and mapping them into the time domain (i.e. identifying the time required for completion of the biological event. This generates event time as a function of these parameters).

1.3 Salmonella PhoPQ System Model

In a Salmonella cell, virulence is produced by the PhoPQ two compartment system that is activated by Mg^{2+} concentration change. We identify the key biological functions involved in the PhoPQ regulatory network (from the sensing of Mg^{2+} at the cell membrane to the expression of virulent genes from the DNA in the cytoplasm). The basic schematic block diagram of the processes which we have identified to capture the sequence of actions is presented first. For each process block, we have some input signal(s) coming into the process and output signal(s) which can be considered as the outcome of the process and can trigger one or more processes (or the same process itself in a feedback mechanism). Fig 1.6 captures the basic high-level biological functions involved.

Mg^{2+} receptor Signaling Process: Normally a biological process is defined by a pathway (experimentally determined by biologists) that shows the cascade of biological functions in time. Currently, many pathway databases have been established maintaining this record for different species which we use to understand this process. With the departure of a Mg^{2+} molecule, the PhoQ protein auto-phosphorylates (kinase activity) by making use of an ATP molecule from the cell. The phosphatase activity of PhoQ regulates the phosphotransfer mechanism to phosphorylate the PhoP protein under micromolar Mg^{2+} concentrations, and dephosphorylates the phosphorylated PhoP molecules under millimolar Mg^{2+} concentrations. Generally, Mg^{2+} concentrations higher than 250 mM stimulate the dephosphorylation of phospho-PhoP. Two independent mechanisms of dephosphorylation of phospho-PhoP occur. One involves the reversion of the reaction that takes place to phosphorylate the response regulator, and the other is a specific phospho-

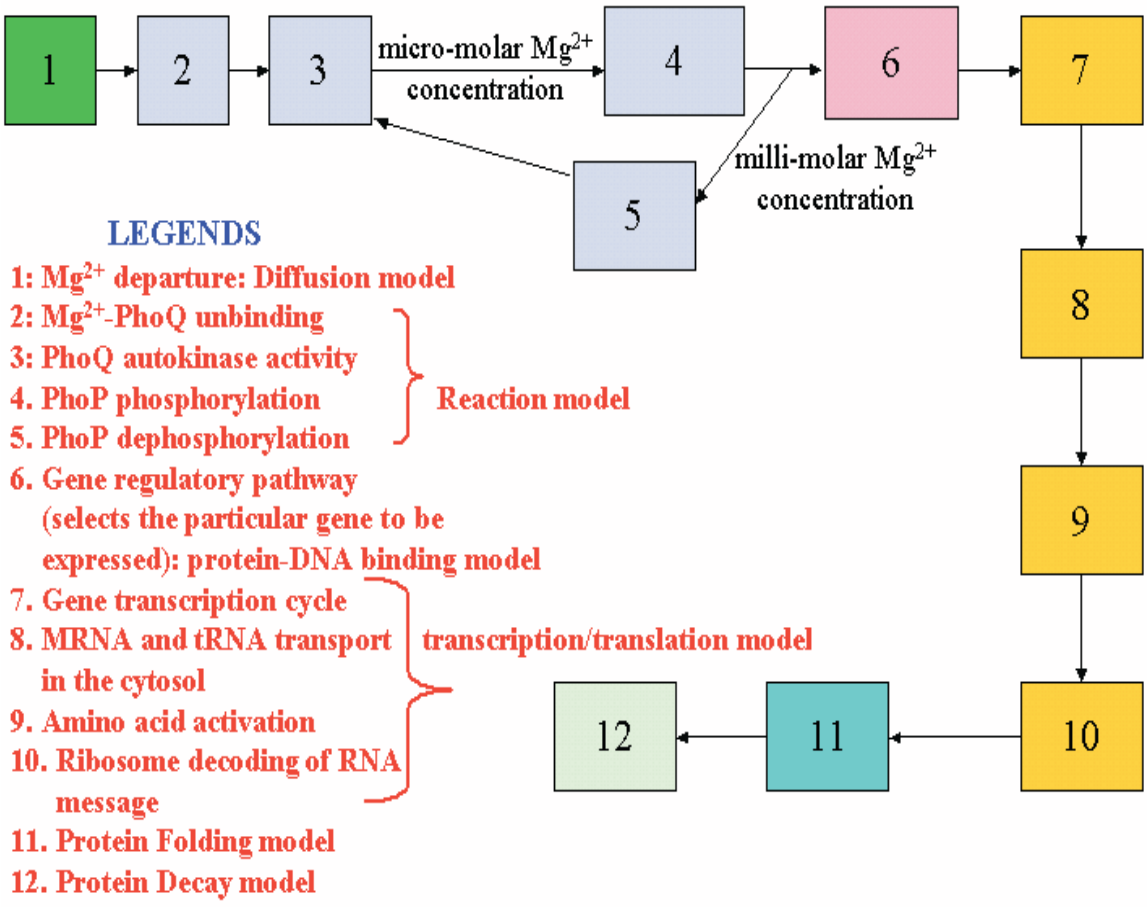

Figure 1.6. Biological Processes involved in the PhoPQ Process in Salmonella.

PhoP phosphatase induced by high concentrations of Mg^{2+} that renders the release of inorganic phosphate.

Thus we can identify the following discrete events from the PhoPQ pathway: with the departure of a Mg^{2+} molecule (event: ion movement from membrane protein), the PhoQ protein autophosphorylates (kinase activity) by making use of an ATP molecule from the cell (event: membrane reaction). The phosphate activity of the PhoQ regulates the phosphotransfer mechanism to phosphorylate the PhoP protein under micro molar Mg^{2+} concentrations, and dephosphorylates the phosphorylated PhoP molecules under millimolar Mg^{2+} concentrations (event: cytoplasmic reaction). The Phospho PhoP

(phoPp) activates the promoter loci and there is only one activation per phoPp. The loci are obtained from the determination of regulatory pathway. PhoPp binding to DNA site is required for transcription (event: DNA protein binding). RNA polymerases are involved in the process of transcription (event: cytoplasmic multi molecule reaction). We also need to consider translation (including steps such as binding of polymerases, regulatory factors, subunits etc) and transport processes.

We identify different biological functions by following this research process to complete the type of event models necessary for discrete event simulation. Each of these models are required to estimate their event time characteristics. The models for cytoplasmic reactions [70, 71, 76], DNA-protein binding [72, 78], protein-ligand docking [73, 77] and molecular transport [74, 75] are reported in this thesis. Based on these models and the protein synthesis model [90] we can complete the simulation of the PhoPQ system.

1.4 Our contributions

Our contributions in this dissertation can be summarized as follows:

- We present an analytical model for the molecular transport biological function driven by concentration and potential gradients. The proposed models meet the accuracy and computational speed requirements for modeling complex biological processes. The models are also parametric and can be used for different cases of molecular transport.
- We develop a method to transform the complex biochemical reactions from the thermodynamic energy plane to the information plane by quantifying the probability of microlevel reaction events. W use the micro events to design a stochastic process that captures the complete event behavior. Finally, we use probability theory to estimate the reaction time from the micro event probability measures. We modify the current collision reaction model to treat the reaction as a discrete

stochastic process. We use a velocity distribution of the molecules to capture the effects of the thermodynamic force field profile of the cell. We define a threshold parameter that the collision has to overcome for successful reactions to capture the effect of molecular binding strength by using the parameter activation energy.

We use this estimation method for two modeling scenarios (a) the single molecule model where a single molecule of one reactant can react with multiple molecules of a second reactant, (b) the batch arrival model of the reactants where a batch of molecules of one reactant suddenly arrives to react with multiple molecules of the second reactant. Our results are statistical parameters and we derive the expressions for the average and second moment of the reaction time.

We outline a method to estimate the reaction time for more complex biochemical reactions like different pathway processes sharing the same molecule.

To ascertain the validity of this model, we compare our results with the existing rate based reaction results that provide the mean reaction rate for glycolysis. We show that considering the chaotic environment of the cell, the reaction time estimate will be stochastic in nature. We also demonstrate that for single molecule interactions an exponential distribution will characterize the nature of the reaction time distribution, but for batch arrival process, the distribution will be a two moment distribution like the gamma distribution.

We also analyze the impact of event serialization on the results from our model. The stochastic event technique serializes the different events (i.e., if two different reactions involve the same reactant and scheduled to occur at the same time, we assume that one reaction occurs after the other). Our analytical results show that the adverse effect of this approximation is reduced with increasing number of reactant molecules in the system.

- Next, we develop stochastic models for the protein-DNA binding event. We consider the binding for both bacterial and eukaryotic transcription factors (TFs) to the DNA assuming that the structure, location on chromatin and other details of target sites on the DNA are known. This data can be found from the existing biological databases (e.g., [43, 42]) or need to be determined from experiments if they are not. In contrast to the existing thermodynamic and diffusion based models, our approach closely follows the biological process that involves a number of discrete microevents. We assume that the TF binding site of the DNA is exposed with a difficulty factor depending upon the location of the site with respect to the nucleosome.

  The main idea is that for bacterial cells, the TF (with matching motif) randomly collides with the DNA and only when it hits the binding site with enough kinetic energy to overcome the energy barrier of the site, can the binding occur. Based on our research focus, we abstract the first micro biological event, collision of the TF to the DNA surface, by using the collision theory model where one collision object is non-spherical. The information measure we compute from this abstraction is the probability of DNA-protein collision. The next microlevel biological event is the binding of a TF to the DNA based on the description of the protein and DNA structures on the chromatin as encountered in the biological process.

  This model is general in nature and computationally very fast. This permits its repeated use in the biological simulation for many TF-DNA binding situations. This method *bypasses* the speed-stability paradox of protein-DNA interactions to allow for a computationally efficient model. Our discrete-event based simulator uses this fast model in a similar way as the rate constants used by the Gillespie simulator [29] to approximate the protein-DNA binding time. The TF sliding mechanism due to thermal gradient, for searching the binding region is also incorporated in our model and we show that not all DNA-TF collisions result in sliding. For eukaryotic cells,

the protein-DNA binding mechanism is achieved in two steps 1) diffusion of the TF to the nucleus of the cell and 2) random collisions of the TF with the DNA (we assume that the TF never comes out of the nucleus) for the binding. Our model computes the entire DNA-protein binding time for bacterial cells and DNA-protein binding time *once the protein has entered the nucleus* for eukaryotic cells. The average time for diffusion of protein molecules to the nucleus can be easily computed from standard diffusion models.

We validate our model for the DNA replication process in prokaryotic cells. We also present some "in silico" results showing the effects of protein-DNA binding on gene expression in prokaryotic cells.

- Next, we introduce a collision theory model to explain the temporal kinetics of ligand-protein docking. This is a simplified model which does not incorporate the effects of electrostatic forces and desolvation directly as parameters of the model but consider their effects through the random molecular motion of the proteins in the binding environment. This simplification of the model makes it a random collision problem within the cell and gives us a fairly accurate but computationally fast model for the docking time estimate. Note that the Gillespie simulator considers the docking process as another rate-based equation (a measured quantity that encapsulates all the kinetic properties of the process during the experiment), whereas our proposed model can incorporate the salient features of the docking process along with the structural and functional properties of the protein-ligand pair. This parametric presentation of the binding process makes the model generic in nature and can be easily used for other cases of protein-ligand binding where the assumptions are valid.

  The results generated by this model are very close to experimental estimates. The main conclusion of our work is that the total time required for docking is mostly

contributed by the repeated collisions of the ligand with the protein. Also because the ligand on arriving inside the cell compartment spends most of the time (for binding) away from the protein (to which it binds), the effects of electrostatic force and desolvation are negligible in the binding time estimation. However, electrostatic force and desolvation play a significant role in the determination of the free energy change of the docked complex [22]. This effect is included in our model in determining the probability of docking.

- Finally, we describe a new markov chain based model to simulate complex biochemical reaction systems with reduced computation and memory overheads. The central idea is to transform the continuous domain chemical master equation (CME) based method into a discrete domain of molecular states with corresponding state transition probabilities and times. Our methodology allows the basic optimization schemes devised for the CME and can also be extended to reduce the computational and memory overheads appreciably at the cost of accuracy. The simulation results for the standard Enzyme-Kinetics and a simple Transcriptional Regulatory biological systems show promising correspondence with the CME based methods and point to the efficacy of our scheme. This simulator can give us the dynamics of complex biochemical systems and can be used to improve the scalability of the discrete-event based stochastic simulator.

1.5 Organization of the dissertation

The dissertation is organized as follows. In Chapter 2, we present the stochastic models for molecular/ionic transport inside a cell. Chapter 3 presents the stochastic models for biochemical reactions in the cell cytoplasm and its possible extensions to model reactions occurring in the cell membrane or nucleus. In Chapter 4, we present the model for protein-DNA binding and show its effects on gene transcription and translation.

Chapter 5 presents the models for protein-ligand docking events in the cell. In Chapter 6 we present a markov chain based stochastic biochemical system simulator which is less accurate than the discrete event based system simulator, but has the potential of being more scalable and memory efficient. Finally, in Chapter 7 we summarize the contributions of this thesis, and discuss some directions of future work.

CHAPTER 2

MOLECULAR TRANSPORT

Our event modeling paradigm requires techniques that maintain a reasonable accuracy of the biological process and also reduces the computational overhead. This objective motivates the use of new methods that can transform the problem from energy and affinity based modeling to information based modeling. To achieve this we transform all dynamics within the cell into a random event in time, which is specified through an information measure like probability distribution. We present a model to compute the information measure of molecular transport by estimating the statistical parameters of inter-arrival time between molecules/ions coming to a cell receptor as external signal. This model transforms the diffusion process into the information measure of stochastic event time to get the distribution of the Mg^{2+} departure events. This chapter is organized as follows. Section 2.1 presents the transient models for molecular transport. Section 2.2 reports the results and analysis for a few simple transport events. In Section 2.3, we present the results from the discrete event simulator to show how different transport rates impact the entire PhoPQ system dynamics. Finally, in Section 2.4 we summarize the findings of this chapter.

2.1 Analytical models for molecular transport

From the PhoPQ system, we find that an important process that we have to model is the movement of molecules (Mg^{2+} ions, phoPp, etc). We have identified the following transport models for biological processes: (a) diffusion of charged ions (e.g., Mg^{2+}) in the cell (to model the Mg^{2+} arrival/departure process); (b) diffusion of non-charged

molecules (to model the transport function of phospho-PhoP in the cytosol); (c) diffusion of charged ions out of the cell (to model the Mg^{2+} departure process out of the cell). This transport model should also consider the breakage of the ionic bond between Mg^{2+} and PhoQ molecules for the diffusion to occur; (d) movement of ions or molecules due to additional energy provided by the pump system. In this thesis, we present the analytical solution of the first two models.

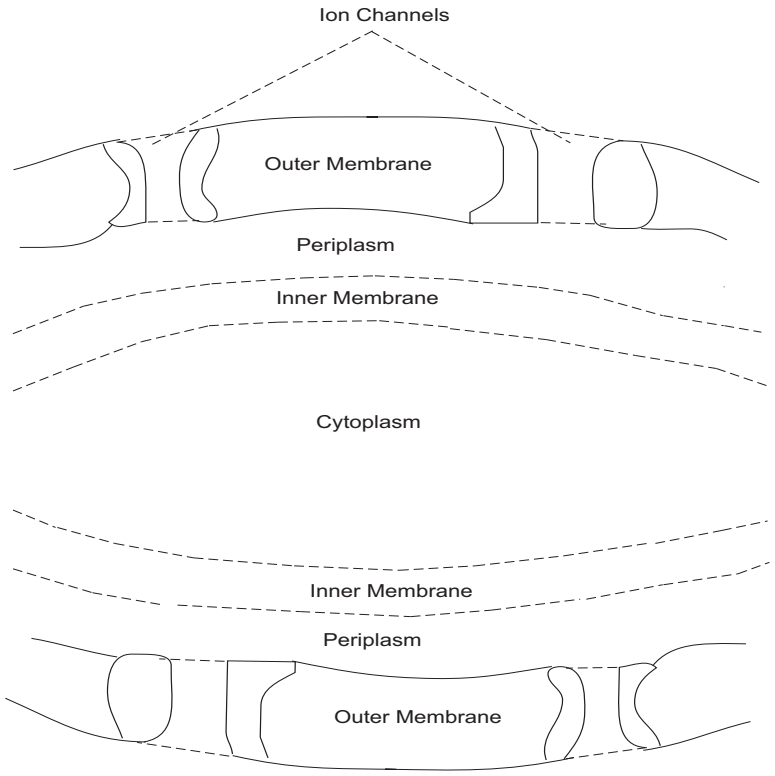

Figure 2.1. Gram-negative bacterial cell showing location of ion channels.

2.1.1 Molecular transport Model 1: Diffusion

The diffusion model of Mg^{2+} ions into the cell is illustrated in Fig 2.1. The diffusion takes place through ion-channels in the outer membrane. PhoQ is located in the cell

membrane and phoP is located in the cytoplasm. The same diffusion model proposed in this chapter can be extended to other cells that have only a single membrane with embedded ion channels such that the ions are transported directly into the cytoplasm.

In [44], the authors have shown that ion transport through ion-channels can be appropriately modeled using standard diffusion equations. We consider the following hypothetical mathematical model: suppose that a long capillary (open at one end) filled with water is inserted into a solution of known chemical concentration C_0, and the chemical species diffuses into the capillary through the open end. The concentration of the chemical species should depend only on the distance down the tube and so is governed by the diffusion equation:

$$\frac{\partial C}{\partial t} = D\frac{\partial^2 C}{\partial x^2}, \qquad 0 < x < \infty, \quad t > 0 \tag{2.1}$$

where for convenience we assume that the capillary is infinitely long. Here, D is the diffusion constant having units $length^2/time$, c is the concentration of the chemical, t is the time and x is the distance traversed inside the capillary by the chemical.

Because the solute bath in which the capillary sits is large, it is reasonable to assume that the chemical concentration at the tip is fixed at $C(0,t) = C_0$, and because the tube is initially filled with pure water, $C(x,0) = 0$.

The solution of this problem is given by [21]:

$$C(x,t) = 2C_0[1 - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{y} exp(-\frac{s^2}{2})ds] \tag{2.2}$$

where $y = \frac{x}{\sqrt{2Dt}}$. We can compute the inter-arrival time between the diffused molecules from the following theorem:

Theorem 1 *The inter-arrival time between the diffusion of the* $(i+1)^{th}$ *and* i^{th} *molecules or ions when the diffusion is based on the concentration gradient only, is given by:*

$$I_{i+1} - I_i = \frac{\pi(2i+1)}{4C_0^2 G^2 D} \tag{2.3}$$

where I_{i+1} and I_i are respectively the times taken for diffusion of the $(i+1)^{th}$ and i^{th} molecules, and G is the cross-sectional area of the capillary.

Proof 1 *The total number of molecules entering the capillary in a fixed time t is given by*

$$N = G\int_0^\infty C(x,t)dx = 2C_0G\sqrt{\frac{tD}{\pi}} \tag{2.4}$$

Thus we get:

$$I_{i+1} = \frac{(i+1)^2\pi}{4C_0^2G^2D},\ \ I_i = \frac{i^2\pi}{4C_0^2G^2D} \Rightarrow I_{i+1} - I_i = \frac{\pi(2i+1)}{4C_0^2G^2D}$$

It is also possible to determine the diffusion coefficient by solving Eqn 2.4 for D:

$$D = \frac{\pi N^2}{4C_0^2G^2t}$$

In [56], this expression was used to measure the diffusion constant in bacteria. With concentration $C_0 = 7 \times 10^7$/ml, and $t = 2, 5, 10, 12.5, 15$ and 20 minutes, they counted $N = 1800, 3700, 4800, 5500, 6700$ and 8000 bacteria in a capillary of length 32 mm with 1 μl total capacity. In addition, with $C_0 = 2.5, 4.6, 5.0$, and 12.0×10^7 bacteria/ml, counts of $1350, 2300, 3400$, and 6200 bacteria were found at $t = 10$ minutes. A value of D in the range of $0.1 - 0.3\ cm^2/hour$ was estimated using Eqn 2.4.

Furthermore, from Eqn 2.2 it can be observed that $C(x,t)/C_0$ is constant on any curve for which z is constant. Thus, $t = x^2/D$ is a level curve for the concentration, and measures how fast the diffusive elements move into the capillary. Here, $t = x^2/D$ is called the diffusion time for the process. Table 6.1 shows typical diffusion times for a variety of cellular structures. Clearly, diffusion is quite effective when distances are short, but totally inadequate for longer distances (e.g. along a nerve axon) and biological systems have to employ other transport mechanisms in such situations. For the sample biological system introduced in Section 1.3, the PhoPp transport to the cytosol process can be

Table 2.1. Estimates of diffusion times for typical cellular structures, computed from the relation $t = x^2/D$ using $D = 10^{-5} cm^2/s$

x	t	Example
10 nm	100 ns	thickness of cell membrane
1 μm	1 ms	size of mitochondrion
10 μm	100 ms	radius of small mammalian cell
250 μm	60 s	radius of squid giant axon
1 mm	16.7 min	half-thickness of frog sartorius muscle
2 mm	1.1 h	half-thickness of lens in the eye
5 mm	6.9 h	radius of mature ovarian follicle
2 cm	2.6 d	thickness of ventricular myocardium
1 m	31.7 yrs	length of a nerve axon

modeled using the diffusion model discussed above. But it is not suited for diffusion of charged molecules, e.g., Mg^{2+}. Also, this is only an approximate model as the source does not ideally replenish itself. So, we will have better results if the initial concentration C_0 is quite high.

2.1.2 Molecular transport Model 2: Diffusion considering ion flux

For better analysis of the diffusion process, we need to consider the ion flux through the membrane of width l (supposing a potential difference exists across it with $\phi(0) = \phi_1$ and $\phi(l) = \phi_2$) created due to movement of positively charged Mg^{2+} ions. We can make a simplifying approximation that the potential gradient through the channel is constant:

$$\frac{\partial \phi}{\partial x} = \frac{\phi_1 - \phi_2}{l} = \frac{V}{l}, \quad V = \phi_1 - \phi_2 \tag{2.5}$$

If the process is in steady state so that the ion flux everywhere in the channel is the same constant, then the total flux, J, can be written as:

$$J = -D[\frac{\partial C(x,t)}{\partial x} + \alpha C(x,t)\frac{V}{l}] \tag{2.6}$$

where, $\alpha = zF/RT$, z = total number of positive charges in Mg^{2+}, F = Faraday's constant, T = absolute temperature, and R = gas constant. Substituting the value of J in the diffusion equation we get:

$$\frac{\partial C}{\partial t} = -\frac{\partial J}{\partial x} = D\frac{\partial^2 C}{\partial x^2} + aD\frac{\partial C}{\partial x}, \quad 0 < x < \infty, \quad t > 0 \tag{2.7}$$

where $a = \alpha V/l$. As it is difficult to achieve a closed form solution of the above equation, we modify the boundary conditions leading to the following theorem:

Theorem 2 *The solution to the diffusion problem outlined in Eqn 2.7 with boundary conditions* $0 < x < l$ *and* $t > 0$ *is given by:*

$$C(x,t) = \sum_{m=1}^{\infty} [\frac{2C_0 m\pi(1-(-1)^m e^{-\frac{zFV}{2RT}})}{(\frac{z^2F^2V^2}{4R^2T^2} + m^2\pi^2)} \times e^{-(\frac{m^2\pi^2}{l^2} + \frac{z^2F^2V^2}{4R^2T^2l^2})Dt - \frac{zFVx}{2RTl}} \sin\frac{m\pi x}{l}] \tag{2.8}$$

Proof 2 *A standard method for solving the above partial differential equation (PDE) is to assume that the variables are separable. Thus, we attempt to find a solution of Eqn 2.7 by putting*

$$C = Y(x)Z(t) \tag{2.9}$$

where Y *and* Z *are functions of* x *and* t*, respectively. Substituting in Eqn 2.7 yields*

$$\frac{Z'(t)}{Z(t)} = D[\frac{Y''(x) + aY'(x)}{Y(x)}] \tag{2.10}$$

such that the left hand side depends on t *only, while the right hand side depends on* x *only. Both sides must therefore be equal to the same constant which is conveniently taken as* $\lambda^2 D$*. We thus have two ordinary differential equations:*

$$\frac{1}{Z}\frac{dZ}{dt} = -\lambda^2 D, \tag{2.11}$$

$$Y''(x) + aY'(x) + \lambda^2 Y(x) = 0 \tag{2.12}$$

The solution for the first equation is given by:

$$Z = e^{-\lambda^2 Dt} \tag{2.13}$$

For the second equation, we make a change of variables to bring it down to a standard form as follows:

$$f'' + (\lambda^2 - \frac{a^2}{4})f = 0, \tag{2.14}$$

$$where, \quad \ln Y = \ln f - \frac{1}{2}\int a dx = \ln f - \frac{ax}{2}$$

The solution for f is given by:

$$f = A\sin x\sqrt{\lambda^2 - \frac{a^2}{4}} + B\cos x\sqrt{\lambda^2 - \frac{a^2}{4}} \tag{2.15}$$

$$f = Ye^{\frac{ax}{2}} \tag{2.16}$$

where A and B are the constants of integration. Thus we can write:

$$Y(x) = e^{-\frac{ax}{2}}[A\sin x\sqrt{\lambda^2 - \frac{a^2}{4}} + B\cos x\sqrt{\lambda^2 - \frac{a^2}{4}}] \tag{2.17}$$

and the concentration at distance x and time t is given by:

$$C(x,t) = Z(t)Y(x) = e^{-(\lambda^2 Dt + \frac{ax}{2})}[A\sin x\sqrt{\lambda^2 - \frac{a^2}{4}} + B\cos x\sqrt{\lambda^2 - \frac{a^2}{4}}] \tag{2.18}$$

Since we are solving a linear equation, the most general solution is obtained by summing solutions of type Eqn 2.18 so that we have:

$$C(x,t) = \sum_{m=1}^{\infty} e^{-(\lambda_m^2 Dt + \frac{ax}{2})} \times [A_m \sin x\sqrt{\lambda_m^2 - \frac{a^2}{4}} + B_m \cos x\sqrt{\lambda_m^2 - \frac{a^2}{4}}] \tag{2.19}$$

The previous capillary model cannot be used in this case to obtain a solution because the solution of this equation is not easily possible. We make a simplified assumption of the system to solve the problem.

We will now consider diffusion out of a plane sheet of thickness l through which the diffusing substance is initially uniformly distributed and the surfaces of which are kept at zero concentration. Mapping this model to our case, the ion channel of length

l is assumed to contain the entire diffusing substance. Every single molecule coming out of this sheet is assumed to enter the cell membrane (Mg^{2+} arrival process). This model thus approximately characterizes the Mg^{2+} diffusion process. The corresponding boundary conditions are as follows:

$$C(x,0) = C_0, \quad 0 < x < l \tag{2.20}$$

$$C(0,t) = 0, \quad C(l,t) = 0 \tag{2.21}$$

where Eqn 2.20 signifies the initial concentration inside the ion channel and Eqn 2.21 signifies the initial concentration (before the start of diffusion) inside the cell membrane. Eqn 2.21 yields:

$$C(0,t) = \sum_{m=1}^{\infty} B_m e^{-\lambda_m^2 Dt} = 0 \Rightarrow B_m = 0$$

Also, substituting $B_m = 0$ in Eqn 2.21 for $x = l$, we get:

$$C(l,t) = 0 = \sum_{m=1}^{\infty} e^{-\lambda_m^2 Dt - \frac{al}{2}} A_m \sin l\sqrt{\lambda_m^2 - \frac{a^2}{4}} \tag{2.22}$$

The solution can be obtained by elimination of variables such that we have:

$$\sin l\sqrt{\lambda_m^2 - \frac{a^2}{4}} = 0 \Longrightarrow \lambda_m = \sqrt{\frac{m^2\pi^2}{l^2} + \frac{a^2}{4}}$$

Substituting these values in Eqn 2.20 we get:

$$C_0 = \sum_{m=1}^{\infty} e^{-\frac{ax}{2}} A_m \sin\frac{m\pi x}{l} \Rightarrow C_0 e^{\frac{ax}{2}} = \sum_{m=1}^{\infty} A_m \sin\frac{m\pi x}{l} \tag{2.23}$$

Multiplying both sides of Eqn 2.23 by $\sin\frac{g\pi x}{l}dx$ and integrating from 0 to l, we get:

$$C_0 \int_0^l e^{\frac{ax}{2}} \sin\frac{g\pi x}{l}\, dx = \sum_{m=1}^{\infty} A_m \int_0^l \sin\frac{m\pi x}{l} \sin\frac{g\pi x}{l}\, dx \tag{2.24}$$

We will use the following identities for the solution of A_m:

$$\int e^{ax} \sin bx\, dx \;=\; e^{ax}[\frac{a}{a^2+b^2}\sin bx - \frac{b}{a^2+b^2}\cos bx]$$

$$\int_0^l \sin\frac{m\pi x}{l}\sin\frac{g\pi x}{l}\,dx = \left\{\begin{array}{ll} 0, & m \neq g \\ \frac{l}{2}, & m = g \end{array}\right\} \tag{2.25}$$

Substituting these identities in Eqn 2.24, we get:

$$A_m = \frac{2C_0 m\pi(1-(-1)^m e^{-\frac{al}{2}})}{l^2(\frac{a^2}{4}+\frac{m^2\pi^2}{l^2})} \tag{2.26}$$

Hence we can write:

$$C(x,t) = \sum_{m=1}^{\infty} \frac{2C_0 m\pi(1-(-1)^m e^{-\frac{zFV}{2RT}})}{(\frac{z^2F^2V^2}{4R^2T^2}+m^2\pi^2)} \times e^{-(\frac{m^2\pi^2}{l^2}+\frac{z^2F^2V^2}{4R^2T^2l^2})Dt-\frac{zFVx}{2RTl}} \sin\frac{m\pi x}{l} \tag{2.27}$$

Thus we get the time domain analysis for the concentration of Mg^{2+} molecules from which we can derive the mean departure rate of Mg^{2+}. The inter-arrival time between the diffused molecules can be computed from the following theorem:

Theorem 3 *The inter-arrival time between the diffusion of the $(i+1)^{th}$ and i^{th} molecules or ions when the diffusion is based on both the concentration and potential gradients across the cell is given by $I_{N-i} - I_{N-i-1}$, where I_{N-i} and I_{N-i-1} are the times taken for diffusion of the i^{th} and $(i+1)^{th}$ molecules/ions respectively and can be solved from the following equations:*

$$N-i-1 = 2C_0G\sum_{m=1}^{\infty} m^2\pi^2\{\frac{1-(-1)^m e^{-\frac{zFV}{2RT}}}{\frac{z^2F^2V^2}{4R^2T^2}+m^2\pi^2}\}^2 \times e^{-(\frac{m^2\pi^2}{l^2}+\frac{z^2F^2V^2}{4R^2T^2l^2})DI_{N-i-1}} \tag{2.28}$$

$$N-i = 2C_0G\sum_{m=1}^{\infty} m^2\pi^2\{\frac{1-(-1)^m e^{-\frac{zFV}{2RT}}}{\frac{z^2F^2V^2}{4R^2T^2}+m^2\pi^2}\}^2 \times e^{-(\frac{m^2\pi^2}{l^2}+\frac{z^2F^2V^2}{4R^2T^2l^2})DI_{N-i}} \tag{2.29}$$

Proof 3 *The total number, N, of molecules/ions present inside the sheet of area G in a fixed time I_N is given by:*

$N = G\int_0^l C(x,t) = 2C_0G\sum_{m=1}^{\infty} m^2\pi^2\{\frac{1-(-1)^m e^{-\frac{zFV}{2RT}}}{\frac{z^2F^2V^2}{4R^2T^2}+m^2\pi^2}\}^2 \times e^{-(\frac{m^2\pi^2}{l^2}+\frac{z^2F^2V^2}{4R^2T^2l^2})DI_N}$

The inter-arrival time can be computed in a straightforward way by noting that diffusion occurs when a molecule/ion goes out off the plane sheet.

Table 2.2. Parameter Estimation for the Numerical Plots

Parameters	Salmonella cell
Diameter of an ion-channel (d)	$10 \times 10^{-10} m$
Cross-sectional area of ion-channel (G')	$4\pi(\frac{d}{2})^2$
Number of ion-channels (N')	100
Cross-sectional area of capillary (G)	$N' \times G'$
Diffusion constant (D)	$10^{-5}\ cm^2/s$
Potential gradient (V)	$60\ mV$

2.2 Numerical results and analysis

Let us present the numerical results for our transport models. Table 6.2 lists the parameters used.

Fig 2.2 plots the inter-arrival time of diffused molecules for molecular concentrations of $10^{-9}, 10^{-6}, 10^{-5}, 10^{-4}$ moles, respectively. As stated earlier, this model is suitable for diffusion of uncharged molecules. The figure shows that the inter-arrival time increases with increasing number of molecules diffused in. This is because the concentration gradient reduces with more molecules diffusing in, resulting in a larger time required for the molecules to move in. It is observed that the larger the initial concentration, the lesser is the inter-arrival time. This is expected due to a higher concentration gradient. Also, it can be observed that the inter-arrival time distribution can be fitted to an exponential distribution.

Fig 2.3 plots the inter-arrival times for diffusion model 2 where the potential gradient is considered. We assume a constant potential gradient of 60mV for the molecules to overcome for diffusion to take place. The inter-arrival times are higher than the first model because the molecules have to overcome the potential gradient as well in order to diffuse. Here, the exponential increase in the inter-arrival times can be observed more

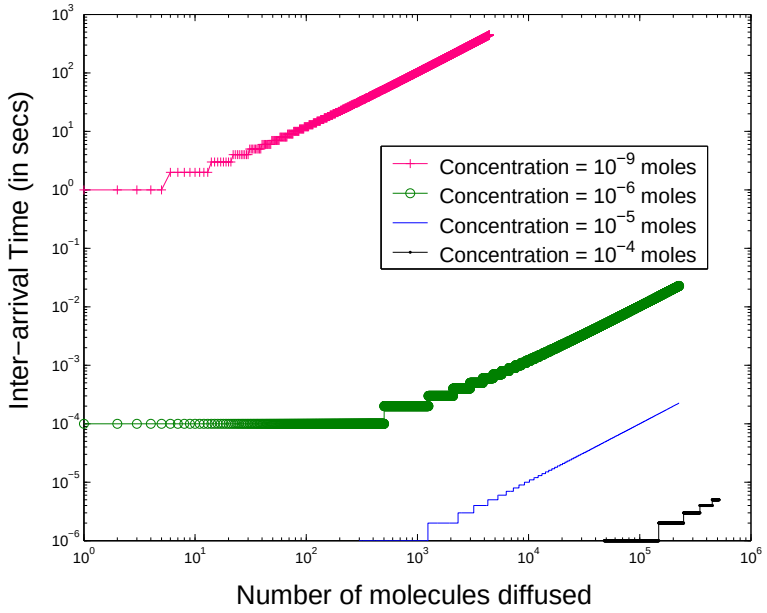

Figure 2.2. Inter-arrival time vs number of molecules for Diffusion Model 1.

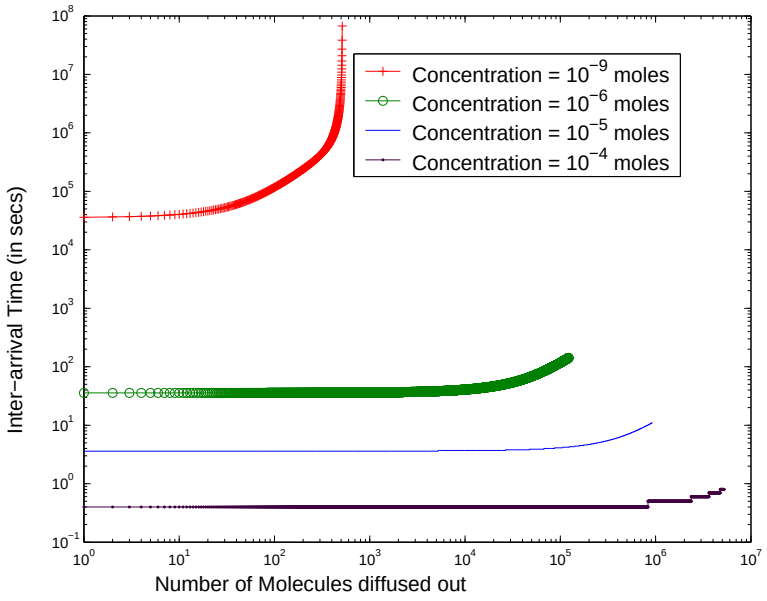

Figure 2.3. Inter-arrival time against the number of molecules for Diffusion Model 2.

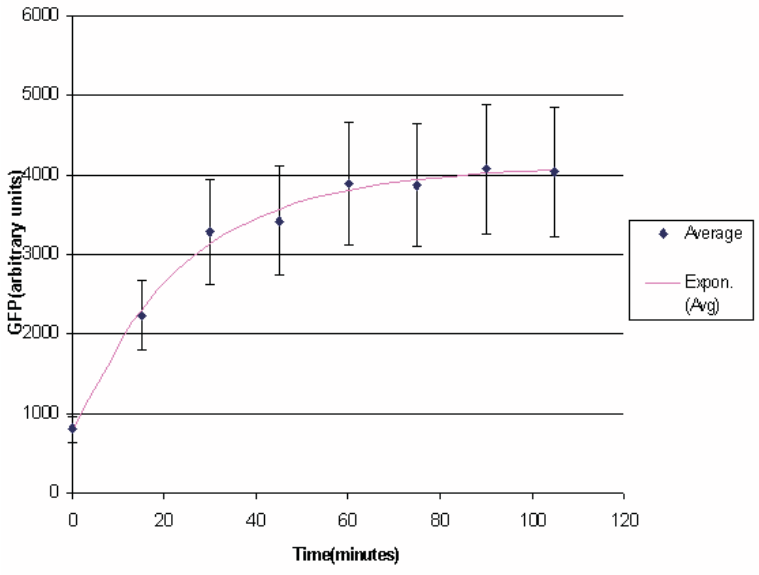

Figure 2.4. Experimental results: phoPp concentration vs time, $Mg^{2+} \sim 10^{-3} moles$.

clearly. This scenario is best depicted by the curve for concentration $10^{-9} moles$ where the results are generated for a large number of molecules diffused out.

Note that model 1 is standard and we estimated the inter-arrival times of Mg^{2+} molecules using it. The transient analysis of model 2 is hard to solve and hence we chose a specific boundary condition (as mentioned before) to derive a closed form expression. The corresponding results compare well with model 1 indicating its validity.

2.3 Simulation results of the PhoPQ system

As the arrival/departure of Mg^{2+} molecules into the periplasm is essentially a stochastic process, a constant diffusion rate is not suitable to trigger the input process of the PhoPQ system. Hence we use an *exponential distribution* (as indicated by the numerical plots) to estimate the inter-arrival times for diffusion of Mg^{2+} (which is considered to be a random variable) to generate the results. The mean of this exponential distribution is obtained from similar plots of inter-arrival times and corresponding curve-fitting. As

mentioned in Section 1.3, the PhoPQ system is triggered at micromolar concentrations of Mg^{2+} outside the cell, i.e., with millimolar Mg^{2+} concentration inside the cell. Thus, it is fair to assume $C_0 \simeq 10^{-3}$ moles. The mean of the inter-arrival times of Mg^{2+} for this concentration is estimated as approximately 10^{-6} secs for Model 1 and 10 msecs for Model 2 respectively. The discrete-event simulation framework correspondingly uses a Poisson distribution with the same mean (as the inter-arrival times follow an exponential distribution) to estimate the departure process of Mg^{2+} triggering the signal transduction cascade.

The simulation framework also uses the holding time estimates of other elementary biological processes such as cytoplasmic reactions [70],[71],[76] (models $2, 3, 4$ and 5 in Fig 1.6), protein-DNA binding [72],[78] (model 6 in Fig 1.6) and gene transcription/translation times (the average time for this process was assumed based on current research results).

In the following, we present the results illustrating the sensitivity of the simulation to the diffusion models used.

Fig 2.4 plots the concentration of phoPp molecules with time as observed in wet lab experiments [95]. At present it is difficult to directly link the results of the simulation to the wet lab experiments data that we have. This is because simulation gives the temporal dynamics in actual molecular count, whereas the fluorescent tag based wet lab experiments only show the sensitivity of the fluorescent light. It was not possible to calibrate the fluorescent tag sensitivity to molecular count per cell in the past. Thus our simulation results validate the similarity of the temporal dynamics of experimental results now, without actual comparison of the molecular count of a cell. Currently more sophisticated experiments like microfluidic based single cell assay [98] allow real time observation of single molecules in a cell. In future, we hope to get molecular level measurements in a cell to validate our results quantitatively. Fig 2.5 plots the change of

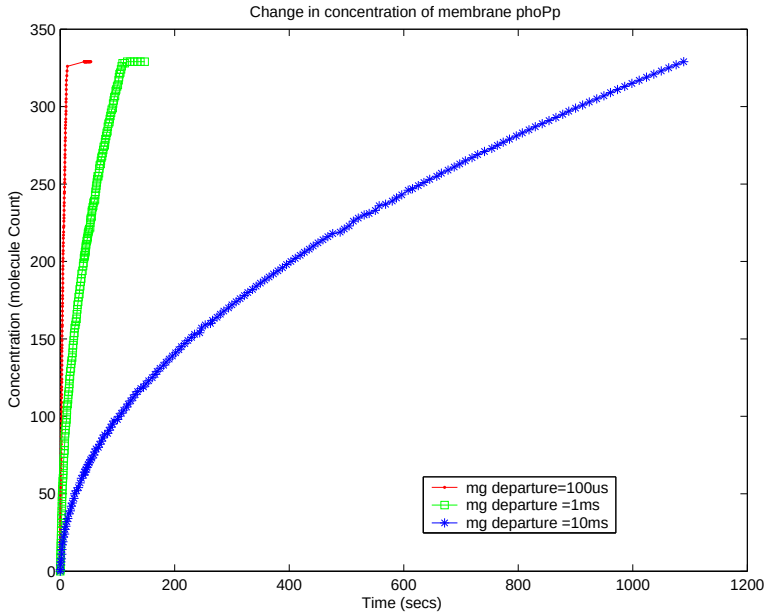

Figure 2.5. Simulation results: phoPp concentration vs time, $Mg^{2+} \sim 10^{-3} moles$.

phoPp concentration from our discrete event simulation framework with three different means for the Mg^{2+} departure process. It can be noted that with mean of $100\mu s$, the phoPp concentration change is quite steep, and it achieves the maximum value of phoPp (observed experimentally) in the cell at about 1 sec. But as the mean is increased to 10 ms, we get acceptable estimates of the phoPp concentration. This outlines the importance of diffusion Model 2 where the mean of the Mg^{2+} departure process is indeed in the range of 10 ms as opposed to the $1\mu s$ range for Model 1. As discussed earlier, Model 1 is suitable for the phoPp transport process in the cytosol.

2.4 Summary

For the in-silico simulation, we need the transformation of biological functions into information measure like probability distributions of event time. We have presented one example of the transformation of the biological function (i.e., molecular transport time) driven by concentration and potential gradients in this chapter. The proposed stochastic

models meet the accuracy and computational speed requirements for modeling complex biological processes. These models are parametric and can be used for different cases of molecular transport. Once the complete set of mathematical models for the different biological functions are in place, it should be possible to reuse these models to construct other biological process models with marginal changes. The models provide for both speed of computation and flexibility that is required to model the dynamics of an entire cell.

CHAPTER 3

REACTION MODELS

This chapter presents the event model of the biochemical reactions between the molecules inside the cytoplasm of a cell where the reaction environment is highly chaotic. We present a mathematical formulation for the estimation of the reaction time between two molecules within a cell based on their discrete states. In particular, we propose two models: 1) The reactant molecules enter the system one at a time to initiate reactions, and 2) The reactant molecules arrive in batches of a certain size. We derive expressions for the average and second moment of the time for reaction. Unlike rate equations, the proposed model does not require the assumption of concentration stability for multiple molecule reactions. The parametric nature of the model makes it generic and useful for diverse studies.

This chapter is organized as follows: Section 3.1 presents the stochastic models for biochemical reactions inside the cell. Section 3.2 reports the results and analysis for a few simple reaction events. In Section 3.3, we discuss how delayed reactions can be handled by our models and also present the limitations of our model. Finally, in Section 3.4 we summarize the findings of this chapter.

3.1 Models for biochemical reactions

We will present a simple reaction pair in this section to explain the discrete event modeling of biochemical reactions. Consider the elementary reaction pair R_1 and R_2 with five types of molecules X_1,X_2,X_3,X_4 and X_5:

$$R_1 : X_1 + X_2 \longrightarrow X_3 \tag{3.1}$$

$$R_2 : X_4 + X_2 \longrightarrow X_5$$

Note that we divide the reaction event into two independent micro-events as follows:

1. Random collisions between the reactants; this allows us to compute the probability of collision (p_c) between the reactant molecules.
2. A reaction will occur only when the kinetic energy of the colliding reactant exceeds the activation energy requirement for the reaction; this allows us to compute the probability of reaction (p_r).

The total probability for reaction after a collision is hence the joint probability of these two events. To model these reactions analytically in the time domain, we consider two different models for the arrivals of X_1 and X_4 types of molecules in the system which, we are assumed to contain a fixed number, n_2, of X_2 molecules.

3.1.1 Model 1: Reactant molecules enter the system one at a time

The molecules of X_1 and X_4 enter the system one at a time to start the reactions. From the principles of collision theory for hard spheres, we model each reactant molecule as a rigid sphere with radii r_1, r_2, r_3, r_4, and r_5 respectively for molecules of types X_1, X_2, X_3, X_4, and X_5 as shown in Fig 4.1.

We define our coordinate system such that molecule X_2 is stationary with respect to molecule X_1 for reaction R_1, so that X_1 moves towards molecule X_2 with a relative velocity U_{12}. Molecule X_1 moves through space to sweep out a collision cross section $A = \pi r_{12}^2$ (as illustrated in Fig 3.2), where r_{12} is the collision radius given by:

$$r_{12} = r_1 + r_2$$

If the center of the X_2 molecule comes within a distance of r_{12} of the center of the X_1 molecule, they will collide. To discretize the system we consider the dynamics of this process within a small time interval Δt. We assume that the temporal reaction process

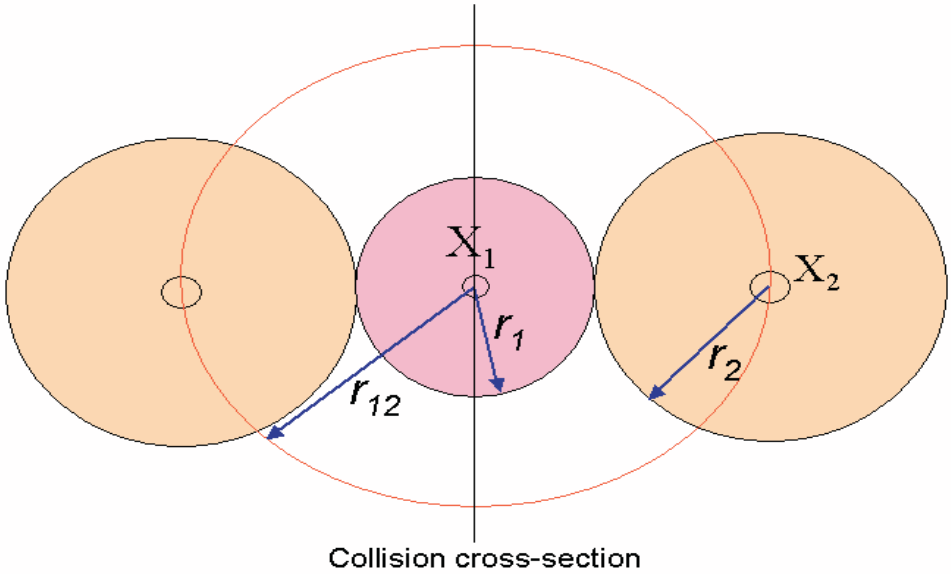

Figure 3.1. Schematic diagram of molecules of types X_1 and X_2.

is an independent sequence of events separated by Δt. In this time interval, the X_1 molecule sweeps out a volume ΔV given by:

$$\Delta V = \pi r_{12}^2 U_{12} \Delta t$$

Now, the probability of X_1 molecules in the collision volume ΔV is $p_{X_1} = 1$ (as one X_1 molecule entered the cell creating a collision volume of ΔV).

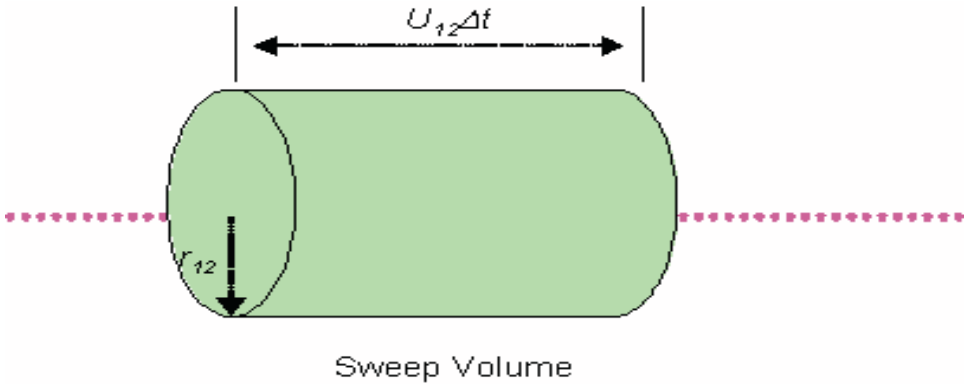

Figure 3.2. Volume swept out by molecule X_1 in time Δt.

The probability of at least one molecule of X_2 being present in an arbitrary uniformly distributed ΔV in V is $p_{X_2} = \Delta V.n_2/V$, where V denotes the cell volume (ideally V should be the volume of the cytoplasm which can be approximated by the entire cell volume). The probability that an X_1 molecule collides with an X_2 molecule in Δt is given by:

$$p_c = p_{X_1} \times p_{X_2} = \Delta V.n_2/V = \frac{n_2 \pi r_{12}^2 U_{12} \Delta t}{V} \tag{3.2}$$

Thus we have a stochastic sequence of events characterized by the probability of collision, and it is important to determine whether the collision will create the reaction. To complete the reaction, the molecules have to bind to each other. Different type of bonds (ionic, covalent, hydrogen, etc.) require different activation energies for binding. We next assume that the colliding molecules must cross an energy threshold, defined by the free energy (E_{Act}), to provide the energy to react. Also, we assume that only the kinetic energy directed along the line of centers contribute to the reaction as the effects of other forces (e.g. coulomb force) have been captured with the velocity distribution. With these two assumptions, we define the probability of another independent event: successful reaction after collision denoted by p_r. The kinetic energy of approach of X_1 towards X_2 with relative velocity U_{12} is $E = \frac{m_{12} U_{12}^2}{2}$, where $m_{12} = \frac{m_1 m_2}{m_1 + m_2}$ = the reduced mass, m_1 = mass (in gm) of molecular species X_1 and m_2 = mass (in gm) of molecular species X_2. We also assume that as E increases above E_{Act}, the number of collisions that result in reaction also increases linearly. Thus the probability, p_r, for a reaction to occur, is given by:

$$p_r = \left\{ \begin{array}{cc} \frac{E - E_{Act}}{E}, & \text{for } E > E_{Act} \\ 0, & \text{otherwise} \end{array} \right\} \tag{3.3}$$

Hence, the joint probability, p, for collision and reaction is given by:

$$p = p(Reaction, Collision) = p_r \times p_c = \left\{ \begin{array}{c} p_c \frac{(E-E_{Act})}{E}, \quad \text{for } E > E_{Act} \\ 0, \qquad \text{otherwise.} \end{array} \right\}$$

The above equations assume a fixed relative velocity U_{12} for the reaction. The velocity distribution of the macromolecules inside a cell capturing the effects of the different forces is obtained from Molecular Dynamic Simulation. It is found to be comparable to the Maxwell-Boltzmann distribution of molecular velocities [38] for a species of mass m, given by:

$$f(U,T)dU = 4\pi(\frac{m}{2\pi k_B T})^{3/2} e^{\frac{-mU^2}{2k_B T}} U^2 dU$$

where k_B = Boltzmann's constant = $1.381 \times 10^{-23} kg\ m^2/s^2/K/molecule$, and T is the absolute temperature at which the reaction occurs. Replacing m with the reduced mass m_{12} of the molecules X_1 and X_2, we get,

$$f(U,T)dU = 4\pi(\frac{m_{12}}{2\pi k_B T})^{3/2} e^{\frac{-m_{12}U^2}{2k_B T}} U^2 dU \tag{3.4}$$

The term on the left hand side of the above equation denotes the fraction of X_1 molecules with relative velocities between U and $(U + dU)$. Summing up the collisions for the X_1 molecules for all velocities we get the probability of reaction, p, as a function of temperature only as follows:

$$p(T) = \int_0^\infty p f(U,T) dU$$

Now, recalling $E = \frac{m_{12}U_{12}^2}{2}$, and hence, $dE = m_{12}U_{12}dU$, and substituting into Eqn 4.3, we get:

$$f(U,T)dU = 4\pi(\frac{m_{12}}{2\pi k_B T})^{3/2} \frac{2E}{U m_{12}^2} e^{\frac{-E}{k_B T}} dE$$

Thus,

$$p = \int_{E_{Act}}^\infty \frac{(E - E_{Act}) 4 n_2 \pi r_{12}^2 \Delta t}{V k_B T} \sqrt{\frac{1}{2\pi k_B T m_{12}}} e^{-\frac{E}{k_B T}} dE = \frac{n_2 r_{12}^2 \Delta t}{V} \sqrt{\frac{8\pi k_B T}{m_{12}}}\ e^{\frac{-E_{Act}}{k_B T}} \tag{3.5}$$

We mentioned that we discretize the temporal reaction process as a Bernoulli trial process. Next we compute the average time taken to complete the reaction with this probability. Let us assume that the molecule composition does not change during the reaction time. This is valid due to the very short time for reaction compared to the time taken for a potential change in the reaction environment for the associated molecules. Let $\Delta t = \tau$ be an infinitely small time step. The molecules try to react through repeated collisions. If the first collision fails to produce a reaction, they collide again after τ time units and so on.

We can interpret p as the probability of a successful reaction in time τ. Thus the *average time of reaction* R_1, T^{DE}_{avg1}, and the corresponding second moment, $T^{DE}_{2_{nd}moment1}$, can be formalized as follows:

Theorem 4 *The average and the second moment of reaction times of type R_1 for one molecule of X_1 reacting with n_2 molecules of X_2 is given by:*

$$\begin{aligned} T^{DE}_{avg1} &= \frac{\tau}{p} \\ T^{DE}_{2_{nd}moment1} &= \frac{(2-p)\tau^2}{p^2} \end{aligned}$$

Proof 4 *The average time of reaction can be approximated by summing up the times taken for reaction with the first collision, second collision and so on. Thus we get:*

$$T^{DE}_{avg1} = p\tau + (1-p)p.2\tau + (1-p)^2p.3\tau + ...$$

We have:

$$T^{DE}_{avg_1} = p\tau S \tag{3.6}$$

where,

$$S = 1 + 2(1-p) + 3(1-p)^2 + ... \tag{3.7}$$

Also, multiplying both sides of this equation by (1-p) we get:

$$(1-p)S = (1-p) + 2(1-p)^2 + 3(1-p)^3 + ... \tag{3.8}$$

Subtracting Eqn 3.8 from Eqn 3.7 we get:

$$pS = 1 + (1-p) + (1-p)^2 + (1-p)^3 + ... = \frac{1}{p} \Rightarrow S = \frac{1}{p^2}$$

Substituting S in Eqn 3.6 we get:

$$T_{avg_1}^{DE} = \frac{\tau}{p}$$

Similarly, for the second moment we have:

$$T_{2_{nd}moment1}^{DE} = p\tau^2 + (1-p)p.(2\tau)^2 + (1-p)^2p.(3\tau)^2 + ... = p\tau^2 S$$

where,

$$S = 1 + 2^2(1-p) + 3^2(1-p)^2 + ... \tag{3.9}$$

Similarly,

$$S(1-p) = (1-p) + 2^2(1-p)^2 + 3^2(1-p)^3 + ... \tag{3.10}$$

Subtracting Eqn 3.10 from Eqn 3.9, we get:

$$pS = 1 + (2^2 - 1^2)(1-p) + (3^2 - 2^2)(1-p)^2 + (4^2 - 3^2)(1-p)^3 + ...$$

Now, we can substitute the terms of the form $a^2 - (a-1)^2$ *in the above equation by* $(2a-1)$ *as follows:*

$$\begin{aligned} pS &= 1 + (2.2-1)(1-p) + (3.2-1)(1-p)^2 + (4.2-1)(1-p)^3 + ... \\ &= [1 + 2\{2(1-p) + 3(1-p)^2 + 4(1-p)^3 + ...\}] \\ &- [(1-p) + (1-p)^2 + (1-p)^3 + ...] \\ &= 1 + 2Y - Z \end{aligned} \tag{3.11}$$

where, $Y = \{2(1-p) + 3(1-p)^2 + 4(1-p)^3 + ...\}$ *and*

$Z = (1-p) + (1-p)^2 + (1-p)^3 + ...$

Z forms an infinite geometric series and we get:

$$Z = \frac{1-p}{p}$$

Next we multiply $(1-p)$ *to* Y *and subtract it from the expression of* Y *to get:*

$$pY = 2(1-p) + (1-p)^2 + (1-p)^3 + ... = 2(1-p) + \frac{(1-p)^2}{p} \Rightarrow Y = \frac{1-p^2}{p^2}$$

Substituting the values for Y *and* Z *in Eqn 3.11 we get:*

$$\begin{aligned} S &= \frac{2-p}{p^3} \\ \Rightarrow T^{DE}_{2_{nd}moment1} &= \frac{(2-p)\tau^2}{p^2} \end{aligned}$$

Note that the computation of T^{DE}_{avg1} and $T^{DE}_{2_{nd}moment1}$ assume that no other reaction (having the same reactant) is overlapping with R_1. If reaction R_2 overlaps with R_1, the average time estimate (of R_1) should increase as R_2 will reduce the number of X_2 molecules available for reaction R_1. The discrete event approach *serializes* such overlapping reactions and hence our estimates of T^{DE}_{avg1} and $T^{DE}_{2_{nd}moment1}$ is independent of the effect of R_2. We next derive expressions for the *adjusted time* required for chemical reactions where such overlapping is considered. Our goal is to show that the average time for serialized chemical reactions is comparable to the adjusted time when the number of reactant molecules is large in the biological system.

3.1.1.1 The actual scenario: conflicting reaction events

The discrete event technique assumes that all events are serialized in time. Thus no two conflicting events occur at exactly the same time. If two conflicting reaction events

are triggered at the same time, one event will be considered to occur before the other one. Let us consider the following three reactions:

$$R_1 : X_1 + X_2 \longrightarrow X_3$$

$$R_2 : X_4 + X_2 \longrightarrow X_6$$

$$R_3 : X_6 + X_5 \longrightarrow X_7$$

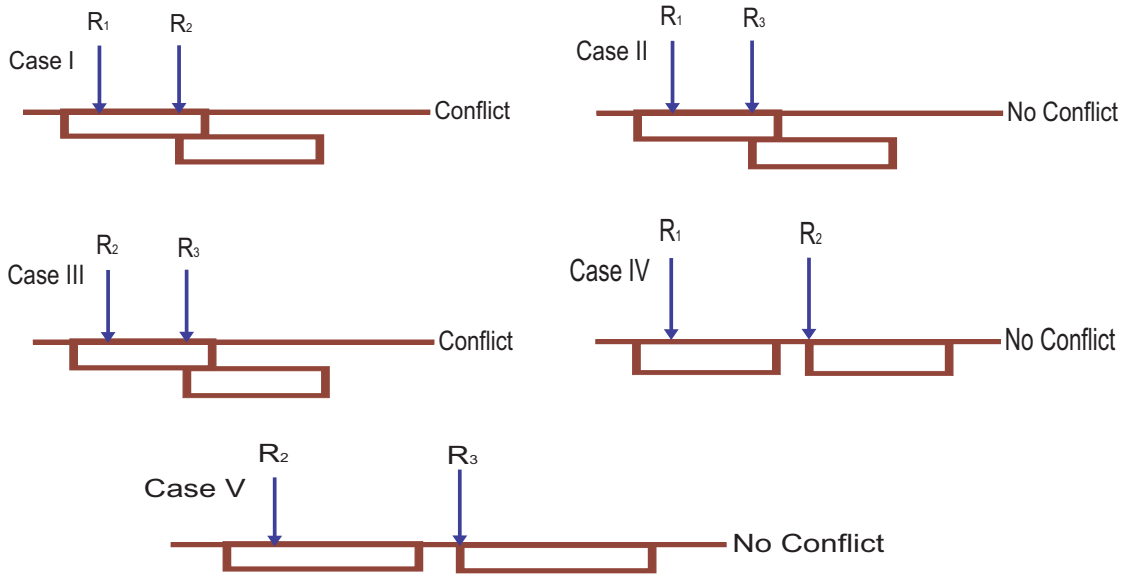

Figure 3.3. Possible Event scenarios with potential conflicts.

We next illustrate the different possible event scenarios in Fig 3.3. The events are shown along the time axis. There is a conflict in Case I because the X_2 molecules are shared by both reactions R_1 and R_2; so they have to be serialized. There are no conflicts between reactions R_1 and R_3 as they do not share any molecules (Case II). Similarly, there is conflict between R_2 and R_3 due to the X_6 molecules (Case III). Also, there are no conflicts between reaction pairs R_1, R_2 (Case IV) and R_2, R_3 (Case V) as the reactions do not overlap. The concept of serialization in discrete event modeling requires that these conflict scenarios should be approximated as one reaction occurring only after the

completion of another one. Thus, in Case III above, we assume that the molecule of X_6 is available to R_3 only after the completion of R_2. Similarly, in Case I, we assume, that the molecule of X_2 is available to R_2 only after the completion of R_1. These assumptions will work well as long as the number of molecules in the system is large. Thus, this marginal adjustment will not significantly distort the reaction time.

We next present the analytical modeling of Case I to show how the adjusted reaction time should be derived.

3.1.1.2 Computing the Adjusted Time of Reaction R_1 for Case I:

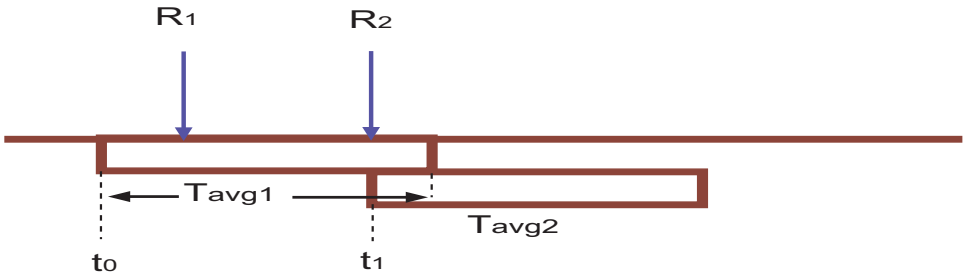

Figure 3.4. Case I Event Scenario and its Timing Details.

In Case I (Fig 3.4), molecules of X_2 are shared by reactions R_1 and R_2. Thus when R_2 starts, there is faster depletion of molecules of X_2. This will obviously result in a larger average time of reaction, R_1. In this section, we will estimate $T_{avg_1}^{adjusted}$, the adjusted time for reaction R_1. It should be noted however, that this adjusted time is also a random variable like $T_{avg_1}^{DE}$. But, unlike $T_{avg_1}^{DE}$, the computation of $T_{avg_1}^{adjusted}$ involves the effect of multiple reactions that share a reactant. Here, we compute $T_{avg_1}^{adjusted}$ for two conflicting reactions, i.e., R_1 and R_2. The computations become a lot more complex if we consider more conflicting reactions. This is exactly the reason why the discrete event

framework only uses the estimates of T^{DE}_{avg1} and $T^{DE}_{2_{nd}moment1}$ to approximately estimate the reaction time.

Let, n_1 be the number of molecules of X_1 at time t_0, n_2 be the number of molecules of X_2 at time t_0 and n_4 = number of molecules of X_4 at time t_1. Also, let p_1 be the probability of reaction of X_1 and X_2 during time $(t_1 - t_0)$. As calculated before, we have

$$p_1 = \frac{n_2 r_{12}^2 \tau}{V} \sqrt{\frac{8\pi k_B T}{m_{12}}}\, e^{\frac{-E_{A12}}{k_b T}}$$

where E_{A12} is the activation energy required for R_1. Let,

$$q = \lceil \frac{t_1 - t_0}{\tau} \rceil$$

Here, q denotes the number of timesteps of length τ between t_1 and t_0. Now, the average number of X_2 molecules in the system at time t_1 is $n_2^{t_1} = n_2 - \sum_{k=1}^{q} p_1(1-p_1)^{k-1}$. Also, let p_2 be the probability of reaction between X_4 and X_2 molecules. Then,

$$p_2 = \frac{n_2^{t_1} r_{42}^2 \tau}{V} \sqrt{\frac{8\pi k_B T}{m_{42}}}\, e^{\frac{-E_{A42}}{k_b T}}$$

where r_{42}, m_{42} and E_{A42} are defined as before for reaction R_2. Now, the probability of reaction R_1 changes from the $(q+1)^{th}$ step onwards as R_2 can reduce the number of X_2 molecules in the system. Thus,

$$p_1^{t_1+h\tau} = (n_2^{t_1} - \sum_{k=1}^{h} p_2(1-p_2)^{k-1}) \frac{r_{12}^2 \tau}{V} \sqrt{\frac{8\pi k_B T}{m_{12}}}\, e^{\frac{-E_{A12}}{k_b T}}$$

Here, $p_1^{t_1+h\tau}$ denotes the probability of reaction R_1 at the h^{th} time-step after time t_1. The adjusted time for reaction R_1, $T^{adjusted}_{avg_1}$, is given by:

$$T^{adjusted}_{avg_1} = \sum_{i=1}^{q} p_1(1-p_1)^{i-1} i\tau + \sum_{j=1}^{\infty} p_1^{t_1+j\tau}(1-p_1)^q \prod_{k=1}^{j-1}(1-p_1^{t_1+k\tau})[(q+j)\tau]$$

Similarly, the adjusted time for R_2 can be calculated. Also, the adjusted second moments for reactions R_1 and R_2 can be calculated easily following the same concepts. In Section 5.4, we report the comparisons between $T^{DE}_{avg_1}$ and $T^{adjusted}_{avg_1}$, and show that the difference between them reduces as the number of X_2 molecules in the cell increases.

3.1.2 Model 2: Reactant molecules enter the cell in fixed size batches

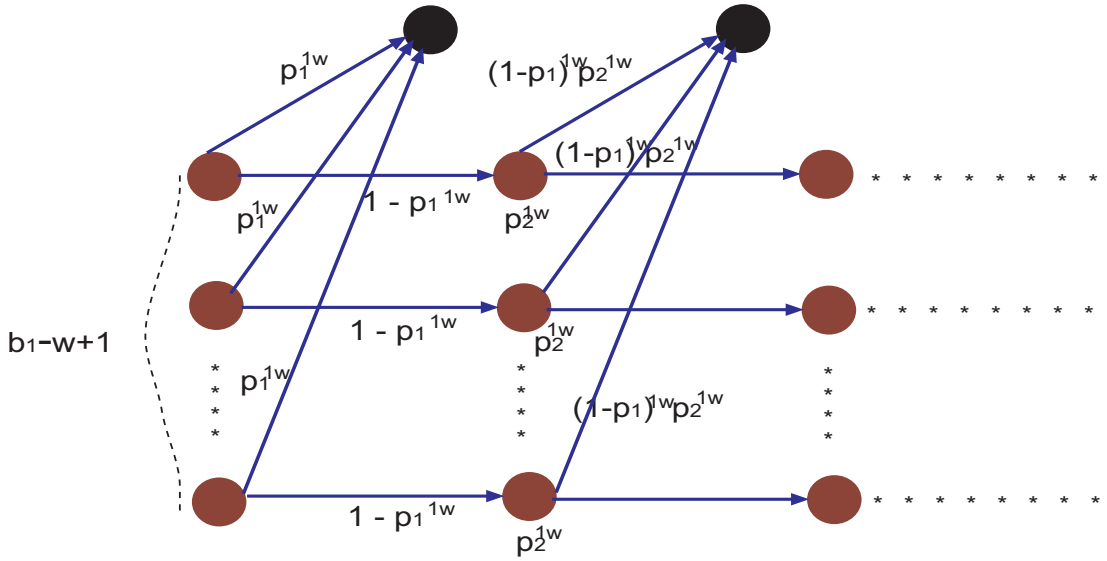

Figure 3.5. State diagram: w^{th} reaction when X_1 molecules arrive in batches.

We assume that the X_1 molecules arrive in batches of size b_1 and X_4 molecules arrive in batches of size b_4 in the system. We will analytically model the average time for reaction R_1 for only the discrete event case, $T_{avg_1}^{batch/DE}$ (i.e., assuming no overlap between reactions involving shared reactants). Fig 3.5 depicts the scenario. Let, p_k^{ij} denote the probability of the j^{th} reaction (we can have a total of b_1 reactions in a batch size of b_1) of type R_i at the k^{th} collision. Thus, the probability of the *first reaction of type* R_1 between one X_1 molecule and an X_2 molecule resulting from the first collision, p_1^{11}, is given by:

$$p_1^{11} = \frac{b_1 n_2 r_{12}^2 \tau}{V} \sqrt{\frac{8\pi k_B T}{m_{12}}}\, e^{\frac{-E_{A12}}{k_b T}}$$

The numerator on the right hand side of the above equation gets multiplied by b_1 to sum up the probability of reaction for all the b_1 molecules that arrived in a single batch. The black circles in Fig 3.5 signify the reaction R_1 occurring from the first collision, second

collision and so on. Hence, the probability of the first reaction of type R_1 from the i^{th} collision ($2 \leq i \leq \infty$), is given by:

$$p_i^{11} = p_1^{11}$$

Also, the probabilities of the w^{th} reaction of type R_1 from the first collision is given by:

$$p_1^{1w} = \frac{(b_1 - w + 1)(n_2 - w + 1)r_{12}^2\tau}{V}\sqrt{\frac{8\pi k_B T}{m_{12}}}\, e^{\frac{-E_{A12}}{k_B T}}$$

Similarly, the probability of the w^{th} reaction of type R_1 from the i^{th} collision, p_i^{1w}, is given by:

$$p_i^{1w} = p_1^{1w}$$

The average time to complete one reaction of type R_1, $T_{avg_1}^{batch/DE}$, and the corresponding second moment, $T_{2_{nd}moment1}^{batch/DE}$, in the discrete event model is given by the following theorem:

Theorem 5 *The average and second moment to complete a reaction of type R_1 in a batch of size b_1 molecules of X_1 and n_2 molecules of type X_2 in the discrete event model is given by:*

$$T_{avg_1}^{batch/DE} = \frac{\sum_{k=1}^{b_1}[\sum_{i=1}^{\infty}(p_i^{1k} i\tau \prod_{j=1}^{i-1}(1 - p_j^{1k}))]}{b_1} = \frac{\sum_{k=1}^{b_1}[\sum_{i=1}^{\infty}(p_1^{1k}(1 - p_1^{1k})^{i-1} i\tau)]}{b_1}$$

$$T_{2_{nd}moment1}^{batch/DE} = \frac{\sum_{k=1}^{b_1}[\sum_{i=1}^{\infty}(p_i^{1k}(i\tau)^2 \prod_{j=1}^{i-1}(1 - p_j^{1k}))]}{b_1} = \frac{\sum_{k=1}^{b_1}[\sum_{i=1}^{\infty}(p_1^{1k}(i\tau)^2(1 - p_1^{1k})^{i-1})]}{b_1}$$

Proof 5 *The time taken for the k^{th} reaction in the batch is computed by adding up the contributions from all the i collisions, (where $1 \leq i \leq \infty$) as follows:*

$$\sum_{i=1}^{\infty}(p_i^{1k} i\tau \prod_{j=1}^{i-1}(1 - p_j^{1k})) \tag{3.12}$$

The average time of any reaction in the batch is simply computed by adding up the times for all the possible b_1 reactions and taking the mean (i.e., dividing by b_1). The second moment can also be calculated in a similar fashion.

The computation of the adjusted time of reaction R_1 in the batch model becomes very cumbersome and is not included here. The comparisons between the discrete event based estimates and adjusted estimates are only shown for Model 1 in Section 5.4. However, the batch model is required when the number of reactions increase significantly in the system, triggering a large number of discrete reaction events in the stochastic simulation. In such scenarios, we can club b_1 such reactions (of type R_1) together as a single event using the batch model. This would automatically reduce the complexity of the system.

3.1.3 Probability of collision calculation for a time-step τ

The probability of collision, p_c, as calculated in Eqn 4.1 will change if τ increases sufficiently. This is because the number of collisions of one molecule of X_1 (under Model 1) with molecules of X_2 in the area ΔV is given by $n_2\frac{\Delta V}{V}$, where $\Delta V = \pi r_{12}^2 U_{12}\tau$ and n_2 is the number of X_2 molecules. We can estimate the number of collisions of the X_1 molecule with a molecule of X_2, Est_{col} for a successful reaction as follows:

$$Est_{col} = p_c.1 + p_c^2.2 + p_c^3.3 + ... = \frac{p_c}{(1-p_c)^2} = n_2(\frac{\Delta V}{V})$$

Solving for p_c from the above quadratic equation and noting that $p_c \leq 1$, we get:

$$p_c = \frac{1 + 2n_2\frac{\Delta V}{V} - \sqrt{1 + 4n_2\frac{\Delta V}{V}}}{2n_2\frac{\Delta V}{V}} \tag{3.13}$$

We can calculate the probability of reaction, p, from this new estimate of p_c as before.

For batch arrivals (Model 2), the estimated number of collisions should be added up for the b_1 molecules of X_1 arriving in a single batch. Thus,

$$Est_{coll}^{batch} = b_1 n_2 \frac{\Delta V}{V} = \frac{p_c}{(1-p_c)^2} \tag{3.14}$$

and hence:

$$p_c^{batch} = \frac{1 + 2b_1n_2\frac{\Delta V}{V} - \sqrt{1 + 4b_1n_2\frac{\Delta V}{V}}}{2b_1n_2\frac{\Delta V}{V}} \tag{3.15}$$

In the numerical results section we shall show that the difference between the adjusted and average (i.e., for the discrete event model) reaction times is minimal as τ grows large. This is due to the higher quantization error in the estimation process. But, we cannot have τ excessively large because that would violate our assumption of only one collision in this time period.

3.1.4 Generalization for other types of reactions

We considered simple reactions of type R_1 where the reaction is activated by single molecules of the reactants. The analysis becomes cumbersome for reactants having more than one molecules participating in the reaction. Nevertheless, such situations can also be modeled with our scheme. Note that, in such cases, only the p_c computation changes. Let us consider the following reaction:

$$R_4 : X_1 + yX_2 \longrightarrow X_3$$

Hence, the probability of collision, p_c for Model 1, and p_c^{batch} for Model 2, can be written as:

$$\begin{aligned} p_c &= \binom{n_2}{y}\frac{\Delta V}{V} \\ p_c^{batch} &= b_1\binom{n_2}{y}\frac{\Delta V}{V} \end{aligned}$$

If more than one X_1 molecule is involved in a reaction, then we can only consider batch arrivals of Model 2. Thus, for reaction R_5 we obtain:

$$\begin{aligned} R_5 &: xX_1 + yX_2 \longrightarrow X_3 \\ p_c^{batch} &= \binom{b_1}{x}\binom{n_2}{y}\frac{dV}{V}, \quad b_1 \geq x \end{aligned}$$

3.1.5 Reactions occurring in the cell membrane or inside the nucleus

Note that the modeling concept presented here can be extended easily to estimate the holding time for membrane reactions and reactions occurring in the nucleus. The only difference between time estimates is governed by the estimates of V. Thus for a reaction occurring inside the nucleus, V will denote the volume of the nucleus. This assumes that the reacting molecules does not come out of the nucleus in course of the reaction event.

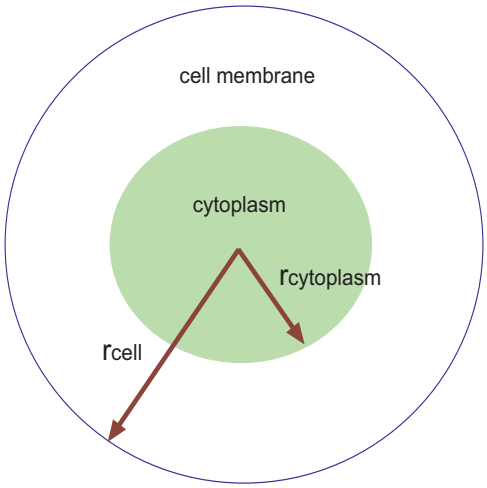

Figure 3.6. Estimation of V for membrane and cytoplasmic reactions.

For membrane reactions the computation of V requires a knowledge of the radius of the cell (r_{cell}) as well as the radius of the cytoplasm denoted by $r_{cytoplasm}$. And, V can be simply estimated assuming the cell to be spherical as:

$$V = \frac{4}{3}\pi[(r_{cell})^3 - (r_{cytoplasm})^3]$$

Clearly, this assumes that the reacting molecules are freely dispersed along the cell membrane. When the molecules are tightly bound to the cell periphery, however, we can use the estimation method outlined in [3].

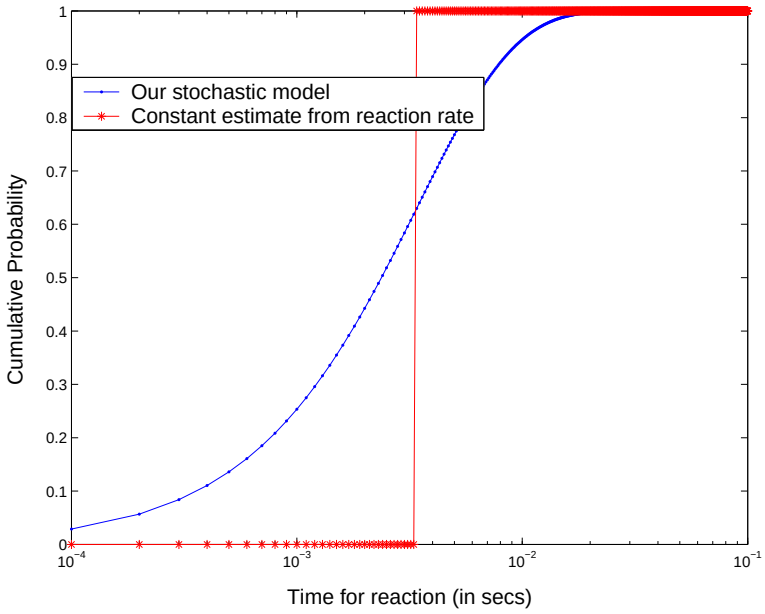

Figure 3.7. Comparison: CDF of Model 1 and rate based equation model.

3.2 Numerical results

In this section, we present the numerical results for our reaction models. After showing the comparisons of our stochastic reaction models with existing rate based equation models, we illustrate the effect of τ on the reaction time. Finally, we show that the average estimate of reaction time (using discrete event serialization) is comparable to the adjusted time of reaction when the number of molecules increase in the system.

3.2.1 Comparison with existing rate based equation model

The rate based model for reactions is a well studied topic. In [38], the authors apply a collision theory based approach to estimate the rate of reaction R_1 *per unit time* and *per unit volume* at absolute temperature T (denoted by $\widetilde{k}(T)$) as:

$$\widetilde{k}(T) = n_1 n_2 r_{12}^2 \sqrt{\frac{8\pi k_B T}{m_{12}}} e^{\frac{-E_{act}}{k_B T}} \tag{3.16}$$

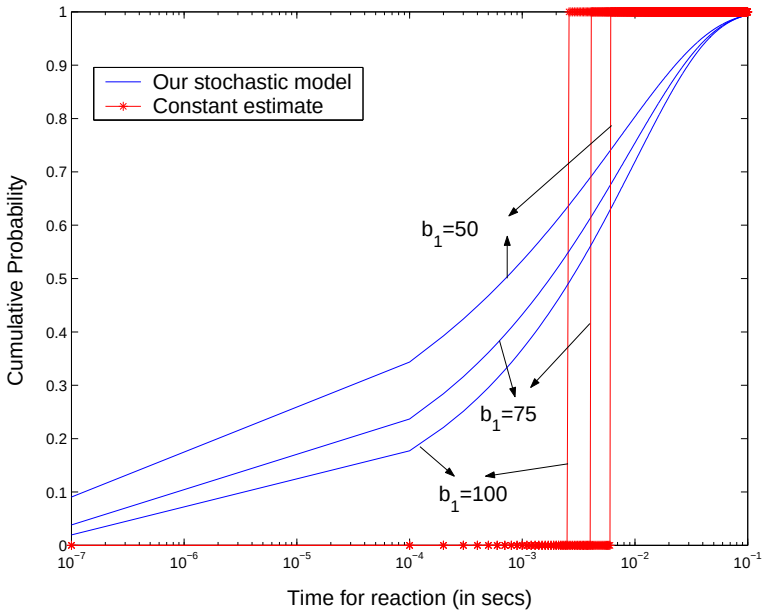

Figure 3.8. CDF of Model 2 vs rate equation model (1200 ATP molecules).

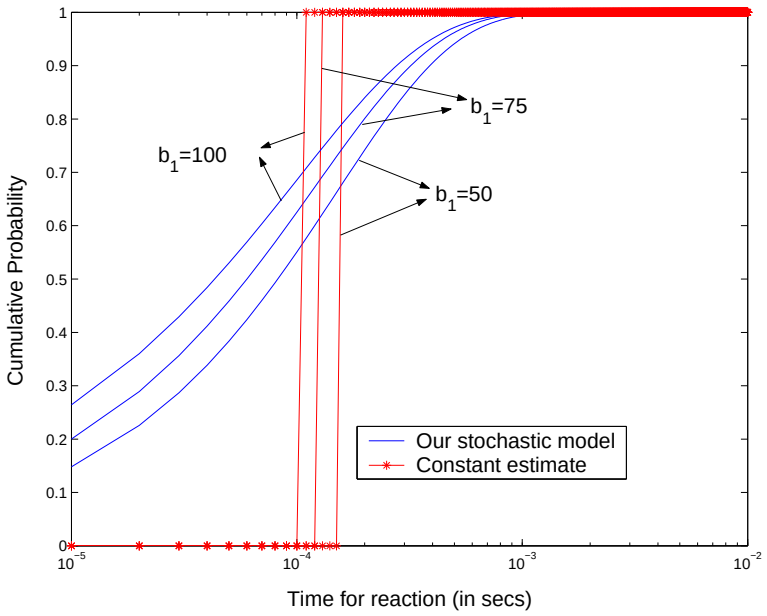

Figure 3.9. CDF of of Model 2 vs rate equation model (1200000 ATP molecules).

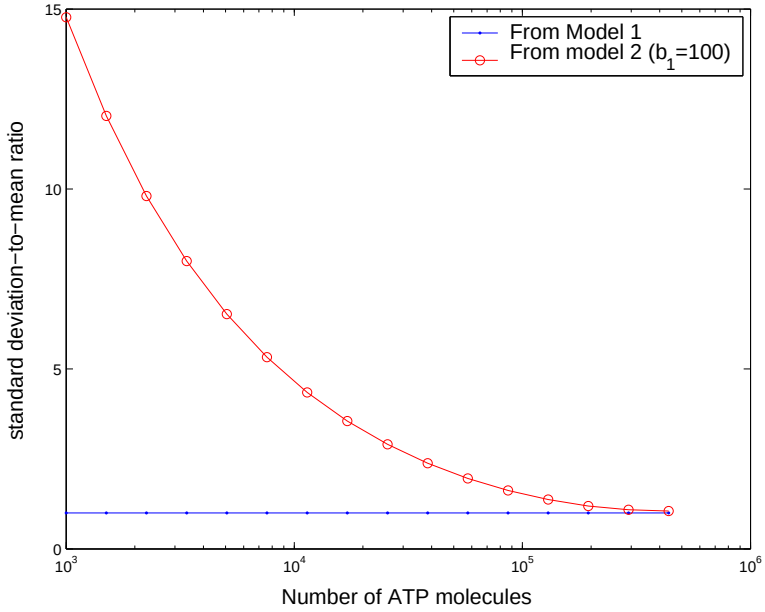

Figure 3.10. Standard deviation to mean ratio vs number of ATP molecules.

Our estimate of T^{DE}_{avg1} can also be written as:

$$T^{DE}_{avg1} = \frac{V}{n_2 r_{12}^2 \sqrt{\frac{8\pi k_B T}{m_{12}}} e^{\frac{-E_{act}}{k_B T}}}; \tag{3.17}$$

Note that, if we compute T^{DE}_{avg1} per unit volume (denoted by $T^{DE/volume}_{avg1}$), we will have $T^{DE/volume}_{avg1} = \frac{1}{\widetilde{k}(T)}$ (we have $n_1 = 1$ as 1 molecule of X_1 enters the cell). This illustrates the validity of our model with the existing rate based model. In particular, we can conclude that the inverse of the reaction rate estimation gives the time required for one reaction of type R_1 in the rate based model, which is *exactly the same as* to the average time for reaction R_1 by a single molecule of type X_1 estimated by our stochastic model. However, the rate constant in the rate based model is a real variable and thus can only return a *constant* time for completion of reaction R_1. But such reactions in the cytoplasm are essentially chaotic [40]. Hence they should be considered as a *stochastic* process and the time required for reaction is actually a random variable.

Table 3.1. Parameter Estimation for Glycolysis reaction

Parameters	Prokaryotic Cell (Salmonella Typhimurium)
V (Volume of the cell)	$4.52 \times 10^{-18} m^3$
r_1 (Radius of Glucose)	4.386 nm
r_2 (Radius of ATP)	0.77 nm
n_2 (number of ATP molecules in the cell)	1200000
E_{act} (Activation energy)	-4.3 kcal/mol
T (Absolute temperature)	300 K
m_1 (mass for Glucose)	180 Dalton
m_2 (mass for ATP)	507.181 Dalton

To generate the numerical results, we consider the glycolysis reaction:

$$R_1 : Glucose + ATP \Longleftrightarrow Glucose6P + ADP \tag{3.18}$$

The corresponding parameters are shown in Table 6.1.

Fig 3.7 plots the cumulative distribution function for the time of reaction R_1 from Model 1 and also that from rate based equations [38]. The time for reaction follows an exponential distribution with mean 0.003422 secs and variance 0.0000113 (note that the standard deviation is nearly equal to the mean). The rate based model, however, gives a constant reaction time of 0.003422 secs. Similar trends are observed for the batch model (Model 2) as depicted in Fig 3.9. In this case, the standard deviation is larger than the mean and hence it is not appropriate to assume any distribution that is based on a single moment (e.g., the exponential distribution). By appropriate Chi-square test, it is possible to fit the mean and standard deviation (for batch model) to a Gamma distribution with appropriate parameters. In the current ODE based stochastic simulations based on Gillespie technique, it is assumed that the reaction rate is constant and only changes to stochastic at lower number of reactant molecules and the distribution is assumed exponential. We have shown the nature of the distribution for different

reaction conditions. Our model indicates that if we consider the chaotic environment of the cell, the reaction time is always stochastic. For single molecule this can be assumed exponentially distributed within the range of interest for cell modeling. But for batch arrival of molecules, a Gamma distribution will be the appropriate model of the reaction time. For the numerical study of batch reactions (Model-2), we consider three batch sizes such as $b_1 = 50, 75$, and 100. The average time for reaction decreases as the batch size increases (because larger the batch size, the larger is the probability of individual reactions in the batch which effectively decreases the average time for any one reaction). For, $b_1 = 50$, the mean is 0.000151 secs and variance is 0.362×10^{-7}. Similarly, for $b_1 = 75$, the mean is 0.000121 secs and variance is 0.264×10^{-7} and for $b_1 = 100$, we have mean of 0.000101 secs and variance of 0.218×10^{-7}. The reaction time from the rate based model however remains constant in all three cases (which is calculated by substituting $n_1 = b_1$ in Eq 3.16). We observe that the constant time for a reaction in the rate-based model is slightly *lower* than the corresponding average time of reaction in the batch model. This is because, the effect of reduction in the probability of reaction for the later reactions in the batch is not considered in the rate-based model.

Fig 3.8 however shows an interesting characteristic for the CDF of Model 2. We consider only 1200 ATP molecules in the system such that the reduction in probability due to the initial set of reactions in the batch of size b_1 is more pronounced. Note that this reduction in probability is because more ATP molecules are being used up by these initial set of reactions. As a result, the time taken for the later reactions in b_1 is more resulting in an overall increase in average reaction time of any reaction in the batch. Hence, the average time estimates increase as the batch size increases. In particular, for $b_1 = 50, 75, 100$, we have mean = 0.007538, 0.009052, and 0.010076 secs respectively. However, for the rate-based equation model, the time for a reaction in a batch of size b_1 is still constant. For different batch sizes, the reaction time for the

rate-based equation model decreases with increase in batch size as normal. However, we observe that the reaction time estimates from rate-based equations are significantly less than that from our batch model as the former does not consider the reduction in probabilities for each reaction in the batch. These effects however, fades off with more number of ATP molecules in the system (typically greater than 1500) for a batch size of 100.

Fig 3.10 plots the standard deviation to mean ratio for the two models presented i.e., $\frac{\sqrt{T^{DE}_{2_{nd}moment1} - (T^{DE}_{avg1})^2}}{T^{DE}_{avg1}}$ (for Model 1) and $\frac{\sqrt{T^{batch/DE}_{2_{nd}moment1} - (T^{batch/DE}_{avg1})^2}}{T^{batch/DE}_{avg1}}$ (for Model 2). We find that for Model 1, the ratio remains constant at 1 for appreciable number of ATP molecules in the system. This corroborates our assumption that the reaction time follows an exponential distribution in most cases. With further increase in the number of ATP molecules (beyond a billion molecules, which is not realistic for many reactions), it starts to decrease resulting in a constant reaction time where the rate equation-based model can be applied. Thus, for fewer number of ATP molecules in the system, our stochastic model gives a better estimate of the reaction time and becomes deterministic (i.e. constant) with further increase in ATP molecules. The plot for the batch model, shows that the ratio is consistently greater than 1 for acceptable estimates of ATP molecules in the cell. This is why we model the reaction time for Model 2 using a Gamma distribution. However, further increase in the number of ATP molecules (beyond a billion molecules) will bring down the ratio to less than 1 and again make the reaction time constant (as given by the rate-based model).

Table 3.2. Parameter estimation for R_1 and R_2 reaction pair in Eq 3.19

Parameters	Estimates
r_1 (Radius of H_2)	1.37 angstrom
r_2 (Radius of O_2)	1.55 angstrom
r_3 (Radius of Cl_2)	1.75 angstrom
E_{act} (same E_{act} assumed for R_1 and R_2)	7 kcal/mol
T (Absolute temperature)	273 K
m_1 (for H_2)	1 gm/mol
m_2 (for O_2)	32 gm/mol
m_3 (for CL_2)	71 gm/mol
b_1 (batch size of O_2 molecules in R_1)	100

3.2.2 Dependence of the reaction time of our stochastic model on τ

Figs 3.11-3.14 plot the performance of our stochastic reaction models. The graphs are generated considering the following simple reactions:

$$R_1 : H + O_2 \rightarrow OH + O$$
$$R_2 : H + Cl_2 \rightarrow HCl + Cl \tag{3.19}$$

The corresponding parameters are depicted in Table 6.2. Note that from our conventions in the previous section, the H molecules are the X_2 type molecules in R_1 and R_2. The reaction time estimates are made *per unit volume*. All the results have been generated assuming 50% overlap between the reactions, i.e., $q = \lceil \frac{T^{DE}_{avg_1} \times 0.5}{\tau} \rceil$.

Fig 3.11 plots the average and adjusted times of reaction for Model 1 against τ. We find that initially they differ, but converge as τ increases. So, our reaction time estimates should consider the time step τ to fall in this range. But τ cannot be increased indefinitely because that would violate our assumption of one collision taking place in one time step. A good estimate of τ would be 10^{-3} secs. Also, we find that the average time of reaction is *independent* of the value of τ. With increasing τ, the probability

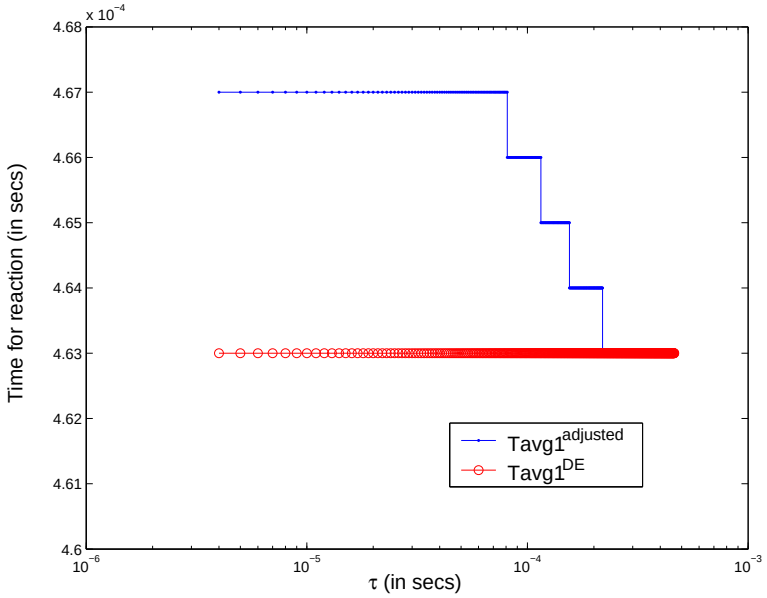

Figure 3.11. Reaction time vs τ for Model 1 .

of reaction increases by an equal amount as the number of collisions required decreases resulting in a constant average reaction time. However, we find that the adjusted reaction time decreases initially before becoming a constant. This is because, with higher τ, the probability of reaction R_1 increases resulting in a lower value for the adjusted time of reaction.

In Fig 3.12, we only show the average time taken for reactions of type R_1 occurring in a single batch of size 100. We can observe that the average time of reaction increases with τ as more time is required for every single collision of the reactant molecules resulting in increased average reaction time. However, the average time becomes constant with $\tau \simeq 10^{-5}$. A point to note is that the adjusted time of reaction should be more than the average time because it captures the actual scenario of reduction in probabilities for conflicting reactions. It should be noted that the batch model requires lower time for reaction than Model 1 because we have calculated the average of the time required to complete all the b_1 reactions corresponding to the b_1 molecules arriving in a single

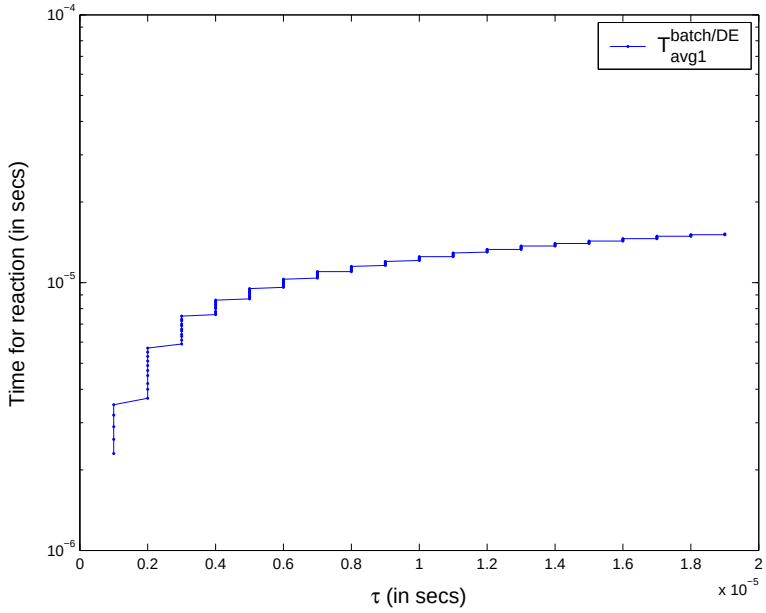

Figure 3.12. Reaction time vs τ for Model 2 .

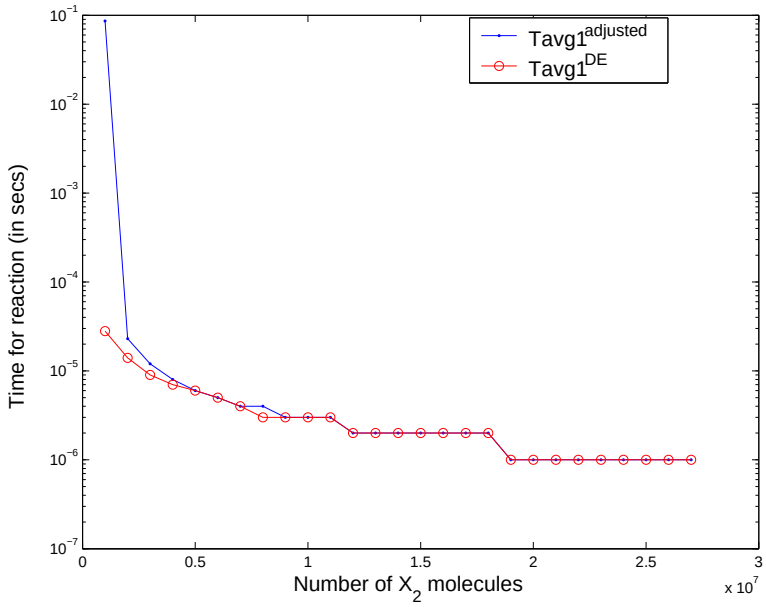

Figure 3.13. Reaction time vs number of X_2 molecules for Model 1 .

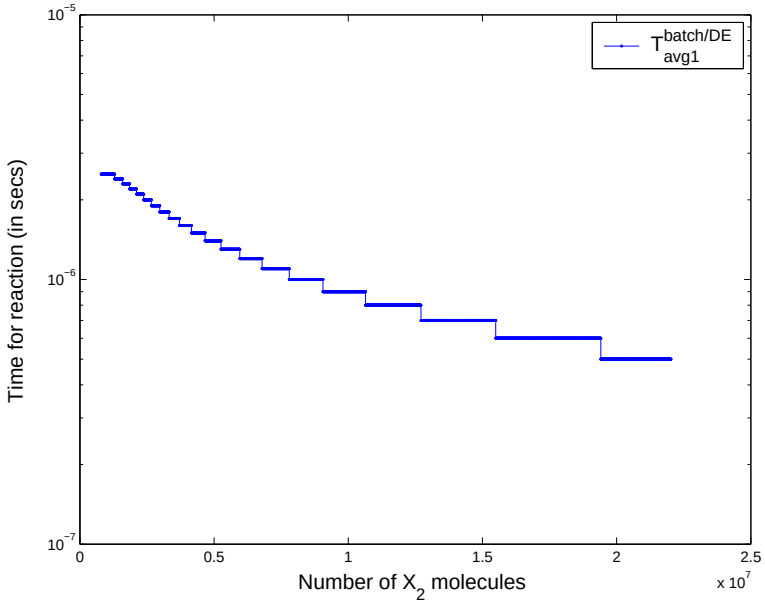

Figure 3.14. Reaction time vs number of X_2 molecules for Model 2 .

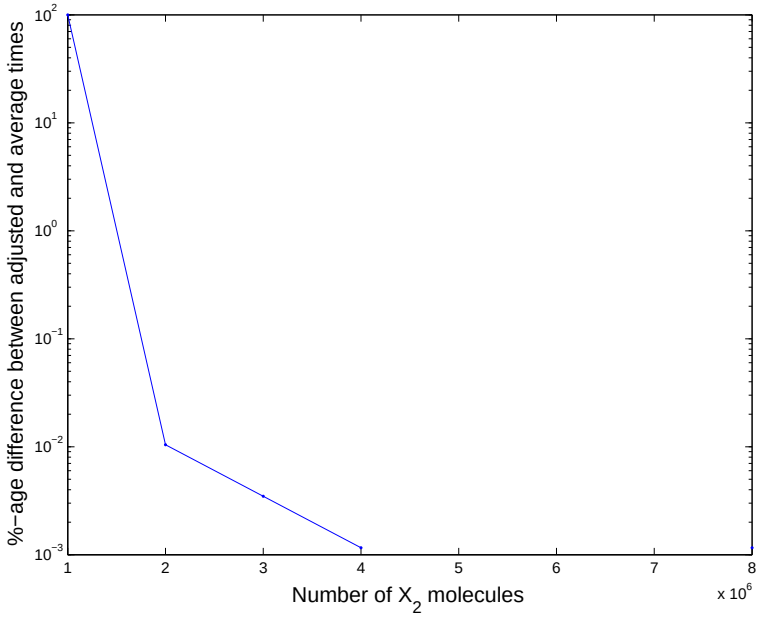

Figure 3.15. Percentage difference between adjusted and actual times of reaction.

batch which intuitively should need less time than a single reaction as in Model 1 as the probability of a single reaction in the batch increases.

3.2.3 Dependence of the reaction time on the number of X_2 molecules

Figs 3.13 and 3.14 plot the average and adjusted times for reaction with increase in the number of X_2 molecules in the system for Model 1 and Model 2. For both plots we find that the average time for reaction reduces with increasing number of molecules for obvious reasons. Initially, there is some difference between the adjusted and average results, but the difference quickly reduces as the number of molecules is increased in the system as illustrated in Fig 3.13. Fig 3.15 plots the difference between the adjusted and average times in percentage. We find that with $\sim 1.5 \times 10^6$ molecules of X_2 the difference becomes negligible. For micro-molar concentrations of the reactant molecules, we find a negligible difference between the adjusted and average results that point to the efficacy of the serialization process in discrete event simulations. Here also, we observe that average time of reaction for Model 2 is slightly lower than that for Model 1. This is because, in the batch model, the effect of increase in time taken for the reaction of the last few molecules of a single batch because of reduction in the number of X_2 molecules is overridden by the increase in probability of each reaction in the batch resulting in an overall lower average time for reaction than Model 1.

3.3 Discussions

3.3.1 Handling delayed reactions

Note that in this chapter, we have equated the time taken to complete a reaction event and the holding time of a discrete reaction event. The underlying assumption is that reactant collisions occur with some probability and once a collision of sufficient energy occurs, a reaction takes place *instantaneously*. Hence, we assume that there is

no holding time of an activated complex. If there is some time delay associated with initiation and completion of the reaction, the probability evolution becomes more complicated [26]. Our reaction models cannot directly handle such delayed reactions which would require comprehensive modeling of the delayed states. Such an attempt was made to model protein-ligand docking events in [73]. However, the reaction models in this chapter were primarily developed for the discrete event simulator that can model delayed reactions easily using the concept of event serialization. Thus, the random reaction time can be convoluted with the random/constant delay to give the total delayed reaction completion time. However, this involves an implicit approximation assuming that the reactant molecules are not available for other reactions once it has entered the present reaction event. But because both the original reaction event time and the delay can be random variables incorporating the probability of successfully completing the delayed reaction event, this approximation should be small. Further analysis is required to study the effect of this approximation.

3.3.2 Limitations of our model

3.3.2.1 Maxwell-Boltzmann distribution of molecular velocities

The Maxwell-Boltzmann distribution gives a good estimate of molecular velocities where we have spatial homogeneity and is widely used in practice. Molecular dynamic (MD) simulation measurements during protein reactions show that the velocity distribution of proteins in the cytoplasm closely match the Maxwell-Boltzmann distribution. However, its application in our model might not give perfect results for cases where the effects of the specific location (e.g. endoplasmic reticulum) of the reactant molecules may violate the assumption of uniform distribution in a volume. Ideally the velocity distribution should incorporate the properties of the reaction space (nucleus/membrane for

reactions occurring in the membrane or nucleus) and the effect on velocity distribution due to the space shape and irregularities. We plan to explore the possibility to improve this velocity distribution by considering the other biological factors that can influence the velocity of the reacting molecules.

3.3.2.2 Activation energy threshold

The activation energy (E_{act}) has been measured for many reactions and we need an estimate of this parameter to be able to predict the nature of the reaction time. The reactions are mostly affected by the amount of energy required to create the different type of bonds used for the reaction. In most cases, however, reactions can be categorized based on the chemical properties of the reactants, and the activation energy for similar kinds of reactions (involving similar reactants) are quite comparable. Comprehensive research results in biochemistry are available and we believe that we should be able to use that knowledge to define this parameter for most of the reactions if we can specify the reaction type.

3.3.2.3 Reverse reactions

We did not consider the reverse reaction conditions in our model because we are interested in the state of single molecules at a time. The state of the molecule can however change through reverse reactions based on the reverse reaction parameters. Say X_1 and X_2 molecules combine to form X_4. The moment the X_4 molecule is formed, a death event of an X_4 molecule can occur at a later time (depending on the reverse reaction parameters), when it decomposes back into X_1 and X_2 molecules. Because of the large number of molecules in the system, and assuming the memoryless property of our simulation technique, we can cause the death of any molecule of X_4 at the occurrence of such death events. During such death events, the system state is changed by killing

any one of the X_4 molecules to create the X_1 and X_2 molecules. The assumption of memory-less property allows for a one-directional reaction model.

3.3.2.4 Reaction neighborhoods

In addition, there is increasing evidence of sub-compartmental (i.e., intra-compartmental localization) in cells, so local neighborhoods of reactions will have higher apparent concentrations than simply the number of molecules divided by the size of the compartment. However, this would require more in depth modeling of the different molecular concentrations inside the cell that reduces the scalability of the simulation framework. Indeed, the Gillespie simulator also fails to address this issue as it requires different rate constants for specific neighborhoods of the reaction type. Nevertheless, our model can be easily extended to incorporate such reaction neighborhoods by limiting the movement of the reactant molecules inside a limited reaction space while computing the probability. However, the applicability of the Maxwell-Boltzmann velocity distribution in the neighborhood requires further research.

3.4 Summary

In this chapter we proposed a model to compute the reaction time for cytoplasmic molecular reactions as a stochastic variable that appropriately reflects the cell environment. The main idea of this modeling is to transform the reaction process from a continuous deterministic process to a discrete random process. This concept allows the transformation of biological reactions to the stochastic domain and make it suitable for discrete event simulation. We have systematically presented the models of different types of reaction situations like single molecule reaction, reaction with a batch of molecules and complex reaction sets sharing common molecule types. In addition, we presented the accuracy impact of reaction serialization that we assume for our discrete event simulation.

We found that the serialization impact on the accuracy of reaction time estimate is minimal when the number of molecules are large in the system which is commonly true for a biological process. This type of random reaction time estimate can differentiate the reaction environments in the cell based on number of molecules available for reaction and the size of the batches that activated the reactions. The average reaction time estimated from this method (for the single molecule model) is exactly same as the reaction rate estimates of kinetic modeling. In addition, we are able to estimate the two moments of the reaction time to capture the stochastic nature of the reaction function. The proposed batch molecule model can significantly reduce the computational complexity when large number of molecules enter a system in a very short time. The discrete modeling framework for biological functions presented here is flexible enough to be extended to create the models for complex biological functions like protein-DNA binding, protein-protein interaction, transcription, translation etc. The stochasticity of the reaction time that is modeled in this chapter is a new metric and current experimental methods are not able to capture this measurement at molecular level. We have shown that the first moment estimate is in agreement with current reaction rate based estimation technique and thus established the validity of the model for the highest moment. At present no experimental data is available to validate the second moment of the reaction time. Newer experiments on single molecule movements in a cell can throw additional light on this aspect of the variance of the reaction time.

CHAPTER 4

PROTEIN-DNA BINDING

This chapter presents a parametric model to determine the execution time of the protein-DNA binding event. Our model considers the actual binding mechanism in conjunction with the approximate model of the protein and DNA structures. We model the effects of thermal and concentration gradients on the binding process using a collision probability. This modeling approach significantly removes the complexity of the classical protein sliding along the DNA model, improves the speed of computation and can bypass the speed-stability paradox. The model produces acceptable estimates of DNA-protein binding time necessary for our event-based stochastic system simulator where the higher order (more than second order statistics) uncertainties can be ignored. The results show good correspondence with available experimental estimates. The parametric nature of the model does not depend on experimentally generated rate constants and permits binding time computation under various conditions. To illustrate the use of this model for "in silico" simulation, we provide the results of the simple model of the protein-DNA binding on gene expression in prokaryotic cells.

This chapter is organized as follows: Section 4.1 discusses some related works on analytical models for protein-DNA binding. Sections 4.2 and 4.3 presents our stochastic model for protein-DNA binding. Section 4.4 reports the results for a few sample transcription factors for human and bacterial cells and also some results from the discrete event simulator that we have built for validating the model. Finally, in Section 4.5 we summarize the findings of this chapter.

4.1 Background on protein-DNA binding models

The transcription factors (TFs) bind DNA at specific sites to initiate the complex transcription machinery of cells. Upon binding to the site, the TF forms a stable protein-DNA complex that can either activate or repress transcription of nearby genes, depending on the actual control mechanism. In this chapter, we focus on models for both bacterial and eukaryotic TFs by assuming that the structure, location on chromatin and other details of cognate (target) sites on the DNA are known from existing experimental data. Such problems of specific binding and binding rates also arise in the context of oligonucleotides-DNA binding [2]. In the proposed model, we do not include the effects of chromatin remodeling and histone modifications [21].

Vast amounts of experimental data available these days provide the structures of protein-DNA complexes at atomic resolution in crystals and in solution [66, 19, 20], binding constants for dozens of native and hundreds of mutated proteins [5, 100], calorimetry measurements [84] and novel single-molecule experiments [67]. Based on these experimental data, a conceptual basis for describing both the kinetics and thermodynamics of protein-DNA interactions was first presented in [80, 69, 81, 79]. The classical model of protein-DNA sliding based on the experimental data reported in [66, 19, 20], however, is quite complicated. The problem faced by the sliding mechanism, if the energetics of protein-DNA interactions are taken into account, is outlined in [60], where the authors introduce a quantitative formalism for protein-DNA interaction.

4.1.1 Protein sliding model along the DNA

The existing TF-DNA binding model involves a combination of both three-dimensional (3-d) and one-dimensional diffusion (1-d) of the TF. The total search process can be considered as a 3-d search followed by binding to the DNA and a round of 1-d diffusion. The TF, upon dissociation from the DNA, continues on a 3-d diffusion until it binds

at a different place on the DNA. The 1-d diffusion along the DNA proceeds along the rough energy landscape of the DNA. A quantitative analysis of the search process in [60] reported the following:

1. The diffusion along the DNA becomes prohibitively slow when the roughness of the binding energy landscape is at least $2k_BT$.
2. The optimal energy[1] of nonspecific binding to the DNA provides the maximal search rate. However, even the optimal combination of 1-d and 3-d diffusions cannot achieve experimental estimates of binding time when the roughness of the landscape is at least $2k_BT$. In the optimal regime of search, the protein spends equal amounts of time diffusing along nonspecific DNA (i.e, 1-d diffusion) and diffusing in the solution (i.e, 3-d diffusion). A fairly smooth landscape (with roughness of the order of k_BT) is required for the 1-d diffusion to achieve experimentally observed and biologically relevant rates.
3. Stability of the protein-DNA complex at the target site requires considerably larger roughness than k_BT where rapid search is impossible, leading to the speed-stability paradox. In fact, the minimal roughness as reported in [60] is $5k_BT$ given a genome size of 10^6 bps. A search-and-fold mechanism for the DNA-binding proteins is proposed in [60] to resolve the paradox.

[1]While the TF diffusion along the DNA is controlled by the specific binding energy (i.e., energy required for the TF to bind to a particular DNA sequence), the dissociation of the TF from the DNA depends on the total binding energy (i.e., on the non-specific binding as well as on the specific one). Moreover, since the dissociation events are much less frequent than the hopping between neighboring base-pairs, the non-specific energy makes a larger contribution to the total binding energy.

4.2 DNA-Protein binding model

We partition this problem into two biological microevents: (i) Collision of the protein molecule to a binding site ($\pm B$) on the DNA surface, i.e., we assume that the TF can slide a distance of B (in either direction) on the DNA before binding and (ii) a protein colliding with DNA at the binding site ($\pm B$) will bind only if it hits it with enough kinetic energy to overcome the energy barrier of the site.

4.2.1 Modeling the first microevent: calculating p_n

In this section we abstract the first microevent by computing the probability, p_n, of collision of the protein (TF) with the binding site ($\pm B$) on the DNA. From the principles of collision theory for hard spheres, we model the protein molecule as a rigid sphere with diameter d and the TF binding region of the DNA as a solid cylinder with diameter D and length $L+2B$ (Fig 4.1). Note that the $2B$ factor is incorporated as the TF can slide in either direction on the DNA.

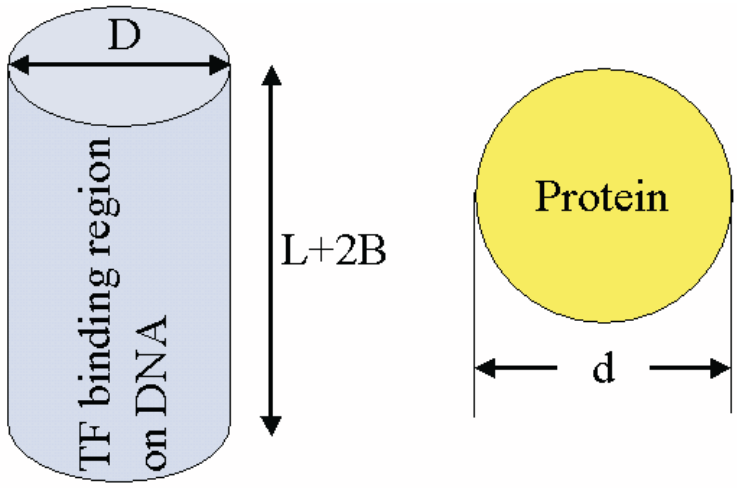

Figure 4.1. Schematic diagram: protein molecule and TF binding region of DNA.

We define our coordinate system such that the DNA is stationary with respect to the protein molecule. This assumption allows the TF to move towards the DNA with relative velocity U. The protein molecule moves through space to sweep out a collision cross section, C. The number of collisions during a time period Δt is determined when a protein molecule will be inside the space created by the motion of the collision cross section over this time period due to the motion of the protein molecule.

4.2.1.1 Average surface area of collision between a sphere and cylinder

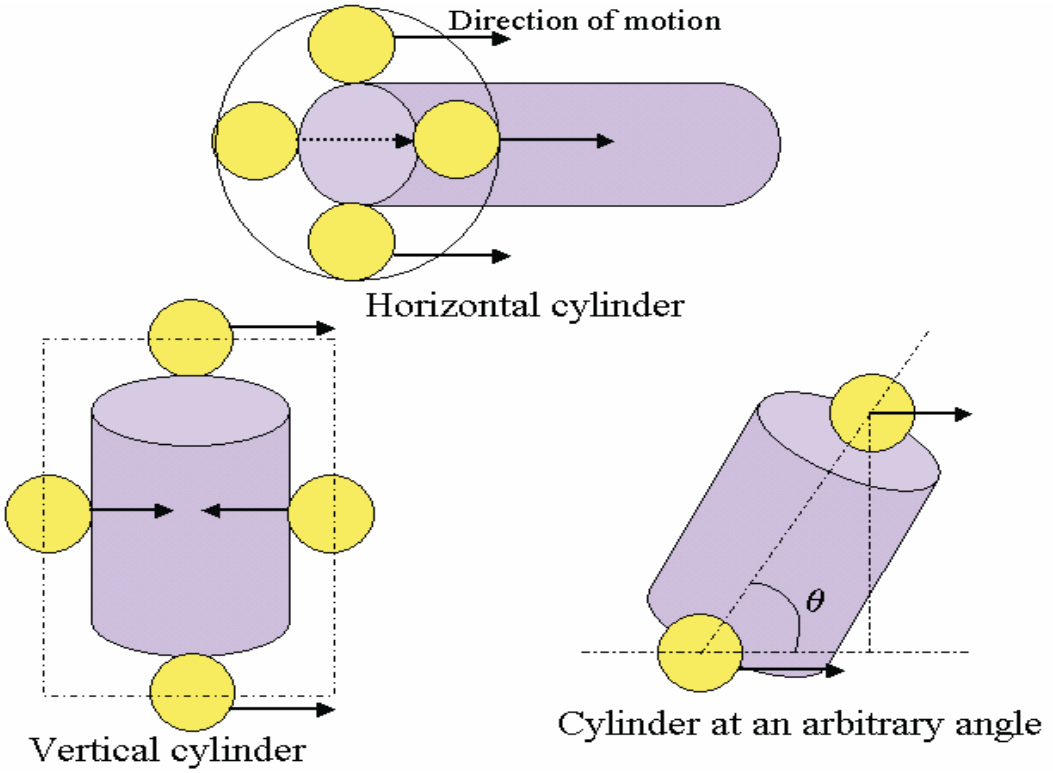

Figure 4.2. Collision of spherical protein and cylindrical DNA TF binding region.

The spherical protein molecule during its motion can encounter the DNA binding sites in three different configurations (1) horizontal cylinder, (2) vertical cylinder

and (3) cylinder at an arbitrary angle, θ, with the direction of motion of the protein (Fig 4.2). For the horizontal cylinder model, the cross-sectional area of collision traces out a circle, whereas for the vertical cylinder model, it is a cylindrical in shape. The third case can be derived from the vertical cylinder model considering a cylindrical collision area of length $(L+2B+d)\sin\theta$. Thus, the cross-sectional area of collision, C, is given by:

$$C = \left\{ \begin{array}{ll} \pi\frac{(d+D)^2}{4}, & \text{for} \quad \theta = 0^0 \\ (L+2B+d)(D+d), & \text{for} \quad \theta = 90^0 \\ (D+d)(L+2B+d)\sin\theta, & \text{otherwise} \end{array} \right\}$$

Thus for any arbitrary θ $(0^0 < \theta < 90^0)$, we can express the cross-sectional area of collision as a function of θ as follows: $C(\theta) = (D+d)(L+2B+d)\sin\theta$.

Note that the border conditions $(\theta = 0^0, 90^0)$ constitute a set of measure zero and for all practical purposes, the whole calculation can be limited to the case where $0^0 < \theta < 90^0$. We assume a uniform density for the occurrence of the different θ's in the range $0^0 \leq \theta \leq 90^0$, i.e. having density $\frac{\theta}{(\pi/2)}$. It is to be noted that ideally θ can take any value in $0^0 \leq \theta \leq 360^0$, but our working range of $0^0 ... 90^0$ suffices for all these cases. Thus the average cross-sectional area, C_{avg}, can be expressed by:

$$C_{avg} = \int_0^{\frac{\pi}{2}} \frac{2}{\pi} C(\theta) d\theta = \frac{2}{\pi}(D+d)(L+2B+d).$$

Note that $\theta = 0^0$ disappears from consideration but we can argue that the probability of that happening is too small to change the expression for C_{avg} significantly. This cross-section C_{avg}, moves in the cytoplasmic space (nucleus for eukaryotes) to create the collision volume for a particular binding site.

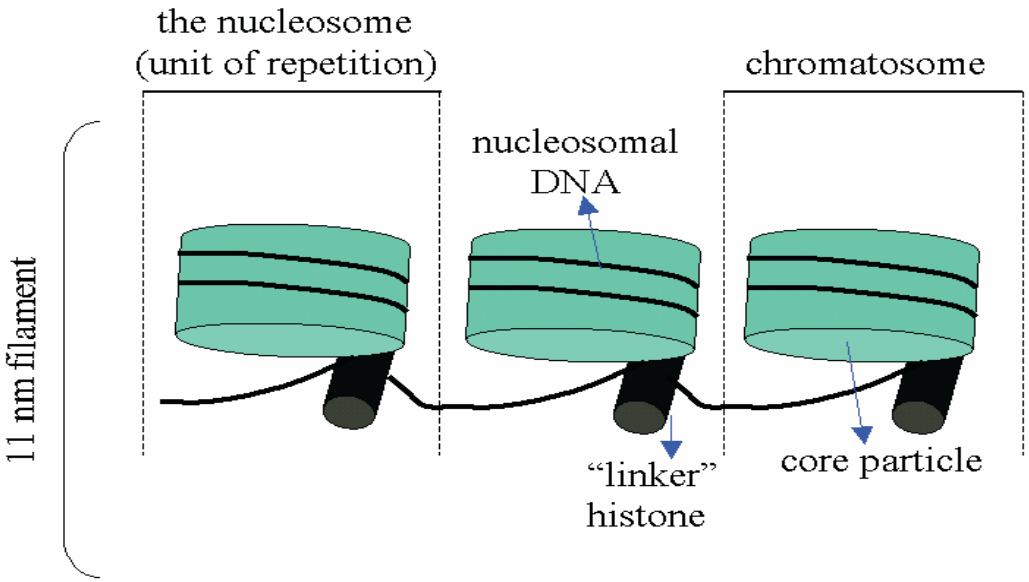

Figure 4.3. DNA packing through nucleosomes.

4.2.1.2 Probability of protein-DNA binding in eukaryotic cells

Fig 4.3 simplistically illustrates how DNA is packed along different cylindrical nucleosomes. We do not include chromatin remodeling and histone modification in the current model as discussed in Section 5.5.1. Thus, L in the expression for C_{avg} denotes the length of the TF binding region and D the diameter of the DNA strand (assumed cylindrical in shape) on a nucleosome cylinder. As single or multiple motifs [23] can be present for a gene in the promoter region, the value of L is adjusted to reflect those conditions. Now, we can have three cases based on where the TF binding region is located on the DNA:

1) Case I: The region entirely lies within the DNA portion on a nucleosome cylinder;

2) Case II: The region lies entirely within the DNA portion that is outside the nucleosome cylinders;

3) Case III: The region is shared between the DNA on a nucleosome cylinder and that outside it.

We analyze each of these case in the following:

Case I: Let the probability that the protein molecule hits the correct nucleosome cylinder given it collided with the DNA with sufficient energy be p_h^c. We have:

$$\begin{aligned} p_h^c &= \frac{\text{length of that nucleosome cylinder}}{\text{length of all nucleosomes + length of all stretches}} \\ &= \frac{l_n}{N_1 l_n + \sum_{i=1}^{N_2} l_s^i} \end{aligned}$$

where, l_n denotes the length of a nucleosome cylinder (assumed fixed for all the cylinders), l_s^i denotes the length of the i^{th} stretch of DNA, i.e., the length of DNA present in between the i^{th} and $(i+1)^{th}$ nucleosome cylinders. Here, N_1 and N_2 denote the number of nucleosome cylinders and that of stretches of DNA respectively. Now, the probability, p_d, of hitting the DNA portion of the nucleosome cylinder, can be estimated from the surface area of the nucleosome cylinder and that of the DNA present in the cylinder as follows:

$$p_d = \frac{\pi D l_d}{\pi D l_d + \pi d_n l_n}$$

where, l_d is the length of the DNA present inside the cylinder and d_n is the diameter of the nucleosome cylinder. Because the DNA is known to make 1.65 turns in a nucleosome cylinder, we have $\frac{l_d}{l_n} = 1.65$. Let, p_f^c designate the probability of colliding with the TF binding region ($\pm B$) in the DNA, given that the protein molecule already collided with the DNA with enough energy and also hit the correct nucleosome cylinder. We have:

$$p_f^c = \frac{(\text{length of TF binding region in the DNA}) + 2\text{B}}{\text{total DNA length in that particular nucleosome}}$$

Also, the particular motif of the colliding protein molecule is of interest to us, as it should come in proximity of the TF binding region ($\pm B$) of the DNA for a binding to occur. So, we need to calculate the probability, p_m, of identifying the motif of the colliding protein molecule, as follows:

$$p_m = \frac{\text{length of the motif region of the protein}}{\text{total length of amino acid chain of the protein}}$$

Thus, the total probability of collision of the TF to the DNA binding site ($\pm B$) is given by:

$$p_n = p_m \times p_h^c \times p_f^c \times p_d$$

Now, because the DNA is wrapped around a particular nucleosome cylinder, some part of it will not be available for the TF to bind to. Thus C_{avg} as calculated above is not entirely available to the TF to bind to. Nucleosomes themselves are stable and show limited mobility. The dynamic characteristics are due to action of nucleosome-modifying and remodeling complexes that restructure, mobilize and eject nucleosomes to regulate access to the DNA. We approximate the impact of this complex process currently through a difficulty parameter α, which denotes the *percentage availability in average collision cross-sectional area*. This parameter represents approximately the percentage of the time the hidden DNA surface is made visible for reaction through histone remodeling (we are currently working on a separate model of histone remodeling to compute this parameter). Thus, the effective cross-sectional area, C_{eff}, available for TF binding can be calculated as follows: $C_{eff} = \alpha \times C_{avg}$.

Case II: In this case, the probability, p_h^s, of hitting the correct stretch of DNA in between the nucleosome cylinders is given by:

$$p_h^s = \frac{l_s^i}{N_1 l_n + \sum_{i=1}^{N_2} l_s^i}$$

where we assume that the TF binding site is located in the i^{th} stretch of DNA. Similarly, let p_f^s designate the probability of colliding with the TF binding region ($\pm B$) in the DNA similarly as before. We have:

$$p_f^s = \frac{\text{(Length of TF binding region on DNA) + 2B}}{\text{total DNA length in that particular stretch}}$$

and, the total probability of collision of the TF to the DNA binding site denoted by p_n is given by:

$$p_n = p_m \times p_h^s \times p_f^s$$

In this case, the entire TF binding region in the DNA is available for the binding process to occur and we have: $C_{eff} = C_{avg}$.

Case III: Because the TF binding region ($\pm B$) is shared between a nucleosome cylinder and an adjoining stretch, the probability calculations become complex for this case. We approximate the calculations in the following way. Suppose the TF binding site ($\pm B$) is shared between the i^{th} nucleosome cylinder and the j^{th} stretch of DNA. Because the cylinder and the stretch has to be side by side, we must have either $j = i$, or $i = j + 1$ depending on whether the first part of the TF binding site is in the cylinder or in the stretch respectively. Let p_w^c and p_w^s denote the probabilities of hitting the TF binding portion in the cylinder and that in the stretch respectively. In this case however, p_f^c and p_f^s computations should change as follows:

$$\begin{aligned} p_f^c &= \frac{\text{(length of TF binding region portion in nucleosome) + B}}{\text{total length of DNA in that particular nucleosome}} \\ p_f^s &= \frac{\text{(length of TF binding region portion in the stretch) + B}}{\text{total length of DNA in that particular stretch}} \end{aligned}$$

Hence we have:

$$p_w^c = p_m \times p_h^c \times p_f^c \times p_d; \quad p_w^s = p_m \times p_h^s \times p_f^s; \quad p_n = p_w^c + p_w^s$$

Thus p_n is the total probability of collision of the TF to the DNA binding site ($\pm B$). Furthermore, the average cross-sectional area calculations become a little different in this case. We break up C_{avg} into C_{avg_1} and C_{avg_2} based on the length of the TF binding region (L_1) in the nucleosome cylinder and that in the adjoining stretch (L_2). We assume for

simplicity that the TF binding region is shared between one stretch and one nucleosome cylinder only, because this region is generally quite small in length compared to the length of the DNA packed inside a nucleosome cylinder. However, if the region is extended to more than one nucleosome cylinder or stretch, we can handle that case in a similar fashion. Thus the effective cross-sectional area of binding is represented as:

$$C_{eff} = \alpha \times C_{avg_1} + C_{avg_2}$$

Thus the total probability, p_n, of collision to one specific TF binding region can be calculated easily for each of the three cases discussed above. But we need to know how exactly the DNA is packed in the nucleosome cylinders to determine p_n and the effective surface area (C_{eff}) required for binding. In particular, we assume that the DNA packing structure in nucleosome cylinders is fixed and hence we can find where the TF binding region is located as described in Cases I, II and III.

4.2.1.3 Approximate mechanism to find the TF binding region

Nucleosomes have 1.65 turns of DNA and a diameter, d_n, of 11 nm. Thus the length of DNA inside a nucleosome cylinder can be approximated as $1.65 \times \pi \times d_n$, where πd_n is the circumference of the nucleosome cylinder. We assume that all the nucleosome cylinders have identical shape and number of turns of DNA in them. We also assume that all the stretches of DNA between nucleosome cylinders are equal in length. Thus, length of DNA in a stretch can be approximated as $(\frac{T_D - N \times (1.65 \times \pi \times d_n)}{N-1})$, where T_D is the total length of the DNA and N is the number of nucleosome cylinders present. The denominator is due to the assumption that there can only be $(N-1)$ stretches of DNA present in between the N nucleosome cylinders. From the complete genomic sequence, we can find out the exact position of the TF binding region along with its length. Thus

we can approximately estimate whether the TF binding region corresponds to Case I, II or III.

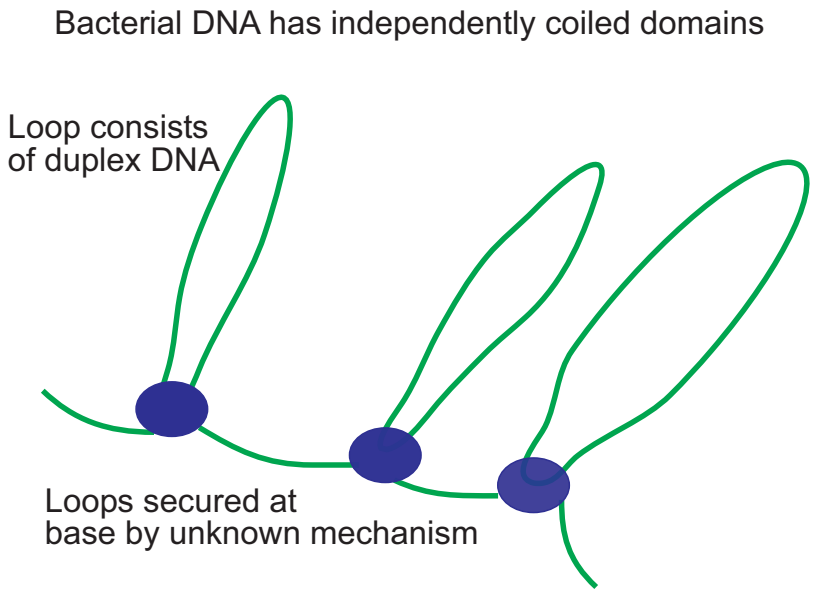

Figure 4.4. Bacterial Genome Structure.

4.2.1.4 Protein-DNA binding probability for bacterial cells

The bacterial genome is supercoiled with a general organization as depicted in Fig 4.4 [12]. Each domain consists of a loop of DNA, the ends of which are secured in some way. Hence, the total probability of collision in this case is simply approximated as:

$$p_n = p_m \times p_w; \text{ where } p_w = \frac{\text{length of TF binding region} + 2\text{B}}{\text{total length of the DNA}}$$

Since the entire surface area of the DNA is available for binding, the effective cross-sectional area of binding is given by: $C_{eff} = C_{avg}$

4.2.2 Modeling the second microevent: calculating p_b

Let p_b denote the probability that the TF collides with the DNA with enough kinetic energy such that it can bind to the DNA. In time Δt, the TF sweeps out a volume ΔV such that:

$$\Delta V = C_{eff} U \Delta t$$

Now, the probability of the protein molecule being present in the collision volume ΔV is $p_P = 1$ given that one protein molecule arrived to create a collision volume of ΔV.
The probability of the DNA being present in an arbitrary uniformly distributed ΔV in the total volume, V is given by $p_D = \frac{\Delta V}{V}$. Note that the prokaryotic cells do not have a nucleus and hence V denotes the total volume of the cell; for eukaryotic cells, however, V will denote the volume of the nucleus.

Thus, probability of the protein molecule to collide with the DNA during time Δt is:

$$p_c = p_P \times p_D = \frac{\Delta V}{V} = \frac{C_{eff} U \Delta t}{V} \tag{4.1}$$

We next assume that the colliding protein molecule must have free energy of at least E_{Act} to bind to the specific DNA transcription factor binding region. This kinetic energy will be required for the rotational motion of the protein molecule such that all the binding points in the protein molecule come close to those in the DNA for the binding to take place successfully. The kinetic energy of approach of the protein towards the DNA with a velocity U is $E = \frac{m_{PD} U^2}{2}$, where $m_{PD} = \frac{m_P . m_D}{m_P + m_D}$ is the reduced mass, m_P is the mass (in gm) of the protein molecule, and m_D is the mass (in gm) of the DNA. It is to be noted that we consider the mass of the entire chromosome and not just the TF binding site of the DNA. This is because the entire chromosome has to undergo rotational motion for the binding process. We also assume that as the kinetic energy, E, linearly

increases above E_{Act}, the number of collisions that result in binding also increases. Thus, the probability for a binding to occur because of sufficient kinetic energy of the protein molecule is given by:

$$p_r = \left\{ \begin{array}{c} \frac{E-E_{Act}}{E}, \quad \text{for } E > E_{Act} \\ 0, \quad \text{otherwise} \end{array} \right\} \tag{4.2}$$

and the overall probability, p_o, for collision with sufficient energy is given by:

$$p_o = p(binding, Collision) = p_r \times p_c = \left\{ \begin{array}{c} p_c \frac{(E-E_{Act})}{E}, \quad \text{for } E > E_{Act} \\ 0, \quad \text{otherwise.} \end{array} \right\}$$

The above equations assume a fixed relative velocity U for the reaction. We will use the Maxwell-Boltzmann distribution of molecular velocities for a species of mass m given by:

$$f(U,T)dU = 4\pi(\frac{m}{2\pi k_B T})^{3/2} e^{\frac{-mU^2}{2k_B T}} U^2 dU$$

where $k_B = 1.381 \times 10^{-23}\ kg\, m^2/s^2/K/molecule$ is the Boltzmann's constant and T denotes the absolute temperature (taken as 273^0 K). Replacing m with the reduced mass m_{PD} of the protein molecule and DNA, we get

$$f(U,T)dU = 4\pi(\frac{m_{PD}}{2\pi k_B T})^{3/2} e^{\frac{-m_{PD}U^2}{2k_B T}} U^2 dU \tag{4.3}$$

The term on the left hand side of the above equation denotes the fraction of this specific protein molecule with relative velocities between U and $(U + dU)$. Summing up the collisions for the protein molecule for all velocities, the probability (p_b) of collision with sufficient energy is obtained as follows:

$$p_b = \int_0^\infty p_o f(U,T)dU$$

Recalling that $E = \frac{m_{PD}U^2}{2}$, i.e., $dE = m_{PD}UdU$ and substituting into Eqn. 4.3, we get:

$$f(U,T)dU = 4\pi(\frac{m_{PD}}{2\pi k_B T})^{3/2} \frac{2E}{U m_{PD}^2} e^{\frac{-E}{k_B T}} dE$$

Thus we get:

$$p_b = \int_{E_{Act}}^{\infty} \frac{(E - E_{Act})4C_{eff}\Delta t}{Vk_BT}\sqrt{\frac{1}{2\pi k_BTm_{PD}}}e^{-\frac{E}{k_BT}}dE$$

$$= \frac{C_{eff}\Delta t}{V}\sqrt{\frac{8k_BT}{\pi m_{PD}}}\, e^{\frac{-E_{Act}}{k_BT}} \tag{4.4}$$

4.2.3 Total binding probability considering different binding regions

Ideally, for any protein molecule, we can have more than one TF binding regions on the DNA. Let G be the number of different TF binding regions on the DNA for the specific TF colliding with the DNA. Also, let p_t^i denote the total probability of binding (combining the first and second microevents) for the i^{th} TF binding region ($1 \leq i \leq G$). Note that the probabilities of the first and second microevents as calculated above will depend on the specific binding site i on the DNA under consideration. We denote these two probabilities as p_n^i and p_b^i for the i^{th} site that can be calculated similarly as shown above. In general, all the binding sites corresponding to a particular TF are identical making $p_n^i = p_n^j$ and $p_b^i = p_b^j$, for $i \neq j$, and $1 \leq i, j \leq G$. Hence,

$$p_t^i = p_n^i \times p_b^i$$

Thus if p denotes the actual probability of binding of the protein with any of these G different regions, we have:

$$p = \sum_{i=1}^{G}[p_t^i \prod_{j=1, i\neq j}^{G}(1 - p_t^j)]$$

This is because the probability of binding to the first TF binding region is given by $p_t^1 \prod_{j=2}^{G}(1 - p_t^j)$; that for the second region is $[p_t^2(1 - p_t^1)(1 - p_t^3)(1 - p_t^4)...(1 - p_t^G)]$; and so on. The total probability, p, is the sum of all these individual cases.

4.3 Time taken for protein-DNA binding

We next estimate the time taken to complete the binding with total binding probability, p. Let $\Delta t = \tau$ be an infinitely small time step. The protein molecules try to bind to the DNA through collisions. If the first collision fails to produce a successful binding, they collide again after τ time units and so on. Note that now we can have a TF-DNA binding in two ways: (a) the TF directly collides and binds to the DNA binding site or (b) the TF collides at a distance ($\leq B$ bps) and slides on the DNA to bind to the site. The average binding time computation requires a probability assignment to these two events. Let *per* denote the probability that the binding occurs due to collision only (point (a) above). Hence, binding occurs with collision and sliding with probability $(1 - per)$. Note that $per = 1$ simplifies to the case where the protein does not slide along the DNA at all, and $per = 0$ boils down to the model in [60] where it is assumed that the TF slides along the DNA at every round. In [60], the authors derived the 1-d diffusion time, τ_{1d}, along the DNA using the mean first passage time (MFPT) from site 0 to B as follows:

$$\tau_{1d}(B) \simeq B^2 e^{\frac{7\sigma^2}{4(k_BT)^2}} (\nu)^{-1} (1 + \frac{\sigma^2}{2(k_BT)^2})^{-\frac{1}{2}}$$

where ν is the effective attempt frequency for hopping to a neighboring site and σ is the roughness of the DNA landscape in units of k_BT. Here τ_{1d} *considers the different energy barriers* on the DNA that the TF has to overcome while sliding whereas E_{act} *is required for the actual binding to the cognate site.* Therefore, the total probability of binding is:

$$p_{binding} = p_{no-sliding}(1 - p) + p(1 - p_{no-sliding}); \text{and, } p_{no-sliding} = |p|_{B=0}$$

where $p_{no-sliding}$ denotes the probability of binding when the sliding along the DNA is not considered altogether. Hence, the average time for protein-DNA binding model (i.e., the first moment) is given by:

$$T_1 = \quad p_{binding}(per \times \tau + (1 - per)(\tau + \tau_{1d}))$$

$$+ \quad (1-p_{binding})p_{binding} \times 2(per \times \tau + (1-per)(\tau + \tau_{1d}))$$

$$+ \quad (1-p_{binding})^2 p_{binding} \times 3(per \times \tau + (1-per)(\tau + \tau_{1d})) + ...$$

$$\Rightarrow T_1 = \frac{(per \times \tau + (1-per)(\tau + \tau_{1d}))}{p_{binding}}$$

The second moment of the binding time is given by

$$T_2 = \frac{(2-p_{binding})(per \times \tau + (1-per)(\tau + \tau_{1d}))^2}{(p_{binding})^2}$$

When no sliding is considered, we find that the time for DNA-protein binding follows an exponential distribution for most ranges of E_{act} (reported in the next section). Moreover, since τ is assumed to be quite small, we can approximate the total time measurements of binding using a continuous (exponential in this case) distribution instead of a discrete geometric distribution. The average time T_1 as calculated above gives the estimated time for protein-DNA binding in bacterial cells. For eukaryotic cells, we should add the average protein transport time from the cytoplasm to the nucleus that can be computed from any standard diffusion model.

4.4 Results and analysis

4.4.1 Problems in validation of our model

Before presenting the numerical results, let us first discuss the difficulty of experimentally validating our model. We compute the average time for protein-DNA binding in this chapter. On the other hand, existing experimental results are based on estimation of the binding rate of any specific TF to the DNA. The experimental estimate of $1 \sim 10$ seconds (secs) is reported from this rate measurement [60]. Hence, the time taken by a TF to bind to the DNA site depends on the number of TFs in the cell. However, our model computes the time taken by any particular TF to bind to the DNA which should be independent of the number of TFs in the cell. It is certainly very difficult to carry

out experiments to track a particular TF and physically compute the time. Also, the stochastic nature of the binding process suggests that the distribution of the time taken will have a very high variance. In other words, in some cases the TF requires time in milliseconds whereas in other cases it might take as long as 100 seconds. The results we present in this section assume that the time taken for any particular TF-DNA binding is $1 \sim 10$ secs even though it is not a true estimate of this event because it is not a molecular level measurement.

4.4.2 Numerical results for $per = 1$ (i.e. no TF sliding is considered)

In this section, we present the numerical results for the theoretical models derived in the chapter. Figs 5.9-4.8 present the results for the PurR TF (having 35 binding sites) on the Escherichia coli (*E. coli*) chromosome. Similarly, Figs 4.9-4.10 illustrate the behavior for eukaryotic cells where we considered the average human cell with 20 μm diameter and the Htrf1 DNA-binding protein. The different parameters assumed for the numerical results are concisely presented in Table 5.1. We used the EcoCyc database [43] for the *E. coli* data and the PDB database [42] for human cell data.

4.4.2.1 Results for prokaryotic cells

Fig 5.9 plots T_1 against different values for Δt. The average time for DNA-protein binding remains constant initially and shoots up exponentially with increasing Δt. The same characteristics are seen for different activation energies, $E_{act} = 10\ k_BT$, $15\ k_BT$ and $20\ k_BT$. The activation energy estimates follow from the change in free energy related to binding that includes the entropic loss of translational and rotational degrees of freedom of the protein and amino acid side chains, the entropic cost of water and ion extrusion from the DNA surface, the hydrophobic effect, etc. as discussed in [35]. The smaller the required E_{act}, the larger is p_b for the protein molecules and hence the smaller is T_1. Note

Table 4.1. Parameter Estimation for Bacterial (pertaining to PurR TF in *E. coli*) and Eukaryotic (pertaining to Htrf1 TF in human) Cells

Parameters	Prokaryotic Cell (from [43])	Eukaryotic Cell (from [42])
V (volume)	$4.52 \times 10^{-18} m^3$ (of cell)	$4.187 \times 10^{-16} m^3$ (of nucleus)
Length of DNA	4.64×10^6 bp (*E. coli*)	3×10^9 bp (*Human cell*)
G (number of binding regions)	35 (for PurR)	35 (assumed for Htrf1)
Length of TF binding site (L)	26	48
Length of protein amino acid chain	341 (for PurR)	53 (Htrf1)
Length of protein motif	26 (for PurR)	48 (Htrf1)
Radius of Amino acid chain	1 nm (for PurR)	1 nm (Htrf1)
Average radius of the protein ($\frac{d}{2}$)	5 $\mathring{A}$ (for PurR)	5 $\mathring{A}$ (Htrf1)
m_P	38.175 Dalton (for PurR)	6635 Dalton (for Htrf1)
Diameter of DNA (D)	2 nm (for *E. coli*)	2 nm (*Human cell*)
Mass of DNA (m_D)	3×10^6 Dalton (*E. coli*)	1.9×10^{12} Dalton (*Human cell*)

that p_b as calculated above also corresponds to the number of collisions in time Δt of the protein molecule with the DNA. And, for our assumption of at most one collision taking place in Δt to hold, we have to make sure that $0 \leq p_b \leq 1$ (this is also true because p_b is a probability). Thus the *regions to the right of the vertical lines* corresponding to each E_{act} plot denotes the forbidden region where $p_b > 1$ even though $0 \leq p \leq 1$. This gives us an estimate of the allowable Δt values for different E_{act}'s such that T_1 indeed remains constant. With increasing Δt, the time taken for successive collisions between the TF and DNA increases, resulting in an overall increase in the average binding time. However, with $\Delta t \leq 10^{-8}$, T_1 remains constant for each E_{act}.

Fig 4.6 plots T_1 against the different possible E_{act} estimates. It shows that the average time for binding increases with increasing E_{act} values. As E_{act} increases, more

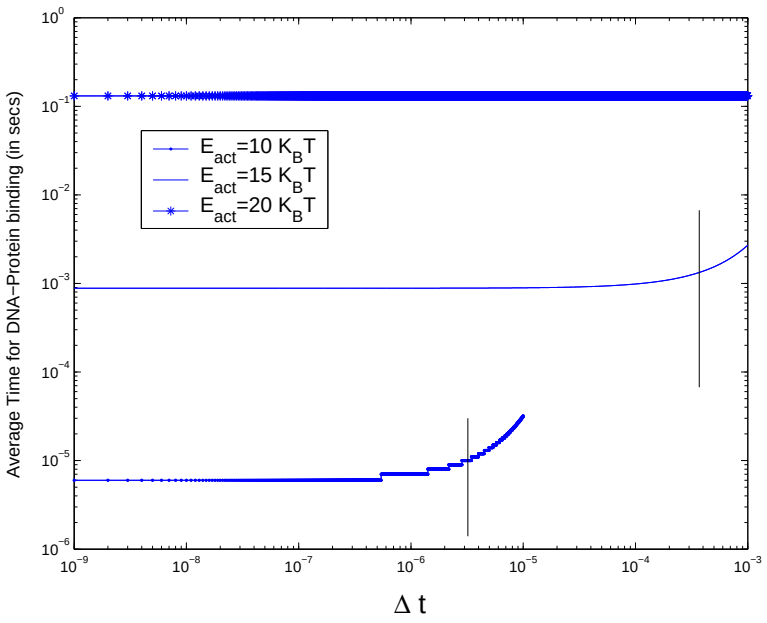

Figure 4.5. Average TF-DNA binding time (T_1) against increasing Δt for *E. coli*.

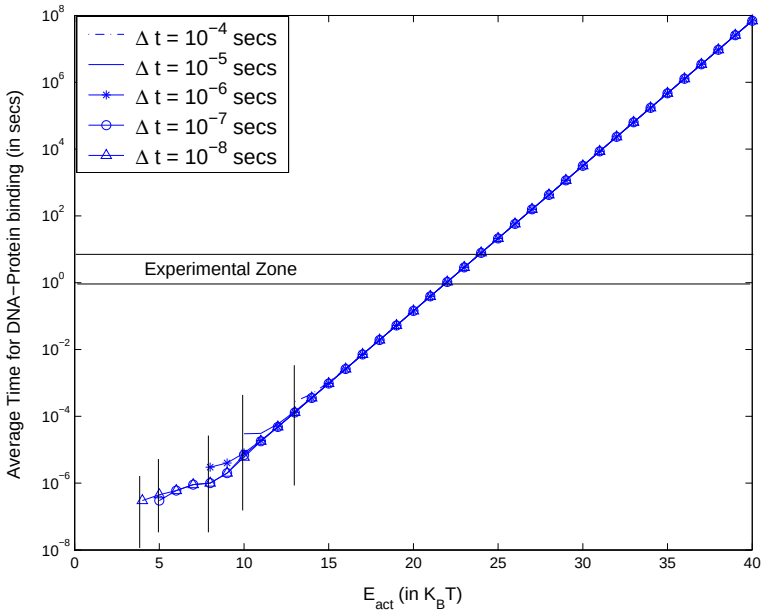

Figure 4.6. T_1 against increasing E_{act} for *E. coli*.

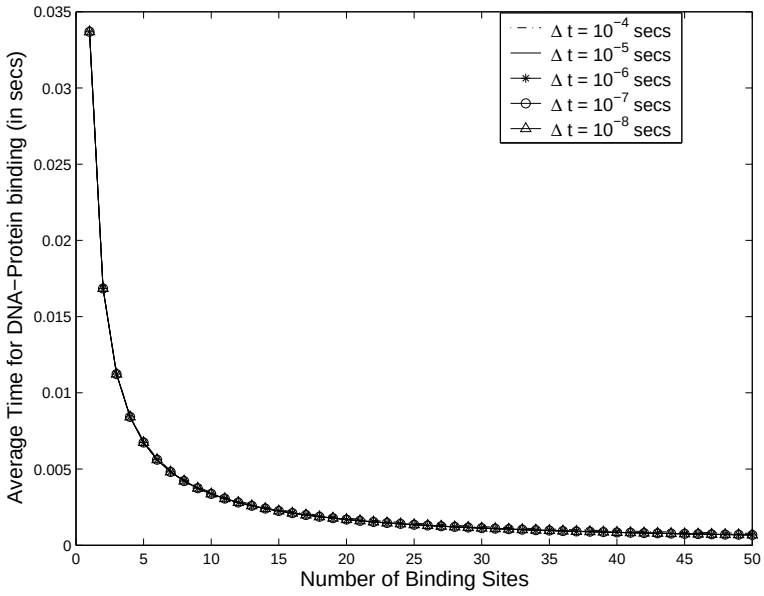

Figure 4.7. T_1 against increasing number of binding sites for *E. coli*.

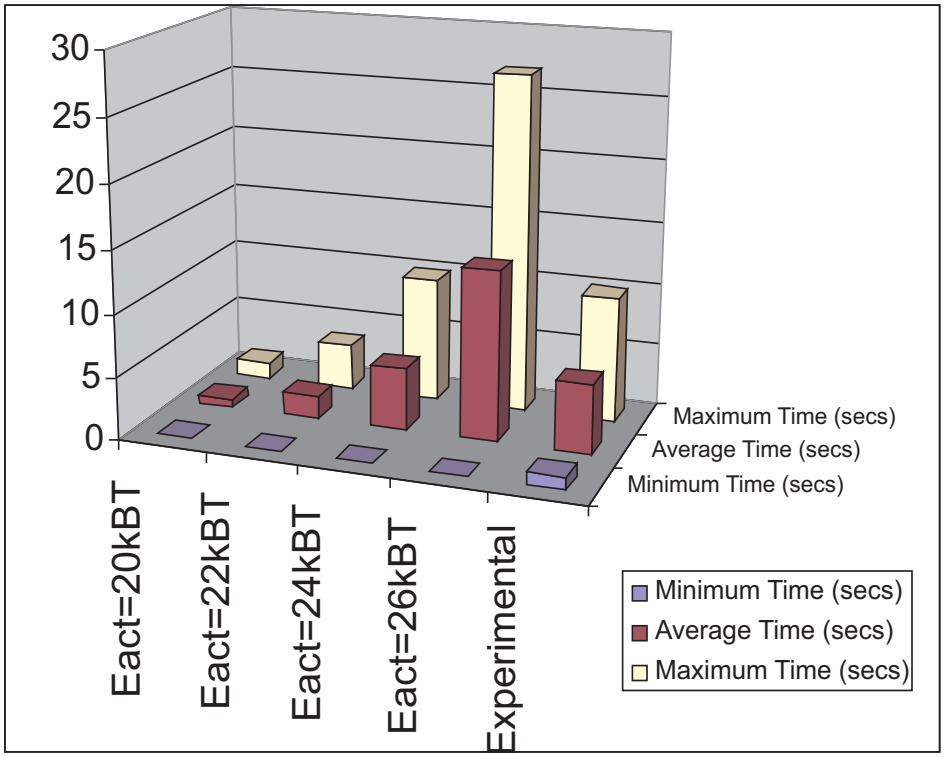

Figure 4.8. Comparison of T_1 with experimental results.

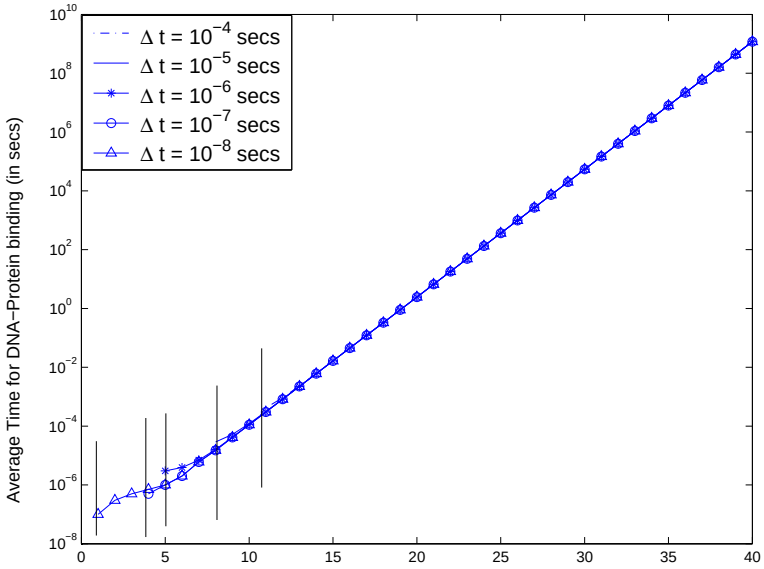

Figure 4.9. T_1 against E_{act} for eukaryotic cells.

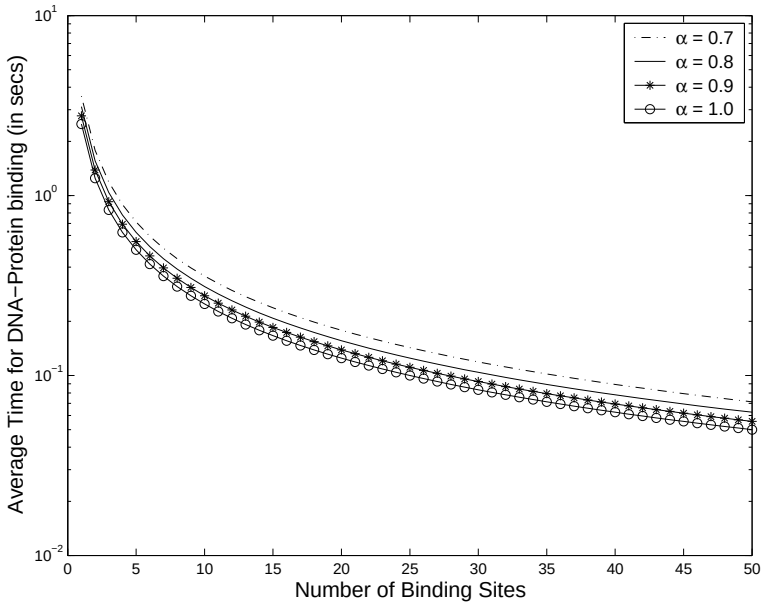

Figure 4.10. T_1 against different α's in eukaryotic cells.

kinetic energy is required by the TFs to achieve stable binding, and only higher molecular velocities can produce that energy. Hence p_b decreases resulting in an overall increase in T_1. However, for very low E_{act}, the binding times tend to increase because the TFs actually has to spend more time to bind to a DNA site due to low kinetic energy requirement. Another interesting feature is that T_1 remains the same for different estimates of Δt as long as $0 \leq p_b \leq 1$. As discussed before, *the regions to the left of the vertical lines* denote the forbidden regions where $p_b > 1$. The speed-stability paradox [60] says that for acceptable average time estimates we should have $\sigma \sim k_B T$, whereas for stable binding we need $\sigma \geq 5k_B T$. Our results show that we can achieve stable binding between $E_{act} = 1k_B T$ for $\Delta t = 10^{-8}s$ and $E_{act} = 13k_B T$ for $\Delta t = 10^{-4}s$. The minimum possible values for E_{act} for different Δt's are reported in Table 4.2. The average time for TF-DNA binding is experimentally measured [60] to be $1 \sim 10s$, which is achieved with $E_{act} \simeq 20k_B T$. Fig 4.8 gives the comparison between the experimental results and our theoretical estimates. We find that for $20k_B T \leq E_{act} \leq 26k_B T$, our results match with the experimental values. The minimum and maximum times for binding reported in the figure for different E_{act} values are calculated assuming 95% confidence interval. Thus, our theoretical model also gives an estimate of the activation energy required for stable binding. It should be noted that E_{act} refers to the total free energy change due to binding and should be higher than σ as calculated in [60]. We also find that in the range $20k_B T \leq E_{act} \leq 26k_B T$, the time of binding follows an *exponential distribution* (as the calculated mean is very close to the standard deviation). In Fig 5.10, we find that T_1 decreases as the number of binding sites G is increased which is again logical as the protein molecules now have more options for binding.

Table 4.2. Allowable E_{act} values against Δt such that $0 \leq p_b \leq 1$

Δt (in secs)	Minimum E_{act} (in k_BT)
10^{-4}	13
10^{-5}	10
10^{-6}	7.6
10^{-7}	5
10^{-8}	1

4.4.2.2 Results for eukaryotic cells

Fig 4.9 shows similar trends for eukaryotic cells. The T_1 values for eukaryotic cells are higher than those for bacterial cells mainly because the volume of the nucleus is larger than the average volume for prokaryotic cells. Also, α decreases the probability of binding appreciably as the DNA is arranged in nucleosome cylinders, thereby reducing the average surface area for collision and hence reducing p_b. Also, the p_d component of p_t results in lesser values of p_t for eukaryotic cells and hence greater values for T_1. Fig 4.10 shows the dependence of T_1 on α. With smaller α, the value of C_{eff} is smaller, and hence T_1 is higher. It can be observed that α does not significantly affect the average time for binding.

Figs 5.10,4.9,4.10 were generated with $E_{act} = 15\ k_BT$. For eukaryotic cells, we consider the average time for binding after the TF has diffused inside the nucleus. Thus, the overall time for DNA-protein binding has to consider the time taken by protein molecules for diffusion. This has been extensively studied and not reported here.

4.4.2.3 Important observations from the $per = 1$ results

1. Our model achieves the experimental estimate of $1 \sim 10$ secs with activation energy in the range: $20k_BT \leq E_{act} \leq 26k_BT$ for prokaryotic cells (obviously the results are

generated for the PurR TF in *E. coli* and we have not tested this range for other TFs as yet). The corresponding range for eukaryotic cells has not been reported here because we need to know the corresponding experimental estimates for human cells.

2. The stochastic nature of protein-DNA binding time can be approximated by an exponential distribution in this range as the observed values for mean and standard deviation of the binding time are comparable.
3. The average time for DNA-protein binding increases for higher E_{act}.
4. The DNA-protein binding time is independent of the value of Δt. The recommended value of Δt is 10^{-8} secs. Figs 5.9-4.6 show the dependence of the average time on Δt and E_{act}. We find that a wider range of E_{act} is available (keeping $p_b \leq 1$) with lesser Δt. The same estimate holds true for eukaryotic cells also.
5. The average time decreases as the number of DNA binding sites increase because the TF has more sites to bind to.
6. The average time is not significantly affected by α, i.e., the percentage availability of average collision cross-sectional area.

Fig 4.11 plots the cumulative distribution function (CDF) for the time of binding with $E_{act} = 22k_BT$ for *E. coli*. Figs 4.12 and 4.13 respectively show the dependence of T_1 on Δt and the number of binding sites for eukaryotic cells.

4.4.3 Validation of DNA replication with no-sliding assumption

We used another model validation exercise having robust measurement data. We build the DNA replication model of *E. coli* that provides the gross measurement data of a large number of DNA nucleotide/protein interaction sequences. We also build the analytical model from the micro-scale DNA nucleotide/protein interaction times to copy the DNA.

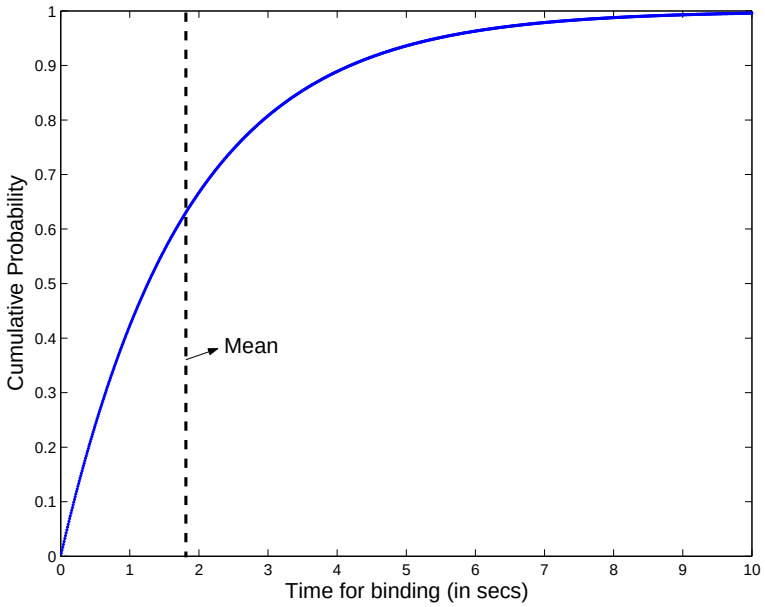

Figure 4.11. CDF of DNA-protein binding time ($E_{act} = 22k_BT$, $\Delta t = 10^{-8}$) in *E. coli*.

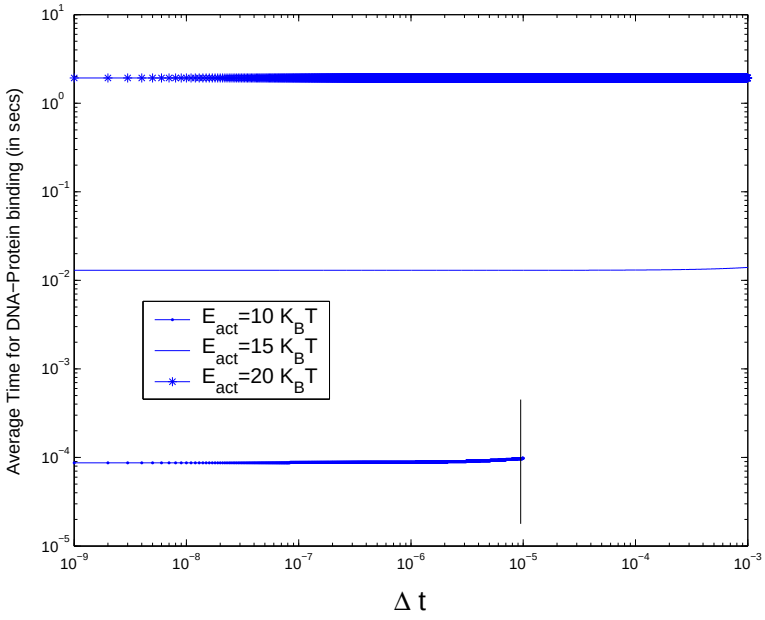

Figure 4.12. T_1 measurements with increasing Δt for eukaryotic cells.

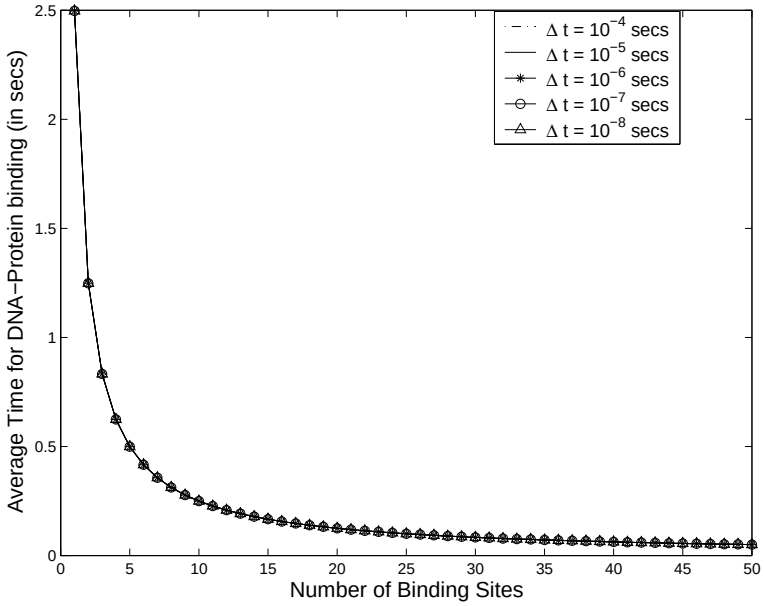

Figure 4.13. Average Time against increasing number of binding sites for eukaryotes.

Table 4.3 presents the parameters used to compute the total time taken for DNA replication using the TF-DNA binding model as the base model and assume that the TFs never slide on the DNA ($per = 1$). We assumed that (i) the rate of replication is the same in both leading strand and lagging strands and (ii) replication stops at the position directly opposite to the OriC in the chromosome. We now estimate the individual time delays in each step of DNA replication mechanism.

1) *Binding of DnaA, initiation proteins, with DNA at OriC:* We consider length of the replication as 245 bps [7] and 20 molecules of DnaA proteins bind with DNA [8] one after another at OriC. The total time delay for the whole process will be 20 times the time taken to bind one molecule of DnaA with DNA at OriC. With $E_{act} = 20k_BT$, $T = 273K$, $L = 245bps$ and $1bp = 0.34 \times 10^{-9}m$, the time taken for a DnaA is 0.133 secs. Hence, the time taken for 20 molecules of DnaA is $S_1 = 2.6565$ secs.

2) *Binding of DnaB (Helicase) with DNA double helix at OriC:* DnaB binds with the complex formed by DnaA molecules. Two molecules of DnaB enzymes will be required

Table 4.3. Parameter Estimation for DNA replication in *E. Coli*

Component	Mol. Weight (Dalton)	Radius (Angstroms)
DnaA	52574	24.5 (Stokes radius)
DnaB	9551	24.5 (assumed)
DnaG	68001	24.5 (assumed)
DNA Poly III holoenzyme	900000.4	60 (Stokes radius)
SSB	140000	45 (Stokes radius)
ATP	507	7.7

for one of the two replication forks. We ignored the role of DnaC (another enzyme that helps loading the DnaB with the complex), since the loading function is not known clearly. The time taken for a DnaB is computed as 0.18 secs. Hence, the time taken for 2 molecules of DnaB is $S_2 = 0.36$ secs.

3) *Binding of DnaG (Primase) with initiation complex:* A molecule of DnaG binds with the complex formed after the previous step. For the two replication forks, two DnaG enzymes will be used. Hence the total time delay is twice the time taken to bind one molecule of DnaG with the complex. We compute the time taken for a DnaG as 0.15 secs. Hence, the time taken for 2 molecules of DnaG is $S_3 = 0.3$ secs.

4) *Binding of DNA polymerase III holoenzyme (Polymerase) complex with replication formed after step 3 in the DNA double helix:* 2 DNA polymerase III holoenzymes are required for the two replication forks. Hence the time delay for this step is twice the time taken for binding one molecule of DNA Polymerase III holoenzyme. We compute the time taken for a DNA Poly Holoenzyme with DNA = 0.363 secs. Hence, the time taken for 2 molecules of DNA Poly Holoenzymes with DNA is $S_4 = 0.726$ secs.

5) *Unwinding of DNA by Helicase by hydrolyzing* 1 *ATP molecule:* Helicase unwinds the double stranded DNA by hydrolyzing ATP and the rate of unwinding is 3 bps by hydrolyzing one ATP molecule to ADP. We compute the time taken for unwinding 3 nucleotides = 0.002736 secs. And, time taken to unwind 33 nucleotides, $S_5 = 0.7$ secs.

Table 4.4. E_{act} and *per* requirements for $n = 100bps$

σ (in k_BT)	E_{act} (in k_BT)	per
5	$20 - 26$	1.0
4	$20 - 26$	1.0
3	$11 - 15$ or $20 - 26$	$0.1 - 0.9$ or 1.0
2	$14 - 17$ or $20 - 26$	$0.1 - 0.9$ or 1.0
1	$20 - 24$ or $20 - 26$	$0.1 - 0.9$ or 1.0

6) *Coating of ssDNA with SSB protein for stabilizing replication process:* We assume one SSB molecule covers ~ 33 nucleotides in the ssDNA [9]. SSB proteins are required in both leading as well as lagging strands. These proteins are continuously attached with ssDNAs before the new DNA strand is synthesized and attached. We compute the time taken for coating 3 nucleotides as 0.002736 secs. Hence, the time taken to coat 33 nucleotides is $S_6 = 0.7$ secs.

7) *Synthesis of new DNA by DNA polymerase III holoenzyme:* DNA polymerase III synthesize new DNA at the rate of 3 nucleotides [10] by hydrolyzing 1 ATP molecule to ADP. The time taken to synthesize 3 nucleotides by DNA Poly III holoenzyme is computed as 0.00275 secs. Therefore, the total time required for the complete DNA is 35.403 min.

Adding the time delays from each of the above steps, the total time required for DNA replication in *E. Coli* from our model is ~ 36 mins which is quite close to the experimental estimate of 42 mins.

4.4.4 Numerical results for the combined model in E. coli ($per \neq 1$)

In [60], the authors presented an experimental estimate of τ_{1d} for different values of sliding distance (denoted by $\overline{n}$) and at different roughness σ for the PurR TF of *E.Coli*

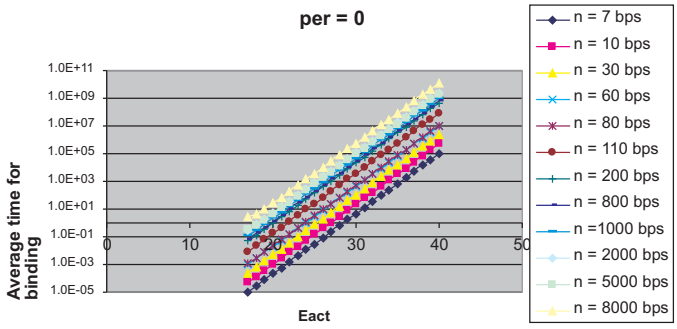

Figure 4.14. Average binding time for purR ($\sigma = 1k_BT$).

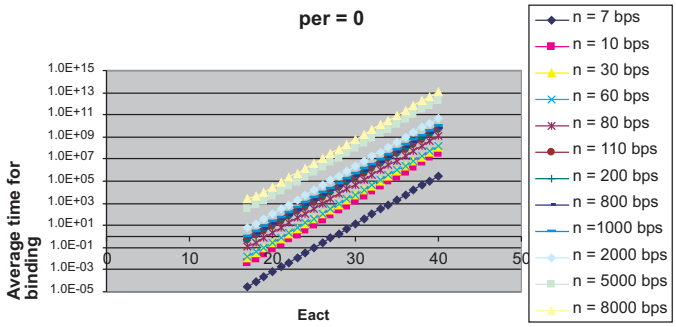

Figure 4.15. Average binding time for purR ($\sigma = 2k_BT$).

with a random and uncorrelated energy profile having standard deviation $\simeq 6.5k_BT$. These τ_{1d} estimates have been used to generate the plots.

Figs 4.14-4.18 plot T_1 for $\sigma = 1, 2, 3, 4, 5\ k_BT$ respectively with $per = 0$ and different values of the sliding distance, n, in bps. The x-axis gives the values for E_{act} and the y-axis is plotted on a logarithmic scale with $E \pm z = 10^{\pm z}$. Note that the average binding time estimates increase with increasing σ.

For $\sigma = 1k_BT$ and $per = 0$, the experimental estimates of $1 \sim 10$ secs can be achieved with $15k_BT \leq E_{act} \leq 20k_BT$, even with $n = 8000bps$. However, the experimental results can be achieved up to $(n = 2000bps, \sigma = 2k_BT)$, $(n = 200bps, \sigma = 3k_BT)$, $(n = 20bps, \sigma = 4k_BT)$ and $(n = 7bps, \sigma = 5k_BT)$. Thus if we assume that every collision of the TF with the DNA is accompanied with a 1-d diffusion, the average number

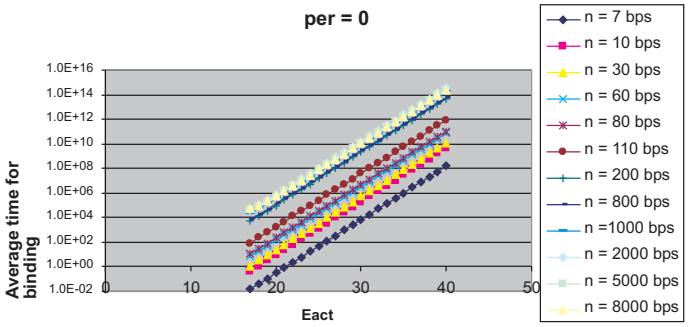

Figure 4.16. Average binding time ($\sigma = 3k_BT$).

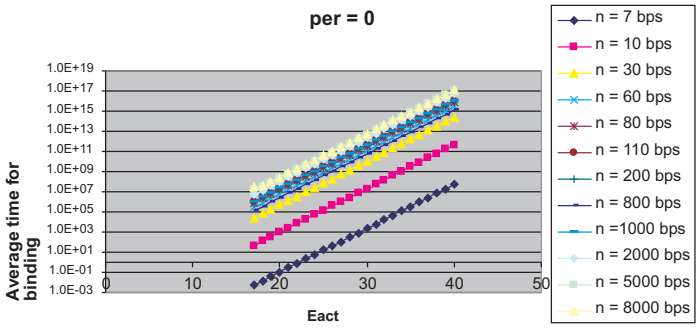

Figure 4.17. Average binding time ($\sigma = 4k_BT$).

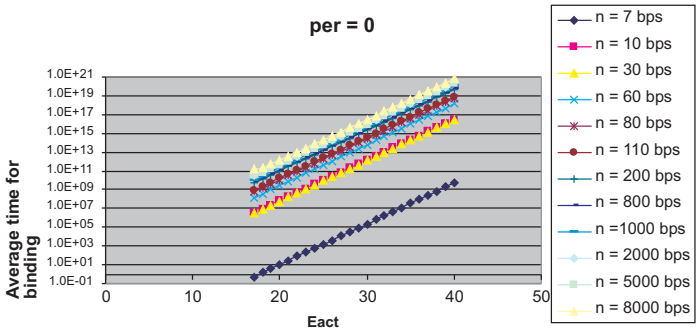

Figure 4.18. Average binding time ($\sigma = 5k_BT$).

Table 4.5. E_{act} and *per* requirements for $n = 50bps$

σ (in k_BT)	E_{act} (in k_BT)	per
5	$20 - 26$	1.0
4	$20 - 26$	1.0
3	$12 - 15$ or $20 - 26$	$0.1 - 0.9$ or 1.0
2	$20 - 24$ or $20 - 26$	$0.1 - 0.9$ or 1.0
1	$22 - 25$ or $20 - 26$	$0.1 - 0.9$ or 1.0

of base pairs that the TF can slide is only 7 bps when $\sigma = 5k_BT$. This is certainly a very low estimate and it is logical to assume that *not every TF-DNA collision involves 1-d diffusion.*

The next step is to find an estimate of *per* ($\neq 0$), that gives binding times in the experimental range even with biologically relevant amounts of sliding. In [60], the authors reported the optimal number of base-pairs that can be searched at $\sigma = 1k_BT$ as 100 bps. We report the maximum σ that can achieve the experimental estimates from our results in Table 4.4 and that for 50 bps in Table 4.5. Thus we can get the bounds on E_{act}, for different combinations of per, σ and n. The above results show the maximum value of σ for which the experimental rate can be achieved. However, for $\sigma = 5k_BT$, we have to consider either $per = 1.0$, i.e., *the TF does not slide on the DNA*, or it can *slide a maximum of 7 bps.*

4.4.5 Simulating the dynamics of protein-DNA binding

In this section, we analyze the dynamics of the protein-DNA binding event at a "systems level" - studying its effect in association with other molecular events involved in a cellular process. In particular, we focus on the effects of TF binding event on the expression of genes in prokaryotic cells.

Stochasticity in prokaryotic gene expression has been extensively studied, both mathematically [50],[41], as well as in experimental systems [53],[51]. Particularly, the *burstiness* in protein production i.e., proteins are produced in random bursts of short duration rather than in a continuous manner, have been shown in single cell experiments conducted on the *lacZ* gene in *E. Coli* [53],[51]. 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 RNA Polymerase (RNAP) - 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. The fine-grained regulation of gene expression by the transcriptional machinery, specifically, transcription initiation frequency controlled by the binding of the transcription factor (TF) upstream of the promoter region, has been quantitatively studied in [4],[27],[92].

In order to quantitatively study the stochastic dynamics of TF-DNA binding on prokaryotic gene expression, we build a discrete-event based simulation environment, as outlined in [88], capturing the key molecular events involved in the process [2]

- 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 is defined by its first and second moments, $\bar{R}_{mR}$ and σ_{mR} respectively, and the time between two transcription events is represented by the random variable $\tau_{transcription}$. This event encompasses the micro-event of TF-DNA binding and includes the average binding time mathematically captured in T_1 and T_2.

[2]The details of the stochastic models for prokaryotic transcription and translation, together with the simulation framework are available in [90]

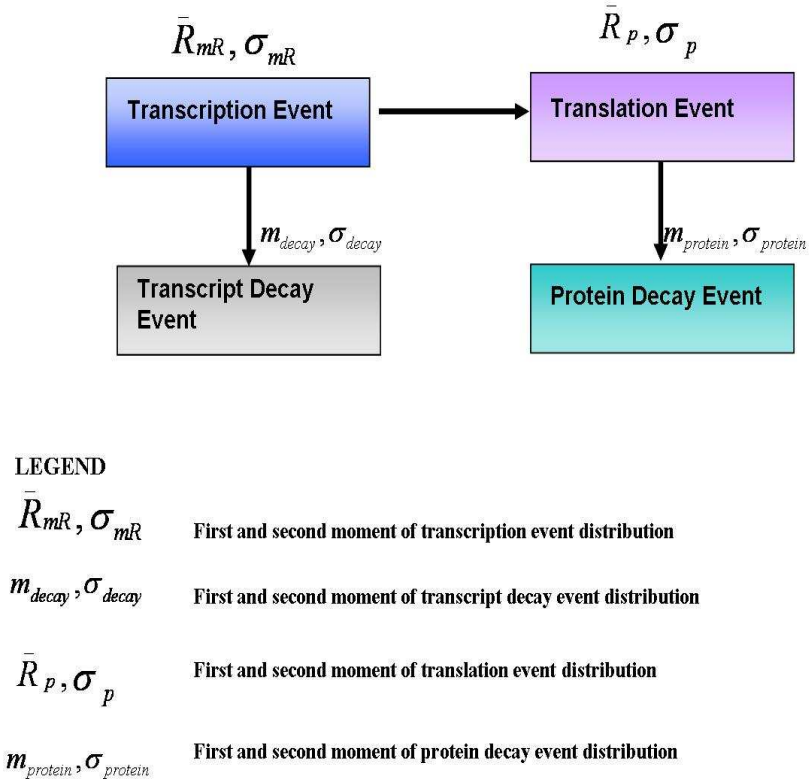

Figure 4.19. Molecular events involved in prokaryotic gene expression.

- 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 [90].
- Translation Event: This event captures the process of protein synthesis from a single mRNA molecule characterized by the probability distribution of its time ($\bar{R}_p$ and σ_p)
- Protein Decay Event: This event represents the decay of a protein characterized by an exponential distribution with half life of $m_{protein}$ obtained from experimental data [90].

The interactions of these molecular events, as captured in Fig 4.19, drives the dynamics of protein production in prokaryotic cells. In order to study the effect

of TF-DNA binding time, as expressed by the parameterized model elucidated in the previous section, on the stochasticity of protein synthesis, we conducted several *in silico* experiments by varying the average binding time for the TF-DNA binding microevent involved in transcription [3].

4.4.5.1 Protein synthesis dynamics with TF-DNA binding time of 10 secs

We conducted simulation studies to validate experimentally observed "bursts" in protein generation of E.Coli. With the TF-DNA binding time of 10 secs (based on experimental observations reported in the previous section), Fig 4.20 shows the temporal dynamics of mRNA and protein molecules together with the noise profile (noise being quantitatively measured as the ratio of the variance to squared mean [50],[90]. As observed from the plots, the burstiness in the number of *LacZ* proteins produced (marked by a corresponding increase in noise) is primarily caused by the low frequency of transcription events (around 1.2 mRNA molecules are produced per cell cycle).

4.4.5.2 Protein synthesis dynamics with TF-DNA binding time of 0.1 secs

As noted in [4], the transcription initiation frequency has a key role in controlling the nature of stochasticity in protein synthesis. In order to analyze the impact of the TF-DNA binding event in this fine-grained regulation, we conducted simulation experiments for the *LacZ* system with different average TF-binding times computed from our model. In Fig 4.21, we show the dynamics of the gene expression process for TF-DNA binding time of 0.1 secs. As seen from the figure, a decrease in the average binding time does not

[3]The simulation was carried for the *lacZ* gene expression in E.Coli to validate with available experimental data. The simulation experiments were conducted for 10 cell cycle times and results represent average value for 50 simulation runs

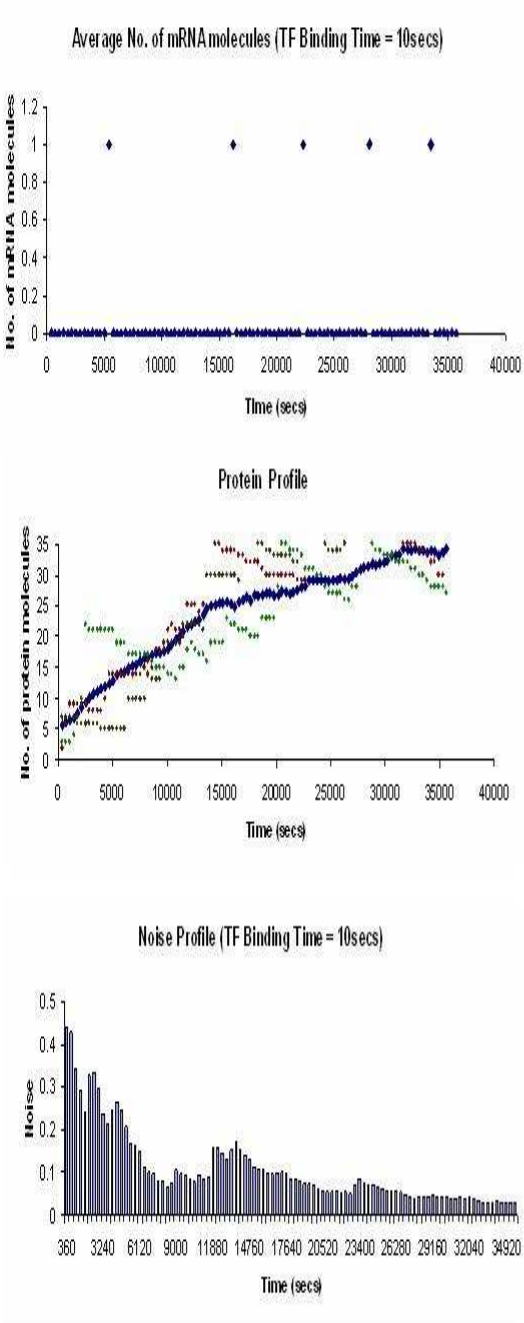

Figure 4.20. Dynamics: *lacZ* gene expression vs experimental TF-DNA binding time.

significantly increase the transcription event rate as observed in the similar protein and noise profiles as reported in the previous simulation case study.

4.4.5.3 Protein synthesis dynamics with TF-DNA binding time of 100 secs

In order to further analyze the effect of the TF-DNA binding event, particularly with increased event time, we conducted a simulation experiment with the TF-DNA binding event time set to 100 secs, an order of magnitude higher than the experimentally reported value. As seen from the protein and noise profiles in Fig 4.22, the TF-DNA binding time causes the transcription event time to decrease (i.e., the number of mRNA molecules released decreases due to the high TF-DNA binding time), thus increasing the "burstiness" in subsequent protein synthesis.

In this section, we have quantitatively captured the effect of the TF-DNA binding event as part of a dynamical system involving the temporal interaction of multiple molecular events associated with gene expression in prokaryotic cells. Our simulation results confirm biological observed burstiness in protein synthesis while providing *in silico* insights into the role of TF-DNA binding on the amplitude of fluctuations (noise) of the gene expression process.

4.4.6 Limitations of our model

Maxwell-Boltzmann distribution of molecular velocities: As mentioned before, the application of the Maxwell-Boltzmann distribution in our collision theory model requires further research.

3-d protein structure: The p_m estimation can be improved by considering the 3-d structure of the protein. Ideally, the motif of the protein molecule is located towards the outer surface such that p_m is actually higher than what we compute.

The actual protein-DNA binding process: The present model does not incorporate

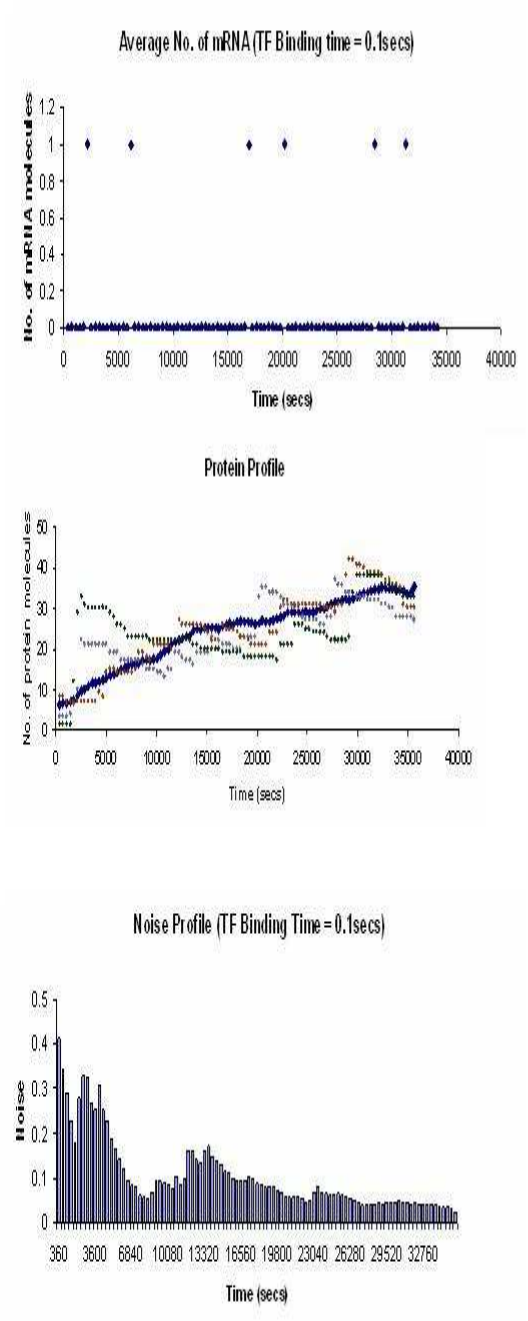

Figure 4.21. Dynamics: *lacZ* gene expression vs decreased TF-DNA binding time.

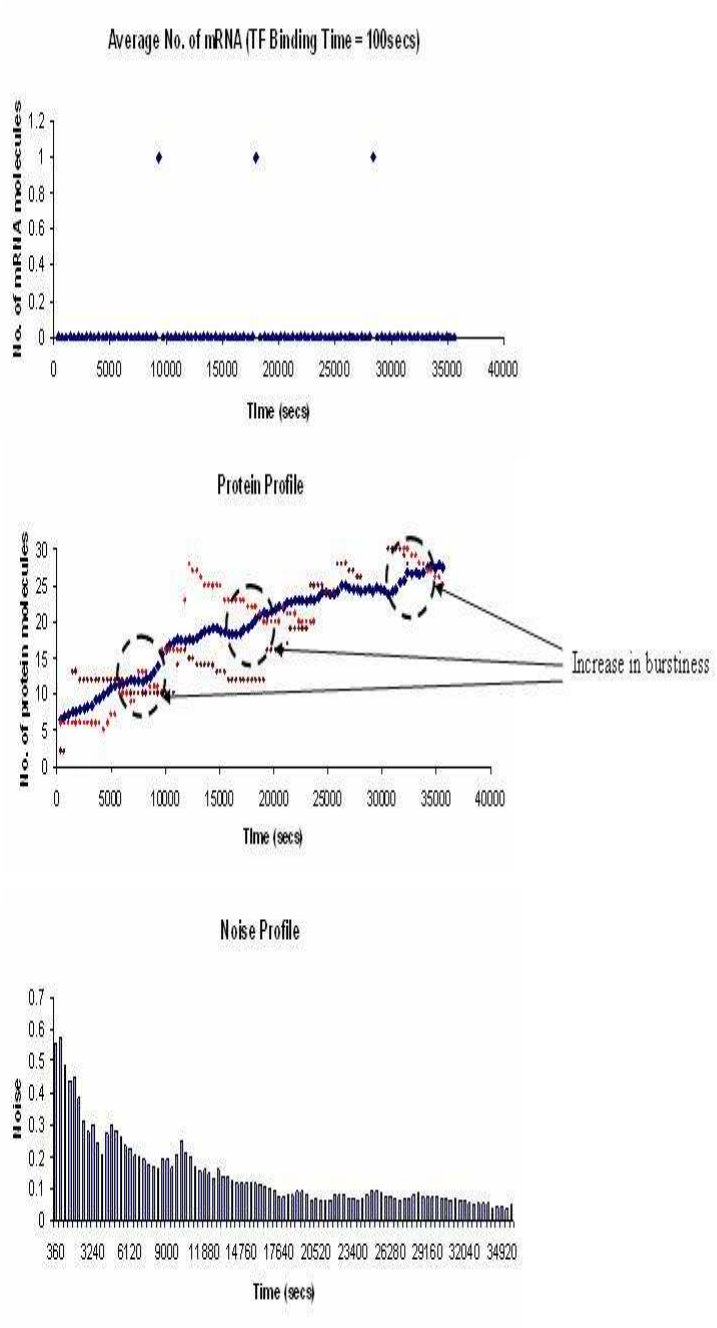

Figure 4.22. Dynamics: *lacZ* gene expression vs increased TF-DNA binding time.

the time required for the actual binding process i.e, how the specific atoms of the protein form chemical bonds with the DNA forming a stable complex. Also, it should be noted that the complex machinery of transcription, especially for eukaryotes, is not completely understood, yet. Many proteins can play a role in regulating one gene that would require further analysis. Our model can serve as a starting point for handling such cases.

Nucleosome dynamics: The long DNA chain in eukaryotes uses a systematic hierarchical compression. In the lowest compaction level the genetic material comprises arrays of coiled DNA around histones (globular octamer of cationic nucleus proteins) [12]. Each of these array elements is called a nucleosome that exhibits the following four dynamics: (1) compositional alternation, (2) covalent modification, (3) translational repositioning, and (4) conformational fluctuation. Compositional alternation is done by some remodeling enzymes to promote gene activation. Post translational modifications including acetylation, methylation, phosphorylation and ubiquitination are among the covalent modifications that can destabilize the histone cores and exploit DNA access to the biological processes. ATP-dependent remodellers use energy derived from ATP hydrolysis to loosen the contacts between the coiled DNA and the histone core. In translational repositioning, the bp position of core particles in the genome change to enhance the target site access. This process can happen both intrinsically or by the aid of remodellers. Conformational fluctuation is a periodic minor change to the conformation of a canonical nucleosome. The model presented in this chapter can help incorporate these factors in a more comprehensive protein-DNA binding model.

4.4.7 Biological implications

Several TFs searching simultaneously: If we consider several TFs searching for their sites on the DNA simultaneously, our results still remain valid. In [60], the authors argue that this may reduce the total search time because the experimental estimate

of $1 \sim 10$ secs is generated from the binding rate of the TFs to the DNA site. Our results, however, are for any specific TF and compute the average time required for this TF to bind to the DNA. Thus increasing the number of TFs should not change the results that we report for any particular TF. In fact, this brings down the experimental estimate of the binding time and hence requires lesser E_{act} for stable binding (as discussed in the next subsection). This may also cause molecular crowding in the cell which can have an impact on the search time. We did not consider molecular crowding on DNA or protein hopping (intersegment transfer) in our model for similar reasons as in [60].

Funnels and local organization of sites: Several known bacterial and eukaryotic sites tend to cluster together. Such clustering or other local arrangement of the sites can create a funnel in the binding energy landscape leading to more rapid binding of cognate sites. Our model assumes no such funnels of energy field. In the present model, the probability of collision is assumed uniform for the entire DNA. Because of local organization of sites, there is bias in the collision site, we can model that effect by changing the uniform distribution by another distribution to represent this bias. Also, due to change in energy landscape if the binding energy requirement changes, the probability of binding will increase in our model and hence will reduce the binding time.

Possible experiments to test our predictions: The search time depends on the activation energy of the TFs, which, in turn, can be controlled by the ionic strength of solution. Also, we show how the binding rate depends on the average collision time between two random segments of DNA, τ. This time measurement (τ) depends on the DNA concentration and the domain organization of DNA. By changing DNA concentration and/or DNA stretching in a single molecule experiment, one can alter τ and thus study the role of DNA packing on the rate of binding. This effect has implications for DNA recognition in vivo, where DNA is organized into domains. Similarly, one can experimentally measure and compare the binding rate, in the presence of other DNA-

binding proteins or nucleosomes.

Biological relevance of our model: Our model suggests that the kinetic energy of the TFs has to exceed E_{act} for successful binding. Is the *kinetic energy* of the TFs greater than this minimum requirement in general ? Theoretically, of course, the energy can be infinitely large for any molecule. Moreover, the bound on E_{act} can be brought down significantly if we incorporate the above factors. Note that the experimental estimate of $1 \sim 10$ secs incorporates the actual binding time. Thus the time for searching a DNA site by a TF should be quite smaller than $1 \sim 10$ secs resulting in a very low requirement of E_{act}. Also, because the experimental results depend on the binding rate, the total search time for 100 copies of a TF searching in parallel for the cognate site in a cell of 1 μm^3 volume is $\simeq 0.1$ s. This estimate further decreases with increasing number of TFs. So, to compute the average time for binding experimentally, we really need to compute the average number of that particular TF in the cell. Thus the model presented in this chapter can be further extended to incorporate these factors.

4.5 Summary

We have presented a simplified model to estimate the DNA-protein binding time by transforming the biological function as a stochastic process of a number of biological microevents. The probabilities of these microevents are used to create the complete stochastic model of the biological event. We used collision theory and Maxwell Boltzmann velocity distribution to calculate this microevent probability. The model is computationally fast and provides two moments for this random binding time. The model is robust as the major factors are captured in a reasonably accurate way for general cell environments. The complexity of DNA packing has been simplified to achieve acceptable estimates of the DNA-protein binding time. We found the range of activation energies of the TFs that are crucial for the robust functioning of gene transcription. The speed-stability paradox

can also be bypassed using the no TF sliding assumption and its effects reduced if we incorporate 1-d diffusion. The proposed mechanism has important biological implications in explaining how a TF can find its site on DNA in vivo, in the presence of other TFs and nucleosomes and by a simultaneous search by several TFs. In addition to providing a quantitative framework for analysis of the kinetics of TF binding (and hence, gene expression), our model also links molecular properties of TFs and the location of the binding sites on nucleosome cylinders to the timing of transcription activation. This provides us with a general, predictive, parametric model for this biological function. These details make the model more versatile compared to the current rate constants used in the Gillespie simulation. Thus, our discrete stochastic modeling can incorporate more parameters in the simulation.

CHAPTER 5

PROTEIN-LIGAND DOCKING

This chapter presents a computationally fast analytical model to estimate the time taken for protein-ligand docking in biological pathways. The model includes the structural details of the ligands, proteins and the binding mechanism, thus permitting its usage in different protein-ligand docking pairs. We use a modification of the collision theory based approach. The model captures the randomness of this problem in discrete time and estimate the first two moments of this process. The numerical results for the first moment show promising correspondence with experimental results and demonstrate the efficacy of our model.

This chapter is organized as follows: Section 5.1 discusses some related works on analytical models for protein-ligand docking. Sections 5.2 and 5.3 presents our stochastic model for protein-ligand docking. Section 5.4 reports the results for a sample protein-ligand pair for bacterial cells. In Section 5.5, we discuss the biological implications of our model and also present its limitations. Finally, in Section 5.6 we summarize the findings of this chapter.

5.1 Background on existing protein-ligand docking models

Most of the work on protein-ligand docking use Brownian dynamic simulations to model the mechanism. From the point of view of kinetics, protein docking should entail distinct kinetic regimes where different driving forces govern the binding process at different times [13, 14, 15]. This is because of the free energy funnel created by the binding site of the protein. The funnel distinguishes three kinetic regimes. First,

nonspecific diffusion (regime I) brings the molecules to close proximity. This is the motion created by the random collision of the molecules. Second, in the recognition stage (regime II), the chemical affinity steers the molecules into relatively well oriented encounter complexes ($\approx 5 \times 10^{-10}$ m), overcoming the mostly entropic barrier to binding. Brownian dynamics simulation of this regime [22] were also found to be consistent with a significant narrowing of the binding pathway to the final bound conformation. Finally, regime III corresponds to the docking stage where short-range forces mold the high affinity interface of the complex structure.

Long-range electrostatic effects can heavily bias the approach of the molecules to favor reactive conditions. This effect was shown to be important for many association processes, including those of proteins with DNA [80], proteins with highly charged small molecules [52], and proteins with oppositely charged protein substrates [86, 55, 36, 82, 62]. These systems have been thoroughly studied and are frequently regarded as typical examples of binding phenomena. Electrostatics is clearly not the only force that can affect the association rate. In addition, the most important process contributing to the binding free energy is desolvation, i.e., the removal of solvent both from nonpolar (hydrophobic) and polar atoms [16]. It is generally accepted that partial desolvation is always a significant contribution to the free energy in protein-protein association, and it becomes dominant for complexes in which the long-range electrostatic interactions are weak [17]. Brownian dynamics simulations to study the effects of desolvation on the rates of diffusion-limited protein-protein association have been reported in [22].

5.2 Proposed analytical model

Let us consider the docking between a protein A and a ligand B. Let the total number of surface binding points in A be n_A and that in B be n_B. The number of surface docking points to produce the AB complex is denoted by n_s, such that:

$$n_s << n_A; \quad n_s << n_B \tag{5.1}$$

We assume that the n_s docking points are all contiguous and if any *three* of the docking

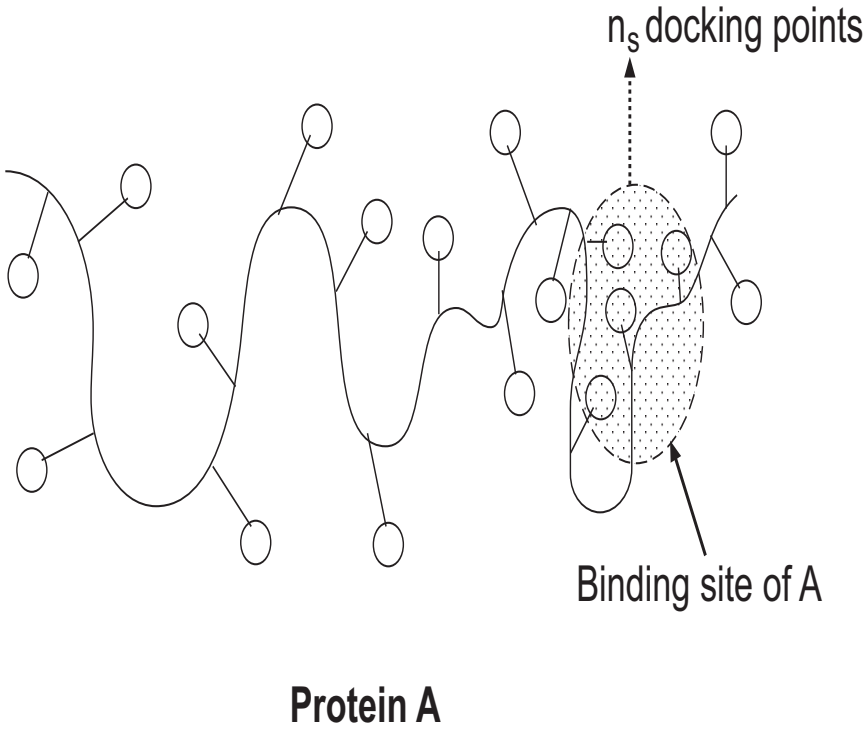

Figure 5.1. The protein docking mechanism.

points is hit by the ligand during a collision, the attractive force of the amino acid side-chain will force the ligand to change orientation so that it can bind to the site. This assumption has a few limitations which we will discuss in Section 5.5. Now, let the total probability of hitting the site during a collision for successful docking be p_f. The

probability of hitting the binding site at only one of the docking points is $p_f^1 = \frac{\binom{n_s}{1}}{\binom{n_A}{1}\binom{n_B}{1}}$.

Similarly, the probability of hitting the binding site at i docking points is given by:

$$p_f^i = \frac{\binom{n_s}{i}}{\binom{n_A}{i}\binom{n_B}{i}}, \quad (1 \leq i \leq n_s) \tag{5.2}$$

Thus p_f can be expressed as follows:

$$p_f = \sum_{i=3}^{n_s} p_f^i = \sum_{i=1}^{n_s} \frac{\binom{n_s}{i}}{\binom{n_A}{i}\binom{n_B}{i}} \tag{5.3}$$

Also, let p_b denote the probability that the ligand collides with the protein A with sufficient kinetic energy for successful docking. Hence, the total probability that the ligand hits the binding site while colliding with the protein, p_t, is given by:

$$p_t = p_b \times p_f \tag{5.4}$$

In general, the process of protein-ligand association can be described by a three-step reaction mechanism as follows:

$$A + B = \underset{k^-}{\overset{k^+}{\Longleftrightarrow}} A...B \underset{k_1^-}{\overset{k_1^+}{\Longleftrightarrow}} A - B \underset{k_{AB}^-}{\overset{k_{AB}^+}{\Longleftrightarrow}} AB, \tag{5.5}$$

where $A...B$ denotes the nonspecific encounter pairs, $A-B$ denotes the precursor state(s) leading to the docked conformation AB [18]. If long-range interactions can be neglected, the first reaction step is the random collision of the protein and ligand (A and B), resulting in a nonspecific encounter complex $A...B$ within the desolvation layer. To a good approximation, the limiting rate k^+ of this first regime is given by the Smoluchowski limit [63], k_{coll}. Indeed, the overall repulsion of the force fields has little effect on k^+. The authors in [22] report that the typical lifetime of a nonspecific encounter complex $A...B$ diffusing within the desolvation layer is about 4 ± 1 ns. This value is consistent with the nonspecific affinity between proteins that is estimated to be $10^2 M^{-1}$ or less [91].

The third reaction step in Eq. 5.5 i.e., the late transition between the favorable intermediate(s) $A - B$ and the bound state AB, substantially differs from the first two steps. The onset of the late transition coincides with the need to remove steric clashes and charge overlaps in the binding mechanism. Although the first two steps are governed by diffusion, the third is a process of induced fit that requires structural rearrangements involving mostly side chains. [22] reports that this late transition is not diffusive. For ligands that bind in a diffusion-controlled (or diffusion limited) reaction, the rate-limiting step must be the diffusive search for the partially desolvated intermediate(s) or precursor state(s) rather than the third step, and thus $k_{AB}^{+} \gg k_{1}^{-}$.

In this chapter, we focus on the kinetics of the total binding process. In particular, the collision theory model incorporates the first two steps together, whereas the Ligand axis rotation model estimates the third step.

5.2.1 Rotation of the ligand axis with respect to protein A

Fig 5.2 shows the rotation of the ligand axis to bring about the final docking configuration. The final orientation can be reached by the rotation of the ligand axis by an angle θ, where $(0 \leq \theta \leq 2\pi)$. However, as we will see in Section 5.4, this angle is often quite small ranging between $(0 \leq \theta \leq \frac{\pi}{2})$. Also, we must have:

$$d_{11'} \leq \gamma, \quad d_{33'} \leq \gamma, \quad d_{55'} \leq \gamma \tag{5.6}$$

where, γ is the threshold distance between any two binding points of A and B respectively for docking to occur.

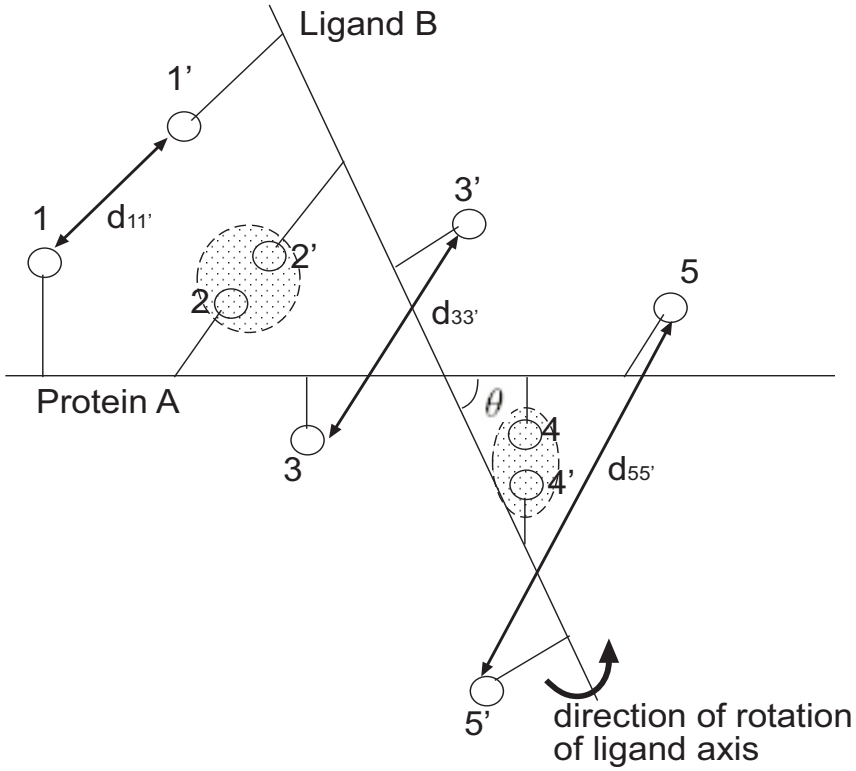

Figure 5.2. The rotation of the ligand axis.

5.2.2 Assumptions

1. Only the ligand rotates, to reach the final docked conformation whereas the protein remains fixed. In particular, we consider the *relative rotation* of the ligand axis with respect to the protein axis.
2. The docking point extends out of the ligand/protein backbones at an angle to the corresponding axis. In the analytical model, we have included both the cases when this angle is equal to $\frac{\pi}{2}$ and otherwise. The subsequent numerical results have been generated assuming an angle equal to $\frac{\pi}{2}$ as this is not yet reflected in the biological databases.
3. The *docking site* on the ligand/protein backbones are approximated as straight lines for ease in calculations. Note that the first step is to find the average angle

(in radians) that the binding site of the ligand axis has to rotate to reach the final docked conformation. We assume that the binding site of the ligand behaves like a rubber handle extending out of the spherical ligand structure. This allows us to compute the average time taken for the rotation of the ligand axis easily.

4. At least 3 docking points in the ligand has to come within the range of the threshold distance of the corresponding 3 docking points in protein A for a successful binding to occur.
5. We consider a 2-d coordinate system to estimate our results. A 3-d coordinate system can be used following the same concept but the equations become quite complicated to solve as discussed later. If 3 docking points are considered, it is always feasible to have the three points on the same plane where the other points are contributing to reduce the rotational threshold energy required for binding for these three 2-d points. Thus a 2-d assumption is appropriate for the model.
6. The docking points extend out of the protein/ligand backbones in a straight line.

The requirement of *at least 3 docking points* to come within the threshold distance of γ allows us to calculate the average angle of rotation, θ_{avg}, that the ligand axis has to rotate for successful docking with Protein A as discussed below.

5.2.3 Finding θ_{avg}

It should be noted that in the subsequent discussion all references to the ligand/protein backbones actually applies to only the docking site of the corresponding backbones (which are assumed as straight lines). Fig 5.3 shows the scenario when the ligand and the protein come within a distance of γ for at least 3 docking points.

5.2.3.1 Conventions

1. There are a total of n_s docking points.

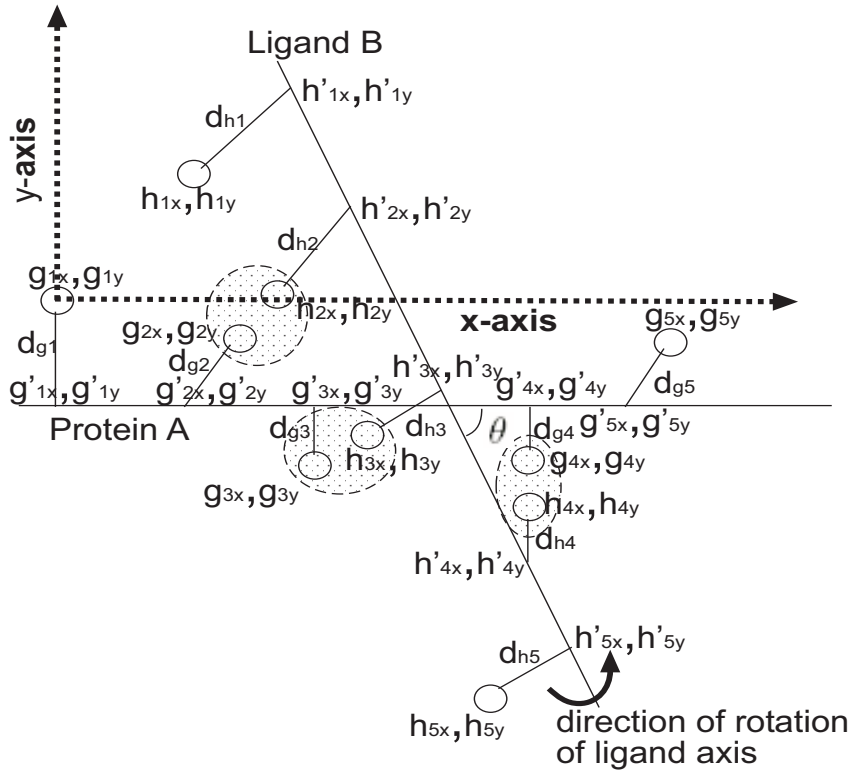

Figure 5.3. Ligand/Protein coming within threshold distance of 3 docking points.

2. The docking points on the *protein* are labelled as (g_{ix}, g_{iy}) to denote the x and y coordinates respectively of the i^{th} docking point.
3. The points on the amino acid backbone of the *protein* corresponding to the i^{th} docking points are denoted by (g'_{ix}, g'_{iy}).
4. The docking points on the *ligand* are labelled as (h_{ix}, h_{iy}) to denote the x and y coordinates respectively of the i^{th} docking point.
5. The points on the amino acid backbone of the *ligand* corresponding to the i^{th} docking points are denoted by (h'_{ix}, h'_{iy}).
6. The origin of our 2-d coordinate system is at (g_{1x}, g_{1y}), i.e, $(g_{1x}, g_{1y}) = (0, 0)$.
7. The distance between the i^{th} docking point and the corresponding point on the *protein* backbone is given by d_{gi}.

8. The distance between the i^{th} docking point and the corresponding point on the *ligand* backbone is given by d_{hi}.
9. The angle between the straight line connecting the i^{th} docking point and the protein backbone and the straight line denoting the protein backbone is denoted by ϕ_i.
10. The angle between the straight line connecting the i^{th} docking point and the ligand backbone and the straight line denoting the ligand backbone is denoted by ψ_i.
11. The docking site on the protein backbone (assumed to be a straight line) is parallel to the x-axis of the 2-d coordinate system. Thus the equation of this straight line is $y = -(d_{g1}) \sin \phi_1$.
12. The distance between the points on the protein backbone corresponding to the i^{th} and j^{th} docking points is denoted by D_{gij}.
13. The distance between the points on the ligand backbone corresponding to the i^{th} and j^{th} docking points is denoted by D_{hij}.

The angles ϕ_i, $(\forall i)$ are measured from the protein axis to the straight line extending out of the axis carrying the docking point in an anti-clockwise direction as shown in Fig 5.4. Similarly, the angles ψ_i, $(\forall i)$ are also computed.

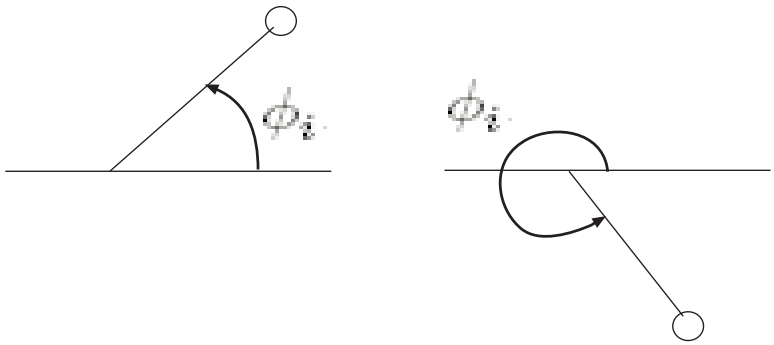

Figure 5.4. Determining the angles between the axis and the docking point.

5.2.3.2 Coordinates of the docking points of the protein backbone

Its fairly easy to compute the coordinates of all the n_s docking points and their corresponding contact points on the protein backbone. We will simplistically sketch the process in this section.

The first docking point on the protein backbone, (g_{1x}, g_{1y}) is considered to be the origin of our coordinate system. Also, because the equation of the straight line denoting the protein axis is known, we can write:

$$(g_{1x}, g_{1y}) = (0, 0); \tag{5.7}$$

$$g'_{1y} = -(d_{g1}) \sin \phi_1; \qquad (g'_{1x})^2 + (g'_{1y})^2 = (d_{g1})^2 \tag{5.8}$$

From, Eq 5.8 we can readily calculate (g'_{1x}, g'_{1y}). Next, we can compute (g'_{ix}, g'_{iy}), $(1 \leq i \leq n_s)$ by solving the following set of equations:

$$g'_{iy} = -(d_{g1}) \sin \phi_1 \tag{5.9}$$

$$(g'_{ix} - g'_{1x})^2 + (g'_{iy} - g'_{1y})^2 = (D_{g1i})^2; \quad 2 \leq i \leq n_s \tag{5.10}$$

Next, we can estimate the coordinates of the docking points of the protein (g_{ix}, g_{iy}), $(2 \leq i \leq n_s)$ by solving the following equation pair:

$$(g'_{ix} - g_{ix})^2 + (g'_{iy} - g_{iy})^2 = (d_{gi})^2; \quad g_{iy} = g'_{iy} + (d_{gi}) \sin \phi_i \tag{5.11}$$

5.2.3.3 Calculating the coordinates of any three docking points on the ligand

The angle θ as shown in Fig 5.3 denotes the angle made by the docking sites of the ligand backbone with the protein backbone (and equivalently the x-axis). As mentioned before, we assume that *any three* docking points on the ligand come within the threshold distance of the corresponding docking points of the protein. Without loss of generality, let us assume that these 3 docking points are denoted by (h_{ix}, h_{iy}), (h_{jx}, h_{jy}) and (h_{kx}, h_{ky})

corresponding to the docking points on the protein denoted by (g_{ix}, g_{iy}), (g_{jx}, g_{jy}) and (g_{kx}, g_{ky}), where $1 \leq i, j, k \leq n_s$ and $i \neq j \neq k$. Thus we can write:

$$(h_{ix} - g_{ix})^2 + (h_{iy} - g_{iy})^2 \leq \gamma^2 \tag{5.12}$$

$$(h_{jx} - g_{jx})^2 + (h_{jy} - g_{jy})^2 \leq \gamma^2 \tag{5.13}$$

$$(h_{kx} - g_{kx})^2 + (h_{ky} - g_{ky})^2 \leq \gamma^2 \tag{5.14}$$

Next, we can find the distance between the docking points (h_{ix}, h_{iy}) and their corresponding points of attachment to the ligand axis (h'_{ix}, h'_{iy}) denoted by d_{hi} (from the PDB database [42]) and hence:

$$(h_{ix} - h'_{ix})^2 + (h_{iy} - h'_{iy})^2 = (d_{hi})^2 \tag{5.15}$$

$$(h_{jx} - h'_{jx})^2 + (h_{jy} - h'_{jy})^2 = (d_{hj})^2 \tag{5.16}$$

$$(h_{kx} - h'_{kx})^2 + (h_{ky} - h'_{ky})^2 = (d_{hk})^2 \tag{5.17}$$

The distances between the corresponding points on the ligand axis can also be estimated (from the PDB database) and we have:

$$(h'_{ix} - h'_{jx})^2 + (h'_{iy} - h'_{jy})^2 = (D_{hij})^2 \tag{5.18}$$

$$(h'_{ix} - h'_{kx})^2 + (h'_{iy} - h'_{ky})^2 = (D_{hik})^2 \tag{5.19}$$

Also, our assumption that the docking points extend out of the ligand backbone in a straight line allows us to formulate the slope of these lines as $\frac{h_{iy}-h'_{iy}}{h_{ix}-h'_{ix}}$, $\frac{h_{jy}-h'_{jy}}{h_{jx}-h'_{jx}}$ and $\frac{h_{ky}-h'_{ky}}{h_{kx}-h'_{kx}}$. And because the corresponding angles of these lines with the ligand axis can be estimated, we have:

$$\left\{ \begin{array}{c} \tan\psi_i = \frac{\frac{h_{iy}-h'_{iy}}{h_{ix}-h'_{ix}} - m}{1 + m\frac{h_{iy}-h'_{iy}}{h_{ix}-h'_{ix}}}, \quad \text{for } \psi_i \neq \frac{\pi}{2} \\ m\frac{h_{iy}-h'_{iy}}{h_{ix}-h'_{ix}} = -1, \quad \text{for } \psi_i = \frac{\pi}{2} \end{array} \right\} \tag{5.20}$$

$$\left\{ \begin{array}{c} \tan\psi_j = \frac{\frac{h_{jy}-h'_{jy}}{h_{jx}-h'_{jx}}-m}{1+m\frac{h_{jy}-h'_{jy}}{h_{jx}-h'_{jx}}}, \quad \text{for } \psi_j \neq \frac{\pi}{2} \\ m\frac{h_{jy}-h'_{jy}}{h_{jx}-h'_{jx}} = -1, \quad \text{for } \psi_j = \frac{\pi}{2} \end{array} \right\} \tag{5.21}$$

$$\left\{ \begin{array}{c} \tan\psi_k = \frac{\frac{h_{ky}-h'_{ky}}{h_{kx}-h'_{kx}}-m}{1+m\frac{h_{ky}-h'_{ky}}{h_{kx}-h'_{kx}}}, \quad \text{for } \psi_k \neq \frac{\pi}{2} \\ m\frac{h_{ky}-h'_{ky}}{h_{kx}-h'_{kx}} = -1, \quad \text{for } \psi_k = \frac{\pi}{2} \end{array} \right\} \tag{5.22}$$

where, m is the slope of the straight line denoting the ligand axis. Note that, in Section 5.4, we assume an angle of $\frac{\pi}{2}$ to generate the results as the corresponding angles are not reported in the biological databases. Finally, because the points (h'_{ix}, h'_{iy}), (h'_{jx}, h'_{jy}) and (h'_{kx}, h'_{ky}) lie on the same straight line (i.e, the ligand backbone), we can write:

$$h'_{ky} - h'_{iy} = (h'_{kx} - h'_{ix})\frac{h'_{jy} - h'_{iy}}{h'_{jx} - h'_{ix}} \tag{5.23}$$

$$h'_{ky} - h'_{jy} = (h'_{kx} - h'_{jx})\frac{h'_{jy} - h'_{iy}}{h'_{jx} - h'_{ix}} \tag{5.24}$$

Thus, in Equations 5.12-5.24, we have 13 equations to solve for the following 13 unknown variables: $h_{ix}, h_{iy}, h_{jx}, h_{jy}, h_{kx}, h_{ky}, h'_{ix}, h'_{iy}, h'_{jx}, h'_{jy}, h'_{kx}, h'_{ky}$ and m. Note that, we need at least 3 docking points to form sufficient number of equations for solving all the unknown variables. To calculate θ from m, we observe that the slope of the ligand axis is given by $\tan(\theta)$, such that we have:

$$\theta = \arctan(m) \tag{5.25}$$

Note that the slope can be both positive or negative resulting in clockwise or anticlockwise rotations of the ligand axis. However, because we are interested in computing the time for rotation of the ligand axis, the direction of rotation is not important for us. Also, because the equations are nonlinear and involve inequalities, we can only make an approximate estimate of the coordinates of the docking points on the ligand.

5.2.3.4 Calculating θ_{avg} from θ

The next step is to estimate the average angle of rotation, θ_{avg}. We will find the angle θ (as outlined above) considering any 3 docking points out of the possible n_s points. This requires a total of $\binom{n_s}{3}$ iterations.

We next find the average angle of rotation considering 3 docking points, θ^3_{avg}, from the $\binom{n_s}{3}$ different θ^3_i's ($1 \leq i \leq \binom{n_s}{3}$) calculated (where, θ^3_i denotes the angle computed using the above equations for the i^{th} combination of 3 docking points). Assuming uniform probability for all these cases, we have:

$$\theta^3_{avg} = \sum_{i=1}^{\binom{n_s}{3}} \frac{\theta^3_i}{\binom{n_s}{3}} \tag{5.26}$$

Note that if greater number of docking points come within the threshold distance, θ^j_{avg} ($4 \leq j \leq n_s$) will continue to decrease. We next consider the case when more than 3 docking points come within the threshold distance. If 4 points come within the distance, we will have an extra 4 variables to solve ($h_{mx}, h_{my}, h'_{mx}, h'_{my}$). Note that our assumptions for this coordinate system is only valid if all of these four points are on the same plane. We will have another 4 equations by adding the equations corresponding to this new point to the Eqs 5.12-5.14, Eqs 5.15-5.17, Eqs 5.18-5.19 and Eqs 5.20-5.22 respectively as follows:

$$(h_{mx} - g_{mx})^2 + (h_{my} - g_{my})^2 = \gamma^2 \tag{5.27}$$

$$(h_{mx} - h'_{mx})^2 + (h_{my} - h'_{my})^2 = (d_{hm})^2 \tag{5.28}$$

$$(h'_{ix} - h'_{mx})^2 + (h'_{iy} - h'_{my})^2 = (D_{him})^2 \tag{5.29}$$

$$\left\{ \begin{array}{c} \tan\psi_m = \frac{\frac{h_{my}-h'_{my}}{h_{mx}-h'_{mx}} - m}{1+m\frac{h_{my}-h'_{my}}{h_{mx}-h'_{mx}}}, \quad \text{for } \psi_m \neq \frac{\pi}{2} \\ m\frac{h_{my}-h'_{my}}{h_{mx}-h'_{mx}} = -1, \quad \text{for } \psi_m = \frac{\pi}{2} \end{array} \right\} \tag{5.30}$$

Next we can calculate the average angle of rotation considering 4 docking points, θ^4_{avg}, in the same way as discussed above assuming uniform probability for all the $\binom{n_s}{4}$ different cases as follows:

$$\theta^4_{avg} = \sum_{i=1}^{\binom{n_s}{4}} \frac{\theta^4_i}{\binom{n_s}{4}} \tag{5.31}$$

This procedure is repeated to calculate θ^j_{avg}, $(4 < j \leq n_s)$ in the same away by adding 4 new equations for each extra docking point considered.

Finally, the average angle of rotation, θ_{avg} can be approximated as;

$$\theta_{avg} = \frac{1}{p_f} \sum_{i=3}^{n_s} p^i_f \times \theta^i_{avg} \tag{5.32}$$

5.2.4 Computing θ_{avg} using a 3-d coordinate system

As mentioned before, a 3-d coordinate system can be used in a similar way to compute θ_{avg}. However, as this increases the number of unknown variables appreciably, we need to assume that at least 15 docking points of the protein/ligand come within the threshold distance. This greatly increases the number of equations that has to be solved as well. Moreover, for small docking sites, the assumption of 15 docking points coming close might not be a practical way of solving the problem. Another disadvantage of the 3-d calculations is that as we need more docking points to come close, the value of θ_{avg} becomes less than what we estimate with the 2-d system, resulting in a further decrease in the estimation of the time for the rotation of the ligand axis. Fig 5.5 plots the rotational energy required (measured in terms of total change in free energy reported in [22]) for different number of docking points coming within threshold distance (varied from 3 to 15). The results were generated for the protein-ligand pair of human leukocyte elastase and OMTKY3 where the optimal configuration corresponds to 15 docking points coming close (as we will have maximum chance of docking in that case) and the subsequent energy requirements were assumed for lesser number of docking points coming close. We

observe that as more docking points come close, the rotational energy required is lesser i.e., the ligand axis has to rotate less to reach the docked conformation indicating that the time required for rotation also decreases.

As we show later, the total protein-ligand docking time is primarily governed by the collision theory component (i.e., the time required for rotation of the ligand axis is negligible in comparison to the time taken by the ligand to collide with the docking site on the protein), and hence the lesser accuracy of the 2-d based computations is not a deterrent in estimating the total docking time. Also, this reduces the number of equations that need to be solved making the model computationally fast which is a basic requirement for our discrete event-based simulator.

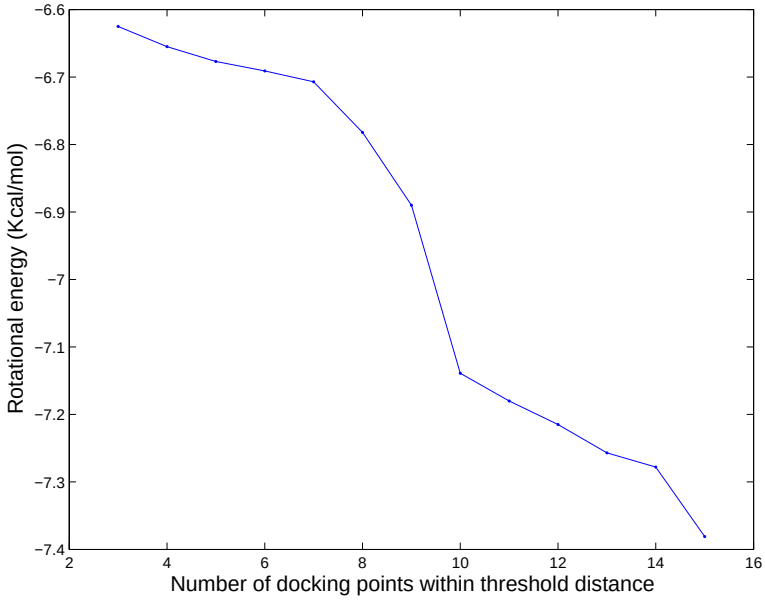

Figure 5.5. Rotational energy vs no. of docking points within threshold distance.

5.2.5 Calculating p_b

We assume that the ligand molecules enter the cell one at a time to initiate the binding. From the principles of collision theory for hard spheres, we model the protein and ligand molecules as rigid spheres with radii r_P and r_L respectively. As mentioned before, p_b denotes the probability of collision of the ligand with the protein with enough kinetic energy for the binding to occur successfully. Let the total volume of the cell be V and n_2 denotes the number of protein molecules present inside the cell. We next assume that the colliding ligand molecule must have free energy E_{Act} or greater to overcome the energy barrier and bind to the specific protein molecule. Let $m_{PL} = \frac{m_P.m_L}{m_P+m_L}$ = the reduced mass where, m_L = mass (in gm) of the ligand molecule and m_P = mass (in gm) of the protein. Thus following the reaction model, we get:

$$p_b = \frac{n_2 r_{PL}^2 \Delta t}{V} \sqrt{\frac{8\pi k_B T}{m_{PL}}}\, e^{\frac{-E_{Act}}{k_b T}}$$

5.3 Computing the time taken for protein-Ligand docking

Now, we are in a position to analytically compute the time taken for ligand-protein docking. This can be divided into two parts: 1) computing the time taken for the ligand to collide with the binding site of the protein molecule with enough activation energy to create a temporary binding and 2) computing the time taken for the rotation of the ligand axis to stabilize the binding to the protein molecule. Note that the first part computes the time for the random collisions until the creation of the precursor state $A - B$ (as shown in Eq. 5.5) and involves the first two steps in Eq. 5.5. The second part computes the time taken for the formation of the final docked complex, AB, from $A - B$.

5.3.1 Estimation of collision time for successful docking

Let $\Delta t = \tau =$ an infinitely small time step. The ligand molecules try to bind to the protein through collisions. If the first collision fails to produce a successful binding , they collide again after τ time units and so on.

We can interpret p_t as the probability of a successful binding in time τ. Thus, the *average time for the ligand to collide with the binding site of the protein molecule with enough activation energy for successful docking* denoted by T_1^c is given by:

$$T_1^c = p_t\tau + p_t(1-p_t)2\tau + p_t(1-p_t)^2 3\tau + ... = \frac{\tau}{p_t}$$

and the corresponding second moment, T_2^c, is given by:

$$T_2^c = p_t(\tau^2) + p_t(1-p_t)(2\tau)^2 + p_t(1-p_t)^2(3\tau)^2 + ... = \frac{(2-p_t)\tau^2}{p_t^2}$$

We find that the time for ligand-protein collisions (which is a random variable denoted by x) follows an exponential distribution for the specific ligand and protein used to generate the results (reported in the next section). It should be noted that as we assume τ to be quite small, we can approximate the total time measurements of binding using a continuous (exponential in this case) distribution instead of a discrete geometric distribution. Thus as reported later, we find $T_1^c \approx T_2^c$, and hence the pdf of the exponential distribution is given by:

$$f_1(x) = \left\{ \begin{array}{c} (\frac{1}{T_1^c})e^{-(\frac{x}{T_1^c})}, \quad \text{for } x \geq 0 \\ 0, \quad \text{otherwise} \end{array} \right\} \tag{5.33}$$

5.3.2 Finding the average time for rotation of ligand axis

Now to rotate the docking site on the ligand about the axis to reach the final docking configuration, we need to have some rotational energy which is contributed by

the total change in free energy in forming the docked complex (denoted by E_f). Thus we have:

$$\frac{1}{2} I_d w_d^2 = E_f \tag{5.34}$$

where, I_d and w_d are respectively the average rotational inertia and angular velocity of the docking site of the ligand. Now the estimates of E_f have been reported extensively in the literature, and our goal is to calculate I_d and w_d.

5.3.2.1 Calculating the average moment of inertia of the ligand, I_d:

The moment of inertia calculation can become tricky as we have to consider the axis of rotation as well as its distance from the ligand axis. Fig 5.6 illustrates the possible orientations of the protein and ligand axis where the dotted line with an arrow signifies the axis of rotation. Note that the protein and ligand axes might not intersect as well in some configurations (Figs 5.6(b),(c),(d)). In such cases, it becomes imperative to calculate the distance of the ligand axis from the point about which it rotates making the moment of inertia calculation quite cumbersome.

We assume that the ligand and protein axes do actually intersect in all cases (i.e. Figs 5.6(b),(c),(d) can never occur). This is a practical consideration because the ligand physically collides with the protein. We also assume that the ligand axis rotates about this point of intersection. Note that this simplifies the average moment of inertia calculation as the intersection point will always be on the ligand axis (and we do not have to compute the distance of the ligand axis from the axis of rotation).

From section 5.2.3.3 we can easily find the equations of the two lines denoting the protein and ligand axes (as the coordinates of at least 3 points on each line is known). Hence the point of intersection can be computed in a straight-forward manner. Let the point of intersection be denoted by (δ_x, δ_y). Also, we can estimate the coordinates of the

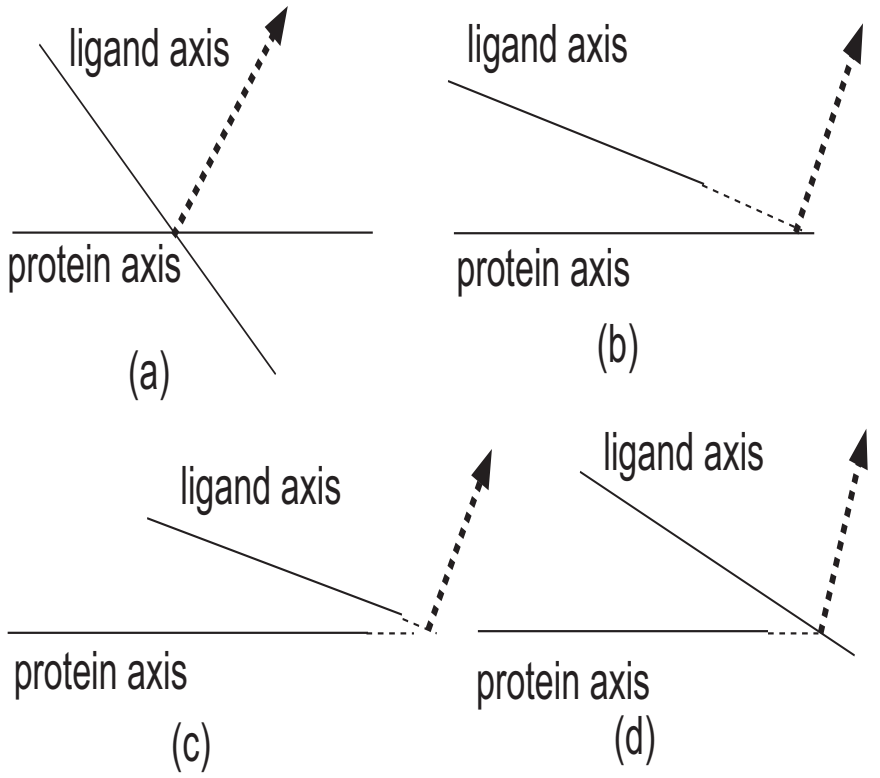

Figure 5.6. 4 possible orientations ((a),(b),(c),(d)) of the protein and ligand axes.

beginning (denoted by (b_x, b_y)) and end (denoted by (e_x, e_y)) points on the ligand axis corresponding to the first and last docking points 1 and n_s.

As explained before, the docking sites of the ligand and protein axes are assumed as straight lines, such that the ligand can be approximated as a sphere (of radius r_L) with a rubber handle (which is the straight line denoting the docking site on the ligand backbone). Fig 5.7 explains the model. This rubber handle on the ligand can be approximated as a cylinder with radius r_d and length $\sqrt{(b_x - e_x)^2 + (b_y - e_y)^2}$. Note that in Section 5.2.5 we had modelled the ligand as a hard sphere. However, the calculation of θ_{avg} and I_d requires the docking site of the ligand axis to be a straight line (for ease in computation). Note that, in general, the docking site is quite small compared to the length of the entire ligand, and thus the rubber handle assumption is quite feasible. The

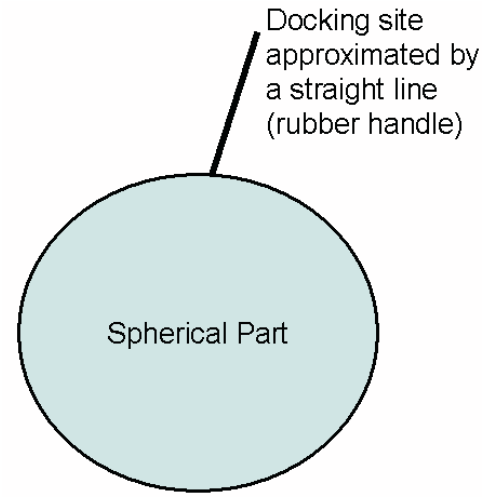

Figure 5.7. Approximate model of the Ligand molecule.

collision theory estimate can still treat the entire ligand as a sphere without taking into account the rubber handle part. However, because the docking site is approximated as a rubber handle, only this part rotates to bind to the corresponding site on the protein and hence I_d is the rotational inertia of the docking site only. We also assume that the docking site on the ligand has uniform density, ρ_d, and cross-sectional area, $A_d = \pi r_d^2$. Thus we can approximate I_d as follows:

$$I_d = \int_{-\sqrt{(\delta_x - e_x)^2 + (\delta_y - e_y)^2}}^{\sqrt{(\delta_x - b_x)^2 + (\delta_y - b_y)^2}} \rho_d A_d x^2 \, dx$$

$$= \frac{\rho_d A_d}{3}([(\delta_x - e_x)^2 + (\delta_y - e_y)^2]^{\frac{3}{2}} + [(\delta_x - b_x)^2 + (\delta_y - b_y)^2]^{\frac{3}{2}}) \tag{5.35}$$

5.3.2.2 Calculating T_1^r:

The average time for rotation of the docking site of the ligand axis (denoted by T_1^r) is given by:

$$T_1^r = \frac{\theta_{avg}}{w_d} \tag{5.36}$$

However, this does not allow us to compute the second moment of the time for rotation. We assume that the time for rotation follows an exponential distribution and hence the second moment of the time for rotation is given by:

$$T_2^r = 2(T_1^r)^2 \tag{5.37}$$

Thus this exponential distribution has both mean and standard deviation as T_1^r and pdf of the form:

$$f_2(x) = \left\{ \begin{array}{c} (\frac{1}{T_1^r})e^{-(\frac{x}{T_1^r})}, \quad \text{for } x \geq 0 \\ 0, \quad \text{otherwise} \end{array} \right\} \tag{5.38}$$

5.3.3 General distribution of the total time for protein-ligand docking

The total time for protein-ligand docking can be computed from the convolution of the two pdf's given in Eqns 5.33 and 5.38 as follows:

$$f(x) = f_1(x) \bigodot f_2(x) = \int_0^x f_1(z) f_2(x-z)\, dz$$

where, $f(x)$ denotes the pdf of the general distribution for the total time and $\bigodot$ is the convolution operator. Hence we get:

$$f(x) = \left\{ \begin{array}{c} \frac{e^{-\frac{x}{T_1^c}} - e^{-\frac{x}{T_1^r}}}{T_1^c - T_1^r}, \quad \text{for } x \geq 0 \\ 0, \quad \text{otherwise} \end{array} \right\} \tag{5.39}$$

Also we have:

$$T_1 = \int_0^\infty x f(x)\, dx = T_1^c + T_1^r; \quad T_2 = \int_0^\infty x^2 f(x)\, dx = 2[(T_1^c)^2 + T_1^c T_1^r + (T_1^r)^2]$$

where, T_1 and T_2 are the first and second moments of the total time taken for protein-ligand docking.

5.4 Results and analysis

5.4.1 Problems in validation of our model

Before presenting the results, we first discuss the difficulty of experimentally validating our model. Note that we compute the average time for protein-ligand binding in this chapter. Existing experimental results are based on estimation of the binding rate of the ligands to a specific protein. We consider the binding of the turkey ovomucoid third domain (OMTKY) ligand to the human leukocyte elastase protein to generate the results. The experimental rate constant of $10^6 M^{-1}s^{-1}$ as reported in [22] is derived from these rate measurements. Hence, the number of ligands in the cell will affect this estimate of time taken by one single ligand to bind to the protein because the rate of reaction incorporates the ligand concentration as well. However, our model computes the time taken by *any particular* ligand to bind to the protein which should be independent of the number of ligands in the cell. It is currently very difficult to carry out experiments to track a particular ligand and physically compute the time. Also, the stochastic nature of the binding process suggests that the distribution of the time taken will have a very high variance. In other words, in some cases the ligand requires time in microseconds whereas in other cases it might take as long as 1 second. The results (for the ligand-protein pair identified above) we present in the next section assume that the time taken for any particular OMTKY-human leukocyte elastase binding has a rate constant of $10^6 M^{-1}s^{-1}$ (as reported in [22]) even though it cannot be a true estimate of this event. Also, note that our model can be easily extended to incorporate the effects of multiple ligands present in the cell on the binding rate as discussed in Section 5.5.2.

5.4.2 Numerical results

In this section, we present the numerical results for the theoretical model derived in the chapter. Figs 5.8-5.12 present the results for OMTKY-Human leukocyte elastase

Table 5.1. Parameter Estimation for an average Human Cell

Parameters	Eukaryotic Cell
V	$4.187 \times 10^{-15} m^3$ (average volume of a human cell)
r_P	23.24×10^{-10} m (for *Human leukocyte elastase*)
r_L	14.15×10^{-10} m (for *Turkey ovomucoid third domain*)
n_s	8
r_d	1 nm
E_f (total change in free energy)	-7 Kcal/mol [22]
m_P	23328.2 Dalton (for *Human leukocyte elastase*)
Number of ligand (OMTKY) molecules	10^5
m_L	6047.9 Dalton (for *Turkey ovomucoid third domain*)
ρ_d	1.44 g/cm^3 (for *Turkey ovomucoid third domain* [37])

binding in an average human cell with 20 μm diameter. Also, the results were generated for $n_s = 8$ docking points on the protein/ligand. The different parameters assumed for the numerical results are concisely presented in Table 5.1. We used actual values from the from the PDB database [42] and some assumptions as reported in [22].

5.4.2.1 Calculation of I_d and w_d

To calculate I_d we need to know the point of intersection of the straight lines denoting the docking sites of the protein and ligand. Because, we need to estimate the average rotational inertia, we consider two cases: (1) the intersecting point is at the center of the docking site on the ligand and (2) the intersecting point is at the end of the

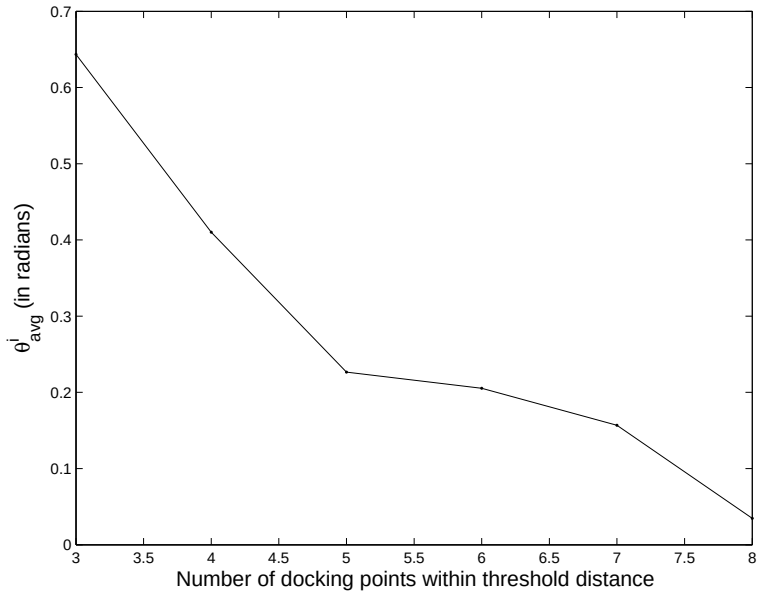

Figure 5.8. θ^i_{avg} against number of docking points within threshold distance.

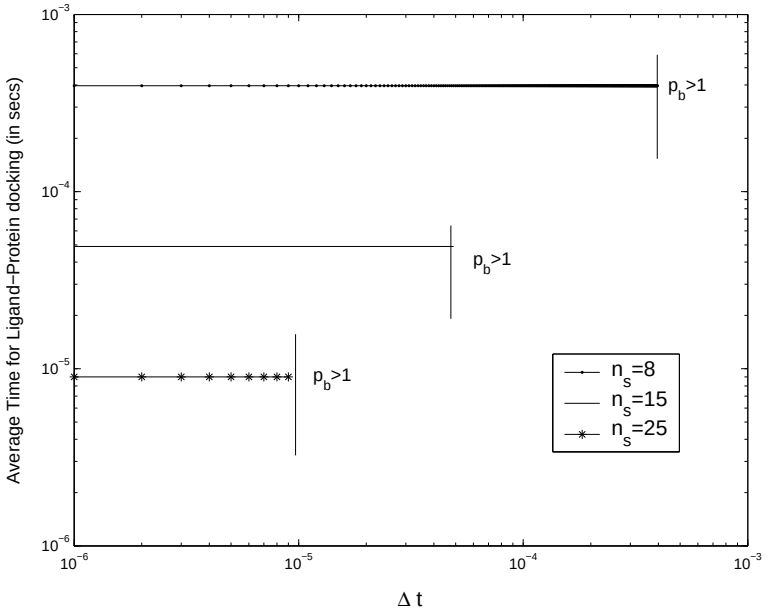

Figure 5.9. Average Time against Δt for different n_s.

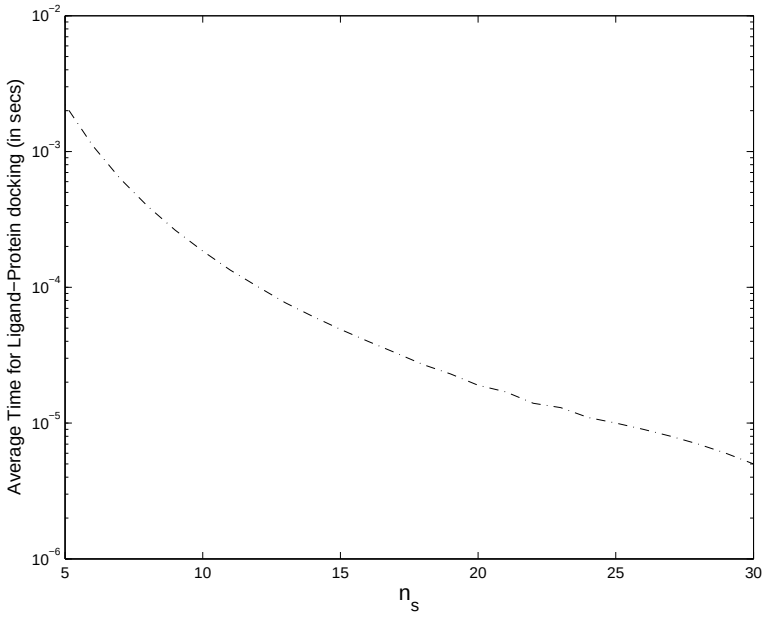

Figure 5.10. Average Time against n_s.

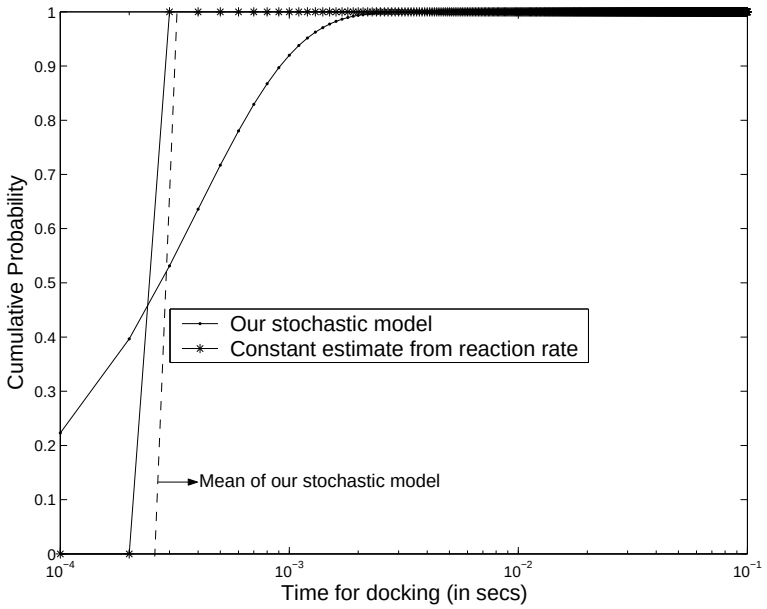

Figure 5.11. Cumulative probability distribution for the ligand-protein docking time.

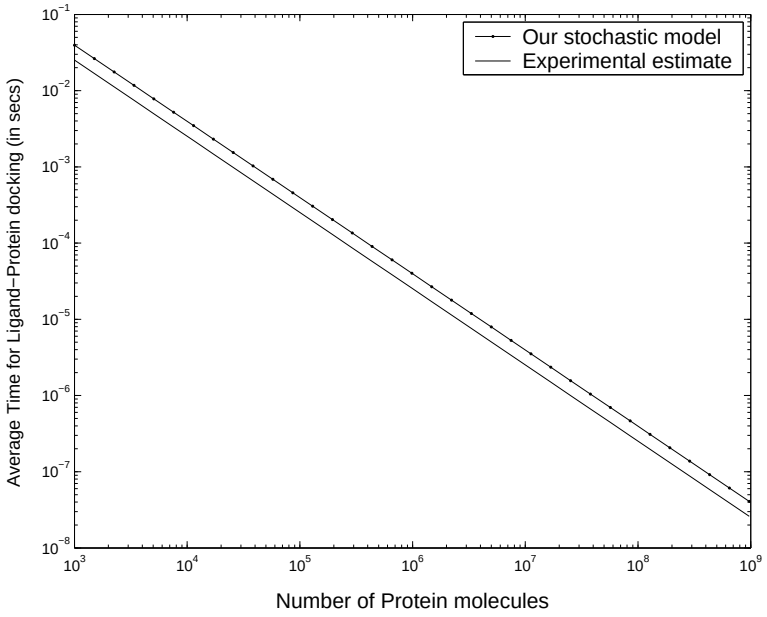

Figure 5.12. Average Time against number of Protein molecules (n_2).

docking site on the ligand. Note that the coordinates of the exact set of docking points and their corresponding points on the protein/ligand backbones have been estimated using the LPC software [96]. Also the density of the ligand molecule is assumed to be 1.44 g/cm^3 as the molecular weight of OMTKY is ≈ 6 KDalton (see [37] for details).

The corresponding values for w_d (assuming $E_f = -7$ Kcal/mol, from [22]) are 63.5×10^9 and 31.75×10^9 radians/sec respectively. Note that, [64] reports that the average angular velocity of a protein molecule is in the range $\approx 10^9$ radians/sec, which is very close to our estimate.

5.4.2.2 Estimation of θ_{avg}

Fig 5.8 plots θ^i_{avg}, $(3 \leq i \leq 8)$, against the number of docking points coming within threshold distance of $\gamma = 2 \times 10^{-10}$ m. Note that instead of averaging out the $\binom{n_s}{i}$ possible cases of choosing i docking points, we assumed that only i contiguous points can come within a distance of γ. This is because for the other combinations, the angle was

too small making the corresponding θ^i_{avg} too low. Thus, Eq 5.26 was modified as follows to generate the results:

$$\theta^i_{avg} = \sum_{j=1}^{n_s-i+1} \frac{\theta^i_j}{n_s - i + 1} \tag{5.40}$$

As expected, we find that the angle reduces as more docking points come within threshold distance. Also, we calculate $\theta_{avg} = 0.643483$ radians for the specific ligand-protein pair under consideration.

5.4.2.3 Estimation of T_1^r

The next step is to estimate the mean of the time for rotation of the docking site of the ligand axis to produce the final docked complex. We obviously get $T_1^r \approx 1 \times 10^{-11}$ and 2×10^{-11} secs for the two w_d estimates reported previously. Thus in general we can say that the time for rotation is too small in comparison to the time for collision, T_1^c as reported subsequently. Thus the total time for ligand-protein docking is dominated by T_1^c which corroborates the results reported in [22].

5.4.2.4 Dependence of T_1 on Δt

Fig 5.9 plots T_1 against different values for Δt. The average time for ligand-protein docking remains constant with increasing Δt. The same characteristics are seen for different number of docking points considered, $n_s = 8, 15, 25$ respectively. Though we have $n_s = 8$ for the ligand-protein pair under consideration, we have reported the plots for different values of n_s to show the dependence of the average binding time on n_s. The activation energy, E_{act} is kept at 0 for the above plots. For, $n_s = 8$, we find $T_1 = 0.000395$ secs as against 0.00025 secs as estimated from the experimental rate constant value of $10^6 M^{-1} s^{-1}$. This is a very important finding from our model. It states that for the process of ligand-protein docking no activation energy is required, i.e.

the ligand molecules do not have to overcome an energy barrier for successful docking. Indeed, biological experiments have indicated that the docking process occurs due to changes in monomer bonds into dimers and the resultant change in free energy is used for the rotational motion of the ligand to achieve the final docked conformation. Thus this finding corroborates the validity of our model. The results were generated assuming an average of 10^5 molecules of OMTKY inside the cell.

Also it can be noted that the average time for binding (= 0.000395 secs) is very high compared to our estimate of T_1^r. Thus it can be inferred that the time taken for the rotational motion of the ligand is negligible in comparison to T_1^c.

It is to be noted that p_b as calculated above also corresponds to the number of collisions in time Δt of the ligand molecule with the protein. And for our assumption of at most one collision taking place in Δt to hold, we have to make sure that $0 \leq p_b \leq 1$ (this is also true because p_b is a probability). Thus the *regions to the right of the vertical lines* corresponding to each n_s plot denotes the forbidden region where $p_b > 1$ even though $0 \leq p \leq 1$. This gives us an estimate of the allowable Δt values for different n_s's such that T_1 indeed remains constant. Out estimates show that with $\Delta t \leq 10^{-8}$, T_1 remains constant for most values of n_s.

5.4.2.5 Dependence of T_1 on n_s

Fig 5.10 plots T_1 against the different possible n_s values and we find that the average time for docking decreases as the total number of docking points n_s is increased. This is again logical as the ligand molecules now have more options for binding resulting in a higher value of p_f and subsequently p_t.

5.4.2.6 The stochastic nature of the docking time

Fig 5.11 plots the cumulative distribution function (CDF) for the total time of binding with $E_{act} = 0$. The time for collision followed an *exponential distribution* (as the calculated mean was very close to the standard deviation). Also, because the T_1^r component is very small in comparison to T_1^c, the overall time for binding can be approximated to follow an exponential distribution given by Eq 5.33. Note that incorporating $T_1^r \ll T_1^c$ in Eq 5.39 we get Eq 5.33 implying that the total time for docking is dominated by the exponential distribution outlined in Eq 5.33.

Fig 5.12 illustrates the dependence of the average time for docking (T_1) on the number of protein (Human Leukocyte elastase) molecules in the cell for a fixed number of ligand (OMTKY) molecules ($\approx 10^5$). The corresponding time of reactions estimated from the experimental rate constant of $10^6 M^{-1}s^{-1}$ have also been reported. The docking time estimates from our theoretical model very closely matches the experimental estimates in the acceptable range of the number of protein molecules (varied from $10^3 - 10^9$ molecules as can be found in any standard human cell).

5.4.3 Important observations

1. Our model achieves the experimental rate constant estimate with *zero* activation energy requirement for the protein-ligand pair under consideration in human cells. The stochastic nature of protein-ligand binding time can be approximated by a general distribution with pdf of the form given in Eq 5.39 and first and second moments given by T_1 and T_2 respectively. However, for this protein-ligand pair, the total docking time can be approximated as an exponential distribution with pdf given by Eq 5.33 as $T_1^r \ll T_1^c$.

2. The average time for DNA-protein binding is independent of Δt and decreases as the length of the docking site increases (i.e., as n_s increases).
3. An acceptable estimate of Δt is 10^{-8} secs. Fig 5.9 shows the dependence of the average time on Δt. We find that a wider range of Δt is available (keeping $p_b \leq 1$) as n_s decreases.
4. The mean of the total docking time (T_1) decreases as the length of the docking site (n_s) increases.
5. The average angle of rotation (θ) for the ligand to reach the final docked conformation is very small. This coupled with the fact that the average angular velocity of the docking site on the ligand axis being very high makes the mean time taken for rotation negligible in comparison to the collision theory component of the docking time.

5.5 Discussion

5.5.1 Limitations of our model

5.5.1.1 Maxwell-Boltzman distribution of molecular velocities

As mentioned before, the Maxwell-Boltzmann distribution gives a good estimate of molecular velocities of proteins in the cytoplasm. However, the velocity distribution should incorporate the properties of the cytoplasm, the protein/ligand structure and also the electrostatic forces that come into play.

5.5.1.2 3-D protein/ligand structure

Another point to note is that the p_f estimation can be improved by considering the 3-D structures of the protein and the ligand. Ideally, the motifs of the protein/ligand molecules are located towards the outer surface such that our straight line assumption of

the docking sites are quite realistic. However, the denominator in the expression for p_f^i considers all possible atoms on the protein/ligand molecules. However, due to their 3-d structure, not all of these molecules are exposed towards the outer protein surface that the ligand can collide to. As a result our estimates of p_f^i is actually a little lower than what should be a good estimate for the same, resulting in a corresponding decrease in p_f and hence p_t and a resultant increase in T_1^c and hence T_1. This might as well explain the slightly greater time reported from our model in comparison to the experimental estimates (recall that the experimental estimate was 0.00025 secs as against the 0.000395 secs reported by our model).

5.5.1.3 Straight line assumption of sites on protein/ligand backbones

As mentioned before, we have approximated the docking site on the protein/ligand axes as straight lines to simplify the computations of the average angle of rotation θ_{avg} and subsequently the average time required for rotation, T_1^r. However, because $T_1^r \ll T_1^c$, the T_1^r component of T_1 is negligible and the results reported from our theoretical model are quite close to experimental estimates. We are working on this aspect to identify a better estimate of T_1^r that models the actual docking process more closely.

5.5.2 Biological implications

5.5.2.1 Several ligands coming into the cell for docking

If we consider several ligands searching for their docking sites on the protein simultaneously, our results still remain valid. Note that as the number of ligands increase in the cell, the *binding rate* will increase. Assuming the docking time to be completely characterized by the collision theory part, an analytical estimate of the binding rate in such cases can be achieved by using the batch model for cytoplasmic reactions. However,

the time taken for any particular ligand to bind to the corresponding protein molecule still remains the same. Thus increasing the number of ligands should not change the results that we report for any particular ligand. In fact, this discrepancy arises because of the definition of the binding rate the inverse of which gives the time required for a successful docking to occur between the protein-ligand pair. Looking into the problem from one specific ligand's perspective (as we do in this thesis), the average time required for docking will be the same assuming there are enough number of protein molecules in the cell. This is a salient feature of our stochastic simulation paradigm where we track the course of events initiated by any particular molecule in the cell to study the dynamics of the entire cell. However, this may cause molecular crowding (of ligands) in the cell which can have an impact on the search time. Further studies are required to cover this aspect of ligand-protein docking.

5.5.2.2 Funnels and local organization of sites

Local arrangement of the binding sites of proteins tend to create a funnel in the binding energy landscape leading to more rapid binding of cognate sites. Our model assumes no such funnels of energy field. If the ligands spend most of their search time far from the cognate site our model will remain valid and no significant decrease in binding time is expected.

5.6 Summary

We have presented a computationally simplified model to estimate the ligand-protein binding time based on collision theory. The model is robust enough as the major contributing factors (molecular motion) are captured in a reasonably accurate way for general cell environments. For an extreme cell environment condition, where the influence of the electrostatic force will be significantly different, the model will not provide

such accuracy. We are exploring the possibility to modify the velocity distribution to capture the effect of this extreme cell environment. However, the model is computationally fast and allows our stochastic simulator to model complex biological systems at the molecular level (i.e., that involves many such docking events). The complexity of the 3-d protein/ligand structures have been simplified in this chapter to achieve acceptable estimates of the holding time of the ligand-protein binding event. We found that no activation energy is required for the docking process and the rotational energy for ligand-protein complex to attain the final docked conformation is contributed by the total change in free energy of the complex. The proposed mechanism has important biological implications in explaining how a ligand can find its docking site on the protein, in vivo, in the presence of other proteins and by a simultaneous search of several ligands. Besides providing a quantitative framework for analysis of the kinetics of ligand-protein binding, our model also links molecular properties of the ligand/protein and the structure of the docking sites on the ligand/protein backbones to the timing of the docking event. This provides us with a general parametric model for this biological function for our discrete-event based simulation framework. Once the model is validated for a few test cases, it can serve as a parametric model that can be used for all ligand-protein binding scenarios where the binding details are available. This may eliminate the necessity of conducting specific experiments for determining the rate constants to model a complex biological process.

CHAPTER 6

MARKOV CHAIN BASED BIOCHEMICAL SYSTEM ANALYSIS

The molecular networks regulating basic physiological processes in a cell are generally converted into rate equations assuming the number of biochemical molecules as deterministic variables. At steady state these rate equations gives a set of differential equations that are solved by a computer using numerical methods. The recent identification of the stochasticity of the biochemical environment motivates us to propose a mathematical framework for analyzing such biochemical molecular networks. The stochastic simulators that solve a system of differential equations is one technique that includes this stochasticity in the model, but suffer from simulation stiffness and require huge computational overheads. This chapter describes a new markov chain based model to simulate such complex biological systems with reduced computation and memory overheads. The central idea is to transform the continuous domain chemical master equation (CME) based method into a discrete domain of molecular states with corresponding state transition probabilities and times. Our methodology allows the basic optimization schemes devised for the CME and can also be extended to reduce the computational and memory overheads appreciably at the cost of accuracy. The simulation results for the standard Enzyme-Kinetics and a simple Transcriptional Regulatory biological systems show promising correspondence with the CME based methods and point to the efficacy of our scheme.

This chapter is organized as follows: Section 6.1 discusses some related works on biochemical system simulation. Sections 6.2 presents our markov-chain based stochastic biochemical system simulator. Section 6.3 reports the results for sample Enzyme-Kinetics

and Transcriptional regulatory systems. In Section 6.4, we discuss the biological implications of our model and its limitations. Finally, in Section 6.5 we summarize the findings of this chapter and discuss about the future improvements.

6.1 Background: stochastic biochemical system analysis

In a stochastic biochemical system, the state of the system at any time is defined by the number of molecules of each type. The transition from one state to another is derived from the probability of the reactions at the current state and the resulting next state is the new molecular state. As the molecular reactions in a biological process occur due to the random collision of the molecules, the state transition parameters are random and the state space is discrete. Let us assume in a stochastic biochemical system there are M elementary (monomolecular or bimolecular) irreversible reaction channels, which react at random times. A monomolecular reaction converts a reactant molecule into one or more product molecules. A bimolecular reaction converts two reactant molecules into one or more product molecules. We can decompose a reaction channel that involves more than two reactant molecules into a cascade of elementary reaction channels and model a reversible reaction channel by two irreversible reaction channels. The state of a stochastic biochemical system at time t is characterized by the M-dimensional random vector

$$Z(t) = [Z_1(t)Z_2(t)...Z_M(t)]^T$$

where $Z_m(t) = z$, if the m^{th} reaction has occurred z times during the time interval $[0, t)$ and T denotes vector or matrix transposition. The random variable $Z_m(t)$ is referred to as the degree of advancement (DA) of the m^{th} reaction [65]. Also $X_n(t)$ denotes the number of molecules of the n^{th} reactant or product species present in the system at time t. By assuming N distinct species, we have

$$X(t) = [X_1(t)X_2(t)...X_N(t)]^T$$

Given that the biochemical system is at state $X(t) = x$ at time t, let $q_m(x)$ be the number of all possible distinct combinations of the reactant molecules associated with the m^{th} reaction channel when the system is at state x. Note that

$$q_m(x) = \left\{ \begin{array}{cc} x_i, & \text{for monomolecular reactions} \\ x_i(x_i-1)/2, & \text{for bimolecular reactions} \\ & \text{with identical reactants} \\ x_i x_j, & \text{for bimolecular reactions} \\ & \text{with different reactants} \end{array} \right\}$$

for some $1 \leq i, j \leq N, i \neq j$. Moreover, let $c_m > 0$ be the probability per unit time that a randomly chosen combination of reactant molecules will react through the m^{th} reaction channel. This probability is known as the specific probability rate constant of the m^{th} reaction. Then, the probability that one m^{th} reaction will occur during a time interval $[t, t+dt)$ will approximately be equal to $\pi_m(x)dt$, for a sufficiently small dt, where

$$\pi_m(x) = c_m q_m(x),\ m \in M = \{1, 2, ..., M\},$$

is known as the propensity function of the m^{th} reaction channel [31, 32]. Note that, given the state $z(t)$ of the biochemical system at time t, we can uniquely determine the state $x(t)$ of the system at time t. This is because

$$X_n(t) = g_n(Z(t)) = x_{0,n} + \sum_{m \in M} s_{nm} Z_m(t), \quad t \geq 0, \tag{6.1}$$

where $x_{0,n}$ is the initial number of molecules of the n^{th} species present in the cell at time $t = 0$ and s_{nm} is the stoichiometric coefficient. This coefficient quantifies the change in the number of molecules of the n^{th} molecular species caused by one occurrence of the m^{th} reaction. The state $z(t)$ cannot be determined from $x(t)$ in general since there might be several states $z(t)$ that lead to the same state $x(t)$. To distinguish $Z(t)$ from $X(t)$, all

existing works on stochastic simulation refer to $Z(t)$ as the hidden state and to $X(t)$ as the observable state and use a hidden markov model to analyze the system.

The discrete-valued random process

$$Z = \{Z(t), t \geq 0\}$$

characterizes the dynamic evolution of the hidden state of a biochemical system. This process is specified by the probability mass function (PMF)

$$P_z(z;t) = Pr[Z(t) = z|Z(0) = 0],$$

for every $t \geq 0$. Simple probabilistic arguments show that $P_z(z;t)$ satisfies the following first-order differential equation [34]:

$$\frac{\partial P_z(z;t)}{\partial t} = \sum_{m \in M} \alpha_m(z - e_m)P_z(z - e_m;t) - \alpha_m(z)P_z(z;t),$$

for $t > 0$, with initial condition $P_z(0;0) = 1$, where e_m is the m^{th} column of the $M \times M$ identity matrix and

$$\begin{aligned} \alpha_m(z) &= \pi_m(g(z)) = c_m q_m(g(z)), \\ g(z) &= [g_1(z)g_2(z)...g_N(z)]^T \end{aligned}$$

This is the well-known forward Kolmogorov differential equation [93, 94, 6] governing the stochastic evolution of a continuous-time Markov chain. In computational biochemistry, Eqn. 6.1 is referred to as the chemical master equation (CME) [65]. It turns out that Z is a multivariate birth process [93, 6] and X is a multivariate birth-death process.

6.2 Our markov chain based formulation

Our approach is to replace the hidden markov model based approach by a Markov Chain based approach to model a composite biochemical system. Note that the system only represents biochemical reactions inside the cell or protein-ligand docking mechanisms. Thus in the Markov Chain, each state transition occurs due to *one* reaction or

docking event. If multiple reaction or docking events are possible, then the state transitions can occur due to any one of those reaction/docking events and hence there can be multiple transition paths to the next state. The *states* in the Markov Chain are defined as the number of molecules of the different components in the biological system that we are considering, i.e., by $X(t) = [X_1(t), X_2(t), ..., X_N(t)]$. For example, consider the following biochemical system:

$$R_1 : X_1 + X_2 \longrightarrow X_3; \qquad R_2 : X_2 + X_4 \longrightarrow X_5$$

where, X_1, X_2, X_4 are proteins and X_3, X_5 denote the docked complexes. Then each state in the Markov Chain will have 5 tuples corresponding to the number of molecules of these 5 components. The corresponding Markov Chain with the possible state transitions is shown in Fig 6.1. Note that each transition signifies either an R_1 or an R_2 type of event.

Thus, the total number of edges coming out of each node is given by the possible number of reaction/docking events (and equivalently the number of differential equations) considered in the system.

6.2.1 The MFPT concept

Assuming first order kinetics, the probability that a particle has reached the final state at some time t is given by $P_f(t) = 1 - e^{-kt}$ where t is the time, k is the rate, and $P_f(t)$ is the probability of having reached a final state by time t. By running many independent simulations shorter than $1/k$, one can estimate the cumulative distribution $P_f(t)$, and hence fit the value for the rate, k. The mean first passage time is the average time when a particle will reach the final state for the first time, given that it is in an initial state at $t = 0$,

$$MFPT = \int_{t=0}^{\infty} (\frac{d}{dt} P_f(t)) t \; dt = \int_{t=0}^{\infty} kte^{-kt} dt = \frac{1}{k}$$

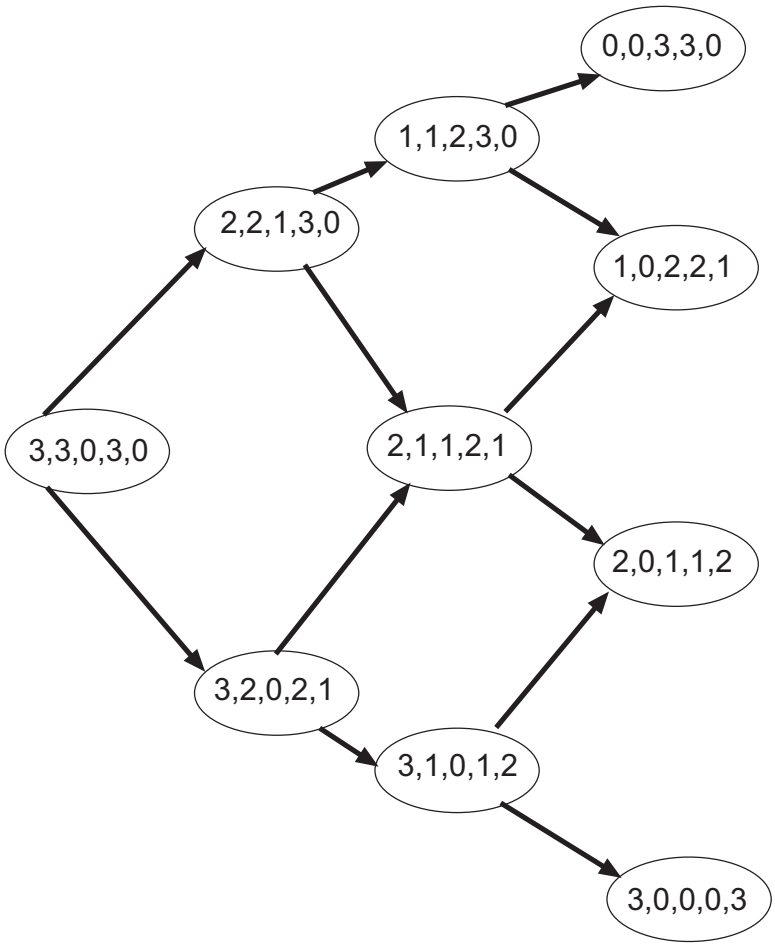

Figure 6.1. Markov Chain: 3 molecules each of X_1, X_2, X_4 and no X_3, X_5 molecules.

6.2.2 Computing the state transition probabilities and times

Note that computing the MFPT requires an estimation of each state transition probability along with the time taken for the transition. Because, each state transition signifies either a reaction or docking, we can estimate the state transition probabilities and times from the *batch models* of the reaction and docking events using concepts from collision theory. For clarity, we are including a brief description of these two models in the appendix. The batch model incorporates the number of molecules of each reactant present before the start of the reaction/docking events. This makes each state transition depend upon the current state that the system is in. Note that the batch model estimates

the time of reaction/docking as a random variable following a Gamma distribution when few reactant molecules are present in the system. However, as the number of reactant molecules increase, the mean-to-standard deviation ratio for time becomes close to 1 signifying an exponential distribution. Also, note that [73] reports that the docking time is primarily affected by the collision theory component. Hence the batch models of [71] which model the stochastic biochemical reaction time are also applicable to the docking events.

6.2.2.1 Monomolecular reactions

The time taken for monomolecular reactions can be simply computed from the experimentally determined reaction rate constant for the reaction. Denoting the reaction rate constant by k_{R_3}, the probability of reactions of type R_3 (denoted by P_{R_3}) is given by:

$$R_3 : \ X_6 \rightarrow X_7 + X_8; \quad P_{R_3} = [X_6] k_{R_3} \tau$$

where $[X_6]$ denotes the concentration of X_6 type of molecules and τ denotes a infinitely small time step (generally in the order of $\sim 10^{-8}$ secs). Note that this definition of the monomolecular reaction probability is exactly the same as that used for solving the CME and can be defined as the probability of a reaction of type R_3 occurring in time τ.

The time taken for completion of R_3 (denoted by T_{R_3}) can also be estimated from the rate constant as follows:

$$T_{R_3} = \frac{1}{[X_6] k_{R_3}}$$

In [71] we have shown that the reaction time is a *random variable* following an exponential distribution when there are sufficient number of molecules in the system. Hence, we assume that the monomolecular reaction completion time also follows an exponential distribution with mean T_{R_3}.

6.2.2.2 Bimolecular reactions

We use the batch model developed in [71] for computing the probability of reaction and first and second moments of the reaction completion times. Considering reaction R_1, the probability and time can be estimated as:

$$P_{R_1} = \frac{n_1 n_2 r_{12}^2 \tau}{V} \sqrt{\frac{8\pi k_B T}{m_{12}}}\, e^{\frac{-E_{A12}}{k_B T}}; \quad T_{R_1} = \frac{\tau}{p_{R_1}}$$

where, n_1, n_2 are the numbers of X_1 and X_2 type molecules present in the cell, r_{12} is the collision radius computed as the sum of the radii of X_1 and X_2 molecules (which are assumed to be spherical), m_{12} is the reduced mass computed as $m_{12} = \frac{m_1 m_2}{m_1 + m_2}$ (where m_1, m_2 are the masses in gm of X_1 and X_2 type molecules), V is the cell volume, T is the temperature (in Kelvin), k_B is the Boltzmann's constant $= 1.381 \times 10^{-23} kg\ m^2/s^2/K/molecule$ and E_{A12} is the activation energy required for reaction R_1. T_{R_1} denotes the *mean* of the reaction completion time which is assumed to follow an exponential distribution. Note that the Gillespie simulator also considers the reaction time to be a random variable following the exponential distribution.

In [70], we have shown that the mean of the reaction time (T_{R_1}) is actually equal to the time reported by the rate equation based model. Hence, denoting the rate of reaction R_1 by k_{R_1}, we have:

$$T_{R_1} = \frac{1}{n_1 n_2 k_{R_1}}$$

Hence the probability of reaction can also be computed if one does not know the activation energy for any specific reaction but the rate constant is known.

As before, reactions involving multiple copies of any molecule type can be represented by a cascade of elementary reactions of the above types.

6.2.2.3 Reversible Reactions

The Gillespie simulator considers reversible reactions as two separate reactions. This increases the complexity of the system as more number of reactions need to be handled. Also, in our Markov Chain based model, a reversible reaction will involve a double edge between any two nodes making the MFPT computations difficult. Hence we can approximately characterize reversible reactions using a simple birth-death model as shown in Fig 6.2.

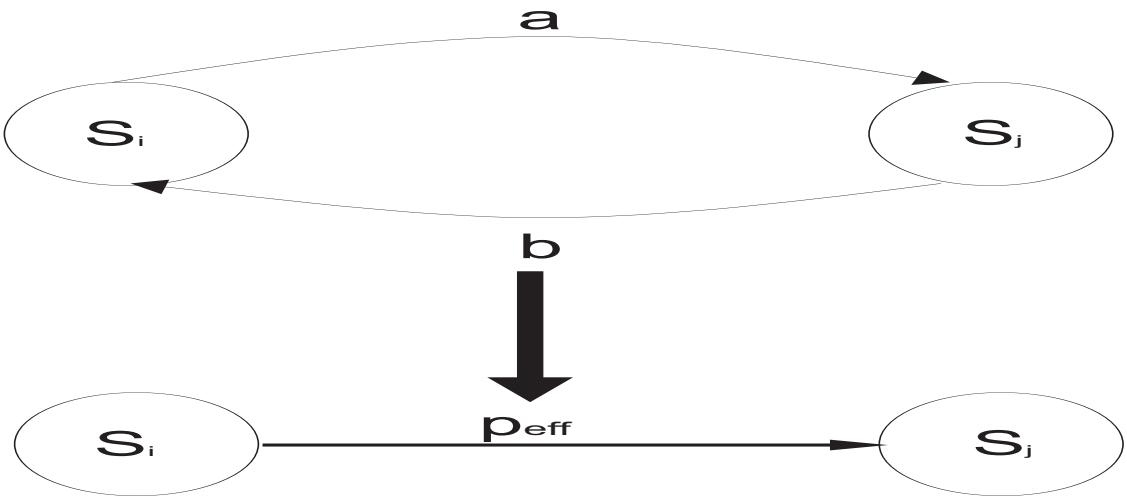

Figure 6.2. A simple birth-death model for reversible reactions.

Let us denote the forward and backward transition probabilities between any two states S_i and S_j by a and b respectively. We need to compute the *effective* probability that the reaction proceeds in the forward direction denoted by P_{eff} such that the double edge can be replaced by a single edge driving the reaction in the forward direction with probability P_{eff}. However, the time for the forward reaction still remains the

same and can be computed as above. The computation of P_{eff} will be different for the monomolecular and bimolecular reaction scenarios. In general, P_{eff} can be expressed by:

$$P_{eff} = P(S_i) \times a - P(S_j) \times b$$

where, $P(S_i)$ and $P(S_j)$ are the probabilities of being in states S_i and S_j respectively. However, $P(S_i)$ and $P(S_j)$ does not simply depend on a and b, but also on the transition probabilities of edges into and out of nodes S_i and S_j making the P_{eff} estimation quite complicated. In the following, we show two approximate schemes of computing P_{eff} for monomolecular and bimolecular reactions.

Monomolecular reactions: Consider reversible reactions of type R_1, i.e., $X_1 + X_2 \leftrightarrow X_3$. In this case, the probabilities of forward and backward reactions (a and b) can be computed as discussed before. We approximate P_{eff} as $P_{eff} = a - b$ in such cases. Note that this approximation assumes that $P(S_i) \approx P(S_j)$ for all the reversible reactions in the system. While this indeed is a gross simplification of the reversible reaction kinetics, the results obtained show that it is not overly restrictive. Moreover, when $a \approx b$, we assume that the reversible reaction attains equilibrium and make node S_i a sink i.e., no further state transitions can originate from this node.

Bimolecular reactions: Consider reversible reactions of type R_4 as follows:

$$R_4 : \quad X_9 + X_{10} \leftrightarrow X_{11} + X_{12}$$

Here also we can use the above approximation of $P(S_i) \approx P(S_j)$ and compute $P_{eff} = a - b$. However, we can slightly change the collision theory model presented in [70] to recompute the probabilities of forward and backward reactions (a and b) such that this approximation is no more required.

From [70, 71], we compute the collision probability for the forward and backward reactions as:

$$p_c^{forward} = \frac{n_9 n_{10} \pi (r_9 + r_{10})^2 U_{9,10} \tau}{V}$$
$$p_c^{backward} = \frac{n_{11} n_{12} \pi (r_{11} + r_{12})^2 U_{11,12} \tau}{V}$$

where, $U_{9,10}$ and $U_{11,12}$ denote the relative velocities between the X_9 and X_{10} molecules and that between the X_{11} and X_{12} molecules respectively. Using the reduced mass $m_{9,10} = m_9 m_{10}/(m_9 + m_{10})$ and $m_{11,12} = m_{11} m_{12}/(m_{11} + m_{12})$ for the forward and backward reactions respectively and the Maxwell-Boltzmann molecular velocity distribution we can estimate the total probability of reaction as:

$$\begin{aligned}
a &= \int_{E_{A_{9,10}}}^{E_{A_{11,12}}} \frac{(E - E_{A_{9,10}}) 4 n_9 n_{10} \pi (r_9 + r_{10})^2 \tau}{V k_B T} e^{-\frac{E}{k_B T}} \times \sqrt{\frac{1}{\pi k_B T (m_{9,10} + m_{11,12})}} dE \\
&= \frac{4 n_9 n_{10} (r_9 + r_{10})^2 \tau}{V} \sqrt{\frac{\pi}{k_B T (m_{9,10} + m_{11,12})}} \left[(E_{A_{9,10}} \right. \\
&+ \left. k_B T - 1) e^{\frac{-E_{A_{9,10}}}{k_B T}} - (E_{A_{11,12}} + k_B T - 1) e^{\frac{-E_{A_{11,12}}}{k_B T}} \right] \\
b &= \int_{E_{A_{11,12}}}^{\infty} \frac{(E - E_{A_{11,12}}) 4 n_{11} n_{12} \pi (r_{11} + r_{12})^2 \tau}{V k_B T} e^{-\frac{E}{k_B T}} \times \sqrt{\frac{1}{\pi k_B T (m_{9,10} + m_{11,12})}} dE \\
&= \frac{4 n_{11} n_{12} (r_{11} + r_{12})^2 \tau}{V} \sqrt{\frac{\pi k_B T}{(m_{9,10} + m_{11,12})}} \, e^{\frac{-E_{A_{11,12}}}{k_B T}}
\end{aligned}$$

Note that in [70, 71], the integrations ranged from the activation energy of that specific reaction to infinity as theoretically the colliding molecules can have infinite activation energy. Also, because we can assume that the forward reaction probability (and equivalently forward reaction rate) is much higher than the backward reaction probability (or rate), we will have $E_{A_{9,10}} \ll E_{A_{11,12}}$. The integration range for computing a is kept between $E_{A_{9,10}}$ and $E_{A_{11,12}}$ to ensure that the backward reactions does not occur. Also, the Maxwell-Boltzmann velocity distribution is normalized with effective reduced mass

given by $(m_{9,10} + m_{11,12})/2$ such that it can be applied to both the forward and backward reactions. Hence we can approximate P_{eff} as:

$$P_{eff} = a + b(1 - \frac{p_c^{backward}}{p_c^{forward}})$$

Note that the second term in the above expression accounts for the number of molecules creating the forward reaction having activation energy greater than $E_{A_{11,12}}$. Also, when the forward and backward collision probabilities are the same, P_{eff} is equal to a, i.e., the adjusted value of the forward reaction probability (with the contribution from the backward reaction deducted in terms of the activation energy).

6.2.3 Pruning the markov chain

As mentioned before, we will estimate the time taken to reach *any* node in the markov chain by using the MFPT. Hence, we consider each node in the chain as a sink to compute its MFPT. Also, it has to be ensured that every node in the Markov Chain is able to reach the sink. Otherwise, since these nodes will have an infinite mean first passage time, calculations done on the Markov Chain will fail. We identify the nodes that can reach the sink by performing a depth first search from the sink over the incoming edges, and marking all nodes that are reachable. The nodes that were not marked can be simply deleted, thus ensuring that all nodes in the Markov Chain can reach a node in the final state. Next, we normalize the probabilities on all the edges so that on each node, the sum of the probabilities for all outgoing edges is one as follows:

$$P_{ij}^{new} = \frac{P_{ij}}{\sum_{edge_k} P_{ik}}$$

The probability on each edge equals the number of times that transition was made divided by the total number of transitions from that node.

6.2.4 Computing the total probability of reaching a final state

The Markov Chain consists of a set of nodes and a set of transitions or edges between these nodes. Each edge has a probability associated with it as well as the time taken to traverse this edge. We define the P_{sink} of a node as the probability that the system starting in the initial state would reach the sink state before reaching the initial state again. Following [68] we will use the Markov Chain to calculate the P_{sink} values. The P_{sink} can be defined conditionally based on the first transition made from the node as follows:

$$P_{sink}(node_i) = \sum_{transition(i,j)} P(transition(i,j) \times P_{sink}(node_i|transition(i,j))$$

where the sum is over all possible transitions (that are mutually exclusive) from $node_i$. The possible transitions from $node_i$ are simply all of the edges leading from $node_i$, and the probability of each of these transitions is the P_{ij} values defined previously. This satisfies the above condition. $P_{sink}(node_i|transition(i,j))$ is simply the P_{sink} of $node_j$ which results in the following equations:

$$\begin{aligned} P_{sink}(node_i) &= \sum_{edge_{ij}} P_{ij} P_{sink}(node_j), \\ P_{sink}(node_i) &= 1, \quad node_i \in sink, \\ P_{sink}(node_i) &= 0, \quad node_i \in source \end{aligned}$$

Thus the probability of reaching any node in the chain can be estimated by a simple recursive procedure that traverses the chain. Note that in the worst case, the chain becomes a tree, where each node can traverse to M *different* new nodes (M being the number of reactions considered). Hence the worst case time complexity of traversing the chain is $O(V+E) \approx O(E)$, where V, E are the number of vertices and edges of the chain. This is because the number of edges is generally greater than the number of vertices in the chain. In the worst case we might have a tree where $E = V - 1$. Also, as the

probability has to be computed for each node in the chain, we have an overall complexity of $O(VE)$.

6.2.5 Computing the MFPT for reaching the final state

We define the mean first passage time (MFPT) of any node in the chain as the average time taken to reach that node (considered the sink) from the first node in the chain. The MFPT is defined conditionally based on the first transition made from any node:

$$MFPT(node_i) = \sum_{transition_{ij}} P(transition(i,j)) \times MFPT(node_i|transition(i,j))$$

where the sum is over all possible transitions from $node_i$. The MFPT of $node_i$ given that a transition to $node_j$ was made is the time it took to get from $node_i$ to $node_j$ added to the MFPT from $node_j$. This leads to the equation for MFPT as follows:

$$MFPT(node_i) = \sum_{edge_{ij}} P_{ij}(time_{ij} + MFPT(node_j)) \tag{6.2}$$

where the sum is over all edges leading from $node_i$. Also, we can define the initial conditions as follows:

$$MFPT(node_i) = \infty, \quad node_i \notin sink$$
$$MFPT(node_i) = 0, \quad node_i \in sink$$

Note that *time* is a random variable, and hence cannot be added as shown in the equations above. Hence we need to compute the convolution of exponential distributions that has to replace a simple addition of this random variable. Equivalently, it should be understood that *the MFPT is no longer fixed, but is also a random variable.*

We need general expressions for the following two types of convolutions of exponential distributions:

1. General expression for $n+1$-fold convolution of exponential variables from an n-fold convolution for the $(time_{ij} + MFPT(node_i))$ component of Eqn 6.2:

$$\begin{aligned} f_n &= a_1^n e^{-\frac{x}{T_1}} + a_2^n e^{-\frac{x}{T_2}} + ... + a_n^n e^{-\frac{x}{T_n}} \\ f_{n+1} &= \frac{T_1}{T_1 - T_{n+1}} a_1^n e^{-\frac{x}{T_1}} + \frac{T_2}{T_2 - T_{n+1}} a_2^n e^{-\frac{x}{T_2}} + ... + \frac{T_n}{T_n - T_{n+1}} a_n^n e^{-\frac{x}{T_n}} \\ &- [\frac{T_1}{T_1 - T_{n+1}} a_1^n + \frac{T_2}{T_2 - T_{n+1}} a_2^n + ... + \frac{T_n}{T_n - T_{n+1}}] e^{-\frac{x}{T_{n+1}}} \\ &\Rightarrow f_{n+1} = a_1^{n+1} e^{-\frac{x}{T_1}} + a_2^{n+1} e^{-\frac{x}{T_2}} + ... + a_{n+1}^{n+1} e^{-\frac{x}{T_{n+1}}} \end{aligned}$$

where, $T_1, T_2, ..., T_n$ denote the means of the reaction times of each edge of the n-fold convolution (convolution of the times for n edges gives an n-fold convolution), and $T_{n+1} = time_{ij}$ in the $(time_{ij} + MFPT(node_i))$ component of Eqn 6.2. While the above expression gives the general distribution for the $n+1$-fold convolution, the first and second moments can also be generically expressed as follows:

$$\text{First Moment} = F^{n+1} = a_1^{n+1}(T_1)^2 + a_2^{n+1}(T_2)^2 + ... + a_{n+1}^{n+1}(T_{n+1})^2$$

$$\text{Second Moment} = S^{n+1} = a_1^{n+1}(T_1)^3 + a_2^{n+1}(T_2)^3 + ... + a_{n+1}^{n+1}(T_{n+1})^3$$

After a few manipulations it can be shown that the first and second moments of this general distribution reduces to:

$$\begin{aligned} F^{n+1} &= T_1 + T_2 + ... + T_{n+1}; \\ S^{n+1} &= S^n + T_{n+1}(\sum_{i=1}^{n+1} T_i); \\ S^1 &= (T_1)^2 \end{aligned}$$

2. General expression for a convolution between an n-fold convolution (f_n) and an m-fold convolution (g_m) for the $(\sum_{edge_{ij}})$ component of Eqn 6.2:

$$f_n \otimes g_m = \sum_{j=1}^{m} \sum_{i=1}^{n} a_i^n a_j^m (\frac{e^{-\frac{x}{T_i^n}} - e^{-\frac{x}{T_j^m}}}{\frac{1}{T_j^m} - \frac{1}{T_i^n}})$$

Note that the above expression contains $m + n$ terms in total and the first and second moments of this general distribution can also be computed in a similar manner as before.

Moreover, because of the simplified expression for the first moment of the MFPT, we can use the same expression as in Eqn 6.2 if one is only interested in the *mean* value of the MFPT itself. In the next section we report the results based on this mean value of the MFPT distribution. However, it is also possible to compute the exact MFPT distribution of each node in the chain.

It should be noted that the above expressions for the general distribution of the MFPT and corresponding first and second moments were derived assuming $T_i \neq T_j$, for all i, j. This will be true for most cases as it is quite unlikely that the mean of the reaction times are equal (because the mean also depends on the concentration of the reactant molecules and most states in the chain will have different concentrations of the particular reactants of the specific reaction). However, in certain cases, the mean reaction times might be equal and we need to add a small δ to make them different such that the above reactions remain valid. Consider a 2-fold convolution of exponentially distributed random variables with means T_1 and T_2. If $T_1 = T_2$, the general distribution takes the form $\frac{xe^{-\frac{x}{T_1}}}{T_1}$, and when $T_1 \neq T_2$, it is of the form $\frac{(e^{-\frac{x}{T_1}} - e^{-\frac{x}{T_2}})}{T_1 - T_2}$. However, with $\delta = T_1 - T_2$, we can show that

$$\lim_{\delta \to 0} \frac{(e^{-\frac{x}{T_1}} - e^{-\frac{x}{T_2}})}{T_1 - T_2} = \frac{xe^{-\frac{x}{T_1}}}{T_1}$$

Hence, smaller the value of δ, the more precise are the results obtained.

6.2.6 Approximations: reducing complexity at the cost of accuracy

In most cases, it is not possible to derive an analytical solution of the CME. The following approximation techniques have been proposed to reduce the complexity of the CME:

1. Langevin approximation (LA) [32]: A useful approximation to the CME is obtained by assuming that there exists a time step dt such that the following two conditions are satisfied:
   - Changes in the hidden system states that occur during any time interval $[t, t+dt)$ do not appreciably affect the propensity functions.
   - The expected number of occurrences of each reaction in a time interval $[t, t+dt)$ is much larger than one.

   It can be shown that, under both conditions, the dynamic evolution of the hidden state process is governed by a simpler system of stochastic differential equations that can be solved by the Monte Carlo estimates.
2. Linear Noise approximation (LNA) [24, 49]: Unfortunately, the LA method does not allow us to obtain an expression for the joint probability density function (PDF) of the hidden states. However, by using additional approximations, the hidden states can be characterized by a multivariate Gaussian PDF that can be solved numerically (e.g., by the standard Euler method) and is faster than the Monte Carlo method. However, both the LA and LNA methods require both conditions (shown above) to be satisfied simultaneously which is not possible in most biological systems.
3. Poisson approximation (PA) [99]: A better approximation of the HMM is obtained by employing a time step dt satisfying the first condition, but may not necessarily satisfy the second one. Since reactions that occur during the time interval $[kdt, (k+1)dt)$ will not appreciably change the values of the propensity functions,

these reactions will occur independently of each other. Moreover, the number of occurrences of the m^{th} reaction during $[kdt, (k+1)dt)$ is assumed to be a Poisson random variable.

4. Mean-Field approximation (MFA) [45]: The PA method does not allow us to derive an expression for the joint PMF of the hidden states. However, it is possible to approximately characterize the hidden states by a PMF by the dynamic evolution of the normal Gibbs distribution. This method is superior to the LNA method for three main reasons:
   - It is based on the more accurate Poisson approximation,
   - its approximation accuracy does not depend on the cellular volume, and
   - it does not require linearization of the underlying propensity functions.
5. Stochastic quasi-equilibrium approximation (SQEA) [46]: Most often, reactions occur on vastly different time scales e.g., the transcription and translation reactions are typically slow reactions, whereas dimerization is a fast reaction. This means that transcription and translation may occur infrequently, whereas, dimerization may occur numerous times within successive occurrences of slow reactions.

   In such cases, the Gillespie algorithm spends most of the time simulating fast reaction events. It may, however, be less important to know the activity of fast reactions in detail since the system's dynamic evolution may be mostly determined by the activity of the slow reactions. Hence, it is possible to approximate the CME by one that involves only slow reactions.

In our Markov model formulation, we do not have any hidden states as the chain can be appropriately characterized by the number of different molecule types present in the system (denoting the states of the chain), and each state transition is characterized by the corresponding reaction/docking events. Hence, most of the above techniques are not directly applicable to this formulation. However, we can employ the SQEA approach to

substantially simplify the markov chain (with lesser number of states) making the MFPT computations faster. In this case, the states of the markov chain will have the same tuples as before, however the *state transitions will only be governed by the slow reactions.* During each state transition, the new state in the chain is computed depending on this slow reaction and also computing how many fast reactions can occur in that time and appropriately updating the molecule counts of the reactants in the fast reactions.

In fact this technique has a direct analogy to Gillespie's tau-leap algorithm, wherein, we can specify a certain time step Δt, and compute how many reactions (both fast and slow) occur within that period. Thus we can compute the next state and the markov chain will become a 1-dimensional chain thereby greatly reducing the complexity. Also the memory requirements for storing the Markov chain can be completely removed as the MFPT can be computed online as the chain progresses in time.

6.3 Results and analysis

6.3.1 Enzyme-Kinetics system

In this section, we present the results for the well known Enzyme Kinetics system governed by the following three elementary reactions:

$$E.S \rightarrow P + E, \quad E + S \leftrightarrow E.S$$

The rate constant for the reversible reaction pair is set at $1s^{-1}$ and that for the first reaction is $0.1s^{-1}$.

Figs 6.3-6.5 show the molecular distributions of the product (P) molecules with time for different number of enzyme (E) and substrate (S) molecules. Note that it is possible to report the exact molecular distributions of any molecule type in the system using our approach. The time axis reports the mean value of the MFPT (which is also a random variable as discussed earlier). Fig 6.8 compares the dependency of mean

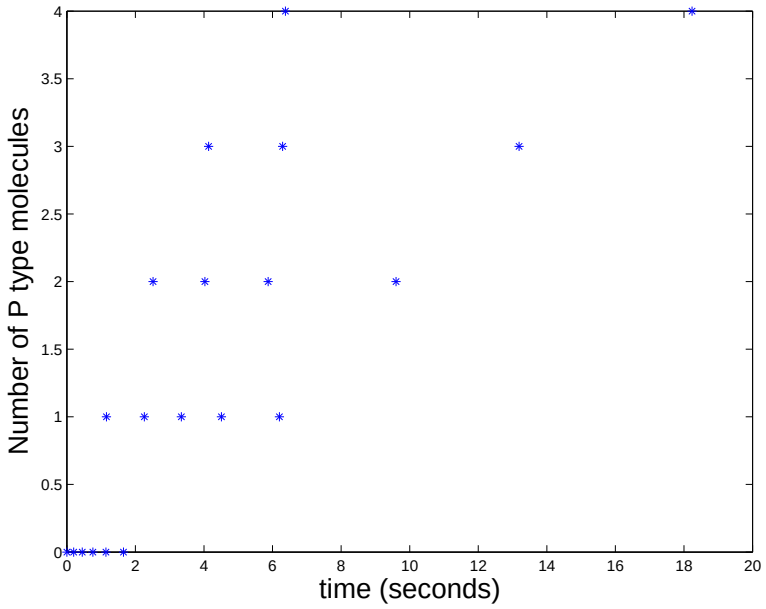

Figure 6.3. Molecular distribution of P type molecules, with E=10, S=5.

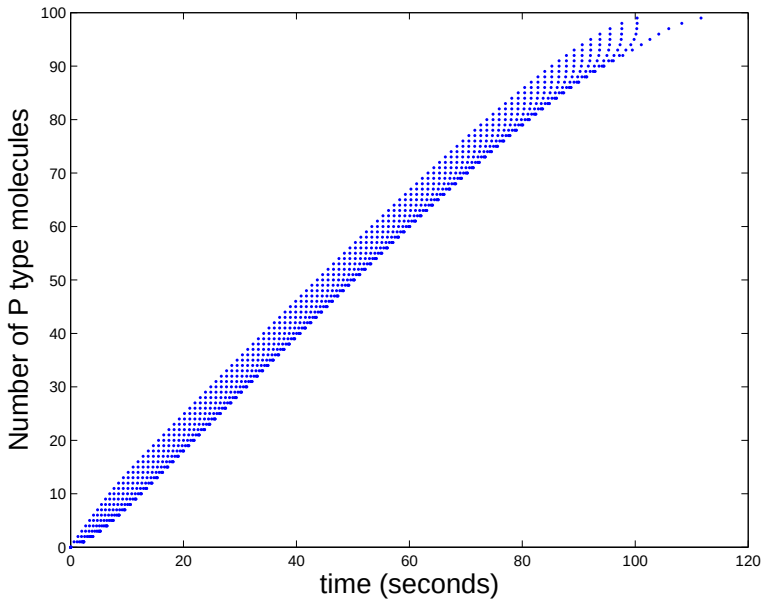

Figure 6.4. Molecular distribution of P type molecules, with E=10, S=100.

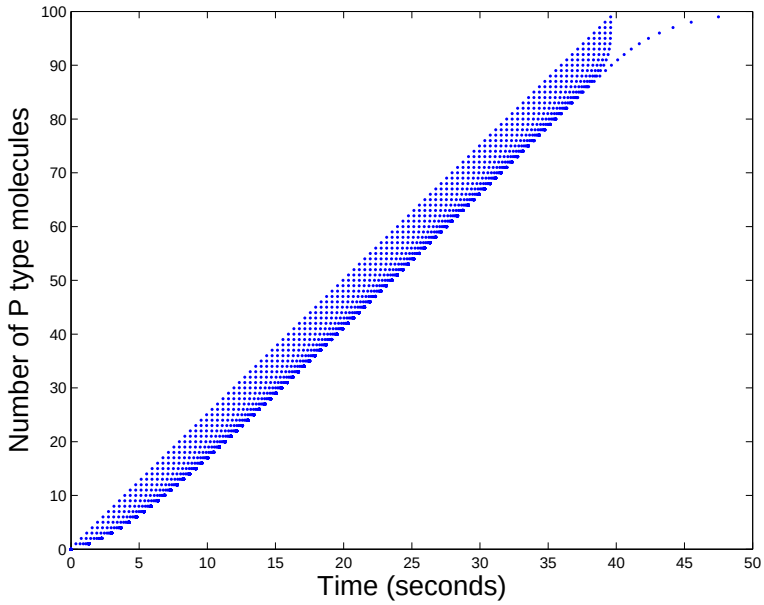

Figure 6.5. Molecular distribution of P type molecules, with E=1000, S=100.

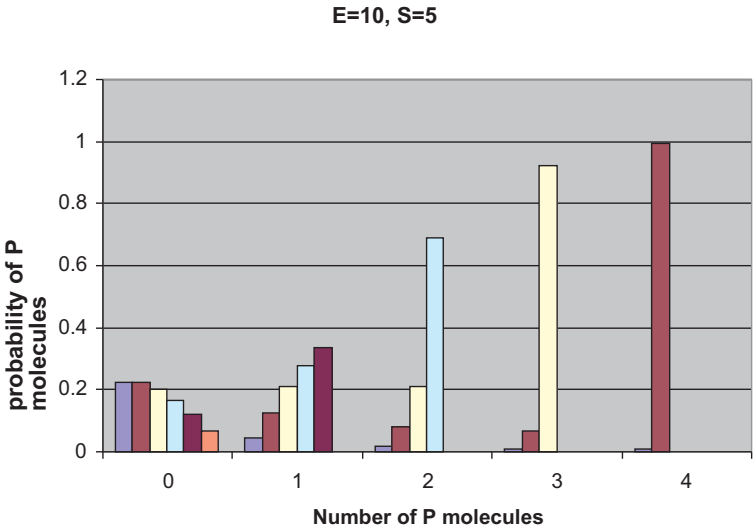

Figure 6.6. Probability distribution of P type molecules, with E=10, S=5.

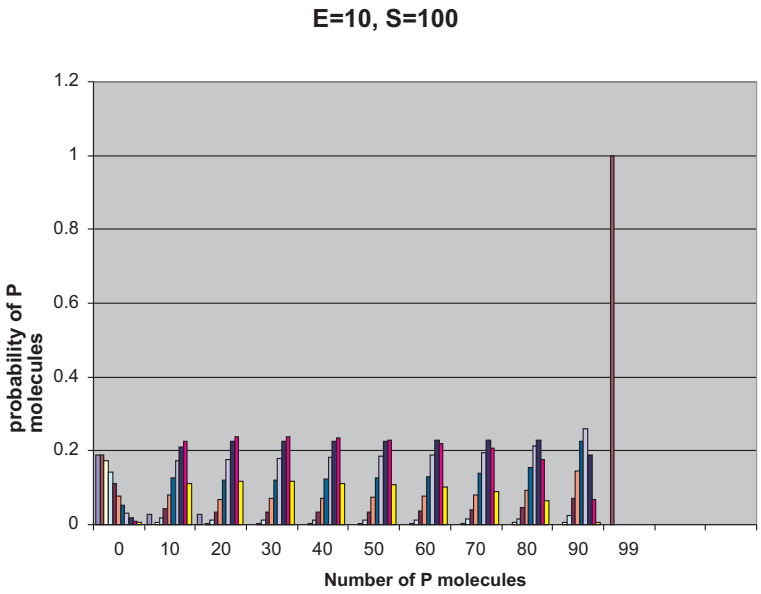

Figure 6.7. Probability distribution of P type molecules, with E=10, S=100.

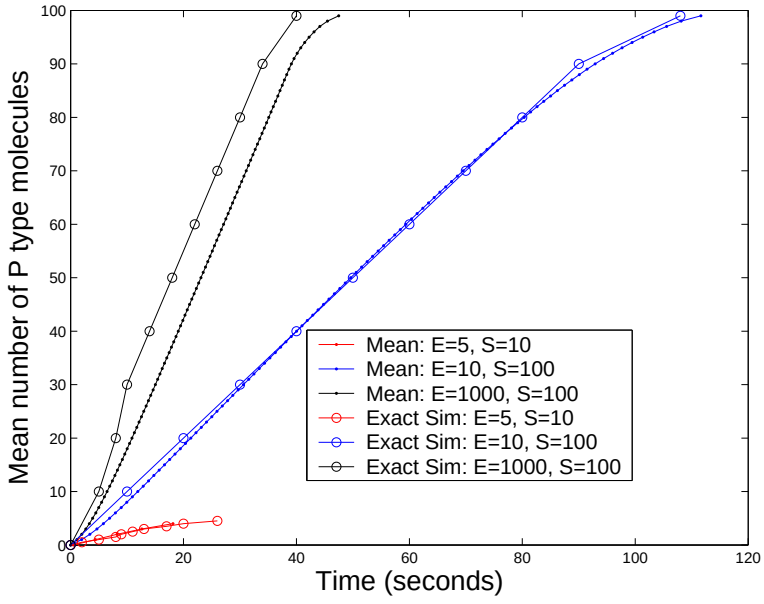

Figure 6.8. Mean number of P type molecules, Our model Vs Exact Simulation.

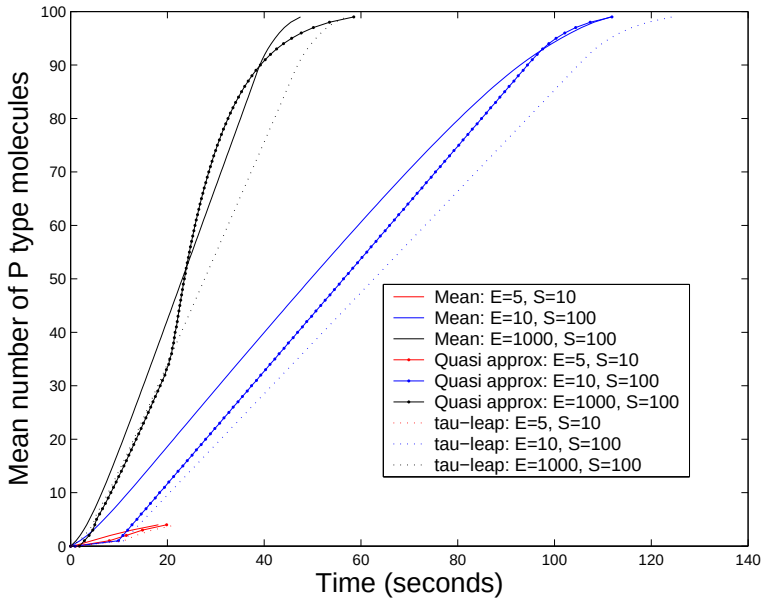

Figure 6.9. Effects of SQEA and Tau-leaping approximations.

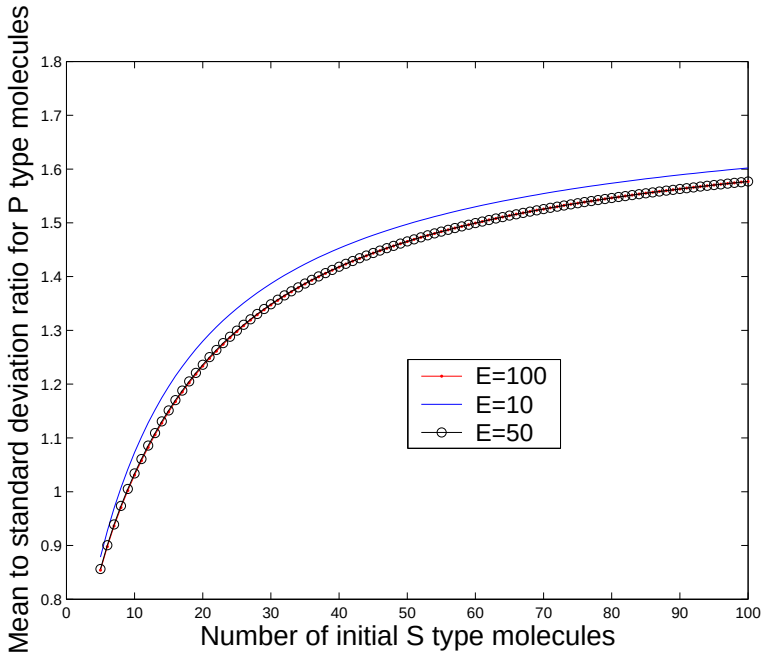

Figure 6.10. Mean to s.d. ratio of P molecules (constant no. of enzymes).

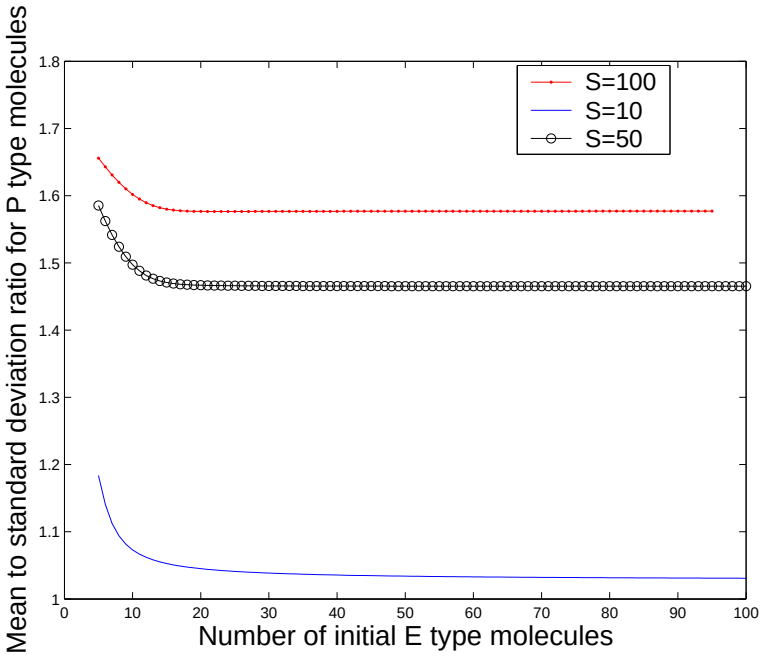

Figure 6.11. Mean to s.d. ratio of P molecules (constant no. of substrates).

number of P type molecules on time with that reported from an exact simulation of the CME (obtained from Monte Carlo simulation of the differential equations in the system). Our results compare very well with the exact simulation for low number of molecules in the system. With large number of enzyme molecules present, the reactions occur very fast and the markov model formulation being driven in discrete time produces less accurate results. Nevertheless, it is computationally very fast and allows the study of more complicated systems (with large number of reactions and molecular types involved).

Figs 6.6-6.7 plots the probability distributions of the product molecules. The different bars at each possible molecular count value of the P type molecules correspond to the probability of reaching different states (from the initial state) in the Markov model having that number of P type molecules (and different molecular count values for the other entities in the system). It is again possible to compute the complete distribution

(not just the first and second moments) of all the different molecule types in the system with our formulation.

Fig 6.9 shows the effects of the SQEA (denoted by "quasi approx") and tau-leap approximations to our markov model. The reversible reactions are considered fast reactions in our analysis. As expected, the SQEA approach provides a very accurate approximation of the mean number of product molecules whereas the tau-leap variation (with $\Delta t = 10^{-3}$ secs) provides the fastest (and most memory efficient) solution at the cost of accuracy.

Figs 6.10-6.11 plot the mean to standard deviation ratio of the molecular distribution of the product molecules with varying number of *substrate* and *enzyme* molecules respectively. With less number of substrates, the stochastic resonance is quite high in the system (as the ratio is less than 1). With higher number of substrates, the ratio saturates at 1.5 implying lesser stochasticity in the system. Also, the stochasticity is not very much dependent on the number of enzyme molecules in the system as depicted in Fig 6.11. Thus from these plots we can infer that the stochastic resonance in the molecular distribution of the product molecules is primarily governed by the number of substrate molecules in the system.

6.3.2 Transcriptional regulatory system

We next show the results of our model for a simple transcriptional regulatory system as shown in Fig 6.12. Protein M, synthesized by transcription of a gene, dimerizes to the transcription factor D, which may bind to the gene's regulatory region at two binding sites, R_1 and R_2. Binding of D at R_1 activates transcription of M. However, binding of D at R_2 excludes the RNA polymerase from binding at the gene's promoter and in this case transcription is repressed. Table 6.1 presents the terminology used for the different components of this example system, whereas Table 6.2 shows the list of reactions involved along with their respective rate constants [45].

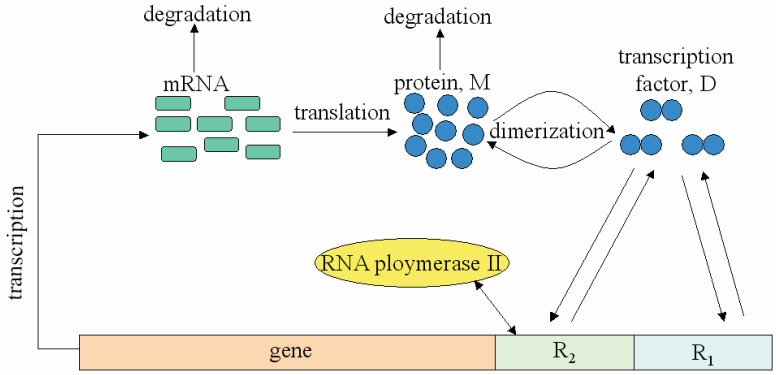

Figure 6.12. A simple transcriptional regulatory system.

Table 6.1. Terminology for the Transcriptional Regulatory System

M	Protein (monomer)
D	Transcription factor (dimer)
RNA	mRNA
DNA	DNA template free of dimers
$DNA.D$	DNA template bound at R_1
$DNA.2D$	DNA template bound at R_1 and R_2

In this system as well, we find very good agreement between the exact simulation results with that from our model. In both the example systems, the reversible reactions used the $P_{eff} = a - b$ approximation for both monomolecular and bimolecular reactions as discussed before. Thus, for reaction-pairs $\{5, 6\}$ $\{7, 8\}$ and $\{9, 10\}$ we choose the forward reactions as 6, 7 and 10 respectively and drive the Markov Chain formulation accordingly. The accuracy of our system suffers from this approximation (hence the difference from the exact simulation results).

It should be noted that these results were generated for a low number of the different molecule types in the system. As the number of molecules increase, the MFPT based results are further off from the exact simulation results because of the approximations. Thus, our model allows for a computationally efficient implementation of a complex bio-

Table 6.2. Reactions Associated with the Transcriptional Regulatory System

	Reaction	Rate Constant
1	$RNA \rightarrow RNA + M$	$0.043s^{-1}$
2	$M \rightarrow \emptyset$	$0.0007s^{-1}$
3	$DNA.D \rightarrow RNA + DNA.D$	$0.0715s^{-1}$
4	$RNA \rightarrow \emptyset$	$0.0039s^{-1}$
5	$DNA + D \rightarrow DNA.D$	$0.02s^{-1}$
6	$DNA.D \rightarrow DNA + D$	$0.4791s^{-1}$
7	$DNA.D + D \rightarrow DNA.2D$	$0.002s^{-1}$
8	$DNA.2D \rightarrow DNA.D + D$	$0.8765 \times 10^{-11}s^{-1}$
9	$M + M \rightarrow D$	$0.083s^{-1}$
10	$D \rightarrow M + M$	$0.5s^{-1}$

chemical system simulation which can give accurate results when the number of molecules of the components in the system are small. It also allows us to reduce the computational overheads appreciably by using many graph-theoretic techniques as we discuss later.

6.4 Discussion

Here we make some comments regarding both the differential equation based and our discrete random process based approach for biological system modeling. The former approach is usually used to model the variations of the concentrations of biomolecules, where the latter models the variations of the number of biomolecules. As for any research problem for which there are a variety of feasible solutions, each of these approaches has its own pros and cons. For example, when the number of biomolecules is extremely large, it may not even be practical to use our discrete random process-based model because of the following reasons:

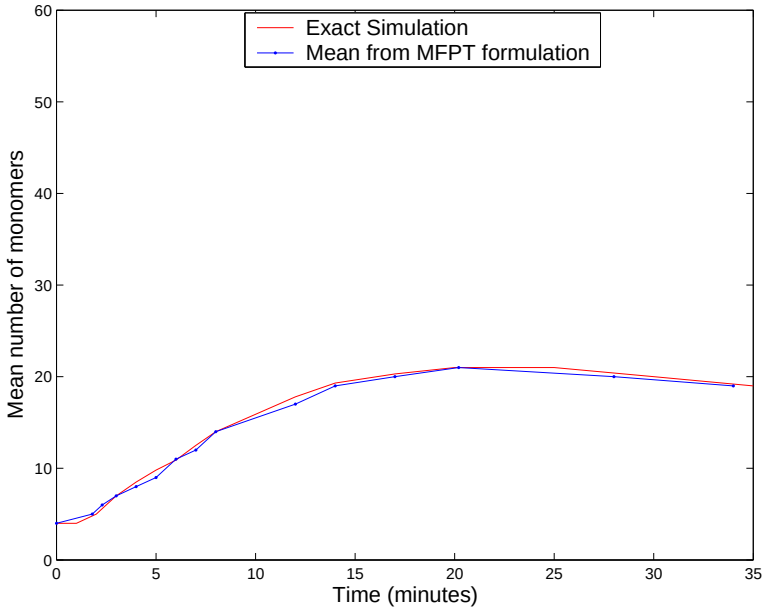

Figure 6.13. Mean number of monomers: Exact Simulation Vs Our Model.

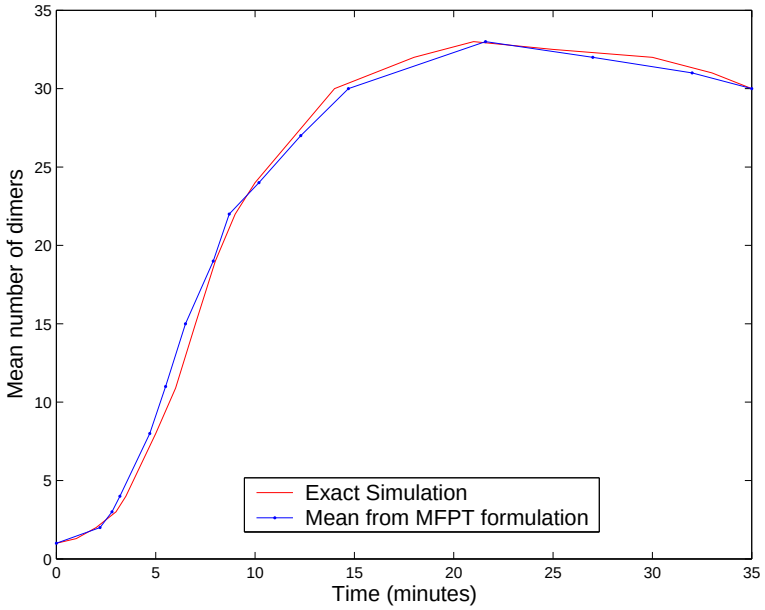

Figure 6.14. Mean number of dimers: Exact Simulation Vs Our Model.

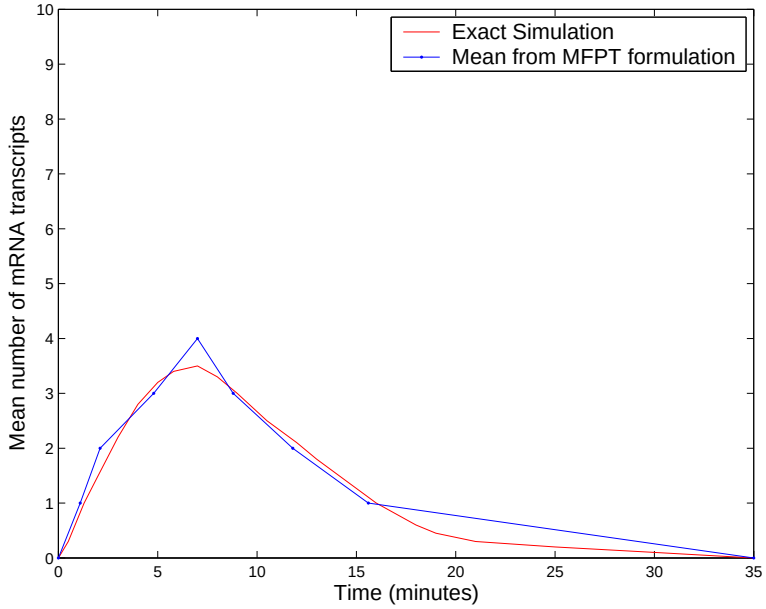

Figure 6.15. Mean number of mRNA transcripts: Exact Simulation Vs Our Model.

1. the number of possible candidate states of a molecular entity, $X(t) \in \{0, 1, ...,$the maximum number of molecules$\}$, could be too huge to handle; and
2. if a discretization strategy is used, then accuracy of the model could be compromised.

No matter which model is used, some of the parameters (e.g., kinetic parameters for the differential equation based models) need to be estimated and the estimates have to be consistent with the reality as much as possible. The parametric models we have introduced for biochemical reactions and docking (shown in the appendix) can estimate these parameters theoretically and can be used once we have sufficient fidelity in these models. However, the Markov model based approach presented in this chapter will work for both cases i.e., by estimating the kinetic parameters through controlled experiments or by using the parametric models.

6.5 Summary and future directions

We have introduced a Markov Chain based analysis technique as an alternative for complex biological process modeling. The main idea of this modeling is to transform the biological processes from a continuous deterministic process to a discrete random process. Because of its simplicity in comparison to solving numerically a large number of differential equations, our framework reduces the computational overhead and increases scalability considerably. We are currently working on a complex pathway model with many molecular types and with large number of molecules of each type to estimate the computational complexity. The main benefit of this analysis is to analyze the stochasticity of many reactions occurring together. Current experimental methods are not able to capture this measurement at a molecular level without special set-up.

The challenge in the model proposed here is the optimization of memory and computational speed of DFS and MFPT algorithms. Note that each node in the Markov Chain has an out-degree of M, where M is the number of reactions/docking considered in the system. The storage of an arbitrary graph with a large number of nodes and out-degree will have memory problems. It is also imperative to find appropriate simplifications and data structures to speed up the process. Can the chain be converted into a tree structure by eliminating/adding pseudo nodes/edges ? This will allow us to traverse the chain (during DFS or MFPT computations) in $O(log_M V)$ time. We have already stated that the tau-leap approximation on the chain reduces it to a 1-dimensional chain and the MFPT computations can be performed online. Also, can the tree structure be converted into a trie wherein the chain is compressed optimally thereby reducing the memory overheads ?

The complete cell model by this analysis may not be feasible due to the large number of molecules in the cell, but we expect that many complex biological systems

can be modeled by this technique. The event statistics thus derived can be used for a discrete event based cell simulation.

CHAPTER 7

CONCLUSION

In this dissertation, we have presented a discrete-event based framework for stochastic modeling and simulating the dynamics of complex biological systems. We have developed stochastic models for a few basic biological events (e.g., molecular transport, biochemical reactions, protein-DNA binding and protein-ligand docking) that form the building blocks for the simulator. We also explained a simple biological system, i.e., the two-component PhoPQ signal transduction system in Salmonella typhimurium to explain the simulation technique and how our stochastic event models fit into the bigger picture. While the models presented here can approximately capture most of the important biological events inside a cell, many other models are required to realize our endeavor of a complete cell simulation. Some of these are as follows:

- Gene expression duration: This model should compute the total time taken for the gene transcription and translation processes. It will play an important role in the quantifying the burstiness in mRNA production and protein generation. Some preliminary works on this can be found in [90].
- Molecular/ionic mobility in a cell: While we have modeled the basic molecular/ionic transport mechanisms (simple diffusion processes) in this thesis, some other transport mechanisms have not been covered. These include transport mechanisms using the active/passive pump system. Also, we have used the Maxwell Boltzmann distribution to model the macromolecular velocity distribution in the cell. However, its applicability outside the cell cytoplasm (e.g., nucleus or membrane) requires further research.

- Transport vesicles: In a very general sense, a transport vesicle could be any vesicle that transports material around the cell. More specifically, transport vesicles usually refer to those vesicles that transport material from the Endoplasmic Reticulum to the Golgi apparatus or from one part of the Golgi to another. Some work is required to analytically model this process.
- Protein folding duration: A lot of research has been done to study the protein folding mechanism as it determines whether a protein is correctly formed. These works are mostly based on molecular dynamic simulation and we need analytical models to consider the protein folding event in the simulation. As of now, we are simply assigning a certain probability to the correct formation of the protein, but certainly more detailed analysis is required of this important biological event.
- Protein life time duration: The protein decay event has been captured in our simulation by an exponentially distribute random variable computed from the protein decay rate (which is experimentally measured). However, analytical models explaining the protein decay process should be able to predict the protein decay rate in the absence of such experimental results.

Also, the individual models presented here can also be improved in terms of accuracy and computational speed. Our goal, however, was to present some simple models for the most important biological processes that enables us to complete the discrete-event simulator as a proof of concept.

We have also presented a markov-chain based biochemical system simulator that is less efficient than the discrete-event based simulator in terms of accuracy, but has the potential of providing higher scalability. Further work is required to analyze the accuracy and effectiveness of this simulator for larger biochemical systems. Once we build sufficient fidelity into this simulator, we can start several graph theoretic optimization techniques to improve its scalability and memory usage.

REFERENCES

[1] A.D. McCulloch and G. Huber. Integrative biological modeling in silico. *Novartis Foundation Symposium*, 2002, ISBN:9780470844809.

[2] A. Lomakin and M. Frank-Kamenetskii. A theoretical analysis of specificity of nucleic acid interactions with oligonucleotides and peptide nucleic acids (PNAs). *J. Mol. Biol.*, 1998, 276:57-70.

[3] A. Mazloom, K. Basu, and S.K. Das, A Random Walk Modeling Approach for Passive Metabolic Pathways in Gram-Negative Bacteria. *proceedings of the 2006 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2006, pp. 1-8.

[4] A. M. Kierzek, J. Zaim and P. Zielenkiewicz. The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *J Biol Chem.* 2001, 276(11):8165-8172.

[5] A. O. Grillo, M. P. Brown and C. A. Royer. Probing the physical basis for Trp repressor-operator recognition. *J. Mol. Biol.*, 1999, 287:539-554.

[6] A. Papoulis and S.U. Pillai. Probability, Random Variables and Stochastic Processes. *fourth ed. New York: McGraw-Hill*, 2002, ISBN:0073660116.

[7] BB 492/592 Spring Term 2003:
http://oregonstate.edu/instruction/bb492/figletters/FigH3.html

[8] BB 492/592 Spring Term 2003:
http://oregonstate.edu/instruction/bb492/lectures/DNAII.html

[9] BB 492/592 Spring Term 2003: http://chem-mgriep2.unl.edu/replic/SSB.html

[10] BB 492/592 Spring Term 2003: http://chem-mgriep2.unl.edu/replic/Helicase.html

[11] BioSpice: open-source biology, *http://biospice.lbl.gov/home.html*

[12] B. Alberts, D. Bray, J. Lewis, M. Raff, K. Roberts and J. D. Watson Molecular Biology of the Cell. *Garland Publishing; 3rd edition,* 1994, ISBN:0815316194.

[13] C. Camacho, Z. Weng, S. Vajda and C. DeLisi. *Biophysical Journal*, 1999, 76:1166-1178.

[14] C. Camacho, C. DeLisi and S. Vajda. *Thermodynamics of the Drug-Receptor Interactions*, 2001, ed. R. Raffa (Wiley, London), ISBN-10:0-471-72042-9.

[15] C. Camacho and S. Vajda. Protein docking along smooth association pathways. *PNAS*, 2001, 98(19):10636-10641.

[16] C. Chothia and J. Janin. Principles of protein-protein recognition. *Nature*, 1975, 256:705-708.

[17] C. Camacho, Z. Weng, S. Vajda and C. DeLisi. Free energy landscapes of encounter complexes in protein-protein association. *Biophisics J.*, 1999, 76:1166-1178.

[18] C. DeLisi and F. Wiegel. Effect of nonspecific forces and finite receptor number on rate constants of ligand-cell-bound-receptor interactions. *Proc. Natl. Acad. Sci.*, 1981, 78:5569-5572.

[19] C. E. Bell and M. Lewis. A closer view of the conformation of the Lac repressor bound to operator. *Nat. Struct. Biol.*, 2000, 7:209-214.

[20] C. E. Bell and M. Lewis. The Lac repressor: a second generation of structural and functional studies. *Curr. Opin. Struct. Biol.*, 2001, 11:19-25.

[21] C. Fall, E. Marland, J. Wagner, and J. Tyson, Computational Cell Biology. *Interdisciplinary Applied Mathematics.*, 2002, ISBN-10:0387953698.

[22] C. J. Camacho, S. R. Kimura, C. DeLisi and S. Vajda. Kinetics of Desolvation-Mediated Protein-Protein Binding. *Biophysical Journal*, 2000, 78:1094-1105.

[23] C. T. Harbison et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*, 2004, 431:99-104.

[24] C. V. Rao, D. M. Wolf and A. P. Arkin. Control, Exploitation and Tolerance of Intracellular Noise. *Nature*, 2002, 420:231-237.

[25] DARPA, BioSpice: open-source biology, *http://biospice.lbl.gov/home.html*

[26] D. Bratsun, D. Volfson, L. Tsimring, & J. Hasty, Delay-induced stochastic oscillations in gene regulation. *Proceedings of the National Academy of Sciences,* 2005, 102(41):14593-14598.

[27] D. Browning and S. Busby. The regulation of bacterial transcription initiation. *Nat. Rev Microbiol*, 2004, 2:5765.

[28] D. Endy, & R. Brent, Modeling cellular behavior. *Nature.*, 2001, 400:391-395.

[29] D. T. Gillespie, Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.*, 1977, 81(25):2340-2361.

[30] D. Gillespie, Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics.*, 2001, 115(4): 1716-1733.

[31] D.T. Gillespie. A Rigorous Derivation of the Chemical Master Equation. *Physica A*, 1992, 188:404-425.

[32] D.T. Gillespie. The Chemical Langevin Equation. *J. Chemical Physics*, 2000, 113(1):297-306.

[33] E. A. Groisman, The Pleiotropic Two-Component Regulatory System PhoP-PhoQ. *Journal of Bacteriology.*, 2001, 183(6):1835-1842.

[34] E.L. Haseltine and J.B. Rawlings. Approximate Simulation of Coupled Fast and Slow Reactions for Stochastic Chemical Kinetics. *J. Chemical Physics*, 2002, 117(15):6959-6969.

[35] G. D. Stormo and D. Fields. Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.*, 1998, 23:109-113.

[36] G. Schreiber and A. Fersht. Rapid, electrostatically assisted association of proteins. *Nature Struct. Biol.*, 1996, 3:427-431.

[37] H. Fischer, I. Polikarpov and A. F. Craievich. Average protein density is a molecular-weight-dependent function. *Protein Science*, 2004, 13:28252828.

[38] H. Fogler and M. Gurmen, Elements of Chemical Reaction Engineering. Chapter 3.1, Equation 13, online at http://www.engin.umich.edu/ cre/03chap/html/collision/

[39] H. Kitano, CellDesigner: A modeling tool of biochemical networks, *http://celldesigner.org/*

[40] H. MacAdams and A. Arkin, It is a noisy business! Genetic regulation at the nanomolar scale. *Trends in Genetics*, 1999, 15:65-69.

[41] H. McAdams and A. Arkins. Stochastic mechanisms in gene expression. *PNAS*, 1997, 94:814-819.

[42] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 2000, 28:235-242.

[43] I. M. Keseler, J. Collado-Vides, S. Gama-Castro, J. Ingraham, S. Paley, I. T. Paulsen, M. Peralta-Gil and P. D. Karp. EcoCyc: A comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 2005, 33:D334-7.

[44] J. Alvarez, and B. Hajek, Ion channels, or stochastic networks with charged customers. *Invited Talk at Stochastic networks Conference*, 2004.

[45] J. Goutsias. A Hidden Markov Model for Transcriptional Regulation in Single Cells. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2006, 3(1):57-71.

[46] J. Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *J. Chemical Physics*, 2005, 122:184102-184116.

[47] J. Hasty, and J. J. Collins, Translating the Noise. *Nature, Genet.*, 2002, 31:13-14.

[48] J. Keener, & J. Sneyd, Mathematical Physiology. *Springer.*, 1 edition, 1998, ISBN:0387983813.

[49] J.M. Raser and E.K. O'Shea. Control of Stochasticity in Eukaryotic Gene Expression. *Science*, 2004, 304:1811-1814.

[50] J. Paulsson. Models of Stochastic Gene Expression. *Phys. Life Rev.*, 2005, 2:157-175.

[51] J. Yu. Probing Gene Expression in Live Cells, One Protein Molecule at a Time. *Science*, 2006, 311(5767):1600-1603.

[52] K. Sharp, R. Fine and B. Honig. Computer simulations of the diffusion of a substrate to an active site of an enzyme. *Science*, 1987, 236:1460-1463.

[53] L Cai. Stochastic protein expression in individual cells at the single molecule level. *Nature Letters*, 2006, 440:358-362.

[54] L. Loew, The Virtual Cell Project. *'In Silico' Simulation of Biological Processes (Novartis Foundation Symposium No. 247).*, Wiley, Chichester, 2002, 207-221.

[55] L. Eltis, R. Herbert, P. Barker, A. Mauk and S. Northrup. Reduction of horse ferricytochrome c by bovine liver ferrocytochrome b_5. Experimental and theoretical analysis. *Biochemistry*, 1991, 30:3663-3674.

[56] L. Segel, I. Chet, and Y. Henis, A simple quantitative assay for bacterial motility. *J. Gen. Microbiol.*, 1977, 98(2):329-337.

[57] M. Nagasaki, A. Doi, & S. Miyano, Cell Illustrator 2.0: A Platform for Biopathway Modeling and Simulation, *www.fqspl.com.pl/life_science/cellillustrator/ci.htm*

[58] M. Rathinam, L. Petzold, & D. Gillespie, Stiffness in Stochastic Chemically Reacting Systems: The Implicit Tau-Leaping Method. *Journal of Chemical Physics.*, 2003, 119(24):12784-12794.

[59] M. Schellersheim, & G. Mack, SIMMUNE, a tool for simulating and analyzing immune system behavior. *CoRR cs.MA/9903017*, 1999.

[60] M. Slutsky and L. A. Mirny. Kinetics of Protein-DNA Interaction: Facilitated Target Location in Sequence-Dependent Potential. *Biophysical Journal*, 2004, 87:4021-4035.

[61] M. Tomita, The E-CELL Project: Towards Integrative Simulation of Cellular Processes. *New Generation Computing.*, 2000, 18(1):1-12.

[62] M. Vijaykumar, K. Wong, G. Schreiber, A. Fersht, A. Szabo and H. Zhou. Electrostatic enhancement of diffusion-controlled protein-protein association: comparison of theory and experiment on Barnase and Barstar. *J. Mol. Biol.*, 1998, 278:1015-1024.

[63] M. V. Smoluchowski. Versuch einer mathematischen Theorie der Koagulationskinetik kolloider Loeschungen. *Z. Phys. Chem.*, 1917, 92:129168.

[64] Nanomedicine, vol. I: Basic Capabilities, http://www.nanomedicine.com/NMI/3.2.5.htm

[65] N.G. vanKampen. Stochastic Processes in Physics and Chemistry. *Amsterdam: Elsevier*, 1992, ISBN-10:0444893490.

[66] N. M. Luscombe, S. E. Austin, H. M. Berman, and J. M. Thornton. An overview of the structures of protein-DNA complexes. *Genome Biol.*, 2000, 1:1-37.

[67] N. Shimamoto. One-dimensional diffusion of proteins along DNA. Its biological and chemical significance revealed by single-molecule measurements. *J. Biol. Chem.*, 1999, 274:15293-15296.

[68] N. Singhal et al. Error analysis and efficient sampling in Markovian state models for molecular dynamics. *Jour. Of Chem. Physics.*, 2005, 123:204909-204921.

[69] O. G. Berg and P. H. von Hippel. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *J. Mol. Biol.*, 1987, 193:723-750.

[70] P. Ghosh, S. Ghosh, K. Basu, S.K. Das, and S. Daefler, An Analytical Model to Estimate the time taken for Cytoplasmic Reactions for Stochastic Simulation of

Complex Biological Systems. *2nd IEEE Granular Computing Conference*, 2006, pp. 79-84.

[71] P. Ghosh, S. Ghosh, K. Basu, S.K. Das, and S. Daefler, Stochastic Modeling of Cytoplasmic Reactions in Complex Biological Systems. *6th IEE International Conference on Computational Science and its Applications (ICCSA)*, 2006, pp. 566-576.

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

[73] P. Ghosh, S. Ghosh, K. Basu, S.K. Das, and S. Daefler, A stochastic model to estimate the time taken for Protein-Ligand Docking. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2006, 1-8.

[74] P. Ghosh, S. Ghosh, K. Basu, and S.K. Das, A Diffusion Model to Estimate Inter-arrival Time of Charged Molecules in Stochastic Event based Modeling of Complex Biological Networks. *proceedings of the IEEE Computational Systems Bioinformatics Conference,* CSB 2005, 2005, pp. 17-18.

[75] P. Ghosh, S. Ghosh, K. Basu, S.K. Das and S. Daefler. Modeling the diffusion process in the PhoPQ signal transduction system: A stochastic event based simulation framework. *Intl. Symp. on Computational Biology & Bioinformatics (ISBB)*, 2006.

[76] P. Ghosh, S. Ghosh, K. Basu and S.K. Das. Holding Time Estimation for Reactions in Stochastic Event-based Simulation of Complex Biological Systems, *Elsevier Simulation Modelling Practice and Theory*, 2007.

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

[78] P. Ghosh, S. Ghosh, K. Basu and S.K. Das. Parametric modeling of protein-DNA binding kinetics: A Discrete Event based Simulation approach. *Journal of Discrete Applied Mathematics (DAM*, special issue on Networks in Computational Biology, 2008.

[79] P. H. von Hippel and O. G. Berg. Facilitated target location in biological systems. *J. Biol. Chem.*, 1989, 264:675-678.

[80] P. H. von Hippel and O. G. Berg. On the specificity of DNA-protein interactions. *Proc. Natl. Acad. Sci.*, 1986, 83:16081612.

[81] R. B. Winter, O. G. Berg and P. H. von Hippel. Diffusion-driven mechanisms of protein translocation on nucleic acids. 3. The Escherichia coli Lac repressor-operator interaction: kinetic measurements and conclusions. *Biochemistry*, 1989, 20:6961-6977.

[82] R. Gabdoulline and R. Wade. Simulation of the diffusional association of barnase and barstar. *Biophisics J.*, 1997, 72:1917-1929.

[83] R. L. Lucas, and C. A. Lee, Unravelling the mysteries of virulence gene regulation in Salmonella Typhimurium. *Journal of Molecular Biology.*, 2000, 36(5):1024-1033.

[84] R. S. Spolar and M. T. Record. Coupling of local folding to sitespecific binding of proteins to DNA. *Science*, 1994, 263:777-784.

[85] R. Steuer, Z. Changsong, & J. Kurths, Constructive effects of fluctuations in genetic and biochemical regulatory systems. *BioSystems,* 2003, 72:241-251.

[86] R. Stone, S. Dennis and J. Hofsteenge. Quantitative evaluation of the contribution of ionic interactions to the formation of thrombin-hirudin complex. *Biochemistry*, 1989, 28:6857-6863.

[87] S.K. Das, F. Sarkar, K. Basu, & S. Madhavapeddy, Parallel Discrete Event Simulation in Star Networks with Applications to Telecommunications. *International*

Workshop on Modeling, Analysis and Simulation of Computer and Telecommunication Systems, 1995, pp. 66-71.

[88] S. Ghosh, P. Ghosh, K. Basu, S. Das, and S. Daefler, iSimBioSys: A Discrete Event Simulation Platform for 'in silico' Study of Biological Systems *Proceedings of IEEE 39th Annual Simulation Symposium*, 2006.

[89] S. Ghosh, P. Ghosh, K. Basu, and S. Das, iSimBioSys: An 'In Silico' Discrete Event Simulation Framework for Modeling Biological Systems. *IEEE Comp. Systems BioInf. Conf.*, 2005, pp. 170-171.

[90] S. Ghosh, P. Ghosh, K. Basu and S. Das. Modeling the stochastic dynamics of gene expression in single cells: A Birth and Death Markov Chain Analysis. *IEEE International Conference on Bioinformatics and Biomedicine*, 2007.

[91] S. H. Northrup and H. P. Erickson. Kinetics of protein-protein association explained by Brownian dynamics computer simulations. *Proc. Natl. Acad. Sci.*, 1992, 89:3338-3342.

[92] S. J. Greive and P. H. von Hippel. Thinking quantitatively about transcriptional regulation. *Nature Reviews Molecular Cell Biology* 2005, 6:221-232.

[93] S. Karlin and H.M. Taylor. A First Course in Stochastic Processes. *second ed. San Diego, Calif.: Academic Press*, 1975, ISBN-10:0123985528.

[94] S. Karlin and H.M. Taylor. A Second Course in Stochastic Processes. *San Diego, Calif.: Academic Press*, 1981, ISBN-10:0123986508.

[95] V. K. Rangavajhala, and S. Daefler, Modeling the Salmonella PhoPQ two component regulatory system. *University of Texas at Arlington, Master's Thesis*, online at http://crewman.uta.edu/dynamic/bone/publications.htm

[96] V. Sobolev, A. Sorokine, J. Prilusky, E. E. Abola and M. Edelman. Automated analysis of interatomic contacts in proteins. *Bioinformatics*, 1999, 15:327332.

[97] Workshop: Making Sense of Complexity *Summary of the Workshop on Dynamical Modeling of Complex Biomedical Systems*, (2002).

[98] Y. Ji, X. Jie, R. Xiaojia, L. Kaiqin, and X. Sunney, Probing Gene Expression in Live Cells, One Protein Molecule at a Time. *Science*, 2006, 311:1600-1603.

[99] Y. Cao, D.T.Gillespie and L.R. Petzold. Avoiding Negative Populations in Explicit Poisson Tau-Leaping. *J. Chemical Physics*, 2005, 123:054104.

[100] Y. Takeda, A. Sarai and V. M. Rivera. Analysis of the sequencespecific interactions between Cro repressor and operator DNA by systematic base substitution experiments. *Proc. Natl. Acad. Sci. USA.*, 1989, 86:439-443.

BIOGRAPHICAL STATEMENT

Preetam Ghosh received his Bachelor of Engineering degree in Computer Science and Engineering from Jadavpur University, Calcutta, India, in 2000. After spending two years in the telecommunication industry in India, he began his graduate studies in the Department of Computer Science and Engineering at the University of Texas at Arlington in Fall 2002, where he received the Master of Science degree in August 2004. His primary research area is stochastic modeling, analysis and simulation of complex biological systems. His other research interests include optical networks, applications of scheduling and game theory, mobile grid computing, wireless mesh networks and multi-player online networked games.