1-1-2011

Optimal Control in Gene Mutation

Juanyi Yu
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

School of Engineering and Applied Science

Department of Electrical and Systems Engineering

Thesis Examination Committee:
Tyzh-Jong Tarn, Chair
Quo-Shin Chi
Norman I. Katz
Jr-Shin Li
Chenyang Lu
Heinz M. Schaettler

Optimal Control in Gene Mutation

by

Juanyi Yu

A dissertation presented to the Graduate School of Arts & Sciences
of Washington University in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2011
Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Optimal Control in Gene Mutation

by

Juanyi Yu

Doctor of Philosophy in Electrical Engineering

Washington University in St. Louis, 2011

Research Advisor: Professor Tzyh-Jong Tarn, Professor Jr-Shin Li

Gene mutations are the radical causes of many diseases, including inheritance diseases and cancers. Current medical treatments usually focus on changing the concentrations of related chemicals or mRNAs at the cellular level to stop protein productions or cell duplications, which can only control the diseases under certain circumstances but cannot cure them. Little research work has been done at the molecular level, the fundamental of inheritance, to search possible ways to cure those severe diseases.

In this dissertation, we propose a molecular level control system view of the gene mutations in DNA replication from the finite field concept. By treating DNA sequences as state variables, chemical mutagens and radiation as control inputs, one cell cycle as a step increment, and the measurements of the resulting DNA sequence as outputs, we derive system equations for both deterministic and stochastic discrete-time, finite-state systems of different scales. Defining the cost function as a summation of the costs of applying mutagens and the off-trajectory penalty, we solve the deterministic and stochastic optimal control problems by dynamic programming algorithm. In

addition, given that the system is completely controllable, we find that the global optimum of both base-to-base and codon-to-codon deterministic mutations can always be achieved within a finite number of steps.

Acknowledgments

I owe my deepest gratitude to my advisors, committee members, colleges, friends and parents. This dissertation would not have been possible without helps and supports from any of them during my study in Washington University in St. Louis.

To my advisors Tzyh-Jong Tarn and Jr-Shin Li, gracious mentors who demonstrated their invaluable guidance, ongoing inspiration and vital support throughout my study.

To my committee members for their encouraging words, precious time and attention.

To my professors for teaching me through challenging curriculum skills and methodologies to solve problems.

To my colleges for their comments on my work, Justin Ruths, Gongguo Tang, Anatoly Zlotnik, Isuru Dasanayake, Pei-Lan Liu, and Dionisis Stefanatos.

To the staff of the Department of Electrical and Systems Engineering for assisting me with the administrative tasks necessary for completing my doctoral program.

To my supportive, generous and dearest friends.

To my parents for their love, support and understanding during the long years of my education.

Juanyi Yu

Washington University in Saint Louis
December 2011

To my family.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Control and automation play critical roles in systems biology, an emerging academic field aiming at system-level understanding of biological systems. The control engineers not only provide new technology and equipment for biologists to design and perform meticulous experiments and take high-throughput measurements, but also work closely with doctors to develop new medical therapies and perform precise manipulations. The wide range of aspects which control and automation have been applied to include but are not limited to, gene regulation [Tanaka et al., 2006, Yildirim and Mackey, 2003], drug delivery [Langer, 1990, Yang et al., 2010], and neuron networks [Feng and Tuckwell, 2003, Moehlis et al., 2006].

Biological systems can be divided into three levels according to their scales: molecular, cellular and tissue level, respectively. Most current research work focuses on either cellular or tissue level systems. Not much work has been done at the molecular level. Understanding biological systems at the molecular level provides instrumental information about radical causes of many diseases and the genetic evidence of evolution. It also helps biologists to gain a better understanding of molecular level interactions, draw a complete blueprint of gene networks, improve existing means and create novel means to cure genetic diseases, and to elaborate on the theory of evolution.

On the other hand, current obstacles in systems biology are obvious. Albeit the progress in molecular biology has enabled us to collect comprehensive data sets on system performance and gain information on the underlying molecules [Kitano, 2002], the structure and dynamics of biological systems are usually unclear, which makes it

difficult to build abstract mathematical models for the given biological systems. Conventionally, biologists perform a series of experiments to identify interactions among related chemicals, construct mathematical models by modifying empirical or heuristic equations, and estimate parameters from experimental data. Random factors generated by experiments and parameter estimations may lead to the inconsistency between theoretical and experimental results.

In this dissertation, we use a new approach to construct an abstract mathematical model at the molecular level directly based on biological theory. With reasonable assumptions, we can avoid the problems caused by empirical or heuristic equations, measurements and parameter estimations. The cost function, a summation of the costs for applying mutagens and the off-trajectory penalty, together with the system equations, formulates the optimal control problem. The optimal control is then solved by the dynamic programming algorithm technique.

This chapter is organized as follows. In §1.1, we give an overview about systems biology. In §1.2, a brief introduction about the research work that has been done for gene regulatory networks is given. §1.3 describes the organization of this dissertation.

1.1 Systems Biology

Most living organisms use identical biopolymers, DNA, as the medium for long-term storage of genetic information, and eukaryotes follow the same rules to express genetic information. Although many organisms share these basic traits, they nevertheless take many different forms, and every individual of the same species is unique. Even though every organism has the same DNA in all cells, gene expressions differ in different functional tissues. Biologists have struggled for years to discover the structure of functional units in gene regulations, but have gained understanding only of specific regulatory units of simple life forms, such as the regulatory network of *lac* operon in Escherichia coli and TTG-bHLH-MYB in Arabidopsis, due to the complexity of gene regulatory networks. Elucidating the structure of functional gene regulatory units through biological experiments is crucial at early stages of research into gene expression and regulation processes, but the efficiency of such discovery can be enhanced

by considering common structures among functional units. For instance, the arabinose utilization network and tryptophan metabolic systems share a similar dynamic structure with the *lac* operon from the viewpoint of complex systems.

Biological systems have many intrinsic properties that are similar to man-made complex systems, including stochasticity, nonlinearity, stability, controllability and reachability. System theorists can construct generalized models to describe biological dynamics in a systematic fashion, while making these models adaptive to the specific characteristics of diverse functional units. Various tools in systems theory can be adaptively applied to model, analyze, control and reconstruct biological systems. With the help of efficient computational algorithms, systems theorists can simulate biological systems under different circumstances and at a low cost by tuning parameters and incorporating random factors. This leads to an understanding of the structures and properties of functional units without tedious repetition of the same experiments many times under the same or different conditions. Systems theorists can use properly designed mathematical models to make inferences about the behavior of biological systems that can lead to the development of novel practical therapies for diseases [Ledzewicz and Schaettler, 2009, Ledzewicz et al., 2010a,b].

In 1953, James D. Watson and Francis Crick discovered the double helix structure of DNA and the rule of base pairing. Five years later, Francis Crick first articulated the central dogma of molecular biology, which describes the transfer of genetic information between macromolecules, specifically DNA, RNA and proteins. Since then, many details of gene expression and regulation processes have been unveiled. System theorists have played an important role in developing various mathematical models for understanding complex biological systems. For instance, Leon Glass and Stuart A. Kauffman applied logical functions to represent biological regulatory units in 1973.

The early development of systems biology started in the late 1940s [Wiener, 1948]. Recent technology in molecular biology, including genome sequencing and high-throughput measurements, has made a system-level analysis of biological system possible. In general, a system-level understanding of a biological system can be derived from insight into four key properties: (1) the system's structure, (2) the system dynamics, (3) the control method, and (4) the design method [Kitano, 2002]. Equivalently, identifying related components and their interactions, gathering qualitative and quantitative

information about the system's evolution under different circumstances, achieving the desired outputs by controlling the input with appropriate definitions of inputs and outputs of the system, and reconstructing analogous systems by eliminating the undesired properties are four essential steps in systems biology done by collaboration among engineers, biologists, and doctors. The major difficulties, in general, lie in identifying the first two properties. Figure 1.1 illustrates the typical method of system construction and verification commonly applied. Control engineers construct models, run simulations, and predict the system behaviors. Biologists design and carry out the experiments and measure the output data. Control engineers revise and verify the models by comparing the predictions and experimental results. Currently, data-driven and hypothesis-driven methods are two main tools broadly applied. Due to the complexity of the systems, the mathematical models are usually formulated by modifying empirical equations or heuristic equations with only partial information available. The parameters of proposed models are obtained by various estimation methods. Although those models can disclose significant details of the system's structure and dynamics, the inconsistency between theoretical and experimental results creates obstacles for control engineers to verify the models, develop optimal controls, and reconstruct systems with desired properties.

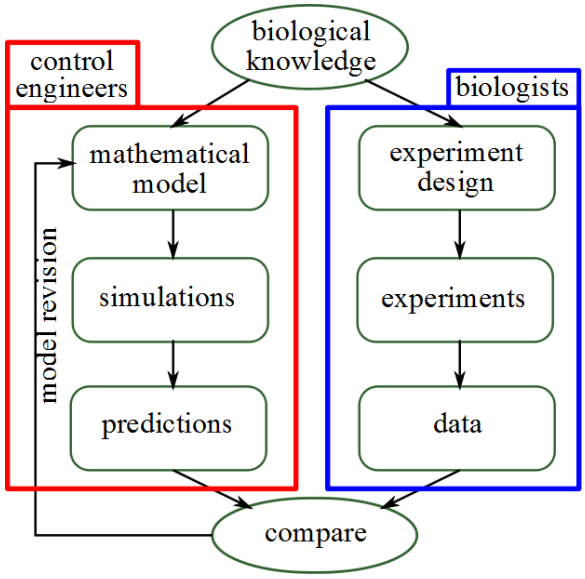

Figure 1.1: Typical analysis of biological systems.

Systems biology is a cross-cutting research area connecting control engineering, biology, and medical science. as shown in Figure 1.2. It aims at understanding the bare

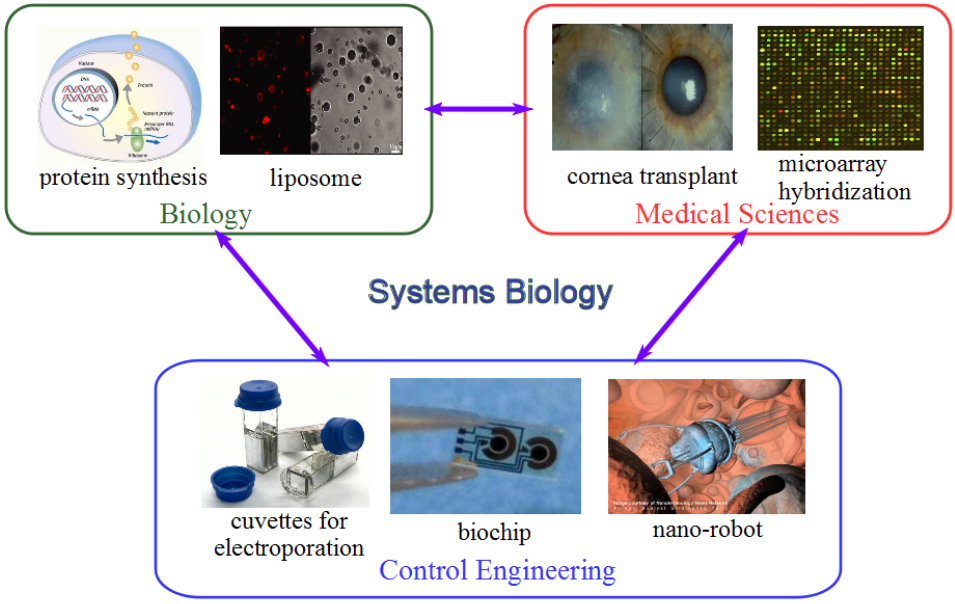

Figure 1.2: Systems biology is a cross-cutting research area connecting control engineering, biology, and medical science.
Sources: protein synthesis `http://www.anticancer.de`, liposome [Lentacker et al., 2009], corneal transplant `http://www.avclinic.com`, microarray hybridization [Reinke, 2006], cuvettes for electroportation `http://en.wikipedia.org`, biochip `http://www.clemson.edu`, nano robot `http://www.molecularlab.it`.

function and integration function of the cell to reconstruct the biological systems with desired features. Control and automation play critical roles in this novel field, not only by providing new technology and equipment for biologists to design and perform meticulous experiments, to take high-throughput measurements, and to analyze experimental data efficiently, but also by offering doctors new medical applications and improve the precision of medical manipulations. The equipment provided by control engineers includes but is not limited to, nano-devices, bio-chips, cuvettes for electroporation, and gene guns. Biologists perform various of biological experiments, such as protein synthesis and virus DNA modifications, to gather measurements for model revisions and verifications, to conclude theoretical and practical results from evidences, and to help medical practice. Doctors use both theoretical and practical results from biologists to perform tissue engineering, such as organ transplants and artificial tissue construction. In addition, engineers develop efficient computational

algorithms to analyze excessive experimental data provided by doctors and biologists. These three groups of scientists collaborate closely to promote the development of this new emerging field.

Biological systems can be divided into three levels according to their scales: molecular (nm), cellular (μm), and tissue level (cm). These levels are analogous in systems theory to part, individual and group, respectively. Consider, for examples, gene regulation systems at different levels. Molecular level research focuses on how, when, where and to what extent a gene is expressed [De Jong, 2002]. The completion of Human Genome Project makes it possible to sketch a complete picture of a gene by identifying the control sequences that govern how its DNA segments are coded and how they interact. The essential goal at the molecular level is to sketch a complete blueprint of gene regulatory network that theoretically describes clearly the function of every gene (approximately 20,000-25,000 in human DNA) and how genes interact, predicts possible results of a mutated gene, and provides researchers clues to introduce human interferences into natural processes to eliminate side effects caused by mutations. Besides gene expression and regulation, gene delivery, focusing on how to delivery and integrate target genes at the right chromosomes and spots by appropriate vehicles, is also a major topic at the molecular level. While changes in complex systems at the cellular or tissue level may be accounted for through intrinsic biological principles, changes in systems at the molecular level certainly affect the behavior of the whole system at the cellular and tissue levels. Therefore, understanding the molecular level systems is of great importance in improving the conventional medical interventions and creating innovative therapies to control or cure them, by identifying the radical causes in genes. Moreover, novel means to rescue endangered species may be found by genetically identifying beneficial and deleterious in the course of evolution.

Research on the cellular level, in general, treats one cell as a plant in classical system theory and investigates its response to a changing environment, especially changes in the concentration of RNA and proteins. State-of-the-art medical therapies primarily based on experimental results at the cellular level, by adjusting the external environment to promote the production of beneficial proteins, repress the expression of deleterious proteins, or to stop the reproduction of bacteria by high concentrations of antibiotics. Researchers also focus on how to delivery the drug to the desire cells by

various means, such as electric polarization, which control the penetration of cell wall or membrane by tuning the intensity of local electric field. Another major topic at the cellular level is to position the master cell of a group of cells, exam how it affects the gene expression and regulation of neighboring cells, and develop optimal control methods to intervene its function. Tissue level research mainly concerns deleterious tissue suppression, tissue reconstruction, artificial tissue substitutes, and tissue function recovery. The cell differentiation process is an important topic at this level, and radiation therapy for cancer is a typical example of a practical application.

Gene regulations are collaboratively controlled by all three level systems, and most complex biological systems have a hierarchical structure similar to gene regulation systems. For instance, the metabolism of lactose in Escherichia Coli is jointly controlled by the *lac* operon, composed by three structural genes, $lacZ$, $lacY$ and $lacA$, a promoter, a terminator, a regulator, and an operator, at the molecular level, and the concentrations of glucose and lactose both in the cell and in the local environment at the cellular level.

In general, complex biological systems at the molecular level have discrete state spaces, because molecules and nucleotide bases are discrete. The time index of molecular level systems processes can be continuous or discrete, depending on the system. Our model for gene mutation in DNA replication is discrete-time, with one cell cycle normalized to the step increment. Because distinct cells have different cell cycles, a continuous-time system can be a more accurate model. While the controls for biological systems at the molecular level are usually ON/OFF controls, and hence are discrete variables, their corresponding systems at the cellular level, in most cases, are continuous-time, continuous-state systems, where the state and control variables are typically concentrations of chemicals, whose derivatives are production and reduction rates. The state space of models for biological systems at the tissue level can be discrete (e.g. the number of cells in a tumor), or continuous (e.g. the size of a tumor), and the control variables are usually piecewise continuous (e.g. scheduled radiation therapy).

1.2 Existing Approaches

Researchers have applied various methods to model, simulate, and control the gene regulation processes. Early attempts to model and simulate gene regulatory systems are summarized in [De Jong, 2002], including direct graphs, Bayesian networks, Boolean networks, ordinary and partial different equations, qualitative differential equations, qualitative differential equations, stochastic equations, and role-base formalisms ([De Jong, 2002]). Other approaches include Petri nets ([Matsuno et al., 2000]), transformational grammars ([Collado-Vides, 1989, Collado-Vides et al., 1998]), and process algebra([Regev et al., 2001]). Some recent work of system-view approach to gene regulation includes [Tanaka and Kimura, 2008], [Layek et al., 2011], [Mayo et al., 2006] and [Zhang et al., 2006]. Three important modeling methods in recent work are compound control models, logic network models and base-to-base molecular level formulations.

Compound control models

Biological systems are always in response to the compound environmental changes [Tanaka and Kimura, 2008]. Various mathematical models have been derived for the gene regulatory units ([Ozbudak et al., 2004, Santillán and Mackey, 2004, Setty et al., 2003, Yildirim and Mackey, 2003]), including *lac* operon and *cis*-regulatory units. [Tanaka et al., 2006] proposed a generalized model that can be adaptive to several gene regulatory units with similar structures at the cellular level, including arabinose utilization network, tryptophan metabolic system, heat shock response system and λ-system.

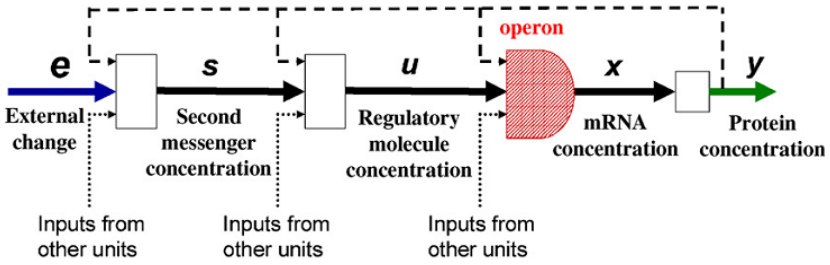

Figure 1.3: General scheme of a regulatory unit [Tanaka and Kimura, 2008]

.

The general scheme of a regulatory unit is illustrated in Figure 1.3, where the operon acts as the plant in the conventional control theory. And the system dynamics can be represented by equations

$$\dot{x} = F(u, y) - \alpha x, \tag{1.1a}$$

$$\dot{y} = G(x, y) - \beta x, \tag{1.1b}$$

$$\dot{u} = H(s, y, u) - \gamma x, \tag{1.1c}$$

$$\dot{s} = K(e, y, s) - \delta x, \tag{1.1d}$$

where x, y, u, s and e are column vectors corresponding to the concentrations of mRNA of the operons, of produced proteins, of regulatory molecules, of second messengers, and of external changes, respectively. α, β, γ and δ represent the degradation rates (together with the growth rates) for x, y, u and s. The four functions F, G, H and K describes the production rates of x, y, u and s, respectively [Tanaka and Kimura, 2008].

However, F, G, H, and K are usually obtained from empirical biomedical kinetic models or curve fitting algorithms based on the experimental results. And real-time measurements of concentrations of different chemicals within the cell and the environment is unavailable. Therefore, the accuracy of such models are not guaranteed.

Logic network models

The early research of using logical functions to represent biological regulatory is presented in [Glass and Kauffman, 1973, Thomas, 1973]. The basic idea of this approach comes from the similarity between gene regulatory networks and digital circuits. Later, researchers extend this idea to various logic networks, including baysian networks, boolean networks, generalized logical networks, Petri net and their probabilistic generalizations.

Figure 1.4 shows an example of directed hypergraph representation of gene regulatory networks. Directed hypergraph is defined as a tuple $< V, E >$ with V, a set of vertices representation genes or other elements, and E, a collection of edges representing the interaction among genes. $+$ in Figure 1.4 indicates the activation, and $-$ indicates inhibition. A directed edge is defined by a tuple $< i, [j_1, j_2, \cdots, j_n], [s_1, s_2, \cdots, s_n] >$, representing the edges from vertices

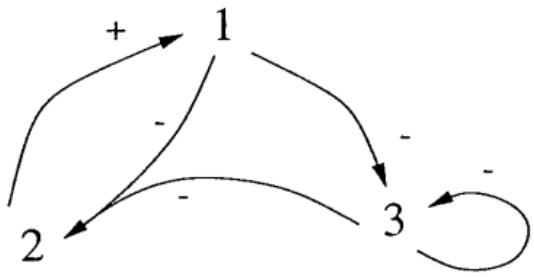

Figure 1.4: An example of directed hypergraph representation of a regulatory network with cooperative interactions [De Jong, 2002].

$[j_1, j_2, \cdots, j_n]$ to vertex i, with $s_m \in \{+, -\}$ corresponding to the pathway from vertex j_m to vertex i, respectively. The definition of V and E in Figure 1.4 can be expressed as

$$V = \{1, 2, 3\},$$
$$E = \{< 2, [1, 3], [-, -] >, < 3, [1], [-] >, < 1, [2], [+] >, < 3, [3], [-] >\}.$$

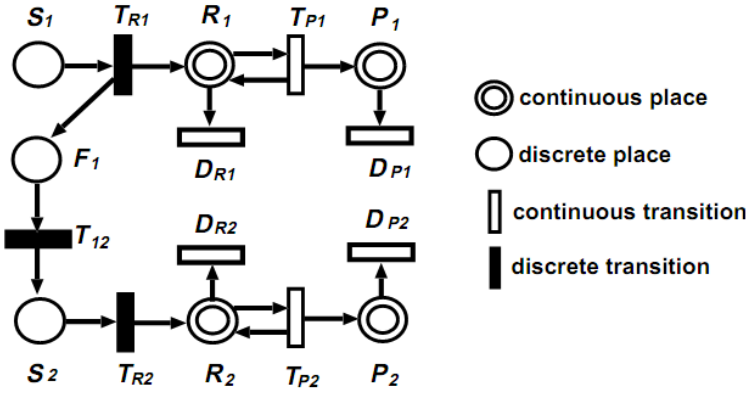

Figure 1.5: Hybrid petri net model for two-genes operon [Matsuno et al., 2000].

Figure 1.5 shows a typical hybrid Petri net model for two-genes operon. A hybrid Petri net is convertionally defined as $Q = (P, T, h, \mathbf{Pre}, \mathbf{Post}, M_0)$, with set of places $P = \{P_1, P_2, \cdots, P_n\}(n \geq 1)$, set of transitions $T = \{T_1, T_2, \cdots, T_m\}(m \geq 1)$, $h : P \cup T \rightarrow \{D, C\}$ indicating every place or transition whether it is discrete or continuous, $\mathbf{Pre}(P_i, T_j)(\mathbf{Post}(P_i, T_j))$ from a place P_i (a transition T_j) to a

transition T_j (a place P_i), a function with weighted arc, and the initial marking M_0 [Matsuno et al., 2000].

The drawback of logic network models is too many details are omitted during the modeling process. As a result, such models can only provide qualitative representations for gene regulatory units. Integration of logic network models and biomedical kinetic models provide more biological details, yet the disadvantages of both approaches still exist.

Base-to-base molecular level formulation

Instead of using biomedical kinetic models, researchers in the field of DNA computation proposed a novel mathematical formulation at molecular level. Initiated by the idea in [Adleman, 1994], [Zhang et al., 2006] proposed three ways, complex number, integer number, and vector representations, to convert the character-base DNA sequences to numerical sequences, as shown in (1.2a), (1.2b) and (1.2c), respectively.

$$f(x) = \begin{cases} 1 & x = A, \\ -1 & x = T, \\ 1 & x = G, \\ -1 & x = C; \end{cases} \tag{1.2a}$$

$$f(x) = \begin{cases} 0 & x = A, \\ 1 & x = T, \\ 2 & x = G, \\ 3 & x = C; \end{cases} \tag{1.2b}$$

$$f(x) = \begin{cases} \begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} & x = A, \\ \begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} & x = T, \\ \begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} & x = G, \\ \begin{bmatrix} 0 & 0 & 0 & 1 \end{bmatrix} & x = C. \end{cases} \tag{1.2c}$$

They also presented the corresponding inverse functions, complementary DNA sequences representations and several important propositions in [Zhang et al., 2006]. Models for DNA hybridizations, RNA self-hybridizations and a unified representation of gene expression can be found in [Gao et al., 2010].

1.3 Organization of This Dissertation

This dissertation is organized as below.

Chapter 2 gives an outline of this dissertation. Our motivations, problem statement, methodology, and potential impacts of our work are discussed in this chapter.

Chapter 3 begins with a brief introduction to biological details of gene regulations, including the central dogma in molecular biology, DNA replication, and gene mutations. Then we go through the details of constructing mathematical models of mutations in DNA replication for both point and multi-sites, stochastic and deterministic cases, followed by several important propositions. In addition, we extend our model to describe mutation involving broken DNA strands.

Chapter 4 shows how we define our optimal control objectives, and apply the dynamic programming algorithm to compute optimal control sequences. Simulation results of base-to-base and codon-to-codon, deterministic and stochastic optimal control problems are illustrated and compared. Codon-to-codon deterministic and stochastic cases are critical as a gene is composed of finite number of codons. Consequently, the solutions to codon-to-codon deterministic and stochastic cases lead directly to the solutions to gene-to-gene deterministic and stochastic mutations.

Chapter 5 summarizes the work has been done and discusses possible future work.

Chapter 2

Problem Statement

This chapter gives an outline of our work. Apart from the previous models introduced in §1.2, we propose a novel model for DNA replication, and solve the optimal control problem based on our model. The motivations are discussed in §2.1. The problem statement and methodology to approach the problem are discussed in §2.2. We discuss the potential impacts of our work in §2.3.

2.1 Motivations

As mentioned in §1.2, although system models at the cellular level, including logic networks, compound controls, and the integration of both, can describe the gene regulatory units, parameter estimations and real-time measurements are two major problems in these conventional modeling methods. The bottleneck of system models at the cellular level can be solved by a better understanding of the corresponding biological systems at the molecular level. The commonalities shared by most life forms at the molecular level makes it essential to construct a rigorous state-space model at the molecular level directly based on biological theories.

Most living organisms use DNA as the medium for long-term genetic information storage. Damage of DNA molecules can lead to deleterious consequences, such as lethal diseases or inheritance diseases. The high-fidelity of DNA molecules, especially the regulatory and coding DNA segments, ensures the productions of functional proteins for metabolism. DNA is usually encapsulated in the chromosomes inside cell nucleus. DNA sequences are less stable during DNA replications and transcriptions

when the double helix structure of DNA molecules is destroyed by breaking the hydrogen bonds between two DNA strands. Consequently, externals disturbances and internal noises can interrupt the DNA sequences during those period with a higher probability than when encapsulated. On the other hand, applying corresponding mutagens at the correct time instance during those periods can restore damaged DNA sequences. DNA replication is inter-cellular, i.e. the genetic information is replicated and passed to the new cell. However, DNA transcription is usually inner-cellular, i.e. the genetic information is transcribed and transported into the cytoplasm of the same cell. Although the proteins produced by one cell can affect protein productions in other cells, the accuracy of DNA replication is particularly important to ensure the correct expression of genes.

Currently, there are two means to restore an abnormal DNA segment back to a normal sequences, mutagens and viral infection. Mutagens, usually chemicals or radiative rays, may cause deleterious consequences to living organisms. However, comparing to viral infection, which may lead to uncontrollable aftermaths, such as cancers or immune system diseases, applying mutagens in the right order and at the correct time instance is a relative safe way to control and cure genetic diseases. Because of the poisonousness of mutagens, we would like to find a way in which we can restore a disrupted DNA segment to a normal sequence at the lowest risk.

In summary, our goal is to derive a robust system model for gene mutation in DNA replication at molecular level, and then to find control sequences to drive an abnormal DNA to a normal state at the minimum cost.

2.2 Problem Statement & Methodology

Figure 2.1 shows the system diagram of restoring an abnormal DNA segment to a normal sequence by applying mutagens during the process of DNA replication. Once obtaining patients' genome, we compare the coding DNA segments with normal DNA segments in our database to figure out the range of disrupted segments. Due to the redundancy in genetic codes, as long as any two DNA segments containing the same number of nucleotide bases can be transcribed and translated to the same amino acid sequence, the distance reference between them is considered to be zero. Therefore,

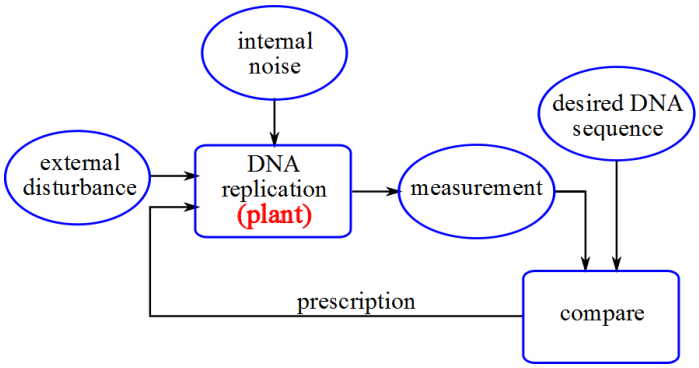

Figure 2.1: System diagram of restoring an abnormal DNA segment back to a normal sequence by applying mutagens during the process of DNA replication.

our final state lies in a set where the distance reference between any sequence and the desired sequence is zero, different from classical systems theory, the final state of which is usually a point or a neighborhood in the state space. We name this set the final desired set. The prescription is then determined by picking the path with the lowest cost from the current measurement of current DNA segment to every sequence in the final desired set.

From the viewpoint of systems theory, a dynamic system is a quintuple, input space, state space, output space, state transition map, read-out map. Here, we give an overview of those five components in our systems.

Although the duration of cell cycles varies among tissues, DNA replication always happens at the beginning of every cell cycle, before cell division. At the molecular level, our target is usually DNA molecules inside one cell or a group of cells in the same tissue, therefore, we normalize the duration of a cell cycle to 1, representing a step increment.

Biologically, gene mutations, both spontaneous and induced, may change the number of nucleotide bases of targeted segment. On the other hand, from the viewpoint of systems theory, it is important to keep the consistency of state space during a system's evolution. Since DNA segments of different lengths belong to different state spaces, we require the length of DNA segment remains the same through the whole evolution. In other word, mutations involving base insertions are ignored in our model, but those involving base deletions can be expressed by filling deleted sites with artificial non-sense bases. Enlightened by Zhang's work [Zhang et al., 2006], we map four nucleotide

bases, together with an artificially added non-sense bases into a finite field composed of five real numbers. Subsequently, nucleotide base sequences of DNA segments are converted to vectors, which are elements of state space. Since DNA segments are of finite length, and there are only five alleles (four nucleotide bases and one artificial nonsense bases) at every spot, the total number of all permutations of a given-length nucleotide sequence is finite. Hence, our state space contains finite number of vectors. The field structure of our state space is very helpful in our modeling process as it possesses nice mathematic properties, such as the multiplication and addition among elements from the field always stay within the same field.

The output, the measurement of current DNA sequence, lies in the same space as the state variables. In reality, a mutagen, especially a chemical mutagen, correspond to one or two major transfer patterns and other minor transfer patterns, which occur at very low probabilities. Therefore, we can view our control space as a set of ON/OFF controls of all available mutagens at every spot of targeted DNA segment, representing whether a mutagen is applied at a particular position. The state transition map describes how the system is driven from input to the output, which are fully discussed in Chapter 3. In this particular problem, the output is exactly the next state, since we assume our measurement is always accurate.

Therefore, our system is a discrete-time dynamic system with finite state space and output space, and a set of ON/OFF switches as controls. Our goal is to optimally drive this system from a given initial state to a desired final set at the lowest cost.

The objective function is defined as an accumulated sum of the cost of applying mutagens, including poisonousness scales, and off-trajectory penalty, a distance reference between current measurement and the desired set at current step, during system's evolution. The chemical and physical properties of animo acid makes it difficult to define a proper metric over the finite field for codons. Alternatively, we define a distance reference, which acts as a penalty if the final state is not in the final desired set. The optimal control sequences are computed in advance to let doctors make medical treatment plans according to the patient's initial condition. In general, the optimal control sequence and the corresponding optimal trajectory are not unique because the bases mutate independently in most cases and the order of mutating different bases

is does not matter if the number of medical treatment sessions is not under a tight restriction.

Additional measurements are taken before and after each treatment if necessary. In deterministic cases, the purpose of taking additional measurements is to verify the result of medical interventions and to eliminate internal and external disturbances. The treatment plan is adjusted if current measurement does not follow the prediction. In stochastic settings, we take the measurement to conquer the randomness cased by both mutagens and other noises. The treatment is then updated accordingly. In the real cases, both the distance reference and the costs of applying mutagens should be defined by doctors or biologists according to the statistics. We apply the dynamic programming algorithm to compute the optimal control sequences.

2.3 Impacts

Our work derive a mathematical framework for gene regulation at molecular level and provide a novel angle to view biological systems in a systematic fashion. Our model can facilitate biologists and doctors in identifying the structure of functional units efficient, analyzing biological systems at the molecular level, and gaining a better understanding of gene expression and regulation at the cellular and tissue levels. Our model and optimal control algorithm are crucial in the improvement and creation of novel medical therapies.

In addition, our model can contribute in sketching a complete map of gene network. In laboratories, biologists can mutate particular sections of DNA on purpose to identify the regulatory genes, coding sequences and non-coding sequences, and discover the interactions among them, follow the optimal control sequences computed in advance. The corresponding cellular and tissue level systems can monitored simultaneously to understand the interaction among different levels.

Moreover, the beneficial and neutral mutations lead to the diversity of species. Geneticists examine compare the differences between DNA from ancient and living animals to trace the evidence of evolution. By eliminating the deleterious mutations in the

natural selection, our model provide a relative safe way to perform those mutations in the lab or hospital to increase the diversity of the gene pool.

Last not not the least, our model also helps the construction of a molecular computer. Though DNA replication and hybridization is efficient, but it is not error-free. Our work provides a possible way to correct computation errors.

Chapter 3

System Equation Formulation

In this chapter, we construct a novel mathematical formulation for mutations occurring during DNA replication at the molecular level. Instead of using chemical concentrations as state variables, we directly use the nucleotide-base sequences of DNA segments as our state variables. Based on the transfer matrix we derived for perfect DNA replication, we discover that with proper assignment, nucleotide bases, together with an artificially introduced non-sense base, can be convert into a finite field composed of five real numbers, under proper definition of addition and multiplication. The system equations for point mutation and large-scale, deterministic and stochastic cases are then developed. In addition, we show that our system equations can be adapted to other biological processes at the molecular level, such as broken DNA strands.

The organization of this chapter is as follows. In §3.1, we give a brief introduction to biological background. Assumptions and their feasibilities are discussed in §3.2. Then we present a basic model for the perfect DNA replication, from which we find a constant transfer matrix and the underlying field of single nucleotide bases when converting them into real numbers. We go through the details of modifying our basic model to express deterministic point mutation by reverse engineering the mutations occurring in DNA replication in §3.3. §3.4 and §3.5 are the extensions of §3.3 to large scale deterministic and stochastic cases.

3.1 Background

3.1.1 The Central Dogma of Molecular Biology

The central dogma of molecular biology, first elaborated in [Crick, 1958] and re-stated in [Crick, 1970], illustrates the detailed residue-by-residue transfer of genetic sequential information. Nowadays, it is widely recognized as the backbone of molecular biology. It describes the genetic information flow between three kinds of biopolymers: DNA, RNA, and protein. In most living organisms, genetic information transfers from DNA to RNA, and then into protein. Though under special conditions, some transfers are reversible, protein always acts as the sink of information flow, as shown in Figure 3.1. In Table 3.1, we list out all nine possible transfers between three kinds of macromolecules, which are classified into three classes: general transfer, special transfer, and unknown transfer. General transfers are normal biological processes. Special transfers exists only in virus or in laboratory. No evidence shows that unknown transfers occur in natural processes or in laboratory up till now.

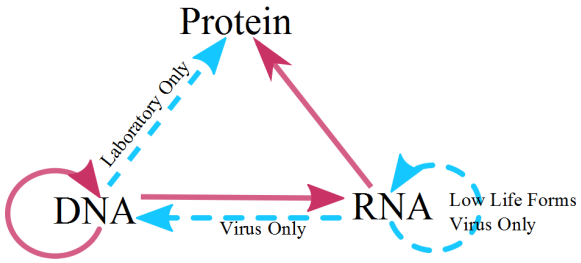

Figure 3.1: Biological information flow in central dogma of molecular biology.

General Transfer	Special Transfer	Unknown Transfer
DNA → DNA	RNA → DNA	Protein → DNA
DNA → RNA	RNA → RNA	Protein → RNA
RNA → Protein	DNA → Protein	Protein → Protein

Table 3.1: Three classes of transfers between three kinds of macromolecules classified by the Central Dogma in Molecular Biology.

Three general transfers in Table 3.1 are named DNA replication (DNA $\rightarrow$ DNA), transcription (DNA $\rightarrow$ RNA), translation (RNA $\rightarrow$ protein), respectively, by biologists.

DNA replication usually happens when a cell prepares to divide. Though there exist several DNA repair mechanisms which can eliminate replication errors [Friedberg et al., 1995], errors that are overlooked may lead to severe genetic diseases including cancers. Our work focus mainly on compensating such errors in genes. More details about DNA replication will be given in §3.1.2.

Transcription and translation are two key steps leading to the expression of genes. Transcription is the process of transfering genetic information from DNA segments into RNA sequences following complementary language, i.e. A (adenine) $\rightarrow U$ (uracil), T (thymine) $\rightarrow A$ and G (guanine) $\leftrightarrow C$ (cytosine). The resulting RNA is called messenger RNA (mRNA) as it contains genetic sequential information from transcribed DNA sequence. Different from the double-helix structure of DNA sequences, RNA is always single-stranded. In transcription, only one strand of DNA serves as the template, and it is read in the direction of $3' \rightarrow 5'$. The other strand is called the coding strand as its sequence is the same as the resulting mRNA sequences except T is replaced by U, if no error occurs. Only DNA segments which direct and regulate protein synthesis and the coding sequences that is translated into protein are transcribed. Noncoding DNA sequence, a large portion of total genome size, is copied to the new DNA strands during DNA replication, but is not involved in the transcription and translation process.

Translation is the process of converting a mRNA sequence into an amino acid polypeptide chain. Translation starts at the start codon of mRNA (usually AUG, sometimes GUG or UUG), attached by a ribosome, under appropriate initiation factors. One codon contains three consecutive nucleotides, and one combination corresponding to a specific amino acid. As there are 4^3 combinations forming 64 different codons but only 20 amino acids plus the stop codon, so there exists degeneracy, i.e. several codons correspond to the same animo acid. But one codon always corresponds to only one amino acid. Table 3.2 lists the genetic codes, the language of translation process, in terms of DNA codons. Transfer RNA (tRNA) brings the corresponding amino acid to each codon as the ribosome moves down the mRNA strand. Translation stops at a

stop codon. The synthesis of the peptide chain ends and the whole chain is released from the ribosome, which folds into the correct conformation. This folding process continues until a natal polypeptide becomes a mature protein.

1^{st} letter \ 2^{nd} letter	T	C	A	G
T	T Phe	T Ser	T Tyr	T Cys
	C Phe	C Ser	C Tyr	C Cys
	A Leu	A Ser	A STOP	A STOP
	G Leu	G Ser	G STOP	G Trp
C	T Leu	T Pro	T His	T Arg
	C Leu	C Pro	C His	C Arg
	A Leu	A Pro	A Gln	A Arg
	G Leu	G Pro	G Gln	G Arg
A	T Ile	T Thr	T Asn	T Ser
	C Ile	C Thr	C Asn	C Ser
	A Ile	A Thr	A Lys	A Arg
	G Met/Start	G Thr	G Lys	G Arg
G	T Val	T Ala	T Asp	T Gly
	C Val	C Ala	C Asp	C Gly
	A Val	A Ala	A Glu	A Gly
	G Val	G Ala	G Glu	G Gly

Table 3.2: Genetic codes (DNA $5' \rightarrow 3'$). A codon consists three consecutive nucleotide bases. The column shows the first base, the row the second, and the letter in the grid the third.

3.1.2 DNA Replication

DNA molecules, encapsulated in chromosomes within cell nucleolus, serve as the medium for long-term genetic information storage, and is the basis of genetic inheritance. It consists four kinds of nucleotide acids, *adenine* (A), *thymine* (T), *guanine* (G) and *cytosine* (C), and backbone made of sugars and phosphate. In 1953, James D. Watson and Francis Crick found the double helix structure of DNA and the rule of

base-pairing, known as Watson-Crick base-pairing [Watson and Crick, 1953, 2003]. A always pairs with T, G always pairs with C, and vice versa.

DNA replication occurs in the interphase of cell cycle. It begins at special locations in genome, called "origins". Double-stranded DNA is unwound at the origin by helicases, forming a replication fork with two prongs. Both strands serve as template and two new double-stranded DNA molecules are formed by adding nucleotides matched to the template strand and a number of associated chemicals, as shown in Figure 3.2. The new DNA molecules are half old half new, with one strand directly from the unwound DNA, and the complementary from linking free nucleotide bases inside nucleolus. In DNA replication, both strands are read in the direction of $3' \to 5'$.

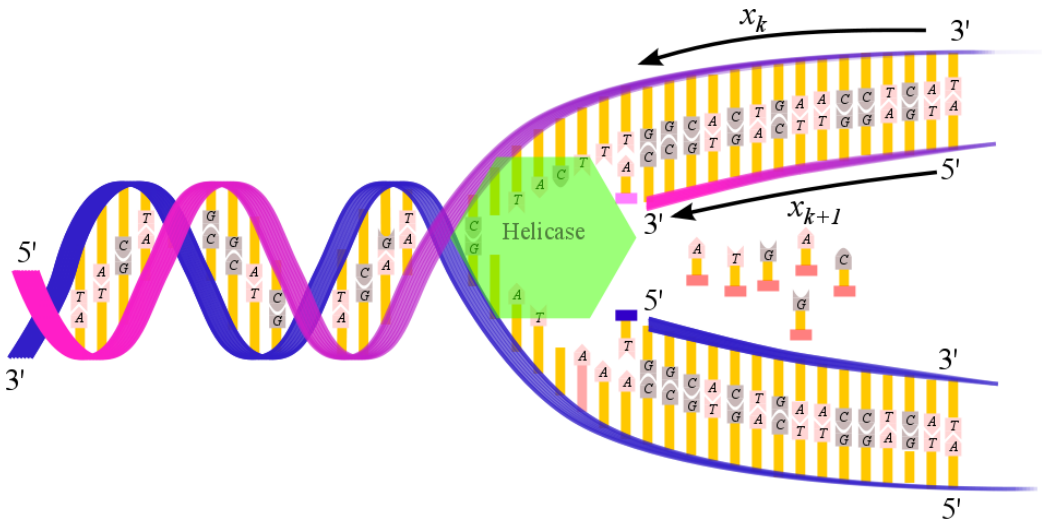

Figure 3.2: Biological information flow in central dogma of molecular biology.

The high fidelity of DNA replication is extremely important as the accuracy of gene expression and regulation are based on precision of nucleotide sequences. In nature, replication errors occur at a very low rate, one error for every 10^7 nucleotides added [McCulloch and Kunkel, 2008]. The redundancy of information caused by the double-helix structure ensures the fidelity of DNA replication. Some DNA self-repair mechanisms, listed in [Friedberg et al., 1995], such as proofreading, also help to eliminate errors during the replication process.

3.1.3 Gene Mutation

Gene mutations are changes in the nucleotide sequence of DNA or RNA. Usually, we focus only on mutations occurring in coding DNA sequences and RNA. Mutations are caused by various reasons. In natural, spontaneous mutation occurs at a relative constant rate. Mutation rate is different from one species to another. Induced gene mutations are brought by mutagenic agents, including chemicals, radiation and viral infection. Three main types of chemical mutagens that can alter base-pair sequences are base analogs, base modifiers and intercalcalating agents. Base analogs have structures similar to DNA bases, thus can substitute normal bases during the replication process. But unlike normal nucleotide bases, they can bind with bases other than the complementary to the one they replaced. Base modifiers can change existing bases and cause them to pair with bases other than the complementary. Intercalcalating agents can interrupt replication and transcription by inserting themselves directly into the DNA helix. Radiation includes ionizing radiation, including α, β, γ and x-rays, which can disrupt normal DNA sequences, usually by knocking out base pairs, and non-ionizing radiation, including ultraviolet light, which can block DNA replication by bonding adjacent T's on a DNA strand and may cause point mutation. Intensive radiation can destroy the cell by damage the backbone of DNA. It is widely applied in cancer therapies, usually combined with chemotherapy. Viral infection can reprogram the genes which regulate cell cycle, and lead to uncontrolled cell division, which is a major characteristic of cancer cells.

Mutations can be classified differently under specific criteria. According to the number of affected bases, we can divide them into small-scale mutations and large-scale mutations. The simplest form of small-scale mutations is point mutation, which substitutes one nucleotide base by another and happens only at one site of a targeted DNA segment. Point mutations can be further divided into transitions ($A \leftrightarrow G$ or $C \leftrightarrow T$) and transversions ($A/G \leftrightarrow C/T$). Figure 3.3 shows an example of transversion, with two bases in the green shaded area not complementary to each other. Transversions are theoretically expected to be twice as frequent as transitions, but transitions may be favored over transversions in coding DNA because they usually result in a more conserved polypeptide sequence [Strachan and Read, 2004]. Based on inheritance, we can classify them into somatic and gametic mutations. The former

occur in body cells and the latter occur in sex cells. In general, gametic mutations can be passed on to offsprings except for organisms reproduce asexually.

In this dissertation, we mainly focus on their impact on the resulting amino acid and protein sequence. Under this criterion, we divide mutations into five classes [Robinson, 2005].

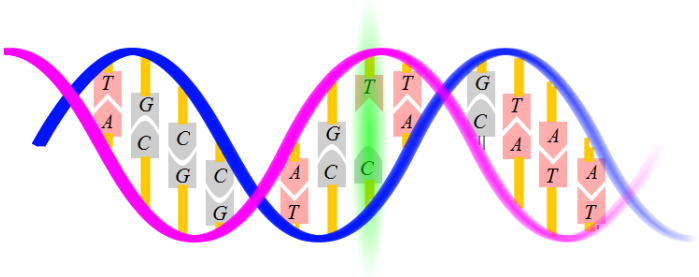

Figure 3.3: An example of point mutation. The area shaded by green is where mutation occurs.

Silent mutation is a mutation that only changes nucleotide sequence but not the resulting amino acid sequence. It may occur in noncoding DNA segments or within a codon but resulting in the same amino acid due to the redundancy of genetic codes in Table 3.2.

Neutral mutation is a mutation that occurs in a codon which results in a different amino acid with similar chemical properties. In addition, the resulting protein can function normally. Neutral mutation may be candidates of natural selection.

Missense mutation replaces the original amino acid with a different one, and as a result, alters the function of the corresponding protein.

Nonsense mutation is a mutation that occurs within a codon causing an early stop of transcription or translation. The protein may be malfunctional or totally nonfunctional depending on the location of nonsense mutation.

Frameshift mutation is a shift in reading frame caused by the addition or deletion of one or more nucleotides. It can totally mess up the downstream sequence from the mutated site. Insertion, deletion or duplication of a number of nucleotides are included in this category.

DNA self-repair mechanismcannot eliminate all mutations in DNA replication since all organisms interact with their environment, which involves various random factors, in a unique and unpredictable way. The results of mutations can beneficial, deleterious or neutral. Deleterious mutations can lead to inheritance diseases or cancers.

Our research focuses on the mutations in coding DNA segments, and we develop a compensation at the lowest cost for gene mutations.

3.2 Assumptions

In our work, we mainly focus on applying chemical mutagens and radiation to restore the original amino acid sequence. Other factors that may affect the gene mutation, including temperature and electroporation, are not within our consideration.

As previously mentioned, mutations involving inserting additional bases into target DNA segment is ignore to maintain the consistency of state space during system's evolution. Mutations involving base deletions can be expressed by using artificial non-sense bases to replace deleted bases.

The following assumptions hold for Chapter 3 and Chapter 4, unless stated otherwise.

Assumption 3.1. *Chemical mutagens or radiation can target one and only one nucleotide base at any predetermined site, despite the technical limitation.*

Due to the limitation of current technology, this assumption is not practically true. Current experimental work can only mix chemical mutagens into DNA solutions, but it is difficult to predict qualitatively and quantitatively how many targeted bases and at which sites of targeted bases will be mutated at the molecular level, which leads to an uncontrollable mutation process. Although gene therapy is a possible practical solution to large-scale mutations, but the high risk caused by virus, common vector to transport desired genes into targeted cells by integrating its programmed DNA segments into the DNA of host cells, is unavoidable. Chemical mutagens or radiation, applied in correct order with proper doses, can reduce the risk of restoring process.

Assumption 3.2. *Every base mutates independently.*

[Koch, 1971] shows some evidence that the mutation rates of neighboring bases may change if one base is mutated. We neglect affiliated effects of neighboring bases in mutations since these cases occur in specific situations. Hence, we assume every base mutates independently and there is no chain effect caused by mutagens.

Assumption 3.3. *Only one base in the targeted DNA segment can be mutated by chemical mutagens or radiation at each step, i.e. at most one chemical mutagen and one radiative ray are orderly applied at each step. The order of applying chemical mutagens or radiation can be random, but they cannot be applied at the same time. And targeted bases response to chemical mutagens and radiation independently.*

Indeed, our state space model can describe multi-site mutations within one step. But in order to avoid chemical reactions between chemical mutagens and the ionization effects of radiation on chemical mutagens, we take this assumption for simplicity and unambiguity. Later in this chapter, we define an order and corresponding time instances to apply chemical mutagens and radiation in a typical cell cycle to construct a generalized model.

Assumption 3.4. *Measurements are always correct.*

The technology used in Human Genome Project makes it possible to determine a DNA sequence in a simple, efficient and reliable manner. Though sometimes this method is unable to distinguish a base analog from a normal base, repetitive measurements can always compensate those situations. Therefore, we assume the measurement is 100% correct.

Assumption 3.5. *DNA replication error, background mutation rate, and other random noise can be eliminated from measurements by considering them as spontaneous mutation.*

We always take current measurement as our new state variable to calculate future control sequences. Replication errors are negligible as mentioned in §3.1.2. Spontaneous mutations can be incorporated into our models by introducing a term similar to chemical mutagens. Without loss of generality, we compensate all noises from measurements.

Assumption 3.6. *Radiation cause random mutations at a much higher rate than chemical mutagens.*

In practice, both radiation and chemical mutagens cause randomness. Radiation, in general, is more difficult and with a relative higher risk, to control the mutation comparing to chemical mutagens. Under deterministic conditions, we assume the induced mutations have no randomness. Under stochastic conditions, randomness is carefully considered and interpreted.

3.3 Base-to-base, Deterministic Model

Denote a targeted DNA segment with n nucleotide bases at k^{th} step by a column vector x_k, as shown in Figure 3.2. x_{k+1} is the state variable at $(k+1)^{th}$ stage. x_k^i denotes the i^{th} element of x_k. Let P be the transfer matrix from x_k to x_{k+1}, $\forall k$, $k \in \mathbb{Z}^+ \cup \{0\}$, without mutation. Then the perfect DNA replication process can be expressed as

$$x_{k+1} = Px_k. \tag{3.1}$$

Claim 3.1. $P = -I$.

Proof. As no mutation occurs, x_{k+1} is completely complementary to x_k by Watson-Crick base pairing rule, and x_{k+2} is completely complementary to x_{k+1}. Therefore, x_{k+2} is exactly the same as x_k.

$$x_{k+2} = Px_{k+1} = P^2x_k \Rightarrow P^2 = I.$$

Since x_k and x_{k+1} are of the same dimension, P is a square matrix. According to §3.1.2, every element of x_{k+1}^i only depends on the corresponding element of x_k^i, thus P is diagonal. In addition, $x_{k+1} \neq x_k$, we conclude $P = -I$. □

We rewrite (3.1) as

$$x_{k+1} = -Ix_k. \tag{3.2}$$

Based on (3.2), we assign values to nucleotide bases set $\{A, G, C, T, O\}$, where O is an artificial non-sense base. Define an equivalence relationship between $\{A, G, C, T, O\}$

and $\{1, 2, -2, -1, 0\}$, i.e. $\{A, G, C, T, O\} \Leftrightarrow \{1, 2, -2, -1, 0\}$, with

$$x_k^{(i)} = \begin{cases} 1 & \text{if } A, \\ 2 & \text{if } G, \\ -2 & \text{if } C, \\ -1 & \text{if } T, \\ 0 & \text{if } O. \end{cases} \tag{3.3}$$

Claim 3.2. $\{1, 2, -2, -1, 0\}$ *is a field under proper definitions of addition and multiplication.*

Proof. $\mathbb{F}_p = \mathbb{Z}/p\mathbb{Z}$ is a field with p a prime number. Let $p = 5$, $\mathbb{F}_5 = \mathbb{Z}/5\mathbb{Z} = \{0, 1, 2, 3, 4\}$ is a field and the multiplication and addition defined on this field by $mod5$.

Initiated by $\mathbb{F}_5$, for $\{1, 2, -2, -1, 0\}$, we define the addition table and multiplication table as follows:

$+$	1	2	-2	-1	0
1	2	-2	-1	0	1
2	-2	-1	0	1	2
-2	-1	0	1	2	-2
-1	0	1	2	-2	-1
0	1	2	-2	-1	0

Table 3.3: Addition table for $\{1, 2, -2, -1, 0\}$.

$\times$	1	2	-2	-1	0
1	1	2	-2	-1	0
2	2	-1	1	-2	0
-2	-2	1	-1	2	0
-1	-1	-2	2	1	0
0	0	0	0	0	0

Table 3.4: Multiplication table for $\{1, 2, -2, -1, 0\}$.

Now check if the set $\{1, 2, -2, -1, 0\}$ satisfies the requirements of a field with our addition and multiplication table.

Closed under addition and multiplication

Satisfied obviously from Table 3.3 and Table 3.4.

Associativity of addition and multiplication

Implicitly satisfied by integer addition and multiplication.

Commutativity of addition and multiplication

Satisfied as Table 3.3 and Table 3.4 are symmetric according to the diagonal.

Additive and multiplicative identity

Additive identity is 0 and multiplicative identity is 1.

Additive and multiplicative inverses

Additive inverses pair: $1 \leftrightarrow -1, 2 \leftrightarrow -2, 0 \leftrightarrow 0$.

Multiplicative inverses pair: $1 \leftrightarrow 1, 2 \leftrightarrow -2, -1 \leftrightarrow -1$.

Distributivity of multiplication over addition

Implicitly satisfied by integer addition and multiplication.

We conclude $\{0, 1, 2, -2, -1\}$ is a field under addition and multiplication defined by Table 3.3 and 3.4. □

From now on, we use $\mathscr{F}$ to denote the field $\{0, 1, 2, -2, -1\}$. And $x_k \in \mathscr{F}^n$ is the state variable representing a DNA segment with n nucleotide bases at k^{th} stage.

As stated in §3.1.3, the simplest mutation is point mutation, which involves only one nucleotide base. Assume a measurement is taken before every duplication process.

If there is a point mutation as shown in Figure 3.3, we modify (3.2) to

$$x_{k+1} = (-I + \Delta s)\, x_k + \Delta w, \tag{3.4}$$

where $x_k, x_{k+1} \in \mathscr{F}$, and $-I$ reduces to -1 as only one base is involved. The corresponding values of Δs and Δw, obtained by reverse engineering with all possible pair of x_k and x_{k+1}, are listed in Table 3.5.

k^{th} \ $(k+1)^{th}$	A	G	C	T	O	
A	2	-2	-1	0	1	
G	-1	2	0	-2	1	$\}\Delta s$
C	-2	0	2	-1	1	
T	0	-1	-2	2	1	
O	1	2	-2	-1	0	$\}\Delta w$

Table 3.5: Possible values of Δs and Δw.

Here, Δs represents mutations from four normal nucleotide bases, and Δw corresponds to mutations from artificial non-sense base, i.e. $\Delta w \neq 0$ only if $x_k = 0$.

Rewriting (3.4) by collecting all values of Δs and Δw in Table 3.5, we get

$$x_{k+1} = \left(-I + \sum_{j=0}^{4} u_k^j s_j\right) x_k + \sum_{j=0}^{4} c_k^j w_j, \tag{3.5a}$$

$$= (-I + u_k s)\, x_k + c_k w, \tag{3.5b}$$

where $\{s_0, s_1, s_2, s_3, s_4\} = \{0, 1, 2, -2, -1\}$, $\{w_0, w_1, w_2, w_3, w_4\} = \{0, 1, 2, -2, -1\}$, $u_k^j, c_k^j \in \{0, 1\}$, representing the on/off controls, $u_k = \begin{bmatrix} u_k^0 & u_k^1 & u_k^2 & u_k^3 & u_k^4 \end{bmatrix}$, $c_k = \begin{bmatrix} c_k^0 & c_k^1 & c_k^2 & c_k^3 & c_k^4 \end{bmatrix}$, and $s = w = \begin{bmatrix} 0 & 1 & 2 & -2 & -1 \end{bmatrix}^T$.

In (3.5a), s_j and w_j are constants for all k. u_k^js and c_k^js, the inputs of our system, are on/off controls for chemical mutagens or radiation. And $\sum_{j=0}^{4} c_k^j = 1$ only if $x_k = 0$. (3.5b) is a simplified version of (3.5a) as we put u_k, s, c_k, w into vector representations. s and w serve as vector basis for base-to-base deterministic model. u_k and c_k are now multi-input controls, each of them contains 5 on/off controls. For a particular k, at most one of u_k^js and c_k^js can be 1, as stated in Proposition 3.1. This is consistent with the fact that every state can be transferred to only one of the five states in the state space $\mathscr{F}$ with corresponding mutagens available.

Proposition 3.1.
It is always $1-1$ *transfer when mutation occurs, i.e. one nucleotide base can only transfer to another one, therefore*

- *If* $x_k = 0$ *and* $c_k^j = 0, \forall j, 0 \le j \le 4$, *or* $c_k^0 = 1, c_k^j = 0, \forall j, 1 \le j \le 4 \Leftrightarrow x_k = 0$ *and* $c_k = 0$, *or* $c_k = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$, *then* $x_{k+1} = 0$.

- *If* $x_k \neq 0$, $c_k^j = 0, \forall j, 0 \le j \le 4$ *and* $\sum_{j=0}^{4} u_k^j = 0$ *or* $1 \Leftrightarrow c_k = 0$ *and* u_k *is either* 0 *or a unit row vector.*

- *If* $x_k = 0$, $u_k^j = 0, \forall j, 0 \le j \le 4$ *and* $\sum_{j=0}^{4} c_k^j = 0$ *or* $1 \Leftrightarrow u_k = 0$ *and* c_k *is either* 0 *or a unit row vector.*

- $\sum_{j=0}^{4} u_k^j + c_k^j = 0$ *or* $1, \forall k \in \mathbb{Z}^+ \cup \{0\} \Leftrightarrow u_k + c_k$ *is either* 0 *or a unit row vector,* $\forall k \in \mathbb{Z}^+ \cup \{0\}$.

Now suppose for some reason, we need to take an addition measurement in the middle of every cell cycle, after the completion of the k^{th} duplication and before the start of the $(k+1)^{th}$. We name this kind of measurement an intermediate state, and denote it by x'_k. Then we have

$$x_{k+1} = \left(I + \Delta s'\right) x'_k + \Delta w', \tag{3.6}$$

where the values of $\Delta s'$ and $\Delta w'$, listed in Table 3.6, are obtained in the same way as Δs and Δw in Table 3.5.

k^{th} \ $(k+1)^{th}$	A	G	C	T	O	
A	0	1	2	−2	−1	
G	2	0	−2	1	−1	$\Delta s'$
C	1	−2	0	2	−1	$\Delta s'$
T	−2	2	1	0	−1	
O	1	2	−2	−1	0	$\Delta w'$

Table 3.6: Possible values of $\Delta s'$ and $\Delta w'$.

Comparing Table 3.5 and 3.6, we find the collection of Δs and $\Delta s'$, Δw and $\Delta w'$, form the same set, respectively. Thus, we continue using s and w when rewriting

(3.6) in the form of (3.5), i.e.

$$x_{k+1} = \left(I + \sum_{j=0}^{4} v_k^j s_j \right) x_k' + \sum_{j=0}^{4} c_k'^j w_j, \tag{3.7a}$$

$$= (I + v_k s)\, x_k' + c_k' w, \tag{3.7b}$$

where v_k, v_k^j, c_k', $c_k'^j$ are the counterparts of u_k, u_k^j, c_k, c_k^j, respectively. And Proposition 3.2 follows.

Proposition 3.2.
Due to the $1-1$ *transfer in mutation, hence*

- *If* $x_k' = 0$ *and* $c_k'^j = 0, \forall j, 0 \le j \le 4$, *or* $c_k'^0 = 1, c_k'^j = 0, \forall j, 1 \le j \le 4 \Leftrightarrow x_k = 0$ *and* $c_k' = 0$, *or* $c_k' = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$, *then* $x_{k+1} = 0$.
- *If* $x_k' \neq 0$, $c_k'^j = 0, \forall j, 0 \le j \le 4$ *and* $\sum_{j=0}^{4} v_k^j = 0$ *or* $1 \Leftrightarrow c_k' = 0$ *and* v_k *is either* 0 *or a unit row vector.*
- *If* $x_k' = 0$, $v_k^j = 0, \forall j, 0 \le j \le 4$ *and* $\sum_{j=0}^{4} c_k'^j = 0$ *or* $1 \Leftrightarrow v_k = 0$ *and* c_k *is either* 0 *or a unit row vector.*
- $\sum_{j=0}^{4} v_k^j + c_k'^j = 0$ *or* $1, \forall k \in \mathbb{Z}^+ \cup \{0\} \Leftrightarrow v_k + c_k'$ *is either* 0 *or a unit row vector,* $\forall k \in \mathbb{Z}^+ \cup \{0\}$.

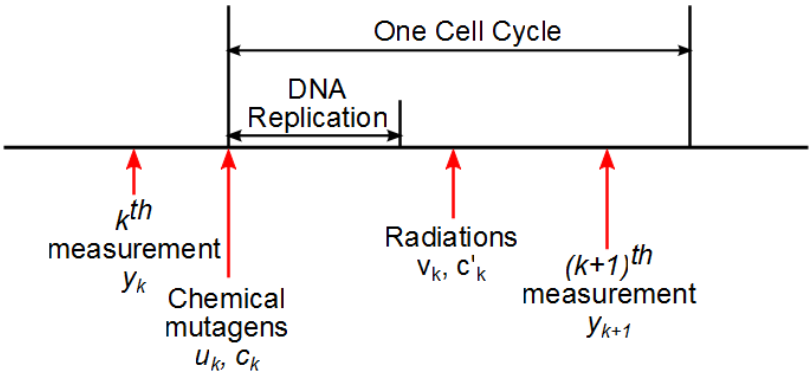

Figure 3.4: The order of taking measurements, applying chemical mutagens and radiation in a cell cycle.

Now take both chemical mutagens and radiation within our consideration. Since the order of applying chemical mutagens and radiation does not effect the result as

stated in Assumption 3.3, without loss of generality, we assume that radiation is always applied after a chemical mutagen in every cell cycle. A chemical mutagen is applied before duplication process starts, radiation is applied in the middle of every cell cycle, and a measurement is taken before every replication starts, as shown in Figure 3.4. Then we can express our system equations as

$$x'_k = \left(-I + \underbrace{u_k s}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from normal bases}}} \right) x_k + \underbrace{c_k w}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from } O}}, \tag{3.8a}$$

$$x_{k+1} = \left(I + \underbrace{v_k s}_{\substack{\text{mutations caused by radiative} \\ \text{rays from normal bases}}} \right) x'_k + \underbrace{c'_k w}_{\substack{\text{mutations caused by radiative} \\ \text{rays from } O}}, \tag{3.8b}$$

$$y_k = x_k, \tag{3.8c}$$

where u_k and v_k are the inputs of the system, and y_k is the measurement. Obviously, (3.8a) is modified from (3.5b), and (3.8b) from (3.7b). u_k, c_k obey Proposition 3.1, and v_k, c'_k obey Proposition 3.2.

The intermediate state x'_k of two-step mutation avoids those ambiguous cases in which x_k is changed to different bases by radiation and chemical mutagens simultaneously. Substituting (3.8a) into (3.8b), we get

$$x_{k+1} = (I + v_k s)(-I + u_k s)x_k + (I + v_k s)c_k w + c'_k w, \tag{3.9a}$$

$$y_k = x_k. \tag{3.9b}$$

Obviously, Proposition 3.1 still holds for u_k and c_k, and Proposition 3.2 holds for v_k and c'_k for (3.9a).

For a deterministic point mutation, we have 20 on/off controls in total for every step k, 10 for chemical mutagens as described before, and the rest for radiation.

3.4 Gene-to-gene, Deterministic Model

In general, several or more bases are involved when mutation happens. In those cases, large scale deterministic model is necessary. Now we show how to extend our model to large scale systems.

Suppose we have a DNA segment with length n, then $x_k \in \mathscr{F}^n$. Since there are integer number of codons, which contains three consecutive bases, in a coding DNA segment, n is generally a multiple of 3. Let x_k^i denote the i^{th} component of x_k. This notation is consistent with the one in §3.3. Again, we take a measurement before every replication starts and apply a chemical mutagen in front of radiation, if applicable, as shown in Figure 3.4. Initiated by the base-to-base deterministic model from §3.3, we write our system equation for large scale system as

$$x_k' = \left(-I + \underbrace{\sum_{i=1}^{n} u_k^i S_k^i}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from normal bases}}} \right) x_k + \underbrace{\sum_{i \in \mathscr{O}_k} c_k^i W_k^i}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from } O}} \quad , \text{(3.10a)}$$

$$x_{k+1} = \left(I + \underbrace{\sum_{i=1}^{n} v_k^i S_k'^i}_{\substack{\text{mutations caused by radiative} \\ \text{rays from normal bases}}} \right) x_k' + \underbrace{\sum_{i \in \mathscr{O}_k'} c_k'^i W_k'^i}_{\substack{\text{mutations caused by radiative} \\ \text{rays from } O}} \quad , \quad \text{(3.10b)}$$

$$y_k = x_k, \quad \text{(3.10c)}$$

where $u_k^i, v_k^i, c_k^i, c_k'^i$ are on/off controls of the i^{th} element, and $S_k^i, S_k'^i$ are $n \times n$ square matrices corresponding to mutations from normal bases induced by chemicals and radiation, respectively, $W_k^i, W_k'^i$ are n dimensional column vectors representing the process of mutations from the artificial non-sense base induced by chemicals and radiation, respectively, and $\mathscr{O}_k = \{i : x_k^i = 0, 1 \le i \le n\}$, $\mathscr{O}_k' = \{i : x_k^i = 0, 1 \le i \le n\}$.

Carefully examining (3.10), we discover S_k^i and $S_k'^i$ should be diagonal matrices according to Assumption 3.2 in §3.2. In addition, the values in Table 3.5 and 3.6 correspond to the diagonal elements of S_k^i and $S_k'^i$, respectively. The last row of Table 3.5 and 3.6 should be assigned to W_k^i and $W_k'^i$ at non-sense base's spots of x_k.

Instead of using step-varying $S_k^i, S_k'^i, W_k^i, W_k'^i$, we would like to find matrix basis to make $u_k^i, c_k^i, v_k^i, c_k'^i$ be the only variables depending on k, as we did for point mtuations. Then we can write (3.10) in a form similar to (3.8).

Define $\mathscr{S} = \{s_j e_i e_i^T, \forall i, j, 0 \leq j \leq 4, 1 \leq i \leq n\}$, a collection of $n \times n$ matrices, where s_j is the same as in (3.5), e_i is the unit column vector of length n with i^{th} component equals to 1 and all other components equal to 0, and $e_i e_i^T$ is the square matrix with only the i^{th} element on the diagonal equals to 1, and 0 otherwise.

S_k^i, an $n \times n$ diagonal matrix with the values of diagonal elements from the first four rows of Table 3.5, can always be written as the linear combination of all terms in $\mathscr{S}$, i.e.

$$S_k^i = \sum_{i=1}^{n} \sum_{j=0}^{4} \vartheta_{ij} s_j e_i e_i^T, \tag{3.11}$$

with $\vartheta_{ij} \in \{0, 1\}$.

Similarly, $S_k'^i$, a square matrix with the values of diagonal elements from the first four rows of Table 3.6 can be written as a linear combination of terms from $\mathscr{S}$,

$$S_k'^i = \sum_{i=1}^{n} \sum_{j=0}^{4} \vartheta'_{ij} s_j e_i e_i^T, \tag{3.12}$$

with $\vartheta'_{ij} \in \{0, 1\}$.

Define $\mathscr{W} = \{w_j e_i, \forall i, j, 0 \leq j \leq 4, 1 \leq i \leq n\}$, where w_j is the same as (3.5). So

$$W_k^i = \sum_{i=1}^{n} \sum_{j=0}^{4} \iota_{ij} w_j e_i, \tag{3.13}$$

with $\iota_{ij} \in \{0, 1\}$, and

$$W_k'^i = \sum_{i=1}^{n} \sum_{j=0}^{4} \iota'_{ij} w_j e_i, \tag{3.14}$$

with $\iota'_{ij} \in \{0,1\}$, are linear combinations of all terms in $\mathscr{W}$.

Substituting ϑ_{ij} by $u_k^{(i,j)}$ in (3.11), ϑ'_{ij} by $v_k^{(i,j)}$ in (3.12), ι_{ij} by $c_k^{(i,j)}$ in (3.13) and ι'_{ij} by $c_k^{\prime(i,j)}$ in (3.14), we rewrite (3.10) as

$$x'_k = \left(-I + \underbrace{\sum_{i=1}^{n}\sum_{j=0}^{4} u_k^{(i,j)} s_j e_i e_i^T}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from normal bases}}} \right) x_k + \underbrace{\sum_{i\in\mathscr{O}_k}\sum_{j=0}^{4} c_k^{(i,j)} w_j e_i}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from } O}} \quad , \tag{3.15a}$$

$$x_{k+1} = \left(I + \underbrace{\sum_{i=1}^{n}\sum_{j=0}^{4} v_k^{(i,j)} s_j e_i e_i^T}_{\substack{\text{mutations caused by radiative} \\ \text{rays from normal bases}}} \right) x'_k + \underbrace{\sum_{i\in\mathscr{O}'_k}\sum_{j=0}^{4} c_k^{\prime(i,j)} w_j e_i}_{\substack{\text{mutations caused by radiative} \\ \text{rays from } O}} \quad , \tag{3.15b}$$

$$y_k = x_k, \tag{3.15c}$$

where $u_k^{(i,j)}, v_k^{(i,j)}, c_k^{(i,j)}, c_k^{\prime(i,j)} \in \{0,1\}$.

If we define $u_k^i, v_k^i, c_k^i, c_k^{\prime i} \in \mathbb{F}_2^{1\times 5}$ as the i^{th} row of u_k, v_k, c_k, c'_k, respectively, with $u_k, v_k, c_k, c'_k \in \mathbb{F}_2^{n\times 5}$, $s = w = \begin{bmatrix} 0 & 1 & 2 & -2 & -1 \end{bmatrix}^T$, the same as in (3.5b), (3.15) can be simplified to

$$x'_k = \left(-I + \sum_{i=1}^{n} u_k^i s e_i e_i^T \right) x_k + \sum_{i\in\mathscr{O}_k} c_k^i w e_i, \tag{3.16a}$$

$$x_{k+1} = \left(I + \sum_{i=1}^{n} v_k^i s e_i e_i^T \right) x'_k + \sum_{i\in\mathscr{O}'_k} c_k^{\prime i} w e_i, \tag{3.16b}$$

$$y_k = x_k. \tag{3.16c}$$

Substituting (3.16a) into (3.16b), we get

$$x_{k+1} = \left(I + \sum_{i=1}^{n} v_k^i s e_i e_i^T\right)\left(-I + \sum_{i=1}^{n} u_k^i s e_i e_i^T\right) x_k + \left(I + \sum_{i=1}^{n} v_k^i s e_i e_i^T\right) \sum_{i \in \mathscr{O}_k} c_k^i w e_i + \sum_{i \in \mathscr{O}_k'} c_k'^i w e_i, \tag{3.17a}$$

$$y_k = x_k. \tag{3.17b}$$

Proposition 3.3.
For large scale deterministic system, u_k, v_k, c_k, c_k' follow the rules below.

- *If $e_i^T x_k = 0$, then $i \in \mathscr{O}_k$.*
- *If $e_i^T x_k = 0$, $c_k^{(i,j)} = 0, \forall j, 0 \le j \le 4$ or $c_k^{(i,0)} = 1, c_k^{(i,j)} = 0, \forall j, 1 \le j \le 4 \Leftrightarrow e_i^T x_k = 0$, $c_k^i = 0$ or $c_k^i = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$, then $i \in \mathscr{O}_k'$.*
- *$\forall i \notin \mathscr{O}_k$, $\sum_{j=0}^{4} u_k^{(i,j)} = 0$ or 1 and $c_k^{(i,j)} = 0, \forall j, 0 \le j \le 4 \Leftrightarrow u_k^i$ is either 0 or a row unit vector and $c_k^i = 0$.*
- *$\forall i \in \mathscr{O}_k$, $\sum_{j=0}^{4} c_k^{(i,j)} = 0$ or 1 and $u_k^{(i,j)} = 0, \forall j, 0 \le j \le 4 \Leftrightarrow c_k^i$ is either 0 a row unit vector and $u_k^i = 0$ for that particular i.*
- *$\forall i \notin \mathscr{O}_k'$, $\sum_{j=0}^{4} v_k^{(i,j)} = 0$ or 1 and $c_k'^{(i,j)} = 0, \forall j, 0 \le j \le 4 \Leftrightarrow v_k^i$ is either 0 or a row unit vector and $c_k'^i = 0$.*
- *$\forall i \in \mathscr{O}_k'$, $\sum_{j=0}^{4} c_k'^{(i,j)} = 0$ or 1 and $v_k^{(i,j)} = 0, \forall j, 0 \le j \le 4 \Leftrightarrow c_k'^i$ is either 0 or a row unit vector and $v_k^i = 0$.*
- *$\forall i, k, 1 \le i \le n, k \in \mathbb{Z}^+ \cup \{0\}$, $\sum_{j=0}^{4} u_k^{(i,j)} + c_k^{(i,j)} = 0$ or 1 and $\sum_{j=0}^{4} v_k^{(i,j)} + c_k'^{(i,j)} = 0$ or $1 \Leftrightarrow u_k^i + c_k^i$ is either 0 or a unit row vector and $v_k^i + c_k'^i$ is either 0 or a unit row vector.*

The generalized mathematical model we proposed in (3.16) or (3.17) is adaptive to other biological systems at the molecular level, for instance, broken DNA strand and transcription process.

Consider, broken DNA strands, as an example. DNA strand breaks due to various reasons. Our system equation can represent this phenomenon by dividing one system into small subsystems. Of course, significant fractured DNA strand is simply eliminated by cell mechanism to ensure the accuracy to DNA replication and gene expression. (3.18) represents a case of breaking one DNA segment into two subsequences by chemical mutagens.

$$\left(\frac{x'_k(1)}{x'_k(2)}\right) = \left(\begin{array}{c|c} -I_m + \sum_{i=1}^{m} u_k^i s e_i e_i^T & 0 \\ \hline 0 & -I_{n-m} + \sum_{i=m+1}^{n} u_k^i s e_i e_i^T \end{array}\right) \left(\frac{x_k(1)}{x_k(2)}\right) + \left(\frac{\sum_{i\in\mathscr{O}_k, 1\le i\le m} c_k^i w e_i}{\sum_{i\in\mathscr{O}_k, (m+1)\le i\le n} c_k^i w e_i}\right), \tag{3.18a}$$

$$x_{k+1}(1) = \left(I_m + \sum_{i=1}^{m} v_k^i s e_i e_i^T\right) x'_k(1) + \sum_{i\in\mathscr{O}'_k, 1\le i\le m} c_k'^i w e_i, \tag{3.18b}$$

$$x_{k+1}(2) = \left(I_{n-m} + \sum_{i=m+1}^{n} v_k^i s e_i e_i^T\right) x'_k(2) + \sum_{i\in\mathscr{O}'_k, (m+1)\le i\le n} c_k'^i w e_i. \tag{3.18c}$$

3.5 Gene-to-gene, Stochastic Model

Mutagens, no matter chemicals or radiation, always cause randomness in gene mutation. A gene-to-gene stochastic model is necessary to describe these conditions.

First, we introduce new random variables, $h_{k,l_1}^{(i,j)}, r_{k,l_2}^{i,j} \in \{0,1\}$, associated with probability, where k is the step index, l_1 is the mutagen index for mutagens inducing mutation from normal bases, l_2 is the mutagen index for mutagens inducing mutation from O, i is the index of DNA segment, and j is the index of set $\mathscr{S}$ and $\mathscr{W}$. We denote the probability associated with $h_{k,l_1}^{(i,j)}$ by $p_{l_1,j}^{(h)}$, and associated with $r_{k,l_2}^{(i,j)}$ by $p_{l_2,j}^{(r)}, \forall i,k, 1 \le i \le n, k \in \mathbb{Z}^+ \cup \{0\}$. Note different mutagens have different probability

assignments, and the probability assignments are only related to the type of mutagens. Our controls, for chemical mutagens, are $u^i_{k,l_1}, c^i_{k,l_2} \in \{0, 1\}$, with 1 representing the mutagen with corresponding index is applied at i^{th} spot of DNA segment at k^{th} generation, and 0 representing the mutagen with corresponding index is not applied at spot i at k^{th} step, similar to §3.3 and §3.4.

In an ideal case, suppose we have 16 kinds of chemical mutagens, each corresponding to a special mutation pattern, as listed in Table 3.7.

Mutagen index (l_1)	Major transfer pattern	Other possible transfers
1	$A \to A$	$A \to G$, $A \to C$, $A \to T$, $A \to O$
2	$A \to G$	$A \to A$, $A \to C$, $A \to T$, $A \to O$
3	$A \to C$	$A \to A$, $A \to G$, $A \to T$, $A \to O$
4	$A \to T$	
5	$G \to A$	$G \to G$, $G \to C$, $G \to T$, $G \to O$
6	$G \to G$	$G \to A$, $G \to C$, $G \to T$, $G \to O$
7	$G \to C$	
8	$G \to T$	$G \to G$, $G \to C$, $G \to T$, $G \to O$
9	$C \to A$	$C \to G$, $C \to C$, $C \to T$, $C \to O$
10	$C \to G$	
11	$C \to C$	$C \to A$, $C \to G$, $C \to T$, $C \to O$
12	$C \to T$	$C \to G$, $C \to C$, $C \to T$, $C \to O$
13	$T \to A$	
14	$T \to G$	$T \to A$, $T \to C$, $T \to T$, $T \to O$
15	$T \to C$	$T \to A$, $T \to G$, $T \to T$, $T \to O$
16	$T \to T$	$T \to G$, $T \to C$, $T \to T$, $T \to O$

Table 3.7: Possible transfer patterns by chemical mutagens.

Clearly, chemical mutagens indexed 4, 7, 10 and 13 are artificially added, since those transfer pairs are complementary to each other, which naturally happen in DNA replication. Therefore, we always assign $p_{4,AT} = p_{7,GC} = p_{10,CG} = p_{13,TA} = 1$ and $p_{4,A\cdot} = p_{7,G\cdot} = p_{10,C\cdot} = p_{13,T\cdot} = 0$ otherwise.

Table 3.8 shows the relationship between $h^{i,j}_{k,l_1}$ with associated probability $p^{(h)}_{l_1,j}$. The index j is the index of $s_j \in \mathscr{S}$ from Table 3.5.

Mutagens (l_1)	$h_{k,l}^{i,j}$ corresponding to major transfer	Probability associated with major transfer	Probability associated with minor transfers
1	$h_{k,1}^{i,2}$	$p_{1,AA}^{(h)} = p_{1,2}^{(h)}$	$p_{1,AG}^{(h)}, p_{1,AC}^{(h)}, p_{1,AT}^{(h)}, p_{1,AO}^{(h)}$
2	$h_{k,2}^{i,3}$	$p_{2,AG}^{(h)} = p_{2,3}^{(h)}$	$p_{2,AA}^{(h)}, p_{2,AC}^{(h)}, p_{2,AT}^{(h)}, p_{2,AO}^{(h)}$
3	$h_{k,3}^{i,4}$	$p_{3,AC}^{(h)} = p_{3,4}^{(h)}$	$p_{3,AA}^{(h)}, p_{3,AG}^{(h)}, p_{3,AT}^{(h)}, p_{3,AO}^{(h)}$
4	$h_{k,4}^{i,0}$	$p_{4,AT}^{(h)} = p_{4,0}^{(h)}$	
5	$h_{k,5}^{i,4}$	$p_{5,GA}^{(h)} = p_{5,4}^{(h)}$	$p_{5,GG}^{(h)}, p_{5,GC}^{(h)}, p_{5,GT}^{(h)}, p_{5,GO}^{(h)}$
6	$h_{k,6}^{i,2}$	$p_{6,GG}^{(h)} = p_{6,2}^{(h)}$	$p_{6,GA}^{(h)}, p_{6,GC}^{(h)}, p_{6,GT}^{(h)}, p_{6,GO}^{(h)}$
7	$h_{k,7}^{i,0}$	$p_{7,GC}^{(h)} = p_{7,0}^{(h)}$	
8	$h_{k,8}^{i,3}$	$p_{8,GT}^{(h)} = p_{8,3}^{(h)}$	$p_{8,GA}^{(h)}, p_{8,GG}^{(h)}, p_{8,GC}^{(h)}, p_{8,GO}^{(h)}$
9	$h_{k,9}^{i,3}$	$p_{9,CA}^{(h)} = p_{9,3}^{(h)}$	$p_{9,CG}^{(h)}, p_{9,CC}^{(h)}, p_{9,CT}^{(h)}, p_{9,CO}^{(h)}$
10	$h_{k,10}^{i,0}$	$p_{10,CG}^{(h)} = p_{10,0}^{(h)}$	
11	$h_{k,11}^{i,2}$	$p_{11,CC}^{(h)} = p_{11,2}^{(h)}$	$p_{11,CA}^{(h)}, p_{11,CG}^{(h)}, p_{11,CT}^{(h)}, p_{11,CO}^{(h)}$
12	$h_{k,12}^{i,4}$	$p_{12,CT}^{(h)} = p_{12,4}^{(h)}$	$p_{12,CA}^{(h)}, p_{12,CG}^{(h)}, p_{12,CC}^{(h)}, p_{12,CO}^{(h)}$
13	$h_{k,13}^{i,0}$	$p_{13,TA}^{(h)} = p_{13,0}^{(h)}$	
14	$h_{k,14}^{i,4}$	$p_{14,TG}^{(h)} = p_{14,4}^{(h)}$	$p_{2,TA}^{(h)}, p_{2,TC}^{(h)}, p_{2,TT}^{(h)}, p_{2,TO}^{(h)}$
15	$h_{k,15}^{i,3}$	$p_{15,TC}^{(h)} = p_{15,3}^{(h)}$	$p_{1,TA}^{(h)}, p_{1,TG}^{(h)}, p_{1,TT}^{(h)}, p_{1,TO}^{(h)}$
16	$h_{k,16}^{i,2}$	$p_{16,TT}^{(h)} = p_{16,2}^{(h)}$	$p_{3,TA}^{(h)}, p_{3,TG}^{(h)}, p_{3,TC}^{(h)}, p_{3,TO}^{(h)}$

Table 3.8: Probability assignments to random variables $h_{k,l_1}^{(i,j)}$s.

If we only apply chemical mutagens to a DNA segment, and ignore mutations from non-sense base O to normal bases, from the viewpoint of stochastic systems, our state space can be written as

$$x_{k+1} = \left(-I + \sum_{\substack{l_1=1 \\ l_1 \neq 4,7,10,13}}^{16} \sum_{i=1}^{n} u_{k,l_1}^{i} \sum_{j=0}^{4} h_{k,l_1}^{(i,j)} s_j e_i e_i^T \right) x_k, \tag{3.19}$$

where u_{k,l_1}^{i} denotes the on/off switch of applying mutagen l_1 at i^{th} spot to k^{th} generation of the DNA segment and $h_{k,l_1}^{(i,j)}$ decides the i^{th} spot nucleotide base of $(k+1)^{th}$ generation if $u_{k,l_1}^{i} = 1$.

Rewriting $h_{k,l_1}^{(i,j)}$ in the vector form, as we did for $u_k^{(i,j)}$ in §3.4, we can eliminate index j. Subsequently, (3.19) becomes

$$x_{k+1} = \left(-I + \sum_{\substack{l_1=1 \\ l_1 \neq 4,7,10,13}}^{16} \sum_{i=1}^{n} u_{k,l_1}^{i} h_{k,l_1}^{i} s e_i e_i^T \right) x_k, \tag{3.20}$$

where $h_{k,l_1}^{i} \in \mathbb{F}_2^{1\times 5}$ and $s = \begin{bmatrix} 0 & 1 & 2 & -2 & -1 \end{bmatrix}^T$.

In practice, we may have several mutagens corresponding to one transfer pair, or no mutagen for one or several transfer pairs. Hence, we replace 16 by l. In addition, we assume our system is completely controllable.

Remark 3.1. *DNA replication systems with system equations proposed as* (3.8)*,* (3.9)*,* (3.15)*,* (3.16)*,* (3.17)*,* (3.21)*,* (3.22) *and* (3.23) *are* ***completely controllable*** *if and only if* $\forall x_0, x_{2k_1}, x_{2k_2+1} \in \mathscr{F}, k_1, k_2 \in \mathbb{Z}^+ \cup \{0\}, \exists$ *at least one path from* x_0 *to* x_{2k_1} *and at least one path from* x_0 *to* x_{2k_2+1} *by applying proper mutagens in the correct order, with* k_1, k_2 *finite.*

Incorporating terms corresponding to mutations from normal bases induced by radiation, and mutations from non-sense base O induced by chemical mutagens and radiation into stochastic system equation, we get (3.21).

$$x'_k = \left(-I + \underbrace{\sum_{l_1=1}^{l} \sum_{i=1}^{n} u^i_{k,l_1} \sum_{j=0}^{4} h^{(i,j)}_{k,l_1} s_j e_i e_i^T}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from normal bases}}} \right) x_k + \underbrace{\sum_{l_2=1}^{m} \sum_{i \in \mathscr{O}_k} c^i_{k,l_2} \sum_{j=0}^{4} r^{(i,j)}_{k,l_2} w_j e_i}_{\substack{\text{mutations caused by chemical} \\ \text{mutagens from } O}}, \tag{3.21a}$$

$$x_{k+1} = \left(I + \underbrace{\sum_{l_3=1}^{l'} \sum_{i=1}^{n} v^i_{k,l_3} \sum_{j=0}^{4} h'^{(i,j)}_{k,l_3} s_j e_i e_i^T}_{\substack{\text{mutations caused by radiative} \\ \text{rays from normal bases}}} \right) x'_k + \underbrace{\sum_{l_4=1}^{m'} \sum_{i \in \mathscr{O}'_k} c'^i_{k,l_4} \sum_{j=0}^{4} r'^{(i,j)}_{k,l_4} w_j e_i}_{\substack{\text{mutations caused by radiative} \\ \text{rays from } O}}, \tag{3.21b}$$

$$y_k = x_k. \tag{3.21c}$$

Simplify (3.21), we have

$$x'_k = \left(-I + \sum_{l_1=1}^{l} \sum_{i=1}^{n} u^i_{k,l_1} h^i_{k,l_1} s e_i e_i^T \right) x_k + \sum_{l_2=1}^{m} \sum_{i \in \mathscr{O}_k} c^i_{k,l_2} r^i_{k,l_2} w e_i, \tag{3.22a}$$

$$x_{k+1} = \left(I + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} v^i_{k,l_3} h'^i_{k,l_3} s e_i e_i^T \right) x'_k + \sum_{l_4=1}^{m'} \sum_{i \in \mathscr{O}'_k} c'^i_{k,l_4} r'^i_{k,l_4} w e_i, \tag{3.22b}$$

$$y_k = x_k, \tag{3.22c}$$

where $h^i_{k,l_1}, r^i_{k,l_2}, h'^i_{k,l_3}, h'^i_{k,l_4} \in \mathbb{F}_2^{1\times 5}$.

Instead of having 20 controls, including 5 $u^{(i,j)}_k$s, 5 $c^{(i,j)}_k$s, 5 $v^{(i,j)}_k$s and 5 $c'^{(i,j)}_k$s for each spot i as in gene-to-gene deterministic cases, we have $(l+m+l'+m')$ on/off controls at every spot i, i.e. l u^i_{k,l_1}s, m c^i_{k,l_2}s, l' v^i_{k,l_3}s and m' c'^i_{k,l_4}s.

Substituting (3.22a) into (3.22b), we have

$$\begin{aligned} x_{k+1} = & \left(I + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} v_{k,l_3}^{i} h_{k,l_3}^{\prime i} s e_i e_i^T \right) \left(-I + \sum_{l_1=1}^{l} \sum_{i=1}^{n} u_{k,l_1}^{i} h_{k,l_1}^{i} s e_i e_i^T \right) x_k \\ & + \left(I + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} v_{k,l_3}^{i} h_{k,l_3}^{\prime i} s e_i e_i^T \right) \sum_{l_2=1}^{m} \sum_{i \in \mathscr{O}_k} c_{k,l_2}^{i} r_{k,l_2}^{i} w e_i \\ & + \sum_{l_4=1}^{m'} \sum_{i \in \mathscr{O}_k'} c_{k,l_4}^{\prime i} r_{k,l_4}^{\prime i} w e_i, \end{aligned} \tag{3.23a}$$

$$y_k = x_k. \tag{3.23b}$$

Proposition 3.4 states the rules $u_{k,l_1}^{i}, h_{k,l_1}^{i}, c_{k,l_2}^{i}, r_{k,l_2}^{i}, v_{k,l_3}^{i}, h_{k,l_3}^{\prime i}, c_{k,l_4}^{\prime i}, r_{k,l_4}^{\prime i}$ in (3.22) and (3.23) need to follow.

Proposition 3.4.
For large-scale stochastic system, $u_{k,l_1}^{i}, h_{k,l_1}^{i}, c_{k,l_2}^{i}, r_{k,l_2}^{i}, v_{k,l_3}^{i}, h_{k,l_3}^{\prime i}, c_{k,l_4}^{\prime i}, r_{k,l_4}^{\prime i}$ *follow the rules below.*

- *If* $e_i^T x_k = 0$, *then* $i \in \mathscr{O}_k$.
- *If* $e_i^T x_k = 0$ *and* $\sum_{l_2=1}^{m} c_{k,l_2}^{i} = 0$, *then* $i \in \mathscr{O}_k'$.
- *If* $e_i^T x_k = 0$, $\sum_{l_2=1}^{m} c_{k,l_2}^{i} = 1$ *and* $r_{k,l_2}^{(i,0)} = 1, r_{k,l_2}^{(i,j)} = 0, \forall j, 1 \le j \le 4 \Leftrightarrow e_i^T x_k = 0$, $\sum_{l_2=1}^{m} c_{k,l_2}^{i} = 1$ *and* $r_{k,l_2}^{i} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \end{bmatrix}$, *then* $i \in \mathscr{O}_k'$.
- $\forall i, k, l_1, 1 \le i \le n, k \in \mathbb{Z}^+ \cup \{0\}, 1 \le l_1 \le l$, *if* $u_{k,l_1}^{i} = 1$, *then* $\sum_{j=0}^{4} h_{k,l_1}^{(i,j)} = 1 \Leftrightarrow$ h_{k,l_1}^{i} *is a unit row vector.*
- $\forall i, k, l_2, 1 \le i \le n, k \in \mathbb{Z}^+ \cup \{0\}, 1 \le l_2 \le m$, *if* $c_{k,l_2}^{i} = 1$, *then* $\sum_{j=0}^{4} r_{k,l_2}^{(i,j)} = 1$ $\Leftrightarrow r_{k,l_2}^{i}$ *is a unit row vector.*
- $\forall i, k, l_3, 1 \le i \le n, k \in \mathbb{Z}^+ \cup \{0\}, 1 \le l_3 \le l'$, *if* $v_{k,l_3}^{i} = 1$, *then* $\sum_{j=0}^{4} h_{k,l_3}^{\prime(i,j)} = 1$ $\Leftrightarrow h_{k,l_3}^{\prime i}$ *is a unit row vector.*
- $\forall i, k, l_4, 1 \le i \le n, k \in \mathbb{Z}^+ \cup \{0\}, 1 \le l_4 \le m'$, *if* $c_{k,l_4}^{\prime i} = 1$, *then* $\sum_{j=0}^{4} r_{k,l_4}^{\prime(i,j)} = 1$ $\Leftrightarrow r_{k,l_4}^{\prime i}$ *is a unit row vector.*

- $\forall i \notin \mathscr{O}_k$, $\sum_{l_1=1}^{l} u_{k,l_1}^{i} = 0$ *or* 1 *and* $c_{k,l_2}^{i} = 0, \forall l_2, 1 \leq l_2 \leq m$.
- $\forall i \in \mathscr{O}_k$, $\sum_{l_2=1}^{m} c_{k,l_2}^{i} = 0$ *or* 1 *and* $u_{k,l_1}^{i} = 0, \forall l_1, 1 \leq l_1 \leq l$.
- $\forall i \notin \mathscr{O}'_k$, $\sum_{l_3=1}^{l'} v_{k,l_3}^{i} = 0$ *or* 1 *and* $c'^{i}_{k,l_4} = 0, \forall l_4, 1 \leq l_4 \leq m'$.
- $\forall i \in \mathscr{O}'_k$, $\sum_{l_4=1}^{m'} c'^{i}_{k,l_4} = 0$ *or* 1 *and* $v_{k,l_3}^{i} = 0, \forall l_3, 1 \leq l_3 \leq l''$.
- $\forall i, k, 1 \leq i \leq n, k \in \mathbb{Z}^{+} \cup \{0\}$, $\sum_{l_1=1}^{l} u_{k,l_1}^{i} + \sum_{l_2=1}^{m} c_{k,l_2}^{i} = 0$ *or* 1 *and* $\sum_{l_3=1}^{l'} v_{k,l_3}^{i} + \sum_{l_4=1}^{m'} c'^{i}_{k,l_4} = 0$ *or* 1.

The probabilities associated with $h_{k,l_1}^{(i,j)}, h'^{(i,j)}_{k,l_3}, r_{k,l_2}^{(i,j)}, r'^{(i,j)}_{k,l_4}$ sum up to 1, respectively, as stated in Proposition 3.5.

Proposition 3.5.

$$\sum_{j=0}^{4} p_{l_1,j}^{(h)} = 1, \forall l_1, 1 \leq l_1 \leq l. \tag{3.24a}$$

$$\sum_{j=0}^{4} p_{l_2,j}^{(r)} = 1, \forall l_2, 1 \leq l_2 \leq m. \tag{3.24b}$$

$$\sum_{j=0}^{4} p_{l_3,j}^{(h')} = 1, \forall l_3, 1 \leq l_3 \leq l'. \tag{3.24c}$$

$$\sum_{j=0}^{4} p_{l_4,j}^{(r')} = 1, \forall l_4, 1 \leq l_4 \leq m'. \tag{3.24d}$$

Similar to the gene-to-gene deterministic model, the gene-to-gene stochastic system we derived is a generalized model and is adaptive to other stochastic biological systems at the molecular level, such as multi-site mutations within one stage, broken DNA strands and transcription process.

Chapter 4

Optimal Control

In this chapter, we formulate objective functions for dynamic systems constructed in Chapter 3, and compute the optimal trajectories and minimum costs to drive the system from initial state to the final desired set.

In §4.1, we present a generalized objective function, consisting of the costs and risks of applying mutagens, and off-trajectory penalty. The major difficulty in formulating an objective function is to define a proper metric over the finite field. By taking the redundancy of genetic codes and both physical and chemical properties of amino acids within our consideration, we define an alternative distance reference which serves as off the preset trajectory penalty in §4.2. §4.3 begins with some basic information about the dynamic programming algorithm, and then we show the details of applying this algorithm to solve the generalized optimization problem.

In §4.4, §4.5 and §4.6, we carefully present the optimal control problems in single-base deterministic case, codon-to-codon deterministic case and codon-to-codon stochastic case, respectively. Examples and simulation results of different scales and parameter assignments are illustrated and compared. When solving optimal control problem for a deterministic point mutation, we find that the global optimal can be reached within finite steps if the system is completely controllable, which is essential in finding the solution to optimal control problem without constraints on the number of steps. This result can be further extended to codon-to-codon deterministic case.

4.1 Objective Function Formulation

As mentioned in Chapter 2, our goal is to find the optimal control policy to drive the system from an initial state to a final desired set, which is generated by a final desired state, follow a predefined trajectory, which minimizes the sum of the total cost (including the risk) of applying mutagens and off-trajectory penalty. To formulate an optimal control problem, we begin with defining a proper objective function to quantitatively describe our physical goal. Moreover, this formulation needs to be adaptive to all kinds of system models we proposed in Chapter 3.

Mathematically, in systems where the controllable parameters of interest are discrete, the objective function is usually a weighted sum representing the number of times that a piece of equipment is turned "on" or "off", or the number of resources needed to execute certain tasks in the frequent cases [Hristu-Varsakelis and Levine, 2005]. Hence, our objective function includes a sum of the total number of times that different mutagens are applied weighted by the corresponding cost (including the risk) of applying them.

In medical practice, a predefined trajectory is of great importance in avoiding hidden risks, for instance, disruption of the cell cycle or early stop of transcription process, which lead of abnormal metabolism or diseases including cancer. In our problem, such trajectory is indeed a collection of trajectories generated from a desired trajectory. In other words, a desired set at each stage is generated by the corresponding state of the desired trajectory, containing all nucleotide sequences that are eventually translated to the same animo acid sequence as the desired state at the same stage, and possible desired trajectories are generated by picking one state from the desired set at each stage and orderly linking them together. Therefore, we do not attempt to compensate any silent mutations.

Obviously, the set of predefined trajectories is updated according to current measurement. If current measurement implies the current state is off-trajectory, a distance reference between current state and the desired set at current stage is added as a penalty cost. Such distance reference is different from common metric defined on finite field, because of the physical and chemical properties of animo acid sequences. For example, the distance reference between any two nucleotide sequences which are

translated to the same amino acid sequence is set to be zero. And the distance reference of sequences is defined based on the distance reference of codons. Details about how to a sample distance reference for codons later are illustrated in §4.2. Practically, this distance reference is defined by doctors or biologists based on experimental results.

The constraint of our optimal control problem is the corresponding system equation. We choose multi-dimensional stochastic system equation as the generalized constraints as it can be reduced to one-dimensional and multi-dimensional deterministic cases by modifying the probability distribution associated with random variables. Moreover, the control sequence of such systems contains information about the order and the types of mutagens applied, which is important in objective function formulation.

Therefore, our objective function can be mathematically expressed as

$$\begin{aligned} J_0(x_0) = \min_{u,c,v,c'} \mathop{\mathbb{E}}_{h,h',r,r'} \Bigg[& \underbrace{\sum_{k=0}^{N-1} \sum_{l_1=1}^{l} \sum_{i=1}^{n} \alpha_{l_1} u_{k,l_1}^{i} + \sum_{k=0}^{N-1} \sum_{l_2=1}^{m} \sum_{i=1}^{n} \beta_{l_2} c_{k,l_2}^{i}}_{\text{cost of applying chemical mutagens}} \\ & + \underbrace{\sum_{k=0}^{N-1} \sum_{l_3=1}^{l'} \sum_{i=1}^{n} \alpha'_{l_3} v_{k,l_3}^{i} + \sum_{k=0}^{N-1} \sum_{l_4=1}^{m'} \sum_{i=1}^{n} \beta'_{l_4} c'^{i}_{k,l_4}}_{\text{cost of applying radiative rays}} \\ & + \underbrace{\sum_{k=0}^{N} d\left(x_k, \left\{x_k^d\right\}\right)}_{\text{tracing cost}} \Bigg], \end{aligned} \tag{4.1}$$

with $x_0, x_k^d \in \mathscr{F}^n, 1 \leq k \leq N$ given. l_1, l_2, l_3, l_4 are the indices of chemical mutagens inducing mutations from normal bases and from O, and radiation inducing mutations from normal bases and from O, respectively. $\alpha_{l_1}, \beta_{l_2}, \alpha'_{l_3}, \beta'_{l_4}$ are the corresponding cost (including the risk) of applying chemical mutagens and radiation indexed l_1, l_2, l_3, l_4, respectively. $\{x_k^d\}$ denotes the set of nucleotide sequences which are eventually translated to the same amino acid sequence as x_k^d, or equivalently, the desired state at k^{th} step. And $d\left(x_k, \left\{x_k^d\right\}\right)$ is the distance reference between current state x_k and the desired set $\{x_k^d\}$ at k^{th} step. The final penalty, the distance reference between final state to the desired set at $k = N$, is included in the last term. u_{k,l_1}^i, c_{k,l_2}^i, v_{k,l_3}^i,

$c'^{i}_{k,l_4} \in \{0,1\}$, inputs of the systems, are the on/off controls, whose physical meanings are the same as defined in §3.5.

In general, $\beta_{l_2}, \beta'_{l_4} \ll \alpha_{l_1}, \alpha'_{l_3}, \forall l_1, l_2, l_3, l_4, 1 \leq l_1 \leq l, 1 \leq l_2 \leq m, 1 \leq l_3 \leq l', 1 \leq l_4 \leq m'$, because O is a set of non-sense bases physically and more information is necessary to convert an O to normal bases, for instance, the cost of identifying the exact element in the set O. Our goal is to drive our system optimally from initial state x_0 to the desired final set $\left\{x_N^d\right\}$ by applying a sequence of mutagens indexed with $\{l_1, l_2, l_3, l_4\}$, at problematic positions i, and in a correct order k.

In (4.1), the first two terms inside the expectation are the portion of costs for transferring a DNA segment from the initial state x_0 to the final state x_N with fixed N by chemical mutagens, and the third and fourth terms are the portion generated from applying radiation. These four terms do not depend on random variables h^i_{k,l_1}, r^i_{k,l_2}, h'^{i}_{k,l_3} and r'^{i}_{k,l_4}, $\forall i, k, l_1, l_2, l_3, l_4$, as the treatment plan is computed based on the initial state x_0. Given y_k, the updated treatment plan is computed accordingly, but still not related to random variables. The last term inside expectation, $\sum_{k=0}^{N} d\left(x_k, \left\{x_k^d\right\}\right)$, is the only term in summation that depends on the distribution of the random variables.

Therefore, we can rewrite our objective function, and formulate our optimal control problem as

$$\begin{aligned} J_0(x_0) = \min_{\{u,c,v,c'\}_{0,1,\cdots,N-1}} \Bigg[& \sum_{k=0}^{N-1} \sum_{l_1=1}^{l'} \sum_{i=1}^{n} \alpha_{l_1} u^i_{k,l_1} + \sum_{k=0}^{N-1} \sum_{l_2=1}^{m} \sum_{i=1}^{n} \beta_{l_2} c^i_{k,l_2} \\ & + \sum_{k=0}^{N-1} \sum_{l_3=1}^{l'} \sum_{i=1}^{n} \alpha'_{l_3} v^i_{k,l_3} + \sum_{k=0}^{N-1} \sum_{l_4=1}^{m'} \sum_{i=1}^{n} \beta'_{l_4} c'^{i}_{k,l_4} \\ & + \sum_{k=0}^{N} \mathop{\mathbb{E}}_{\{h,r,h',r'\}_{0,1,\cdots,N-1}} \left[d\left(x_k, \left\{x_k^d\right\}\right) \right] \Bigg], \end{aligned} \tag{4.2}$$

subject to

$$x_{k+1} = \left(I + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} v_{k,l_3}^{i} h_{k,l_3}^{\prime i} s e_i e_i^T \right) \left(-I + \sum_{l_1=1}^{l} \sum_{i=1}^{n} u_{k,l_1}^{i} h_{k,l_1}^{i} s e_i e_i^T \right) x_k + \left(I + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} v_{k,l_3}^{i} h_{k,l_3}^{\prime i} s e_i e_i^T \right) \sum_{l_2=1}^{m} \sum_{i \in \mathscr{O}_k} c_{k,l_2}^{i} r_{k,l_2}^{i} w e_i + \sum_{l_4=1}^{m'} \sum_{i \in \mathscr{O}_k'} c_{k,l_4}^{\prime i} r_{k,l_4}^{\prime i} w e_i, \tag{4.3a}$$

$$y_k = x_k, \tag{4.3b}$$

with $x_0, x_k^d \in \mathscr{F}^n, 1 \leq k \leq N$ given, $\alpha_{l_1}, \beta_{l_2}, \alpha_{l_3}', \beta_{l_4}' \in \mathbb{R}, \forall l_1, l_2, l_3, l_4, 1 \leq l_1 \leq l, 1 \leq l_2 \leq m, 1 \leq l_3 \leq l', 1 \leq l_4 \leq m'$, $d : \mathscr{F}^n \times \mathscr{F}^n \to \mathbb{R}^+ \cup \{0\}$, $x_k, y_k \in \mathscr{F}^n$ with $n \equiv 0 \pmod 3$, $u_{k,l_1}^i, c_{k,l_2}^i, v_{k,l_3}^i, c_{k,l_4}^{\prime i} \in \{0,1\}, \forall i, k, l_1, l_2, l_3, l_4$, $s = w = \begin{bmatrix} 0 & 1 & 2 & -2 & -1 \end{bmatrix}^T$, and $h_{k,l_1}^i, r_{k,l_2}^i, h_{k,l_3}^{\prime i}, r_{k,l_4}^{\prime i} \in \{e_j \in \mathbb{R}^5, e_j \text{ unit column vector}, 1 \leq j \leq 5\}, \forall i, k, l_1, l_2, l_3, l_4$.

Discussions and examples of base-to-base deterministic case, codon-to-codon deterministic case and codon-to-codon stochastic case are demonstrated in §4.4, §4.5 and §4.6, respectively.

4.2 Distance Reference

As mentioned in §4.1, we need to define a proper distance reference to quantitatively describe the relationship between codons. In this section, we mainly focus on codons composed of normal nucleotide bases, and codons containing the artificial base O are omitted since their chemical and physical properties cannot be found in literature. We first define the distance reference between codons, and then extend it to DNA sequences.

Use $d(\varphi_1, \varphi_2), \varphi_1, \varphi_2 \in \mathscr{F}^3$ to denote the distance reference between two codons, φ_1 and φ_2. The distance reference needs to fulfill the biological requirements as below.

1. (Non-negativity) The distance reference between any two codons is either positive or zero. Mathematically, $d : \mathscr{F}^3 \times \mathscr{F}^3 \to \mathbb{R}^+ \cup \{0\}$, $d(\varphi_1, \varphi_2) \geq 0$.

2. The distance reference between two codons corresponding to the same amino acid is zero.

3. (Symmetry) The distance reference from codon φ_1 to codon φ_2 equals to the distance reference from codon φ_2 to codon φ_1, i.e. $d(\varphi_1, \varphi_2) = d(\varphi_2, \varphi_1)$.

4. The distance reference between two codons corresponding to different amino acids should reveal the chemical and physical differences between two amino acids.

5. The distance references from stop codons to all other codons is much larger than those between other codons as early termination of amino acid sequences is more deleterious than other forms of mutations.

All the existing metric defined on the finite field cannot achieve all the requirements above. The second requirement violates the identity of indiscernible, i.e. $d(\varphi_1, \varphi_2) = 0$ if and only if $\varphi_1 = \varphi_2$. The redundancy in genetic codes implies $d(\varphi_1, \varphi_2) = 0$ if those two amino acids, φ_1 and φ_2, correspond to the same amino acids according to the genetic codes. In addition, the triangular inequality is not necessarily true, according to the underlying physical meanings. For instance, it is impossible to judge if one codon is closer to the stop codon than another. Therefore, we take the assumption that stop codons are of the same distance reference from and to all other codons.

Important physical and chemical properties are listed in Table 4.1. The polarity property is the opposite of hydrophobicity, i.e. polar amino acids are hydrophilic, and non-polar are hydrophobic. The last two columns of Table 4.1 are related to each other, and only of them is considered when defining the distance reference.

From Table 4.1, we can see all codons are divided into different sets with each set corresponding to the same amino acid. The size and the elements in one codon set vary from one amino acid to another. This implies that the costs of driving one codon to the desired final set generated by the desired final state might be different from the costs of driving the complementary codon to the desired final set generated by the complementary of desired final state. In other words, there is no symmetric

Amino Acid	Abbrev.	Codon(s)	Polarity	PH	Avg. Mass(Da)	Size
Alanine	Ala	GCT, GCC, GCA, GCG	non-polar	6.01	89.09404	tiny
Arginine	Arg	$CGA, CGG, CGC, CGT, AGA, AGG$	polar	10.76	174.20274	normal
Asparagine	Asn	AAC, AAT	polar	5.41	132.11904	small
Aspartic acid	Asp	GAT, GAC	polar	2.85	133.10384	small
Cysteine	Cys	TGT, TGC	non-polar	5.05	121.15404	small
Glutamine	Gln	CAA, CAG	polar	5.65	146.14594	normal
Glutamic acid	Glu	GAA, GAG	polar	3.15	147.13074	normal
Glycine	Gly	GGA, GGG, GGC, GGT	non-polar	6.06	75.06714	tiny
Histidine	His	CAC, CAT	polar	7.60	155.15634	normal
Isoleucine	Ile	ATA, ATC, ATT	non-polar	6.05	131.17464	normal
Leucine	Leu	$TTA, TTG, CTA, CTG, CTC, CTT$	non-polar	6.01	131.17464	normal
Lysine	Lys	AAA, AAG	polar	9.60	146.18934	normal
Methionine	Met	ATG	non-polar	5.74	149.20784	normal
Phenylalanine	Phe	TTC, TTT	non-polar	5.49	165.19184	normal
Proline	Pro	CCA, CCG, CCC, CCT	non-polar	6.30	115.13194	small
Serine	Ser	$TCA, TCG, TCC, TCT, AGT, AGC$	polar	5.68	105.09344	tiny
Threonine	Thr	ACT, ACC, ACA, ACG	polar	5.60	119.12034	small
Tryptophan	Trp	TCC	non-polar	5.89	204.22844	normal
Tyrosine	Tyr	TAC, TAT	polar	5.64	181.19124	normal
Valine	Val	GTA, GTG, GTC, GTT	non-polar	6.00	117.14784	small
Stop codon	Term	TAA, TAG, TGA	-	-	-	-

Table 4.1: Properties of amino acids.

relationship between the sets generated by complementary codons. More discussions about this issue are presented in §4.5.

The distance reference between any two codons can be defined by any reasonable functions. Here, we use a weighted sum of the differences between physical and chemical properties as an example. And the distance reference between two DNA sequences are defined as a weighted sum of distance reference between the corresponding pair of codons. The biological statics plays a crucial rule to define this distance function practically.

An example of the distance reference can be expressed as

$$d(\xi_1,\xi_2) = \zeta_{polarity} polarity(\xi_1,\xi_2) + \zeta_{PH} PH(\xi_1,\xi_2) + \zeta_{size} size(\xi_1,\xi_2), \quad (4.4)$$

$$polarity(\xi_1,\xi_2) = \begin{cases} 0 & \text{if } \xi_1,\xi_2 \text{ are both polar or non-polar,} \\ 1 & \text{if one of } \xi_1,\xi_2 \text{ is polar, and the other non-polar;} \end{cases}$$

$$PH(\xi_1,\xi_2) = |\text{PH value of } \xi_1 - \text{PH value of } \xi_2|;$$

$$size(\xi_1,\xi_2) = \begin{cases} 0 & \text{if } \xi_1,\xi_2 \text{ are both tiny, small, or normal,} \\ \sigma_1 & \text{if one of } \xi_1,\xi_2 \text{ is tiny, and the other small,} \\ \sigma_2 & \text{if one of } \xi_1,\xi_2 \text{ is tiny, and the other normal,} \\ \sigma_3 & \text{if one of } \xi_1,\xi_2 \text{ is small, and the other normal,} \end{cases}$$

where ξ_1, ξ_2 are two amino acids.

The last term in (4.4) can be substituted by $\zeta_{mass} mass(\xi_1,\xi_2)$ with

$$mass(\xi_1,\xi_2) = |\text{average mass of } \xi_1 - \text{average mass of } \xi_2|.$$

In Table 4.2, we show an example of distance reference computed by (4.4) with $\zeta_{polarity} = 8$, $\zeta_{PH} = 3$, $\zeta_{size} = 1$, $\sigma_1 = 2$, $\sigma_2 = 5$ and $\sigma_3 = 3$.

The distance reference is then assigned to pairs of codons according to the genetic codes listed in Table 3.2, which is used for simulations in §4.4, §4.5 and §4.6.

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val	Term
Ala	0.00	27.25	11.80	19.48	4.88	14.08	21.58	0.15	17.77	5.12	5.00	22.15	5.81	6.56	2.87	8.99	11.23	5.36	14.11	2.03	50.00
Arg	27.25	0.00	19.05	26.73	28.13	15.33	22.83	27.10	9.48	22.13	22.25	5.10	23.06	23.81	24.38	20.24	18.48	22.61	15.36	25.28	50.00
Asn	11.80	19.05	0, 00	7.68	9.08	3.72	9.78	11.95	9.57	12.92	12.80	13.95	11.99	11.24	10.67	2.81	0.57	12.44	3.69	9.77	50.00
Asp	19.48	26.73	7.68	0.00	14.60	11.40	3.90	19.63	17.25	20.60	20.48	21.63	19.67	18.92	18.35	10.49	8.25	20.12	11.37	17.45	50.00
Cys	4.88	28.13	9.08	14.60	0.00	12.80	16.70	5.03	18.65	6.00	5.88	23.03	5.07	4.32	3.75	11.89	9.65	5.52	12.77	2.85	50.00
Gln	14.08	15.33	3.72	11.40	12.80	0.00	7.50	14.23	5.85	9.20	9.08	10.23	8.27	8.48	12.95	5.09	3.15	8.72	0.03	12.05	50.00
Glu	21.58	22.83	9.78	3.90	16.70	7.50	0.00	21.73	13.35	16.70	16.58	17.73	15.77	15.02	20.45	12.59	10.35	16.22	7.47	19.55	50.00
Gly	0.15	27.10	11.95	19.63	5.03	14.23	21.73	0.00	17.62	5.03	5.15	22.00	5.96	6.71	2.72	9.14	11.38	5.51	14.26	2.18	50.00
His	17.77	9.48	9.57	17.25	18.65	5.85	13.35	17.62	0.00	12.65	12.77	4.38	13.58	14.33	14.90	10.76	9.00	13.13	5.88	15.80	50.00
Ile	5.12	22.13	12.92	20.60	6.00	9.20	16.70	5.03	12.65	0.00	0.12	17.03	0.93	1.68	3.75	14.11	12.35	0.48	9.23	3.15	50.00
Leu	5.00	22.25	12.80	20.48	5.88	9.08	16.58	5.15	12.77	0.12	0.00	17.15	0.81	1.56	3.87	13.99	12.23	0.36	9.11	3.03	50.00
Lys	22.15	5.10	13.95	21.63	23.03	10.23	17.73	22.00	4.38	17.03	17.15	0.00	17.96	18.71	19.28	15.14	13.38	17.51	10.26	20.18	50.00
Met	5.81	23.06	11.99	19.67	5.07	8.27	15.77	5.96	13.58	0.93	0.81	17.96	0.00	0.75	4.68	13.18	11.42	0.45	8.30	3.78	50.00
Phe	6.56	23.81	11.24	18.92	4.32	8.48	15.02	6.71	14.33	1.68	1.56	18.71	0.75	0.00	5.43	13.57	11.33	1.20	8.45	4.53	50.00
Pro	2.87	24.38	10.67	18.35	3.75	12.95	20.45	2.72	14.90	3.75	3.87	19.28	4.68	5.43	0.00	11.86	10.10	4.23	12.98	0.90	50.00
Ser	8.99	20.24	2.81	10.49	11.89	5.09	12.59	9.14	10.76	14.11	13.99	15.14	13.18	13.57	11.86	0.00	2.24	13.63	5.12	10.96	50.00
Thr	11.23	18.48	0.57	8.25	9.65	3.15	10.35	11.38	9.00	12.35	12.23	13.38	11.42	11.33	10.10	2.24	0.00	11.87	3.12	9.20	50.00
Trp	5.36	22.61	12.44	20.12	5.52	8.72	16.22	5.51	13.13	0.48	0.36	17.51	0.45	1.20	4.23	13.63	11.87	0.00	8.75	3.33	50.00
Tyr	14.11	15.36	3.69	11.37	12.77	0.03	7.47	14.26	5.88	9.23	9.11	10.26	8.30	8.45	12.98	5.12	3.12	8.75	0.00	12.08	50.00
Val	2.03	25.28	9.77	17.45	2.85	12.05	19.55	2.18	15.80	3.15	3.03	20.18	3.78	4.53	0.90	10.96	9.20	3.33	12.08	0.00	50.00
Term	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	0.00

Table 4.2: Sample distance reference between different amino acids.

4.3 Dynamic Programming

Dynamic programming is an optimization method to solve multi-stage complex problems by breaking them down into simpler steps at different time points. It deals with the tradeoff between the desire for low present cost with the undesirability of high future costs [Bertsekas, 1995]. This method was first articulated in [Bellman, 1952] by Richard Bellman. Later, a central result of dynamic programming for discrete time systems, the recursive relationship between the value functions in two consecutive periods, together with its constraints, is named Bellman equation. The key step of this optimization method is to find the optimal control policy to compute the best possible value of the objective based on the knowledge given.

Our optimal problem, mathematically expressed as (4.2), is a multi-stage optimization problem and can be broken down into simpler steps by the measurement $y_k, 1 \leq k \leq N$. In addition, it satisfies the two principal features stated in [Bertsekas, 1995]: (1) an underlying discrete-time dynamic system, and (2) a cost function that is additive over time. Therefore, we apply dynamic programming to solve our optimal control problem.

The derivation in this section follows closely to the one of applying dynamic programming to solve basic model in [Bertsekas, 1995].

To simplify our expression, we rewrite our optimal control problem as

$$J_0(x_0) = \min_{u,c,v,c'} \mathop{\mathbb{E}}_{h,r,h',r'} \left[g_N(x_N) + \sum_{k=0}^{N-1} g(x_k, u_k, c_k, v_k, c'_k, h_k, r_k, h'_k, r'_k) \right], \tag{4.5}$$

subject to

$$x_{k+1} = f(x_k, u_k, c_k, v_k, c'_k, h_k, r_k, h'_k, r'_k), \qquad k = 0, 1, \cdots, N-1, \tag{4.6}$$

where

$$g_N(x_N) = d\left(x_N, \left\{x_N^d\right\}\right),$$

$$\begin{aligned} g(x_k, u_k, c_k, v_k, c'_k, h_k, r_k, h'_k, r'_k) = &\sum_{l_1=1}^{l}\sum_{i=1}^{n}\alpha_{l_1}u^i_{k,l_1} + \sum_{l_2=1}^{m}\sum_{i=1}^{n}\beta_{l_2}c^i_{k,l_2} + \sum_{l_3=1}^{l'}\sum_{i=1}^{n}\alpha'_{l_3}v^i_{k,l_3} \\ &+ \sum_{l_4=1}^{m'}\sum_{i=1}^{n}\beta'_{l_4}c'^i_{k,l_4} + d\left(x_k, \left\{x_k^d\right\}\right), \end{aligned}$$

and $f(x_k, u_k, c_k, v_k, c'_k, h_k, r_k, h'_k, r'_k)$ is the right hand side of (4.3a). The probability distributions associated with the random variables $h^i_{k,l_1}, r^i_{k,l_2}, h'^i_{k,l_3}, r'^i_{k,l}$ are $p^{(h)}_{l_1}(\cdot|x^i_k, u^i_{k,l_1})$, $p^{(r)}_{l_2}(\cdot|x^i_k, c^i_{k,l_2})$, $p^{(h')}_{l_3}(\cdot|x^i_k, v^i_{k,l_3})$ and $p^{(r')}_{l_4}(\cdot|x^i_k, c'^i_{k,l_4})$, respectively. The probability distribution only depends on the indices of mutagens l_1, l_2, l_3, l_4, but is irrelevant to step index k and spot index i.

Use $U_k(x_k)$ to denote the collection of all possible controls can be applied to the given x_k. Clearly, $U_k(x_k)$ is a subset of the control space, since the control space is the collection of all controls that can be applied to at least one state in the state space. This implies that for the given DNA segment at k^{th} instance, some mutagens cannot be applied to induce mutation at a specific spot, which is physically consistent with practical conditions because each mutagen corresponding to a specific transfer pattern by assumption.

In the dynamic programming algorithm, the crucial step to generate the optimal trajectory is to find the set of admissible control policies, and then pick the optimal from the set. Therefore, we first define the collection of admissible policies by

$$\pi = \{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$$

with $\mu_k(x_k) = \left\{\{u^{i_1}_{k,l_1}, c^{i_2}_{k,l_2}, v^{i_3}_{k,l_3}, c'^{i_4}_{k,l_4}\}|x_k, \forall i_1, i_2, i_3, i_4, l_1, l_2, l_3, l_4, 1 \le i \le n, 1 \le l_1 \le l, 1 \le l_2 \le m, 1 \le l_3 \le l', 1 \le l_4 \le m', 1 \le i_1, i_2, i_3, i_4 \le n\right\}$, a mapping from state space to control space, and $\mu_k(x_k) \in U_k(x_k), \forall x_k \in F^n$. The collection of all admissible policies is denoted by Π.

Rewriting our optimal control problem in (4.5) and (4.6), we get

$$J_0(x_0) = \min_{u,c,v,c'} \mathop{\mathbb{E}}_{h,r,h',r'} \left[g_N(x_N) + \sum_{k=0}^{N-1} g(x_k, \mu_k(x_k), h_k, r_k, h'_k, r'_k) \right], \tag{4.7}$$

subject to

$$x_{k+1} = f(x_k, \mu_k(x_k), h_k, r_k, h'_k, r'_k), \qquad k = 0, 1, \cdots, N-1. \tag{4.8}$$

Our goal is to find the $\pi^* = \{\mu_0^*, \mu_1^*, \cdots, \mu_{N-1}^*\}$, such that

$$J_0(x_0) = \mathop{\mathbb{E}}_{h,r,h',r'} \left[g_N(x_N) + \sum_{k=0}^{N-1} g(x_k, \mu_k^*(x_k), h_k, r_k, h'_k, r'_k) \right],$$

and π^* satisfies the constraints such that

$$x_{k+1} = f(x_k, \mu_k^*(x_k), h_k, r_k, h'_k, r'_k), \qquad k = 0, 1, \cdots, N-1.$$

The dynamic programming technique is based on the Bellman's Principle of Optimality, as stated below.

Principle of Optimality
An optimal policy has the the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision [Bellman, 2003].

Translated it mathematically. Let $\pi^* = \{\mu_0^*, \mu_1^*, \cdots, \mu_{N-1}^*\}$ be an optimal policy for the basic problem, and assume that when using π^*, a given state x_q occurs at time q with positive probability. Consider the subproblem whereby we are at x_q at time q and which to minimize the "cost-to-go" from time q to time N,

$$\mathop{\mathbb{E}}_{h,r,h',r'} \left\{ g_N(x_N) + \sum_{k=q}^{N-1} g(x_k, \mu_k(x_k), h_k, r_k, h'_k, r'_k) \right\}.$$

Then the truncated policy $\{\mu_q^*, \mu_{q+1}^*, \cdots, \mu_{N-1}^*\}$ is optimal for this subproblem [Bertsekas, 1995].

The principle of optimality implies that an optimal policy is constructed backward, starting from building up the subproblem involving the last period. And then the optimal policy is expended to the one involving last two periods. Continuing in the same manner, till the optimal policy for the entire problem is found. The dynamic programming algorithm is derived from this idea.

Define

$$J_q(x_q) = \min_{u,c,v,c'} \mathop{\mathbb{E}}_{h,r,h',r'} \left[g_N(x_N) + \sum_{k=q}^{N-1} g(x_k, u_k, c_k, v_k, c'_k, h_k, r_k, h'_k, r'_k) \right], \tag{4.9}$$

$$\begin{aligned} &= \min_{\{u,c,v,c'\}_{q,q+1,\cdots,N-1}} \left\{ \sum_{k=q}^{N-1} \sum_{l_1=1}^{l} \sum_{i=1}^{n} \alpha_{l_1} u_{k,l_1}^i + \sum_{k=q}^{N-1} \sum_{l_2=1}^{m} \sum_{i=1}^{n} \beta_{l_2} c_{k,l_2}^i \right. \\ &\quad + \sum_{k=q}^{N-1} \sum_{l_3=1}^{l'} \sum_{i=1}^{n} \alpha'_{l_3} v_{k,l_3}^i + \sum_{k=q}^{N-1} \sum_{l_4=1}^{m'} \sum_{i=1}^{n} \beta'_{l_4} c'^i_{k,l_4} \\ &\quad \left. + \sum_{k=q}^{N} \mathop{\mathbb{E}}_{\{h,r,h',r'\}_{q,q+1,\cdots,N-1}} \left[d\left(x_k, \left\{x_k^d\right\}\right) \right] \right\}, \end{aligned} \tag{4.10}$$

as the "tail problem" for any given $x_q, \forall q, 0 \le q \le N$. According the principle of optimality, if $\pi^* = \{\mu_0^*, \mu_1^*, \cdots, \mu_{N-1}^*\}$ is the optimal policy to J_0, then $\{\mu_q^*, \mu_{q+1}^*, \cdots, \mu_{N-1}^*\}$ is optimal for J_q.

Dynamic Programming Algorithm ([Bertsekas, 1995])
For every initial state x_0, the optimal cost $J_0(x_0)$ proceeds backward in time from period $N-1$ to period 0:

$$J_N(x_N) = g_N(x_N), \tag{4.11}$$

$$\begin{aligned} J_q(x_q) = \min_{\{u_q, c_q, v_q, c'_q\} \in U_q(x_q)} \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} &\left[g(x_q, u_q, c_q, v_q, c'_q, h_q, r_q, h'_q, r'_q) \right. \\ &\left. + J_{q+1}\left(f(x_q, u_q, c_q, v_q, c'_q, h_q, r_q, h'_q, r'_q) \right) \right], \\ q = 0, 1, \cdots, N-1, \end{aligned} \tag{4.12}$$

where the expectation is taken with respect to the probability distribution of h_k, r_k, h'_k, r'_k, which depends on state x_k and controls u_k, c_k, v_k, c'_k. Furthermore, if $\{u_k^*, c_k^*, v_k^*, c_k'^*\}$ $= \mu_k^*(x_k)$ minimizes the right side of (4.12) for each x_k and k, the policy $\pi^* = \{\mu_0^*, \mu_1^*, \cdots, \mu_{N-1}^*\}$ is optimal.

Proof of this algorithm can achieve the optimal can be found in §A.1.

Writing our optimal control problem explicitly in the form of (4.11) and (4.12), we get initialization and iterative equations for our generalized optimal control problem as

$$J_N(x_N) = d\left(x_N, \left\{x_N^d\right\}\right), \tag{4.13}$$

$$\begin{aligned} J_q(x_q) = & \min_{u_q, c_q, v_q, c'_q} \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} \Bigg[\sum_{l_1=1}^{l} \sum_{i=1}^{n} \alpha_{l_1} u_{q,l_1}^i + \sum_{l_2=1}^{m} \sum_{i=1}^{n} \beta_{l_2} c_{q,l_2}^i \\ & + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} \alpha'_{l_3} v_{q,l_3}^i + \sum_{l_4=1}^{m'} \sum_{i=1}^{n} \beta'_{l_4} c_{q,l_4}'^i + d\left(x_q, \left\{x_q^d\right\}\right) + J_{q+1}(x_{q+1}) \Bigg], \\ = & \min_{u_q, c_q, v_q, c'_q} \Bigg\{ \sum_{l_1=1}^{l} \sum_{i=1}^{n} \alpha_{l_1} u_{q,l_1}^i + \sum_{l_2=1}^{m} \sum_{i=1}^{n} \beta_{l_2} c_{q,l_2}^i + \sum_{l_3=1}^{l'} \sum_{i=1}^{n} \alpha'_{l_3} v_{q,l_3}^i \\ & + \sum_{l_4=1}^{m'} \sum_{i=1}^{n} \beta'_{l_4} c_{q,l_4}'^i + \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} \left[d\left(x_q, \left\{x_q^d\right\}\right) + J_{q+1}(x_{q+1}) \right] \Bigg\}, \\ & q = 0, 1, \cdots, N-1. \end{aligned} \tag{4.14}$$

4.4 Base-to-base, Deterministic Case

Similar to Chapter 3, we start with the simplest case, base-to-base deterministic mutation. Therefore, we do not need to evaluate the expected value. For one-dimensional optimal control problem, the desired set $\left\{x_k^d\right\}$ along the trajectory for every $k, 0 \leq k \leq N$, always contains a single element. Therefore, for simplicity, we use x_k^d instead of $\left\{x_k^d\right\}$ to denote the desired state at k^{th} stage. In addition, we require that $x_N = x_N^d$, and tracing cost along the trajectory $k, 0 \leq k \leq N-1$ is ignored. Since the distance reference defined in §4.2 cannot be applied in the single base mutations,

we define the distance reference between bases as

$$d\left(x_N, x_N^d\right) = \begin{cases} 0 & \text{if } x_N = x_N^d, \\ \infty & \text{if } x_N \neq x_N^d, \end{cases} \tag{4.15}$$

with $x_N, x_N^d \in \mathscr{F}$.

Therefore, the distance reference between different nucleotide bases is much larger than the costs of applying mutagens, and the corresponding term can be removed from our objective function. Instead, we add $x_N = x_N^d$ as another constraint to our single base deterministic optimal control problem.

We consider applying chemical mutagens only because the randomness caused by applying radiation is much larger and more difficult to control. We also ignore the mutations between a normal base and O as explained in §4.1, β_{l_2} is, in general, much larger than α_{l_1}, and more information is necessary to convert O to normal bases.

In sum, our optimal control problem for deterministic single base mutation can be formulated as

$$J_0(x_0) = \min_{u_{k,l_1}, 0 \leq k \leq N, 1 \leq l_1 \leq l} \left\{ \sum_{k=0}^{N-1} \sum_{l_1=1}^{l} \alpha_{l_1} u_{k,l_1} \right\}, \tag{4.16}$$

subject to

$$x_{k+1} = \left(-I + \sum_{l_1=1}^{l} u_{k,l_1} s \right) x_k, \tag{4.17a}$$

$$x_N = x_N^d, \tag{4.17b}$$

with x_0 given, $x_k \in \mathscr{F}_{\setminus\{0\}}$, where $\mathscr{F}_{\setminus\{0\}}$ denotes the set $\mathscr{F}$ excluding the element 0.

The definition of tail problem $J_q(x_q), q = 0, 1, \cdots, N-1$, and the update equation of control policy can be written as

$$J_q(x_q) = \min_{u_{k,l_1}, q \le k \le N-1, 1 \le l_1 \le l} \left\{ \sum_{k=q}^{N-1} \sum_{l_1=1}^{l} \alpha_{l_1} u_{q,l_1} \right\}, \tag{4.18}$$

$$= \min_{u_{q,l_1}, 1 \le l_1 \le l} \left\{ \sum_{l_1=1}^{l} \alpha_{l_1} u_{q,l_1} + J_{q+1} \left(\left(-I + \sum_{l_1=1}^{l} u_{q,l_1} s \right) x_q \right) \right\}. \tag{4.19}$$

Clearly, $\sum_{l_1=1}^{l} u_{q,l_1} = 0$ or 1, according to the Proposition 3.4.

If in addition, we assume that we have $l_1 = 12$ kinds of mutagens, each corresponding to a specific transfer pattern, as listed in Table 4.3. The corresponding controls and costs of applying a specific control are listed in Table 4.4 and 4.5.

Index (l_1)	1	2	3	4	5	6
Transfer Pattern	$A \to A$	$A \to G$	$A \to C$	$G \to A$	$G \to G$	$G \to T$
Index (l_1)	7	8	9	10	11	12
Transfer Pattern	$C \to A$	$C \to C$	$C \to T$	$T \to G$	$T \to C$	$T \to T$

Table 4.3: An example of chemical mutagens and their corresponding transfer patterns in deterministic mutations.

k^{th} \ $(k+1)^{th}$	A	G	C	T
A	u_{AA}	u_{AG}	u_{AC}	u_{AT}
G	u_{GA}	u_{GG}	u_{GC}	u_{GT}
C	u_{CA}	u_{CG}	u_{CC}	u_{CT}
T	u_{TA}	u_{TG}	u_{TC}	u_{TT}

Table 4.4: Controls corresponding to transfer between bases within one step. The leftmost column denotes the state k^{th} step, and the upmost row denotes the $(k+1)^{th}$ state.

k^{th} \ $(k+1)^{th}$	A	G	C	T
A	α_{AA}	α_{AG}	α_{AC}	α_{AT}
G	α_{GA}	α_{GG}	α_{GC}	α_{GT}
C	α_{CA}	α_{CG}	α_{CC}	α_{CT}
T	α_{TA}	α_{TG}	α_{TC}	α_{TT}

Table 4.5: Corresponding step cost of controls as shown in Table 4.4.

The elements along the anti-diagonal of Table 4.4, u_{AT}, u_{GC}, u_{CG} and u_{TA}, are artificially added, because transfers between complementary bases naturally happen and no mutagen is necessary. Therefore, the costs along the anti-diagonal of Table 4.5 are all zero, i.e. $\alpha_{AT} = \alpha_{GC} = \alpha_{CG} = \alpha_{TC} = 0$. We use nucleotide bases as subscripts in Table 4.4 and 4.5 because this representation is more straightforward. Otherwise, we can define a map $l_1 : \{A, T, G, C\} \times \{A, T, G, C\} \to \{\text{integers from 1 to 12}\}$, as listed in Table 4.3. Equivalence relationship between subscription in two nucleotide bases and in integer l_1 is defined by Table 4.3.

Under the above assumptions, we can rewrite (4.19) explicitly as

$$\begin{aligned} J_q(x_q) = \min_{u_{q,l_1}} \Big\{ & \alpha_{x_q A} + J_{q+1}(A), \alpha_{x_q G} + J_{q+1}(G), \\ & \alpha_{x_q C} + J_{q+1}(C), \alpha_{x_q T} + J_{q+1}(T) \Big\}, \qquad (4.20a) \\ = \min_{u_{q,l_1}} & \left\{ \alpha_{x_q \psi} + J_{q+1}(\psi), \forall \psi \in \{A, T, G, C\} \Leftrightarrow \mathscr{F}_{\setminus\{0\}} \right\}. \qquad (4.20b) \end{aligned}$$

Claim 4.1, 4.2, 4.3 and 4.4 are based on two conditions: (1) the optimal control problem and constraints follows (4.16) and (4.17), (2) all available mutagens (controls), the corresponding transfer patterns and costs are listed in Table 4.3, 4.4, and 4.5.

Claim 4.1. *For the same x_N^d, $J_q(\psi) \leq J_{q+1}(\overline{\psi}), \forall q, 0 \leq q \leq N-1, \forall \psi \in \{A, T, G, C\}$, where $\overline{\psi}$ denotes the complementary base of ψ in character-based notation, which is equivalent to $-\psi$ in numerical notation..*

Proof. This fact is due to the zero cost for the transfers between complementary bases in consecutive steps.

For any $0 \leq q \leq N-1$, the relationship between minimal costs in consecutive steps is shown in (4.20). Since $\psi \in \{A, T, G, C\}$, $\alpha_{\psi\overline{\psi}} + J_{q+1}(\overline{\psi})$ is one of the four elements in the set from which the $J_q(\psi)$ is picked. Moreover, $\alpha_{\psi\overline{\psi}} = 0$, therefore, $J_{q+1}(\overline{\psi})$ is one of the four elements in the set. Since $J_q(\psi)$ is the minimum picking for a set containing $J_{q+1}(\overline{\psi})$, we conclude that $J_q(\psi) \leq J_{q+1}(\overline{\psi})$. □

Claim 4.2. *$\forall N \geq 6$, $J_{N-6}(x_{N-6})$ is guaranteed to be the global optimal for every pair of $(x_{N-6}, x_N^d) \in \{A, T, G, C\} \times \{A, T, G, C\}$. This global optimal can be achieved by applying at most three different kinds of chemical mutagens, i.e. at most 3 u_{k,l_1}s ($1 \leq l_1 \leq 12$), $N-6 \leq k \leq N-1$, equal to 1, in at most 6 steps.*

If $1 \leq N \leq 5$, $\exists J_0(x_0)$ for every particular x_N^d.

Claim 4.2 can be proved by the brute force method as shown in §A.2. Since the costs of mutagens are represented by symbols, we cannot compare them numerically. However, we can eliminate all paths containing circular subunits to reduce the set where the minimum is picked from.

According to Claim 4.2, with N free, we are guaranteed to reach the global minimum within 6 steps from the initial state to the target state. For completely controllable systems, there always exists an M, the first time instance that the global minimum is achieved.

Proof. (**Existence of M**)
Since the system is completely controllable, $\exists k_1, k_2, \in \mathbb{Z} \cup \{0\}$, s.t. there exists at least one path from x_0 to x_{2k_1} and at least one path from x_0 to x_{2k_2+1} by applying proper mutagens in the correct order, with k_1, k_2 finite, $\forall x_0, x_{2k_1}, x_{2k_2+1} \in \{A, G, C, T\}$. Incorporating this fact into the optimal paths in our example, instead of completing each transfer within one steps, we use k_1 or k_2 steps transfers to substitute those one step transfers. Though the value of k_1, k_2 varies according to the initial and final state, the total steps needed to reach the global optimal is always finite. Therefore, M exists and is always finite. □

The existence of M implies that for without restriction on the number of steps, we can reach the global optimal in $N-M$ steps, 6 steps in our example. In addition, the

objective function remains optimal for $x_q = x_M = \psi$ and $x_N^d, \forall q, 0 \leq q \leq M, M-q \equiv 0 \pmod 2$, and $x_q = \overline{x_M} = \overline{\psi}, \forall q, 0 \leq q \leq M, M-q \equiv 1 \pmod 2, \psi \in \{A, T, G, C\}$, which leads to Claim 4.3.

Claim 4.3. *If the system is completely controllable, $\exists M$, the first instance that the global minimum is reached, $J_M(\psi)$ is the global minimum and $\forall q, 0 \leq q \leq M, J_q(\psi) = J_M(\overline{\psi})$ if $M-q \equiv 1 \pmod 2$, and $J_q(\psi) = J_M(\psi)$ if $M-q \equiv 0 \pmod 2$. In addition, $\forall q, 2 \leq q \leq M$,*

$$J_q(\psi) = J_{q-2}(\psi) = J_{q-1}(\overline{\psi}), \psi \in \{A, T, G, C\} \Leftrightarrow \{1, -1, 2, -2\} = \mathscr{F}_{\setminus\{0\}},$$

with the same x_N^d. In our example, $M \geq N - 6$.

Proof. For $q = M$, $J_M(\psi)$ is the global minimum. From Claim 4.1, $J_{M-1}(\overline{\psi}) \leq J_M(\psi)$, therefore, $J_{M-1}(\overline{\psi}) = J_M(\psi)$ for the same x_N^d. Therefore, $J_{M-1}(\overline{\psi})$ is also a global minimum.

By backward induction, suppose for $q = q_1$, the statement is true, i.e. $J_{q_1-1}(\overline{\psi}) = J_{q_1}(\psi)$ is the global optimal either from $x_{q_1-1} = \overline{\psi}$ or $x_{q_1} = \psi$ to x_N^d. Obviously, for $q = q_1 - 1$, the statement still true. Therefore, $J_q(\psi) = J_{q-2}(\psi) = J_{q-1}(\overline{\psi}), \psi \in \{A, T, G, C\}, \forall q, 2 \leq q \leq M$. □

In the proof of global minimum can be reached in the finite step in Claim 4.2, we also discover Proposition 3.2. Here, $J_q(x_q, x_d^N)$ denotes the optimal cost from x_q to x_d^N.

Claim 4.4. *Given two single base mutation optimal control problems, with the same fixed N, and the desired final states complementary to each other. If $J_M(\psi, x_N^d)$ is the global minimum, then $J_M(\overline{\psi}, \overline{x_N^d})$ is also the global minimum, i.e. the global minimum of both systems is reach at the same stage M. Moreover, $\forall q, 0 \leq q \leq M$,*

$$J_q(\psi, x_N^d) = J_q(\overline{\psi}, \overline{x_N^d}), \psi, x_N^d \in \{A, T, G, C\}.$$

Physically, Claim 4.4 states that the optimal can be achieve at the same step from a pair of complementary bases to another pair of complementary bases at the same cost. It is implicitly shown in the proof of Claim 4.2 in §A.2. However, this fact is

true only for base-to-base deterministic mutations, because the distance reference is well-defined in (4.15).

Now, we show an example with simulation results.

The costs of applying different mutagens are listed in Table 4.6. It is a numerical assignment to Table 4.5. Since we apply mutagens before every replication starts, u_{AA} actually transfer A to T and then to A by replication. For simplicity, we just use the k^{th} and $(k+1)^{th}$ step states as subscripts to represent the corresponding control and cost. The costs of transitions is lower than the costs of transversions. Therefore, $\alpha_{AC}, \alpha_{CA}, \alpha_{GT}, \alpha_{TG}$ have smaller values than other mutagens, except artificial ones.

x_k \ x_{k+1}	A	G	C	T
A	5.21	6.60	2.33	0
G	6.15	8.95	0	3.82
C	4.61	0	9.17	7.24
T	0	0.64	5.09	10.28

Table 4.6: Sample step costs.

If we use χ to denote the costs of mutagens as listed in Table 4.6, then

$$\chi = \begin{bmatrix} 5.21 & 6.60 & 2.33 & 0 \\ 6.15 & 8.95 & 0 & 3.82 \\ 4.61 & 0 & 9.17 & 7.24 \\ 0 & 0.64 & 5.09 & 10.28 \end{bmatrix} = \begin{bmatrix} \alpha_{AA} & \alpha_{AG} & \alpha_{AC} & \alpha_{AT} \\ \alpha_{GA} & \alpha_{GG} & \alpha_{GC} & \alpha_{GT} \\ \alpha_{CA} & \alpha_{CG} & \alpha_{CC} & \alpha_{CT} \\ \alpha_{TA} & \alpha_{TG} & \alpha_{TC} & \alpha_{TT} \end{bmatrix} \Leftrightarrow \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & 0 \\ \alpha_4 & \alpha_5 & 0 & \alpha_6 \\ \alpha_7 & 0 & \alpha_8 & \alpha_9 \\ 0 & \alpha_{10} & \alpha_{11} & \alpha_{12} \end{bmatrix}.$$

Here, we use $J_q(x_q, x_N^d)$ to denote the optimal cost from x_q to x_N^d as we did in Claim 4.4. Then

$$J_q = \begin{bmatrix} J_q(A,A) & J_q(A,G) & J_q(A,C) & J_q(A,T) \\ J_q(G,A) & J_q(G,G) & J_q(G,C) & J_q(G,T) \\ J_q(C,A) & J_q(C,G) & J_q(C,C) & J_q(C,T) \\ J_q(T,A) & J_q(T,G) & J_q(C,C) & J_q(T,T) \end{bmatrix}.$$

The path to reach the optimal cost is denoted by

$$P_q = \begin{bmatrix} P_q(A,A) & P_q(A,G) & P_q(A,C) & P_q(A,T) \\ P_q(G,A) & P_q(G,G) & P_q(G,C) & P_q(G,T) \\ P_q(C,A) & P_q(C,G) & P_q(C,C) & P_q(C,T) \\ P_q(T,A) & P_q(T,G) & P_q(T,C) & P_q(T,T) \end{bmatrix},$$

where $P_q(x_q, x_N^d)$ is the $(q+1)^{th}$ state from x_q to x_N^d, i.e. $x_{q+1} = P_q(x_q, x_N^d)$.

We run the dynamic programming algorithm for every pair of $(x_q, x_N^d) \in \{A,T,G,C\} \times \{A,T,G,C\}$, $N = 9$. The simulation results are shown as below, including optimal costs for all possible transfer pairs $(x_q, x_N^d) \in \{A,T,G,C\} \times \{A,T,G,C\}$, $J_q, 0 \leq q \leq 8$, graphical representation in Figure 4.1, and optimal path $P_q, 0 \leq q \leq 7$.

$$J_0 = \begin{bmatrix} 5.21 & 5.09 & 0.64 & 0 \\ 6.15 & 6.79 & 0 & 3.82 \\ 3.82 & 0 & 6.79 & 6.15 \\ 0 & 0.64 & 5.09 & 5.21 \end{bmatrix} \qquad J_1 = \begin{bmatrix} 0 & 0.64 & 5.09 & 5.21 \\ 3.82 & 0 & 6.79 & 6.15 \\ 6.15 & 6.79 & 0 & 3.82 \\ 5.21 & 5.09 & 0.64 & 0 \end{bmatrix}$$

$$J_2 = \begin{bmatrix} 5.21 & 5.09 & 0.64 & 0 \\ 6.15 & 6.79 & 0 & 3.82 \\ 3.82 & 0 & 6.79 & 6.15 \\ 0 & 0.64 & 5.09 & 5.21 \end{bmatrix} \qquad J_3 = \begin{bmatrix} 0 & 0.64 & 5.09 & 5.21 \\ 3.82 & 0 & 6.79 & 6.15 \\ 6.15 & 6.79 & 0 & 3.82 \\ 5.21 & 5.09 & 0.64 & 0 \end{bmatrix}$$

$$J_4 = \begin{bmatrix} 5.21 & 5.09 & 0.64 & 0 \\ 6.15 & 6.79 & 0 & 3.82 \\ 3.82 & 0 & 6.79 & 6.15 \\ 0 & 0.64 & 5.09 & 5.21 \end{bmatrix} \qquad J_5 = \begin{bmatrix} 0 & 0.64 & 5.09 & 5.21 \\ 3.82 & 0 & 6.79 & 6.15 \\ 6.15 & 6.79 & 0 & 3.82 \\ 5.21 & 5.09 & 0.64 & 0 \end{bmatrix}$$

$$J_6 = \begin{bmatrix} 5.21 & 5.09 & 0.64 & 0 \\ 6.15 & 6.79 & 0 & 3.82 \\ 3.82 & 0 & 7.88 & 6.15 \\ 0 & 0.64 & 5.09 & 5.21 \end{bmatrix} \qquad J_7 = \begin{bmatrix} 0 & 0.64 & 5.09 & 5.21 \\ 3.82 & 0 & 8.48 & 6.15 \\ 6.15 & 7.88 & 0 & 3.82 \\ 5.21 & 5.09 & 0.64 & 0 \end{bmatrix}$$

$$J_8 = \begin{bmatrix} 5.21 & 6.60 & 2.33 & 0 \\ 6.15 & 8.95 & 0 & 3.82 \\ 4.61 & 0 & 9.17 & 7.24 \\ 0 & 0.64 & 5.09 & 10.28 \end{bmatrix}$$

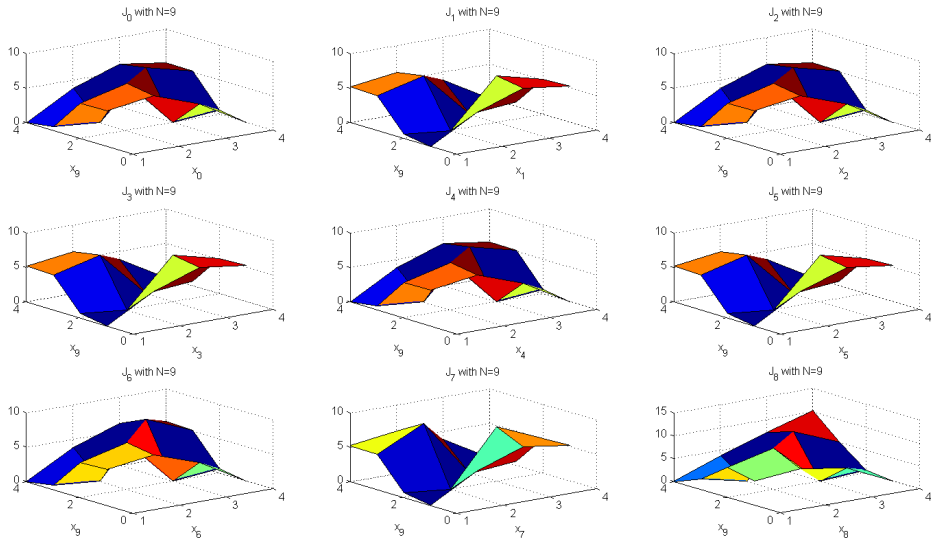

Figure 4.1: Graphical representation of $J_q, 0 \leq q \leq 8, N = 9$, in single base deterministic mutation example. The x-axis and y-axis represent x_q and x_N^d, respectively. $J_q(x_q, x_N^d)$ are represented by 16 isolated points. Those discrete points are connected together to show the surface.

$$P_0 = \begin{bmatrix} A,T & T & T & T \\ A,C & A,C & C & C,T \\ G & G & G & G \\ A & A,G & A,C & A \end{bmatrix} \qquad P_1 = \begin{bmatrix} T & T & T & A,T \\ C,T & C & A,C & A,C \\ G & G & G & G \\ A & A,C & A,G & A \end{bmatrix}$$

$$P_2 = \begin{bmatrix} A,T & T & T & T \\ A,C & A,C & C & C,T \\ G & G & G & G \\ A & A,G & A,C & A \end{bmatrix} \qquad P_3 = \begin{bmatrix} T & T & T & A,T \\ C,T & C & A,C & A,C \\ G & G & G & G \\ A & A,C & A,G & A \end{bmatrix}$$

$$P_4 = \begin{bmatrix} A,T & T & T & T \\ A,C & A,C & C & C,T \\ G & G & G & G \\ A & A,G & A,C & A \end{bmatrix} \qquad P_5 = \begin{bmatrix} T & T & T & A,T \\ C,T & C & A & A,C \\ G & G & G & G \\ A & A,C & A,G & A \end{bmatrix}$$

$$P_6 = \begin{bmatrix} A,T & T & T & T \\ A,C & A & C & C,T \\ G & G & T & G \\ A & A,G & A,C & A \end{bmatrix} \qquad P_7 = \begin{bmatrix} T & T & T & A \\ T & C & A & A \\ G & T & G & G \\ A & C & G & A \end{bmatrix}$$

For simplicity, we use 1 to represent A, 2 to G, 3 to C and 4 to T in graphical interpretation. From Figure 4.1, we can see clearly that the optimal cost decreases as q decreases in the first few steps, and then optimal cost remains at the global minimum. This phenomena obeys Claim 4.1 and 4.2. In this example, global optimal is reach at $M = 6$ for all pairs of initial and final states as $J_7 \neq J_5 = J_3$ and $J_8 \neq J_6 = J_4$. So the global minimum is achieved before we reach $N - 6 = 3$ in this particular case. This also implies that with N free, we can reach the desired final state in 3 steps from the given initial state.

Observing $J_q, 0 \leq q \leq 6$, we find that J_{q-1} equals to J_q by exchanging the first and the last column, and the second and the third column. Or we can exchange the first and the last row, and the second and the third row of J_q to obtain J_{q-1}. J_{q_1} and J_{q_2} are the same for $q_1, q_2 \leq M = 6$ for $q_1 - q_2 = 0 \pmod 2$. This obeys Claim 4.3 and 4.4.

The optimal trajectories are generated from $P_q(x_q, x_N^d)$. For example, given $x_2 = T$, and the final state $x_9^d = G$, we want to generate the optimal trajectories.

$x_3 = P_2(T, G) = A, G.$
If $x_3 = A$, $x_4 = P_3(A, G) = T$; if $x_3 = G$, $x_4 = P_3(A, G) = C$.
If $x_4 = T$, $x_5 = P_4(T, G) = A, G$; if $x_4 = C$, $x_5 = P_4(C, G) = G$.
If $x_5 = A$, $x_6 = P_5(A, G) = T$; if $x_5 = G$, $x_6 = P_5(G, G) = C$.
If $x_6 = T$, $x_7 = P_6(T, G) = A, G$; if $x_6 = C$, $x_7 = P_6(C, G) = G$.
If $x_7 = A$, $x_8 = P_7(A, G) = T$; if $x_7 = G$, $x_8 = P_7(G, G) = C$.

So the optimal routes are

$$T \to A \to T \to A \to T \to A \to T \xrightarrow[\alpha_{TG}]{u_{TG}} G$$
$$T \to A \to T \to A \to T \xrightarrow[\alpha_{TG}]{u_{TG}} G \to C \to G$$
$$T \to A \to T \xrightarrow[\alpha_{TG}]{u_{TG}} G \to C \to G \to C \to G$$
$$T \xrightarrow[\alpha_{TG}]{u_{TG}} G \to C \to G \to C \to G \to C \to G$$

Consequently, the optimal cost is $J_2(T, G) = 0.64 = \alpha_{TG}$.

Optimal trajectories for other pairs of initial and final states can be obtained in the same manner.

4.5 Codon-to-Codon, Deterministic Case

Similar to single base deterministic case, we ignore transfers from or to the non-sense base O, and apply chemical mutagens only. In addition, we ignore the tracing costs along the trajectory, but keep the final penalty.

Hence, for codon-to-codon deterministic mutations, we formulate our optimal control problem as

$$J_0(x_0) = \min_{\substack{u^i_{k,l_1}, 0\le k\le N,\\ 1\le l_1\le l, 1\le i\le 3}} \left\{ \sum_{k=0}^{N-1} \sum_{l_1=1}^{l} \sum_{i=1}^{3} \alpha_{l_1} u^i_{k,l_1} + d\left(x_N, \left\{x_N^d\right\}\right) \right\}, \tag{4.21}$$

subject to

$$x_{k+1} = \left(-I + \sum_{l_1=1}^{l} \sum_{i=1}^{3} u^i_{k,l_1} s e_i e_i^T \right) x_k, \tag{4.22}$$

with $x_0, x_N^d \in \mathscr{F}^3_{\setminus\{0\}}$ given, $x_k \in \mathscr{F}^3_{\setminus\{0\}}, \forall k, 0 \le k \le N$, and $d(\varphi_1, \varphi_2), \varphi_1, \varphi_2 \in \mathscr{F}^3_{\setminus\{0\}}$ as defined in §4.2.

The definition of tail problem $J_q(x_q), q = 0, 1, \cdots, N-1$, and the update equation of control policy can be written as

$$J_q(x_q) = \min_{\substack{u^i_{k,l_1}, q\le k\le N,\\ 1\le l_1\le l, 1\le i\le 3}} \left\{ \sum_{k=q}^{N-1} \sum_{l_1=1}^{l} \sum_{i=1}^{3} \alpha_{l_1} u^i_{k,l_1} + d\left(x_N, \left\{x_N^d\right\}\right) \right\}, \tag{4.23}$$

$$= \min_{u^i_{q,l_1}, 1\le l_1\le l, 1\le i\le 3} \left\{ \sum_{l_1=1}^{l} \sum_{i=1}^{3} \alpha_{l_1} u^i_{q,l_1} + J_{q+1}(x_{q+1}) \right\}, \tag{4.24}$$

with $x_q, x_N^d \in \mathscr{F}^3_{\setminus\{0\}}$ and $u^i_{q,l_1} \in \{0,1\}, \forall q, l_1, i, 0 \le q \le N-1, 1 \le l_1 \le l, 1 \le i \le 3$.

If in addition, we assume all available mutagens, their corresponding transfer pair and applying cost are as listed in Table 4.3, 4.4 and 4.5. Then we can rewrite (4.24) as

$$J_q(x_q) = \min_{u^i_{q,l_1}, 1\le l_1\le l, 1\le i\le 3} \left\{ \alpha_{x^1_q\psi_1} + J_{q+1}\left(\begin{bmatrix}\psi_1\\ \overline{x^2_q}\\ \overline{x^3_q}\end{bmatrix}\right), \alpha_{x^2_q\psi_2} + J_{q+1}\left(\begin{bmatrix}\overline{x^1_q}\\ \psi_2\\ \overline{x^3_q}\end{bmatrix}\right), \right.$$
$$\left. \alpha_{x^3_q\psi_3} + J_{q+1}\left(\begin{bmatrix}\overline{x^1_q}\\ \overline{x^2_q}\\ \psi_3\end{bmatrix}\right), \psi_1, \psi_2, \psi_3 \in \{A, T, G, C\} \Leftrightarrow \mathscr{F}_{\backslash\{0\}} \right\}, \quad (4.25)$$

with $x^i_q \in \{A, T, G, C\} \Leftrightarrow \mathscr{F}_{\backslash\{0\}}, 1 \le i \le 3$ denotes the i^{th} element of $x_q \in \mathscr{F}^3_{\backslash\{0\}}$, and $\overline{x^i_q}$ denotes the complementary base of x^i_q.

The optimal control sequences depends on the numerical values of α_{l_1}s and $d(\varphi_1, \varphi_2)$, $\varphi_1, \varphi_2 \in \mathscr{F}^3_{\backslash\{0\}}$. Although the values we assigned to α_{l_1}s and $d(\varphi_1, \varphi_2)$ are not actual practical values, we can always draw conclusions from simulation results by assigning different sets of numerical values to those parameters.

Below are the simulation results of three different assignments of α_{l_1}s, $\chi, 5\chi$, and 0.5χ, and the same $d(\varphi_1, \varphi_2)$. Here, χ the same as in §4.4.

The graphical interpretation of three assignments are shown in Figure 4.2, Figure 4.3 and Figure 4.4, respectively. The x-axis and y-axis denote x_q and x^d_N, respectively. For a codon $\begin{bmatrix}\psi_1 & \psi_2 & \psi_3\end{bmatrix}^T, \psi_1, \psi_2, \psi_3 \in \{A, T, G, C\} \Leftrightarrow \mathscr{F}_{\backslash\{0\}}$, its index is calculated by

$$4^2(\psi_1 - 1) + 4(\psi_2 - 1) + \psi_3,$$

where $\psi_i = 1$ if A, $\psi_i = 2$ if G, $\psi_i = 3$ if C and $\psi_i = 4$ if T, $1 \le i \le 3$, for the simplicity of graphical interpretation. Since a codon has $4^3 = 64$ permutations, there are 64^2 pairs of initial and final desired states, and there are 64×21 pairs of initial state and final desired set. The surface is generated by connecting 64×64 discrete points together. J_q is calculated following the same procedure as in base-to-base deterministic cases. The value of optimal cost can be read directly from graphical interpretation, and the optimal path can be generated from path matrix P_q, similar

to base-to-base deterministic case. Both J_q and $P_q, \forall q, 0 \leq q \leq N$ are of 64×64 dimension.

N	α_{l_1}	$d(\varphi_1, \varphi_2)$	First q when the global minimum is reached	Global minimum
19	0.5χ	Table 4.2	$q = 12$	$J_{12}(x_{12})$
19	χ	Table 4.2	$q = 13$	$J_{13}(x_{13})$
19	5χ	Table 4.2	$q = 15$	$J_{15}(x_{15})$

Table 4.7: Simulation results with different α_{l_1} assignments and the first q where the global optimal is reached.

From the graphical interpretation and Table 4.7, we find that the value of q where the global minimum is reach at the first time decreases as α_{l_1} decreases. And the surface generated by J_0 is more similar to the one generated J_{18} with $\alpha = 5\chi$ than with $\alpha = \chi$ or $\alpha = 0.5\chi$. This implies that if $d(\varphi_1, \varphi_2)$ is the deterministic term in our objective function, then the treatment plan is made to drive the final state as close to the desired set as possible; if the costs of applying mutagens is the deterministic term in the objective function, then the treatment plan tends to stay in the original state and applying less mutagens; if they are of equal weight, then the treatment plan deals with this tradeoff.

Moreover, no matter how the numerical values of final penalty and the costs of applying mutagens change in our objective function, there is always a $M, M \leq N - 18$, $J_M(x_M)$ is global minimum. This property is stated in Claim 4.5.

Claim 4.5. *Given an optimal control problem with objective function in the form of* (4.21)*, constraints in the form of* (4.22)*, and all available chemical mutagens, their corresponding transfer pairs and costs as listed in Table 4.3, 4.4 and 4.5, then* $J_{N-18}(x_{N-18})$ *is guaranteed to be the global optimal with fixed* $\left\{x_N^d\right\}$ *and* $N \geq 18$*.*

Proof. The objective function in (4.21) can be written as a summation of three separate single base mutation systems, and the distance reference between the final state

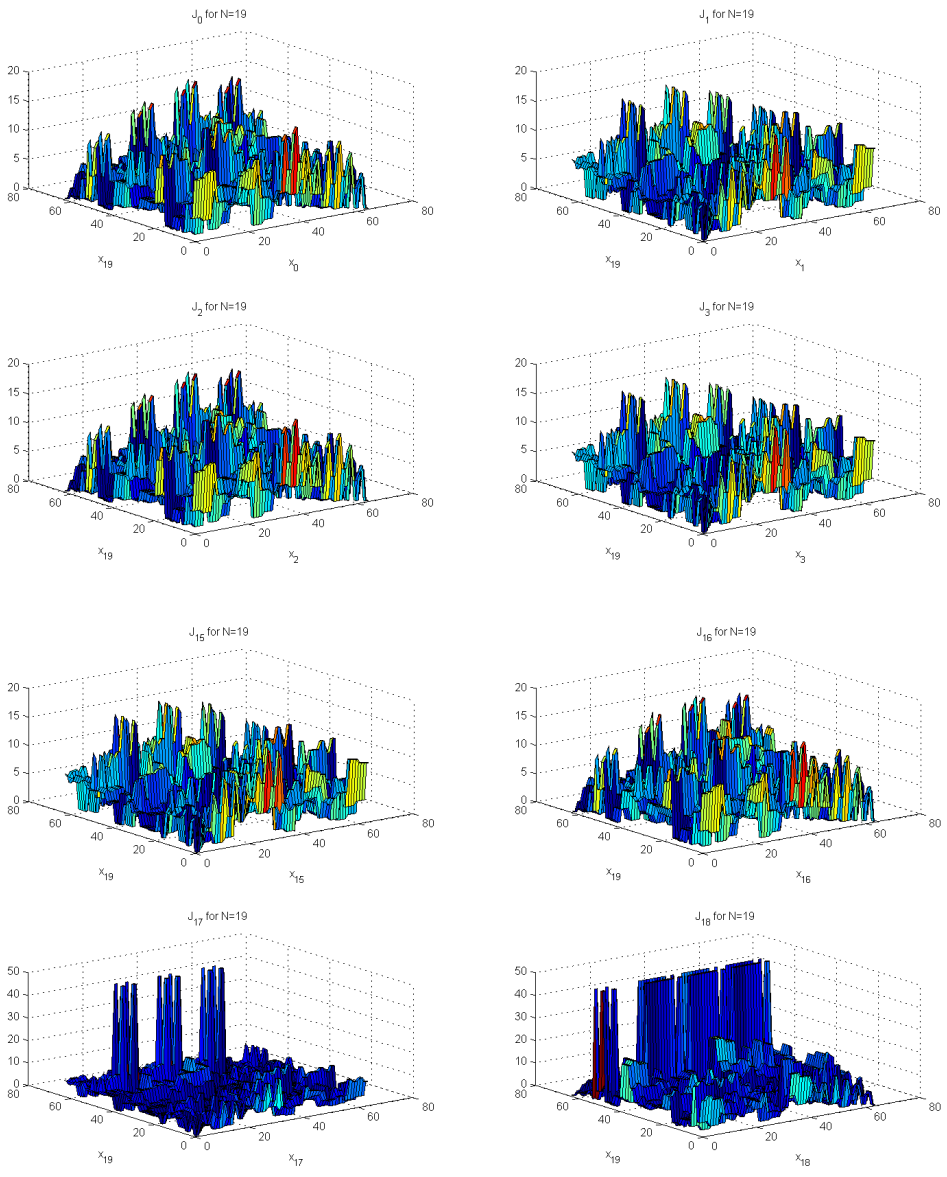

Figure 4.2: Graphically representation of $J_q, q = 0, 1, 2, 3, 15, 16, 17, 18$ for codon-to-codon deterministic mutations, with $\alpha_{l_1} = \chi$, $d(\cdot, \cdot)$ as listed in Table 4.2, $N = 19$.

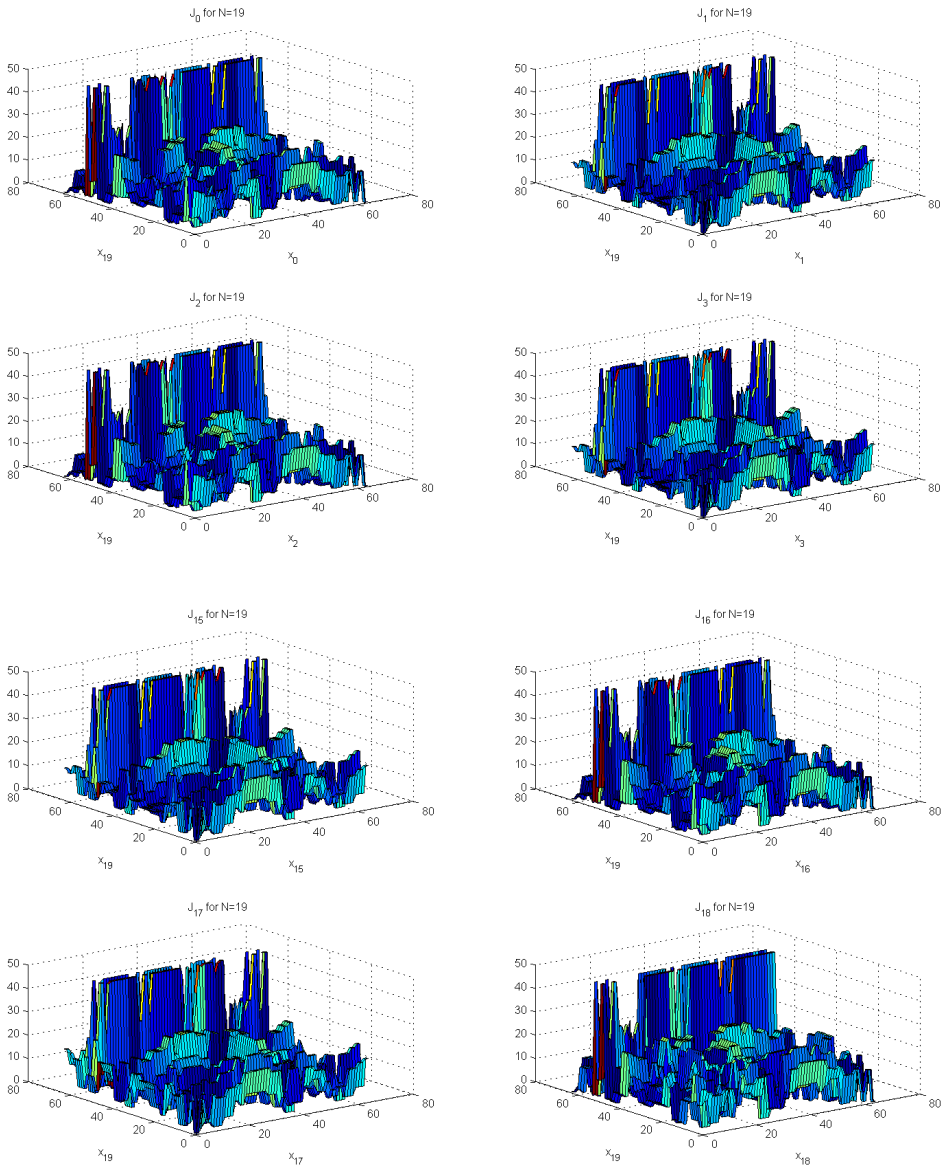

Figure 4.3: Graphically representation of $J_q, q = 0, 1, 2, 3, 15, 16, 17, 18$ for codon-to-codon deterministic mutations, with $\alpha_{l_1} = 5\chi$, $d(\cdot, \cdot)$ as listed in Table 4.2, $N = 19$.

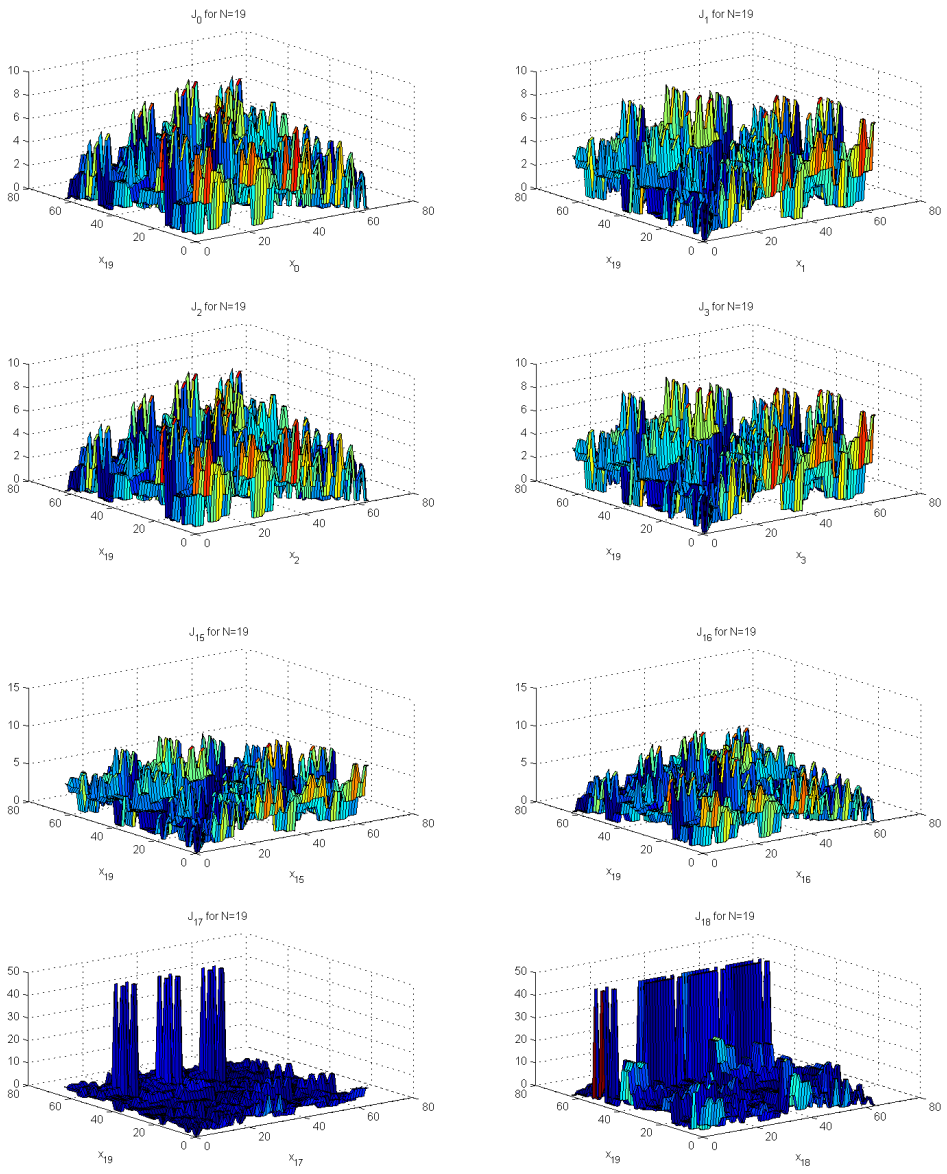

Figure 4.4: Graphically representation of $J_q, q = 0, 1, 2, 3, 15, 16, 17, 18$ for codon-to-codon deterministic mutations, with $\alpha_{l_1} = 0.5\chi$, $d(\cdot, \cdot)$ as listed in Table 4.2, $N = 19$.

and final desired set, i.e.

$$J_q(x_q) = \min_{\substack{n_1,n_2,n_3\geq 0 \\ 2N\leq n_1+n_2+n_3\leq 3N-1}} \Bigg\{ \underbrace{J_{n_1}(x_{n-1})\left(x_q^1,\psi_1\right)}_{\substack{\text{optimal costs of base-to-base} \\ \text{deterministic optimal control} \\ \text{problem formed by the } 1^{st} \text{ base}}} + \underbrace{J_{n_2}(x_{n-2})\left(x_q^2,\psi_2\right)}_{\substack{\text{optimal costs of base-to-base} \\ \text{deterministic optimal control} \\ \text{problem formed by the } 2^{nd} \\ \text{base}}}$$

$$+ \underbrace{J_{n_3}(x_{n-3})\left(x_q^3,\psi_3\right)}_{\substack{\text{optimal costs of base-to-base} \\ \text{deterministic optimal control} \\ \text{problem formed by the } 3^{rd} \\ \text{base}}} + d\left(\begin{bmatrix}\psi_1\\ \psi_2\\ \psi_3\end{bmatrix}, \left\{x_N^d\right\}\right), \Bigg\} \tag{4.26}$$

where $N-q=(N-n_1)+(N-n_2)+(N-n_3)$.

According to Claim 4.2, $J_{N-6}(x_{N-6})$ is guaranteed to be the global optimal for single base mutations. Therefore, optimal costs corresponding to three single base mutation systems, $J_{n_1}(x_{n-1})\left(x_q^1,\psi_1\right)$, $J_{n_2}(x_{n-2})\left(x_q^2,\psi_2\right)$, $J_{n_3}(x_{n-3})\left(x_q^3,\psi_3\right)$ is guaranteed to reach their own global optimal at $n_1=n_2=n_3=N-6$ with all possible combinations of $\psi_1,\psi_2,\psi_3\in\{A,T,G,C\}$. Therefore, $q=N-18$ is a guaranteed global optimal. □

Claim 4.5 is a three-dimensional extension of Claim 4.2. We choose $N=19$ in our example based on Claim 4.5. Indeed, Claim 4.5 is a quite loose condition. The q values where the first global optimal is reached at $M\geq N-18=1$ with different parameter assignments in our example are listed in Table 4.7. Moreover, if $J_M(x_M)$ is a global optimal from x_M to $\{x_N^d\}$, then it is also a global optimal from x_M to any final state in the set of $\{x_N^d\}$ since the final desired set generated by every element from $\{x_N^d\}$ is $\{x_N^d\}$.

We can also extend Claim 4.3 to codon-to-codon deterministic cases.

Claim 4.6. *Similar to base-to-base deterministic mutations, after the global minimum is reached at $J_M(x_M)$ for fixed x_N^d, $\forall q, 2 \leq q \leq M$,*

$$J_q\left(\begin{bmatrix}\psi_1\\ \psi_2\\ \psi_3\end{bmatrix}\right) = J_{q-1}\left(\begin{bmatrix}\overline{\psi_1}\\ \overline{\psi_2}\\ \overline{\psi_3}\end{bmatrix}\right) = J_{q-2}\left(\begin{bmatrix}\psi_1\\ \psi_2\\ \psi_3\end{bmatrix}\right).$$

Therefore, $J_q\left(\begin{bmatrix}\psi_1 & \psi_2 & \psi_3\end{bmatrix}^T\right)$ is the global minimum if $M - q \equiv 0 \pmod 2$, from $x_q = x_M = \begin{bmatrix}\psi_1 & \psi_2 & \psi_3\end{bmatrix}^T$ to $\left\{x_N^d\right\}$ with $x_N^d \in \mathscr{F}^3_{\setminus\{0\}}$; $J_q\left(\begin{bmatrix}\overline{\psi_1} & \overline{\psi_2} & \overline{\psi_3}\end{bmatrix}^T\right)$ is the global minimum if $M - q \equiv 1 \pmod 2$, from $x_q = \overline{x_M} = \begin{bmatrix}\overline{\psi_1} & \overline{\psi_2} & \overline{\psi_3}\end{bmatrix}^T$ to the same final desired set $\left\{x_N^d\right\}$.

The proof of Claim 4.6 is similar to the one of Claim 4.3.

Graphically, the indices of complementary codons, $\begin{bmatrix}\psi_1 & \psi_2 & \psi_3\end{bmatrix}^T$ and $\begin{bmatrix}\overline{\psi_1} & \overline{\psi_2} & \overline{\psi_3}\end{bmatrix}^T$, sum up to 65, i.e.

$$\begin{aligned}
&\big(16(\psi_1 - 1) + 4(\psi_2 - 1) + \psi_3\big) + \Big(16(\overline{\psi_1} - 1) + 4(\overline{\psi_2} - 1) + \overline{\psi_3}\Big)\\
&= \big(16(\psi_1 - 1) + 4(\psi_2 - 1) + \psi_3\big) + \big(16((5 - \psi_1) - 1) + 4((5 - \psi_2) - 1) + (5 - \psi_3)\big),\\
&= 65.
\end{aligned}$$

Therefore, J_q and J_{q-1}, $1 \leq q \leq M$, are symmetric about the plane $x = 32.5$, J_{q-2} and J_q, $2 \leq q \leq M$, are the same, as shown in Figure 4.2, Figure 4.3 and Figure 4.4.

However, as mentioned in §4.2, Claim 4.4 cannot be extended to codon-to-codon deterministic case due to the redundancy of genetic codes, i.e. the set of codons translated to the same amino acid varies from one amino acid to another as shown in Table 4.1. The simulation results show that the costs a pair of complementary codons to two final desired set generated by a pair of complementary final desired codons are different, i.e. $J_q\left(x_q, \left\{x_N^d\right\}\right) \neq J_q\left(\overline{x_q}, \left\{\overline{x_N^d}\right\}\right)$, in general, for any q. Graphically, the optimal cost profile J_q is not symmetric about the plane $y = 32.5$ for J_{q-1}, $1 \leq q \leq M$. Therefore, the doctors need to make treatment plans for both strands and choose the

one with lower cost. This also implies that in large scales cases, for instance, a gene contains hundreds of nucleotide bases, the doctors should make the treatment plan according to the strand with lower optimal cost than the other.

4.6 Codon-to-Codon, Stochastic Case

In this section, we formulate optimal control problem for codon-to-codon stochastic mutations, derive update equation by the dynamic programming algorithm, and demonstrate simulation results with different assignments of parameters.

Again, we take the same assumptions as we did for base-to-base and codon-to-codon deterministic cases. Assume mutagens with possible transfer patterns involving the non-sense base O, and radiation are unavailable. Also, we only keep the final penalty cost and ignore the tracing cost along the path.

The optimal control problem of codon-to-codon stochastic mutations can be written as

$$J_0(x_0) = \min_{\substack{u^i_{k,l_1},0\le k\le N-1\\ 1\le l_1\le l,1\le i\le 3}} \left\{ \sum_{k=0}^{N-1}\sum_{l_1=1}^{l}\sum_{i=1}^{3} \alpha_{l_1} u^i_{k,l_1} + \mathop{\mathbb{E}}_{\substack{h^i_{k,l_1},0\le k\le N-1\\ 1\le l_1\le l,1\le i\le 3}} \left[d\left(x_N, \left\{ x^d_N \right\} \right) \right] \right\}, \tag{4.27}$$

subject to

$$x_{k+1} = -Ix_k + \sum_{l_1=1}^{l}\sum_{i=1}^{3} u^i_{k,l_1} h^i_{k,l_1} s e_i e_i^T x_k, \tag{4.28}$$

with $x_0, x^d_N \in \mathscr{F}^3_{\setminus\{0\}}$ given, $x_k \in \mathscr{F}^3_{\setminus\{0\}}$.

The definition of tail problem $J_q(x_q), q = 0, 1, \cdots, N-1$, and the update equation of control policy can be written as

$$J_q(x_q) = \min_{\substack{u^i_{k,l_1}, q\le k\le N-1 \\ 1\le l_1\le l, 1\le i\le 3}} \left\{ \sum_{k=q}^{N-1}\sum_{l_1=1}^{l}\sum_{i=1}^{3} \alpha_{l_1} u^i_{k,l_1} + \mathop{\mathbb{E}}_{\substack{h^i_{k,l_1}, q\le k\le N-1 \\ 1\le l_1\le l, 1\le i\le 3}} \left[d\left(x_N, \left\{ x^d_N \right\} \right) \right] \right\}, \quad (4.29)$$

$$= \min_{u^i_{q,l_1}, 1\le l_1\le l, 1\le i\le 3} \left\{ \sum_{l_1=1}^{l}\sum_{i=1}^{3} \alpha_{l_1} u^i_{q,l_1} + \mathop{\mathbb{E}}_{h^i_{q,l_1}, 1\le l_1\le l, 1\le i\le 3} \left[J_{q+1}(x_{q+1}) \right] \right\}. \quad (4.30)$$

with $x_q, x^d_N \in \mathscr{F}^3_{\setminus\{0\}}$ given, and $u^i_{q,l_1} \in \{0,1\}, \forall q, l_1, i, 0 \le q \le N-1, 1 \le l_1 \le l, 1 \le i \le 3$.

The major difference between deterministic and stochastic systems is the random variable, h^i_{k,l_1}, is incorporated into our system equation (4.28). We denote the probability associated with h^i_{k,l_1}s by $p^{(h)}_{l_1,\psi_1\psi_2}$ with $\psi_1, \psi_2 \in \{A, T, G, C\}$.

Ideally, we assume that we have $l_1 = 12$ kinds of mutagens, each corresponding to one major transfer pattern, as listed in Table 4.8.

Index(l_1)	From \ To	A	G	C	T	Major transfer pattern
1	A	$p^{(h)}_{1,AA}$	$p^{(h)}_{1,AG}$	$p^{(h)}_{1,AC}$	$p^{(h)}_{1,AT}$	$A \to A$
2	A	$p^{(h)}_{2,AA}$	$p^{(h)}_{2,AG}$	$p^{(h)}_{2,AC}$	$p^{(h)}_{2,AT}$	$A \to G$
3	A	$p^{(h)}_{3,AA}$	$p^{(h)}_{3,AG}$	$p^{(h)}_{3,AC}$	$p^{(h)}_{3,AT}$	$A \to C$
4	G	$p^{(h)}_{4,GA}$	$p^{(h)}_{4,GG}$	$p^{(h)}_{4,GC}$	$p^{(h)}_{4,GT}$	$G \to A$
5	G	$p^{(h)}_{5,GA}$	$p^{(h)}_{5,GG}$	$p^{(h)}_{5,GC}$	$p^{(h)}_{5,GT}$	$G \to G$
6	G	$p^{(h)}_{6,GA}$	$p^{(h)}_{6,GG}$	$p^{(h)}_{6,GC}$	$p^{(h)}_{6,GT}$	$G \to T$
7	C	$p^{(h)}_{7,CA}$	$p^{(h)}_{7,CG}$	$p^{(h)}_{7,CC}$	$p^{(h)}_{7,CT}$	$C \to A$
8	C	$p^{(h)}_{8,CA}$	$p^{(h)}_{8,CG}$	$p^{(h)}_{8,CC}$	$p^{(h)}_{8,CT}$	$C \to C$
9	C	$p^{(h)}_{9,CA}$	$p^{(h)}_{9,CG}$	$p^{(h)}_{9,CC}$	$p^{(h)}_{9,CT}$	$C \to T$
10	T	$p^{(h)}_{10,TA}$	$p^{(h)}_{10,TG}$	$p^{(h)}_{10,TC}$	$p^{(h)}_{10,TT}$	$T \to G$
11	T	$p^{(h)}_{11,TA}$	$p^{(h)}_{11,TG}$	$p^{(h)}_{11,TC}$	$p^{(h)}_{11,TT}$	$T \to C$
12	T	$p^{(h)}_{12,TA}$	$p^{(h)}_{12,TG}$	$p^{(h)}_{12,TC}$	$p^{(h)}_{12,TT}$	$T \to T$

Table 4.8: 12 kinds of mutagens, each corresponding to major transfer patterns, and probability assignment of different mutagens on different transfer patterns.

Then we can write (4.30) explicitly as

$$J_q(x_q) = \min_{u^i_{q,l_1}, 1\le l_1\le l, 1\le i\le 3} \Bigg\{ \alpha_{x^1_q\psi_1} + \mathop{\mathbb{E}}_{h^1_{q,l_1(x^1_q\psi_1)}} \left[J_{q+1}\left(\begin{bmatrix} \cdot \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right], \\ \alpha_{x^2_q\psi_2} + \mathop{\mathbb{E}}_{h^2_{q,l_1(x^2_q\psi_2)}} \left[J_{q+1}\left(\begin{bmatrix} \overline{x^1_q} \\ \cdot \\ \overline{x^3_q} \end{bmatrix} \right) \right], \alpha_{x^3_q\psi_3} + \mathop{\mathbb{E}}_{h^3_{q,l_1(x^3_q\psi_3)}} \left[J_{q+1}\left(\begin{bmatrix} \overline{x^1_q} \\ \overline{x^2_q} \\ \cdot \end{bmatrix} \right) \right], \\ \psi_1, \psi_2, \psi_3 \in \{A, T, G, C\} \Leftrightarrow \mathscr{F}_{\setminus\{0\}} \Bigg\}, \tag{4.31}$$

where

$$J_q(x_q) = \min_{u^i_{q,l_1}, 1\le l_1\le l, 1\le i\le 3} \Bigg\{ \alpha_{x^1_q\psi_1} + \mathop{\mathbb{E}}_{h^1_{q,l_1(x^1_q\psi_1)}} \left[J_{q+1}\left(\begin{bmatrix} \cdot \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right], \\ \alpha_{x^2_q\psi_2} + \mathop{\mathbb{E}}_{h^2_{q,l_1(x^2_q\psi_2)}} \left[J_{q+1}\left(\begin{bmatrix} \overline{x^1_q} \\ \cdot \\ \overline{x^3_q} \end{bmatrix} \right) \right], \alpha_{x^3_q\psi_3} + \mathop{\mathbb{E}}_{h^3_{q,l_1(x^3_q\psi_3)}} \left[J_{q+1}\left(\begin{bmatrix} \overline{x^1_q} \\ \overline{x^2_q} \\ \cdot \end{bmatrix} \right) \right], \\ \psi_1, \psi_2, \psi_3 \in \{A, T, G, C\} \Leftrightarrow \mathscr{F}_{\setminus\{0\}} \Bigg\}, \tag{4.32}$$

where

$$\mathop{\mathbb{E}}_{h^1_{q,l_1(x^1_q\psi_1)}} \left[J_{q+1}\left(\begin{bmatrix} \cdot \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right] = p^{(h)}_{l_1(x^1_q\psi_1), x^1_q A} \left[J_{q+1}\left(\begin{bmatrix} A \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right] + p^{(h)}_{l_1(x^1_q\psi_1), x^1_q G} \left[J_{q+1}\left(\begin{bmatrix} G \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right] \\ + p^{(h)}_{l_1(x^1_q\psi_1), x^1_q C} \left[J_{q+1}\left(\begin{bmatrix} C \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right] + p^{(h)}_{l_1(x^1_q\psi_1), x^1_q T} \left[J_{q+1}\left(\begin{bmatrix} T \\ \overline{x^2_q} \\ \overline{x^3_q} \end{bmatrix} \right) \right],$$

where $x_q^i \in \{A,T,G,C\} \Leftrightarrow \mathscr{F}_{\setminus\{0\}}, 0 \le q \le N-1, 1 \le i \le 3$, denotes the i^{th} element of $x_q \in \mathscr{F}_{\setminus\{0\}}^3$, $\overline{x_q^i}$ denotes the complementary base of x_q^i, and $l_1 : \psi_1\psi_2 \in \{A,T,G,C\} \times \{A,T,G,C\} \to \{\text{integers from 1 to 12}\}$, the mapping from major transfer pattern $\psi_1 \to \psi_2$ to mutagen index, as shown in Table 4.8. The mathematical expression of $\underset{h^2_{q,l_1(x_q^2\psi_2)}}{\mathbb{E}} \left[J_{q+1} \left(\left[\overline{x_q^1} \quad \cdot \quad \overline{x_q^3} \right]^T \right) \right]$ and $\underset{h^3_{q,l_1(x_q^3\psi_3)}}{\mathbb{E}} \left[J_{q+1} \left(\left[\overline{x_q^1} \quad \overline{x_q^2} \quad \cdot \right]^T \right) \right]$ is similar to $\underset{h^1_{q,l_1(x_q^1\psi_1)}}{\mathbb{E}} \left[J_{q+1} \left(\left[\cdot \quad \overline{x_q^2} \quad \overline{x_q^3} \right]^T \right) \right]$ as shown above.

Table 4.8 is the same as Table 3.8 by deleting the four artificial mutagens corresponding to natural transfers and adjusting the indices accordingly. So all elements in Table 4.8 obeys Proposition 3.5.

In order to run simulations, we assign numerical values to probabilities in Table 4.8, as illustrated in Table 4.9.

The same as in §4.5, we use three different assignments for α_{l_1}s, $\chi, 5\chi$, and 0.5χ, respectively, and the distance reference, $d(\varphi_1,\varphi_2), \varphi_1,\varphi_2 \in \mathscr{F}_{\setminus\{0\}}^3$, in Table 4.2. The indices of codons remain the same as those in §4.5, and optimal cost profile J_q with selected q values, for every pair of (x_q, x_N^d), is graphically interpreted in Figure 4.5, Figure 4.6, Figure 4.7, respectively, with $N = 29$.

Index (l_1)	From \ To	A	G	C	T
1	A	0.90	0.05	0.03	0.02
2	A	0.11	0.58	0.21	0.10
3	A	0.14	0.16	0.42	0.28
4	G	0.85	0.07	0.03	0.05
5	G	0.02	0.02	0.92	0.04
6	G	0.10	0.09	0.22	0.59
7	C	0.79	0.13	0.04	0.04
8	C	0.01	0.02	0.87	0.10
9	C	0.04	0.12	0.09	0.75
10	T	0.13	0.76	0.05	0.06
11	T	0.07	0.03	0.62	0.28
12	T	0.08	0.04	0.25	0.63

Table 4.9: Sample probabilities with respect to different mutagens and different transfer patterns.

The simulation results of codon-to-codon stochastic case are somewhat similar to the ones in §4.5. The profile of J_0 is more similar to J_{29} when α_{l_1}s are assigned 5χ than χ or 0.5χ. This implies in codon-to-codon stochastic mutations, the optimal control sequence behaves in a similar way as codon-to-codon deterministic cases, i.e. the system tends to getting as close as possible to the final desired set if α_{l_1}s are much smaller than $d(\varphi_1, \varphi_2)$, and the system tends to remain in the same state with minor mutations when α_{l_1}s are relatively larger than $d(\varphi_1, \varphi_2)$. And Claim 4.1 is still valid in codon-to-codon stochastic case with $x_q = \begin{bmatrix}\psi_1 & \psi_2 & \psi_3\end{bmatrix}^T \in \mathscr{F}^3_{\setminus\{0\}}$, $\psi_1, \psi_2, \psi_3 \in \{A, T, G, C\}$ and $x_N^d \in \mathscr{F}^3_{\setminus\{0\}}$.

However, in stochastic cases, we cannot reach a global minimum because of the randomness caused by mutagens. In general, there exists no stationary global minimum, hence an error tolerance ϵ is necessary to stop the dynamic programming algorithm with N free. In other words, if $|J_q(\psi) - J_{q-1}(\overline{\psi})| \leq \epsilon$ with the same $\{x_N^d\}$, then the dynamic programming algorithm stops iterating. Otherwise, we need to proceed to calculate $J_{q-2}(\psi)$. The value of ϵ is decided based on doctors' experience. Obviously, the smaller ϵ is, the better treatment plan is. Observing Figure 4.5, Figure 4.6 and

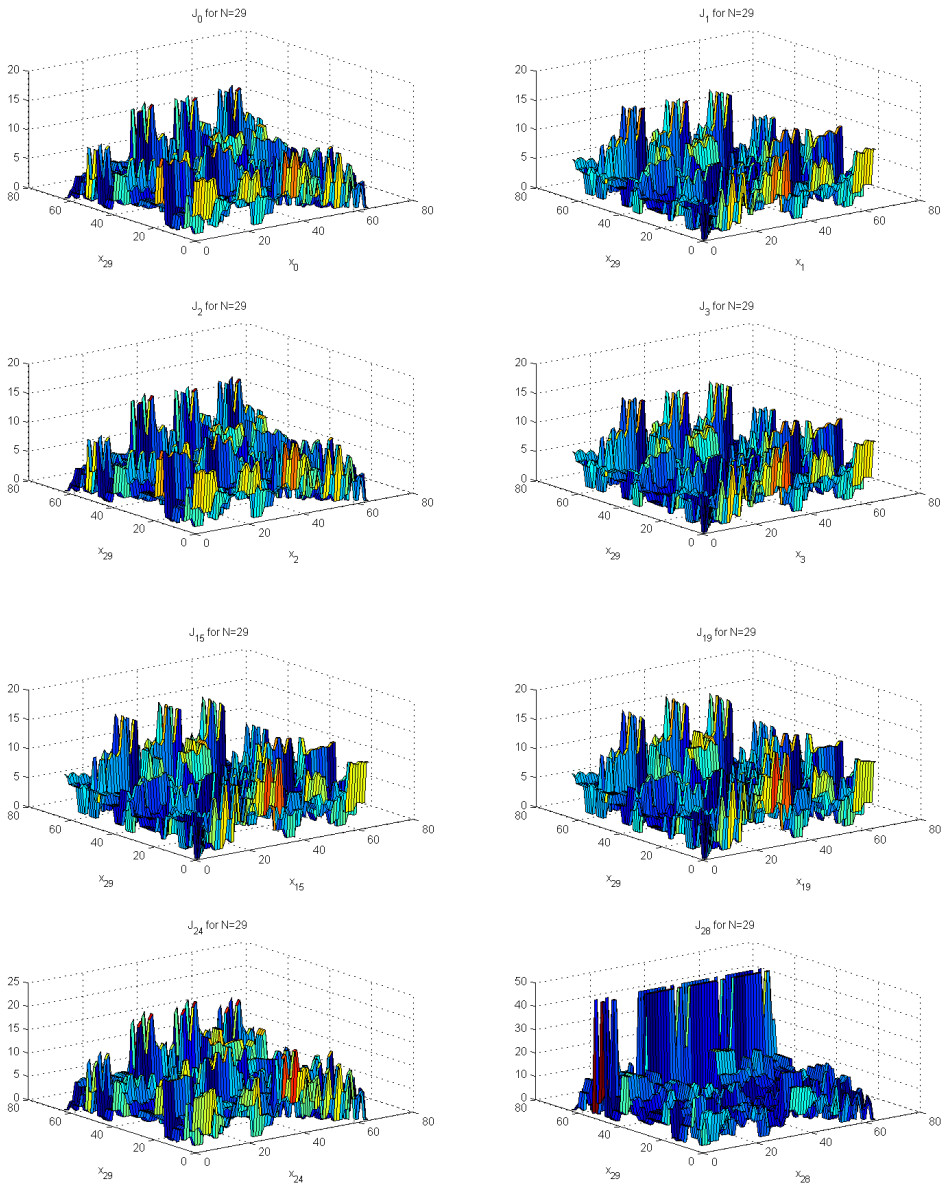

Figure 4.5: Graphically representation of $J_q(x_q), q = 0, 1, 2, 3, 15, 19, 24, 28$ for codon-to-codon stochastic mutations, with $\alpha_{l_1} = \chi$, probability assignment as in Table 4.9, $d(\cdot, \cdot)$ as listed in Table 4.2, $N = 29$.

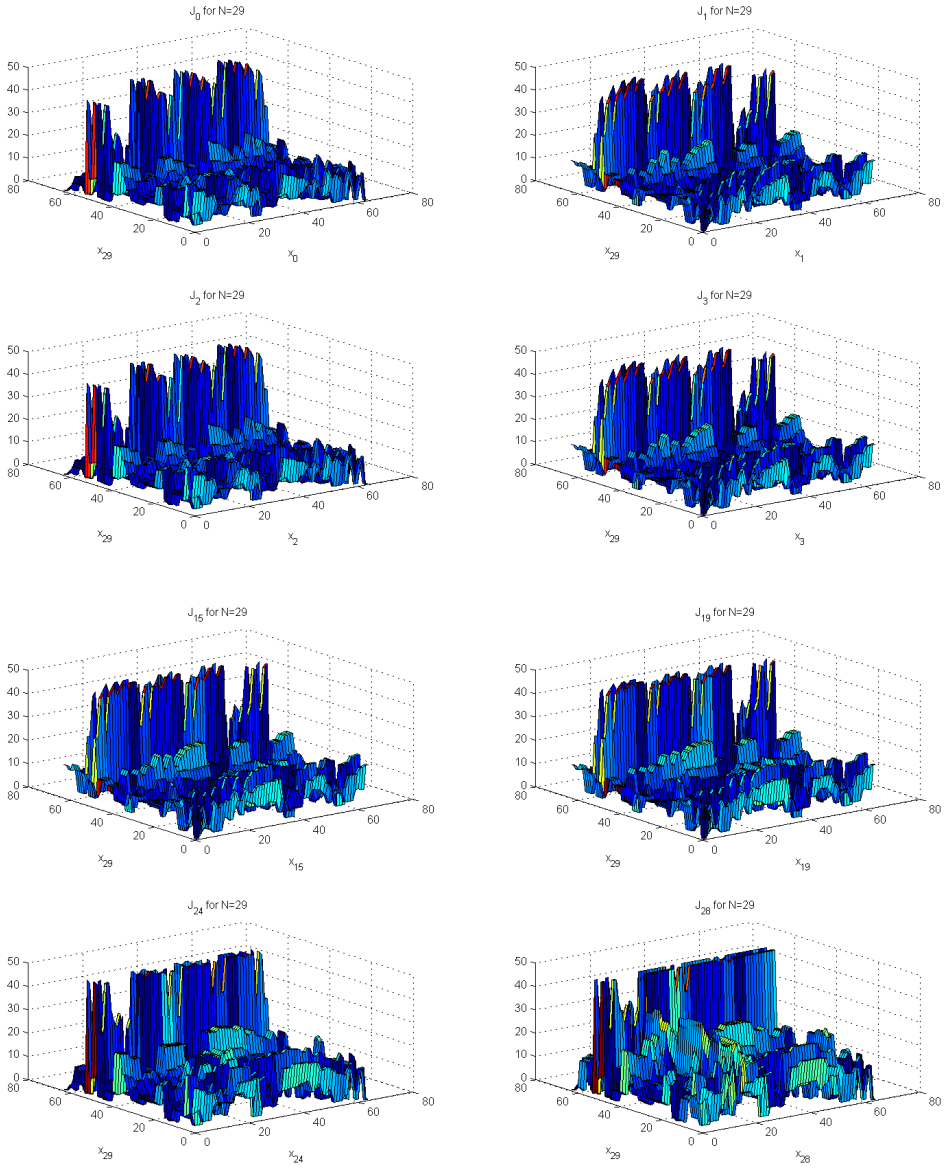

Figure 4.6: Graphically representation of $J_q(x_q), q = 0, 1, 2, 3, 15, 19, 24, 28$ for codon-to-codon stochastic mutations, with $\alpha_{l_1} = 5\chi$, probability assignment as in Table 4.9, $d(\cdot, \cdot)$ as listed in Table 4.2, $N = 29$.

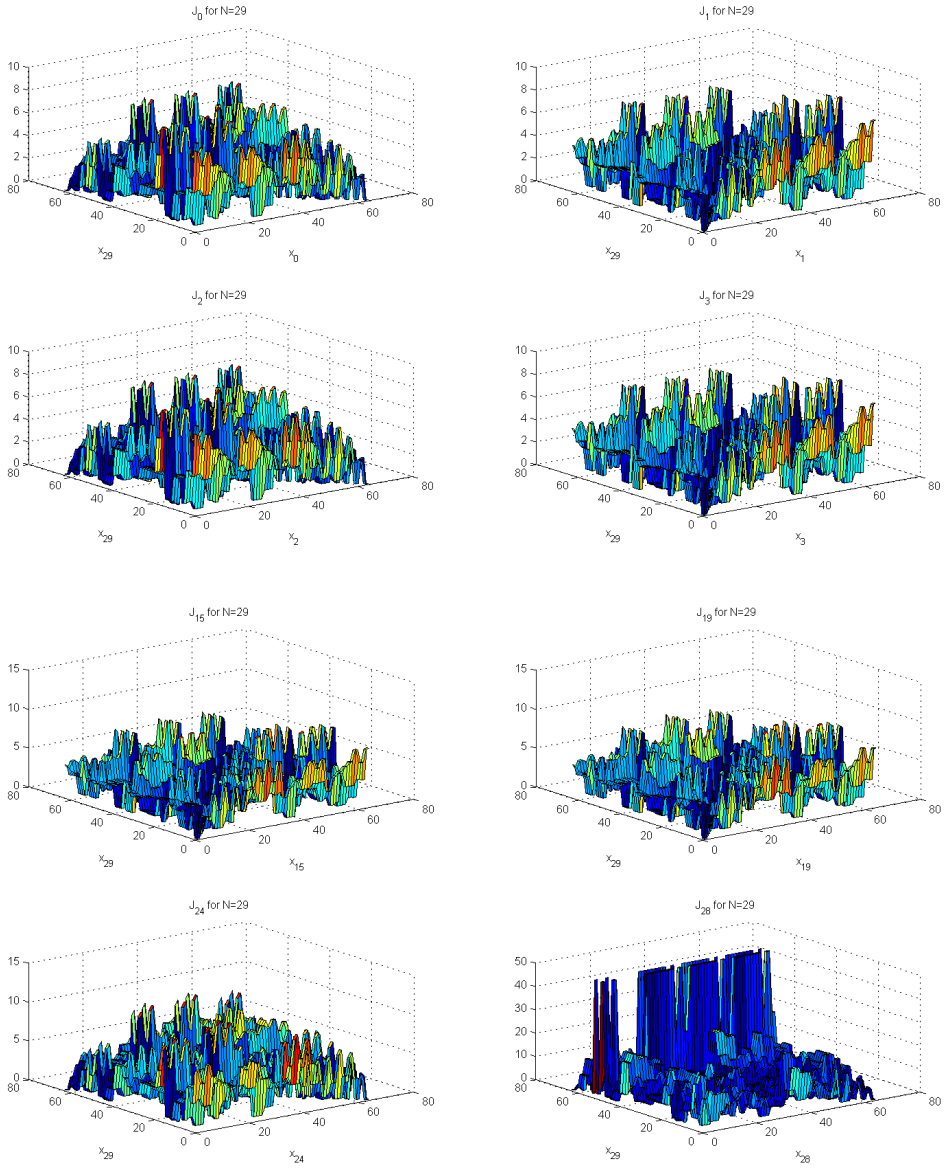

Figure 4.7: Graphically representation of $J_q(x_q), q = 0, 1, 2, 3, 15, 19, 24, 28$ for codon-to-codon stochastic mutations, with $\alpha_{l_1} = 0.5\chi$, probability assignment as in Table 4.9, $d(\cdot, \cdot)$ as listed in Table 4.2, $N = 29$.

Figure 4.7, we conclude that J_0 and J_2 are almost of the same shape in all three different parameter assignments.

Higher dimensional optimal control problems, gene-to-gene stochastic mutations, can be solved by decomposing it into a series of cascade codon-to-codon stochastic optimal control problems.

Chapter 5

Conclusion & Future Work

5.1 Conclusion

In this dissertation, we derive a mathematical model to restore the abnormal gene to a normal nucleotide sequence from the viewpoint of dynamic systems. Different from existing models, our model is constructed directly from basic biological theories, the central dogma in molecular biology and the complementary base pairing rule of DNA molecules with double helix structure. It describes in detail how the induced mutations affect a targeted DNA segment at the molecular level. It provides instrumental information for gene mutations at the molecular level to support research work at the cellular and tissue level systems. Our model is adaptive to point and multi-site, deterministic and stochastic mutations, as shown in Chapter 3. Though we emphasize that we target at the induced mutations during the process of DNA replication in our work, this model can be extended to other biological process at molecular level, such as transcription process and broken DNA strands.

In our optimal control problem, the objective function includes two factors: the risk/cost of applying mutagens and the off-trajectory penalty. Under the optimal control policy, the summation of those two factors are minimized by the dynamic programming algorithm, to propose a low-risk treatment plan. We define the distance reference following the chemical and physical properties of amino acids, representing the off-trajectory penalty. Our objective is to drive the system from a given initial state to the final desired set, generated by the final desired state, at the lowest cost. We define the final desired set since redundancy in genetic codes give us additional

options of final desired state to further reduce the cost and to ignore silent mutations. The dynamic programming algorithm ensures the optimality of the solution. We also discuss optimal control problems of three different small-scale systems, and demonstrate the simulation results of examples in Chapter 4.

The optimal control problems of base-to-base and codon-to-codon deterministic mutations are of theoretical importance. As shown in the subsequent claims, the global optimal can be reached within finite steps if the system is completely controllable. If the step limit N is larger than the number of step that global optimal can be achieved, then we have some flexibility in our treatment plan. In addition, there exist multiple optimal paths with the same total cost for some pairs of initial state and final desired set.

The optimal control problem for codon-to-codon stochastic mutations is of practical importance, since codon is the basic component form nucleotide sequence of genes. The step limit N is decided by doctors according to patients' conditions and the treatment plan is made according to the initial state and step limit. As the doctors constantly take measurement to see the effects of mutagens, the treatment plan is updated according to the current measurement. The optimal control sequence computed for codon-to-codon stochastic mutations is crucial in solving the optimal control problem for gene-to-gene stochastic mutations practically.

Our work contributes to several aspects of systems biology. The optimal control sequences generated by the dynamic programming algorithm make it possible for biologists and doctors to mutate certain sections of a gene on purpose at a relative low cost and low risk in laboratory, an essential step to identify the structure of functional units, to exam the interactions among different segments, and to find healthy, deleterious and lethal nucleotide bases combinations. All those results are beneficial in gene network construction.

The fundamental details of gene mutation at the molecular level help biologists to elaborate on biological theories at the cellular and tissue levels, such as the theory of evolution. By our method, biologists can distinguish the deleterious and beneficial mutations, and induce beneficial mutations during the evolution process in a proper way, which greatly helps to save rare species in danger.

In addition, our solution to the optimal control problem proposed provides a new medical intervention to genetic diseases. Comparing to existing gene therapy, treatments by mutagens are safer because the side effects caused virus infection are avoided.

Our work also contributes to the construction of a DNA computer. Calculation errors, mispairings in the process of two single-stranded DNA segments, can be compensated at the lowest cost by applying a different mutagens in an orderly manner.

5.2 Future Work

Future work can be done in several aspects.

Extending codon-to-codon stochastic optimal control problem to gene-to-gene stochastic mutations is one possible direction. The distance reference between DNA segments with equal length can be defined as a weighted sum of the distance references between codons. Since certain combinations of amino acids are deleterious or lethal, those high-risk states should be avoided. This goal can be achieved by either defining a collection of preset trajectories, or adding extra constraints to avoid high-risk sequences in the state space.

Also, we can examine the system's behavior with noisy measurements. Under this condition, spontaneous mutation can be modeled as an additional random factor in our state update equations. Another random noise is added to output equation representing the random factor incorporated in measurements.

In addition, mutations caused by deletions or insertions can be formulated by our method. As those cases involve the change of state space, we need adept theoretical results in information theory in the modeling process.

Moreover, our mathematical model can be applied to transcription process to control the speed and amount of protein production. Medical interventions at the cellular level can be created by controlling the number of mRNA copies in cytoplasm, which requires to a combination of systems at the molecular and cellular levels.

Appendix A

Proofs

In this appendix, we proof some results in Chapter 4.

A.1 Proof of the Optimality of Dynamic Programming Algorithm

The random variables $h^i_{k,l_1}, r^i_{k,l_2}, h'^i_{k,l_3}, r'^i_{k,l_4}$ takes a finite number of values, and the expected values of all terms in the expression of the cost function (4.7) are well defined and finite for every admissible policy π, therefore we can proof the optimality of the dynamic programming algorithm for our generalized optimal control problems. The proof follows the one to the optimality of the dynamic programming algorithm in [Bertsekas, 1995].

Proof. ([Bertsekas, 1995])

For any admissible policy $\pi = \{\mu_0, \mu_1, \cdots, \mu_{N-1}\}$ and each $q = 0, 1, \cdots, N-1$, denote $\pi^q = \{\mu_q, \mu_{q+1}, \cdots, \mu_{N-1}\}$. Let $J_q(x_q)$ be the optimal cost for the $(N-q)$-stage problem that starts at state x_q and time q, and ends at time N, i.e.

$$\hat{J}_q(x_q) = \min_{\pi_q} \underset{\{h,r,h',r'\}_{q,q+1,\cdots,N-1}}{\mathbb{E}} \left\{ g_N(x_N) + \sum_{k=q}^{N-1} g\left(x_k, \mu_k(x_k), h_k, r_k, h'_k, r'_k\right) \right\}. \quad \text{(A.1)}$$

For $q = N$, we define $\hat{J}_N(x_N) = g_N(x_N)$. We proof that $\hat{J}_q = J_q$, where J_q is generated by the dynamic programming algorithm described in §4.3. Therefore, when $q = 0$, we get the desired result.

By definition, $\hat{J}_N = J_N = g_N$. Suppose for some q and all x_{q+1}, we have $\hat{J}_{q+1}(x_{q+1}) = J_{q+1}(x_{q+1})$. Since $\pi^q = (\pi_q, \pi^{q+1})$, then $\forall x_q$,

$$\begin{aligned}
\hat{J}_q(x_q) = & \min_{(\pi_q, \pi_{q+1})} \left\{ \mathop{\mathbb{E}}_{\{h,r,h',r'\}_{q,q+1,\cdots,N-1}} \left[g\left(x_q, \mu_q(x_q), h_q, r_q, h'_q, r'_q\right) \right.\right. \\
& \left.\left. + g_N(x_N) + \sum_{k=q+1}^{N-1} g\left(x_k, \mu_k(x_k), h_k, r_k, h'_k, r'_k\right) \right] \right\} && \text{(A.2a)} \\
= & \min_{\mu_k} \left\{ \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} \left[g\left(x_q, \mu_q(x_q), h_q, r_q, h'_q, r'_q\right) \right.\right. \\
& + \min_{\pi^{q+1}} \left\{ \mathop{\mathbb{E}}_{\{h,r,h',r'\}_{q+1,\cdots,N-1}} \left[g_N(x_N) \right.\right. \\
& \left.\left.\left. + \sum_{k=q+1}^{N-1} g\left(x_k, \mu_k(x_k), h_k, r_k, h'_k, r'_k\right) \right] \right\} \right\} && \text{(A.2b)} \\
= & \min_{\mu_k} \left\{ \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} \left[g\left(x_q, \mu_q(x_q), h_q, r_q, h'_q, r'_q\right) \right.\right. \\
& \left.\left. + \hat{J}_{q+1}\left(f(x_q, \mu_q(x_q), h_q, r_q, h'_q, r'_q)\right) \right] \right\} && \text{(A.2c)} \\
= & \min_{\mu_k} \left\{ \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} \left[g\left(x_q, \mu_q(x_q), h_q, r_q, h'_q, r'_q\right) \right.\right. \\
& \left.\left. + J_{q+1}\left(f(x_q, \mu_q(x_q), h_q, r_q, h'_q, r'_q)\right) \right] \right\} && \text{(A.2d)} \\
= & \min_{\{u_q, c_q, v_q, c'_q\} \in U_q(x_q)} \left\{ \mathop{\mathbb{E}}_{h_q, r_q, h'_q, r'_q} \left[g\left(x_q, u_q, c_q, v_q, c'_q, h_q, r_q, h'_q, r'_q\right) \right.\right. \\
& \left.\left. + J_{q+1}\left(f(x_q, u_q, c_q, v_q, c'_q, h_q, r_q, h'_q, r'_q)\right) \right] \right\} && \text{(A.2e)} \\
= & J_q(x_q). && \text{(A.2f)}
\end{aligned}$$

In (A.2b), we moved the minimum over π^{k+1} inside the braced expression, using the fact that the probability distributions of $h_k, r_k, h'_k, r'_k, k = q+1, \cdots, N-1$, depend

only on x_k and $\{u_k, c_k, v_k, c'_k\}$, respectively. In (A.2c), we used the definition of $\hat{J}_{q+1}$ as in (A.1), and in the fourth equation, we used the inductions hypothesis. In (A.2e), we converted the minimization over μ_q to a minimization over $\{u_q, c_q, v_q, c'_q\}$, using the fact that for any function F of x and u, we have

$$\min_{\mu \in M} F\left(x, \mu(x)\right) = \min_{u \in U(x)} F(x, u),$$

where M is the set of all functions $\mu(x)$ such that $\mu(x) \in U(x), \forall x$. □

A.2 Proof of Claim 4.2

In this section, we proof Claim 4.2 by the brute force method. Use $J_q(x_q, x_N^d)$ to denote the optimal cost from x_q to x_N^d. Then we get

$\mathbf{q = N - 1}$
$\mathbf{J_{N-1}(A, A)} = \alpha_{AA}$
$\mathbf{J_{N-1}(G, A)} = \alpha_{GA}$
$\mathbf{J_{N-1}(C, A)} = \alpha_{CA}$
$\mathbf{J_{N-1}(T, A)} = 0$

$\mathbf{J_{N-1}(A, G)} = \alpha_{AG}$
$\mathbf{J_{N-1}(G, G)} = \alpha_{GG}$
$\mathbf{J_{N-1}(C, G)} = 0$
$\mathbf{J_{N-1}(T, G)} = \alpha_{TG}$

$\mathbf{J_{N-1}(A, C)} = \alpha_{AC}$
$\mathbf{J_{N-1}(G, C)} = 0$
$\mathbf{J_{N-1}(C, C)} = \alpha_{CC}$
$\mathbf{J_{N-1}(T, C)} = \alpha_{TC}$

$\mathbf{J_{N-1}(A, T)} = 0$
$\mathbf{J_{N-1}(G, T)} = \alpha_{GT}$

$\mathbf{J_{N-1}}(\mathbf{C}, \mathbf{T}) = \alpha_{CT}$
$\mathbf{J_{N-1}}(\mathbf{T}, \mathbf{T}) = \alpha_{TT}$

$\mathbf{q = N - 2}$
$\mathbf{J_{N-2}}(\mathbf{A}, \mathbf{A}) = 0$
$\mathbf{J_{N-2}}(\mathbf{G}, \mathbf{A}) = \min\{\alpha_{GA} + \alpha_{AA}, \alpha_{GT}, \alpha_{GG} + \alpha_{GA}, \alpha_{CA}\}$
$\mathbf{J_{N-2}}(\mathbf{C}, \mathbf{A}) = \min\{\alpha_{CA} + \alpha_{AA}, \alpha_{CT}, \alpha_{GA}, \alpha_{CC} + \alpha_{CA}\}$
$\mathbf{J_{N-2}}(\mathbf{T}, \mathbf{A}) = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{TG} + \alpha_{GA}, \alpha_{TC} + \alpha_{CA}\}$

$\mathbf{J_{N-2}}(\mathbf{A}, \mathbf{G}) = \min\{\alpha_{AA} + \alpha_{AG}, \alpha_{TG}, \alpha_{AG} + \alpha_{GG}, \alpha_{AC}\}$
$\mathbf{J_{N-2}}(\mathbf{G}, \mathbf{G}) = 0$
$\mathbf{J_{N-2}}(\mathbf{C}, \mathbf{G}) = \min\{\alpha_{CA} + \alpha_{AG}, \alpha_{CT} + \alpha_{TG}, \alpha_{GG}, \alpha_{CC}\}$
$\mathbf{J_{N-2}}(\mathbf{T}, \mathbf{G}) = \min\{\alpha_{AG}, \alpha_{TT} + \alpha_{TG}, \alpha_{TG} + \alpha_{GG}, \alpha_{TC}\}$

$\mathbf{J_{N-2}}(\mathbf{A}, \mathbf{C}) = \min\{\alpha_{AA} + \alpha_{AC}, \alpha_{TC}, \alpha_{AG}, \alpha_{AC} + \alpha_{CC}\}$
$\mathbf{J_{N-2}}(\mathbf{G}, \mathbf{C}) = \min\{\alpha_{GA} + \alpha_{AC}, \alpha_{GT} + \alpha_{TC}, \alpha_{GG}, \alpha_{CC}\}$
$\mathbf{J_{N-2}}(\mathbf{C}, \mathbf{C}) = 0$
$\mathbf{J_{N-2}}(\mathbf{T}, \mathbf{C}) = \min\{\alpha_{AC}, \alpha_{TT} + \alpha_{TC}, \alpha_{TG}, \alpha_{TC} + \alpha_{CC}\}$

$\mathbf{J_{N-2}}(\mathbf{A}, \mathbf{T}) = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{AG} + \alpha_{GT}, \alpha_{AC} + \alpha_{CT}\}$
$\mathbf{J_{N-2}}(\mathbf{G}, \mathbf{T}) = \min\{\alpha_{GA}, \alpha_{GT} + \alpha_{TT}, \alpha_{GG} + \alpha_{GT}, \alpha_{CT}\}$
$\mathbf{J_{N-2}}(\mathbf{C}, \mathbf{T}) = \min\{\alpha_{CA}, \alpha_{CT} + \alpha_{TT}, \alpha_{GT}, \alpha_{CC} + \alpha_{CT}\}$
$\mathbf{J_{N-2}}(\mathbf{T}, \mathbf{T}) = 0$

$\mathbf{q = N - 3}$
$\mathbf{J_{N-3}}(\mathbf{A}, \mathbf{A}) = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TC}+\alpha_{CA}, \alpha_{AG}+\alpha_{GT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+ \alpha_{CA}, \alpha_{AC} + \alpha_{CT}, \alpha_{AC} + \alpha_{GA}, \alpha_{AC} + \alpha_{CC} + \alpha_{CA}\}$
$\mathbf{J_{N-3}}(\mathbf{G}, \mathbf{A}) = \min\{\alpha_{GA}, \alpha_{GT} + \alpha_{AA}, \alpha_{GT} + \alpha_{TT}, \alpha_{GT} + \alpha_{TC} + \alpha_{CA}, \alpha_{GG} + \alpha_{GT}, \alpha_{GG} + \alpha_{CA}, \alpha_{CA} + \alpha_{AA}, \alpha_{CT}, \alpha_{CC} + \alpha_{CA}\}$
$\mathbf{J_{N-3}}(\mathbf{C}, \mathbf{A}) = \min\{\alpha_{CA}, \alpha_{CT}+\alpha_{AA}, \alpha_{CT}+\alpha_{TT}, \alpha_{CT}+\alpha_{TG}+\alpha_{GA}, \alpha_{GA}+\alpha_{AA}, \alpha_{GT}, \alpha_{GG}+ \alpha_{GA}, \alpha_{CC} + \alpha_{CT}, \alpha_{CC} + \alpha_{GA}\}$
$\mathbf{J_{N-3}}(\mathbf{T}, \mathbf{A}) = 0$

$\mathbf{J_{N-3}(A,G)} = \min\{\alpha_{AA}+\alpha_{TG}, \alpha_{AA}+\alpha_{AC}, \alpha_{AG}, \alpha_{TT}+\alpha_{TG}, \alpha_{TG}+\alpha_{GG}, \alpha_{TC}, \alpha_{AC}+\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GG}, \alpha_{AC}+\alpha_{CC}\}$
$\mathbf{J_{N-3}(G,G)} = \min\{\alpha_{GA}+\alpha_{AA}+\alpha_{AG}, \alpha_{GA}+\alpha_{TG}, \alpha_{GA}+\alpha_{AC}, \alpha_{GT}+\alpha_{AG}, \alpha_{GT}+\alpha_{TT}+\alpha_{TG}, \alpha_{GT}+\alpha_{TC}, \alpha_{GG}, \alpha_{CA}+\alpha_{AG}, \alpha_{CT}+\alpha_{TG}, \alpha_{CC}\}$
$\mathbf{J_{N-3}(C,G)} = 0$
$\mathbf{J_{N-3}(T,G)} = \min\{\alpha_{AA}+\alpha_{AG}, \alpha_{TG}, \alpha_{AG}+\alpha_{GG}, \alpha_{AC}, \alpha_{TT}+\alpha_{AG}, \alpha_{TT}+\alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GG}, \alpha_{TC}+\alpha_{CC}\}$

$\mathbf{J_{N-3}(A,C)} = \min\{\alpha_{AA}+\alpha_{TC}, \alpha_{AA}+\alpha_{AG}, \alpha_{AC}, \alpha_{TT}+\alpha_{TC}, \alpha_{TG}, \alpha_{TC}+\alpha_{CC}, \alpha_{AG}+\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{GG}, \alpha_{AG}+\alpha_{CC}\}$
$\mathbf{J_{N-3}(G,C)} = 0$
$\mathbf{J_{N-3}(C,C)} = \min\{\alpha_{CA}+\alpha_{AA}+\alpha_{AC}, \alpha_{CA}+\alpha_{TC}, \alpha_{CA}+\alpha_{AG}, \alpha_{CT}+\alpha_{AC}, \alpha_{CT}+\alpha_{TT}+\alpha_{TC}, \alpha_{CT}+\alpha_{TG}, \alpha_{GA}+\alpha_{AC}, \alpha_{GT}+\alpha_{TC}, \alpha_{GG}, \alpha_{CC}\}$
$\mathbf{J_{N-3}(T,C)} = \min\{\alpha_{AA}+\alpha_{AC}, \alpha_{TC}, \alpha_{AG}, \alpha_{AC}+\alpha_{CC}, \alpha_{TT}+\alpha_{AC}, \alpha_{TT}+\alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{GG}, \alpha_{TG}+\alpha_{CC}\}$

$\mathbf{J_{N-3}(A,T)} = 0$
$\mathbf{J_{N-3}(G,T)} = \min\{\alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{AC}+\alpha_{CT}, \alpha_{GT}, \alpha_{GG}+\alpha_{GA}, \alpha_{GG}+\alpha_{CT}, \alpha_{CA}, \alpha_{CT}+\alpha_{TT}, \alpha_{CC}+\alpha_{CT}\}$
$\mathbf{J_{N-3}(C,T)} = \min\{\alpha_{CA}+\alpha_{AA}, \alpha_{CA}+\alpha_{TT}, \alpha_{CA}+\alpha_{AG}+\alpha_{GT}, \alpha_{CT}, \alpha_{GA}, \alpha_{GT}+\alpha_{TT}, \alpha_{GG}+\alpha_{GT}, \alpha_{CC}+\alpha_{CA}, \alpha_{CC}+\alpha_{GT}\}$
$\mathbf{J_{N-3}(T,T)} = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{AG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{CT}, \alpha_{TC}+\alpha_{CA}, \alpha_{TC}+\alpha_{GT}, \alpha_{TC}+\alpha_{CC}+\alpha_{CT}\}$

$\mathbf{q = N-4}$
$\mathbf{J_{N-4}(A,A)} = 0$
$\mathbf{J_{N-4}(G,A)} = \min\{\alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{TG}+\alpha_{GA}, \alpha_{GA}+\alpha_{AC}+\alpha_{CT}, \alpha_{GA}+\alpha_{AC}+\alpha_{GA}, \alpha_{GT}, \alpha_{GG}+\alpha_{GA}, \alpha_{GG}+\alpha_{CT}, \alpha_{CA}, \alpha_{CT}+\alpha_{AA}, \alpha_{CT}+\alpha_{TT}, \alpha_{CT}+\alpha_{TG}+\alpha_{GA}, \alpha_{CC}+\alpha_{CT}, \alpha_{CC}+\alpha_{GA}\}$
$\mathbf{J_{N-4}(C,A)} = \min\{\alpha_{CA}+\alpha_{AA}, \alpha_{CA}+\alpha_{TT}, \alpha_{CA}+\alpha_{TC}+\alpha_{CA}, \alpha_{CA}+\alpha_{AG}+\alpha_{GT}, \alpha_{CA}+\alpha_{AG}+\alpha_{CA}, \alpha_{CT}, \alpha_{GA}, \alpha_{GT}+\alpha_{AA}, \alpha_{GT}+\alpha_{TT}, \alpha_{GT}+\alpha_{TC}+\alpha_{CA}, \alpha_{GG}+\alpha_{GT}, \alpha_{GG}+\alpha_{CA}, \alpha_{CC}+\alpha_{CA}, \alpha_{CC}+\alpha_{GT}\}$
$\mathbf{J_{N-4}(T,A)} = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TC}+\alpha_{CA}, \alpha_{AG}+\alpha_{GT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+\alpha_{CA}, \alpha_{AC}+\alpha_{CT}, \alpha_{AC}+\alpha_{GA}, \alpha_{AC}+\alpha_{CC}+\alpha_{CA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{GG}+\alpha_{CA}, \alpha_{TG}+$

$\alpha_{CT}, \alpha_{TG}+\alpha_{CC}+\alpha_{CA}, \alpha_{TC}+\alpha_{GT}, \alpha_{TC}+\alpha_{GG}+\alpha_{GA}, \alpha_{TC}+\alpha_{CC}+\alpha_{CT}, \alpha_{TC}+\alpha_{CC}+\alpha_{GA}\}$

$\mathbf{J_{N-4}(A, G)} = \min\{\alpha_{AA} + \alpha_{AG}, \alpha_{AA} + \alpha_{TC}, \alpha_{TG}, \alpha_{AG} + \alpha_{GG}, \alpha_{AC}, \alpha_{TT} + \alpha_{AG}, \alpha_{TT} + \alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GG}, \alpha_{TC}+\alpha_{CC}, \alpha_{AG}+\alpha_{GT}+\alpha_{AG}, \alpha_{AG}+\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{CA} + \alpha_{AG}, \alpha_{AG} + \alpha_{CC}\}$
$\mathbf{J_{N-4}(G, G)} = 0$
$\mathbf{J_{N-4}(C, G)} = \min\{\alpha_{CA} + \alpha_{AA} + \alpha_{TG}, \alpha_{CA} + \alpha_{AA} + \alpha_{AC}, \alpha_{CA} + \alpha_{AG}, \alpha_{CA} + \alpha_{TT} + \alpha_{TG}, \alpha_{CA}+\alpha_{TC}, \alpha_{CT}+\alpha_{AA}+\alpha_{AG}, \alpha_{CT}+\alpha_{TG}, \alpha_{CT}+\alpha_{AC}, \alpha_{CT}+\alpha_{TT}+\alpha_{AG}, \alpha_{CT}+\alpha_{TT}+\alpha_{TC}, \alpha_{GA} + \alpha_{AA} + \alpha_{AG}, \alpha_{GA} + \alpha_{TG}, \alpha_{GA} + \alpha_{AC}, \alpha_{GT} + \alpha_{AG}, \alpha_{GT} + \alpha_{TT} + \alpha_{TG}, \alpha_{GT} + \alpha_{TC}, \alpha_{GG}, \alpha_{CC}\}$
$\mathbf{J_{N-4}(T, G)} = \min\{\alpha_{AA} + \alpha_{TG}, \alpha_{AA} + \alpha_{AC}, \alpha_{AG}, \alpha_{TT} + \alpha_{TG}, \alpha_{TG} + \alpha_{GG}, \alpha_{TC}, \alpha_{AC} + \alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GG}, \alpha_{AC}+\alpha_{CC}, \alpha_{TT}+\alpha_{AC}, \alpha_{TG}+\alpha_{GA}+\alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{CT} + \alpha_{TG}, \alpha_{TG} + \alpha_{CC}\}$

$\mathbf{J_{N-4}(A, C)} = \min\{\alpha_{AA} + \alpha_{AC}, \alpha_{AA} + \alpha_{TG}, \alpha_{TC}, \alpha_{AG}, \alpha_{AC} + \alpha_{CC}, \alpha_{TT} + \alpha_{AC}, \alpha_{TT} + \alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{GG}, \alpha_{TG}+\alpha_{CC}, \alpha_{AC}+\alpha_{CT}+\alpha_{AC}, \alpha_{AC}+\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GA} + \alpha_{AC}, \alpha_{AC} + \alpha_{GG}\}$
$\mathbf{J_{N-4}(G, C)} = \min\{\alpha_{GA} + \alpha_{AA} + \alpha_{TC}, \alpha_{GA} + \alpha_{AA} + \alpha_{AG}, \alpha_{GA} + \alpha_{AC}, \alpha_{GA} + \alpha_{TT} + \alpha_{TC}, \alpha_{GA} + \alpha_{TG}, \alpha_{GT} + \alpha_{AA} + \alpha_{AC}, \alpha_{GT} + \alpha_{TC}, \alpha_{GT} + \alpha_{AG}, \alpha_{GT} + \alpha_{TT} + \alpha_{AC}, \alpha_{GT} + \alpha_{TT} + \alpha_{TG}, \alpha_{GG}, \alpha_{CA} + \alpha_{AA} + \alpha_{AC}, \alpha_{CA} + \alpha_{TC}, \alpha_{CA} + \alpha_{AG}, \alpha_{CT} + \alpha_{AC}, \alpha_{CT} + \alpha_{TT} + \alpha_{TC}, \alpha_{CT} + \alpha_{TG}, \alpha_{CC}\}$
$\mathbf{J_{N-4}(C, C)} = 0$
$\mathbf{J_{N-4}(T, C)} = \min\{\alpha_{AA} + \alpha_{TC}, \alpha_{AA} + \alpha_{AG}, \alpha_{AC}, \alpha_{TT} + \alpha_{TC}, \alpha_{TG}, \alpha_{TC} + \alpha_{CC}, \alpha_{AG} + \alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{GG}, \alpha_{AG}+\alpha_{CC}, \alpha_{TT}+\alpha_{AG}, \alpha_{TC}+\alpha_{CA}+\alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GT} + \alpha_{TC}, \alpha_{TC} + \alpha_{GG}\}$

$\mathbf{J_{N-4}(A, T)} = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{AG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{CT}, \alpha_{TC}+\alpha_{CA}, \alpha_{TC}+\alpha_{GT}, \alpha_{TC}+\alpha_{CC}+\alpha_{CT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+\alpha_{GG}+\alpha_{CT}, \alpha_{AG}+\alpha_{CA}, \alpha_{AG}+\alpha_{CC}+\alpha_{CT}, \alpha_{AC}+\alpha_{GA}, \alpha_{AC}+\alpha_{GG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CC}+\alpha_{CA}, \alpha_{AC}+\alpha_{CC}+\alpha_{GT}\}$
$\mathbf{J_{N-4}(G, T)} = \min\{\alpha_{GA}, \alpha_{GT} + \alpha_{AA}, \alpha_{GT} + \alpha_{TT}, \alpha_{GT} + \alpha_{AG} + \alpha_{GT}, \alpha_{GT} + \alpha_{TC} + \alpha_{CA}, \alpha_{GT} + \alpha_{TC} + \alpha_{GT}, \alpha_{GG} + \alpha_{GT}, \alpha_{GG} + \alpha_{CA}, \alpha_{CA} + \alpha_{AA}, \alpha_{CA} + \alpha_{TT}, \alpha_{CA} + \alpha_{AG} + \alpha_{GT}, \alpha_{CT}, \alpha_{CC} + \alpha_{CA}, \alpha_{CC} + \alpha_{GT}\}$
$\mathbf{J_{N-4}(C, T)} = \min\{\alpha_{CA}, \alpha_{CT} + \alpha_{AA}, \alpha_{CT} + \alpha_{TT}, \alpha_{CT} + \alpha_{AC} + \alpha_{CT}, \alpha_{CT} + \alpha_{TG} + \alpha_{GA}, \alpha_{CT}+\alpha_{TG}+\alpha_{CT}, \alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{AC}+\alpha_{CT}, \alpha_{GT}, \alpha_{GG}+\alpha_{GA}, \alpha_{GG}+$

$\alpha_{CT}, \alpha_{CC} + \alpha_{CT}, \alpha_{CC} + \alpha_{GA}\}$
$\mathbf{J_{N-4}}(\mathbf{T}, \mathbf{T}) = 0$

$\mathbf{q = N - 5}$
$\mathbf{J_{N-5}}(\mathbf{A}, \mathbf{A}) = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TC}+\alpha_{CA}, \alpha_{AG}+\alpha_{GT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+$
$\alpha_{CA}, \alpha_{AC}+\alpha_{CT}, \alpha_{AC}+\alpha_{GA}, \alpha_{AC}+\alpha_{CC}+\alpha_{CA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{GG}+\alpha_{CA}, \alpha_{TG}+$
$\alpha_{CT}, \alpha_{TG} + \alpha_{CC} + \alpha_{CA}, \alpha_{TC} + \alpha_{GT}, \alpha_{TC} + \alpha_{GG} + \alpha_{GA}, \alpha_{TC} + \alpha_{CC} + \alpha_{CT}, \alpha_{TC} + \alpha_{CC} +$
$\alpha_{GA}, \alpha_{AG} + \alpha_{GG} + \alpha_{CT}, \alpha_{AG} + \alpha_{CC} + \alpha_{CT}, \alpha_{AG} + \alpha_{CC} + \alpha_{GA}, \alpha_{AC} + \alpha_{GG} + \alpha_{GT}, \alpha_{AC} +$
$\alpha_{GG} + \alpha_{CA}, \alpha_{AC} + \alpha_{CC} + \alpha_{GT}\}$
$\mathbf{J_{N-5}}(\mathbf{G}, \mathbf{A}) = \min\{\alpha_{GA}, \alpha_{GT} + \alpha_{AA}, \alpha_{GT} + \alpha_{TT}, \alpha_{GT} + \alpha_{TC} + \alpha_{CA}, \alpha_{GT} + \alpha_{AG} +$
$\alpha_{GT}, \alpha_{GT} + \alpha_{AG} + \alpha_{CA}, \alpha_{GT} + \alpha_{TC} + \alpha_{GT}, \alpha_{GG} + \alpha_{GT}, \alpha_{GG} + \alpha_{CA}, \alpha_{CA} + \alpha_{AA}, \alpha_{CA} +$
$\alpha_{TT}, \alpha_{CA}+\alpha_{TC}+\alpha_{CA}, \alpha_{CA}+\alpha_{AG}+\alpha_{GT}, \alpha_{CA}+\alpha_{AG}+\alpha_{CA}, \alpha_{CT}, \alpha_{CC}+\alpha_{CA}, \alpha_{CC}+\alpha_{GT}\}$
$\mathbf{J_{N-5}}(\mathbf{C}, \mathbf{A}) = \min\{\alpha_{CA}, \alpha_{CT} + \alpha_{AA}, \alpha_{CT} + \alpha_{TT}, \alpha_{CT} + \alpha_{TG} + \alpha_{GA}, \alpha_{CT} + \alpha_{AC} +$
$\alpha_{CT}, \alpha_{CT}+\alpha_{AC}+\alpha_{GA}, \alpha_{CT}+\alpha_{TG}+\alpha_{CT}, \alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{TG}+\alpha_{GA}, \alpha_{GA}+$
$\alpha_{AC} + \alpha_{CT}, \alpha_{GA} + \alpha_{AC} + \alpha_{GA}, \alpha_{GT}, \alpha_{GG} + \alpha_{GA}, \alpha_{GG} + \alpha_{CT}, \alpha_{CC} + \alpha_{CT}, \alpha_{CC} + \alpha_{GA}\}$
$\mathbf{J_{N-5}}(\mathbf{T}, \mathbf{A}) = 0$

$\mathbf{J_{N-5}}(\mathbf{A}, \mathbf{G}) = \min\{\alpha_{AA} + \alpha_{TG}, \alpha_{AA} + \alpha_{AC}, \alpha_{AG}, \alpha_{TT} + \alpha_{TG}, \alpha_{TG} + \alpha_{GG}, \alpha_{TC}, \alpha_{AC} +$
$\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GG}, \alpha_{AC}+\alpha_{CC}, \alpha_{TT}+\alpha_{AC}, \alpha_{TG}+\alpha_{GA}+\alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+$
$\alpha_{CT} + \alpha_{TG}, \alpha_{TG} + \alpha_{CC}, \alpha_{AC} + \alpha_{CT} + \alpha_{AC}, \alpha_{AC} + \alpha_{GA} + \alpha_{TG}, \alpha_{AC} + \alpha_{GA} + \alpha_{AC}\}$
$\mathbf{J_{N-5}}(\mathbf{G}, \mathbf{G}) = \min\{\alpha_{GA}+\alpha_{AA}+\alpha_{AG}, \alpha_{GA}+\alpha_{AA}+\alpha_{TC}, \alpha_{GA}+\alpha_{TG}, \alpha_{GA}+\alpha_{AC}, \alpha_{GA}+$
$\alpha_{TT} + \alpha_{AG}, \alpha_{GA} + \alpha_{TT} + \alpha_{TC}, \alpha_{GT} + \alpha_{AA} + \alpha_{TG}, \alpha_{GT} + \alpha_{AA} + \alpha_{AC}, \alpha_{GT} + \alpha_{AG}, \alpha_{GT} +$
$\alpha_{TT}+\alpha_{TG}, \alpha_{GT}+\alpha_{TC}, \alpha_{GT}+\alpha_{TT}+\alpha_{AC}, \alpha_{GG}, \alpha_{CA}+\alpha_{AA}+\alpha_{TG}, \alpha_{CA}+\alpha_{AA}+\alpha_{AC}, \alpha_{CA}+$
$\alpha_{AG}, \alpha_{CA} + \alpha_{TT} + \alpha_{TG}, \alpha_{CA} + \alpha_{TC}, \alpha_{CT} + \alpha_{AA} + \alpha_{AG}, \alpha_{CT} + \alpha_{TG}, \alpha_{CT} + \alpha_{AC}, \alpha_{CT} +$
$\alpha_{TT} + \alpha_{AG}, \alpha_{CT} + \alpha_{TT} + \alpha_{TC}, \alpha_{CC}\}$
$\mathbf{J_{N-5}}(\mathbf{C}, \mathbf{G}) = 0$
$\mathbf{J_{N-5}}(\mathbf{T}, \mathbf{G}) = \min\{\alpha_{AA} + \alpha_{AG}, \alpha_{AA} + \alpha_{TC}, \alpha_{TG}, \alpha_{AG} + \alpha_{GG}, \alpha_{AC}, \alpha_{TT} + \alpha_{AG}, \alpha_{TT} +$
$\alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GG}, \alpha_{TC}+\alpha_{CC}, \alpha_{AG}+\alpha_{GT}+\alpha_{AG}, \alpha_{AG}+\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+$
$\alpha_{CA} + \alpha_{AG}, \alpha_{AG} + \alpha_{CC}, \alpha_{TC} + \alpha_{CA} + \alpha_{TC}, \alpha_{TC} + \alpha_{GT} + \alpha_{AG}, \alpha_{TC} + \alpha_{GT} + \alpha_{TC}\}$

$\mathbf{J_{N-5}}(\mathbf{A}, \mathbf{C}) = \min\{\alpha_{AA} + \alpha_{TC}, \alpha_{AA} + \alpha_{AG}, \alpha_{AC}, \alpha_{TT} + \alpha_{TC}, \alpha_{TG}, \alpha_{TC} + \alpha_{CC}, \alpha_{AG} +$
$\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{GG}, \alpha_{AG}+\alpha_{CC}, \alpha_{TT}+\alpha_{AG}, \alpha_{TC}+\alpha_{CA}+\alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+$
$\alpha_{GT} + \alpha_{TC}, \alpha_{TC} + \alpha_{GG}, \alpha_{AG} + \alpha_{GT} + \alpha_{AG}, \alpha_{AG} + \alpha_{CA} + \alpha_{TC}, \alpha_{AG} + \alpha_{CA} + \alpha_{AG}\}$
$\mathbf{J_{N-5}}(\mathbf{G}, \mathbf{C}) = 0$

$\mathbf{J_{N-5}(C,C)} = \min\{\alpha_{CA}+\alpha_{AA}+\alpha_{AC}, \alpha_{CA}+\alpha_{AA}+\alpha_{TG}, \alpha_{CA}+\alpha_{TC}, \alpha_{CA}+\alpha_{AG}, \alpha_{CA}+\alpha_{TT}+\alpha_{AC}, \alpha_{CA}+\alpha_{TT}+\alpha_{TG}, \alpha_{CT}+\alpha_{AA}+\alpha_{TC}, \alpha_{CT}+\alpha_{AA}+\alpha_{AG}, \alpha_{CT}+\alpha_{AC}, \alpha_{CT}+\alpha_{TT}+\alpha_{TC}, \alpha_{CT}+\alpha_{TG}, \alpha_{CT}+\alpha_{TT}+\alpha_{AG}, \alpha_{GA}+\alpha_{AA}+\alpha_{TC}, \alpha_{GA}+\alpha_{AA}+\alpha_{AG}, \alpha_{GA}+\alpha_{AC}, \alpha_{GA}+\alpha_{TT}+\alpha_{TC}, \alpha_{GA}+\alpha_{TG}, \alpha_{GT}+\alpha_{AA}+\alpha_{AC}, \alpha_{GT}+\alpha_{TC}, \alpha_{GT}+\alpha_{AG}, \alpha_{GT}+\alpha_{TT}+\alpha_{AC}, \alpha_{GT}+\alpha_{TT}+\alpha_{TG}, \alpha_{GG}, \alpha_{CC}\}$

$\mathbf{J_{N-5}(T,C)} = \min\{\alpha_{AA}+\alpha_{AC}, \alpha_{AA}+\alpha_{TG}, \alpha_{TC}, \alpha_{AG}, \alpha_{AC}+\alpha_{CC}, \alpha_{TT}+\alpha_{AC}, \alpha_{TT}+\alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{GG}, \alpha_{TG}+\alpha_{CC}, \alpha_{AC}+\alpha_{CT}+\alpha_{AC}, \alpha_{AC}+\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GA}+\alpha_{AC}, \alpha_{AC}+\alpha_{GG}, \alpha_{TG}+\alpha_{GA}+\alpha_{TG}, \alpha_{TG}+\alpha_{CT}+\alpha_{AC}, \alpha_{TG}+\alpha_{CT}+\alpha_{TG}\}$

$\mathbf{J_{N-5}(A,T)} = 0$

$\mathbf{J_{N-5}(G,T)} = \min\{\alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{AC}+\alpha_{CT}, \alpha_{GA}+\alpha_{TG}+\alpha_{GA}, \alpha_{GA}+\alpha_{TG}+\alpha_{CT}, \alpha_{GA}+\alpha_{AC}+\alpha_{GA}, \alpha_{GT}, \alpha_{GG}+\alpha_{GA}, \alpha_{GG}+\alpha_{CT}, \alpha_{CA}, \alpha_{CT}+\alpha_{AA}, \alpha_{CT}+\alpha_{TT}, \alpha_{CT}+\alpha_{AC}+\alpha_{CT}, \alpha_{CT}+\alpha_{TG}+\alpha_{GA}, \alpha_{CT}+\alpha_{TG}+\alpha_{CT}, \alpha_{CC}+\alpha_{CT}, \alpha_{CC}+\alpha_{GA}\}$

$\mathbf{J_{N-5}(C,T)} = \min\{\alpha_{CA}+\alpha_{AA}, \alpha_{CA}+\alpha_{TT}, \alpha_{CA}+\alpha_{AG}+\alpha_{GT}, \alpha_{CA}+\alpha_{TC}+\alpha_{CA}, \alpha_{CA}+\alpha_{TC}+\alpha_{GT}, \alpha_{CA}+\alpha_{AG}+\alpha_{CA}, \alpha_{CT}, \alpha_{GA}, \alpha_{GT}+\alpha_{AA}, \alpha_{GT}+\alpha_{TT}, \alpha_{GT}+\alpha_{AG}+\alpha_{GT}, \alpha_{GT}+\alpha_{TC}+\alpha_{CA}, \alpha_{GT}+\alpha_{TC}+\alpha_{GT}, \alpha_{GG}+\alpha_{GT}, \alpha_{GG}+\alpha_{CA}, \alpha_{CC}+\alpha_{CA}, \alpha_{CC}+\alpha_{GT}\}$

$\mathbf{J_{N-5}(T,T)} = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{AG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{CT}, \alpha_{TC}+\alpha_{CA}, \alpha_{TC}+\alpha_{GT}, \alpha_{TC}+\alpha_{CC}+\alpha_{CT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+\alpha_{GG}+\alpha_{CT}, \alpha_{AG}+\alpha_{CA}, \alpha_{AG}+\alpha_{CC}+\alpha_{CT}, \alpha_{AC}+\alpha_{GA}, \alpha_{AC}+\alpha_{GG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CC}+\alpha_{CA}, \alpha_{AC}+\alpha_{CC}+\alpha_{GT}, \alpha_{TG}+\alpha_{GG}+\alpha_{CA}, \alpha_{TG}+\alpha_{CC}+\alpha_{CA}, \alpha_{TG}+\alpha_{CC}+\alpha_{GT}, \alpha_{TC}+\alpha_{GG}+\alpha_{GA}, \alpha_{TC}+\alpha_{GG}+\alpha_{CT}, \alpha_{TC}+\alpha_{CC}+\alpha_{GA}\}$

$\mathbf{q = N-6}$

$\mathbf{J_{N-6}(A,A)} = 0$

$\mathbf{J_{N-6}(G,A)} = \min\{\alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{TG}+\alpha_{GA}, \alpha_{GA}+\alpha_{AC}+\alpha_{CT}, \alpha_{GA}+\alpha_{AC}+\alpha_{GA}, \alpha_{GA}+\alpha_{TG}+\alpha_{CT}, \alpha_{GT}, \alpha_{GG}+\alpha_{GA}, \alpha_{GG}+\alpha_{CT}, \alpha_{CA}, \alpha_{CT}+\alpha_{AA}, \alpha_{CT}+\alpha_{TT}, \alpha_{CT}+\alpha_{TG}+\alpha_{GA}, \alpha_{CT}+\alpha_{AC}+\alpha_{CT}, \alpha_{CT}+\alpha_{AC}+\alpha_{GA}, \alpha_{CT}+\alpha_{TG}+\alpha_{CT}, \alpha_{CC}+\alpha_{CT}, \alpha_{CC}+\alpha_{GA}\}$

$\mathbf{J_{N-6}(C,A)} = \min\{\alpha_{CA}+\alpha_{AA}, \alpha_{CA}+\alpha_{TT}, \alpha_{CA}+\alpha_{TC}+\alpha_{CA}, \alpha_{CA}+\alpha_{AG}+\alpha_{GT}, \alpha_{CA}+\alpha_{AG}+\alpha_{CA}, \alpha_{CA}+\alpha_{TC}+\alpha_{GT}, \alpha_{CT}, \alpha_{GA}, \alpha_{GT}+\alpha_{AA}, \alpha_{GT}+\alpha_{TT}, \alpha_{GT}+\alpha_{TC}+\alpha_{CA}, \alpha_{GT}+\alpha_{AG}+\alpha_{GT}, \alpha_{GT}+\alpha_{AG}+\alpha_{CA}, \alpha_{GT}+\alpha_{TC}+\alpha_{GT}, \alpha_{GG}+\alpha_{GT}, \alpha_{GG}+\alpha_{CA}, \alpha_{CC}+\alpha_{CA}, \alpha_{CC}+\alpha_{GT}\}$

$\mathbf{J_{N-6}(T,A)} = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TC}+\alpha_{CA}, \alpha_{AG}+\alpha_{GT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+\alpha_{CA}, \alpha_{AC}+\alpha_{CT}, \alpha_{AC}+\alpha_{GA}, \alpha_{AC}+\alpha_{CC}+\alpha_{CA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{GG}+\alpha_{CA}, \alpha_{TG}+$

$\alpha_{CT}, \alpha_{TG}+\alpha_{CC}+\alpha_{CA}, \alpha_{TC}+\alpha_{GT}, \alpha_{TC}+\alpha_{GG}+\alpha_{GA}, \alpha_{TC}+\alpha_{CC}+\alpha_{CT}, \alpha_{TC}+\alpha_{CC}+\alpha_{GA}, \alpha_{AG}+\alpha_{GG}+\alpha_{CT}, \alpha_{AG}+\alpha_{CC}+\alpha_{CT}, \alpha_{AG}+\alpha_{CC}+\alpha_{GA}, \alpha_{AC}+\alpha_{GG}+\alpha_{GT}, \alpha_{AC}+\alpha_{GG}+\alpha_{CA}, \alpha_{AC}+\alpha_{CC}+\alpha_{GT}, \alpha_{TG}+\alpha_{CC}+\alpha_{GT}, \alpha_{TC}+\alpha_{GG}+\alpha_{CT}\}$

$\mathbf{J_{N-6}(A, G)} = \min\{\alpha_{AA}+\alpha_{AG}, \alpha_{AA}+\alpha_{TC}, \alpha_{TG}, \alpha_{AG}+\alpha_{GG}, \alpha_{AC}, \alpha_{TT}+\alpha_{AG}, \alpha_{TT}+\alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GG}, \alpha_{TC}+\alpha_{CC}, \alpha_{AG}+\alpha_{GT}+\alpha_{AG}, \alpha_{AG}+\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{CA}+\alpha_{AG}, \alpha_{AG}+\alpha_{CC}, \alpha_{TC}+\alpha_{CA}+\alpha_{TC}, \alpha_{TC}+\alpha_{GT}+\alpha_{AG}, \alpha_{TC}+\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{CA}+\alpha_{TC}\}$
$\mathbf{J_{N-6}(G, G)} = 0$
$\mathbf{J_{N-6}(C, G)} = \min\{\alpha_{CA}+\alpha_{AA}+\alpha_{TG}, \alpha_{CA}+\alpha_{AA}+\alpha_{AC}, \alpha_{CA}+\alpha_{AG}, \alpha_{CA}+\alpha_{TT}+\alpha_{TG}, \alpha_{CA}+\alpha_{TC}, \alpha_{CA}+\alpha_{TT}+\alpha_{AC}, \alpha_{CT}+\alpha_{AA}+\alpha_{AG}, \alpha_{CT}+\alpha_{AA}+\alpha_{TC}, \alpha_{CT}+\alpha_{TG}, \alpha_{CT}+\alpha_{AC}, \alpha_{CT}+\alpha_{TT}+\alpha_{AG}, \alpha_{CT}+\alpha_{TT}+\alpha_{TC}, \alpha_{GA}+\alpha_{AA}+\alpha_{AG}, \alpha_{GA}+\alpha_{AA}+\alpha_{TC}, \alpha_{GA}+\alpha_{TG}, \alpha_{GA}+\alpha_{AC}, \alpha_{GA}+\alpha_{TT}+\alpha_{AG}, \alpha_{GA}+\alpha_{TT}+\alpha_{TC}, \alpha_{GT}+\alpha_{AA}+\alpha_{TG}, \alpha_{GT}+\alpha_{AA}+\alpha_{AC}, \alpha_{GT}+\alpha_{AG}, \alpha_{GT}+\alpha_{TT}+\alpha_{TG}, \alpha_{GT}+\alpha_{TC}, \alpha_{GT}+\alpha_{TT}+\alpha_{AC}, \alpha_{GG}, \alpha_{CC}\}$
$\mathbf{J_{N-6}(T, G)} = \min\{\alpha_{AA}+\alpha_{TG}, \alpha_{AA}+\alpha_{AC}, \alpha_{AG}, \alpha_{TT}+\alpha_{TG}, \alpha_{TG}+\alpha_{GG}, \alpha_{TC}, \alpha_{AC}+\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GG}, \alpha_{AC}+\alpha_{CC}, \alpha_{TT}+\alpha_{AC}, \alpha_{TG}+\alpha_{GA}+\alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{CT}+\alpha_{TG}, \alpha_{TG}+\alpha_{CC}, \alpha_{AC}+\alpha_{CT}+\alpha_{AC}, \alpha_{AC}+\alpha_{GA}+\alpha_{TG}, \alpha_{AC}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{CT}+\alpha_{AC}\}$

$\mathbf{J_{N-6}(A, C)} = \min\{\alpha_{AA}+\alpha_{AC}, \alpha_{AA}+\alpha_{TG}, \alpha_{TC}, \alpha_{AG}, \alpha_{AC}+\alpha_{CC}, \alpha_{TT}+\alpha_{AC}, \alpha_{TT}+\alpha_{TG}, \alpha_{TG}+\alpha_{GA}+\alpha_{AC}, \alpha_{TG}+\alpha_{GG}, \alpha_{TG}+\alpha_{CC}, \alpha_{AC}+\alpha_{CT}+\alpha_{AC}, \alpha_{AC}+\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GA}+\alpha_{AC}, \alpha_{AC}+\alpha_{GG}, \alpha_{TG}+\alpha_{GA}+\alpha_{TG}, \alpha_{TG}+\alpha_{CT}+\alpha_{AC}, \alpha_{TG}+\alpha_{CT}+\alpha_{TG}, \alpha_{AC}+\alpha_{GA}+\alpha_{TG}\}$
$\mathbf{J_{N-6}(G, C)} = \min\{\alpha_{GA}+\alpha_{AA}+\alpha_{TC}, \alpha_{GA}+\alpha_{AA}+\alpha_{AG}, \alpha_{GA}+\alpha_{AC}, \alpha_{GA}+\alpha_{TT}+\alpha_{TC}, \alpha_{GA}+\alpha_{TG}, \alpha_{GA}+\alpha_{TT}+\alpha_{AG}, \alpha_{GT}+\alpha_{AA}+\alpha_{AC}, \alpha_{GT}+\alpha_{AA}+\alpha_{TG}, \alpha_{GT}+\alpha_{TC}, \alpha_{GT}+\alpha_{AG}, \alpha_{GT}+\alpha_{TT}+\alpha_{AC}, \alpha_{GT}+\alpha_{TT}+\alpha_{TG}, \alpha_{GG}, \alpha_{CA}+\alpha_{AA}+\alpha_{AC}, \alpha_{CA}+\alpha_{AA}+\alpha_{TG}, \alpha_{CA}+\alpha_{TC}, \alpha_{CA}+\alpha_{AG}, \alpha_{CA}+\alpha_{TT}+\alpha_{AC}, \alpha_{CA}+\alpha_{TT}+\alpha_{TG}, \alpha_{CT}+\alpha_{AA}+\alpha_{TC}, \alpha_{CT}+\alpha_{AA}+\alpha_{AG}, \alpha_{CT}+\alpha_{AC}, \alpha_{CT}+\alpha_{TT}+\alpha_{TC}, \alpha_{CT}+\alpha_{TG}, \alpha_{CT}+\alpha_{TT}+\alpha_{AG}, \alpha_{CC}\}$
$\mathbf{J_{N-6}(C, C)} = 0$
$\mathbf{J_{N-6}(T, C)} = \min\{\alpha_{AA}+\alpha_{TC}, \alpha_{AA}+\alpha_{AG}, \alpha_{AC}, \alpha_{TT}+\alpha_{TC}, \alpha_{TG}, \alpha_{TC}+\alpha_{CC}, \alpha_{AG}+\alpha_{GT}+\alpha_{TC}, \alpha_{AG}+\alpha_{GG}, \alpha_{AG}+\alpha_{CC}, \alpha_{TT}+\alpha_{AG}, \alpha_{TC}+\alpha_{CA}+\alpha_{TC}, \alpha_{TC}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GT}+\alpha_{TC}, \alpha_{TC}+\alpha_{GG}, \alpha_{AG}+\alpha_{GT}+\alpha_{AG}, \alpha_{AG}+\alpha_{CA}+\alpha_{TC}, \alpha_{AG}+\alpha_{CA}+\alpha_{AG}, \alpha_{TC}+\alpha_{GT}+\alpha_{AG}\}$

$\mathbf{J_{N-6}(A,T)} = \min\{\alpha_{AA}, \alpha_{TT}, \alpha_{AG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CT}, \alpha_{TG}+\alpha_{GA}, \alpha_{TG}+\alpha_{GG}+\alpha_{GT}, \alpha_{TG}+\alpha_{CT}, \alpha_{TC}+\alpha_{CA}, \alpha_{TC}+\alpha_{GT}, \alpha_{TC}+\alpha_{CC}+\alpha_{CT}, \alpha_{AG}+\alpha_{GG}+\alpha_{GA}, \alpha_{AG}+\alpha_{GG}+\alpha_{CT}, \alpha_{AG}+\alpha_{CA}, \alpha_{AG}+\alpha_{CC}+\alpha_{CT}, \alpha_{AC}+\alpha_{GA}, \alpha_{AC}+\alpha_{GG}+\alpha_{GT}, \alpha_{AC}+\alpha_{CC}+\alpha_{CA}, \alpha_{AC}+\alpha_{CC}+\alpha_{GT}, \alpha_{TG}+\alpha_{GG}+\alpha_{CA}, \alpha_{TG}+\alpha_{CC}+\alpha_{CA}, \alpha_{TG}+\alpha_{CC}+\alpha_{GT}, \alpha_{TC}+\alpha_{GG}+\alpha_{GA}, \alpha_{TC}+\alpha_{GG}+\alpha_{CT}, \alpha_{TC}+\alpha_{CC}+\alpha_{GA}, \alpha_{AG}+\alpha_{CC}+\alpha_{GA}, \alpha_{AC}+\alpha_{GG}+\alpha_{CA}\}$

$\mathbf{J_{N-6}(G,T)} = \min\{\alpha_{GA}, \alpha_{GT}+\alpha_{AA}, \alpha_{GT}+\alpha_{TT}, \alpha_{GT}+\alpha_{AG}+\alpha_{GT}, \alpha_{GT}+\alpha_{TC}+\alpha_{CA}, \alpha_{GT}+\alpha_{TC}+\alpha_{GT}, \alpha_{GT}+\alpha_{AG}+\alpha_{CA}, \alpha_{GG}+\alpha_{GT}, \alpha_{GG}+\alpha_{CA}, \alpha_{CA}+\alpha_{AA}, \alpha_{CA}+\alpha_{TT}, \alpha_{CA}+\alpha_{AG}+\alpha_{GT}, \alpha_{CA}+\alpha_{TC}+\alpha_{CA}, \alpha_{CA}+\alpha_{TC}+\alpha_{GT}, \alpha_{CA}+\alpha_{AG}+\alpha_{CA}, \alpha_{CT}, \alpha_{CC}+\alpha_{CA}, \alpha_{CC}+\alpha_{GT}\}$

$\mathbf{J_{N-6}(C,T)} = \min\{\alpha_{CA}, \alpha_{CT}+\alpha_{AA}, \alpha_{CT}+\alpha_{TT}, \alpha_{CT}+\alpha_{AC}+\alpha_{CT}, \alpha_{CT}+\alpha_{TG}+\alpha_{GA}, \alpha_{CT}+\alpha_{TG}+\alpha_{CT}, \alpha_{CT}+\alpha_{AC}+\alpha_{GA}, \alpha_{GA}+\alpha_{AA}, \alpha_{GA}+\alpha_{TT}, \alpha_{GA}+\alpha_{AC}+\alpha_{CT}, \alpha_{GA}+\alpha_{TG}+\alpha_{GA}, \alpha_{GA}+\alpha_{TG}+\alpha_{CT}, \alpha_{GA}+\alpha_{AC}+\alpha_{GA}, \alpha_{GT}, \alpha_{GG}+\alpha_{GA}, \alpha_{GG}+\alpha_{CT}, \alpha_{CC}+\alpha_{CT}, \alpha_{CC}+\alpha_{GA}\}$

$\mathbf{J_{N-6}(T,T)} = 0$

$\mathbf{q = N-7}$

$\forall \psi_1, \psi_2 \in \{A, T, G, C\}$,

$$J_{N-7}(\psi_1, \psi_2) = J_{N-6}(\overline{\psi_1}, \psi_2) = J_{N-6}(\psi_1, \overline{\psi_2}) = J_{N-7}(\overline{\psi_1}, \overline{\psi_2}). \tag{A.3}$$

$\mathbf{q \leq N-7}$

$\forall \psi_1, \psi_2 \in \{A, T, G, C\}$,

$$J_q(\psi_1, \psi_2) = J_{q+1}(\overline{\psi_1}, \psi_2) = J_{q+1}(\psi_1, \overline{\psi_2}) = J_q(\overline{\psi_1}, \overline{\psi_2}). \tag{A.4}$$

Remark A.1. *For $q \geq N-6$, $J_q(\psi, x_N^d) = \min \Gamma_q(\psi, x_N^d)$ and $J_{q+1}(\overline{\psi}, x_N^d) = \min \Gamma_{q+1}(\overline{\psi}, x_N^d)$, where $\Gamma_q(\psi, x_N^d)$ denotes the set from which $J_q(\psi, x_N^d)$ is selected. By checking every possible pair of $(\psi, x_N^d) \in \{A, T, G, C\} \times \{A, T, G, C\}$, we conclude that*

$$\Gamma_{q+1}(\overline{\psi}, x_N^d) \subset \Gamma_q(\psi, x_N^d),$$

therefore, $J_q(\psi, x_N^d) \leq J_{q+1}(\overline{\psi}, x_N^d)$.

In addition, for $q \leq N-7$,

$$\Gamma_q(\psi, x_N^d) = \Gamma_{q+1}(\overline{\psi}, x_N^d),$$

therefore, $J_q(\psi, x_N^d) = J_{q+1}(\overline{\psi}, x_N^d)$, *for* $q \leq N-7$. *This proves Claim 4.1, 4.2 and 4.3.*

Remark A.2. *The iterative equation* (A.4), *together with* (A.3), *prove Claim 4.3 and 4.4.*

Remark A.3. *For* $N - q = 0 \pmod 2$,

$$J_q(\psi_1, \psi_1) = 0, \forall \psi_1 \in \{A, T, G, C\};$$

For $N - q = 1 \pmod 2$,

$$J_q(\psi_1, \overline{\psi_1}) = 0, \forall \psi_1 \in \{A, T, G, C\},$$

since complementary transfer happens naturally without any mutagen cost.

Remark A.4. *We can generate at least one optimal path from every cost in* $\Gamma_q(\psi, x_N^d)$. *For instance,* $\alpha_{CT} + \alpha_{AC} + \alpha_{GA}, \alpha_{CT} + \alpha_{TG} + \alpha_{GA} \in \Gamma_{N-6}(C, T)$, *the optimal paths generate by this cost is listed in Table A.1.*

Cost	x_{N-6}	x_{N-5}	x_{N-4}	x_{N-3}	x_{N-2}	x_{N-1}	x_N
$\alpha_{CT} + \alpha_{AC} + \alpha_{GA}$	$C \xrightarrow[\alpha_{CT}]{u_{CT}}$	$T \rightarrow$	$A \xrightarrow[\alpha_{AC}]{u_{AC}}$	$C \rightarrow$	$G \xrightarrow[\alpha_{GA}]{u_{GA}}$	$A \rightarrow$	T
$\alpha_{CT} + \alpha_{TG} + \alpha_{GA}$	$C \xrightarrow[\alpha_{CT}]{u_{CT}}$	$T \xrightarrow[\alpha_{TG}]{u_{TG}}$	$G \xrightarrow[\alpha_{GA}]{u_{GA}}$	$A \rightarrow$	$T \rightarrow$	$A \rightarrow$	T
	$C \xrightarrow[\alpha_{CT}]{u_{CT}}$	$T \xrightarrow[\alpha_{TG}]{u_{TG}}$	$G \rightarrow$	$C \rightarrow$	$G \xrightarrow[\alpha_{GA}]{u_{GA}}$	$A \rightarrow$	T
	$C \xrightarrow[\alpha_{CT}]{u_{CT}}$	$T \rightarrow$	$A \rightarrow$	$T \xrightarrow[\alpha_{TG}]{u_{TG}}$	$G \xrightarrow[\alpha_{GA}]{u_{GA}}$	$A \rightarrow$	T
	$C \rightarrow$	$G \rightarrow$	$C \xrightarrow[\alpha_{CT}]{u_{CT}}$	$T \xrightarrow[\alpha_{TG}]{u_{TG}}$	$G \xrightarrow[\alpha_{GA}]{u_{GA}}$	$A \rightarrow$	T

Table A.1: Paths generated by two elements from $\Gamma_{N-6}(C, T)$.

References

L.M. Adleman. Molecular computation of solutions to combinatorial problems. *Science*, 266(5187):1021, 1994.

R. Bellman. On the theory of dynamic programming. *Proceedings of the National Academy of Sciences of the United States of America*, 38(8):716, 1952.

R.E. Bellman. *Dynamic programming*. Dover Pubns, 2003.

D.P. Bertsekas. *Dynamic Programming and Optimal Control, vol. 1, 2*. Athena Scientific, 1995. ISBN 1886529116.

J. Collado-Vides. A transformational-grammar approach to the study of the regulation of gene expression. *Journal of theoretical biology*, 136(4):403–425, 1989.

J. Collado-Vides, R.M. Gutièrrez-Ríos, and G. Bel-Enguix. Networks of transcriptional regulation encoded in a grammatical model. *BioSystems*, 47(1-2):103–118, 1998.

F. Crick. Central dogma of molecular biology. *Nature*, 227(5258):561–563, 1970.

F.H.C. Crick. On protein synthesis. *Symposium of the Society for Experimental Biology XII*, pages 139–163, 1958.

H. De Jong. Modeling and simulation of genetic regulatory systems: a literature review. *Journal of computational biology*, 9(1):67–103, 2002.

J. Feng and H.C. Tuckwell. Optimal control of neuronal activity. *Physical review letters*, 91(1):18101, 2003.

E.C. Friedberg, G.C. Walker, and W. Siede. *DNA repair and mutagenesis*. ASM press, 1995. ISBN 1555810888.

R. Gao, J. Yu, M. Zhang, T.J. Tarn, and J.S. Li. Systems theoretic analysis of the central dogma of molecular biology: Some recent results. *NanoBioscience, IEEE Transactions on*, 9(1):59–70, 2010.

L. Glass and S.A. Kauffman. The logical analysis of continuous, non-linear biochemical control networks. *Journal of Theoretical Biology*, 39(1):103–129, 1973.

D. Hristu-Varsakelis and W.S. Levine. *Handbook of networked and embedded control systems*. Birkhauser, 2005.

H. Kitano. Systems biology: a brief overview. *Science*, 295(5560):1662, 2002.

R.E. Koch. The influence of neighboring base pairs upon base-pair substitution mutation rates. *Proceedings of the National Academy of Sciences of the United States of America*, 68(4):773, 1971.

R. Langer. New methods of drug delivery. *Science*, 249(4976):1527, 1990.

R. Layek, A. Datta, M. Bittner, and E.R. Dougherty. Cancer therapy design based on pathway logic. *Bioinformatics*, 27(4):548, 2011.

U. Ledzewicz and H. Schaettler. Singular controls and chattering arcs in optimal control problems arising in biomedicine. *Control and Cybernetics*, 38(4), 2009.

U. Ledzewicz, J. Marriott, H. Maurer, and H. Sch
"attler. Realizable protocols for optimal administration of drugs in mathematical models for anti-angiogenic treatment. *Mathematical Medicine and Biology*, 27(2): 157, 2010a.

U. Ledzewicz, H. Maurer, and H. Sch
"attler. Minimizing tumor volume for a mathematical model of anti-angiogenesis with linear pharmacokinetics. *Recent Advances in Optimization and its Applications in Engineering*, pages 267–276, 2010b.

I. Lentacker, B. Geers, J. Demeester, S.C. De Smedt, and N.N. Sanders. Design and evaluation of doxorubicin-containing microbubbles for ultrasound-triggered doxorubicin delivery: cytotoxicity and mechanisms involved. *Molecular Therapy*, 18(1): 101–108, 2009.

H. Matsuno, A. Doi, M. Nagasaki, and S. Miyano. Hybrid petri net representation of gene regulatory network. In *Pacific Symposium on Biocomputing*, volume 5, page 87, 2000.

A.E. Mayo, Y. Setty, S. Shavit, A. Zaslaver, and U. Alon. Plasticity of the cis-regulatory input function of a gene. *PLoS biology*, 4(4):e45, 2006.

S.D. McCulloch and T.A. Kunkel. The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases. *Cell Research*, 18(1):148–161, 2008. ISSN 1001-0602.

J. Moehlis, E. Shea-Brown, and H. Rabitz. Optimal inputs for phase models of spiking neurons. *Journal of computational and nonlinear dynamics*, 1:358, 2006.

E.M. Ozbudak, M. Thattai, H.N. Lim, B.I. Shraiman, and A. Van Oudenaarden. Multistability in the lactose utilization network of escherichia coli. *Nature*, 427 (6976):737–740, 2004.

A. Regev, W. Silverman, and E. Shapiro. Representation and simulation of biochemical processes using the-calculus process algebra. In *Pacific symposium on biocomputing*, volume 6, pages 459–470, 2001.

V. Reinke. Germline genomics. *WormBook*, 2006.

T.R. Robinson. *Genetics for dummies*. Wiley Publishing, Inc., 2005. ISBN 0764595547.

M. Santillán and M.C. Mackey. Influence of catabolite repression and inducer exclusion on the bistable behavior of the lac operon. *Biophysical journal*, 86(3): 1282–1292, 2004.

Y. Setty, AE Mayo, MG Surette, and U. Alon. Detailed map of a cis-regulatory input function. *Proceedings of the National Academy of Sciences of the United States of America*, 100(13):7702, 2003.

T. Strachan and A.P. Read. *Human molecular genetics 3*. Garland Science, 2004. ISBN 9780815341840.

R.J. Tanaka and H. Kimura. Mathematical classification of regulatory logics for compound environmental changes. *Journal of theoretical biology*, 251(2):363–379, 2008.

R.J. Tanaka, H. Okano, and H. Kimura. Mathematical description of gene regulatory units. *Biophysical journal*, 91(4):1235–1247, 2006.

R. Thomas. Boolean formalization of genetic control circuits* 1. *Journal of Theoretical Biology*, 42(3):563–585, 1973.

J.D. Watson and F.H.C. Crick. Molecular structure of nucleic acids. *Nature*, 171 (4356):737–738, 1953.

J.D. Watson and F.H.C Crick. A structure for deoxyribose nucleic acid. *A century of Nature: twenty-one discoveries that changed science and the world*, page 82, 2003.

N. Wiener. *Cybernetics*. J. Wiley, 1948.

R. Yang, T.J. Tarn, and M. Zhang. Data-driven feedforward control for electroporation mediated gene delivery in gene therapy. *Control Systems Technology, IEEE Transactions on*, 18(4):935–943, 2010.

N. Yildirim and M.C. Mackey. Feedback regulation in the lactose operon: a mathematical modeling study and comparison with experimental data. *Biophysical Journal*, 84(5):2841–2851, 2003.

M. Zhang, M.X. Cheng, and T.J. Tarn. A mathematical formulation of DNA computation. *NanoBioscience, IEEE Transactions on*, 5(1):32–40, 2006. ISSN 1536-1241.

Vita

Juanyi Yu

Date of Birth April 20, 1984

Place of Birth Zhejiang, China

Degrees Ph.D. Washington University in St. Louis, Electrical Engineering, December 2011
M.S. Washington University in St. Louis, Systems Science and Mathematics, August 2008
B.E. The Chinese University of Hong Kong, Hong Kong, Information Engineering, July 2006

Professional Societies IEEE Student Member
IEEE Robotics and Automation Society

Publications Juanyi Yu, Jr-Shin Li and Tyzh-Jong Tarn (2011). Optimal Control in Molecular-level Gene Manipulation. *Proceeding of the* 9^{th} *World Congress on Intelligent Control and Automation (WCICA 2012)*, Beijing, China. Submitted.

Juanyi Yu, Jr-Shin Li and Tyzh-Jong Tarn (2011). Optimal Control in Gene Mutation in DNA Replication. *Journal of Biomedicine and Biotechnology, Applications of Synthetic Biology in Microbial Biotechnology 2011.* Preprint.

Rui Gao, Juanyi Yu, Mingjun Zhang, Tzyh-Jong Tarn and Jr-Shin Li (2010). Systems Theoretic Analysis of the Central Dogma of Molecular Biology: Some Recent Results. *IEEE Transactions on NanoBioscience,* Volume 9, Issue 1, pp. 59-70.

Rui Gao, Juanyi Yu, Mingjun Zhang and Tzyh-Jong Tarn (2009). A Preliminary Study on Mathematical Modeling of Protein Synthesis Process. *Proceedings of IEEE International*

Conference on Intelligent Computing and Intelligent Systems (ICIS 2009), Shanghai, China.

Rui Gao, Juanyi Yu, Mingjun Zhang and Tzyh-Jong Tarn (2009). Mathematical Models of Protein Secondary Structures and Gene Mutation. *Proceedings of International Conference on Mechatronics and Automation (ICMA 2009),* Changchun, China.

Rui Gao, Juanyi Yu, Mingjun Zhang, Tzyh-Jong Tarn and Jr-Shin Li. A Mathematical Formulation of the Central Dogma of Molecular Biology. *Nanomedicine: A system Engineering Approach,* Mingjun Zhang, Ning Xi (Eds.), Pan Stanford Publishing Pte. Ltd.

December 2011

Optimal Control in Gene Mutation, Yu, Ph.D. 2011