Carnegie Mellon University

CARNEGIE INSTITUTE OF TECHNOLOGY

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF Doctor of Philosophy

TITLE Biophysical Modeling and Optimal Control of DNA Amplification

PRESENTED BY Karthikeyan Marimuthu

ACCEPTED BY THE DEPARTMENT OF

Chemical Engineering

RAJ CHAKRABARTI	4/25/14
ADVISOR	DATE
LORENZ BIEGLER	4/25/14
DEPARTMENT HEAD	DATE

APPROVED BY THE COLLEGE COUNCIL

VIJAYAKUMAR BHAGAVATULA	4/25/14
DEAN	DATE

Biophysical Modeling and Optimal Control of DNA Amplification

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Chemical Engineering

Karthikeyan Marimuthu

B.Tech., Chemical Engineering, A.C.Tech, Anna University

M.Tech., Chemical Engineering, Indian Institute of Technology, Madras

Carnegie Mellon University

Pittsburgh, PA

May, 2014

In the memory of

my father, A. Marimuthu, my brother, M. Boobalan

my grandfather, A. Annamalai and my uncle, A. Kandasamy

To

my mom, Saraswathi, my brothers, Saravanan, and Gopal,

my aunt, Sangunthala and my cousins, Mohana Madhavan, and Hemalatha.

Acknowledgements

I am extremely grateful to have worked with my advisor Prof. Raj Chakrabarti for my PhD thesis. I thank him for his guidance and support for the last five years. Apart from research, he has morally supported me when I transferred from Purdue to Carnegie Mellon University with him. I consider him as an example for being generous and professional.

I would like to thank my committee members, Profs. Lorenz Biegler, Nikolaos Sahinidis and Robert Tilton for their valuable suggestions and feedback.

I greatly acknowledge the financial support from the School of Chemical Engineering, Purdue University and the Department of Chemical Engineering, Carnegie Mellon University.

I sincerely thank our research group members, Andy Koswara, Bradley J Ridder, Vaibhav Bhutoria, Amit Sabne and Aravind Chiruvelli for many exciting research discussions.

I am indebted to Prof. Shankar Narasimhan, IIT Madras who has inspired and motivated me to pursue my doctoral research in the USA. I always consider that I am very fortunate to have met him in my life and worked with

him for my master's thesis. I would like to thank Dr. Niket Kaisare, Principal scientist, ABB corporate research, Dr. Skand Saksena, R&D Director for Water (Global) and Head, R&D Operations (Mumbai Site) at Hindustan Unilever Limited, and Prof. Venkat Venkatasubramanian, Columbia University, for their advice and motivation that helped me to pursue my doctoral research.

I began my doctoral research at Purdue University where I completed my course work requirements. Prof. David Corti's lectures on Advanced Thermodynamics created special interest on this subject and he is one of the best teachers from whom I learned science. I would like to thank him for his awesome lectures.

I would like to thank Ashwin Ravichandran, Aravind Chandrasekaran, Athinthra Sethurajan Krishnaswamy, Indreesh Badrinarayanan and Maya Subramanian, undergraduate students from A.C.Tech, Anna University, who worked with me on the development of the educational softwares.

I sincerely thank my high school Mathematics teachers, A. Nagarajan, T.S. Poomalai and M.P. Kavitha for their inspirational teaching and motivation which I believe help me to solve mathematical problems in my research even today.

My undergraduate research work with Profs. M. Velan and K. Muthukumar at A.C.Tech, Anna University, motivated me to pursue higher studies. I sincerely thank them for providing this opportunity. I am extremely fortunate to have got mentors, Dr. Hariprasad Subramani, Flow assurance engineer at

Chevron corporation and Prof. A. Arunagiri, National Institute of Technology, Trichy, India. I am thankful to them for their guidance and support throughout my educational career.

Being an international student, my family here is my friends, Dhariya Mehta, Harshavardhan Choudhari and Dr. Krishnaraj Sambath (Graduate students at Purdue University). Their constant support during tough times made me to continue my research comfortably. They are certainly, an example of the proverb, "A friend in need is friend indeed". I thank my other apartment-mates Jayachandran Devaraj and Sudarshan Narayanan for their care and support. I would like to thank my other Purdue University friends, Kaiwalya Sabnis, Sumeet Thete, Dr. Anand Venkatesan, Vinod Kumar Venkatakrishnan, Dr. Dharik Sanchan, Dr. Santosh Appathurai, Dr. Anirudh Shenvi, Atish Parekh, Aniruddha Kelkar, Shankali Pradhan, Anuj Verma, Gautham M. R., and Anshu Gupta .

I am grateful to my friends A. Balaguru, D.Shanmuga Priya, Uma Madhavan, T. Dhanalakshmi and T. Aarthi who have supported and motivated me to pursue my higher studies. I am extremely thankful to my friend Saranya Ganapathy who has been morally supporting me through my PhD. I thank Priyanandini Balasubramanian, Viswanath Balasubramaniyan, and Rajalakshmi Vengadasamy, my colleagues at General Electric, for their help and support especially during the initial stage of my PhD.

I am indebted to my mother, Saraswathi Marimuthu, brothers, Saravanan Marimuthu and Gopalakrishnan Marimuthu, sisters in law, Kalaivani and

Prema Priya, cousins, Mohana Madhavan and Hemalatha, nieces, Mohitha and Sivarshini and nephew Kumaran, for their unconditional love, care and support.

Abstract

DNA amplification or the Polymerase Chain Reactions (PCR) is the workhorse of nearly every modern molecular biology laboratory, as well as the burgeoning discipline of personalized medicine. Despite the apparent simplicity of the PCR reaction, the method is often fraught with difficulties that can decrease the cycle efficiency or result in competitive amplification of undesired side products. The focus of this work is to derive an optimal reaction condition for a given PCR using the engineering discipline of control theory that can automatically derive prescriptions for the optimal temperature cycling protocols of a PCR reaction, if a suitable kinetic model exists.

We first developed a theoretical model to estimate the sequence and temperature dependent rate parameters of a oligonucleotide hybridization or annealing reaction. Rate constants that were estimated using our model is in good agreement with the experimentally estimated rate constants of the same oligonucleotide hybridization reaction. Using the theory of enzyme processivity the kinetic parameters of enzyme binding and extension reactions were estimated experimentally. Thus, a first sequence-dependent biophysical model

for DNA amplifications has been developed. It is shown that amplification efficiency is affected by dynamic processes that are not accurately represented in simplified models of DNA amplification that are the basis of conventional temperature cycling protocols. Based on this analysis; a modified temperature protocol that improves the PCR efficiency is suggested. Use of this sequence-dependent kinetic model in a control theoretic framework to determine the optimal dynamic operating conditions of DNA amplification reactions, for any specified amplification objective, is discussed.

Using these control systems, we demonstrate that there exists an optimal temperature cycling strategy for geometric amplification of any DNA sequence and formulate optimal control problems that can be used to derive the optimal temperature profile. Strategies for the optimal synthesis of the DNA amplification control trajectory are proposed. Analogous methods can be used to formulate control problems for more advanced amplification objectives corresponding to the design of new types of DNA amplification reactions. Finally, a PCR optimal control problem is solved and an optimal temperature profile that maximizes the desired DNA concentration as well as minimizes the total reaction is obtained.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

DNA amplification is the process of geometric growth of the number of double-stranded DNA (dsDNA) molecules in solution through repeated replication of single-stranded DNA (ssDNA) templates. Due to the universal need to amplify DNA for applications ranging from molecular cloning to DNA sequencing, such methods have arguably become the central technology of modern molecular biology. The polymerase chain reaction (PCR), the most common DNA amplification reaction, is a cyclic amplification process that can produce millions of copies of double-stranded DNA molecules starting from a single molecule. The traditional three-step PCR reaction cycle consists of

1. Denaturation of double stranded DNA into single-stranded DNAs.

2. Oligonucleotide primer annealing to the resulting ssDNAs.

3. Polymerase-mediated extension steps to produce two dsDNA molecules.

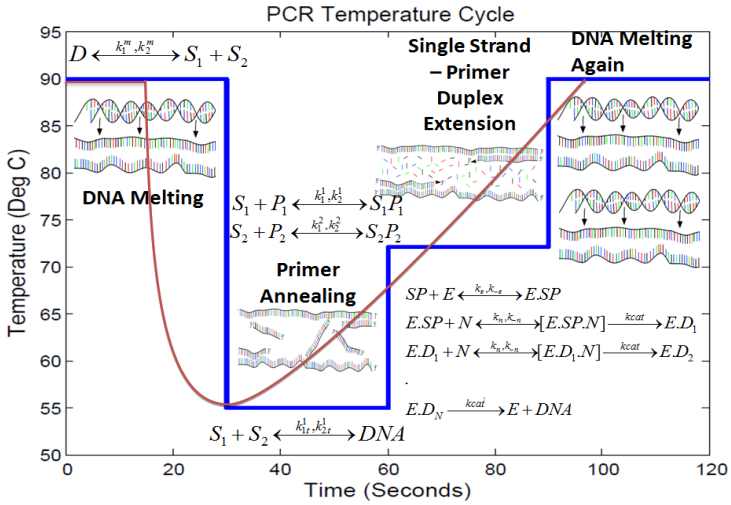


Figure 1.1: Schematic of PCR. Temperature vs Time profile of a typical PCR is represented in blue line.

This cycle is repeated 20-30 times, resulting in geometric growth of the number of DNA molecules. The base of the exponent for geometric growth is termed the amplification efficiency of a cycle. Fig. 1.1 shows the typical PCR temperature profile.

1.1 Background

Despite the fact that the notion of thermal cycling is based on a dynamic picture of DNA amplification, there are currently no models of DNA amplification kinetics that are capable of predicting the evolution of reaction products for general sequences and operating conditions. Without such a model, the optimal temperature cycling protocol for the reaction which is sequence specific - cannot be computed, and reductions in cycle efficiency (either through decreased reaction yield or specificity compared to the theoretical maximum

values) can occur. Due to geometric growth, reductions in the cycle efficiency can result in dramatically diminished efficiency of the overall reaction, and substantial efforts have hence been dedicated to improving the efficiency of DNA amplification reactions (Chakrabarti and Schutt (2001) ,Chakrabarti and Schutt (2001)).

In the absence of predictive models for DNA amplification, the operating conditions for PCR reactions are typically selected based on qualitative analysis of their kinetics and thermodynamics, given the desired amplification objective. Over the past two decades, many variants of DNA amplification have been invented based on the notions of DNA denaturation, annealing and polymerization, each tailored to a particular amplification objective. Each such reaction (which is typically assigned its own acronym) is based on a temperature cycling protocol determined through analysis of reaction thermodynamics and a qualitative analysis of kinetics. A simple example of a temperature cycling protocol that involves modifications to the conventional prescription is the use of two-step PCR cycles (Skladny et al. (1999)), wherein annealing and extension occur simultaneously at a properly chosen temperature.

1.2 Systems Biology Approach

A general approach to kinetic modeling of DNA amplification has applications to the design of new types of amplification reactions, in addition to enhancement of existing reactions. In the language of systems engineering,

the selection of the optimal trajectory of a manipulated input variable such as temperature is referred to as dynamic optimization or optimal control (Stengel (1994)). This work is concerned with the establishment of a foundation for the dynamic optimization of DNA amplification reactions, which can be used for the automated computation (rather than qualitative selection) of temperature cycling protocols. To date, quantitative sequence-dependent modeling of DNA amplification has been largely restricted to the thermodynamics of the reaction. Prior reports of kinetic models for PCR have proven inadequate for the purposes of dynamic PCR optimization. For example, Rychlik et al. (1990) developed an empirical equation to determine an optimal annealing temperature that maximizes the final DNA concentration. Fig. 1.2 illustrates the different types of PCR optimization. Using a probabilistic PCR kinetic model, Stolovitzky and Cecchi (1996) developed a method to calculate the cycle efficiency for PCR quantification. Velikanov and Kapral (1999) proposed a Markov process approach to optimize the extension step of PCR . Yang et al. (2005) discussed the effect of annealing temperature on the concentration of different targets in a multiplex reaction and gave the temperature vs. concentration profile for all the targets. Though the above-developed approaches predict the PCR efficiency, they have several fundamental limitations. For example, Rychlik et al. (1990)'s model does not have a theoretical foundation for prediction of the optimal annealing temperature and their empirical correlation is purely based on the limited number of experiments. Stolovitzky and Cecchi (1996) and Velikanov and Kapral (1999)'s, kinetic model did not

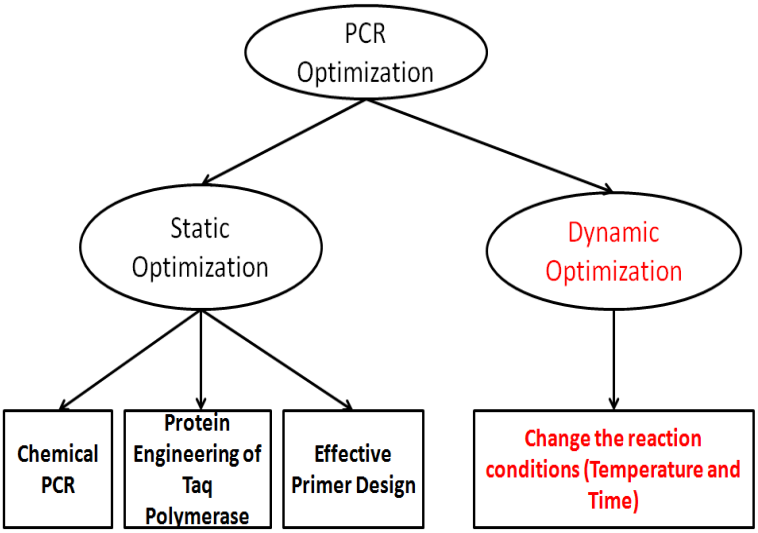


Figure 1.2: Types of PCR optimization

account for the sequence dependence of amplification kinetics or were limited to a single step of the reaction.

A so-called state space model is required for dynamic optimization of DNA amplification. State space models are systems of differential equations that, when solved, describe the dynamics of the system, along with algebraic constraints and specified parameters (e.g., rate parameters such as activation energies and pre-exponential factors) whose values are either predicted based on first-principles theory, independently measured in off-line experiments, or indirectly estimated through on-line measurement of observable quantities during the evolution of the system. Across the published literature, proposed state space models (Mehra and Hu (2005); Hsu et al. (1997); Gevertz et al. (2006); Lee et al. (2006); Griep et al. (2006)) have been incapable of predicting PCR amplification efficiencies. This is because no generalization has been made regarding the dependence of kinetic parameters on both i) the DNA sequence

and ii) temperature. None of these kinetic models are both sequence- and temperature-dependent. It is quite evident from the nearest neighbor method which can be used to calculate the DNA annealing reaction free energy equilibrium constant of DNA hybridization is temperature and sequence-dependent. Data and LiCata (2003) reported temperature-dependent equilibrium dissociation constants for the enzyme binding reaction. Hung et al. (1992) and Innis et al. (1988) reported temperature-dependent enzyme extension reaction rates. Therefore, the kinetic parameters of the three steps of PCR are highly dependent on the sequence composition and temperature of the reaction. Accurate and computationally efficient sequence-dependent state space models, which are essential to solve such problems, require a combination of fundamental biophysical modeling with dynamical systems theory. This notion of sequence-dependent modeling of the kinetics of biochemical reaction networks, which has various applications in dynamical systems biology, is introduced here as one of the contributions of the present work.

1.3 Research Objectives

In this work, we develop the first sequence-dependent kinetic model for PCR reactions that is suitable for engineering control, validating this state space model through comparison to experimental data. First principles models are essential for proper prediction of DNA kinetic rate parameters for any arbitrary DNA sequence. The model introduced herein is based on quantitative

biophysical modeling of DNA melting, annealing, and polymerization, which together enable a mapping of a given DNA sequence and polymerase enzyme onto temperature-dependent kinetic rate constants for the DNA amplification reaction. Such sequence-dependent modeling of amplification kinetics has been enabled, based on the theory of relaxation kinetics. One benefit of such complete state space models for PCR is the ability to achieve similar or enhanced amplification efficiencies or specificities in greatly reduced time, through the exploitation of dynamic processes such as simultaneous annealing and extension - that are not represented in simplified models of DNA amplification, but which play a major role in determining the evolution of chemical species. Prospects for the application of these sequence-dependent models in the formulation of optimal control problems that can enable the computation of optimal cycling strategies, for any specified objective, are discussed. Further, we formulate a PCR optimal control problem and solve it to obtain the optimal reaction conditions. Fig. 1.3 illustrates our systems biology approach for PCR optimization.

1.4 Thesis Outline

This thesis is organized as follows

1. In Chapter 2, a theoretical approach to the prediction of the sequence and temperature-dependent rate constants for oligonucleotide hybridization reactions has been developed based on the theory of relaxation kinet-

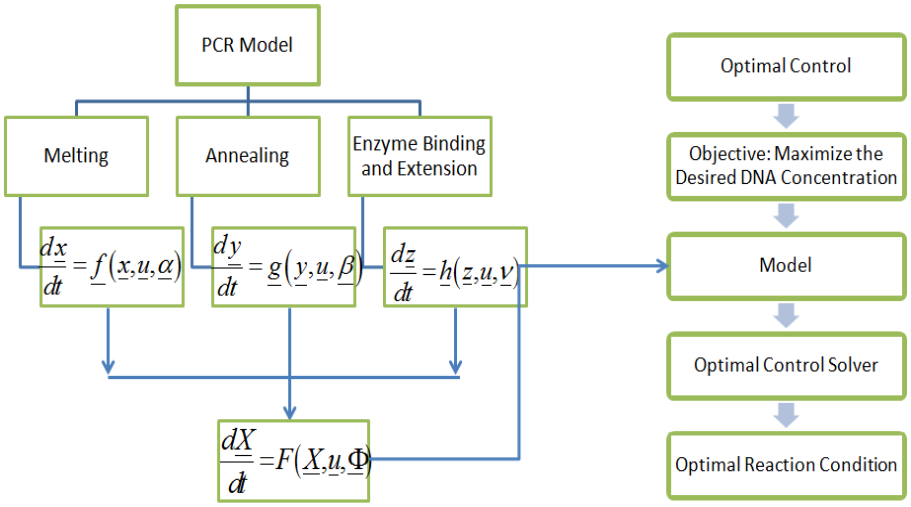


Figure 1.3: Systems Biology Approach for PCR optimization

ics. The model-predicted rate constants are compared to experimentally determined rate constants for the same oligonucleotide hybridization reactions.

2. In Chapter 3, a theoretical framework for prediction of the dynamic evolution of chemical species in DNA amplification reactions, for any specified sequence and operating conditions, is reported. Using the Polymerase Chain Reaction (PCR) as an example, we developed a sequence- and temperature-dependent kinetic model for DNA amplification using first principles biophysical modeling of DNA hybridization and polymerization.

3. In Chapter 4, a theoretical framework for determination of the optimal dynamic operating conditions of DNA amplification reactions, for any specified amplification objective, has been developed based on first-principles biophysical modeling and control theory.

4. In Chapter 5, a numerical approach to solve a PCR optimal control problem is provided and it is implemented to obtain an optimal temperature profile. A strategy to solve a time optimal control problem for PCR is discussed along with its results.

5. In Chapter 6, a summary of the main results and future direction of this work is provided.

Chapter 2

Sequence-Dependent Oligonucleotide Hybridization Kinetics

2.1 Introduction

Oligonucleotide hybridization is a reversible chemical reaction in which two short single-stranded DNA/RNA molecules (generally < 50 base pairs) hybridize to give a double-stranded DNA/RNA molecule. This reaction is an integral part of a wide range of molecular biology technologies ranging from microarrays to the Polymerase Chain Reaction (PCR) (Fish et al. (2007); Horne et al. (2006); Hadiwikarta et al. (2012)). Ever since these techniques were invented, DNA hybridization reactions have been studied extensively (Hadiwikarta et al. (2012)). In particular, the thermodynamics of DNA hy-

bridization reactions has been well-studied both experimentally and theoretically (Chalikian et al. (1999); SantaLucia Jr and Hicks (2004); SantaLucia (1998); Breslauer (1994); Breslauer et al. (1986); Garel and Orland (2004)). However, kinetic study of DNA hybridization has received only limited attention, with the majority of prior work being focused on homopolymers. For applications such as PCR that are time-dependent, the kinetics of DNA hybridization reactions can play an important role. Methods for the prediction of sequence-dependent rate constants could enable, for example, the optimal choice of temperature cycling protocols for PCR. In the late 1960s and early 1970s, a few experimental methods for the estimation of kinetic parameters of oligonucleotide hybridization reactions were proposed and using these, the rate constants of a series of A_nU_n oligonucleotides were estimated at a specified set of temperatures (Craig et al. (1971); Breslauer and Bina-Stein (1977); Pörschke and Eigen (1971)). Though these studies addressed the kinetics of a few homogeneous oligonucleotide hybridization reactions, there was no generalization made to the rate constants of arbitrary homogeneous/heterogeneous oligonucleotide hybridization reactions at specific temperatures. More recent work on modeling hybridization kinetics (Wang and Drlica (2003)) was based on simplified models that, due to the nature of the applications, did not demand physical consistency with hybridization thermodynamics.

With methodologies such as PCR being applied to different oligonucleotide sequences every day in modern molecular biology laboratories for various applications (Pfaffl (2001)), it is not feasible to experimentally estimate the ther-

modynamic and kinetic parameters for all possible DNA/RNA sequences of interest. This is one of the main reasons for the development of the Nearest Neighbor (NN) method, which can be used to calculate the thermodynamic properties of the oligonucleotide hybridization reactions (SantaLucia Jr and Hicks (2004); SantaLucia (1998); Breslauer et al. (1986)). However, thermodynamic properties are insufficient to analyze and optimize reactions such as PCR. Therefore, it is essential to develop a theoretical model to estimate the rate parameters for oligonucleotide hybridization reactions.

Although numerous experimental studies of the kinetics of oligonucleotide hybridization have been reported (Gao et al. (2006); Bloomfield et al. (2000); Chen et al. (2007); Bonnet et al. (1998)), to the best of our knowledge, no theoretical methods for estimation of the rate constants of arbitrary oligonucleotide sequences based on the fundamental biophysics of DNA hybridization have been developed to date. In this work we have developed the first theoretical method to predict sequence- and temperature-dependent rate parameters that is applicable to either homogeneous or heterogeneous oligonucleotide hybridization reactions.

Experimentally, the rate constants of an oligonucleotide hybridization reaction are determined by measuring its a) relaxation time using a temperature jump apparatus, b) equilibrium melting curve, a relationship between the equilibrium conversion and temperature of the DNA hybridization reaction (Craig et al. (1971); Breslauer and Bina-Stein (1977); Pörschke and Eigen (1971)). The relaxation time of an equilibrium reaction is a measure of time needed

for the reaction to return to equilibrium when it is perturbed from its original equilibrium state. Theoretical estimation of hybridization rate constants is based on estimation of the relaxation time and the equilibrium melting curve. The equilibrium melting curve of an oligonucleotide sequence is determined theoretically using the NN model. In this work we have developed a theoretical model to estimate the relaxation time of any oligonucleotide hybridization reactions at a specified temperature. We have combined our model for relaxation time prediction with the NN model to estimate the hybridization rate constants of arbitrary oligonucleotide sequences. In order to develop a theoretical model for relaxation time prediction, we have considered two different hybridization reaction mechanisms: one-sided melting (Wetmur and Davidson (1968); Applequist and Damle (1965); Pörschke and Eigen (1971); Craig et al. (1971)) and two-sided melting (Schwarz Jr and Poland (1975, 1976); Anshelevich et al. (1984)). We have analyzed these two reaction mechanisms and proposed a modified reaction mechanism that is consistent with the laws of chemical reaction kinetics. Our kinetic model can be used not only to estimate appropriate reactions conditions for DNA hybridization, but also for systematic optimization and control of PCR reactions.

In this chapter we

- Describe our theoretical approach to estimate the temperature and sequence-dependent rate parameters of a oligonucleotide hybridization
- Develop a theoretical method to estimate the relaxation times of oligonu-

cleotide hybridization reactions;

- Estimate the model parameters and experimentally validate our theoretical model;
- Analyze our model and discuss the properties of its solutions;
- Consider illustrative applications of kinetically controlled DNA hybridization reactions.

2.2 Theoretical Model for Annealing Kinetics

The forward and reverse rate constants of an equilibrium reaction can be expressed as a function of its equilibrium constant (K_{eq}) and relaxation time (τ) (Espenson (1995)) as shown below.

2.2.1 Prediction of Equilibrium Properties

Consider the following oligonucleotide hybridization reaction

$$S_1 + S_2 \underset{k_r}{\overset{k_f}{\rightleftharpoons}} D \qquad (R_1)$$

Where k_f and k_r are forward and backward reaction rate constants. D and S represent the double stranded and single stranded oligonucleotide. The hybridization reaction, (R_1), is assumed to follow 'all or none' or 'two state' hybridization (Chalikian et al. (1999); SantaLucia Jr and Hicks (2004); San-

taLucia (1998); Breslauer (1994); Breslauer et al. (1986); Garel and Orland (2004); Craig et al. (1971); Breslauer and Bina-Stein (1977); Pörschke and Eigen (1971)). Typically a DNA with less than 50 base pairs follows two state hybridization (Koehler and Peyret (2005)). Reaction (R_1) can be assumed to be an elementary reaction with respect to the fully single-stranded and double-stranded DNA (Craig et al. (1971); Breslauer and Bina-Stein (1977); Pörschke and Eigen (1971)); hence the rate expression is written

$$\frac{d[D]}{dt} = k_f[S_1][S_2] - k_r[D], \tag{2.1}$$

where [ ] represents the molar concentration of respective species. Since usually the DNA hybridization reaction mixture is dilute, the following relationship between the equilibrium concentration, equilibrium constant, forward and backward rate constants is written (Smith et al. (2005))

$$K_{eq}(T) = exp\left(\frac{-\Delta G(T)}{RT}\right) \tag{2.2a}$$

$$= \frac{\left({}^{[D]}/_{[C_0]}\right)}{\left({}^{[S_1]}/_{[C_0]}\right)\left({}^{[S_2]}/_{[C_0]}\right)} \tag{2.2b}$$

$$= \frac{k_f}{k_r[C_0]} \tag{2.2c}$$

where K_{eq} is an equilibrium constant, C_0 is a reference concentration whose value is 1M (Smith et al. (2005); Jost and Everaers (2009)) and ΔG is the Gibbs free energy of reaction (R_1) estimated based on the NN model. (SantaLucia (1998); SantaLucia Jr and Hicks (2004); Koehler and Peyret (2005)). We use

unified NN model parameters to estimate ΔG of reaction (R_1) (SantaLucia (1998)). In Section 2.2.3 we have explained about the NN model.

2.2.2 Relationship between Relaxation time and Rate constants

As per the definition of the relaxation time, when the reaction (R_1) is disturbed from its equilibrium, the following equation is written.

$$\frac{d[D_{eq}+\delta]}{dt} = k_f[S_{1eq}-\delta][S_{2eq}-\delta] - k_r[D_{eq}+\delta] \tag{2.3}$$

δ is a small perturbation from the equilibrium due to an external disturbance. Expanding Eq. (2.3) both sides and neglecting δ^2 term, the following rate equation for the perturbation parameter δ is obtained

$$\frac{d[\delta]}{dt} = -[\delta](k_f[S_{2eq}+S_{1eq}]+k_r) \tag{2.4}$$

Above equation represents the first order kinetics of the perturbation parameter δ with a characteristic time constant which is called relaxation time. Thus, an expression for the relaxation time is obtained as follows

$$\tau = \frac{1}{(k_f[S_{2eq}+S_{1eq}]+k_r)} \tag{2.5}$$

If the left hand side of Eq. (2.2) and (2.5) is known, then these equations can be solved to obtain k_f and k_r. A detailed derivation of the expression that relates relaxation time and equilibrium concentration for a self complementary and non-self complementary sequences have been provided in Appendix F.

2.2.3 Nearest Neighbor Model

Nearest Neighbor parameters are a set of enthalpy and entropy parameters for all possible neighboring base pairs in a DNA. The overall enthalpy and entropy of formation of a double stranded DNA is expressed as a sum of enthalpy and entropy of the pairs of neighboring base pairs. From the enthalpy and entropy of formation, Gibbs free energy of a DNA hybridization (or DNA formation) reaction is calculated as follows.

$$\Delta G\left(T\right) = \Delta H - T\Delta S \tag{2.6}$$

We use the unified NN parameters (SantaLucia (1998)) to predict ΔH and ΔS of a oligonucleotide hybridization. Thermodynamic properties of the Oligonucleotides of length up to 50 base pairs are predicted using NN parameters (Koehler and Peyret (2005)). The accuracy of the prediction increases when the length of a sequence decreases. For sequence length from 4 to 16 base pairs, the melting point is predicted with an average deviation of 1.6 ^{0}C (SantaLucia Jr and Hicks (2004)) and for a sequence length from 6 to 24 base pairs the melting point is predicted with an average deviation of 2.3 ^{0}C (SantaLucia Jr

and Hicks (2004)). These predictions are considered to be the best among all other available methods that predict DNA melting thermodynamic properties (SantaLucia (1998)). In Table B.1 NN parameters have been given and in Appendix B, an example is presented to explain how to calculate ΔG for a oligonucleotide hybridization reaction using NN method. All the NN parameters are estimated at a salt concentration of $1M$. Usually, for example in PCR, the salt concentration is around $50mM$. Therefore these NN parameters need to be corrected for a specific salt concentration. Owczarzy *et al* presented a set of empirical equations to correct the NN parameters for a specific KCl or $NaCl$ and $MgCl_2$ salt concentration (Owczarzy et al. (2008, 2004)). These equations have been provided in Appendix C.1 and D. Though usually it is assumed that ΔH and ΔS of reaction R_1 are independent of temperature, Rouzina and Bloomfield (Rouzina and Bloomfield (1999)) reported the effect of change in heat capacity with respect temperature on the estimation of δH of a hybridization reaction.

2.3 Theoretical Estimation of Relaxation Time: Unimolecular Elementary Step Model

The relaxation time of a reaction can be determined if its elementary steps are known. In this section we consider two reaction mechanisms for reaction (R_1) to estimate the rate of the elementary steps and then estimate the relaxation

time of reaction (R_1).

2.3.1 Nucleation of Oligonucleotide Hybridization

The kinetics of reaction (R_1) depends on the following two processes:

- A pre-equilibrium step that forms the first few base pairs to create a stable nucleus (Wetmur and Davidson (1968); Craig et al. (1971); Pörschke and Eigen (1971)).
- Once a stable nucleus is formed, rapid addition of base pairs to complete the hybridization reaction.

In a DNA hybridization reaction, each base pair has a specific stability (stacking + base pairing free energy). The following equations express the stability constants of AT and GC base pairs as functions of temperature (Wetmur and Davidson (1968)):

$$s_{AT}(T) = 1.04 \times 10^{-5} exp\left(\frac{8000}{RT}\right) \tag{2.7}$$

$$s_{GC}(T) = 1.04 \times 10^{-5} exp\left(\frac{8935}{RT}\right) \tag{2.8}$$

Unlike NN models, these relations approximate the stabilities of different types of base pairs without specification of the identities of the neighboring base pairs. The formation of the first or first few base pairs is quite different from

the formation of other base pairs (Wetmur and Davidson (1968); Craig et al. (1971); Pörschke and Eigen (1971)). Once the first base pair is formed, in order for the second base pair to form, the single strands must align in a specific manner. The strands can easily dissociate until a critical number of base pairs (termed a stable "nucleus") has formed. The stability of a duplex with one or few base pairs is thus smaller than the product of their stability constants. In order to represent the reduction in stability of such duplexes, a parameter σ (different literature uses different notations and we use σ here) is multiplied (Wetmur and Davidson (1968); Craig et al. (1971); Pörschke and Eigen (1971)) with the product of stability constants. Thus, σ accounts for the resistance to the formation of first few base pairs and hence the following reaction mechanism for a oligomer hybridization reaction was proposed (Wetmur and Davidson (1968); Craig et al. (1971); Pörschke and Eigen (1971)):

$$S_1 + S_2 \underset{k_{-0}}{\overset{\sigma_1 k_0}{\rightleftharpoons}} D_1 \underset{k_{-1}}{\overset{\sigma_2 k_1}{\rightleftharpoons}} \ldots\ldots\ldots\ldots \underset{k_{-(N-2)}}{\overset{\sigma_{N-2} k_{N-2}}{\rightleftharpoons}} D_{N-1} \underset{k_{-(N-1)}}{\overset{\sigma_{N-1} k_{N-1}}{\rightleftharpoons}} D_N \qquad (R_2)$$

In the above reaction, N denotes the total number of base pairs and D_i is a hybridized duplex with i base pairs in the double-stranded state. k_i and k_{-i} are the forward and reverse rate constants of the i^{th} step of (R_2). At a fixed temperature, k_i is roughly constant for sequences of a given length and it can be assumed to be the same for all the steps; hence we denote it by k_1 (Wetmur and Davidson (1968); Suyama and Wada (1984)). The ratio of k_1 and k_{-i} is

defined as a stability constant, s_i of i^{th} base pair.

$$s_i = \frac{k_1}{k_{-i}} \tag{2.9}$$

As we explained above, once a stable nucleus is formed, it facilitates subsequent base pair formation. Let x be the number of base pairs that melt before forming a stable nucleus; then

$$\lim_{i \to x} \sigma_i = 1 \implies \prod_{i=x}^{N-1} \sigma_i (T) = 1 \tag{2.10}$$

The overall stability constant of reaction (R_1) in terms of individual stability constants of base pairs and σ_i is expressed as

$$K_{eq}(T) = exp\left(\frac{-\Delta G(T)}{RT}\right) = \prod_{i=0}^{N-1} \sigma_i (T)\, s_i (T) \tag{2.11}$$

Substituting Eq. (2.10) in Eq. (2.11), we obtain

$$K_{eq}(T) = \prod_{i=0}^{x-1} \sigma_i (T) \prod_{i=0}^{N-1} s_i (T) \tag{2.12}$$

In Eq. (2.12) there are x unknown parameters (σ_i) that account for the resistance to form a stable nucleus. If these parameters are clubbed together, they can be represented as a single parameter and assigned as a resistance to the formation of the first base pair. Since all the resistances are then assigned to the first base pair formation, this represents a worst case scenario for for-

mation of a stable nucleus. Thus, using equations (2.7), (2.8) and (2.11) and assuming that all resistance to formation of a stable nucleus is associated with the first base pair, the following equation for σ can be written.

$$\prod_{i=0}^{x-1} \sigma_i(T) = \sigma(T) = K_{eq}(T) \left(\prod_{i=0}^{N-1} s_i \right)^{-1} \tag{2.13}$$

2.3.1.1 Estimation of Sigma

Eq. (2.13) is rewritten to represent a general form of the equilibrium relationship as follows:

$$K_{eq}(T) = \sigma(T) \left(\prod_{i=0}^{N-1} s_i(T) \right) \tag{2.14}$$

K_{eq} is calculated based on the overall Gibbs free energy of hybridization for the sequence, which can be decomposed as follows:

$$\Delta G_{seq} = \Delta G_{nuc} + \underbrace{\Delta G_{st} + \sum_{i=0}^{N-1} \Delta G_{bp}(i)}_{\Delta G_{dup}} \tag{2.15}$$

where ΔG_{nuc} is the Gibbs free energy of nucleation (Manyanga et al. (2009)) and ΔG_{dup} is the Gibbs free energy of a duplex. ΔG_{dup} can be decomposed into the Gibbs free energy for stacking ΔG_{st} and the Gibbs free energy for the formation $\Delta G_{bp}(i)$ of i^{th} base pair. NN methods provide the length-independent nucleation enthalpy (ΔH_{nuc}) and entropy (ΔS_{nuc}) and they are usually called initiation enthalpy and entropy, respectively (Manyanga et al. (2009)). Whereas for DNA, ΔG_{nuc} depends on the terminus at which nucle-

ation occurs, for RNA it is sequence and terminus-independent (Xia et al. (1998)). For self-complementary sequences, an additional term ΔG_{sym} must be added to the right hand side of Eq. (2.15) (SantaLucia Jr and Hicks (2004)).

There are several issues that complicate the use of NN parameters in kinetic models for oligonucleotide hybridization. First, although DNA melting can occur simultaneously from both sides of a sequence, ΔG_{nuc} assumes that the first base pair formation occurs at one or the other terminus. ΔG_{nuc} parameters reported in the literature are base pair-specific and available only for nucleation at one or the other end of the duplex. This reflects the fact that models for DNA hybridization have typically assumed the hybridization is nucleated at the termini of the polymer. Parameters have thus been estimated under this assumption. However, as will be discussed below, DNA hybridization can in fact be nucleated at any position in the sequence. Typically, the sum of ΔG_{nuc} for the two termini is used to calculate the nucleation free energy, and we follow this precedent here in the context of our kinetic models. Second, it is not possible to obtain the stability constants for each base pair in a sequence using NN parameters, since there are only $n-1$ NN parameters for an n base pair sequence. This issue does not arise in thermodynamic models for DNA hybridization since one is only interested in the total hybridization free energy, not the stability of each base pair. In kinetic models, these base pair stabilities are required for the computation of sequence-dependent rate constants, as will be shown below. Due to the computation of ΔG_{seq} as sum of $n-1$ pairwise NN parameters rather than a sum of n individual $\Delta G_{bp}(i)$'s

plus a sum of $n-1$ stacking interactions $\Delta G_{st}(i)$,$i>0$, in the NN approach to modeling free energies, there is no unique way to assign the free energy change due to the hybridization of each successive base. Free energy changes are only available for sequences of two neighboring bases. [1]

In the absence of this information, we apply an approach that approximates ΔG_{dup} in Eq. (2.15) as follows: $\Delta G_{dup} \approx \sum_i \Delta G_{st}(i) + \Delta G_{bp}(i)$, where $\Delta G_{st}(i) + \Delta G_{bp}(i) = -RT \ln s_i$, with the s_i calculated according to Eqs. (2.7) and (2.8). $\sum_i \Delta G_{st}(i) \neq \Delta G_{st,NN}$ because the sequence dependence of stacking free energies is not represented in Eqs. (2.7) and (2.8). We then use the accurate NN parameters to compute ΔG_{seq} and ΔG_{nuc} in Eq (2.15). Introducing the notation

$$\Delta G_{\sigma} \equiv \Delta G_{nuc} + \Delta G_{st} - \sum_i \Delta G_{st}(i), \tag{2.16}$$

we can rewrite Eq. (2.15) as

$$\Delta G_{seq} = \Delta G_{\sigma} + \sum_i \Delta G_{st}(i) + \Delta G_{bp}(i) \tag{2.17}$$

Since the product of stability constants is calculated based on $\sum_i \Delta G_{st}(i) + \Delta G_{bp}(i)$, the ratio $K_{eq}/\prod_{i=0}^{N-1} s_i$ accounts for a combined stacking and nucle-

[1] Some cancellation of errors induced by the non-unique assignment of NN free energies to individual base pair hybridization events can occur in kinetic models for DNA hybridization that allow hybridization to nucleate at any position along the sequence and assign free energy changes in a consistent fashion

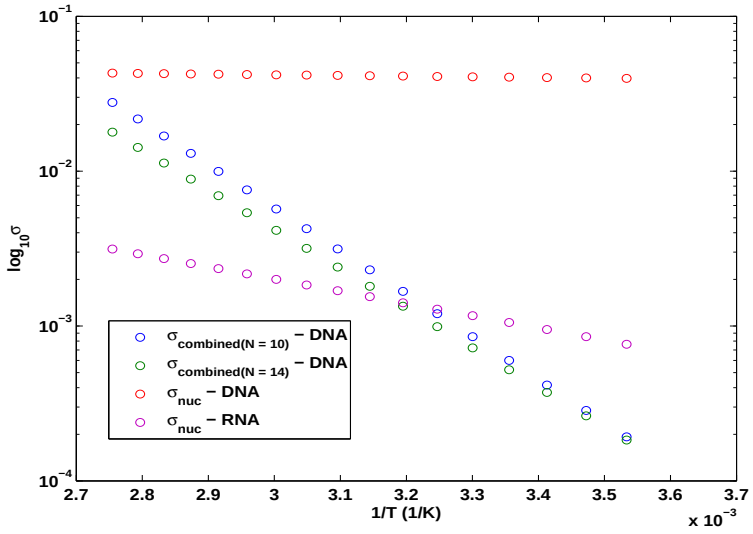


Figure 2.1: Effect of temperature and length of a DNA sequence on nucleation factor, σ_{nuc} and combined nucleation and stacking factor σ. Nucleation/Initiation free energy for AU, AT and GC base pairs have been obtained from NN model. To calculate σ we assumed two different DNA sequences with 10 and 14 base pairs. K_{eq} for these sequences have been calculated using NN model and stability constants have been calculated based on Eqs. (2.7) and (2.8).

ation factor, σ. This can be seen as follows:

$$K_{eq} = \underbrace{\sigma_{nuc}\sigma_{st,corr}}_{\sigma} \prod_{i=0}^{N-1} s_i \tag{2.18}$$

where $\sigma_{st,corr} = \exp\left[-\frac{\Delta G_{st,NN} - \sum_i \Delta G_{st}(i)}{RT}\right]$. We denote by σ the statistical weight of the product of stacking and nucleation factors. σ calculated according to (2.18), with K_{eq} computed according to (2.2) using ΔG_{seq} obtained from NN parameters, will be used in our kinetic models to represent the resistance to formation of the first base pair. Since $\Delta G_{st}(i)$ is much smaller than $\Delta G_{bp}(i)$, we assume the former is the same for all base pairs. This approach maintains thermodynamic consistency between Eqs. (2.14) and (2.15). These approxi-

mations are validated below through comparison to experimental kinetic data.

Fig 2.1 compares the nucleation factor σ_{nuc} calculated based on ΔS_{nuc} and ΔH_{nuc} to the combined nucleation and stacking factor σ for sequences of length 10 and 14 base pairs. There are only three possible values for σ_{nuc}, since the first base pair for DNA/RNA can be any one of $A-U$, $A-T$ and $G-C$, whereas, σ is length-dependent. The difference between σ and σ_{nuc} is due to $\Delta G_{st,NN} - \sum_i \Delta G_{st}(i)$, which is generally < 0, since the sequence-independent stacking energies represented in Eqs (2.7) and (2.8) generally overestimate the magnitude of the actual stacking free energies. This adds a positive ΔH_σ to the ΔG_σ, which contributes to the apparent temperature dependence of sigma. When N increases, x, the number of base pairs that melt before forming a stable nucleus, also increases. As a result of this, the overall resistance to form a stable nucleus increases. Hence σ decreases or resistance increases when N increases. $\Delta G_{st,NN} - \sum_i \Delta G_{st}(i)$ also contributes to the observed length dependence of σ.

2.3.2 One-Sided Melting

In one-sided melting mechanisms, hybridization/ melting progresses from any one side of a DNA and we assign σ to the first step. Fig. 2.2 and the reaction (R_2) (Wetmur and Davidson (1968); Craig et al. (1971); Baldwin (1968)) shows the mechanism of one-sided hybridization. Schwarz and Poland (Schwarz Jr and Poland (1975)) presented the same mechanism as a biased continuous

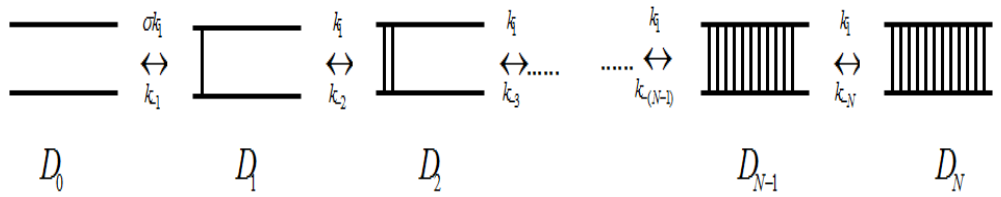


Figure 2.2: Reaction Mechanism of One-sided Hybridization. k_1 is the forward rate constant for each base pair formation and k_{-i} is a dissociation rate constant. σ, a nucleation/stacking parameter, accounts for the resistance to the formation of the first base pair. D_i represents the state of a hybridized duplex with i base pairs in double-stranded form.

time one dimensional random walk (CTRW) with partially reflecting boundary conditions (Gardiner (2004)). The walk is biased because the forward and reverse reaction rates are not equal.

2.3.2.1 Estimation of Relaxation Time Under the Steady-State Approximation

Using the steady state approximation (Baldwin (1968)), the concentrations of the intermediate components in (R_2) can be assumed to be equal to the same constant value. Therefore, the rate of formation of double helix is equal to the rate of disappearance of single strands. Hence, the rate of formation of double helix is the net forward rate at each intermediate step. To estimate the relaxation time, a perturbation δ is introduced to the equilibrium. The net forward rate, r of i^{th} step is

$$r = \begin{cases} \sigma k_1[D_{0eq}][\delta_i] - k_{-i}[\delta_{i+1}], & i = 0 \\ k_1[\delta_i] - k_{-i}[\delta_{i+1}], & \forall 1 \leq i \leq N-1 \end{cases} \tag{2.19}$$

The association of the single strands in the first step of (R_2) is a second-order reaction and the rate equation for this step in terms of δ and σ is obtained by linearizing the actual rate equation around the equilibrium concentration of the single strands, which is denoted as $[D_{0eq}]$ (the factor 2 that was obtained after linearizing has been clubbed together with $[D_0]_{eq}$). Manipulation of Eq. (2.19) leads to the following equation for the rate r in terms of δ_0 (a detailed derivation is provided in the Appendix A).

$$\frac{r}{k_1} = \frac{\delta_0}{N} C \tag{2.20}$$

where

$$C = \left(C_1 - \left[\frac{C_1 C_2 - N\sigma[D_{0eq}]}{C_1 - N \prod_{i=0}^{N-1} \frac{1}{s_i}} \right] \right),$$

$$C_1 = \left[1 + \sum_{i=0}^{N-2} \prod_{j=0}^{i} \frac{1}{s_j} \right],$$

$$C_2 = \sigma[D_{0eq}] + \frac{1}{s} - 1.$$

Relaxation time is defined as

$$\tau^{-1} = -\frac{1}{\delta_0} \frac{d\delta_0}{dt} \tag{2.21}$$

Substituting $r = -\frac{d\delta_0}{dt}$ in equation Eq. (2.21), an equation for the relaxation time is obtained as follows.

$$\tau^{-1} = \frac{k_1}{N} C \tag{2.22}$$

The stability constant s in Eq. (2.22) is a geometric mean of the stability constants of individual base pairs (Wetmur and Davidson (1968)):

$$s = (s_1 s_2 .....s_N)^{\frac{1}{N}} \tag{2.23}$$

Thus, for a given DNA sequence and temperature, the relaxation time can be found using Eq. (2.22).

2.3.2.2 Estimation of Relaxation Time: State space representation for non-steady state solution

Baldwin (Baldwin (1968)) indicates the following requirements for validity of the steady state approximation:

- $\Pi_{i=0}^{n-1} s_i = \Pi_{i=0}^{n-1} \frac{k_i}{k_{-i}} >> n$
- $\sigma << 1$

For short sequences, it may be possible that the product of individual stability constants less than n. Although the value of σ for the oligomer hybridization is always less than 1, in a long DNA melting problem, nucleation is not required for the hybridization of all domains, and hence $\sigma = 1$ for these domains. For these as well as other reasons discussed below, the steady state model is not always useful and a non-steady state solution for the relaxation time is essential as will be shown below.

A detailed unsteady state balance for the chemical reaction network (R_2) yields a set of coupled first order differential equations after linearization around the equilibrium:

$$\frac{dx}{dt} = Ax,\ x(0) = x_{\text{init}}, \tag{2.24}$$

where $x = [D_0, \cdots, D_n]^T$. The state space matrix A is

$$A = \begin{bmatrix} -\sigma[D_{0eq}]k_1 & k_{-0} & 0 & 0 & \cdots \\ \sigma[D_{0eq}]k_1 & -k_1 - k_{-0} & k_{-1} & 0 & \cdots \\ 0 & k_1 & -k_1 - k_{-1} & k_{-2} & \cdots \\ 0 & 0 & k_1 & -k_{-2} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix} \tag{2.25}$$

A is a banded diagonal (tridiagonal) asymmetric matrix ($A \neq A^T$). The solution to the initial value problem (2.24) (or equivalently, the solution to the Chapman-Kolmogorov equation (Gardiner (2004)) for the corresponding discrete state space stochastic process) is

$$\begin{aligned} x(t) &= \exp(At)x(0) \\ &= \exp\left[V\Lambda t W^\dagger\right],\ W^\dagger V = I \\ x(t) &= V\exp(\Lambda t)W^\dagger x(0) \\ &= V\text{diag}\{\exp(\lambda_0 t), \cdots, \exp(\lambda_n t)\}W^\dagger x(0) \end{aligned} \tag{2.26}$$

$W = [w_0, \cdots w_N]$, $w_i^\dagger A = \lambda_i w_i^\dagger$ and $V = [v_0, \cdots, v_N]$, $Av_i = \lambda_i v_i$ denote the matrices of left, right eigenvectors of the fundamental matrix A, respectively, since A is an asymmetric matrix. $\Lambda = \text{diag}\{\lambda_0, \cdots, \lambda_N\}$ is the matrix of eigenvalues. Let $P(t) = \exp\left[V\Lambda t W^\dagger\right]$, which is called the dynamical propagator or the transition probability matrix. The eigenvectors are modes of motion of the dynamical system. These eigenvectors form a complete set since A is a normal matrix. In the non-orthogonal basis W, we have

$$\begin{aligned} W^\dagger x(t) &= \text{diag}\{\exp(\lambda_0 t), \cdots, \exp(\lambda_N t)\} W^\dagger x(0) \\ \tilde{x}(t) &= \text{diag}\{\exp(\lambda_0 t), \cdots, \exp(\lambda_N t)\} \tilde{x}(0). \end{aligned} \tag{2.27}$$

In this basis, the solution for the evolution of each species is

$$\begin{aligned} \tilde{x}_i(t) &= \exp(\lambda_i t)\tilde{x}_i(0), \ i = 0, \cdots, N \\ &= \exp[\lambda_i)t]\tilde{x}_i(0) \end{aligned} \tag{2.28}$$

One examines the mode with eigenvalue $\text{Re}(\lambda_{i,\max}) \leq 0$ to identify the relaxation time for hybridization as the associated time constant:

$$\tau = \tau_{i,\max} = -\frac{1}{\text{Re}(\lambda_{i,\max})}. \tag{2.29}$$

The above expressions are simplified if we exploit the specific structure of A as a chemical kinetics transition rate matrix. Further properties of the solution

(2.26) are considered below.

Schwarz and Poland (Schwarz Jr and Poland (1975)) considered transient dynamics of the homopolymer (but not heteropolymer) melting problem using state space and Kolmogorov master equation methods, for reasons of analytical tractability. As will be discussed below, homopolymer approximations are not typically capable of accurately predicting relaxation times for heteropolymers. The master equation for the continuous-time random walk is a system of partial differential equations for the occupation probability $p(i,t)$ equivalent to the state space system of ordinary differential equations, where i denotes the number of hybridized bases. Master equation solutions (obtained, e.g., by the method of characteristic functions (Gardiner (2004))) are expressed in terms of modified Bessel functions rather than a sum of exponential functions, and are not convenient for calculation of relaxation times or rate constants.

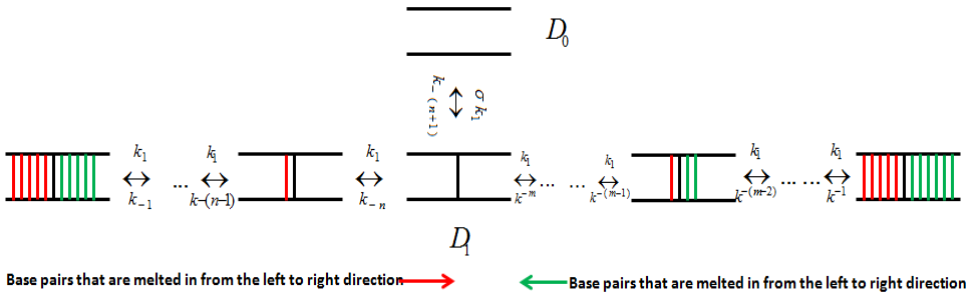
2.3.2.3 Comparison between Transient and Steady State(SS) model

The relaxation time of a DNA hybridization reaction varies from 10^{-1} to 10^{2} seconds (Craig et al. (1971)). Since this reaction is very fast, steady-state assumptions that do not account for the initial transient behavior may be inaccurate. The system of algebraic equations solved in the steady state formulation is inconsistent when the additional constraint corresponding to total mass balance (conservation of probability) is applied. This inconsistency would be observed if one were to solve for the concentrations of the intermediate species. In order for the steady state approximation to be valid,

$\tau_{ss} \approx \tau_{\max} >> \tau_i,\ i \neq \max$ should hold, where τ_{ss} denotes the relaxation time obtained in the steady state formulation, and τ_{max} should be associated with a mode of motion that converts x_0 to x_n. The validity of the steady state approximation can thus be evaluated based on the eigenvalues of the transient model. Properties of the eigenvectors and eigenvalues are considered further in Section 2.4.

2.3.3 Two-Sided Melting

An obvious issue with one-sided melting models is the arbitrariness in the choice of the end of the sequence at which initiation occurs. Schwarz and Poland (Schwarz Jr and Poland (1975)) and Anshelevich *et al.*(Anshelevich et al. (1984)) proposed a two-sided melting reaction mechanism for (R_1) and analyzed it for a homogeneous DNA. According to this model, DNA melting/hybridization can be initiated anywhere on the single strands and subsequent hybridization can occur simultaneously on both sides of the first formed base pair. In this section, we modify this mechanism and develop a transient model for the heterogeneous oligonucleotide melting/hybridization reactions. Fig. 2.3 illustrates two-sided melting, and the reaction mechanism of the two-sided melting reaction R_3 is given below. For clarity, we present the mechanism for $N = 2$.



$$m + n \leq N, 1 \leq m \leq N \;\&\; 1 \leq n \leq N$$

Figure 2.3: Two-sided melting reaction mechanism. m is the number of base pairs melted from the left side and n is the number of base pairs melted from the right side. k^{-i} and k_{-i} represent the dissociation rate constant for left and right side melting respectively. N - total number of base pairs of the given DNA sequence.

$$\begin{array}{ccccc} D_{0,0} & \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} & D_{0,1} & \underset{\sigma k_1}{\overset{k_{-2}}{\rightleftharpoons}} & D_{0,2} \\ \Big\downarrow{\scriptstyle k_1} \Big\uparrow{\scriptstyle k^{-1}} & & \Big\downarrow{\scriptstyle \sigma k_1} \Big\uparrow{\scriptstyle k^{-1}} & & \\ D_{1,0} & \underset{\sigma k_1}{\overset{k_{-1}}{\rightleftharpoons}} & D_{1,1} & & \\ \Big\downarrow{\scriptstyle \sigma k_1} \Big\uparrow{\scriptstyle k^{-2}} & & & & \\ D_{2,0} & & & & \end{array} \qquad (R_3)$$

$D_{m,n}$ represents the state of a DNA molecule where m and n denote the number of base pairs melted from the left and right side of the DNA molecule, respectively. In reaction R_3, the horizontal direction represents right-sided melting and the vertical direction represents left-sided melting. The total number of possible states in an N base pair hybridization reaction is $(N+1)\times(N+2)/2$. Out of these, there are N states that have only one base pair hybridized, in N different positions. The σ parameter is associated with the rate constant for formation of each of these N states. The number of states in the one-sided

melting model is of $O(N)$, whereas in the two-sided melting model, it is of $O(N^2)$. Using the convention that states with the same m value are grouped together, the state vector for two-sided melting is

$$x = [D_{0,0}, D_{0,1}, D_{0,2}, D_{1,0}, D_{1,1}, D_{2,0}]^T$$

and the state space matrix for the reaction (R_3) is given as

$$A = \begin{bmatrix} -(k^{-1}+k_{-1}) & k_1 & 0 & k_1 & 0 & 0 \\ k_{-1} & -(k_{-2}+k^{-1}+k_1) & \sigma[D_{0,2eq}]k_1 & 0 & \sigma[D_{1,1eq}]k_1 & 0 \\ 0 & k_{-2} & -\sigma[D_{0,2eq}]k_1 & 0 & 0 & 0 \\ k^{-1} & 0 & 0 & -(k_{-1}+k^{-2}+k_1) & \sigma[D_{1,1eq}]k_1 & \sigma[D_{2,0eq}]k_1 \\ 0 & k^{-1} & 0 & k_{-1} & -2\sigma[D_{1,1eq}]k_1 & 0 \\ 0 & 0 & 0 & k^{-2} & 0 & -\sigma[D_{2,0eq}]k_1 \end{bmatrix} \tag{2.30}$$

Each diagonal block of this two-sided melting matrix corresponds to melting from the right side with a constant number m of base pairs dissociated from the left side. In the two-sided melting mechanism, there are $N+1$ states representing fully dissociated single-stranded DNA ($m+n = N$). These can be represented using one notation, $D_{0,N}$. By doing this we also correct the rates that were counted twice in the above state space matrix (2.30). For example, in reaction R_3, $D_{1,0}$ gives $D_{1,1}$ and $D_{2,0}$, which represent the single-stranded state. Hence the disappearance rate of the molecule $D_{1,0}$ has been counted twice. Thus, from the above state matrix A, two redundant states (in general N states) should be removed and the rates that were counted twice should also be corrected to obtain the correct state space matrix for two-sided melting. Note the total number of states is $(N+1) \times (N+2)/2 - N$. Hence, the Corrected Reaction Mechanism (CRM) R_4 for two-sided melting (considering $N = 2$) is given below.

$$
\begin{array}{ccccc}
D_{0,0} & \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} & D_{0,1} & \underset{\sigma k_1}{\overset{k_{-2}}{\rightleftharpoons}} & D_{0,2} \\
\Big\Downarrow{\scriptstyle k_{-1}}\Big\Uparrow{\scriptstyle k_1} & & & & \\
D_{1,0} & & & & \\
\Big\Downarrow{\scriptstyle k_{-2}}\Big\Uparrow{\scriptstyle \sigma k_1} & & & & \\
D_{0,2} & & & &
\end{array}
\qquad (R_4)
$$

The state space matrix for two-sided melting ($N = 2$) based on the above

modified reaction mechanism is thus given by

$$A = \begin{bmatrix} -(k^{-1} + k_{-1}) & k_1 & k_1 & 0 \\ k_{-1} & -(k_{-2} + k_1) & 0 & \sigma[D_{0,2eq}]k_1 \\ k^{-1} & 0 & -(k^{-2} + k_1) & \sigma[D_{0,2eq}]k_1 \\ 0 & k_{-2} & k^{-2} & -2\sigma[D_{0,2eq}]k_1 \end{bmatrix} \tag{2.31}$$

$$x = [D_{0,0}, D_{0,1}, D_{1,0}, D_{0,2}]^T$$

Just as we have calculated the relaxation time for one sided melting from its state space matrix, the relaxation time for the two sided melting can be calculated from the above state space matrix (2.31). Anshelevich *et al.* (Anshelevich et al. (1984)) considered the double counting model, but not the more accurate single counting model, and only in the absence of σ, possibly due to the greater simplicity of the former model; due to the omission of σ, the solutions therein are not applicable to oligonucleotide hybridization, and moreover, double counting can result in measurable differences in the predicted reaction times (data not shown).

2.4 Properties of Solutions to the Transient Melting Models

In this section we analyze the state space matrices and give a physical interpretation to the solutions to the state space models. To simplify our analysis we

assume that size of the state space system is n (the corresponding number of states can be back calculated based on one- and two-sided melting) and starting in this section the indices i, j, k will run from 1 to n. The transition rate matrices A for both one- and two-sided melting have rank $n-1$, since the rows of these matrices are subject to the single algebraic constraint $\sum_i A_{ij} = 0$, $\forall j$. In addition, these matrices satisfy the property of detailed balance:

$$A_{ij}\pi_j = A_{ji}\pi_i, \quad \forall\, i, j \in \{1, \cdots, n\} \tag{2.32}$$

for some state π. Since

$$\sum_{j=1}^{n} A_{ij}\pi_j = \pi_i \sum_{j=1}^{n} A_{ji} = 0,$$

due to the transition rate property of A, the state π, which is called the stationary distribution, is the eigenvector associated with the null eigenvalue, which we denote by $\lambda_1 = 0$. Now since

$$\begin{aligned} P(t) &= \exp(At) \\ &= \exp\left[V \Lambda t W^{\dagger}\right] \\ &= V \mathrm{diag}\{\exp(0t), \exp(\lambda_2 t), \cdots, \exp(\lambda_n t)\} W^{\dagger} \\ &= V \mathrm{diag}\{1, \exp(\lambda_2 t), \cdots, \exp(\lambda_n t)\} W^{\dagger}, \end{aligned} \tag{2.33}$$

the transition probability matrix $P(t)$ has one eigenvector (π, a probability vector that is scalar multiple of equilibrium state) with a unit eigenvalue and is a stochastic matrix. If the system starts in this state, it does not evolve. We now derive the general solution for $x(t)$, including properties of the eigenvectors and eigenvalues. First note that the detailed balance condition Eq. (2.32) or $\frac{A_{ij}}{A_{ji}} = \frac{\pi_i}{\pi_j}$ implies

$$\mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n})A = A^T \mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n}).$$

since

$$\begin{aligned}
\frac{A_{ik}}{\pi_i} &= \sum_j \delta_{ij} \frac{1}{\pi_i} A_{jk} \\
&= \left[\mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n})A \right]_{ik} \\
&= \frac{A_{ki}}{\pi_k} \quad \text{(detailed balance)} \\
&= \sum_j A_{ji} \delta_{jk} \frac{1}{\pi_j} \\
&= \left[A^T \mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n}) \right]_{ik}.
\end{aligned}$$

Let $M = \mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n})$. Then

$$MA = M \sum_{i=1}^{n} \lambda_i v_i w_i^\dagger = A^\dagger M = \sum_{i=1}^{n} \lambda_i^* w_i v_i^\dagger M. \tag{2.34}$$

Since $MA = A^\dagger M$, Mv_i is a (right) eigenvector of $A^\dagger$ with eigenvalue λ_i and hence we have $w_i = Mv_i$ for the i-th left eigenvector of A, i.e. the eigendecomposition of A can be written:

$$A = \left(\sum_{i=1}^{n} \lambda_i v_i v_i^\dagger \right) M. \tag{2.35}$$

w_i is a right eigenvector of A with eigenvalue λ_i^*, and a left eigenvector of A with eigenvalue λ_i. Since $W^\dagger V = W^\dagger W = I$, the left and right eigenvectors must be associated with the same eigenvalue and hence $\lambda_j = \lambda_j^*$. We can write $W = MV$ for the matrix of left eigenvectors.

The negativity of eigenvalues $\lambda_2, \cdots, \lambda_n$ also follows from the properties of the transition rate matrix A. It can be shown that the Kullback-Leibler (K-L) distance between $x(t)$ and π, namely $K(t) = \sum_i x_i(t) \ln \frac{x_i(t)}{\pi}$, decreases monotonically with time since

$$\begin{aligned} \frac{d}{dt} K(t) &= \sum_{i,j \neq i} A_{ij} x_j(t) \left\{ \ln \left(\frac{x_i(t)}{\pi_i} \frac{\pi_j}{x_j(t)} \right) + 1 - \frac{x_i(t)}{\pi_i} \frac{\pi_j}{x_j(t)} \right\} \\ &\leq 0, \end{aligned}$$

according to the transition rate property of A and the properties of the K-L distance. Based on these properties, the solution to the state space model can

then be expressed

$$\begin{aligned} x(t) &= V\mathrm{diag}(1, \exp(\lambda_2 t), \cdots, \exp(\lambda_n t))W^\dagger x(0) \\ &= \pi \sum_{j=1}^{n} x_j(0) + \sum_{i=2}^{n} v_i v_i^\dagger \mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n}) \exp(\lambda_i t) x(0), \end{aligned} \tag{2.36}$$

$\lambda_i \in \mathbb{R}^-,\ i = 1, \cdots, n$. Further, due to the stochastic property of the transition probability matrix $P(t)$, we have

$$\begin{aligned} \sum_{j=1}^{n} x_j(t) &= \sum_{j=1}^{n} x_j(0) + \sum_{j=1}^{n}\sum_{i=2}^{n} v_{ij} v_i^\dagger \mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n}) \exp(\lambda_i t) x(0) \\ &= \lim_{t\to\infty} \sum_{j=1}^{n} x_j(t) = \sum_{j=1}^{n} x_j(0), \end{aligned}$$

where v_{ij} denotes the j^{th} element of the i^{th} eigenvector, and hence

$$\sum_{j=1}^{n}\sum_{i=2}^{n} v_{ij} v_i^\dagger \mathrm{diag}(\frac{1}{\pi_1}, \cdots, \frac{1}{\pi_n}) \exp(\lambda_i t) x(0) = 0,\ \forall x(0). \tag{2.37}$$

From this it follows that $\sum_{j=1}^{n} v_{ij} = 0, \forall\ i$ in (2.36). A given mode of motion corresponds to a single step of reaction scheme (R2) if all but two elements ($j, j+1$ in the one-sided melting model) of v_i are equal to 0. If $x(0)$ is an initial state corresponding to a perturbation of the equilibrium (R1) by, for example, a temperature jump or addition of a small amount of either duplex or single-stranded DNA, Eq. (2.36) describes the time evolution of concentration perturbations of the various partially hybridized states D_i as the system attains a new equilibrium consisting of a perturbation of the original equilibrium

distribution of these hybridized states represented by π.

2.5 Estimation of the model parameters

In Section 2.3 we have developed a functional relationship between relaxation time and measurable thermodynamic and kinetics parameters of reaction (R_1) using two different reaction mechanisms and model formulations. Generally such a functional relationship is represented as follows

$$\tau = f([D]_{eq}, \sigma, s_{AT}, s_{GC}, k_1, N) \tag{2.38}$$

While σ, s_{AT}, and s_{GC}, depend on the thermodynamics of reaction R_1, k_1 depends on the kinetics. In Section 2.3.1, the estimation of σ, s_{AT}, and s_{GC} has been considered (In Appendix G we have provided temperature dependent stability constant,s and its Arrhenius relationship). In this section we discuss the estimation of k_1.

2.5.1 Estimation of k_1

Estimation of k_1 is a nonlinear parameter estimation problem in which a functional relationship between model parameters and experimentally measured dependent variables (2.38) is available. For a known set of relaxation time vs equilibrium concentration of single strands data, a least square objective function equal to the sum of squared differences between actual and predicted

relaxation times can be minimized to estimate k_1. Given a set of τ vs. $[D_{eq}]$ data, such a parameter estimation problem can formulated and k_1 can be estimated at a specific temperature.

However, at the current time, only limited experimental τ vs. $[D]_{eq}$ data is available for oligonucleotides. At a specified temperature, only a single relaxation time data is available. Due to this limitation, we solved (2.38) algebraically for k_1 given τ for a specified sequence, and used the relaxation time data available for the other sequence for out-of-sample prediction.

Under the steady-state assumption, Craig *et al* (Craig et al. (1971)) derived the following equation that relates τ (here expressed in terms of the hybridization rate constant k_f) to s, N, σ and k_1:

$$k_1 = \frac{k_f}{2\,(N-1)\,\sigma s} \tag{2.39}$$

From Eq. (2.39) it is evident that k_1 is a function of N and as long as N is of the same order of magnitude, k_1 does not vary much with respect to length of a sequence. In Section 2.3.1, we have shown that σ decreases up to an order of magnitude when N increases from 8 to 14 base pairs. Further, Wetmur and Davidson (Wetmur and Davidson (1968)) and Porschke and Eigen (Pörschke and Eigen (1971)) showed that k_f increases with respect to the length of a sequence. Hence, for a fixed temperature, when N increases k_1 should increase. This can be explained with more physical arguments and reasoning as follows. For a fixed temperature, a sequence with more base pairs is more stable than

a sequence with fewer base pairs. The individual stability constant s_i is equal to the ratio $\frac{k_1}{k_{-i}}$. Though at a fixed temperature s is a constant since a shorter sequence dissociates more quickly than a longer sequence, k_{-i} for a shorter sequence should be higher, and k_{-1} should be lower than those of a long sequence.

k_1 for the transient models can be calculated by using the characteristic equation of the state space matrix A, with the negative reciprocal of the relaxation time τ substituted in it as the eigenvalue λ, for the functional relationship (2.38). Since the state space matrix is relatively large, we implemented a bisection method that is described below to solve for k_1 based on the transient models.

- Assume an interval, $[a, b]$ for k_1. Note that at a, the eigenvalue of the A should be greater than the actual eigenvalue (λ_{actual}) and at b it should be less than the actual value.
- Let $c = \frac{a+b}{2}$.
- Find the eigenvalue of A at c.
- If $(\lambda_c - \lambda_{actual} < 0)$, $b = c$, else $a = c$
- Repeat until $(\lambda_c - \lambda_{actual}) \approx 0$

Using the above algorithm for the transient models and Eq. (2.39) for the steady-state model, we have determined k_1 for four different sequences, A_4U_4, A_5U_5, A_6U_6, and A_7U_7, using their relaxation time and initial concentration

of single strands data that were provided by Craig *et al.* (Craig et al. (1971)). Henceforth, all our analysis is based on the reaction conditions given in Craig *et al* (Craig et al. (1971)), which are similar to the standard PCR reaction conditions. Fig. 2.4 shows the variation of k_1 with respect to temperature for the sequences A_5U_5 and A_7U_7. The variation of apparent rate constant k_1 with respect to temperature follows an Arrhenius relationship and it is similar to k_f with a negative apparent activation energy that is a function of actual activation energy and the heat of base pair formation (Breslauer and Bina-Stein (1977); Pörschke and Eigen (1971)). At a fixed temperature, the difference between k_1 values that were calculated based on one and two-sided melting theories is negligible. However, there is an appreciable difference between k_1 values that are calculated based on the steady-state and transient models. Between the one and two-sided models, the negative activation energy of k_1 based on one-sided melting theory is lower than that from the two-sided melting theory.

Fig. 2.5 shows the variation of the activation energy with respect to N. As we explained above, when N increases, k_1 increases, and hence the negative apparent activation energy decreases. For a fixed N, though we expect that there will be a difference in apparent activation energy for homogeneous and heterogeneous sequences, this difference may may lie within an acceptable error bound. Thus, we use the same k_1 that was estimated using the homogeneous sequences for heterogeneous sequences as long as their lengths are the same. In the absence of experimentally measured relaxation time data for different

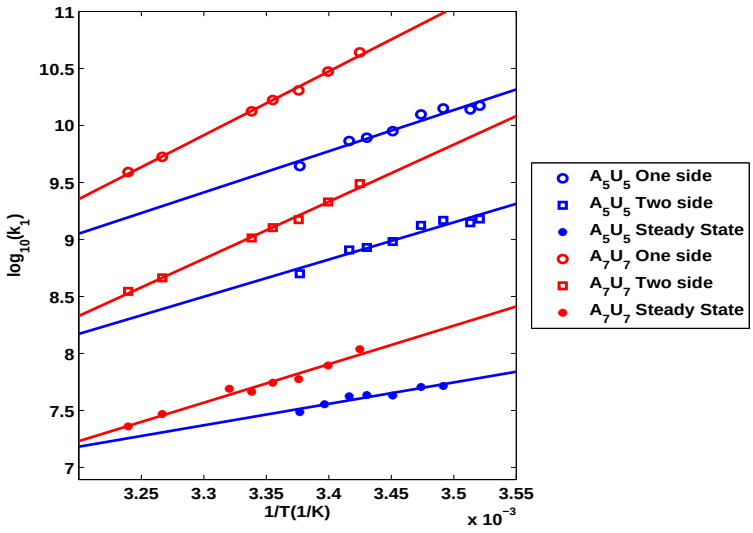


Figure 2.4: The effect of temperature on forward rate constant,k_1, of each step of oligonucleotide hybridization reaction. There is no appreciable difference between k_1 estimated based on the one and two-sided melting mechanisms. k_1 estimated based on steady-state models differs that predicted based on the transient models.

N, we can extrapolate the reaction parameters for different lengths of DNA sequences. However, more experiments can be conducted with different compositions (see, e.g. Williams et al. (1989)) and lengths of DNA sequences to estimate k_1 using least squares estimation rather than nonlinear root finding.

2.6 Experimental Validation of relaxation time models

In order to validate our theoretical model, we considered two sequences, $A_{10}U_{10}$ (Pörschke and Eigen (1971)) and $(A_7U_7)_2$ (Breslauer and Bina-Stein (1977)), which are either completely different or studied independently at a different reaction conditions from those that were used to estimate the model parame-

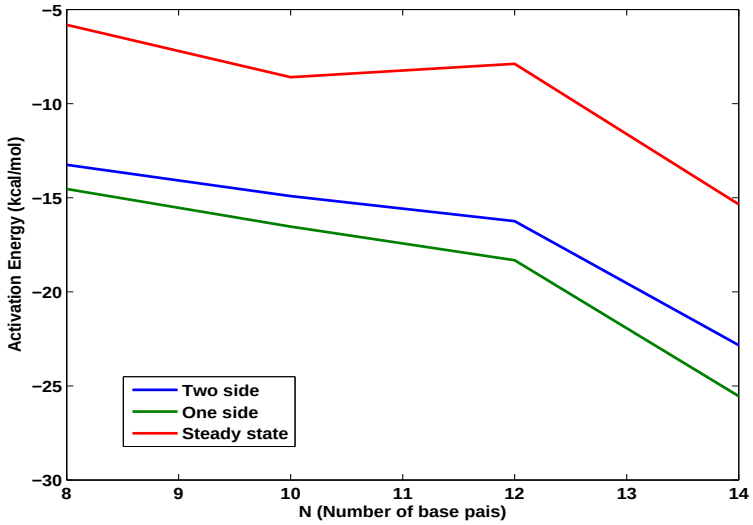


Figure 2.5: Length dependent activation energy for k_1 that was determined based on two sided melting theory. At a fixed temperature, when the number of base pairs of a oligonucleotide sequence increases, the magnitude of the apparent activation energy increases.

ters. To estimate the relaxation time of $A_{10}U_{10}$, we used k_1 that was estimated from the relaxation time of $(A_5U_5)_2$. It should be noted that $(A_5U_5)_2$ (self complementary sequence) and $A_{10}U_{10}$ (non self complementary) are different from each other. Fig.2.6 compares the relaxation time values that are estimated using, experiments, one sided melting theory, two sided melting theory and the steady state model. While one and two sided melting models are in agreement with the experimental data within an acceptable experimental error limit, the steady state model deviates much from the experimental results. Therefore, the steady state assumption may not be valid always and the exact model is recommended. As we have seen in Section 2.5, since there is no difference in the model parameter, k_1 that was estimated using one and two sided melting theory, there is no difference in the relaxation time that was estimated using these two models.

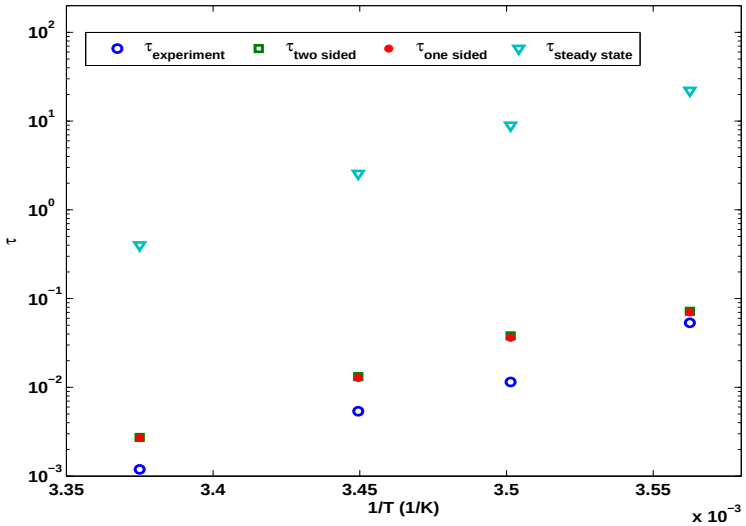


Figure 2.6: Comparison between the experimentally and theoretically estimated relaxation time for $A_{10}U_{10}$. k_1 that is used to estimate the relaxation time of $A_{10}U_{10}$ has been estimated using the relaxation time of a completely different oligomer sequence, $(A_5U_5)_2$ (note that $(A_5U_5)_2$ is a self complementary oligomer sequence and it is different from the non self complementary sequence $A_{10}U_{10}$). There is no appreciable difference between the experimentally estimated relaxation time and theoretically estimated relaxation time using one and two sided melting theory. The steady state model is not in agreement with the experimental results.

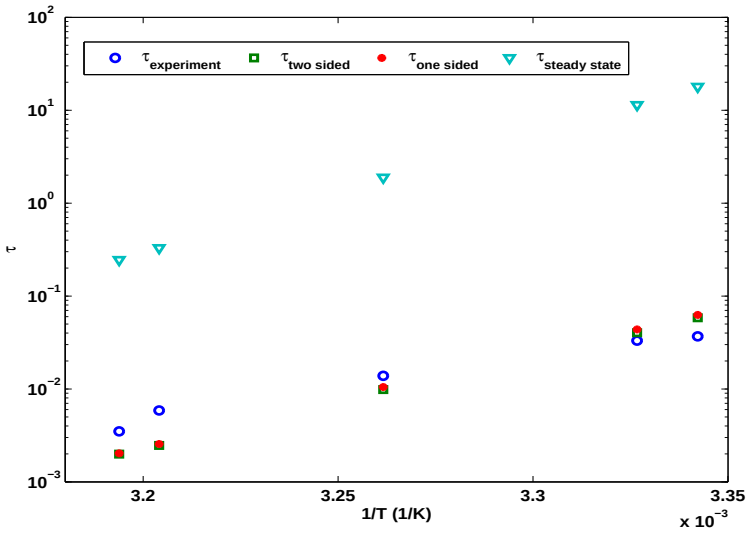


Figure 2.7: Comparison between experimentally and theoretically estimated relaxation time for $(A_7U_7)_2$. The experimental data for $(A_7U_7)_2$ is obtained from (Breslauer and Bina-Stein (1977)). Though we have used k_1 that was estimated from $(A_7U_7)_2$ (Craig et al. (1971)), these two study were entirely independent and conducted at a different set of temperatures. As it was observed in $A_{10}U_{10}$, in this case also, there is no appreciable difference between the experimentally estimated relaxation time and theoretically estimated relaxation time using one and two sided melting theory. The steady state model is not in agreement with the experimental results.

Fig.2.7 compares the experimentally and theoretically estimated relaxation for $(A_7U_7)_2$. Though we have used k_1 that was estimated from $(A_7U_7)_2$ (Craig et al. (1971)), these two study were conducted entirely independently at a different set of temperatures. As it was observed in $A_{10}U_{10}$, in this case also, there is no appreciable difference between the experimentally estimated relaxation time and theoretically estimated relaxation time using one and two sided melting theory. The steady state model is not in agreement with the experimental results. Thus, our theoretical model matches with the experimental results.

2.7 Relaxation time of a Heterogeneous oligonucletide

In this section we predict the relaxation times of heterogeneous DNA using one-sided and two-sided melting theory and compare these predictions. Since we have established in Section 2.6 that the steady state model is not useful, we neglect it. We consider the effects of N, σ, temperature, and GC content on the relaxation time. As we have mentioned in Section 2.3, in oligonucleotide hybridization the σ parameter accounts for the resistance to formation of first base pair. In a specific case of domain melting of long DNA, σ is fixed to be 1. Although we did not undertake a thorough analysis of domain melting here, we investigated the effect of setting the σ to 1 to understand the effect of the nucleation parameter on relaxation times for DNA melting. The sequences of different lengths and GC content that were used in our study are given in Table 2.1. The initial concentration of single strands for this study is $1\mu M$. We have shown that the kinetic parameter k_1 varies both with respect to length of a sequence and temperature. We therefore use k_1 that was estimated for a A_nU_n sequence of the same length. The effects of GC content

Table 2.1: Heterogeneous Oligonucleotide Sequences

Sequence	N	% GC
GATTGTGTAGATAA	14	57
GACTGTGTAGCTCC	14	29
GACTGTGC	8	63
GATTGTGT	8	38

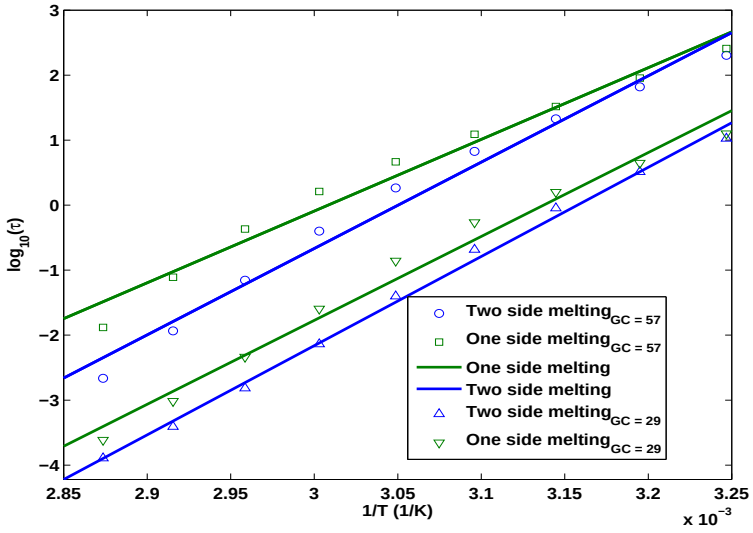


Figure 2.8: Comparison between the one and two-sided melting theory for heterogeneous oligonucleotide hybridization. Number of base pairs is 14 and the initial concentration of the single strands is $1\mu M$.

and temperature on the relaxation time for N=14 are shown in Fig. 2.8. For a fixed GC content, when the temperature increases, the relaxation time decreases. At a fixed temperature and concentration, when the GC content of a sequence increases, the relaxation time also increases. When the overall stability of a sequence increases, the relaxation time increases. In contrast to the homopolymers studied in the previous section, the relaxation times predicted by the one and two-sided models differ for the heteropolymers considered here. The difference between the relaxation times that were obtained using one and two-sided melting theories increases when the temperature increases. All of the above analysis was repeated for $\sigma = 1$ and the results are shown in Fig. 2.9. When $\sigma = 1$, there is no resistance to the formation of the first base pair. The qualitative change of behavior in relaxation time with respect to temperature and GC content for $\sigma = 1$ is similar to that observed for oligonu-

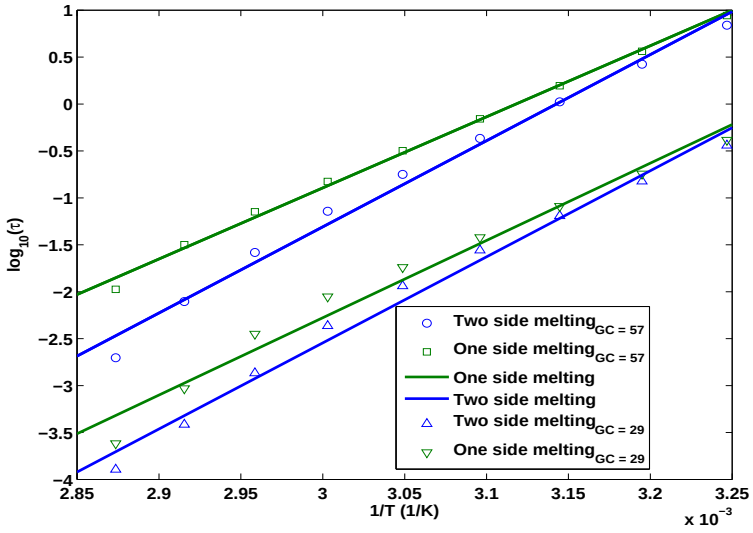


Figure 2.9: Comparison between the one and two-sided melting theory for heterogeneous oligonucleotide for $\sigma = 1$. Number of base pairs is 14 and the initial concentration of the single strands is $1\mu M$.

cleotide hybridization. At a fixed temperature and GC content, however, the relaxation time for $\sigma = 1$ is less than that for oligonucleotide hybridization. We also estimated the relaxation time for a short sequence with 8 base pairs using one and two-sided melting theories (Fig. 2.10). At a fixed temperature and GC content, the relaxation time of a short sequence is lower than that of a long sequence. Unlike the long oligonucleotides, for short oligonucleotides there is a significant difference between the predictions of one and two-sided melting theories at all temperatures.

2.7.1 Unimolecular Hybridization Dynamics of Heterogeneous Oligonucleotides

In this section we study the properties of solutions to the more general two-sided melting model derived in Section 2.4 to systematically analyze the dy-

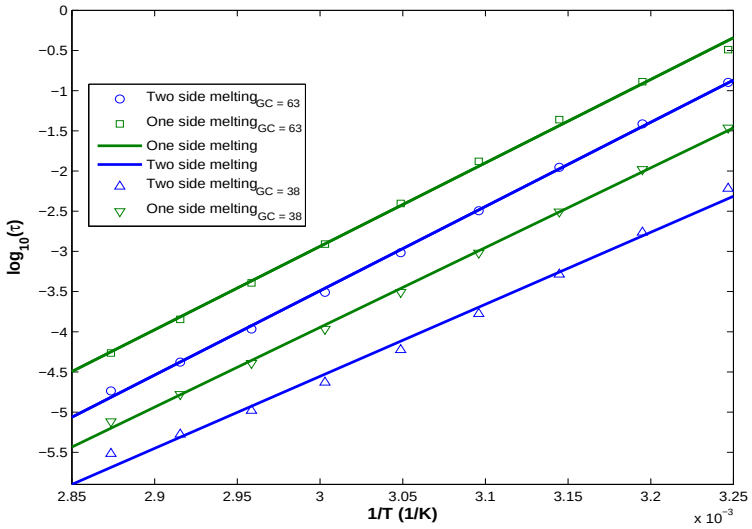


Figure 2.10: Comparison between the one and two-sided melting theory for heterogeneous oligonucleotide. Number of base pairs is 8 and the initial concentration of the single strands is $1\mu M$.

namics of CRM (R_4). With relaxation time alone, it is impossible to determine whether a particular step of CRM (R_4) controls the kinetics of reaction (R_1). The evolution of $\tilde{x}_i(t)$ given by Eq. (2.27) can be used to identify the rate limiting mode. We have shown that one of the eigenvalues of the state space matrix should be zero and it represents the equilibrium. The $\tilde{x}_i(t)$ corresponding to this eigenvalue, π, does not evolve with respect to time. The $\tilde{x}_i(t)$ corresponding to all other eigenvalues converge to zero. $\tilde{x}_i(t)$ corresponding to the largest eigenvalue will converge most slowly. If the difference between the time values at which $\tilde{x}_i(t)$ corresponds to the largest and the second largest eigenvalues converges is considerable compared to the reaction time, a single mode controls the kinetics of reaction (R_1). According to Eqs. (2.36, 2.37), this mode corresponds to a single reaction step of CRM (R_4) only if elements of the associated eigenvector v_i that are coupled in CRM (R_4) are nonzero.

We considered the second sequence that is given in Table 2.1 and performed the above explained study for

- Oligonucleotide hybridization
- In unimolecular DNA melting processes (domain melting), the nucleation parameter $\sigma = 1$.

2.7.1.0.1 Case 1: σ corresponds to oligonucleotide hybridization, Low temperature Fig. 2.11 shows the evolution of $\tilde{x}(t)$ for oligonucleotide hybridization at 40 ^{0}C. Except a $\tilde{x}_i(t)$ that corresponds to the largest nonzero eigenvalue, all other $\tilde{x}_i(t)$ converge very fast. Note that because these $\tilde{x}_i(t)$ converge very fast compared to $\tilde{x}_i(t)$ corresponds to the largest eigenvalue, they superimpose on one another in Fig. 2.11 over the chosen time scale. It is clear from Fig. 2.11 that the $\tilde{x}_i(t)$ corresponding to the largest eigenvalue controls the corrected reaction network (R_4). Furthermore, it can be noted that the time at which $\tilde{x}_i(t)$ corresponding to the largest eigenvalue converges to zero is approximately equal to the relaxation time of a chosen sequence at a chosen temperature. Nonetheless, the associated v_i in Eq. (2.36) has multiple nonzero elements (data not shown), indicating that no single reaction step in CRM (R_4) is rate-limiting. This is consistent with nucleation occurring at any position along the sequence, unlike the case of one-sided melting, where nucleation is associated with only one reaction step in CRM (R_4).

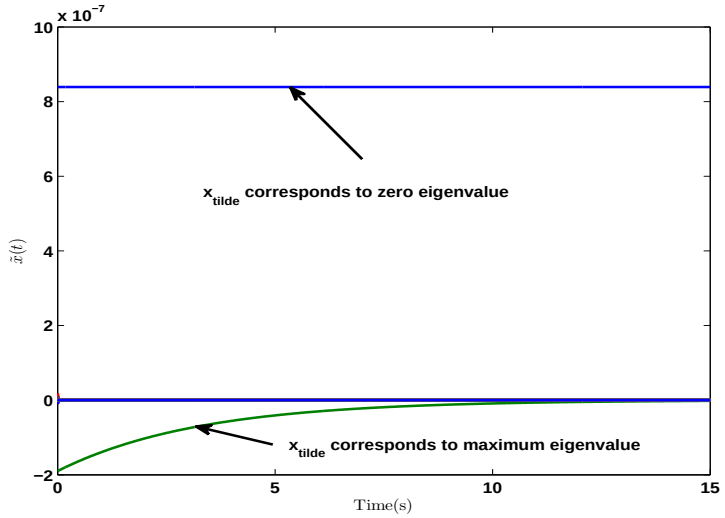


Figure 2.11: Convergence of $\tilde{x}(t)$(described in Eq. (2.27)) for oligonucleotide hybridization at 40 ^{0}C. Equilibrium was perturbed by adding $\delta = 0.01\mu M$ with the equilibrium concentration of single strands.

2.7.1.0.2 Case 2: σ corresponds to oligonucleotide hybridization, High temperature The oligonucleotide that is considered in this study melts at 70 ^{0}C. Therefore, the reverse direction of reaction (R_1) is dominant compared to forward direction. Unlike the hybridization reaction, the melting reaction does not have a resistance in the first step. Therefore, no single mode completely dominates the kinetics of CRM (R_4) at high temperatures. Hence, almost all $\tilde{x}_i(t)$ converge to zero at a same time and this time is approximately equal to the relaxation time at 70 ^{0}C.

2.7.1.0.3 Case 3: $\sigma = 1$ We analyzed the reaction mechanism of reaction (R_1) with $\sigma = 1$. As expected, both at higher and lower temperatures, $\tilde{x(t)}$ corresponding to all the intermediates converges to zero at a same time very quickly. This confirms that no single mode controls the kinetics when $\sigma = 1$.

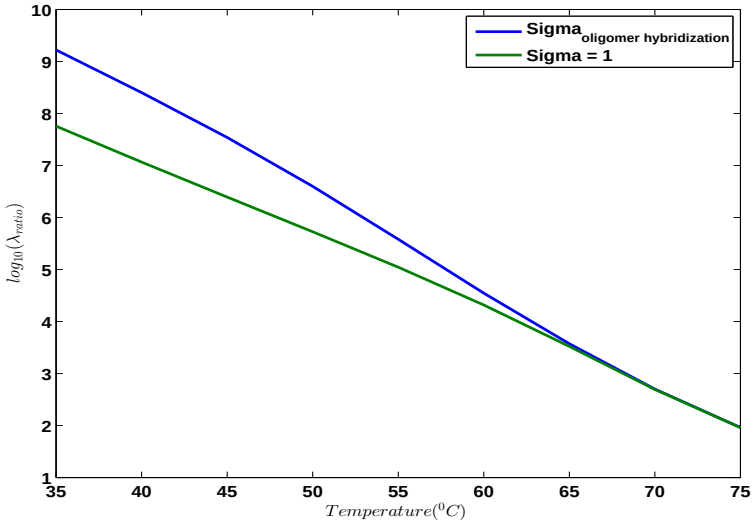


Figure 2.12: Variation of the ratio of largest and the second largest non zero eigenvalues of state space matrix with respect to temperature for oligonucleotide hybridization and domain melting. If this ratio is very high then a specific step of reaction R_4 determine its rate otherwise all the steps participates in determining the rate of reaction R_4.

2.7.1.1 Eigenvalues Analysis

Above analysis can also be done using the eigenvalues of the state space matrix. We know that the largest non negative eigenvalue of the state space matrix determines the rate of the reaction R_4. If other eigenvalues are in the same order of magnitude as the largest eigenvalues, then largest eigenvalue alone does not control the kinetics. Essentially, if the ratio between the largest and the second largest eigenvalue is very high then a specific step of reaction R_4 controls the kinetics, otherwise all the steps plays an equal role in determining the rate of reaction R_4. Figure 2.12 shows the variation of the above discussed eigenvalue ratio with respect to temperature for oligonucleotide hybridization as well as domain melting. In both cases, at lower temperatures, the eigenvalue ratio is very high and when temperature increases it reaches to 1.

2.8 Experimental Validation of Hybridization model

In this Section, we calculate oligonucleotide hybridization rate constants by solving Eq. (2.2) and (2.5) for A_5U_5 and compare them with those estimated experimentally. Relaxation times were obtained from the results in Section 2.6. To calculate K_{eq}, we did not apply the NN method because experimentally estimated ΔG_{seq} were available. Fig. 2.14 compares theoretically and experimentally estimated rate constants for $A_{10}U_{10}$. While the sign of the apparent activation energy is negative for the forward rate constant, it is positive for the reverse rate constant. The apparent activation energy for the forward rate constant is a function of an actual activation energy and reaction enthalpy of base pair formation (Craig et al. (1971)). Since the base pair formation is an elementary reaction, the actual activation energy would be positive and small. In other words, the temperature dependence of rate constants is mainly due to the Gibbs free energy. The reverse rate constant, however, shows a strong dependence on the temperature. At any chosen temperature, the difference between experimentally and theoretically estimated rate constants is small.

We have already shown in Section 2.6 that the difference between the variation of theoretically and experimentally estimated relaxation times with respect to temperature lies within an acceptable error bound. Each experimental data point compared here was estimated by fitting Eq. (2.5) for various total concentrations, whereas, theoretical rate constants are calculated by solving

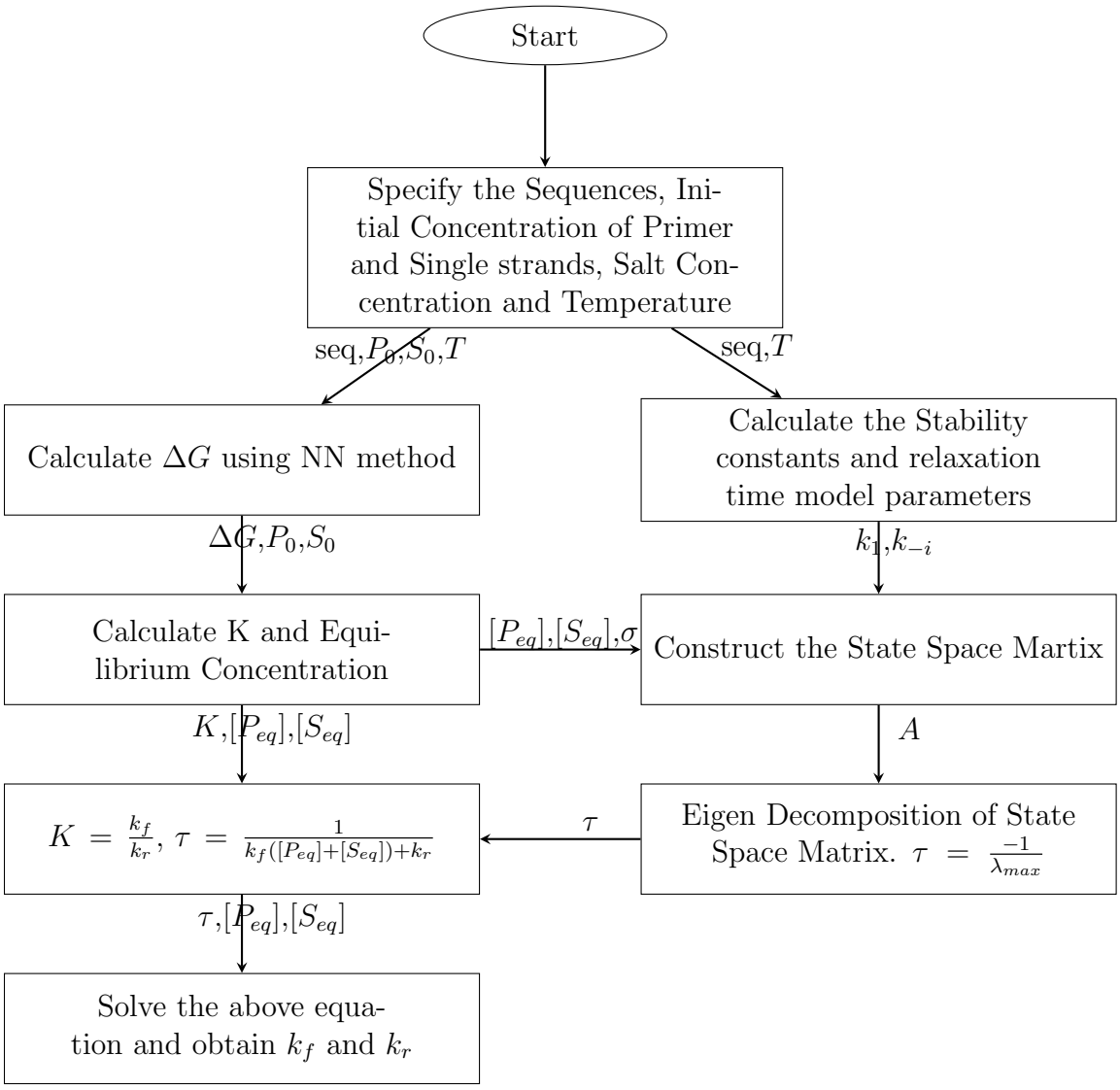


Figure 2.13: Theoretical Modeling of DNA Hybridization Reaction

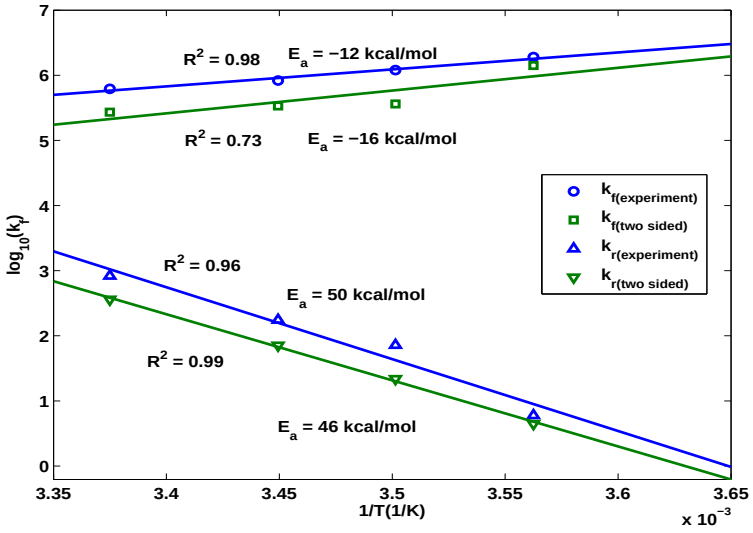


Figure 2.14: Comparison between theoretically and experimentally estimated forward(k_f) and reverse rate constant(k_r) of reaction (R_1) for the hybridization of $A_{10}U_{10}$.

Eq. (2.5) with one relaxation time and a corresponding total concentration, because the different total concentration values that were used in the experiments are not available. This could be a reason for the difference between experimental and theoretical results. Assessment of the accuracy of sequence-dependent rate constant prediction for a broad spectrum of sequences should be carried out systematically with larger datasets including sequences with relaxation times spanning a larger range. The availability of more comprehensive experimental relaxation time data, especially for non-self-complementary sequences with a range of GC contents, should facilitate the development of more accurate parameter sets for oligonucleotide hybridization kinetics. The current work is focused on providing a systematic theory for sequence-dependent oligonucleotide hybridization kinetics with a few representative examples.

We also investigated the robustness of our model by studying the effect

of the difference between theoretically and experimentally estimated rate constants on the evolution of reaction products. Fig. 2.15 shows the evolution of product, D, of reaction (R_1) that was predicted using the experimentally and theoretically estimated rate constants. It can be observed from this Figure that only at low temperatures there is a small deviation between experimental and theoretical prediction. At all other temperatures there is almost no difference. Within the given temperature interval in which the most of the hybridization occurs, the difference between experimentally and theoretically predicted evolution of D of reaction (R_1) is negligible. The given dimensionless concentration in Fig. 2.15 should be multiplied with the micromolar or nanomolar units to get the actual concentration; hence, the difference between the experimental and theoretically predicted concentrations is well within experimental error. Although in this case, the system reaches equilibrium within few seconds, it is shown below that there are different cases of hybridization where sequence-dependent kinetics have a strong influence on the evolution of product over experimentally relevant time scales.

2.9 Application: PCR Simplex Hybridization Reaction

In the context of PCR, oligonucleotide hybridization is called annealing. Annealing reactions can be competitive or noncompetitive depending on the stage

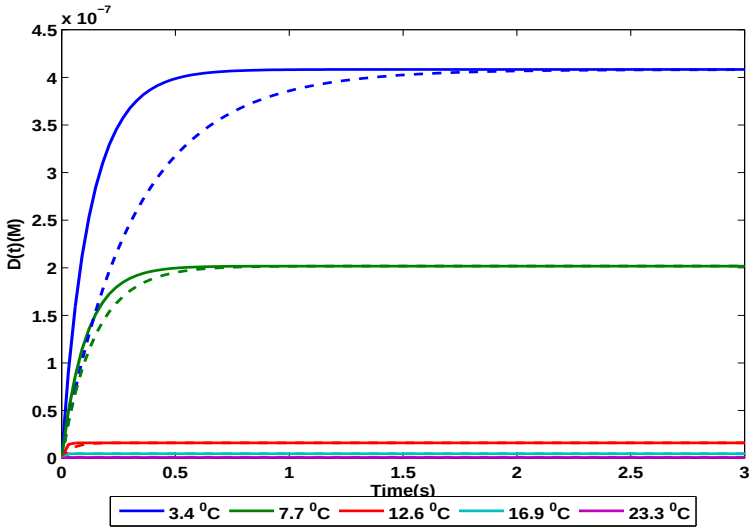


Figure 2.15: Robustness of our kinetic model. The deviation between the evolution of DNA which has been calculated using the theoretically (dotted line) and experimentally (solid line) estimated hybridization rate parameters.

of PCR. In the annealing step of a simplex PCR, there are several following annealing reactions can occur.

1. Two primer molecules bind on a specific part of respective templates separately.

2. Single strands that are formed by melting the template can also anneal each other as they are complementary.

3. Both primers can hybridize on each other.

4. Hybridization of primer 2 on the single strand 1.

5. Hybridization of primer 1 on the single strand 2.

Further, in this study, we assume that stability of the items 3, 4 and 5 are too low to be neglected. Thus the following reaction scheme is written to represent

the simplex PCR annealing reaction.

$$S_1 + P_1 \underset{k_r^1}{\overset{k_f^1}{\rightleftharpoons}} S_1P_1 \qquad (R_6)$$

$$S_2 + P_2 \underset{k_r^2}{\overset{k_f^2}{\rightleftharpoons}} S_2P_2 \qquad (R_7)$$

$$S_1 + S_2 \underset{k_r^3}{\overset{k_f^3}{\rightleftharpoons}} DNA \qquad (R_8)$$

In a typical PCR conditions, the primer concentration is always higher than the template concentration and due to this, the primers annealing reactions R_6 and R_7 don't allow the single strands to hybridize R_8. This is, however, always not true as the template concentration increases and primer concentration decreases during the course of PCR. In microarray applications concentration of both primers and single strands are the same. In this section we investigate the kinetics of reactions R_6 to R_7 for various initial concentrations of single strands and primers. In order to study the kinetics of reactions R_6, R_7 and R_8 we represent the state equations for the reactions species in dimensionless form. For a fixed initial single strand concentration, a dimensionless quantity, the ratio of the initial primer and single strand concentration (PS ratio), can be used to study the evolution of the reaction species.

$$\frac{d[S_iP_i]^*}{dt^*} = k_f^{i*}[S_i]^*[P_i]^* - k_r^{i*}[S_iP_i]^* \quad \forall i = 1, 2 \qquad (2.40)$$

$$\frac{d[DNA]^*}{dt^*} = -k_r^{3*}[DNA]^* + k_f^{3*}[S_1]^*[S_2]^* \qquad (2.41)$$

where

$$[S_i] = [S_{i0}] - [S_iP_i] - [DNA] \quad \forall i = 1, 2$$

$$[P_i] = [P_{i0}] - [S_iP_i] \quad \forall i = 1, 2$$

$$[S_i]^* = \frac{[S_i]}{[S_{10}]}, [P_i]^* = \frac{[P_i]}{[S_{10}]}, [S_iP_i]^* = \frac{[S_iP_i]}{[S_{10}]}$$

$$[DNA]^* = \frac{[DNA]}{[S_{10}]}, t^* = \frac{t}{t_{total}}$$

$$k_f^{i*} = t_{total}k_f^i[S_{10}], k_r^{i*} = t_{total}k_r^i$$

$[S_{10}]$ - Initial Concentration of single strand 1

t_{total} - Total reaction time

In order to solve the state equations 2.40 and 2.41 the algebraic constraints for $[S_i]$ and $[P_i]$ needs to be included. At the start of a PCR, the concentration of both single strands is the same. Though indirectly we include the effect of initial concentration of primers in the form of the mole balance of single strands and by non dimensionalizing the rate constants, it is convenient to use initial PS ratio as a single parameter to study the kinetics of annealing reaction. In this section we study

- Effect of PS ratio on annealing reaction kinetics.
- Effect of GC content on annealing reaction kinetics.
- Effect of mismatch on annealing reaction kinetics.

2.9.1 Effect of PS ratio on annealing dynamics

In order to study the effect of PS ratio on annealing dynamics, initial concentration of the primer is fixed to be $10^{-6}M$ and the concentration of single strand is varied from $10^{-14} - 10^{-6}M$. For our simulations we have considered the following primer molecules.

$$Primer1 : GCTAGCTGTAACTG$$

$$Primer2 : GTCTGCTGAAACTG$$

All our simulations have been conducted at one single hybridization temperature. To select such a temperature, the equilibrium melting curve is generated for both primers at all the above mentioned SP ratios. Melting curves for primer 1 is presented in figure 2.16. Since the GC content for the both primers are same the melting profile is also expected to be the same for them. Therefore, we present the melting profile for only primer 1. To ensure the 100 % hybridization, we have selected 20 ^{0}C at which the equilibrium conversion is always 1 for all PS ratios, as an annealing temperature.

At high PS ratios, reaction R_8 can be neglected as the primer molecules can easily hybridize on the the single strands and will now allow single strands to hybridize each other. When the primer ans single strands concentration is comparable, reaction R_8 needs to considered along with reactions R_6 and R_7. While the rate constants for the reactions R_6 and R_7 have been estimated as

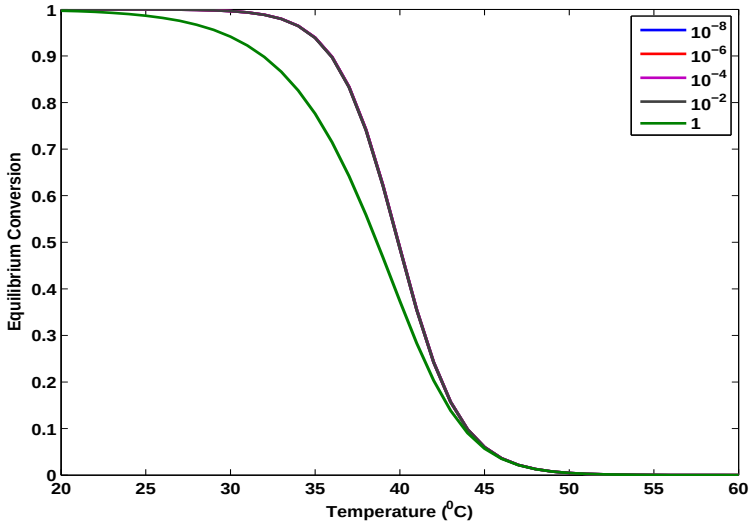


Figure 2.16: Melting Curve for Primer 1 at different S/P ratios. Primer concentration is fixed to be $1\mu M$ and the single strand concentration is increased from $10^{-14}M$. For all SP ratio > 1 melting point is almost the same. In this figure we show the melting curves for only few SP ratios for the better clarity of the plot

explained in Flow chart 2.13, the forward and reverse rate constant for the reaction R_8 should be the maximum that of the rate constants of reactions R_6 and R_7 for the following reasons.

It should be noted that primers 1 and 2 will be similar to some part of the single strands 2 and 1 respectively. Let us say the initial concentration of the primer and single strands are the same. Therefore, there is an equal opportunity for the annealing of S_1 and P_1 and S_1 and a part of S_2 which is has the same sequence composition of P_1. Hence, the rate constant for the annealing of S_1 and the part of S_2 which has the same sequence of P_1 should be equal to that of annealing of S_1 and P_1. If this rate constant is higher than that of reaction 2, then, annealing of S_1 and the part of S_2 which has the same sequence of P_1 will occur first and then the annealing of S_2 and the spot on S_1

which has the same sequence of P_2 will follow. On the other hand, if the rate constant of reaction 2 is higher than reaction R_6, then the annealing between S_2 and the part of S_1 which has the similar sequence of P_2 will occur first. Once any one annealing occurs between S_1 and S_2 then those molecules are not available for P_1 and P_2. Thus, the forward rate constant for the reaction R_8 should be maximum that of reactions R_6 and R_7. It can be argued that during the single strands annealing, one should consider all parts of the single strands rather than the part at which primer binds. It should be, however, observed that there is no competition for the annealing of all such parts and annealing of those will be faster. If the primer concentration is in excess or low compared to the single strand concentration then, the rate constants $k_f P_1$, $k_r P_1$ $k_f P_2$, $k_r P_1$, $k_f t$ and $k_r t$ should be calculated separately in two stages.

- Calculate the rate constants $k_f P_1, k_r P_1, k_f P_2$ and $k_r P_1$ for a known primer and single strands concentration and those rate constants should be used for reactions R_1 and R_2.

- Calculate these rate constants by assuming primer concentration is equal to that of single strands and calculate them. Then, find the maximum forward rate constant value and use it for reaction R_8.

The above procedure will avoid the confusion that to what extend of length of single strands that needs to be considered to find the Gibbs free energy of reaction R_8.

2.9.1.1 Case 1: High PS ratio

At high PS ratio the formation of $[DNA]$ can be neglected and hence reaction R_8 as well. Under this condition, the resulting state equations are independent to each other. Further, the change in concentration of primer through out the reaction is also negligible. Therefore, a pseudo first order kinetics can be assumed for reactions R_6 and R_7.Thus, the state equations for high PS ratios can be rewritten as follows

$$\frac{d[S_iP_i]^*}{dt^*} = k_f^{i*}[S_i]^*[P_{i0}]^* - k_r^{i*}[S_iP_i]^* \quad \forall i = 1, 2 \tag{2.42}$$

Equations 2.42 has been solved and the following analytical solution for the evolution of $[S_iP_i]$ is obtained.

$$[S_iP_i\,(t)]^* = \frac{c_1\,(1 - exp\,(-c_2t))}{c_2} \tag{2.43}$$

where

$$c_1 = k_f^{i*}[P_{i0}]^*[S_{10}]^*$$

$$c_2 = k_f^{i*}[P_{i0}]^* + k_r^{i*}$$

From Equation 2.43, an equation to calculate time required to produce a specific concentration of $[S_iP_i^*]$ is obtained as follows

$$t = -\frac{1}{c_2}\left(1 - ln\left(\frac{c_2[S_iP_i^*]}{c_1}\right)\right) \tag{2.44}$$

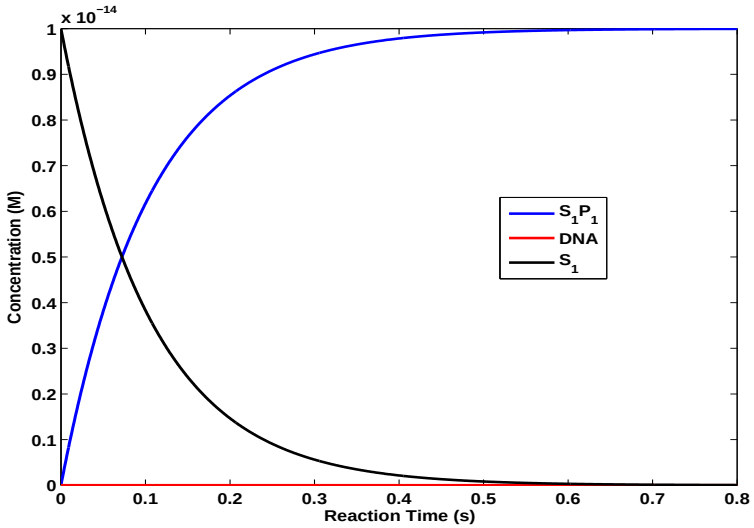


Figure 2.17: Dynamics of Annealing Reaction for a high PS ratio. Annealing Temperature is 20 ^{0}C. Primer concentration is $1\mu M$ and the single strand concentration is $0.01pM$. As expected the reaction reached the equilibrium state within a second due to a high concentration of primers. The evolution of the reaction species when the single strand concentration is varied from $0.1pM$ to $10nM$ is exactly same as this case.(Hence, Pesudo first order kinetics assumption is valid)

Figure 2.17 presents the evolution of single strands, SP and DNA molecules. Note that though we derived an analytical expression for the evolution of SP molecules by assuming a pseudo first order reaction kinetics, we solved equations 2.40 and 2.41 and presented the evolution of the reaction molecules in Figure 2.17. In this way, the pseudo first order kinetics assumption is validated. Annealing temperature and primer concentration is fixed to be 20 ^0C and $1\mu M$. The single strand concentration is varied from $0.01pM$ to $10nM$. As expected, due to a high primer concentration, reaction reaches equilibrium in less than a second. Though in Figure 2.17 we present the evolution for a single strand concentration of $0.01pM$, evolution of reaction species for all other single strands concentration up to $10nM$ is exactly same as this.

2.9.1.2 Case 2: PS ratio = 1

In this section we study the kinetics of the reaction scheme represented by R_6 to R_8 for an equimolar concentration of single strands and primers. While we keep the PS ratio to be 1 and we vary the initial concentration of the primers from $1nm$ to $1\mu M$. Figure 2.18 and Figure 2.19 present the evolution of the reaction species for the primer concentrations $1nM$ and $1\mu M$. When PS ratio is 1, single strands molecules are subjected to a two ways competitive annealing. Even though the primer and single strands concentration is equal since the order of the reaction R_6 to R_8 is 2, their kinetics depends on the initial concentration of the primer. At a higher concentration of primer, the reaction reaches the equilibrium faster compared to at a lower concentration of primer. Note that y-axis is dimensionless and it experess the relative concentration with respect to initial concentration of single strands.

2.9.2 Effect of GC content on Annealing dynamics

In this section we have investigated the effect of GC content on the annealing reaction dynamics for a fixed PS ratio and annealing temperature. Initial concentration of Primer and single strands is assume to be $1nM$. Figure 2.20 presents the effect of GC content on the annealing dynamics at annealing temperature of 20 ^{0}C. When GC content increases then the reaction takes less time to reach the equilibrium. The high GC content sequence are more stable and the equilibrium constant of the hybridization reaction of GC rich

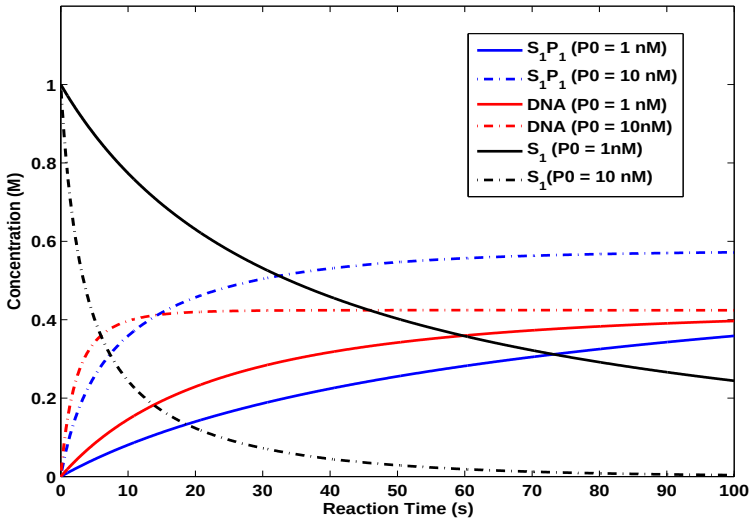


Figure 2.18: Dynamics of Annealing Reaction for a PS ratio = 1. Annealing Temperature is 20 ^{0}C. Primer concentration is varied from $1nM$ to $10nM$. Due to competitive annealing at equal concentration of the single strands and primers, there is a formation of DNA. When the concentration of primer decreases the reaction takes long time to reach equilibrium. Note that y-axis is dimensionless and it expresses the relative concentration with respect to the initial concentration of single strands.

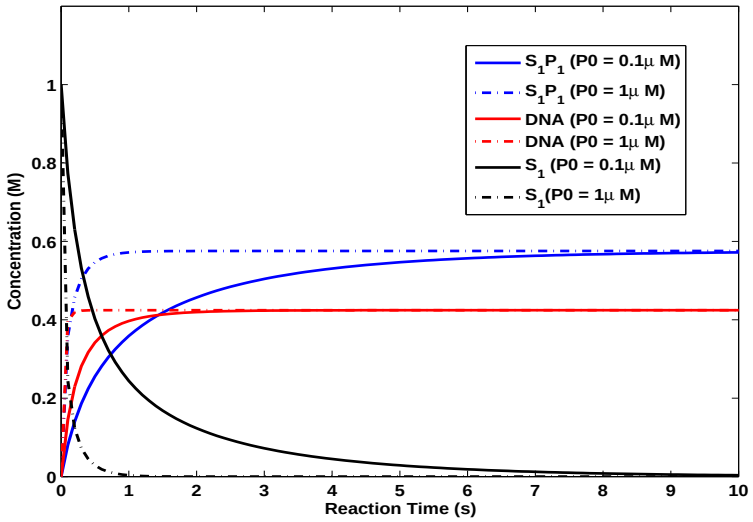


Figure 2.19: Dynamics of Annealing Reaction for a PS ratio = 1. Annealing Temperature is 20 ^{0}C. Primer concentration is varied from $0.1\mu M$ to $1\mu M$. Since the initial concentration of single strand is relatively higher, the reaction reaches equilibrium within 10 seconds.

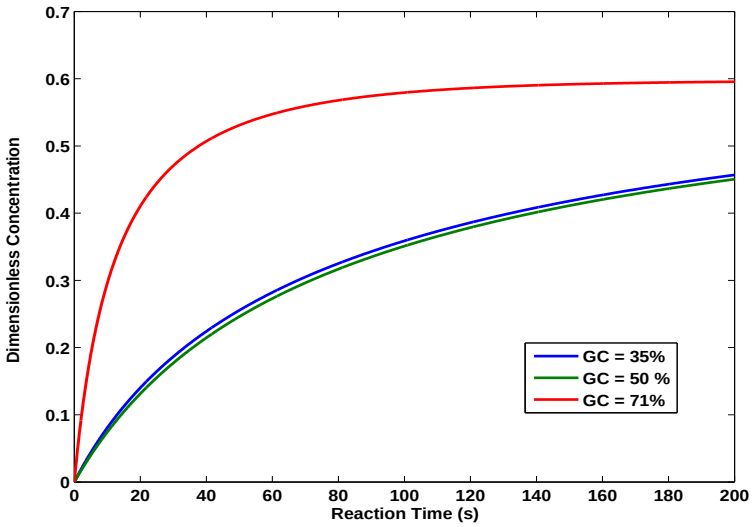


Figure 2.20: Effect of GC content on Annealing Dynamics. When the GC content increases, the reaction takes less time to reach the equilibrium. Initial concentration of Primer and single strands is $1nM$.

sequences is higher compared to low GC content. Therefore, sequences that has more GC content reaches equilibrium faster.

2.9.3 DNA hybridization with mismatching

So far, we have discussed the hybridization of Watson and Crick type DNA sequences which are complementary to each other. While the complementary sequences posses the maximum stability, sequences with few mismatches are also reasonable stable and it is sufficient for them to hybridize. Typically, in genome sequencing techniques such as PCR, primers (probe) are designed in such a way that, they will hybridize on a specific location of the single strand molecules.It is, however possible for a primer molecule to hybridize on a different location of the same single strand with a reasonable stability. Thus, once a non desired hybridization occurs,this can reduce the formation of the

Table 2.2: Perfectly and mismatched sequences

Type	Primer	Complementary Sequence
Perfect Match 1	GCTAGCTGTAACTG	CGATCGACATTGAC
Mismatch	GCTAGCTGTAACTG	CGAT**T**GACATTGAC
Perfect Match 2	GTCTGCTGAAACTG	CAGACGACTTTGAC

desired perfectly matched product. Therefore, it is important to analyze the formation and evolution of the mismatched products during the formation of a perfect hybridization. In this section this phenomena is systematically analyzed and evolution of perfectly and mismatched product is studied. Though there are nearest neighbor parameters to predict the Gibbs free energy for the single mismatches (Allawi and SantaLucia (1998); Peyret et al. (1999); Allawi and SantaLucia (1998); SantaLucia Jr and Hicks (2004)), they are not available to predict the Gibbs free energy of tandem mismatch hybridization. Bourdélat-Parks and Wartell (Bourdélat-Parks and Wartell (2004)) investigated the thermodynamic stability of DNA tandem mismatches for 9 different hairpin DNAs. Fish et al (Fish et al. (2007)) estimated the ΔG for DNA sequences with tandem mismatches. In this work we have considered only an internal mismatch. Table C in Appendix C provides the NN parameters for a single mismatch. Table 2.2 provides the sequences with mismatches.

Reactions R_9 to R_{12} represents a scenario of a mismatching that can occur in a PCR. Primer 1 molecules can form both perfectly matched as well as mismatched products with single strand 1. S_1^{Pr} and S_1^M represent the locations where P_1 can hybridize to form the perfectly matched and mismatched

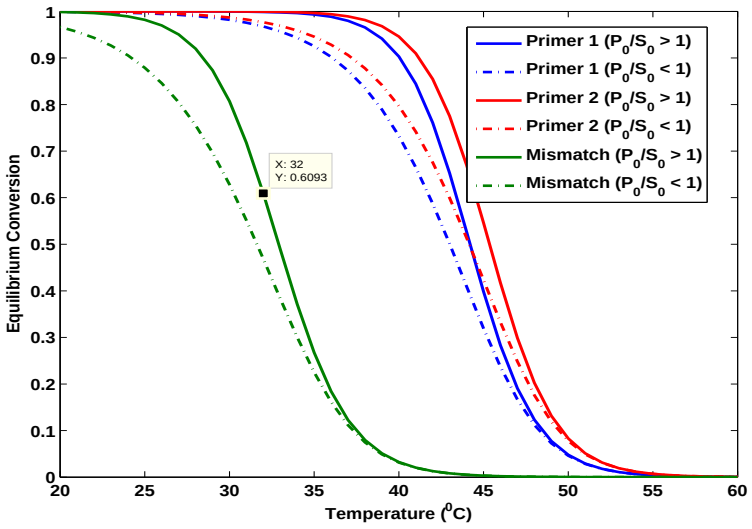


Figure 2.21: Melting curve of Perfectly and Mismatched products. Though mismatched product is not as stable as the perfectly matched products, if the annealing temperature is chosen to be 32 ^{0}C at which perfectly matched products hybrize 100%, the equilibrium conversion of mismatched hybridization is 60%.

products.

$$S_1^{Pr} + P_1 \underset{k_r S_1^{Pr}}{\overset{k_f S_1^{Pr}}{\rightleftharpoons}} S_1^{Pr} P_1 \qquad (R_9)$$

$$S_1^{M} + P_1 \underset{k_r S_1^{M}}{\overset{k_f S_1^{M}}{\rightleftharpoons}} S_1^{M} P_1 \qquad (R_{10})$$

$$S_1^{Pr} + S_2 \underset{k_r S_1^{Pr} S_2}{\overset{k_f S_1^{Pr} S_2}{\rightleftharpoons}} S_2^{Pr} S_2 \qquad (R_{11})$$

$$S_2 + P_2 \underset{k_r P_2}{\overset{k_f P_2}{\rightleftharpoons}} S_2 P_2 \qquad (R_{12})$$

In a PCR usually the length of single strands are much higher than the primers. Therefore, on the same single strand molecules, primers can bind on multiple locations based on the composition of a single strand. Hence, the number of single strand molecules available for both perfectly and mismatched hybridization remains the same. In other words, there is no competitive hybridization

between perfectly and mismatched product. If a specific single strand is hybridized both on a perfectly and mismatched location, it is assumed to be a mismatched product. The same strategy which has been followed to to calculate the rate constants of perfectly matched hybridization reaction is followed for the mismatched hybridization as well.

Here S_1^{Pr} and $S_1^{M_1}$ represent the location on the single strands that can form perfect, mismatched hybridized product.

2.9.3.1 Scenario 1 - P/S ratio > 1

Figure 2.22 presents kinetics of reaction R_9 to R_{12} for a PS ratio > 1. The PS ratio and the annealing temperature is fixed to be 10^8 and 30 ^{0}C with $1\mu M$ initial primer concentration. Since the primer molecules are in excess compared to the single strand molecule, initially a part of primer molecules binds on a location to form a perfectly matched product. If the reaction is allowed to proceed further, the excess primer molecules binds on a location that gives mismatched product. Thus, when the reaction time increases, formation of the mismatched product increases.

2.9.3.2 Scenario 2 - P/S ratio = 1

If the PS ratio is 1, then there is a competitive hybridization between single strands and primers. Figure 2.23 presents the evolution of reaction species for PS ratio of 1. Since the rate constants of reaction R_9 and R_{11} higher than reaction R_{10}, initially there is a competition between reaction R_9 and

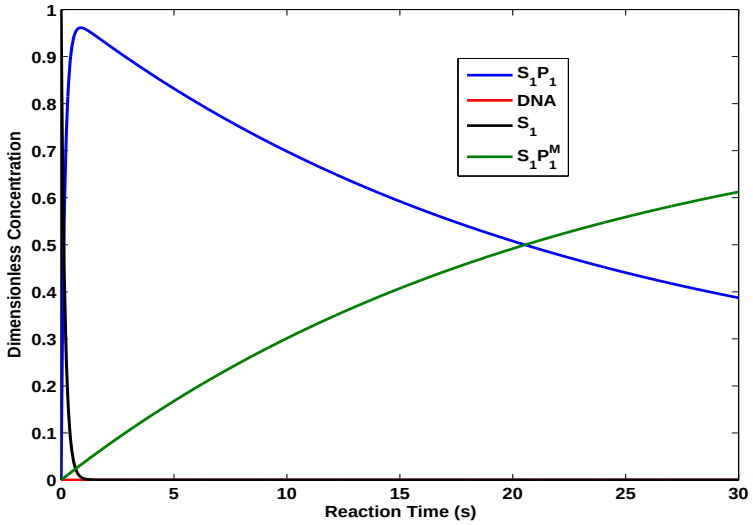


Figure 2.22: Evolution of mismatched product when PS ratio is > 1. If the reactions are allow to carry for long time, with excess primers available in the reaction mixture, mismatched products are formed.

R_{11}. Almost all the products has been formed with in a few seconds then the available primer molecules bind on the a different location that forms a mismatched products. Due to competitive annealing, not all the single strands available for the primer hybridization and that is why there is less formation of mismatched product.

2.10 Conclusion

In this work we have presented a theoretical approach to the prediction of sequence and temperature-dependent rate constants for oligonucleotide hybridization reactions using the equilibrium and relaxation theories of chemical reactions. We considered one-sided and two-sided melting theories to determine the relaxation times of oligonucleotide hybridization reactions, with important modifications to those proposed in past literature on homopoly-

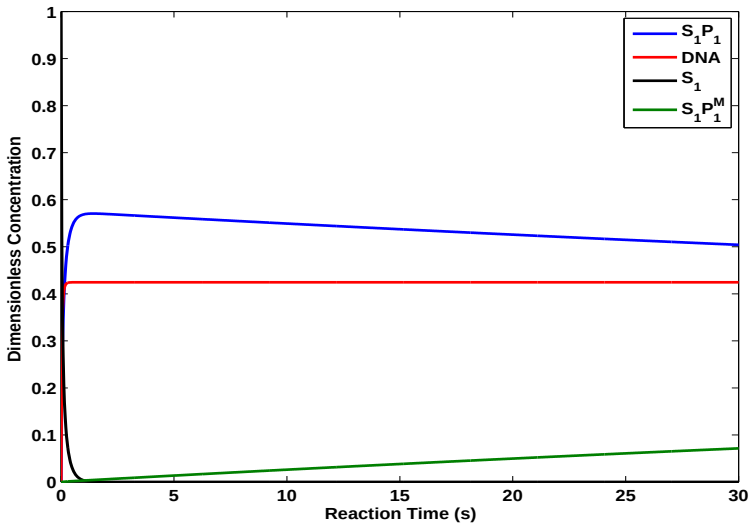


Figure 2.23: Evolution of mismatched product when PS ratio is 1. Due to competitive annealing between primers and single strands, mismatched products is not formed.

mers. The two-sided melting theory that is proposed in the past literature has been modified so that it is consistent with the laws of reaction kinetics. The significance of model parameters has been analyzed in detail and methods for the estimation of these parameters using available experimental data have been presented. Steady-state models were found to be inaccurate, whereas the proposed one-sided and two-sided theoretical models were found to be consistent with the available experimental data. The difference between the predicted and experimental rate constants lies within an acceptable error bound and the difference between the predicted and measured activation energies is negligible. Important differences between one and two-sided melting model predictions are observed for heteropolymers, especially for shorter sequences, indicating that the more physically realistic two-sided models are generally preferred. Transient dynamical analysis of two-sided models shows

that no single elementary reaction step is rate-determining. Finally, we studied the applications of oligonucleotide hybridization reactions in the context of PCR and showed the importance of hybridization kinetics in PCR annealing reactions. Our kinetic model is suitable for the purpose of optimal control studies of PCR, among many other applications.

Chapter 3

Sequence-dependent Modeling of DNA Amplification

3.1 Introduction

In this chapter, a theoretical framework for determination of the optimal dynamic operating conditions of DNA amplification reactions, for any specified amplification objective, has been developed based on first-principles biophysical modeling and control theory. Amplification of DNA is formulated as a problem in control theory with optimal solutions that can differ considerably from strategies typically used in practice. Using the Polymerase Chain Reaction (PCR) as an example, sequence-dependent biophysical models for DNA amplification are cast as control systems, wherein the dynamics of the reaction are controlled by a manipulated input variable. Using these control systems, we demonstrate that there exists an optimal temperature cycling strategy for

geometric amplification of any DNA sequence and formulate optimal control problems that can be used to derive the optimal temperature profile. Two optimal control problems corresponding to the most common objectives of maximization of amplification efficiency for a fixed reaction time and minimization of reaction time for a specified amplification efficiency are formulated. Finally, strategies for the optimal synthesis of the DNA amplification control trajectory are proposed. Analogous methods can be used to formulate control problems for more advanced amplification objectives corresponding to the design of new types of DNA amplification reactions.

3.2 Kinetic Model for PCR

As shown in Fig. 1.3 a kinetic model of PCR consists of kinetic models of melting, annealing, enzyme binding and extension reactions. In this work we have developed a sequence- and temperature dependent state space model for PCR and analyzed its kinetics.

3.2.1 Annealing Kinetics

Reaction $A - R_1$ represents an annealing reaction between the single strands (S) and primers (P).

$$S + P \underset{k_r}{\overset{k_f}{\rightleftharpoons}} SP \qquad (A - R_1)$$

In Chapter 2 we developed a sequence and temperature dependent method to estimate the annealing reaction rate constants. Here we summarize our method that needs to be followed to estimate annealing rate constants.

1. Determine the overall Gibbs free energy and hence the equilibrium constant $K_{annealing}$ for a given sequence at the chosen annealing temperature using the Nearest Neighbor model.

$$K_{annealing} = \frac{k_f}{k_r} = exp\left(\frac{-\Delta G_{annealing}}{RT}\right) \tag{3.1}$$

2. Determine the relaxation time, a characteristic time constant that determines the evolution of reaction coordinates toward equilibrium, at a chosen temperature using either one - or two-sided melting. Here we briefly review our method for calculating the relaxation time of such systems, including the more general case of two-sided heteropolymer melting.

   - For the given length of primer write the reaction mechanism, for example, for $N = 2$.

$$\begin{array}{ccccc} D_{0,0} & \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} & D_{0,1} & \underset{\sigma k_1}{\overset{k_{-2}}{\rightleftharpoons}} & D_{0,2} \\ \Big\Updownarrow{\scriptstyle k_1}{\scriptstyle k_{-1}} & & & & \\ D_{1,0} & & & & \\ \Big\Updownarrow{\scriptstyle \sigma k_1}{\scriptstyle k_{-2}} & & & & \\ D_{0,2} & & & & \end{array} \tag{R_4}$$

- Obtain the values of the rate constants and other parameters as explained in Chapter 2 and form the following state space matrix based on the above reaction mechanism

$$A = \begin{bmatrix} -(k^{-1} + k_{-1}) & k_1 & k_1 & 0 \\ k_{-1} & -(k_{-2} + k_1) & 0 & \sigma[D_{0,2eq}]k_1 \\ k^{-1} & 0 & -(k^{-2} + k_1) & \sigma[D_{0,2eq}]k_1 \\ 0 & k_{-2} & k^{-2} & -2\sigma[D_{0,2eq}]k_1 \end{bmatrix} \tag{3.2}$$

- Calculate the Eigenvalues λ_i of A.
- Calculate the relaxation as per the following equation.

$$\tau = \frac{1}{max\,(\lambda_i)}$$

3. Relaxation time for the reaction $(A - R_1)$ in terms of k_f and k_r can be expressed as

$$\tau = \frac{1}{k_f\,([C_{Seq}] + [C_{Peq}]) + k_r}$$

$[Seq]$ and $[Peq]$ should be determined based on the initial concentration of single strands and primers that are used to determine relaxation time in step 2.

4. Solve the two equations in steps 1 and 3 to determine k_f and k_r.

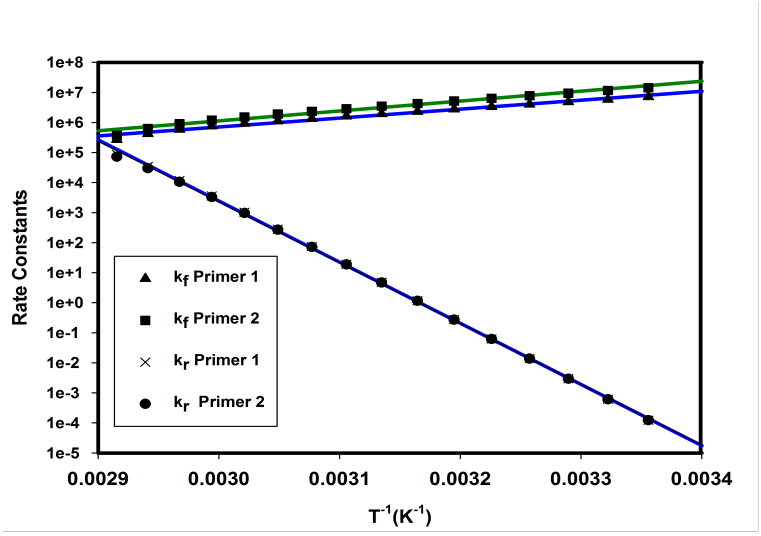


Figure 3.1: Arrhenius plot of the forward (k_f) and reverse (k_r) rate constants for the primer set 1

Fig. 2.13 shows the above procedure as a flowchart.

3.2.2 Example

Using the above steps we have estimated the forward and reverse rate constants for a set of primers and given the reaction parameters in Table 3.1. Arrhenius plots for the annealing rate constants k_f and k_r have been given in Fig. 3.1. From the melting curve (it is not shown here) of the chosen primers

Table 3.1: Rate parameters of primers. Subscript f and r denotes forward and reverse rate constant, respectively

Sequence	$\left(\frac{E_a}{R}\right)_f (K^{-1})$	$\left(\frac{E_a}{R}\right)_r (K^{-1})$	$(k_0)_f$	$(k_0)_f$
GCTAGCTGTAACTG	-7385	46341	2×10^{-4}	5×10^{63}
GTCTGCTGAAACTG	-8202	45926	2×10^{-5}	10^{63}

it can be inferred that 100 % equilibrium conversion for reaction $(A - R_1)$ can be achieved at any temperature less than 30^0C. A very low annealing temper-

ature could form mismatched products, therefore, the annealing temperature is be fixed to be 30^0C and it also reduces the range of the PCR operating temperature that may lead to a long transition time between the annealing and other two steps of the PCR reaction.

3.2.3 DNA Melting

$$DNA \underset{k_{-m}}{\overset{k_m}{\rightleftharpoons}} S_1 + S_2 \qquad (M - R_1)$$

DNA melting is the reverse part of the DNA annealing reaction $M - R_1$. The kinetics of a short DNA melting can be modeled using the 'all or none' or 'two state' model (Craig et al. (1971); Pörschke and Eigen (1971)). Based on this model a DNA is assumed to be either in a single strand or double stranded state and this assumption is valid only when the number of base pairs is less than 50 (Koehler and Peyret (2005)). Long DNA melting obeys co-operative melting in which different regions of a DNA melt simultaneously in a different manner. Poland - Scheraga (PS) model (Garel and Orland (2004); Jost and Everaers (2009); Richard and Guttmann (2004)) can be used to predict this behavior and identify the different regions that can melt independently. To the best of our knowledge the kinetics of long DNA melting has not been investigated. Mehra and Hu (2005) assumed a rate constant that corresponds to the melting of a short DNA. Gevertz et al. (2006) and Stolovitzky and Cecchi (1996) assumed that DNA melting is always 100% efficient and neglected the melting step in the overall PCR model. Unlike in the annealing step where both

annealing, enzyme binding and extension reactions can occur simultaneously, in the melting step, due to very high temperature, only the melting reaction can occur. Moreover, as long as the given DNA does not form a secondary structure, it can melt completely. Based on these assumptions, in our study we have also assumed that the DNA melting reaction is always 100% efficient in our PCR model. This approach allows us to simplify the treatment of the melting step; however, it does not consider the effect of template melting during the extension step at 72^0C which is moderately a high temperature at which a long DNA may melt. Furthermore, in applications like COLD PCR, the melting temperature is lower than the typical PCR DNA melting temperatures (Li et al. (2008)). In order to account all these factors, the temperature-dependent melting rate constants need to be estimated. Although we did not consider these effects in the numerical simulations of the present study, we present a method that can be used to model the kinetics of long DNA melting.

3.2.3.1 Statistical Mechanical Model for the Kinetics of the Melting of a long DNA

As mentioned above, long DNA molecules melts based on co-operative melting. Using the Poland - Scheraga (PS) model, a given DNA sequence can be divided in to the following 5 discrete domains.

1. Internal Loops.

2. Ends

3. Pre Existing coils

4. Expansion of loops

5. Coalescence of neighboring loops

There are many numerical methods and software such as MELTSIM (Blake and Delcourt (1998); Blake et al. (1999); Dwight et al. (2011)) developed to identify the above mentioned domains and solve P-S model for a given long DNA sequence. Once these domains are found, for each domain an overall stability constant is calculated based on the following equation.

$$K_{loop} = \sigma_c f(N) \prod_{i=1}^{N} s_i \tag{3.3}$$

In the above Eq. (3.3), σ_c is referred as a co-operative parameter (Blake et al. (1999)) and it is different from the nucleation parameter σ that has been discussed in annealing model. σ_c is a penalty to statistical weight for melting of domain due to free energy cost of dissociating an internal base pair. Both σ_c and $f(N)$ have been universally estimated and Blake et al. (1999) provided the expressions to calculate them. Each of the above regions melts independently based on the two-sided melting theory (Azbel (1979)). These regions can be identified and the overall stability constant can be estimated as explained above. Once the regions have been identified, the relaxation time of the melting of overall DNA can be found as follows.

3.2.3.2 Relaxation Time of a long DNA

As shown in Section 3.2.1 an exact state space model for the melting of each base pair in each domain can be formulated. Since we know all the domains based on P-S model, now the state space system of each domain can be connected to find the state space system of overall DNA melting. The state space matrix of each domain will be coupled to only one other block, and in the following wayassuming that domain $i+1$ melts after domain i, only the fully molten domain i, fully annealed $i+1$ state will be coupled to all the fully molten i, single base dissociated $i+1$ states. Based on the type of a domain, the state space system for each domain can be modeled using one-sided or two-sided melting. Once the state space matrix of the overall DNA melting is formed, from the largest eigenvalue of the state space matrix, the relaxation time can be estimated. The following steps can be used to find the rate constants of long DNA melting.

1. Using the MELTSIM identify different domains for a given DNA sequence.

2. Order the domains based on their melting temperatures (ascending order).

3. Construct the state space matrices based on two sided or one sided matrix for each domain.

4. Associate the σ. the nucleation parameter with individual states in the

last block according to the method described in Chapter 2.

5. Diagonalize each block and rank order eigenvalues of all blocks.

6. For each domain find the relaxation time using $\frac{-1}{max(\lambda_i)}$

7. Compare all the relaxation time and maximum value of the relaxation time is the relaxation time of the whole DNA sequence

3.2.4 Enzyme Binding Kinetics

In extension step, enzymatic addition of nucleotides converts the duplex (SP) into a complete dsDNA. Both deterministic as well as stochastic (Velikanov and Kapral (1999)) approaches have been proposed to develop a model for the extension reaction. Velikanov and Kapral (1999) presented the following chemical master equation (CME), a probabilistic description of the extension reaction system, with its analytical solution:

$$\frac{\partial P\left(l,t\right)}{\partial t} = w_{l-1\rightarrow l}P\left(l-1,t\right) - w_{l\rightarrow l+1}P\left(l,t\right) \tag{3.4}$$

where $P\left(l,t\right)$ denotes the probability distribution of the duplexes with l base pairs inserted in the extension reaction, and $w_{l-1\rightarrow l}$ denotes the transition probability rate of addition from $l-1$ to l base pairs. Even though, the solution of the above CME can provide the time required to complete the extension reaction, this formulation is not useful in the present context because

- It omits enzyme dissociation/processivity as one of its major drawbacks;

hence it only applies to perfectly processive polymerases (see below for discussion of processivity). Since thermo-stable enzymes are not perfectly processive, it cannot be applied to PCR.

- It cannot be integrated with the models for other steps of PCR. Therefore, it is impossible to analyze annealing, enzyme binding and extension reactions simultaneously. Therefore, we consider an alternative approach with an appropriate reaction mechanism to develop a model for enzyme binding and extension reaction.

3.2.4.1 Reaction Mechanism:

There are several reaction mechanisms proposed for the enzyme binding and extension reactions (Patel et al. (1991); Boosalis et al. (1987); Mendelman et al. (1990); Huang et al. (1992)) and Fig. 3.2 represents a general reaction mechanism (Brown and Suo (2009)). Note that the rate constants in this reaction mechanism are different from those presented in Section 3.2 and Chapter 2. In step 1, enzyme binds with D_i molecule to form a binary complex $E.D_i$. In step 2, a deoxynucleotide triphosphate (dNTP) binds with $E.D_i$ to form a ternary complex, $E.D_i.dNTP$, which undergoes a protein conformational change in step 3 and forms $E'.D_i.dNTP$. In step 4, a the nucleotide is incorporated and a pyrophosphate molecule is released from D_i and as a result $E'.D_{i+1}.PPi$ is formed. $E'.D_{i+1}.PPi$ undergoes a conformational change in step 5 leading to the formation of $E.D_{i+1}.PP_i$. In step 6, PP_i is completely

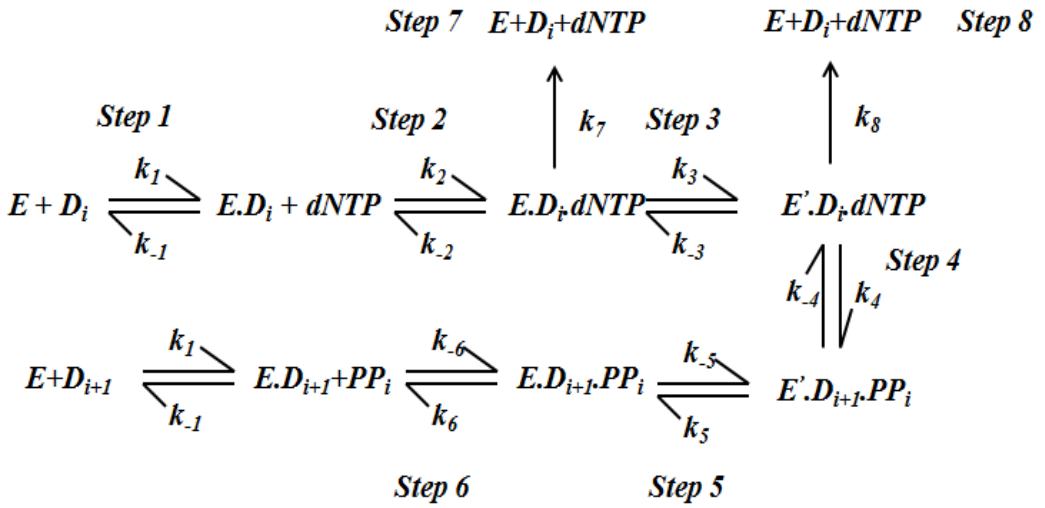


Figure 3.2: A general reaction mechanism of Enzymatic Primer Extension reaction. Note that the rate constants in this reaction mechanism are different from those presented in Section 3.2 and Chapter 2

released from $E.D_{i+1}.PP_i$ and $E.D_{i+1}$ is formed. Finally the dissociation of $E.D_{i+1}$ produces D_{i+1} and E. Besides these steps, there are parallel dissociation reactions represented by step 7 and 8 would also possible to occur. Kuchta et al. (1987), Patel et al. (1991), Brown and Suo (2009), Capson et al. (1992), Fiala and Suo (2004) studied the extension reaction kinetics for *DNA polymerase I Klenow*, *T7 DNA polymerase*, *S. solfataricus P2 DNA polymerase B1*, *T4 gene 43 protein*, and *S. solfataricus P2 DNA polymerase IV*, respectively, at either 20^0C or 37^0C. Using their rate constant data, we simplify the above reaction mechanism. Step 6 is the last step of the reaction mechanism that produces $E.D_{i+1}$. According to Patel et al. (1991), $k_6 = 1000\ s^{-1}$ and $k_{-6} = 0.5(\mu M)^{-1}s^{-1}$. These rate constant values suggest that the association of $E.D_{i+1}$ with PP_i is impossible. In addition to this comparing k_{-6} with $k_1(11(\mu M)^{-1}s^{-1})$ and $k_2(> 50(\mu M)^{-1}s^{-1})$, it can be negligible. Hence, the

final step 6 is irreversible with a rate constant k_6. Step 3, 4 and 5 which are all a first order reversible reactions represent the conformational change of a ternary complex and their rate constants values are higher than k_2 (Patel et al. (1991)) and the forward rate constants of each step is higher than the reverse rate constant (Patel et al. (1991)). Hence, the overall dynamics is controlled by step 2 which forms a ternary complex, $E.D_i.dNTP$ and the final step is irreversible. Thus, as proposed by Boosalis et al. (1987); Mendelman et al. (1990); Huang et al. (1992), the above reaction mechanism can be represented using the simplified reaction schemes given by reaction $(E - R_n)$.

$$E + D_i \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} E.D_i + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.D_i.N] \overset{k_{cat}}{\rightarrow} ED_{i+1} \underset{k_1}{\overset{k_{-1}}{\rightleftharpoons}} E + D_{i+1} \qquad (E - R_n)$$

We now show how the kinetic parameters k_1, k_{-1}, k_{cat} and $K_N = (k_{-2}+k_{cat})/k_2$ in this reaction scheme can be estimated for any polymerase using polymerase processivity and initial rate experiments.

3.2.4.2 Single-hit conditions

In single hit conditions, enzyme concentrations are sufficiently low that the probability of re-association is approximately zero. Therefore, they do not allow enzyme re-association. Hence enzyme-template association occurs only during the initial equilibration of enzyme with SP. Thus, the following reac-

tion scheme for the addition of n base pairs is written

$$E + D_0(SP) \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} E.D_0 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.SP.N] \overset{k_{cat}}{\rightarrow} ED_1 \qquad (E - R_1)$$

$$E + D_i \overset{k_{-1}}{\longleftarrow} E.D_i + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.D_i.N] \overset{k_{cat}}{\rightarrow} ED_{i+1} \quad \forall i = 1, 2, ...n-2 \quad (E - R_i)$$

$$E + D_{n-1} \overset{k_{-1}}{\longleftarrow} E.D_{n-1} + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.D_{n-1}.N] \overset{k_{cat}}{\rightarrow} E.D_n \overset{k'_{cat}}{\rightarrow} E + D_n \quad (E - R_{n-1})$$

Single hit conditions are used to estimate polymerase processivity parameters (with an appropriate kinetic model and associated rate constants or equations relating the rate constants). Processivity is defined as the number of nucleotide incorporated per DNA-enzyme binding event.

3.2.4.3 Processivity of an enzyme

Let i index be the sequence positions on the template. In a Markov chain formulation of dissociation, the index i at which dissociation occurs is called the stopping time/index and is denoted i_{off}. Let p denote the conditional probability of the polymerase not dissociating at position/time i, given that is was bound to the template at position/time $i-1$. The probability of dissociation at position/time i is then

$$p_{off}(i) = (1-p)\, p^{i-1} \qquad (3.5)$$

p is called the microscopic processivity parameter. The expected position of dissociation of the polymerase (expected stopping time) sometimes called the processivity, can be written as

$$E\left[i_{off}\right] = \frac{1}{(1-p)} \tag{3.6}$$

$E\left[i_{off}\right]$ is sometimes reported as the processivity instead of the microscopic processivity parameter. The above expression is derived for a template of infinite length. Usually, in processivity experiments long templates are used to estimate p. For finite length,

$$p_{off}\left(n\right) = 1 - \sum_{i=0}^{n-1} \left(1-p\right) p^{i-1} \tag{3.7}$$

For heterogeneous templates, p will vary with position. From processivity experiments, one can obtain the p at each position since we will have

$$p_{off}\left(n\right) = \left(1-p_i\right) \prod_{j=0}^{i-1} p_j \quad \forall i \tag{3.8}$$

These equations can be used to solve uniquely for each p_i. However, it is impractical to do processivity experiments for each new template. Hence, one can do processivity experiments on templates with different types of nearest neighbor motifs (including hairpins) for a given polymerase, and then nearest neighbor processivity parameters can be used in modeling of an arbitrary sequence.

3.2.4.4 Relationship between Processivity and Enzyme binding and Extension rate constants

Now, at a fixed temperature, we seek a relationship between processivity of an enzyme and the rate constants of the reaction scheme $(E - R_i)$. In order to do this, we write the state space model for the reaction scheme $(E - R_i)$. We omit $E.D_n \overset{k'_{cat}}{\rightarrow} E + D_n$. from the state space model for simplicity as it does not affect equilibrium and we are not estimating the corresponding rate constant. Since the substrate, dNTP, concentration is always in excess compared to enzyme, Michaelis-Menten (MM) kinetics is valid and hence the steady state assumption for the intermediate concentration is valid. Therefore,

$$\frac{d}{dt}\left[E.D_i.N\right] = -\left[E.D_i.N\right]\left(k_{-2} + k_{cat}\right) + k_2\left[E.D_i\right]\left[N\right] = 0 \tag{3.9}$$

$$\implies \left[E.D_i.N\right] = \frac{\left[E.D_i\right]\left[N\right]}{K_N} \tag{3.10}$$

Let $k = \frac{k_{cat}}{K_N}\left[N\right]$, then the state space matrix A of the reaction scheme $(E - R_i)$ is given as

$$\frac{dx}{dt} = Ax \implies \frac{dx}{d\left(t\left(k^1 + k_{-1}\right)\right)} = \frac{1}{k + k_{-1}}A \implies \frac{dx}{dt} = A'x, \tag{3.11}$$

where

$$x = [E.D_0, D_0, E.D_1, D_1, ...E.D_i, D_i, ..., D_{n-1}, E.D_n]$$

$$A' = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{k_{-1}}{k_{-1}+k} & 0 & 0 & 0 & 0 & \cdots & 0 \\ \frac{k}{k_{-1}+k} & 0 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{k_{-1}}{k_{-1}+k} & 0 & 0 & \cdots & 0 \\ 0 & 0 & \frac{k}{k_{-1}+k} & 0 & -1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 0 & 0 & \cdots & 0 \end{bmatrix} \tag{3.12}$$

$k_{-1}dt$: conditional probability of transition from state $E.D_i \rightarrow E + D_i$ in time dt

kdt : conditional probability of transition from state $E.D_i \rightarrow E.D_{i+1}$ in time dt

In a single molecule continuous time Markov chain formulation, Eq. (3.11) can be written in terms of the probability distribution of states $[\rho_{D_0}, \rho_{E.D_0}, ...]$ instead of the vector of species concentrations. An equivalent master equation formulation is:

$$\begin{aligned} \frac{\partial}{\partial t}\rho\left(0,t\right) &= \left(k + k_{-1}\right)\rho\left(0,0\right) \\ \frac{\partial}{\partial t}\rho\left(1,t\right) &= k_{-1}\rho\left(0,0\right) \\ \frac{\partial}{\partial t}\rho\left(2,t\right) &= k\rho\left(0,0\right); \rho\left(0,0\right) = 1; \end{aligned} \tag{3.13}$$

where $\rho\left(i,t\right)$ denotes the probability of the polymerase being in state i at time t. The equilibrium distribution of this master equation can be obtained by

solving for the generalized eigenvectors of the state space system Eq. (3.11). The eigenvector corresponds to the zero eigenvalue that represent the equilibrium distribution of the reaction species is given below.

$$\pi = \left[0, \frac{k_{-1}}{k+k_{-1}}, 0, \frac{kk_{-1}}{(k+k_{-1})^2}, 0, \frac{k^2k_{-1}}{(k+k_{-1})^3}, ..., 1-\sum_{i=1}^{n-1}\frac{k^{i-1}k_{-1}}{(k+k_{-1})^i}\right]^T \tag{3.14}$$

It is found that this distribution has the form specified by Eq. (3.5) with $p_{off}(i) = p_{off}(i, t=\infty)$and the following value of the microscopic processivity parameter:

$$p = \frac{\frac{k_{cat}}{K_N}[N]}{\frac{k_{cat}}{K_N}[N]+k_{-1}} \tag{3.15}$$

Hence,

$$\frac{k_{-1}}{\frac{k_{cat}}{K_N}[N]+k_{-1}} = (1-p) \implies k_{-1} = \frac{\frac{k_{cat}}{K_N}[N](1-p)}{p} \tag{3.16}$$

and Eq. (3.16) can be written as follows in terms of processivity:

$$k_{-1} = \frac{k_{cat}}{K_N}[N]\left(\frac{1}{E[i_{off}]-1}\right) \tag{3.17}$$

As per Eq. (3.16), if $\frac{k_{cat}}{K_N}$ and processivity of a polymerase is known at a specific temperature, it is possible to estimate k_{-1} for any polymerase. Eq. (3.16) is valid under the approximations applied in the derivation of reaction $(E-R_n)$. A comprehensive analysis of these approximations and more general models

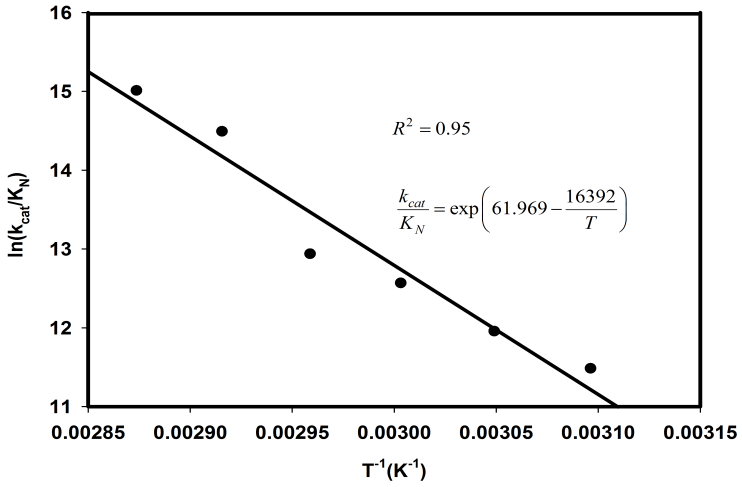


Figure 3.3: Temperature dependence of Extension reaction rate constant for Taq polymerase

.

are considered in a separate work. For each such model, an equation for k_{-1} analogous to reaction $(E-R_n)$ can be derived based on the associated single hit A matrix, in terms of processivity and other model parameters. We estimated $\frac{k_{cat}}{K_N}$ for Taq polymerase at various temperatures based on a bireactants MM kinetics formulation. Fig. 3.3 shows the temperature dependent extension rate constant $\frac{k_{cat}}{K_N}$. In Eq. (3.17) k_{-1} and$\frac{k_{cat}}{K_N}$ are concentration independent terms and hence, the processivity $E\left[i_{off}\right]$ or the conditional probability p depends on $[N]$. In order to use Eq. (3.17) to estimate k_{-1}, $[N]$ and $E\left[i_{off}\right]$ should be consistent or one should use the value $[N]$ at which $E\left[i_{off}\right]$ is estimated. Wang et al. (2004) and Davidson et al. (2003) determined the value of at a specific temperature and nucleotide concentration. The following Table 3.2 provides the values of $E\left[i_{off}\right]$ and the conditions at which they are measured. Using the above values, $\frac{kcat}{KN}$ and Eq. (3.16), we have estimated k_{-1} at 60 ^{0} C and

Table 3.2: Processivity of Taq polymerase

Reference	Temperature (^0C)	$[N]$ (μM)	$E\,[i_{off}]$
Wang et al. (2004)	72	250	22
Davidson et al. (2003)	60	800	50-80

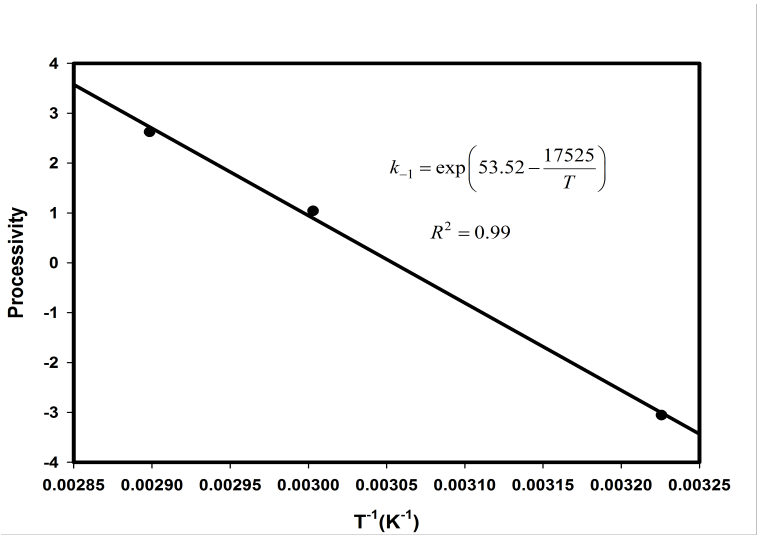


Figure 3.4: Temperature dependence of Enzyme dissociation rate constant for Taq polymerase

.

72 ^{0}C, respectively. We have the k_{-1} for *S. solfataricus* P2 DNA *polymerase* B1(36) at 37 ^{0} C and we use the same value for Taq polymerase enzyme as their equilibrium constants are of the same order of magnitude. Thus, we could obtain k_{-1} at three different temperatures and hence an Arrhenius relationship is fitted as shown in Fig. 3.4 to estimate the temperature dependent dissociation rate constant k_{-1}. Fig. 3.6 explains the steps involved in enzyme binding and extension model parameter estimation. Equilibrium thermodynamic analysis for enzyme binding reaction has been done extensively with *Thermus aquaticus* enzyme by Data and LiCata (2003). They estimated the temperature dependent equilibrium constant which is a ratio of $\frac{k_1}{k-1}$. There-

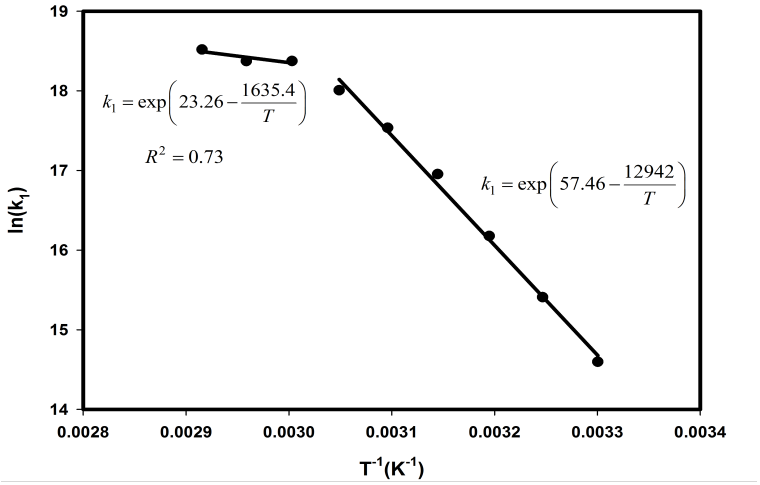


Figure 3.5: Temperature dependence of Enzyme binding rate constant for Taq polymerase

.

fore, based on the temperature dependence of k_{-1} and $K_{binding}$, k_1 has been estimated and its Arrhenius plot is shown in Fig. 3.5. Fig. 3.6 explains the steps involved in enzyme binding and extension model parameter estimation.

3.3 Analysis of PCR Kinetics

Using the kinetic model developed in Section 3.2, we seek an optimal temperature vs. time profile. In order to do this, we first analyze the kinetic model in this Section to assess the importance of such simulations in making accurate predictions of the amplification efficiency of PCR reactions. In particular, we demonstrate that the conventional picture of PCR kinetics, which assumes a single reaction is rate-limiting for each step, and which presumably estimates reaction temperatures and times for each step based on this assumption without solving the associated state equations, is highly inaccurate. Except for

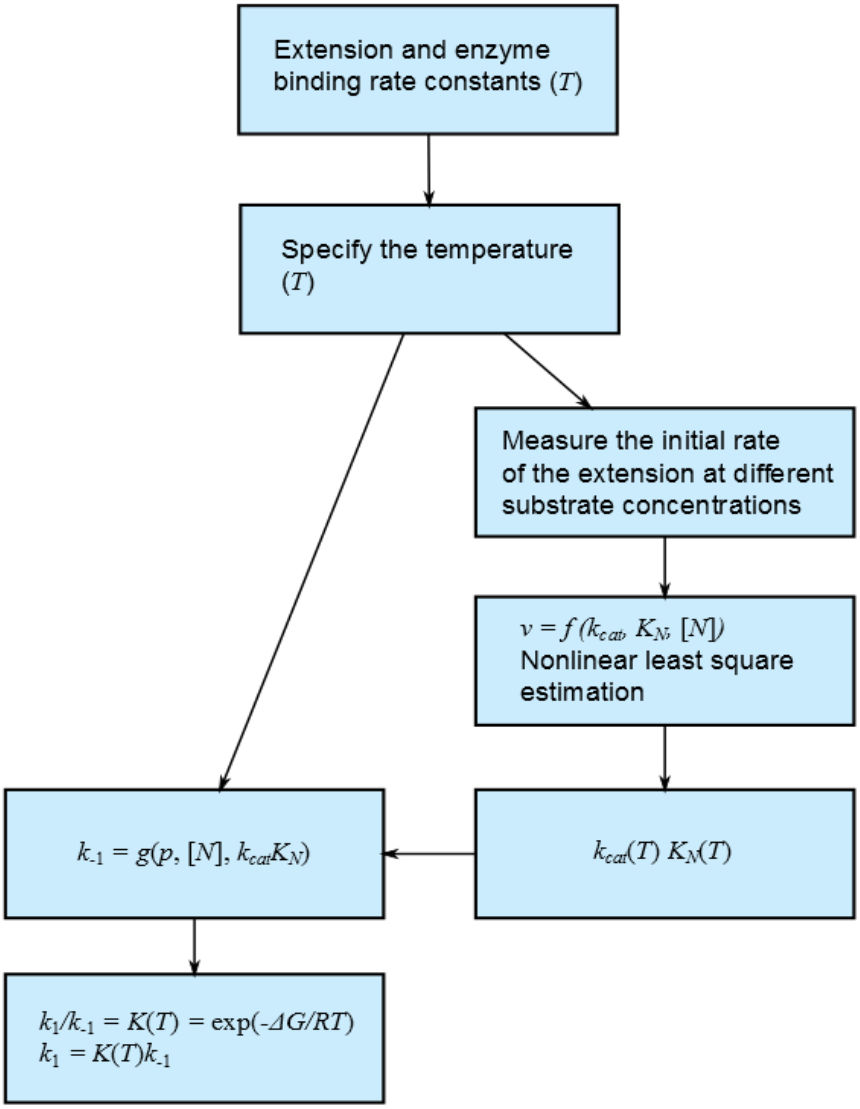


Figure 3.6: Estimation of enzyme binding and extension rate constants

.

the annealing temperature, the reaction conditions are the same as the typical PCR conditions recommended by, for example, Invitrogen (2006). Reactions $(A - R_1)$, $(M - R_1)$, and $(E - R_n)$ have been written for a simplex PCR reaction and they are given along with their state equations below.

3.3.1 PCR Reactions

3.3.1.1 Melting Reaction

$$DNA \underset{k_{-m}}{\overset{k_m}{\rightleftharpoons}} S_1 + S_2 \qquad (M_1)$$

3.3.1.2 Annealing Reactions

$$S_1 + P_1 \underset{k_{rS_1}}{\overset{k_{fS_1}}{\rightleftharpoons}} S_1P_1 \qquad (A_1)$$

$$S_2 + P_2 \underset{k_{rS_2}}{\overset{k_{fS_2}}{\rightleftharpoons}} S_2P_2 \qquad (A_2)$$

$$S_1 + S_2 \underset{k_{-m}}{\overset{k_m}{\rightleftharpoons}} DNA \qquad (A_3)$$

3.3.1.3 Extension Reactions

$$S_1P_1 + E \underset{k_e}{\overset{k_{-e}}{\rightleftharpoons}} E.S_1P_1 \qquad (E_1)$$

$$S_2P_2 + E \underset{k_e}{\overset{k_{-e}}{\rightleftharpoons}} E.S_2P_2 \qquad (E_2)$$

$$E.S_1P_1 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.S_1P_1.N] \overset{k_{cat}}{\rightarrow} ED_1^1 \underset{k_{-e}}{\overset{k_e}{\rightleftharpoons}} D_1^1 + E \qquad (E_3)$$

$$E.S_2P_2 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.S_2P_2.N] \overset{k_{cat}}{\rightarrow} ED_1^2 \underset{k_{-e}}{\overset{k_e}{\rightleftharpoons}} D_1^2 + E \qquad (E_4)$$

$$ED_1^1 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.D_1^1.N] \overset{k_{cat}}{\rightarrow} ED_2^1 \underset{k_{-e}}{\overset{k_e}{\rightleftharpoons}} D_2^1 + E \qquad (E_5)$$

$$ED_1^2 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [ED_1^2.N] \overset{k_{cat}}{\rightarrow} ED_2^2 \underset{k_{-e}}{\overset{k_e}{\rightleftharpoons}} D_2^2 + E \qquad (E_6)$$

$\vdots$

$$ED_{n-1}^1 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.D_{n-1}^1.N] \overset{k_{cat}}{\rightarrow} ED_n \qquad (E_7)$$

$$ED_{n-1}^2 + N \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} [E.D_{n-1}^2.N] \overset{k_{cat}}{\rightarrow} ED_n \qquad (E_8)$$

$$ED_n \overset{k_{cat}^{'}}{\rightarrow} E + DNA \qquad (E_9)$$

3.3.2 State Equations

3.3.2.1 Melting

$$\frac{d[DNA]}{dt} = -k_m[DNA] + k_m[S_1][S_2] \qquad (3.18)$$

3.3.2.2 Annealing

$$\frac{d[S_iP_i]}{dt} = k_{fP_i}[S_i][P_i] - k_{rP_i}[S_iP_i] \qquad \forall i = 1,2 \tag{3.19}$$

$$-\frac{d[P_i]}{dt} = \frac{d[S_iP_i]}{dt} \qquad \forall i = 1,2 \tag{3.20}$$

$$\frac{d[DNA]}{dt} = -k_m[DNA] + k_{-m}[S_1][S_2] \tag{3.21}$$

$$\frac{d[S_i]}{dt} = -k_{fP_i}[S_i][P_i] + k_{rP_i}[S_iP_i] - \frac{d[S_1S_2]}{dt} \qquad \forall i = 1,2 \tag{3.22}$$

3.3.2.3 Extension

$$\frac{d[S_iP_i]}{dt} = -k_e[E][S_iP_i] + k_{-e}[E.S_iP_i] \qquad \forall i = 1,2 \tag{3.23}$$

$$\frac{d[E.S_iP_i]}{dt} = k_e[E][S_iP_i] - k_{-e}[E.S_iP_i] - \frac{k_{cat}}{K_N}[E.S_iP_i][N] \qquad \forall i = 1,2 \tag{3.24}$$

$$\frac{d[E.T_i]}{dt} = \frac{k_{cat}}{K_N}\left([E.T_{i-1}] - [E.T_i]\right)[N] + k_e[E][T_i] - k_{-e}[E.T_i] \tag{3.25}$$

$$T_i = D_1^j, D_2^j, ...D_{n-1}^j \& j = 1,2$$

$$\frac{d[D_i^j]}{dt} = -k_e[E][D_i^j] - k_{-e}[E.D_i^j] \qquad \forall i = 1,2,....n-1 \& j = 1,2 \tag{3.26}$$

$$\frac{d[E.D_n]}{dt} = \frac{k_{cat}}{K_N}\left([E.D_{n-1}^1] + [E.D_{n-1}^2]\right)[N] - k_{cat}^{`}[E.D_n] \tag{3.27}$$

$$\frac{d[DNA]}{dt} = k_{cat}^{`}[E.D_n] \tag{3.28}$$

$$\frac{d[N]}{dt} = -\frac{K_{cat}}{K_N}[N]\left(\sum_{T=S_1P_1}^{D_{n-1}^1}[ET] + \sum_{T=S_2P_2}^{D_{n-1}^2}[E.T]\right) \tag{3.29}$$

$$\frac{d[E]}{dt} = -k_e[E]\left\{[S_1P_1] + [S_2P_2]\right\} + k_{-e}\left\{[E.S_1P_1] + [E.S_2P_2]\right\} + \frac{d[DNA]}{dt} \tag{3.30}$$

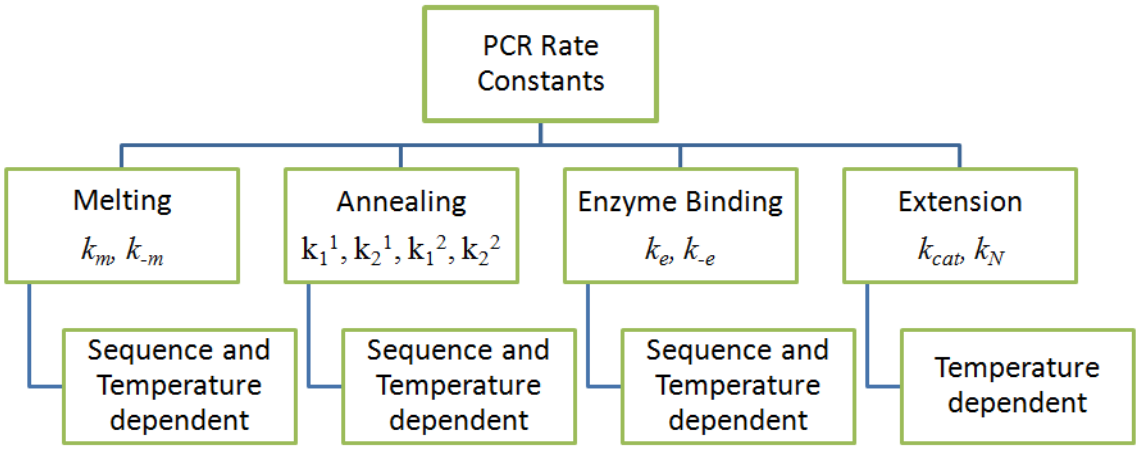


Figure 3.7: Sequence and Temperature dependent PCR Model (Rate Constants)

.

In Fig. 3.7.we have shown which rate constants are sequence and temperature dependent. We summarize the simulation results as follows (data not shown here).

- If the ratio of single strand concentration to primer concentration (SP ratio) $S0/P0$ is ¡1, then, the annealing reaction nearly instantaneous.
- If the SP ratio is close to 1, then there is a transient behavior in the evolution of S_1P_1.

When the SP ratio is low, since the primer concentration is very high compared to the single strand concentration, the primer molecule easily binds on the single strand molecule and does not allow single strands to anneal to each other. On the other hand, when the SP ratio is 1, since the primer and single strand molecules are equal in concentration, there is a competition between them to

anneal on their respective complementary sequences. Since the single strands participate in a two-way competition with both primers, they eventually lose in this competition. Thus, the annealing reaction is not the rate-limiting step during the early stages of PCR but it may be the rate limiting step towards the end of PCR. In PCR, study of the annealing reaction separately may be misleading in deriving conclusions about the optimal annealing time. From Section 3.2.4, it is evident that enzyme binding can occur at the annealing temperatures. This can affect the annealing and hence the overall dynamics of the PCR. Data and LiCata (2003) experiments reveal that the Gibbs free energy of the enzyme binding reaction goes through the minimum at around 50 ^{0}C. In a conventional PCR model, however the enzyme binding is considered to occur during the extension reaction (Mehra and Hu (2005); Hsu et al. (1997)). In the next section we motivate full time-varying models by showing that simultaneous annealing and extension reactions, result in significant differences in reaction efficiency that can be exploited through such modeling. These effects are not captured in conventional models of PCR kinetics, as will be considered in more generality in the next section.

3.3.3 Combined Annealing and Extension

At any instant, since all the reactants for annealing, enzyme binding and extension reactions are available in the reaction mixture, these reactions in principle can occur simultaneously. The combined annealing and extension model

allows us to simulate arbitrary PCR reaction cycling protocols that do not follow the standard 3-step scheme, hence extending beyond the types of on-off behavior commonly assumed in models of PCR. Due to this phenomenon, molecular biologists often run PCR reaction with only two steps per cycle - one step for melting and one step for annealing/extension. However, there is no quantitative prescription available for the temperatures of the annealing/extension steps. In order to provide such prescriptions, in this section we do not distinguish between the annealing and extension steps and solve the kinetic equations corresponding to all these steps together for the overall reaction time. This is one of the main reasons for the need of temperature-dependent rate parameters. The reaction conditions are the same as those in a typical PCR. Annealing and extension times are fixed to be 45 and 30 seconds, respectively, and the length of the target DNA is assumed to be 500 base pairs (bp). The extension reaction temperature is 72 ^{0}C. At a given time since any one of the 3 steps of a PCR kinetically dominating, the rates of the all the reactions are not uniform. This difference in reaction rate creates a stiff state space system that needs to be solved carefully. We used a MATLAB routine *ode15s* (MATLAB (2012)) to solve this system of stiff differential equations. Though the annealing reaction is very fast at low temperatures, its efficiency is determined by the kinetics of the enzyme binding reaction. Therefore, even at high annealing temperatures at which the equilibrium conversion of the annealing reaction in the absence of enzyme binding is lower, it is possible to obtain 100% overall efficiency. The evolution of single strands, single strand-primer

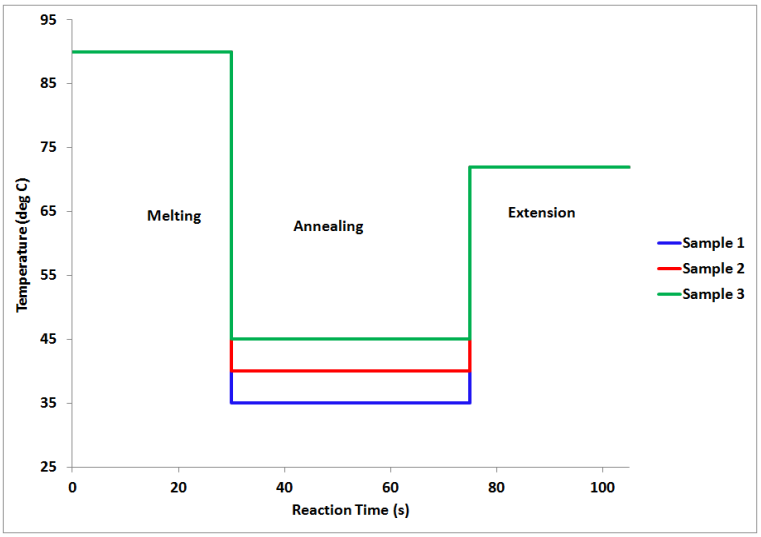


Figure 3.8: Three different temperature cycling samples.

duplex and final DNA for a single cycle at the annealing temperatures 35 ^{0}C, 40 ^{0}C and 45 ^{0}C are presented in Fig 3.9. The temperature cycling profile for each annealing temperature is shown in Fig 3.9. At annealing temperature of 40 ^{0}C the annealing reaction can't reach 100% equilibrium conversion. The equilibrium conversion of the enzyme binding reaction at this temperature is less than 70%. Nevertheless, the overall PCR conversion is nearly 100% at 40 ^{0}C. This is due to the combined annealing, enzyme binding and extension reactions. As soon as the Taq polymerase binds to the SP duplexes, the extension reaction or $E.D_i$ dissociation can occur, and hence the equilibrium of the enzyme binding reaction is disturbed and more enzymes bind to the SP duplexes. The extension reaction rate increases when temperature increases.

Thus, during the annealing step all SP is converted into $E.D_i$ molecules which then dissociate into D_i. Fig. 3.9 shows the concentration profile of sum of concentration of D_i. During the annealing step, since the enzyme dis-

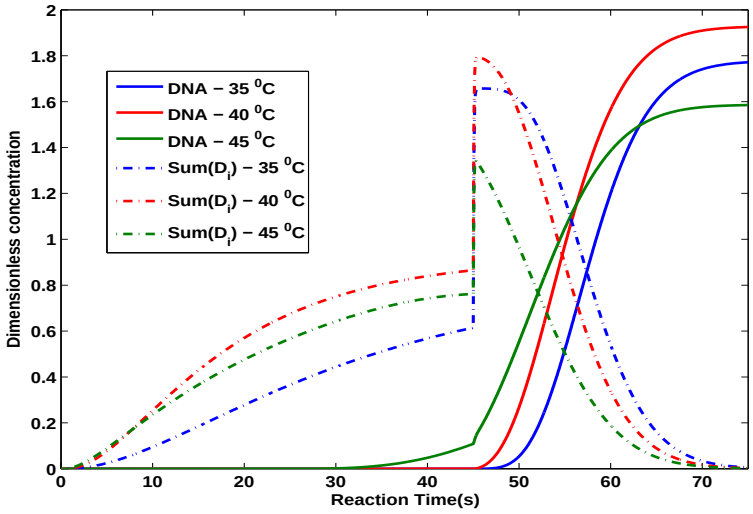


Figure 3.9: Transient behavior of reaction constituents (Di and DNA molecules) for Primer set 1. The annealing temperatures are 35 ^{0}C, 40 ^{0}C and 45 ^{0}C and the length of the target is 1000 bp. Annealing time is 45 seconds and extension time is 30 seconds. Primer, enzyme, and dNTP concentrations are 0.2 μM, 10 nM, 800 μM, respectively.

.

sociation rate constant is comparable to the extension rate constants, $E.D_i$ molecules dissociate. When the temperature of the reaction is increased to 72 ^{0}C during the extension step, equilibrium of $E.D_i$ dissociation is disturbed by the rapid nucleotide addition and eventually all D_i molecules are converted to $E.D_i$ which are in turn converted into target DNA. The melting temperature of the primer in this study is less than 35 ^{0}C and as a result of this combine annealing and extension 40 ^{0}C annealing time produced more DNA. However, very high annealing temperature such as 45 ^{0}C didn't produce more DNA. Even though the equilibrium conversion of annealing at 30 and 35 ^{0}C greater than 40 ^{0}C, low temperature annealing didn't yield more DNA due to slower enzyme binding reaction rate at these temperatures. Thus, it is evident that annealing temperature shouldn't be fixed based on the equilibrium

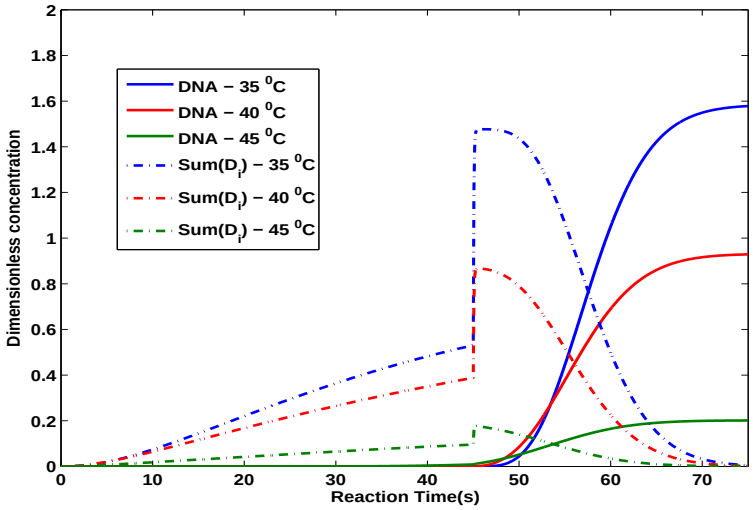


Figure 3.10: Transient behavior of reaction constituents for Primer set 2. The annealing temperatures are 35 ^{0}C, 40 ^{0}C and 45 ^{0}C and the length of the target is 500 bp. Annealing time is 45 seconds and extension time is 30 seconds. Primer, enzyme, and dNTP concentrations are 0.2 μM, 10 nM, 800 μM, respectively.

.

melting temperature and it should be determined by considering the kinetics of the PCR. We have repeated the above analysis for a different set of primer sequences of the same length and same reaction conditions. The sequences are given below and the rate parameters have been calculated as explained in Section 2. Fig. 3.10 shows the evolution of DNA and sum of D_i molecules. Unlike in the above case, in this case the favorable annealing temperature is 35 ^{0}C. Furthermore for the same overall reaction time the overall conversion is different for the two cases. This suggests that the sequence-dependent kinetic model is important.

Forward primer - AATAGCTGTAACTG.

Reverse primer - TTCTTCTGAAACTG.

In the above study, the enzyme concentration is in excess compared to the

single strand concentration. If this condition does not hold, which is the case for the later stages of PCR, the kinetics could be very different. Also, during every PCR cycle, the target DNA concentration increases. Due to this the overall number of nucleotide additions will also increase. Therefore, the reaction conditions that have been maintained during the initial stages of PCR may not be appropriate for the later stages of the PCR. This effect is more pronounced for longer sequences. It should be noted that above discussed results are applicable to the first cycle of a PCR. Furthermore, note that even though our model considers the melting of SP molecules during the extension reaction, it didn't consider the melting of D_i molecules during the extension reaction for the following reasons:

- Our simulation results suggest that even though $\sum_i D_i$ is considerable at the end of annealing step as shown in Fig. 3.9 and 3.10, the summation of the concentrations of D_1 to D_{15} molecules is negligible. Therefore, we assume that these molecules arent present in the reaction mixture.

- The stability/ melting temperature of a duplex increases when the number of base pairs increases. In the present study we have considered a primer of length 14 base pairs. Therefore, the minimum number of base pairs in D_i is greater than 30 and we assume that these duplexes are stable at 72 ^{0}C.

Thus in this section we have established that

- Even during the annealing step, enzyme binding and extension reactions

can occur simultaneously. Hence the state equations of annealing, enzyme binding and extension reaction should be solved together. As will be shown below, it is possible to exploit these simultaneous reactions to improve PCR reaction efficiency through appropriate model-implied choices of temperature cycling strategies.

- As an example, the kinetic model for annealing and extension can provide a quantitative prescription for two-step PCR (melting and combined annealing/extension).

- There should be an optimal annealing temperature at which the reaction is fastest and reaches 100% completion. Importantly, this temperature cannot be computed based on primer melting temperatures alone.

- When the length of the target DNA increases, or in other words, when more nucleotides must be added, the reaction temperature should be higher and reaction time should be increased. Again, the kinetic model for the annealing and extension can provide quantitative prescription for annealing time and temperature.

3.3.4 Geometric Growth

For a fixed annealing and extension time of 45 and 30 seconds, respectively, the state equations of the PCR reaction scheme have been solved at three different temperatures. The temperature cycling profile is shown in Fig. 3.11. Fig. 3.12 shows the geometric growth of DNA concentration. When the annealing

temperature is 40 ^{0}C, the DNA concentration saturates at around 35 nM after 25 cycles. On the other hand, at 45 ^{0}C annealing temperature, after 29 cycles, the DNA concentration is approximately equal to 35 nM. Even though the efficiency at 35 ^{0}C annealing temperature is higher than at 45 ^{0} in the first cycle as shown in Fig. 3.9, when the number of cycles increases, the DNA concentration differs. At 50 ^{0}C, the DNA concentration is much lower than that at other temperatures. Furthermore, it should be noted that when the target DNA concentration is comparable with enzyme concentration, the dynamics of the PCR reaction depends on the temperature and concentration. Ideally, the maximum concentration of the target DNA should be equal to primer concentration. Therefore, during the beginning stage of the PCR, target DNA concentration is the limiting reactant. Once the target DNA concentration exceeds the enzyme concentration, the latter is the limiting reactant. From Fig. 3.12 it is clear that in the second stage the PCR efficiency is lower and a different reaction condition needs to be maintained to improve the efficiency.

3.4 Conclusion

In this chapter, we have developed the first sequence- and temperature-dependent kinetic model for DNA amplification, through biophysical modeling of coupled DNA melting and polymerization processes. Using this model, the kinetics of PCR have been analyzed for various temperature cycling strategies. Based on the results of this kinetic analysis, the need for systematic optimization of

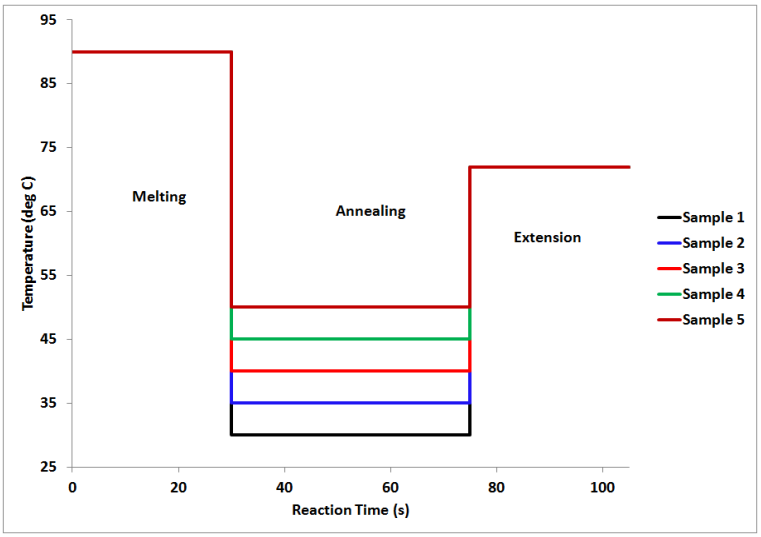


Figure 3.11: Temperature profile for the first cycle at five different annealing temperatures. The same temperature profile is followed for all other cycles.

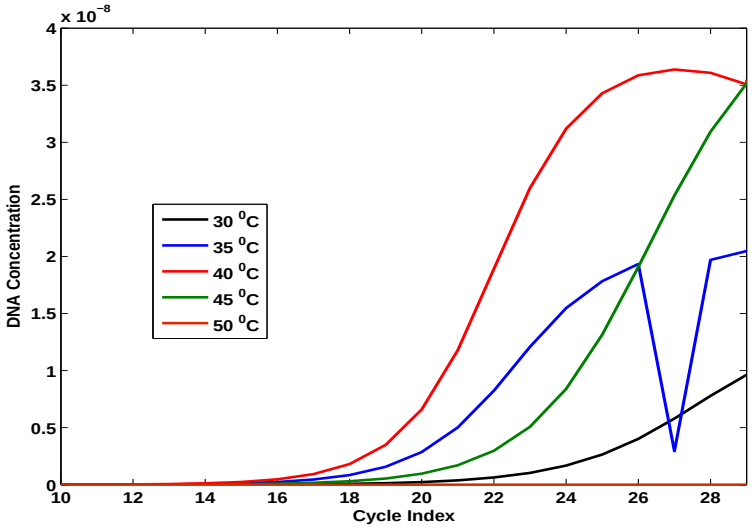


Figure 3.12: Geometric Growth of DNA. Annealing and Extension reaction time in each step of the PCR is 45 and 30 seconds, respectively. The extension temperature is fixed to be 72 ^{0}C. Initial concentration of template, primer, enzyme, and nucleotide is 0.2 μM, 10 nM, 800 μM, respectively.

temperature cycling strategies has been established. The theory of optimal control of dynamical systems (4) provides a framework for the computation of the optimal temperature cycling protocols for DNA amplification. Use of the proposed sequence-dependent kinetic model in a control-theoretic framework should enable determination of the optimal dynamic operating conditions of DNA amplification reactions, for any specified amplification objective. Through the application of this kinetic state space model, it may be possible to i) improve the overall amplification efficiency of the reaction by orders of magnitude for the same number of cycles; ii) substantially reduce the overall time of the reaction compared to conventional PCR protocols.

Chapter 4

Theory of Sequence-dependent DNA Amplification reaction Dynamics and Optimal Control

4.1 Introduction

This chapter is concerned with the establishment of a foundation for the optimal control of DNA amplification reactions, which can be used for the automated computation (rather than qualitative selection) of temperature cycling protocols. This is accomplished through the use of sequence-dependent state space models for DNA amplification dynamics. We use these sequence-dependent kinetic models to formulate amplification of DNA as a problem in control theory with optimal solutions that can differ considerably from strategies typically used in practice. First, the notion of sequence-dependent control

systems of biochemical reaction networks, which have various applications in dynamical systems biology, is introduced. Control systems corresponding to several DNA amplification models are then formulated and compared. Next, we demonstrate by simulation that through the application of sequencedependent kinetic state space models,

- The overall amplification efficiency of the reaction can be improved by orders of magnitude for the same number of cycles.
- The overall time of the reaction can be substantially reduced compared to conventional PCR protocols.

We then formulate an optimal control problem based on the sequence-dependent kinetic model for DNA amplification and specified objectives. Two control problems are formulated: maximization of amplification efficiency for a fixed reaction time and minimization of reaction time for a specified amplification efficiency. The general control problem of DNA amplification is stated and then the specific case of temperature control is discussed.

Besides sequence dependence, another novel feature of control of DNA amplification is the cyclic nature of optimal temperature control strategies. Typically these are assumed to be periodic, with the same control strategy applied every cycle. However, geometric growth leads to dramatic changes in the relative magnitudes of state variables over time. We show based on kinetic modeling why the optimal control strategy for DNA amplification will change depending on the stage of the reaction and how the control problem be

formulated to enable prediction of aperiodic manipulated input functions. A staged approach to optimal control of DNA amplification is hence presented. Control systems for each stage are formulated and strategies for the optimal synthesis are proposed. Control system formulations that reduce the computational complexity of the problem are then considered. Next, it is shown that methods for updating parameter estimates using dynamic filtering can be accommodated within the proposed framework, control system formulations suitable for accurate online parameter estimation are introduced. Finally, prospects for the application of the methodology to other amplification objectives, including the automated design of new types of PCR reactions, through proper specification of the objective function and augmentation of the state space system, are discussed.

4.2 Dynamic models and control systems for DNA amplification

In this section we establish a control theoretic framework for dynamic optimization of DNA amplification. Several classes of control systems, differing in their representations of the sequence and temperature dependence of chemical kinetic rate constants, are introduced. The most general control system is based on the sequence- and temperature-dependent state space model for DNA amplification described in the Methods section. Other control systems

are based on simplified PCR models previously proposed in the literature Mehra and Hu (2005); Hsu et al. (1997); Gevertz et al. (2006); Stolovitzky and Cecchi (1996) and correspond to approximations to the model described in Methods.

For these latter models, we only consider the previously developed model structures and not the rate constants. Marimuthu *et al* (Marimuthu, K., Jing, C., and Chakrabarti, R. Submitted to Biophysical Journal) have shown that the rate constants that were used in the previously proposed models are thermodynamically inconsistent. Therefore, we use the rate constants that are estimated based on the approach described in Methods together with the previously developed model structures.

Denoting by u_i the i^{th} rate constant, which can be manipulated as a function of time, and by x the vector of species concentrations, a general state space model for chemical reaction kinetics can be written as follows:

$$\frac{dx}{dt} = f(x,u) = g_0(x) + \sum_{i=1}^{m} u_i g_i(x) \quad x(0) = x_0 \tag{4.1}$$

where

$$x \in \mathbb{R}^n, u \in \mathbb{R}^m, g_i(x) : \mathbb{R}^n \to \mathbb{R}^n$$

The above form is called a *control-affine system* and $g_0(x)$ is called the drift vector field. The g_i's associated with u_i's are referred to as control vector fields. The notation u_i is conventionally used to denote manipulated input

variables in control theory. However, in what follows we will use the notation k_i since we will be dealing exclusively with chemical rate constants. For PCR amplification reactions, a general dynamic model (assuming all g_i's are control vector fields coupled to time-varying rate constants) can be written

$$\frac{dx(t)}{dt} = \sum_{i=1}^{10} k_i g_i\left(x\left(t\right)\right) \tag{4.2}$$

$$k \in \mathbb{R}^{10};\ k \geq 0, \quad x, g_i\left(x\right) \in \mathbb{R}^{4n+9},\ x \geq 0, \quad h\left(x\right) \in \mathbb{R}^{5},\ h\left(x\right) = 0$$

$$k = \left[k_m \ \ k_{-m} \ \ k_1^1 \ \ k_2^1 \ \ k_1^2 \ \ k_2^2 \ \ k_e \ \ k_{-e} \ \ \frac{k_{cat}}{K_N} \ \ {k'}_{cat}\right]$$

where n denotes the number of base pairs, $g_1\left(x\right)$ to $g_2\left(x\right)$ represent the melting reaction alone, $g_3\left(x\right)$ to $g_6\left(x\right)$ represent the annealing reactions alone and $g_7\left(x\right)$ to $g_{10}\left(x\right)$ represent the extension reactions and $h\left(x\right)$ denotes a set of 5 independent nonlinear constraints enforcing chemical mass balance, which cause the system to evolve on a state manifold $X \subset \mathbb{R}^{4n+9}$ that is of dimension $4n+4$. Hence the dimension of the state manifold is also sequence-dependent. Additional constraints, including equality constraints, may also be applied to the vector of rate constants k, due e.g. to their dependence on a single manipulated input variable, temperature. As shown in Marimuthu *et al* (Marimuthu, K., Jing, C., and Chakrabarti, R. Submitted to Biophysical Journal) except for the extension reaction rate constants, all other rate constants are sequence-dependent. Therefore, the equality and inequality constraints on controls are sequence-dependent. As a result of this the $k_i g_i$ in Eq. (4.2) are also sequence-

dependent. x, $g(x)$ and $h(x)$ for a simplex PCR with $n = 2$ are presented below.

$$\left.\begin{aligned} g_1(x) &= [x_{13}\ 0\ 0\ 0\ 0\ 0\ x_{13}\ 0\ 0\ 0\ 0\ 0\ -x_{13}\ 0\ 0\ 0\ 0]^T \\ g_2(x) &= [-x_1x_7\ 0\ 0\ 0\ 0\ 0\ -x_1x_7\ \ 0\ 0\ 0\ 0\ x_1x_7\ 0\ 0\ 0\ 0]^T \end{aligned}\right\} \text{melting (Sequence-Dependent (SD))}$$

$$\left.\begin{aligned} g_3(x) &= [-x_1x_{14}\ x_1x_{14}\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ -x_1x_{14}\ 0\ 0\ 0]^T \\ g_4(x) &= [x_2\ -x_2\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ x_2\ 0\ 0\ 0]^T \\ g_5(x) &= [0\ 0\ 0\ 0\ 0\ 0\ -x_7x_{15}\ x_7x_{15}\ 0\ 0\ 0\ 0\ 0\ 0\ -x_7x_{15}\ 0\ 0]^T \\ g_6(x) &= [0\ 0\ 0\ 0\ 0\ 0\ x_8\ -x_8\ 0\ 0\ 0\ 0\ 0\ 0\ x_8\ 0\ 0]^T \end{aligned}\right\} \text{annealing (SD)}$$

$$\left.\begin{aligned} g_7(x) &= x_{17}\,[0\ -x_2\ x_2\ -x_4\ x_4\ 0\ 0\ -x_8\ x_8\ -x_{10}\ x_{10}\ 0\ 0\ 0\ 0\ 0\ -(x_2+x_4+x_8+x_{10})]^T \\ g_8(x) &= [0\ x_3\ -x_3\ x_5\ -x_5\ 0\ 0\ x_9\ -x_9\ x_{11}\ -x_{11}\ 0\ 0\ 0\ 0\ 0\ (x_3+x_5+x_9+x_{11})]^T \end{aligned}\right\} \text{enzyme binding (SD)}$$

$$\left.\begin{aligned} g_9(x) &= x_{16}\,[0\ 0\ -x_3\ 0\ x_3-x_5\ x_5\ 0\ 0\ -x_9\ 0\ x_9-x_{11}\ x_{11}\ 0\ 0\ 0\ (x_3+x_5+x_9+x_{11})\ 0]^T \\ g_{10}(x) &= [0\ 0\ 0\ 0\ 0\ -x_6\ 0\ 0\ 0\ 0\ 0\ -x_{12}\ 0\ 0\ 0\ 0\ x_6+x_{12}]^T \end{aligned}\right\} \text{extension}$$

$$x = \left[S_1, S_1P_1, E.S_1P_1, D_1^1, E.D_1^1, E.D_2^1, S_2, S_2P_2, E.S_2P_2, D_1^2, E.D_1^2, E.D_2^2, DNA, P_1, P_2, N, E\right]$$

Additional equality constraints for the state variables are given below

$$h_1(x) = x_{14}^0 - x_1^0 + x_1 - x_{14} \tag{4.3}$$

$$h_2(x) = x_{15}^0 - x_7^0 + x_7 - x_{15} \tag{4.4}$$

$$h_3(x) = x_1^0 + x_7^0 - \left(\sum_{i=1}^{13} x_i\right) - \left(\frac{x_{16}}{K_N}\right)\left(\sum_{i=3,5,9,11} x_i\right) \tag{4.5}$$

$$h_4(x) = \frac{x_{16}^0 - \left(\sum_{i=4,5,10,11} x_i\right) - 2\left(x_6 + x_{12} + x_{13}\right)}{\left(1 + \frac{1}{K_N}\right)\left(\sum_{i=5,11} x_i\right)} - x_{16} \tag{4.6}$$

$$h_5(x) = x_{17}^0 - \left(1 + \frac{x_{16}}{K_N}\right)\left(\sum_{i=3,5,9,11} x_i\right) - x_6 - x_{12} \tag{4.7}$$

We now introduce several types of control systems that are based on the above general formulation, but differ in terms of additional constraints they apply to the controls k_i and the approximations they make regarding the sequence and temperature dependence of these controls.

4.2.1 Staged Time Invariant (On-Off) DNA Amplification Model (STIM)

Mehra and Hu Mehra and Hu (2005) treated melting, annealing and extensions steps independently and did not consider the dissociation of $E.D_i$ molecules into E and D_i in reactions E_5 to E_8 (Methods section). The state equations of the melting, annealing, and extension reactions were solved independently. Also, the rate constants of the respective reactions were held constant Mehra and Hu (2005). The initial condition for the annealing step has been generated

based on the assumption that the melting reaction is 100% efficient. The initial condition of the SP molecule concentrations used to solve the state equations of the extension step is essentially the same as its value at the final time of the annealing step. We call this kind of PCR model a 'Staged' or 'On-Off' time invariant DNA amplification model, since the control vector fields are assumed to be either turned on or off at each step of a PCR cycle. Table 4.1 presents its model structure. The system is controlled by manipulating the switching times between steps.

4.2.2 Time Invariant DNA Amplification Model (TIM)

Stolovitzky and Cecchi Stolovitzky and Cecchi (1996) combined melting, annealing and extension together and formed a state space system for an overall reaction time (summation of the reaction times of melting, annealing, and extension). The rate constants are, however, held constant; i.e., k is not a function of time. For example, annealing (primer hybridization) rate constants are not changed based on the annealing and extension temperatures. We call this kind of PCR model a time invariant DNA amplification model and its model structure is given in Table 4.1. The definitions of x_i, k_i and g_i are the same as those given in the staged time invariant DNA amplification model. Note that, unlike the STIM, the TIM system is formally not a control system since there are no manipulated input variables.

4.2.3 Time Varying DNA Amplification Model with Drift (TVMD)

As we have discussed above, Mehra and Hu Mehra and Hu (2005) developed a staged time invariant model. For the same model structure, one might vary the rate constants in each step to develop a staged time varying model. The staged time varying PCR model is a special case of a time varying model with drift. In a state space model, if a specific set of control variables are kept constant, then the type of state space system is called a time-varying model with drift. During the melting step, the control variables corresponding to the other two steps can be kept constant and the corresponding approximations can be made in the annealing and extension steps as well. Table 4.1 presents such a model structure. The accuracy of drift approximations can be evaluated by comparison of the relative magnitudes of the reaction rate constants at specified temperatures, which are presented in Fig. 4.1. It can be seen that if drift is set to zero in the above model, it reduces to a staged time varying PCR model. Note that drift vector fields for DNA amplification control systems are sequence-dependent due to sequence dependence of the associated k_i's.

4.2.4 Time Varying DNA Amplification Model (TVM)

The time-varying PCR model allows the melting, annealing, and enzyme binding/extension vector fields to be applied simultaneously and allows time varying manipulation of the corresponding temperature-dependent rate constants

described in Methods. $x(t)$ and $g_i(x(t))$ are the same as in the time invariant model. The only change here is that $k(t)$ is a function of time. Since the rate constants vary with respect to time, this model structure is useful for optimal control calculations that provide the optimal temperature profile by considering the whole state space. Hence, fully time-varying models do not require specification of annealing and extension steps in advance.

Table 4.1: Classification of DNA amplification control systems

	STIM	TIM	TVM	TVMD	
Reaction Steps	$f(x,k)$	$f(x,k)$	$f(x,k)$	$g_0(x)$	$f(x,k)$
Melting	$\sum_{i=1}^{2} k_i g_i(x(t))$	$\sum_{i=1}^{10} k_i g_i(x(t))$	$\sum_{i=1}^{10} k_i(t)\, g_i(x(t))$	$\sum_{i=3}^{10} k_i g_i(x(t))$	$\sum_{i=1}^{2} k_i g_i(x(t))$
Annealing	$\sum_{i=3}^{6} k_i g_i(x(t))$	$\sum_{i=1}^{10} k_i g_i(x(t))$	$\sum_{i=1}^{10} k_i(t)\, g_i(x(t))$	$\sum_{i=1,2,7}^{10} k_i g_i(x(t))$	$\sum_{i=3}^{6} k_i g_i(x(t))$
Extension	$\sum_{i=7}^{10} k_i g_i(x(t))$	$\sum_{i=1}^{10} k_i g_i(x(t))$	$\sum_{i=1}^{10} k_i(t)\, g_i(x(t))$	$\sum_{i=1}^{8} k_i g_i(x(t))$	$\sum_{i=7}^{10} k_i g_i(x(t))$

For STIM, TIM and TVM for all the three steps, $g_0(x) = 0$

4.3 Comparison of model-predicted DNA amplification dynamics

Several of the control systems introduced above (STIM, TIM, TVMD) make approximations regarding the magnitudes and temperature variation of the rate constants k_i that enable the corresponding vector fields to be treated either as drift or to be turned off during certain steps of PCR. In temperature-controlled DNA amplification, where $k_i = k_i(T)$, these approximations may not be valid for arbitrary DNA sequences. The variation of melting, annealing, enzyme binding and extension rate constants with respect to time in each step of a standard PCR protocol, depicted in Fig. 4.1, clearly indicates simultaneous annealing, enzyme binding and extension. The effects of 2 ^{0}C changes in the temperatures of each step are depicted to assess the validity of drift approximations. Fig. 4.2 compares the evolution of the DNA concentration in a first PCR cycle based on the above three models. It is clear from Fig. 4.2 that the time invariant model is unrealistic as it completes the annealing and extension within 20 s during the annealing step. The predictions of the time-varying model are consistent with the general Real-Time PCR temperature cycling prescription. The staged time invariant model does not account for the E.Di dissociation and enzyme binding during the annealing step. As a result of this, it can be seen in Fig. 4.2 that the whole extension reaction is completed within 5 seconds and this contradicts the general Real-Time PCR experimental conditions.

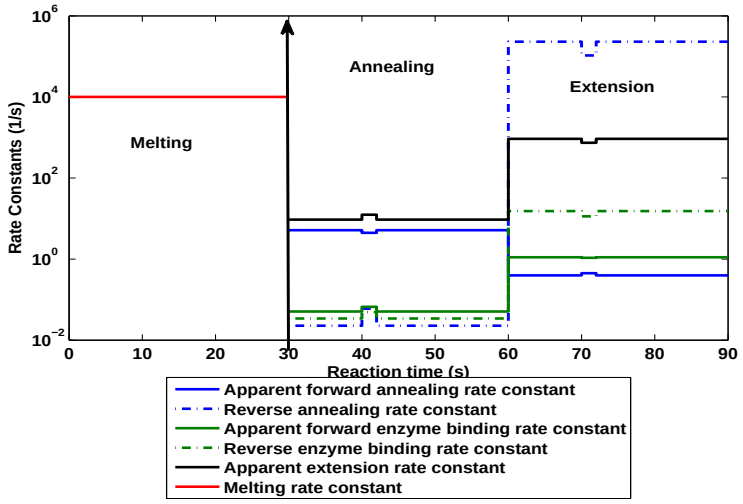


Figure 4.1: Temperature Variation of the DNA Amplification Rate constants. Annealing rate constants have been obtained for a primer with 15 base pairs. The second order forward annealing rate constant, forward enzyme binding rate constant and extension rate constants have been multiplied with Primer concentration (1 μM), Enzyme concentration (10 nM) and Nucleotide Concentration (800 μM). Annealing and extension temperatures are assumed to be 35 and 72 ^{0}C, respectively. The step changes in rate constants depict the effects of change in 2 ^{0}C in temperature. During the melting step all other rate constants have been assumed to be zero and the melting rate constant is assumed to be 10^4.

Therefore, we conclude that the time-varying PCR model is required for dynamic optimization of reaction conditions that can exploit the simultaneous reactions that have been demonstrated to play an important role in PCR dynamics (Marimuthu, K., Jing, C., and Chakrabarti, R. Submitted to Biophysical Journal). Proper modeling of temperature-controlled DNA amplification requires equality constraints to be applied to the components of the vector of rate constants k, based on biophysical modeling of the temperature dependence of the rate constants. In the absence of an accurate sequence- and temperature-dependent kinetic model for DNA amplification that provides these equality constraints, prior work was not able to assess the validity of the approximations made in the STIM, TIM and TVMD models. In the Discussion section, we impose on the components of k in the TVM model several sequence-specific equality constraints, which follow from the temperature dependence of the respective rate constants and are required for optimal control of DNA amplification through manipulation of the temperature.

4.4 Suboptimally Controlled Geometric Amplification of DNA

Standard PCR cycling strategies are based on approximations (like the STIM in Table 4.1) to the true dynamics of PCR. In this section, we show that these cycling strategies are suboptimal given the actual sequence- and temperature-

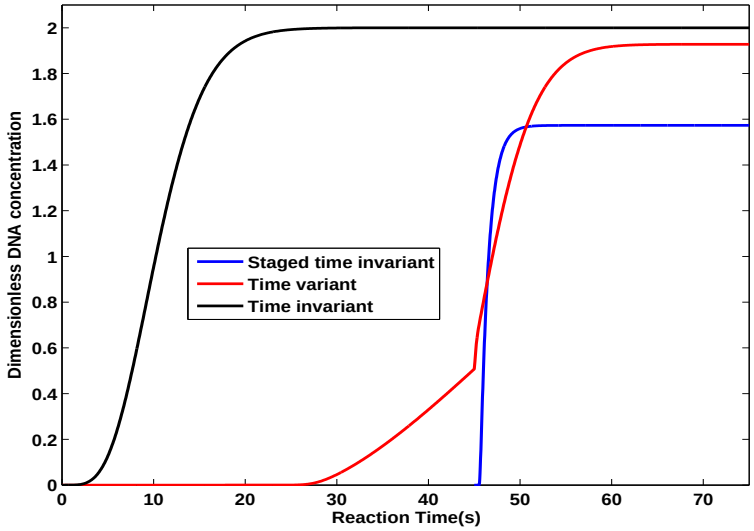


Figure 4.2: Comparison between staged time-invariant (on-off), time-invariant and time-varying models. Annealing time and temperature are fixed to be 45 seconds and 30 ^{0}C. Extension time and temperature are fixed to be 30 seconds and 72 ^{0}C. Since the melting step is neglected in this simulation, the reaction time is sum of the annealing and extension times only. The time-varying model with drift is not shown since the choice of drift vector field is not unique.

dependent dynamics. We simulate the geometric growth of DNA concentration, using the sequence- and temperature-dependent TVM model for DNA amplifi cation, for various choices of manipulated inputs, establishing lower bounds on the margin of improvement that can be achieved through use of optimal cycling strategies. We consider the amplification of a gene for which the primers and kinetic parameters are given in the Materials and Methods section. For every PCR simulation, the number of cycles and length of the target DNA are fixed to be 30 and 500 bp respectively. For fixed annealing and extension reaction times, the annealing temperature is varied from 30 ^{0}C to 50 ^{0}C with an increment of 5 ^{0}C. This simulation is carried out at two different annealing times, 45 seconds (low reaction time) and 120 seconds (high reaction time). The concentration profile of DNA in a typical PCR and the effects of

annealing time and temperature on the final DNA concentration are studied in order to demonstrate the suboptimality of conventional PCR cycling strategies and the need for optimal control of DNA amplification. Figs 4.4 and 4.6 depict the temperature cycling protocols considered. In (Marimuthu, K., Jing, C., and Chakrabarti, R. Submitted to Biophysical Journal), the importance of simultaneous annealing and extension (vector fields g_3 to g_6 and g_7 to g_{10}, respectively) was demonstrated through transient dynamics simulations of single PCR cycles. Generally, lower annealing temperatures increase the primer hybridization efficiency, but higher annealing temperatures increase the polymerase binding and extension rates. The initial species concentrations and reaction time determine which effect dominates. Here, we examine the net effect of these vector fields on the geometric growth rate of DNA concentration using the aforementioned choices of manipulated inputs, illustrating the role played by the initial species concentrations at the start of a cycle (which vary with the stage of PCR) in determining the growth rate. DNA concentrations at the end of each PCR cycle at annealing temperatures 30 ^{0}C to 50 ^{0}C are presented in Figs. 4.5,4.7. Usually, the reaction conditions for the final cycle of PCR are different from those for the other cycles, due to the use of a prolonged extension step followed by low-temperature cooling to ensure the hybridization of any unreacted ssDNA. Hence, for our kinetic study we consider only 29 cycles.

4.4.1 Geometric growth of DNA at low reaction time

The geometric growth of DNA concentration in Fig. 4.5 is similar to that of a typical real time PCR. In Fig. 4.5, the aim is to maximize the concentration of DNA at a specified time by modifying the temperatures of the PCR reaction steps. When the annealing and extension times are 45 and 30 seconds, respectively, there is a difference in DNA concentration at the end of every cycle for an annealing temperature range of 30 to 50 ^{0}C. At 50 ^{0}C, as shown in Fig. 4.5, the efficiency is much lower than that at 30 to 45 ^{0}C, due to the lower melting temperature of the primer (equilibrium melting curves for the primers are shown in Fig. 4.3). When the enzyme concentration is in large excess compared to the DNA concentration during the initial stage of PCR, the enzyme binding reaction is a pseudo-first order reaction. Therefore, the evolution of the DNA concentration in a particular cycle in this stage does not depend on enzyme concentration. While at 35 and 40 ^{0}C, the DNA concentration profile enters the plateau phase at the 24^{th} cycle, at 30 ^{0}C it enters at the 29^{th} cycle. Due to the comparable enzyme and target DNA concentrations at this stage of PCR, which leads to a second-order enzyme binding reaction after 24 cycles, the lower enzyme binding and extension rates reduce the efficiency at 30 ^{0}C annealing temperature. At 40 ^{0}C more DNA is formed after 29 cycles (35 nM @ 40 ^{0}C vs. 20 nM @ 35 ^{0}C). At 45 ^{0}C, the primer hybridization efficiency is lower; this effect dominates and decreases the geometric growth rate in the initial stage of PCR, but the higher polymerase binding and extension rates

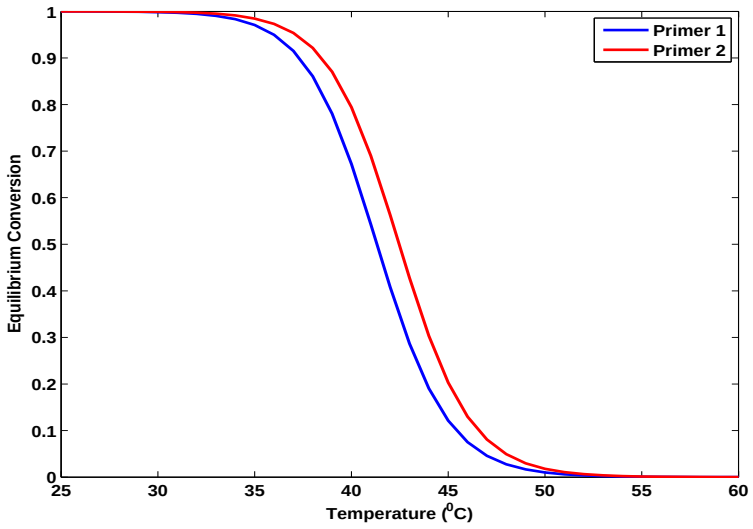


Figure 4.3: Melting curves of the primers

.

dominate in the later cycles of PCR, resulting in a final DNA concentration similar to that at 40 ^{0}C after 29 cycles. Once the DNA concentration is equal to the enzyme concentration in the second stage of PCR, the latter will become the limiting reactant. In order to convert all the ssDNA that are available at the beginning of the every cycle, an equivalent amount of enzyme is needed. The enzyme concentration, however, is fixed at the start of PCR. Given more reaction time, the enzyme could bind to and polymerize excess SP templates after it dissociates from fully-extended dsDNA.

4.4.2 Geometric growth of DNA at high reaction time

The maximum amount of DNA that can be obtained from PCR is equal to the initial concentration of primers in a PCR. When the overall reaction time is low, as shown above, the maximum DNA concentration that is obtained, however, is less than the initial concentration of primers in the final stage of

PCR. This suggests that there can be some improvement in the final DNA concentration if the reaction times are also changed appropriately. Therefore, the annealing time was increased to 120 seconds. As presented in Fig. 4.7 , with this change, the DNA concentration after 25 cycles is higher than that of the previous study. A maximum concentration of 70 nM at 40 ^{0}C is obtained at the end of 25 cycles when annealing time is increased to 120 seconds. As shown in Fig. 4.5., when the reaction time is increased at high annealing temperature (here, 45 ^{0}C) more DNA is produced than at lower temperature.

Once the enzyme has become the limiting reactant, unless the enzyme molecules are released after converting the equivalent number of ssDNA into dsDNA, all the ssDNA cannot be converted into dsDNA. As noted above, since the enzyme binding and extension reactions are faster at higher annealing temperatures, these temperatures produce more DNA than lower annealing temperatures. Note that once the reaction turns to extension mode, even though the enzyme molecules can be released in this step, they cannot bind with additional SP molecules, since these melt at the extension temperature.

Thus, from this study, we conclude that

- The typical temperature cycling protocol followed (based, e.g., on the STIM model in Table 4.1) often does not maximize target DNA concentration.

- Lower annealing temperature and shorter annealing time are preferred

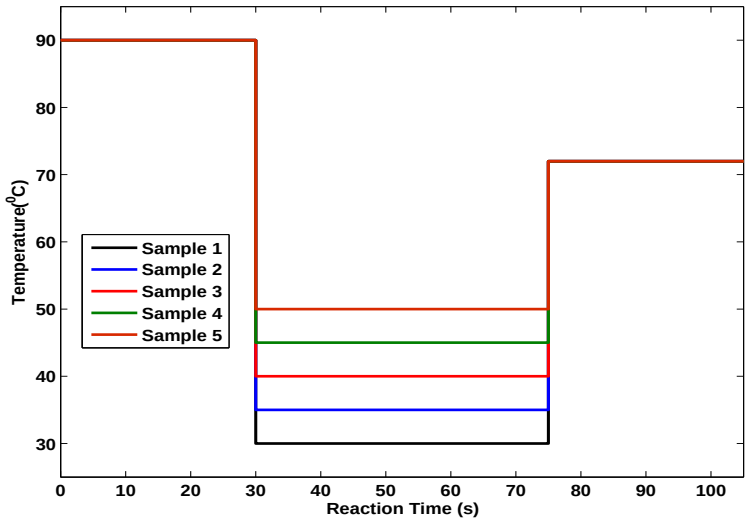


Figure 4.4: Temperature vs Time profile for an annealing time of 45 seconds and extension time of 30 seconds

.

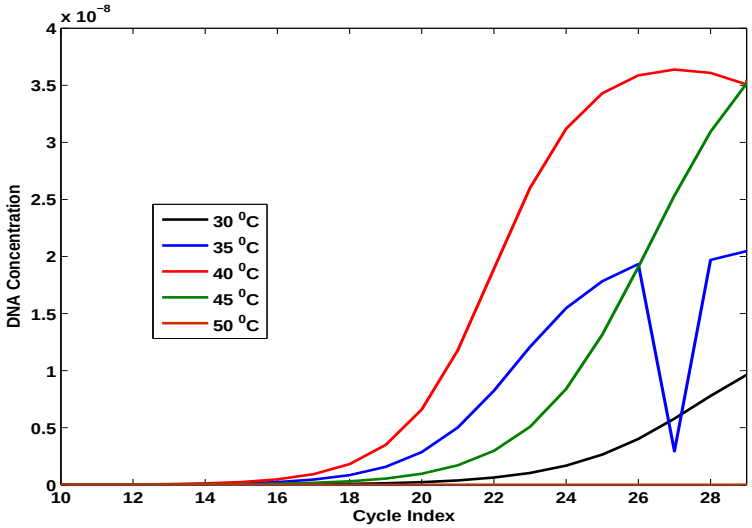


Figure 4.5: Initial concentrations of template, primer, enzyme, and nucleotide are 0.2 μM, 10 nM, 800 μM, respectively and the annealing temperature is varied from 30 ^{0}C to 50 ^{0}C using grid based sampling. Geometric growth of DNA with shorter reaction reaction time (45 s of annealing time). In the geometric growth curves that do not monotonically increase with cycle number, the decreases in DNA concentration are due to denatured single strands that did not react with enzyme to form dsDNA

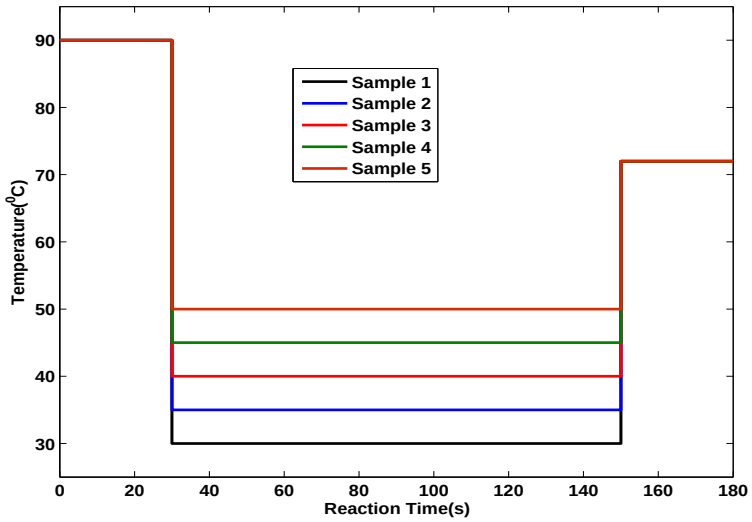


Figure 4.6: Temperature vs Time profile for an annealing time of 120 seconds and extension time of 30 seconds

.

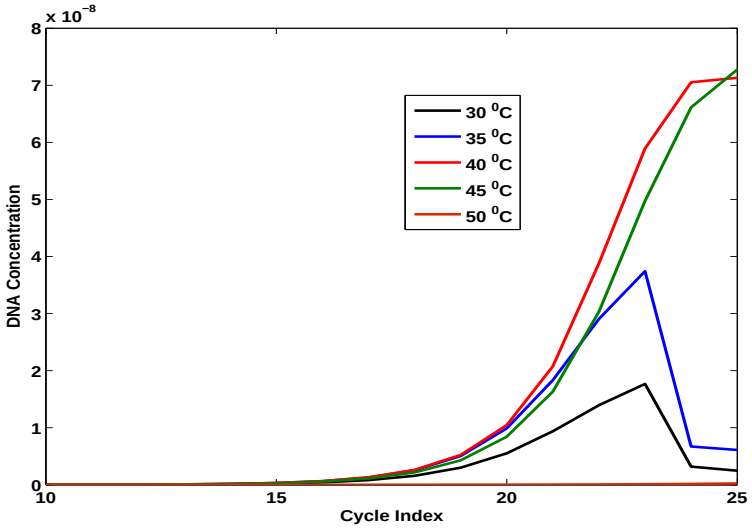


Figure 4.7: Initial concentrations of template, primer, enzyme, and nucleotide are 0.2 μM, 10 nM, 800 μM, respectively and the annealing temperature is varied from 30 ^{0}C to 50 ^{0}C using grid based sampling. Geometric growth of DNA with prolonged reaction reaction time (120 s of annealing time). In the geometric growth curves that do not monotonically increase with cycle number, the decreases in DNA concentration are due to denatured single strands that did not react with enzyme to form dsDNA

for the initial stages of PCR. High annealing temperature and longer annealing time are preferred for the final stages of PCR.

- A systematic procedure should be formulated to obtain optimal annealing time and temperature. Due to geometric growth, there can be a considerable difference between the final DNA concentrations produced by control strategies ($T(t)$) based on misspecified and properly specified DNA amplification models. We show below how to formulate optimal control problems that can provide optimal annealing temperature and times for the objectives of a) maximizing DNA concentration for a specified reaction time; b) minimizing reaction time for a specified target DNA concentration.

4.5 Multistep PCR: development of new cycling strategies using dynamic DNA amplification models

New cycling strategies can be designed based on kinetic models for DNA amplification. For example, we have shown that after a certain number of cycles, annealing reaction time needs to be increased in order to maximize the DNA concentration, since enzyme must rebind to excess SP molecules after dissociating from fully extended dsDNA. Though it is possible to consume all the ssDNA by fixing a long annealing time, this does not exploit the higher rate of

extension at higher temperatures.Therefore, in a single PCR cycle, we propose to conduct annealing and extension multiple times in order to obtain arbitrarily high cyclic efficiency. For example, if the target concentration is 10 nM, the total ssDNA concentration will be 20 nM and therefore, we need 20 nM of enzyme to convert all the ssDNA in to target DNA. Since we have only 10 nM of enzyme in the reaction mixture, we need to conduct 2 annealing and 2 extension steps in this cycle. At the end of the 1st extension reaction, we will have produced nearly 2 nM of target and the enzyme molecules will be free to bind to the unreacted ssDNA. Therefore, after the 1st extension, if we reduce the temperature to the annealing temperature, the required enzyme binding will occur very quickly. Once all the enzyme molecules bind to single strand primer duplex, the temperature of the reaction can be increased to the extension temperature to produce the target DNA. If the initial concentration of the target DNA at the beginning of a cycle is D_0, which is greater than or equal to the enzyme concentration, then the number of annealing and extension steps that needs to be conducted within a given cycle to double the concentration of D_0 is given as

$$N_s = \frac{2D_0}{E_0} \tag{4.8}$$

Based on Eq. (4.8), the number of annealing and extension steps required to double the target concentration is $\mathcal{O}(D_0/E_0)$. By using the appropriate number of annealing and extension steps per cycle, we can reduce the overall PCR reaction time required to achieve a specified DNA concentration. Fig.

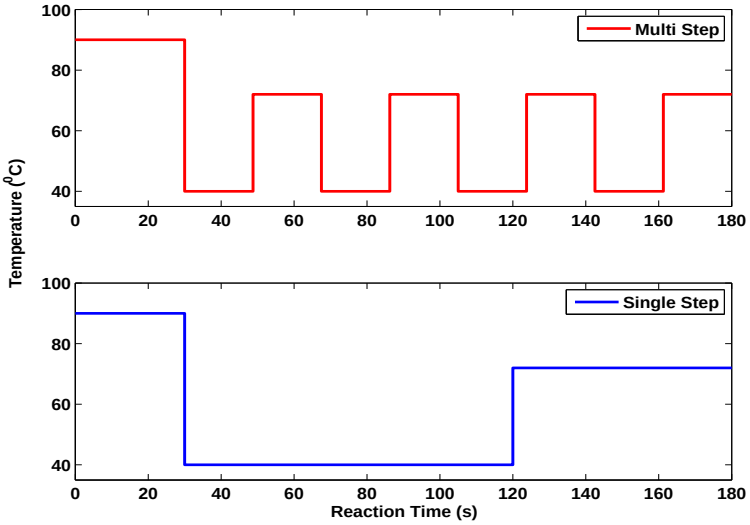


Figure 4.8: Multistep PCR - Temperature profile for multiple annealing step (multi-step) and single annealing step PCR

4.9 shows the evolution of DNA and enzyme molecules in a single annealing step and multistep PCR for a cycle in which the initial concentration of target and enzyme are 10 and 20 nM, respectively and the goal is to achieve 40 nM DNA concentration. Fig. 4.8 compares the temperature profiles for regular and multistep PCR.

The optimal number of annealing and extension steps per cycle cannot be identified without the use of computational kinetic models. Also, in the above simulation, we have arbitrarily fixed the annealing and extension times for the multistep PCR. These can be optimized along with temperatures using control theory in order to minimize the overall reaction time, as described in the next section.

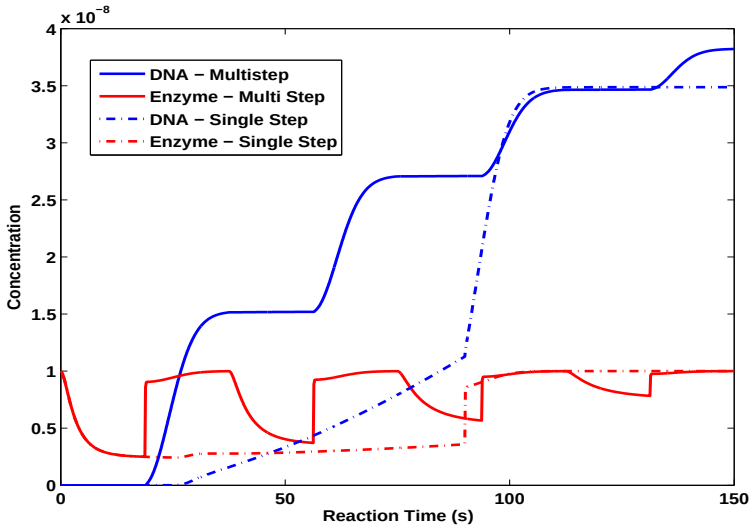


Figure 4.9: Multistep PCR - DNA concentration profile in multi-step PCR. In multi-step PCR, in a PCR cycle, annealing and extension was repeated four times. In a single annealing step PCR, a long annealing time was maintained. Enzyme and template concentrations are 10 nM and 20 nM, respectively, the primer concentration is 200 nM and the nucleotide concentration is 800 μ M.

4.6 Control Strategy

In the previous section, we introduced control systems for DNA amplification and studied the dynamics of geometric growth of DNA for several types of manipulated inputs. We showed that when the annealing time and temperature were varied, due to geometric growth, there was a considerable difference in the final DNA concentration after 25 cycles. Therefore, the determination of optimal time and temperature is essential. In this section we show how to formulate and solve an optimal control problem to obtain not just the optimal annealing time and temperature, but the optimal PCR temperature cycling strategy for a particular amplification objective.

4.6.1 Optimal Control Problem formulation

In the previous section of this article, we have motivated the need for optimal control of DNA amplification. In this section, we formulate optimal control problems for PCR. In optimal control, the optimal evolution of a manipulated variable is sought by minimizing or maximizing a desired objective function. In a PCR, the manipulated variable is the reaction temperature and the desired objective is to maximize the target DNA concentration at the end of every cycle. We classify PCR optimal control problems into two types as follows:

- *Fixed time optimal control problem* - Here we obtain the optimal temperature profile that maximizes the desired DNA concentration profile in a given fixed reaction time. If the cycle time is known in advance, instead of following a grid based optimization approach as shown in Fig. 4.5, it is possible to obtain the optimal temperature profile by solving a fixed time optimal control problem.

- *Minimal time optimal control problem* - Here we obtain the optimal temperature profile that minimizes the overall reaction time for achievement of a specified level of amplification. This allows automated determination of the cycle time and avoids the need for sampling cycle times as was shown in Figs. 4.4 and 4.6.

The desired objective can be expressed in the general form:

$$J = F(x(t_f)) + \int_0^{t_f} L(x(t), u(t))dt \tag{4.9}$$

$F(x(t_f))$ is referred to as an endpoint or Mayer cost and $\int_0^{t_f} L(x(t), u(t))dt$ is referred to as a Lagrange cost.

4.6.2 Fixed-time DNA amplification optimal control problem

In Fig. 4.5, for a fixed reaction time, we varied the reaction temperature to maximize the DNA concentration. This grid-based sampling approach is not an efficient way to obtain the optimal temperature profile. Therefore, we solve an optimal control problem with a desired objective function that maximizes the DNA concentration. Typically a lower limit (T_{min}) for the annealing temperature is applied to reduce the operating temperature range. The upper limit of the operating temperature range could be the maximum melting temperature (T_{max}). Hence lower and upper limits for the PCR reaction temperature can be applied and this can be expressed as an inequality constraint for the optimal control problem. Thus the following optimal control problem can be formulated to maximize the target DNA concentration at the end of

each cycle:

$$\min_{T(t)} J = \sum_{i=1}^{4n+4} x_i\left(t_f\right) \tag{4.10}$$

$$s.t. \quad \frac{dx(t)}{dt} = \sum_{i=1}^{10} k_{i,seq}\left(T\left(t\right)\right) g_i\left(x\left(t\right)\right), \quad x(0) = x_0 \tag{4.11}$$

$$T_{min} \leq T(t) \leq T_{max} \tag{4.12}$$

$$x,\ g_i\left(x\right) \in \mathbb{R}^{4n+9},\ x \geq 0, \tag{4.13}$$

$$h\left(x\right) \in \mathbb{R}^{5},\ h\left(x\right) = 0 \tag{4.14}$$

where $k_{i,seq}(T)$ denote the sequence- and temperature-dependent rate constants described in Material and Methods section. i in the objective function (Eq. (4.10)) indexes all the state variables except for DNA, P_1, P_2, E, and N. Note that due to the equality constraints $h(x) = 0$ on the state vector, $\sum_{i=1}^{4n+4} x_i\left(t_f\right)$ uniquely determines x_{4n+5}, which is the DNA concentration, such that minimization of the former maximizes the latter. Eq. (4.10) specifies a Mayer functional with $F(x(t_f)) = \sum_{i=1}^{4n+4} x_i(t_f)$ and $L(x(t), u(t)) = 0$. The solution of the above optimal control problem provides the optimal trajectory for temperature $(T^*\left(t\right))$ and concentration $(x^*\left(t\right))$ profiles with respect to time. Note that in this formulation, the control (manipulated input) variable is the temperature T, since in PCR reactions the rate constants are varied through temperature cycling. This variable appears in the state equations through the rate constants. Eq. (4.10) assumes a fixed total reaction time for a given cycle. Since the objective function is a Mayer functional, there are multiple solutions

possible for the above optimal control problem.

4.6.2.1 Rate constant as a control variable: Relationship between PCR Rate Constants

The relationship between the temperature and the rate constants is typically given by the Arrhenius equation (Eq. (4.15))

$$k_i = k_{0i} \exp\left(\frac{-E_{ai}}{RT}\right) \tag{4.15}$$

E_a is the activation energy and k_0 is the pre-exponential factor. E_a represents the apparent activation enthalpy of the reaction, whereas k_0 is a function of the apparent activation entropy of the reaction. In general, these thermodynamic parameters may vary with temperature. Temperature-dependent apparent activation enthalpy and entropy can be accommodated through generalizations of Eq. (4.15). However, we have shown that Eq. (4.15) can provide a reasonable approximation for the temperature variation of PCR reaction rate constants. Though in Eq. (4.11) we specified temperature as a control variable, as we see in Eq. (4.15) it appears in the exponential term. This may introduce a strong nonlinearity and hence causes computational difficulty to solve this optimal control problem. In order to avoid this issue, it is possible to use all these 5 rate constants as the control variables and find their optimal evolution. In this

formulation, the optimal control problem can be re-written as

$$\min_{k(t)} J = \sum_{i=1}^{4n+4} x_i\left(t_f\right) \tag{4.16}$$

$$\frac{dx(t)}{dt} = \sum_{i=1}^{10} k_i(t) g_i(x(t)), \quad x(0) = x_0 \tag{4.17}$$

$$x,\ g_i\left(x\right) \in \mathbb{R}^{4n+9},\ x \geq 0 \tag{4.18}$$

$$k \in \mathbb{R}^{10},\ k \geq 0,\ k_{1,seq}^{min} \leq k_1\left(t\right) \leq k_{1,seq}^{max} \tag{4.19}$$

$$k_j = \alpha_{j,seq} k_1^{\beta_{j,seq}}, \quad j = 2, ..., 10 \tag{4.20}$$

$$\alpha_{j,seq} = \left(k_{0j,seq}/k_{01,seq}\right)^{\beta_{j,seq}},\ \beta_{j,seq} = E_{aj,seq}/E_{a1,seq}$$

$$h\left(x\right) \in \mathbb{R}^5,\ h\left(x\right) = 0 \tag{4.21}$$

Where $k_{1,seq}^{min}$ and $k_{1,seq}^{max}$ are the lower and upper bounds for the rate constants $k_{1,seq}$ correspond to T_{min} and T_{max}. The subscript seq indicates the sequence dependency. The control system defined by Eq. (4.17) is a special case of Eq. (4.1). Thus, the optimal control and concentration trajectory is denoted as $\left(k_{seq}^*\left(t\right), x_{seq}^*\left(t\right)\right)$. This trajectory can then be mapped onto the optimal trajectory $(T_{seq}^*(t), x_{seq}^*(t))$ using Eq. (4.15).

4.6.3 DNA amplification in minimal time: Time Optimal Control

In Fig. 4.5 and Fig. 4.7 an arbitrary reaction time has been chosen to obtain the DNA concentration profile. Even though the amplification efficiency

at certain temperatures is nearly 100 %, the cycle reaction time and hence the overall reaction time (sum of reaction time of all the PCR cycles) can be further reduced. To achieve a specified level of DNA amplification in the shortest possible time, the optimal control problem should be formulated in such a way that the solution minimizes the overall reaction time. In practice it is desirable to minimize the overall reaction time and there were several attempts to do this in the past and we summarize them below. Two-step PCR, which uses a combined annealing/extension step at a temperature that lies between the standard annealing and extension temperatures, was an early attempt to decrease total reaction time by reducing cycle time and relying on geometric growth to achieve high overall efficiency. Rapid-cycle PCR (Wittwer and Herrmann (1999); Wittwer and Garling (1991)) is based on fast thermal cycling with short hold times for each step. It also aims to exploit simultaneous annealing and extension. In some incarnations, the temperature is continuously varied. These attempts to reduce PCR reaction time have been empirical studies, with no quantitative basis for the choice of temperatures and times. In particular, use of overly short thermal cycles will result in an increase in overall reaction time. Prior literature provides no mathematical basis for the computation of the optimal cycle switching time. Minimization of the overall reaction time for DNA amplification can be achieved through a minimum time optimal control framework. In particular, the time optimal control framework provides prescriptions for the optimal cycle switching time on a formal mathematical basis. In time optimal control, in addition to state

variables, evolution time also must be optimized. A Lagrange cost of the form $\int_0^{t_f} dt$ corresponding to the evolution time is used as an objective function and state vector is constrained to achieve a specified level of DNA amplification at the end of a cycle as follows:

$$\min_{k(t)} J = \int_0^{t_f} dt \tag{4.22}$$

$$s.t. \frac{dx(t)}{dt} = \sum_{i=1}^{10} k_i(t) g_i(x(t)), \quad x(0) = x_0 \tag{4.23}$$

$$x_{4n+5}(t_f) = x_{4n+5,f}, \quad x_{4n+5,f} \geq 2^m x_{4n+5}(0), \quad m \in \mathbb{Z}^{0+} \tag{4.24}$$

$$x,\ g_i(x) \in \mathbb{R}^{4n+9},\ x \geq 0, \tag{4.25}$$

$$k \in \mathbb{R}^{10},\ k \geq 0,\ k_{1,seq}^{min} \leq k_1(t) \leq k_{1,seq}^{max} \tag{4.26}$$

$$k_j = \alpha_{j,seq} k_1^{\beta_{j,seq}}, \quad j = 2, ..., 10 \tag{4.27}$$

$$\alpha_{j,seq} = \left(k_{0j,seq}/k_{01,seq}\right)^{\beta_{j,seq}},\ \beta_{j,seq} = E_{aj,seq}/E_{a1,seq}$$

$$h(x) \in \mathbb{R}^5,\ h(x) = 0 \tag{4.28}$$

Note that $x_{4n+5} = x_1^0 + x_{2n+3}^0 - \sum_{i=1}^{4n+4} x_i$. Where x_1^0 and x_{2n+3}^0 are the initial concentrations of first and second single strand. m denotes the number of cycles for which the time optimal solution is required. Thus, if $m = 1$, $x_{4n+5,f} \geq 2x_{4n+5}(0)$, which implies that the time is minimized for amplification efficiency greater than 1, which in turn implies that more than one cycle will be generated. The single cycle time optimal solution is obtained by considering $k(t)$ up to the switching time between cycles. In addition to the regular first order conditions for optimality, additional optimality conditions

must be satisfied for the time optimal control problem (Eq. (4.22)) and these are described in Stengel (1994).

4.6.4 A Strategy for Optimal Synthesis of the DNA Amplification Control Trajectory

From the enzyme concentration and the evolution of DNA concentration throughout the amplification reaction, it can be separated into two stages. Stage 1 corresponds to a resource-unlimited environment for sequence replication. Stage 2 begins when environmental resource limitations affect the probability of sequence replication.

4.6.4.1 Stage 1

This stage is comprised of all the cycles at which the enzyme concentration is higher than the target DNA concentration by 2 orders of magnitude so that a pseudo first order kinetics for the enzyme binding reaction can be assumed. Using grid-based sampling of experimental conditions, we have shown that the annealing time for this stage could be around 45 s at a specific annealing temperature. In order to find the optimal time and temperature protocol that improves upon the results in Fig. 4.5 (cycles 1 to 15) in a systematic manner, a fixed time optimal control problem can be formulated (Eq. (4.16)). Alternatively, it is possible to formulate the time optimal control problem (Eq. (4.22)) wherein the objective is the minimization of the reaction time for a specified

target DNA concentration. For both problems, an important feature of the solution $k^*(t)$ for multiple cycles is that they are periodic under the pseudo-first order approximation. This means that problem (Eq. (4.22)) need only be solved once. Here we consider the solution to such a problem where the temperatures as well as switching times for the denaturation and annealing steps are constrained such that $k(t) = f(t)$. Note that application of this constraint results in a control system of the STIM type, wherein the only manipulated input parameter is the switching time between the cycles. However, the vector fields applied in each step are $\sum_{i=1}^{10} k_i(T_1)g_i(x)$, $\sum_{i=1}^{10} k_i(T_2)g_i(x)$ $\sum_{i=1}^{10}, k_i(T_3)g_i(x)$, respectively, where T_1, T_2, T_3 denote the temperatures of the melting, annealing and extension steps, rather than the vector fields listed for the STIM in Table 4.1.

4.6.4.2 Minimal extension time for control of geometric growth

In this section, for a given temperature profile and hence rate constant trajectory $k(t)$ we find the optimal switching time between cycles. For example, in Fig 4.5 for a fixed reaction time of 105 seconds, at three different temperatures we have estimated the target DNA concentration profile. Now we will analyze those DNA concentration profiles and estimate the optimal cycle time that minimizes the overall reaction time in stage 1 of the PCR. Since the optimal control profile is periodic in stage 1, the optimal cycle time for all the cycles in this stage will be the same. Let $0 \leq \eta \leq 1$ be the efficiency in the first cycle

and $n(\eta)$ be the number of cycles in stage 1, then

$$(1+\eta)^{n(\eta)} = y \tag{4.29}$$

Where

$$y = \frac{[DNA]_n}{[DNA]_0}$$

$[DNA]_n$ - Concentration of DNA after n cycles and $[DNA]_0$ - Initial concentration of DNA. From Eq. (4.29), the number of cycles $n(\eta)$ can be expressed as follows:

$$n(\eta) = \frac{log(y)}{log(1+\eta)} \tag{4.30}$$

Let $t(\eta)$ be the time required for a cycle to achieve an efficiency of η and the overall reaction time for stage 1 be $t_{total}(\eta)$; then

$$t_{total}(\eta) = n(\eta)\, t(\eta) = t(\eta)\, \frac{log(y)}{log(1+\eta)} \tag{4.31}$$

Now we consider the following optimization problem to minimize the overall reaction time t_{total}:

$$\min_{\eta}\ t_{total}(\eta) = \min_{\eta}\ t(\eta)\, \frac{log(y)}{log(1+\eta)} \tag{4.32}$$

Hence we seek η such that

$$\frac{d\left(t(\eta)\, \frac{log(y)}{log(1+\eta)}\right)}{d\eta} = 0 \quad \Longrightarrow \quad \frac{1}{t(\eta)}\frac{dt}{d\eta} = \frac{1}{(1+\eta)\,(log(1+\eta))} \tag{4.33}$$

We solve the above equation graphically by plotting the left hand side and the right hand side with respect to η. From the given DNA concentration profile, it is possible to estimate the optimal efficiency ($\eta_{\min}$) and hence, $t\left(\eta_{\min}\right)$ that minimizes the overall reaction time for geometric growth in stage 1, subject to the specified constraints. Note that the computation of the optimal cycle switching time with a given STIM model, as above, does not provide the minimal cycle time that can be achieved with a TVM model. Let $t(\eta, T^*)$, where T^* is time optimal input obtained from solution of problem (Eq. (4.22)), denote the optimal single cycle time for stage 1 as a function of specified cycle efficiency η. Hence $t\left(\eta\right) \geq t\left(\eta, T^*\right)$. Since $t\left(\eta\right) \geq t\left(\eta, T^*\right)$, $\eta_{min} = \arg\,\min t(\eta)$ computed above for the suboptimal $t\left(\eta\right)$ is smaller than η_{min} for $t\left(\eta, T^*\right)$; i.e., the cycle should be run at least as long as computed based on the suboptimal $t\left(\eta\right)$ using the above model, and

$$\min_{\eta}\ t_{total}(\eta) \geq \min_{\eta}\ t\left(\eta, T^*\right) \frac{log\left(y\right)}{log\left(1+\eta\right)} \tag{4.34}$$

4.6.4.2.1 Minimal reaction time example We consider a PCR reaction with primers and reaction conditions that are considered in Fig 4.5 and Fig 4.7. We consider annealing temperatures, 35 and 40 ^{0}C from Fig. 4.5 and 40 ^{0}C from Fig. 4.7. It should be noted that cycle time including melting step in Fig. 4.5 is 105 seconds and in Fig. 4.7 it is 270 seconds. Fig. 4.10 shows the optimal efficiency for these three conditions and from this, in each case the overall reaction time to reach the specific DNA concentration (100 nM)

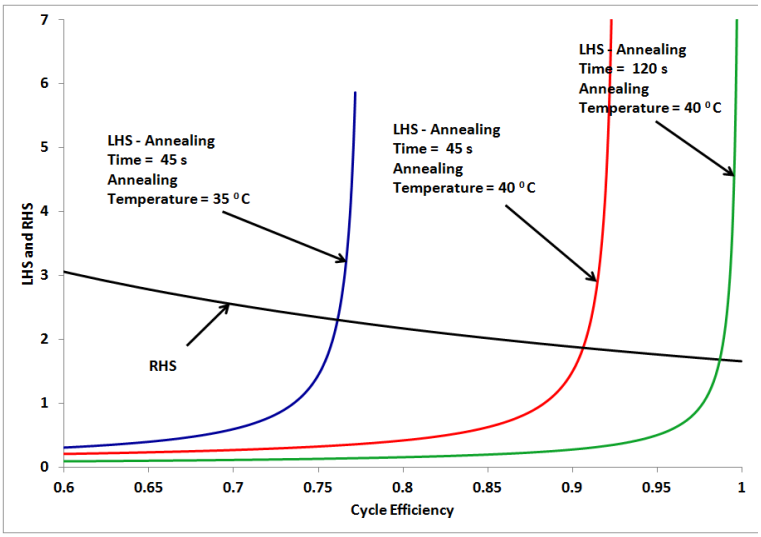


Figure 4.10: Optimal cycle efficiency at three different reaction conditions. Temperature and time in the legend represent the annealing temperature and time. Y axis represents the LHS and RHS of Eq. (4.33). The intersection points specify the optimal switching times between cycles.

has been calculated. This is a constrained optimization version of the time optimal problem (Eq. (4.22)) with the additional constraint that $k(t) = f(t)$. The problem is hence only to find t_f (since $k(t)$ does not change). Table 4.2 compares the overall reaction time under these three conditions and it can be observed that given this sample set, it is possible to reduce the overall reaction time at the specific reaction temperatures. Fig. 4.12 compares the geometric

Table 4.2: Comparison between optimal cyclic efficiency under different reaction conditions

Annealing Temperature (^{0}C)	Annealing reaction time (s)	Optimal Efficiency (%)	Optimal reaction time	Number of of cycles	Overall reaction (s) time (s)
35	45	76.09	102.2	16.2	1664
40	45	90.62	99.5	14.2	1421
40	120	98.68	165.3	13.4	2218

growth of DNA with respect to reaction time under the above three reaction

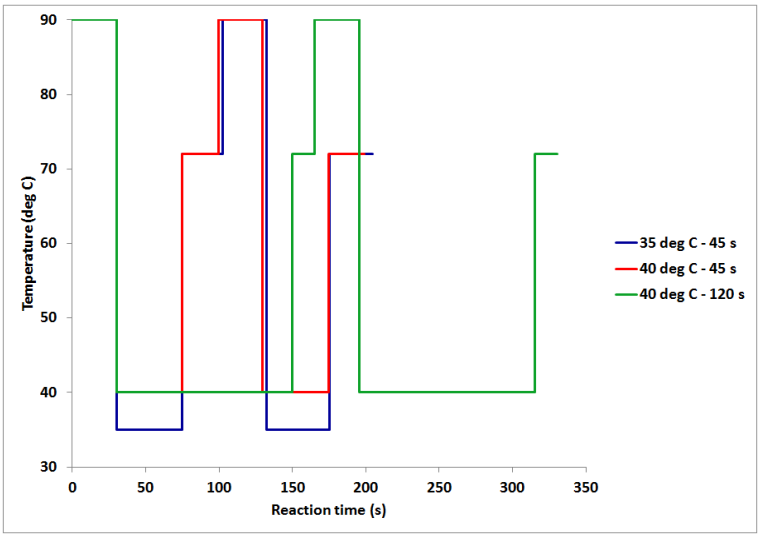


Figure 4.11: Optimized cycling protocol at three different reaction conditions computed based on the cycle switching time criterion depicted in Fig. 4.10. The legends refer to annealing temperatures and times; b)Geometric growth of DNA for three different conditions

conditions. Fig. 4.11 illustrates the modified temperature protocol based on this analysis.

Thus, we have shown that for a specified control vector trajectory $k(t)$ for a cycle, it is possible to reduce the overall reaction time by application of the time optimal cycle switching criterion for geometric growth. In the above case, the optimal choice of annealing temperature is 40 ^{0}C. In the above study, however, we did not consider optimization of the annealing step for the minimization of overall reaction time and hence this method naturally optimizes only the extension reaction time at the end of each cycle. This constraint will be relaxed in future work to solve the time optimal control problem (Eq. (4.22)) using a TVM model.

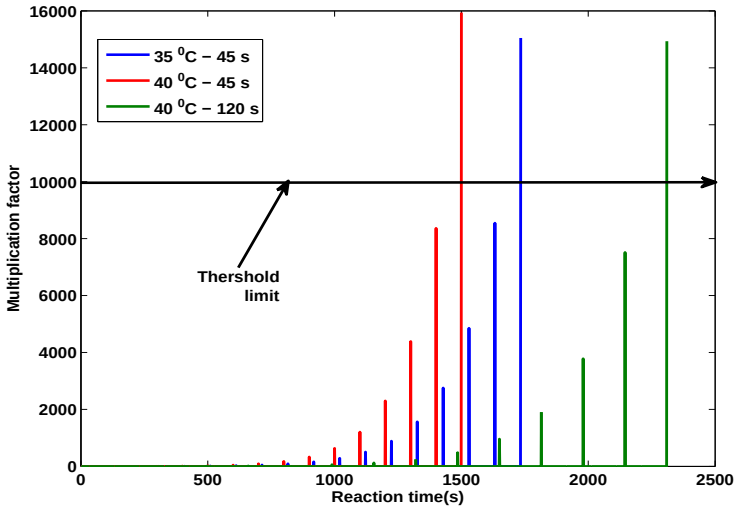


Figure 4.12: Geometric growth of DNA for three different conditions

4.6.4.3 Bilinear time-varying PCR model with and without drift

In stage 1, the primer and nucleotide concentrations are always much higher than the target DNA concentration and the change in their values is negligible compared to the primer and nucleotide concentrations. Therefore, the primer and nucleotide concentration can be treated as constant and hence the annealing and extension reactions are pseudo-first order. The enzyme binding reaction can also assumed to be a pseudo first order reaction, but not for all the cycles in stage 1. In the last few cycles of this stage, enzyme concentration can be comparable with the target DNA concentration; therefore, except for the last two cycles, it can be assumed that even enzyme binding reaction also is a pseudo first order reaction. Furthermore, it is also possible to assume that the melting reaction is irreversible, as hybridization of ssDNA is negligible due to large excess of primers. Thus, the whole PCR model can be expressed as a linear time variant first order state space system. Since both control and

state variables vary with respect to time, to be precise, the PCR model can be expressed as a bilinear control system as follows

$$\frac{dx\left(t\right)}{dt}=\left(A+\sum_{i=1}^{9}B_{i}k_{i}\left(t\right)\right)x\left(t\right) \tag{4.35}$$

$$k\in\mathbb{R}^{9};\ k\geq 0,\quad x,g_{i}\left(x\right)\in\mathbb{R}^{4n+5},\ x\geq 0,\quad \Phi x=0$$

where

$$\Phi=[S_{10}/S_{1}-1,-1,-c,-1,-c,-1,S_{20}/S_{2}-1,-1,-c,-1,-c,-1,-1]$$

$$c=\left(1+\frac{[N_{0}]}{K_{N}}\right)$$

Since the association of ssDNA is neglected, the corresponding rate constants have been eliminated and therefore, the total number of rate constants is 9 and the total number of states is $4n+5$. If the whole system is time variant then $A=0$ and B_{i} is a $4n+5\times 4n+5$ matrix and

$$B_{i}x\left(t\right)=g_{i}\left(x\left(t\right)\right) \tag{4.36}$$

$$x_{1}=[S_{1}],x_{2}=[S_{1}P_{1}],x_{3}=[E.S_{1}P_{1}],x_{2i+2}=\left[D_{i}^{1}\right]$$

$$x_{2i+3}=\left[E.D_{i}^{1}\right]\ \ \forall i=1,2,...n-1$$

$$x_{2n+3}=[S_{2}],x_{2n+4}=[S_{2}P_{2}],x_{2n+5}=[E.S_{2}P_{2}]$$

$$x_{2i+2n+4}=\left[D_{i}^{1}\right],x_{2i+2n+5}=\left[E.D_{i}^{1}\right]\ \ \forall i=1,2,...n-1$$

$$x_{2n+2} = \left[E.D_n^1\right], x_{4n+4} = \left[E.D_n^2\right], x_{4n+5} = [DNA]$$

$$k = \left[k_m \;\; k_1^1 \;\; k_2^1 \;\; k_1^2 \;\; k_2^2 \;\; k_e \;\; k_{-e} \;\; \frac{k_{cat}}{K_N} \;\; k'_{cat}\right]$$

B_1 represents the melting reaction. B_2 and B_3 represent the annealing of the 1st single strand and primer. B_4 and B_5 represent the annealing of the 2nd single strand and primer. B_6 and B_7 represent the enzyme binding reactions. B_8 and B_9 represent the extension reaction. Eq. (4.35) specifies a TVM model for stage 1.

- If the melting and annealing rate constant alone is varied (TVMD for annealing step) then

$$\frac{dx\left(t\right)}{dt} = \left(A + \sum_{i=1}^{5} B_i k_i\left(t\right)\right) x\left(t\right) \tag{4.37}$$

  Here A is a $4n+5 \times 4n+5$ matrix and

$$Ax\left(t\right) = f\left(x\left(t\right)\right) = \left(\sum_{i=6}^{9} k_i B_i\right) x\left(t\right) \tag{4.38}$$

- If the enzyme binding and extension rate constant alone is varied (TVMD for extension step) then

$$\frac{dx\left(t\right)}{dt} = \left(A + \sum_{i=6}^{9} B_i k_i\left(t\right)\right) x\left(t\right) \tag{4.39}$$

Here A is a $4n+5 \times 4n+5$ matrix and

$$Ax\left(t\right) = f\left(x\left(t\right)\right) = \left(\sum_{i=1}^{5} Ak_i\right) x\left(t\right) \tag{4.40}$$

For temperature control, constraints analogous to those in Eq. (4.15) are imposed on the vector of rate constants. The linear structure of the PCR state equations in Stage 1 is convenient for refinement of kinetic parameter estimates via linear filtering Stengel (1994), which will be discussed in future work Moreover, it enables analytical solution of the state equations for each step of PCR. For one single strand, primer annealing and extension reactions, the bilinear matrices have been given below. Note that for this case the number of bilinear matrices will be 7.

$$B_1 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}$$

$$B_2 = \begin{bmatrix} -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_3 = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_4 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_5 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_6 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$B_7 = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

Here, B_1 represents the melting reaction. B_2 and B_3 represent the annealing of the 1st single strand and primer. B_4 and B_5 represent the enzyme binding reactions. B_6 and B_7 represent the extension reaction.

4.6.5 Stage 2

This stage is comprised of all the cycles at which the enzyme concentration is comparable to or less than the target DNA concentration. We have presented a multi-step approach in Fig. 4.9 for these conditions and shown that the efficiency can be improved. For both the fixed time problem (Eq. (4.16)) and minimal time problem (Eq. (4.22)), a multi-step strategy is generally a property of the optimal solution. In each of cycle of this stage, in order to double to concentration of the template DNA, annealing and extension needs to be conducted multiple times. In minimum time optimal control of geometric growth we have shown that 100% efficiency need not be achieved. In such cases the number of steps in a cycle will be less than or equal to the number of steps that was predicted by Eq. (4.8). In order to determine the optimal number of steps, the minimal time optimal control problem described above needs to be solved, either for a single cycle (if the desired single cycle efficiency is known) or for $m \geq 1$ (for more than 1 cycle). For both the fixed time and minimal time problems, the optimal solutions $k^*(t)$ for stage 2 are not periodic.

4.7 Conclusion

In this chapter , the dynamics of DNA amplification reactions have been formulated from a control theoretic standpoint. An optimal control problem for maximization of a desired target DNA concentration has been specified and a strategy for the optimal synthesis of the temperature cycling protocol has been presented.

Chapter 5

Optimal Control of DNA Amplification

5.1 Introduction

In Chapter 4 we have formulated the PCR optimal control problem and discussed the strategies for solving it. In this chapter, we implement those strategies and solve a simple PCR optimal control problem. The foundational concepts of Optimal Control were developed as early as the 1750s in classical mechanics as the principle of least action, a variational approach that minimizes the action (the time integral of a Lagrange cost) of a moving particle to derive the equations of motion in Hamilton's form. A more general and formal version of the optimal control problem, especially from a control engineering perspective, was developed by Pontryagin, who also derived the conditions for the optimality which now comprise Pontryagin Maximum Principle

(PMP) (Stengel (1994); Bryson and Ho (1975)). Optimal control is used in a diverse applications ranging from satellite control to chemical plant control, from quantum control to biological process control, etc. As a result of this, there are several classes of optimal control problems such as constrained optimal control, time optimal control, etc. The complexity of optimal control problems pose challenging issues in developing efficient numerical schemes. There are a tremendous research efforts directed towards to develop various numerical schemes to solve various optimal control problem. Though the primary objective of our work is to solve a PCR optimal control problem, in this chapter we also briefly review several numerical approaches that are available to solve a large scale dynamic optimization problem so that it will be useful in future development of an efficient software to solve PCR optimal control problems. Development of an efficient numerical scheme for the optimal control of PCR is beyond the scope of this work. The primary goal of this work is the development of a foundation for the derivation of optimal control laws for amplification of DNA sequences through application of the PMP to sequence-dependent models of DNA amplification. When applied to a given DNA sequence, these laws prescribe the optimally controlled dynamics of amplification of that sequence.

5.2 Solution of the Optimal Control problem

In order to solve the optimal control problem that we have formulated in (Eq. (4.10)) Chapter 4 we keep temperature (T) as control variable. Even though the upper and lower limits for the temperature are specified as inequality constraints in Eq. (4.12), the optimum lies in the interior of the constraint set; hence, if the algorithm does not sample manipulated variables that are outside the constraint set, the constraints can be ignored. Hence, we solve an unconstrained optimization without any bounds on the temperatures. Moreover, by appropriately applying the equality constraints $h(x)$ to eliminate variables, we can reduce the number of state variables by 5 and solve the control problem on a reduced dimensional state manifold. The conditions for optimality for the above optimal control problem are given below as per the PMP.

5.2.1 Solution of the Optimal Control Problem

Necessary conditions for optimal solutions to the above optimal control problem are derived using Pontryagin's Maximum principle, and they are given as

$$H\left[x(t), \phi(t), T(t)\right] = \phi^{Tr}(t) \sum_{i=1}^{10} k_{i,seq}(T(t)) g_i(x(t)) \tag{5.1}$$

$$\frac{d\phi}{dt} = -\frac{\partial H}{\partial x(t)} = -\sum_{i=1}^{10} k_{i,seq}(T(t)) \frac{\partial g_i(x(t))^{Tr}}{\partial x(t)} \phi(t) \tag{5.2}$$

$$\phi(t_f) = \left[\frac{\partial F(x(t))}{\partial x(t)}\right]_{t=t_f} = \left[\frac{\partial}{\partial x(t)} \sum_{i=1}^{4n+4} x_i(t)\right]_{t=t_f} \tag{5.3}$$

$$\frac{\partial H}{\partial T(t)}\left[x^*(t), \phi^*(t), T^*(t)\right] = \phi^{*,Tr}(t) \sum_{i=1}^{10} \left.\frac{\partial k_{i,seq}(T(t))}{\partial T(t)}\right|_{T^*(t)} g_i(x^*(t)) = 0 \tag{5.4}$$

$$\forall t \in [0, t_f]$$

$H[x(t), \phi(t), T(t)]$ is the PMP-Hamiltonian function. ϕ represents the co-state (a function of time similar to a Lagrange multiplier and analogous to momentum in Hamiltonian mechanics). Equations 4.11,5.2 and 5.4 represent a system of differential algebraic equations (DAE) with the initial and terminal boundary conditions for state and co-state variables. Eq. (5.4) specifies the first order condition for optimality that must be satisfied by the solution $T^*(t)$. An optimal control $T^*(t)$ must also satisfy the second order (Legendre-Clebsch) conditions for optimality, $\frac{\partial^2 H}{\partial T(t) \partial T(t')} \geq 0$, i.e., that the Hessian of the Hamiltonian must be positive semidefinite at T^*. For Mayer cost functionals like that in Eq. (4.10), the Hessian is generally semidefinite at T^* due to the existence of a level set of solutions. This DAE should be solved simultaneously to obtain the optimal temperature profile T. If the target length is n, the number of state equations is $4n + 4$ and typically n will be on the order of 10^3. Thus, the PCR optimal control problem is a large scale optimization problem which requires an efficient numerical algorithm to be solved.

5.2.2 Classification of Optimal Control Problem Solution Strategies

While we developed the necessary conditions for the optimality in the previous section, in this section several numerical strategies to solve an optimal control problem are briefly discussed. Numerical schemes that were developed to solve optimal control problems are broadly classified as *Optimize then Discretize* and *Discretize then Optimize* (Biegler (2010)). These two methods can be further classified into many types as shown in Figure 5.1. In the first type, the optimal conditions for the optimal control problem is derived using PMP principle using the variational approach and then the resulting differential algebraic system of equations is solved using an appropriate numerical solver. In the second type, the given optimization problem is discretized to convert it into a Nonlinear Linear Programming Problem (NLP) which is then solved using efficient solvers.

One variational approach to solving optimal control problems is the so-called shooting approach. In shooting algorithms one solves for the optimal control $u^*(t)$ in terms of $x(t),\ \phi(t)$ using the algebraic constraint from the PMP optimality condition and then integrates the $x,\ \phi$ ODEs simultaneously in terms of known $x(0)$ and unknown $\phi(0)$ initial conditions. Although $\phi(0)$ is unknown, terminal boundary conditions are available in terms of either $x(t_f) = x_f$ (Lagrange functionals) or $\phi(t_f) = \nabla_x F(x(t_f))$ (Bolza or Mayer functionals). This is known as a system of differential equations with split

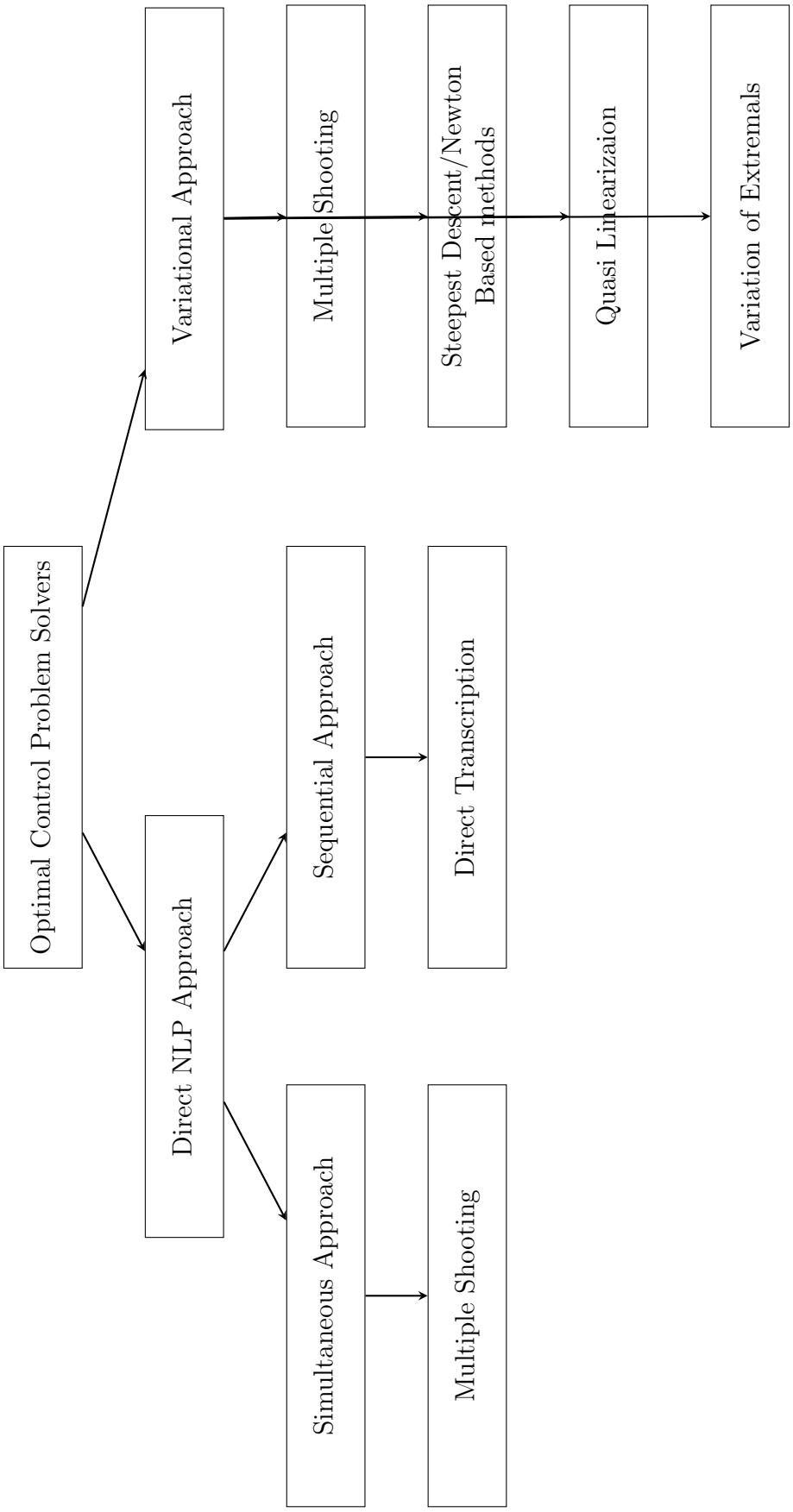


Figure 5.1: Numerical Schemes to solve Optimal Control Problem (This figure is adopted from Biegler (2010))

boundary conditions or a two-point boundary value problem. For nonlinear problems, $\frac{\partial H}{\partial u(t)}$ is an explicit function of $u(t)$ and hence one can solve for $u(t)$ in terms of $x(t)$,$\phi(t)$, even for Mayer functionals. There are multiple solutions to this problem and one can find an approximate solution via line search. Shooting algorithms converge iteratively upon the target $x(t_f) = x_f$or $\phi(t_f)$ vector by making successive changes in the initial conditions $\phi(0)$; i.e., they shoot from x(0),$\phi(0)$, trying to hit the terminal boundary conditions. Numerical algorithms for this problem are typically based on a combination of Newton-Raphson and Runge-Kutta (RK) ODE integration; RK is used to integrate the state/costate ODEs at each step, given $x(0)$ and a guess for $\phi(0)$; NR is used to solve for roots of the boundary condition equations, i.e., $\phi(t_f) - \phi_f = 0$ or $x(t_f) - x_f = 0$. Denoting these by f_i and the initial costate parameters by $\phi_i(0) = c_i$, the NR step is $\delta c = \alpha J^{-1} F(c)$, where the elements of the Jacobian J are $J_{ij} = \frac{\partial f_i}{\partial c_j}$. Shooting algorithms are particularly convenient for Lagrange functionals for which $u(t)$ can be expressed analytically in terms of $x(t), \phi(t)$. Shooting algorithms applied to nonlinear problems wherein $u(t)$ cannot be expressed analytically in terms of $x(t)$, $\phi(t)$ require integration of the state,co-state equations simultaneously with nonlinear root finding to obtain u at each time step. Hence we did not apply them here. In this work we used the variational control vector iteration (CVI) method to solve a fixed time simplex PCR optimal control problem. The software implementation of this algorithm was developed fully by the authors. CVI is commonly used for Bolza or Mayer functionals since $\phi(t_f) = \nabla F(x(t_f))$ is available from solution

of the state equations and an expression for $u = g(x(t), \phi(t))$ is not required.

5.2.3 Solution to Fixed time optimal control problem for PCR

For stage 1, we solve a fixed time optimal control problem (Eq. (4.10)) for a short reaction time (45 seconds of annealing time and 30 seconds of extension time). In stage 1, pseudo first order kinetics for the primer annealing and enzyme binding reaction is applicable. In addition to this, nucleotide concentration can also be assumed to be a constant throughout the cycle. Hence, the algebraic mass balance equations $h(x)$ for these variables need not be solved along with the minimum number of state equations Since the concentrations of primers, enzyme and nucleotides are constants, they can be multiplied with the second order rate constant. Note that the minimum number of state equations is $4n + 4$. Under these conditions, the costate and first order optimality conditions in the PMP become:

$$\frac{d\phi}{dt} = -\frac{\partial H}{\partial x(t)} = \sum_{i=1}^{10} k_{i,seq}(T(t)) B_i^{Tr} \phi(t) \tag{5.5}$$

$$\frac{\partial H}{\partial T(t)} = \sum_{i=1}^{10} \frac{\partial k_{i,seq}(T(t))}{\partial T(t)}\bigg|_{T^*(t)} B_i x^*(t) = 0, \ \forall t \in [0, t_f]. \tag{5.6}$$

We use the methods of optimize and discretize (Variational approach) which gives a set of Differential Algebraic Equations (DAE) that provide a solution for the optimal control problem. Due to the reasons mentioned above, we use

Control Vector Iteration (CVI) (Biegler (2010); Stengel (1994)), which is also called the self consistent iterative method (D'Alessandro (2008)) to solve the above DAE. This iterative process starts with the specification of the initial conditions for the state variables and an assumed manipulated variable profile. The state equations are now integrated for the fixed reaction time from t_0 to t_f. The boundary condition for the co-state variables are obtained as per Eq. (5.2) and the co-state equations are integrated backwards.With the available state and co-state variables, the manipulated variable is updated based on the steepest descent method such that the nominal control profile satisfies Eq. (5.4). The following equation describes the steepest descent method.

$$T_{i+1}(t) = T_i(t) - \alpha \frac{\partial H}{\partial T(t)} \tag{5.7}$$

Where α is a steepest descent parameter which is generally assumed to be a constant in a conventional CVI (Stengel (1994)), In order to speed up the iterative process, α can be estimated through line search via the following minimization.

$$\min_{\alpha} \; F(\alpha(t)) = H\left[x(t), \phi(t), T(t) - \alpha \frac{dH}{dt}(t)\right] \tag{5.8}$$

We used a golden section search algorithm for the above minimization. The following steps and Fig. 5.2 explains the implementation of CVI.

1. For a given initial trajectory of $T_0(t)$ and initial condition x_0, the state

equations are integrated from t_0 to t_f and $x(t)$ is obtained.

2. Boundary condition for the co-state vector is obtained using Eq. (5.4).

3. Co-state equations are integrated backwards from t_f to t_0 and $\phi(t)$

4. Using $x(t)$, $\phi(t)$ and $T_0(t)$, at each t, $\frac{\partial H}{\partial T(t)}$ was found.

5. Using $x(t_f)$, the objective function value J was found.

6. If $J \leq J_{desired}$ and $\|\frac{\partial H}{\partial T(t)}\| \leq \epsilon$, the iterative process was stopped at this point, otherwise the iteration was continued to the next step.

7. Golden section line search method was followed to estimate the optimal $\alpha(t)$.

8. Control profile is updated as per Eq. (5.7)

9. With the updated control profile, step 1 is started again and other steps are followed till the iteration satisfied the convergence criteria, $J \leq 0.05$ and $\|\frac{\partial H}{\partial T(t)}\| \leq 10^{-4}$

5.3 Fixed time optimal control profile

A fixed time PCR optimal control problem with the primer sequences mentioned in Chapter 3.2 and the same reaction conditions (length of target = 500 base pairs) has been solved using the CVI. The reaction time is fixed to be 75 seconds. Figs. 5.3 and 5.4 shows the optimal temperature profile and

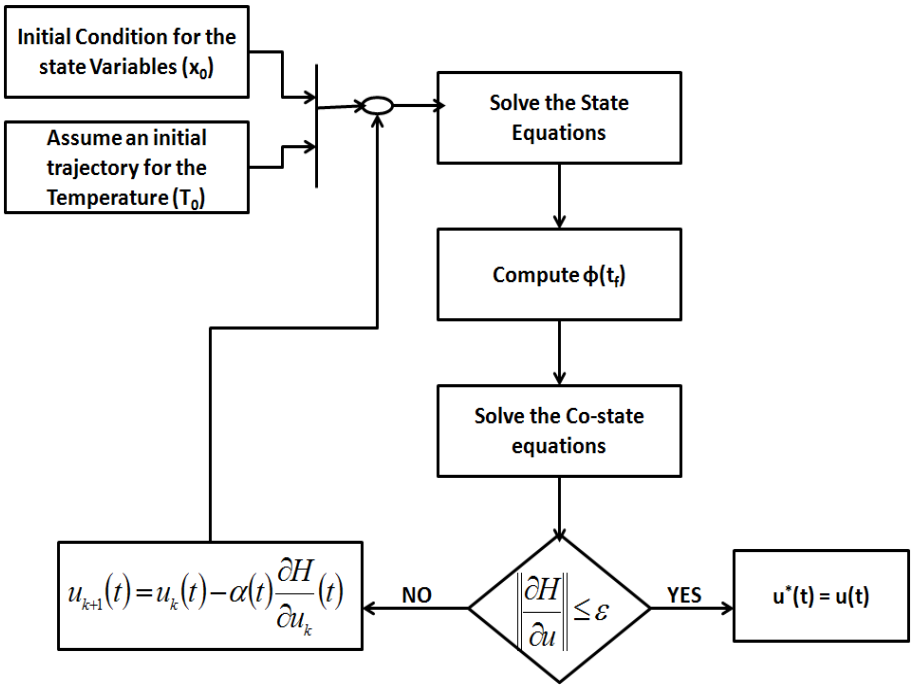


Figure 5.2: Control Vector Iteration to solve optimal control problem

the corresponding optimal DNA concentration profile. The efficiency at the optimal operating condition is 96%. As discussed in Chapter 4 the optimal control solutions $T^*(t)$ for stage 1 are periodic. Hence the $T^*(t)$ from Fig. 5.4 can be applied repeatedly for the first 14 PCR cycles. These are optimal PCR cycling strategies for fixed cyclic time. The same algorithm can be applied to multi-step PCR cycles by appropriate choice of t_f in Eq. (4.10). Since the experimental parameter estimation for the denaturation model in Chapter 3.2 has not yet been completed, we don't solve the multi-step fixed time problem here. Note that since the cycle time is arbitrarily fixed in Fig. 4.8, greater efficiency could be achieved in a specified reaction time. In the next section we consider how to obtain the optimal reaction time t_f.

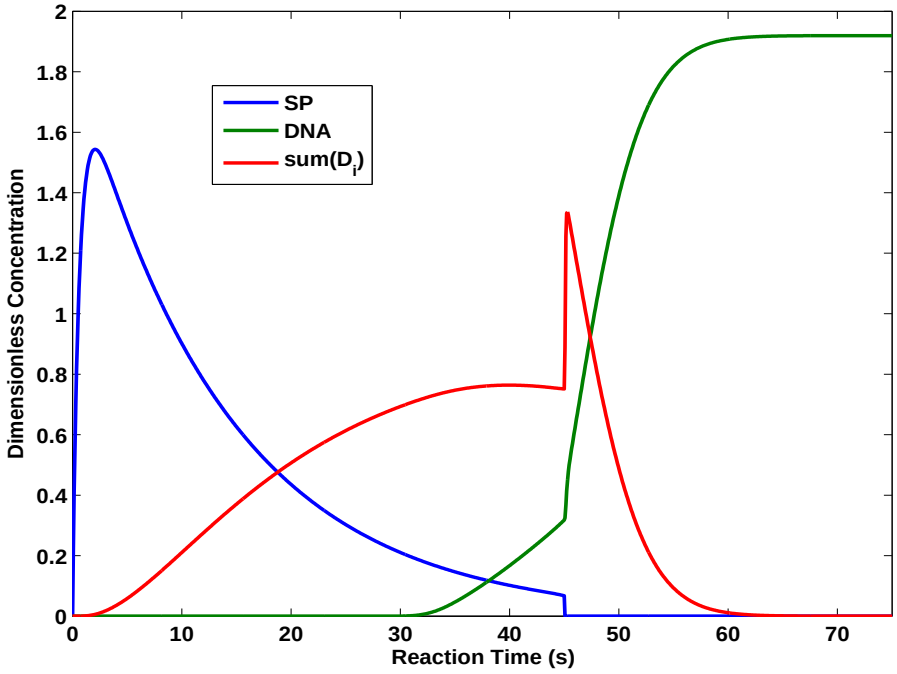


Figure 5.3: Optimal DNA concentration profile that was obtained by solving PCR optimal control problem using control vector iteration.

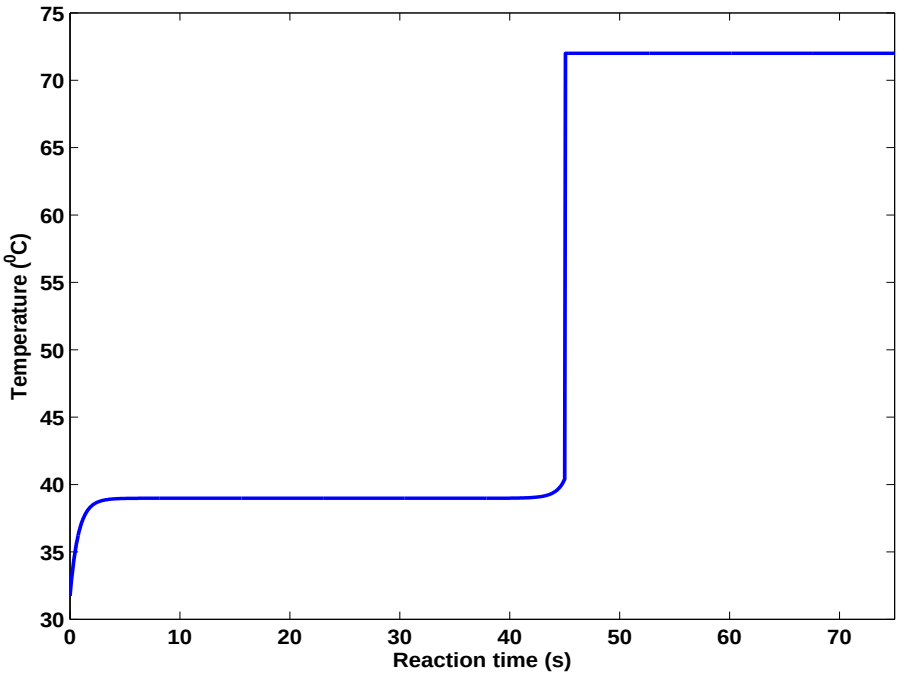


Figure 5.4: Optimal temperature profile that was obtained by solving PCR optimal control problem using control vector iteration.

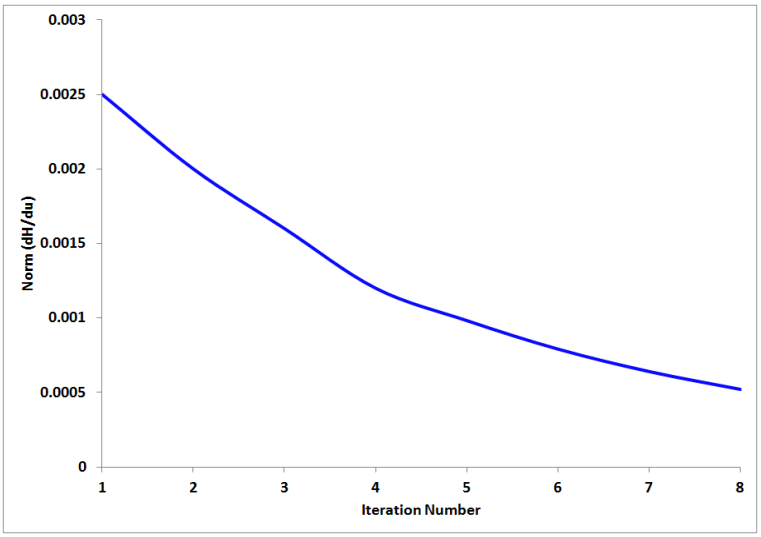


Figure 5.5: Variation of the $\| \frac{dH}{du}(t) \|$ with respect to number of iterations of control vector iteration

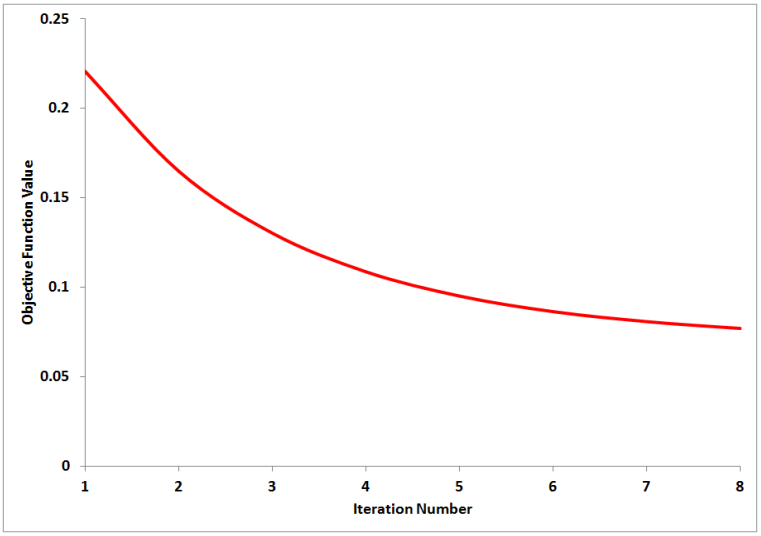


Figure 5.6: Variation of the Objective function value with respect to number of iterations of control vector iteration

5.4 Time optimal control problem

In order to obtain the optimal cycle time for PCR, a multiple cycle time optimal problem of the form (Eq. 4.22) can be solved. In particular, this provides the optimal switching time between cycles. In stage 1, for the reasons discussed in Chapter 4, solution of Eq. (4.10) for $m = 2$ provides the optimal cycling strategy for all cycles in the stage. The PMP-Hamiltonian for time optimal control problem (4.22), expressed using temperature as the control variable, is:

$$\begin{aligned} H\left[x(t), T(t), \phi(t)\right] &= L\left[x(t), T(t)\right] + \phi^{Tr}(t) \sum_{i=1}^{10} k_{i,seq}(T(t)) g_i(x(t)) \\ &= 1 + \phi^{Tr}(t) \sum_{i=1}^{10} k_{i,seq}(T(t)) g_i(x(t)) \end{aligned} \tag{5.9}$$

The first-order conditions for the minimal time problem are:

$$\frac{d\phi}{dt} = -\frac{\partial H}{\partial x(t)} = -\sum_{i=1}^{10} k_{i,seq}(T(t)) \frac{\partial g_i(x(t))^{Tr}}{\partial x(t)} \phi(t) \tag{5.10}$$

$$\frac{\partial H}{\partial T(t)}\left[x^*(t), \phi^*(t), T^*(t)\right] = \phi^{*,Tr}(t) \sum_{i=1}^{10} \left.\frac{\partial k_{i,seq}(T(t))}{\partial T(t)}\right|_{T^*(t)} g_i(x^*(t)) = 0 \tag{5.11}$$

$$\forall t \in [0, t_f]$$

In addition to the $4n + 4$ initial conditions, the following $4n + 4$ terminal boundary conditions are applied to the co-state and state to provide the two-

point boundary value problem:

$$\phi_i(t_f) = 0, \quad i = 1, ..., 4n + 4 \tag{5.12}$$

$$\sum_{i=1}^{4n+4} x_i(t_f) = S_{10} + S_{20} - x_{4n+9,f}, \quad x_{4n+9,f} > 2x_{4n+9}(0). \tag{5.13}$$

We choose $x_{4n+9,f} = 2$ to obtain the time optimal stage 1 cycle protocol; the resulting t_f corresponds to less than 2 cycles unless the time optimal cyclic efficiency is $> \sqrt{2}$. The PMP first-order optimality conditions for time optimal control include an additional so-called algebraic transversality condition for the final time t_f:

$$H\Big|_{t_f} = 0 \implies \phi^{Tr}(t_f) \sum_{i=1}^{10} k_{i,seq}(T(t_f)) g_i(x(t_f)) = -1. \tag{5.14}$$

For stage 1, Eqs. 5.10 and 5.14 are replaced by

$$\frac{d\phi}{dt} = -\sum_{i=1}^{10} k_{i,seq}(T(t)) B_i^T \phi(t)$$

$$\phi^{Tr}(t_f) \sum_{i=1}^{10} k_{i,seq}(T(t_f)) B_i x(t_f) = -1.$$

Various optimization algorithms are available for solving the DAE corresponding to this two-point boundary value problem subject to the algebraic and transversality constraints Eqs. and 5.14. The transversality constraint can be associated with a new Lagrange multiplier. The solution provides both the optimal control function $T^*(t)$ and the minimal time t_f, which is greater than

the period of the function in stage 1. The minimal time cycling strategy for stage 1 is the period of the function. A property of the solution to this problem is that it will satisfy the time optimal cycle switching conditions for geometric growth discussed in Chapter 4. Recall from Chapter 4 that the following bound applies:

$$\min_{\eta} \; t_{total}(\eta) \geq \min_{\eta} \; t\left(\eta, T^{*}\right) \frac{log\left(y\right)}{log\left(1+\eta\right)} \tag{5.15}$$

Since experimental data for estimation of DNA denaturation model parameters are not yet available, we do not solve the multi-step time optimal problem here. In the absence of a denaturation model, assuming a fixed denaturation time, solution of the stage 1 time optimal problem requires tracking of possible cycle efficiencies and computation of $t^{*}\left(\eta\right)$ to identify the optimal cyclic efficiency. Due to the computational complexity of this problem, here we use an approximate approach to the optimal time synthesis for a single cycle for each possible η.

5.4.1 Approximate approach to time optimal control: high processivity assumption

In Section 5.3 we have obtained the optimal control profile for a fixed reaction time. In this section we seek the temperature profile that minimizes the overall reaction time for a target level of amplification. In Chapter 4, the time optimal cycle switching analysis was introduced for a TIM model (equality constrained

controls). Here, we consider the true time optimal cycle switching analysis for the fully time-varying PCR control system. Generation of the full cycle switching curve requires computation of the function $t(\eta)$ for $0 \leq \eta \leq 1$, an interval over which $\eta(t)$ is injective. In an approximate approach to generation of this curve, the efficiency of PCR is defined based on the disappearance of single strand molecules (S_1 and S_2) and single strand primer duplexes (S_1P_1 and S_2P_2). Under the assumption of very high polymerase processivity (where enzyme does not dissociate from partially extended primer-template duplexes), disappearance of SP molecules is the rate limiting step. Indeed, the traditional 3-step PCR cycling protocol can be justified based on the assumption of high polymerase processivity. Since nucleotide addition at the optimal extension temperature is faster than enzyme binding, we define the efficiency of a PCR cycle as

$$\eta = 1 - \frac{S_1P_1 + S_1}{S_{10}} - \frac{S_2P_2 + S_2}{S_{20}} \tag{5.16}$$

We repeated the minimum time optimal control analysis (in Section 4.6.4.2 of Chapter 4.6) with the above defined efficiency to generate the cycle time switching curve and estimate the optimal cycle efficiency - and hence the overall optimal reaction time. Fig. 5.7 shows existence of an optimal cyclic efficiency (92.3%) and the corresponding optimal reaction time is 94.1 seconds. According to this, the total number of PCR cycles needed to multiply the initial concentration of the DNA into 10^4 time is 14 and the corresponding overall reaction time is 1325 seconds. This is 100 seconds less than what has

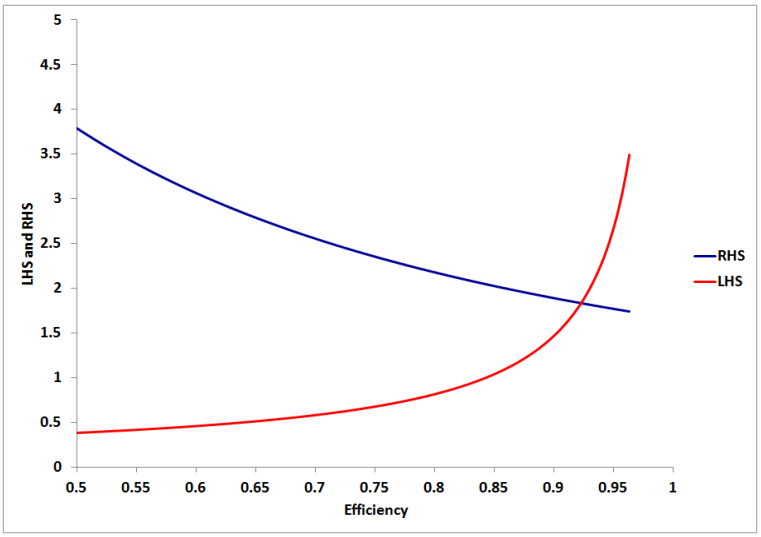


Figure 5.7: Optimal PCR cycle efficiency that minimizes the overall reaction time. Y-axis represents the LHS and RHS of the Eq. (4.33) and the point of intersection is the solution of the Eq. (4.33). $t(\eta)$ used to calculate the LHS and RHS was computed at each η via the approximation method described in the text. The intersection point specifies the optimal switching time between cycles.

been estimated to be the optimal reaction time using the grid-based sampling approach. Note that in this analysis we have fixed the extension time based on our minimum extension time analysis in Section 4.6.4.2 of Chapter 4.6. Fig. 5.8 shows the optimal temperature profile that minimizes the reaction time. Though the performance of the grid-based sampling approach is comparable to that of the above minimal time strategy, the former is computationally very expensive. Furthermore, these kind of optimal temperature profiles can be used as initial guesses for the multi-step PCR fixed time and time optimal control problem. Note that bound (5.15) still applies to the $t(\eta)$ generated by this approach.

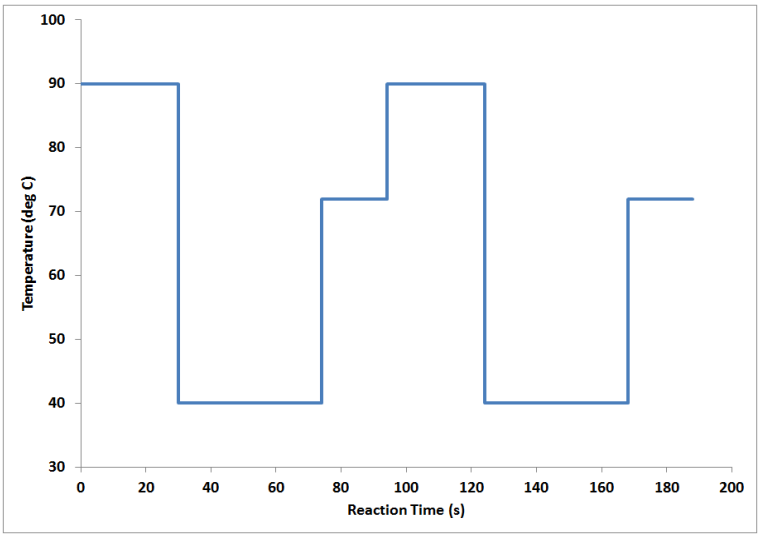


Figure 5.8: Optimal temperature profile that minimizes the overall reaction time by cycling switching at the intersection point in Fig. 2.6.

5.5 Conclusion

In this chapter we have solved a fixed time PCR optimal control problem and obtained the optimal temperature profile. Furthermore, we have proposed an approximate solution for the time optimal control problem and formulated a formal time optimal control problem with it's solution.

Chapter 6

Conclusion and Future Directions

6.1 Notable Contributions

DNA amplification reactions have arguably become the central technology of modern molecular biology. The global PCR market is projected to reach around US$27.4 billion by 2015, with a Compounded Annual Growth Rate (CAGR) of 13.9% for the analysis period, 2009-2015. (Quoted from on-line summary of Polymerase Chain Reaction (PCR) In Medical Application - An Analytical Report, 2009-2015 Published in June 2013). There are many modified versions of traditional PCR available today in the market for various application ranging from diagnosis to genome sequencing. In order to systematically invent new types of PCR, fundamental understanding of the reaction mechanism and kinetics of these reactions is required. In this work, we have de-

veloped the first sequence and temperature dependent kinetic model for PCR and, using the concepts of systems biology, shown that our kinetic model can be used to automatically generate PCR reaction protocols.

Our first contribution in this work was on the development of the sequence-dependent theory of oligonucleotide hybridization kinetics. We considered several kinetic modeling approaches with different reaction mechanisms and compared the predictive power of these approaches using experimental data on hybridization reaction rates. These kinetic models are parameterized not only by the standard nearest neighbor parameters that are used to model DNA hybridization thermodynamics, but also a kinetic parameter associated with the elementary step of the hybridization reaction. We have formulated the problem of estimating this additional parameter, based on our kinetic models and available experimental kinetic data. The methodology presented in the current work establishes a foundation for the automated selection of PCR annealing protocols for any DNA sequence. The methodology developed was illustrated through its application to the modeling of PCR annealing reactions.

Using the theory of enzyme processivity and Michaelis Menten kinetics, we have developed for the first time a sequence and temperature dependent kinetic model for thermostable polymerase enzyme binding and extension reactions. These models have been interfaced with the sequence dependent hybridization kinetics model and hence first sequence and temperature dependent PCR model has been developed. Using our model, we have reported the first theoretical framework for prediction of the dynamic evolution of chemical species

in DNA amplification reactions, for any specified sequence and operating conditions. This establishes a foundation for the dynamic optimization of DNA amplification reactions, which can be used for the automated computation (rather than qualitative selection) of temperature cycling protocols. The model introduced herein is based on quantitative biophysical modeling of DNA melting, annealing, and polymerization that together enable a mapping of a given DNA sequence and polymerase enzyme onto temperature-dependent kinetic rate constants. This framework is based on the notion of sequence-dependent modeling of the kinetics of biochemical reaction networks, which has various applications in dynamical systems biology and is another contribution of the present work.

Based on this biophysical model, it has been shown that DNA amplification efficiency is affected by dynamic processes that are not accurately represented in simplified models of DNA amplification, which are the basis of conventional temperature cycling protocols. Use of this sequence-dependent kinetic model in a control-theoretic framework to determine the optimal temperature cycling protocol of a DNA amplification reaction, for any specified amplification objective, has been discussed. This would enable the automated design of new types of amplification reactions, in addition to enhancement of existing reactions.

We demonstrated by simulation that through the application of sequence dependent kinetic state space models,

- The overall amplification efficiency of the reaction can be improved by orders of magnitude for the same number of cycles
- The overall time of the reaction can be substantially reduced compared to conventional PCR protocols.

We then introduced control systems based on the sequence-dependent kinetic model for DNA amplification. Using these control systems, we have demonstrated that there exists an optimal temperature cycling strategy for geometric amplification of any DNA sequence and formulated optimal control problems that can be used to derive the optimal temperature profile that maximizes the DNA amplification efficiency. Strategies for the optimal synthesis of the DNA amplification control trajectory have been proposed. Finally we have solved a PCR optimal control problem and obtained the optimal temperature profile that maximizes the DNA concentration and minimizes the overall reaction time.

6.2 Future Work

6.2.1 Software Interface

In our current work we have used the control vector iteration technique to solve the PCR optimal control problem. This technique is computationally very expensive. There are very efficient dynamic optimization solvers as reviewed in Fig. 5.1 available. We plan to analyze those methods and choose an

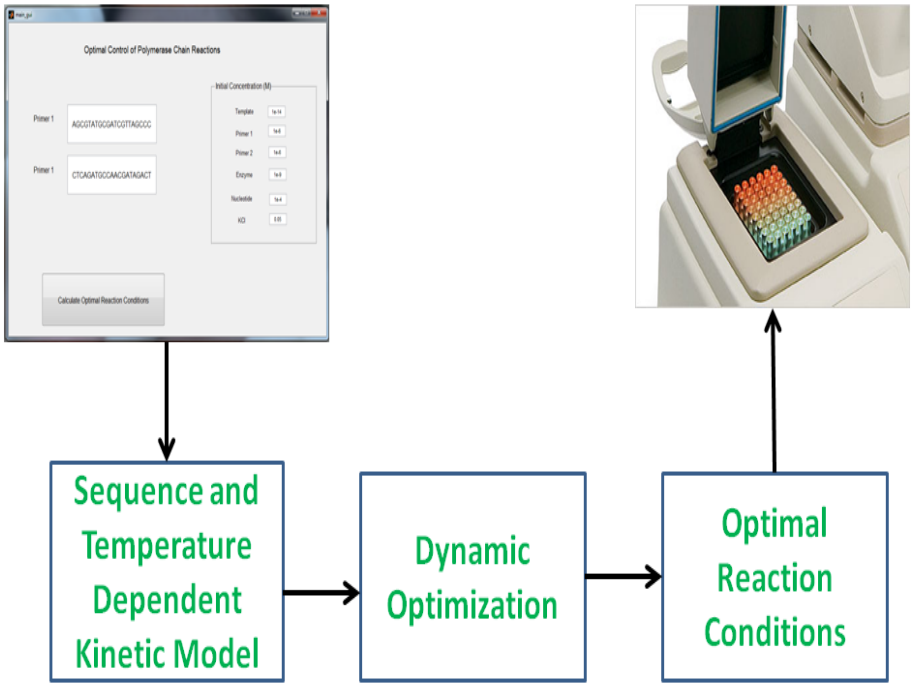


Figure 6.1: On-line prediction of optimal reaction condition for PCR

appropriate dynamic optimization solver that can solve the PCR optimal control problem in a reasonably short time. With this solver we envision creating an on-line software such as those available to predict DNA melting thermodynamic properties, protein sequence alignment, etc, to estimate the optimal reaction conditions for a given template and primer set; Fig.6.1 shows such interface. Since there is no on-line software to predict the optimal PCR reaction conditions, we expect that our work would be potentially useful within the PCR community.

6.2.2 Robust Control Analysis

We plan to create a larger database of heterogeneous DNA sequences and experimentally validate our model parameters to enhance the generality of our

methods. We will also use the methods of robustness analysis to test the sensitivity of our model predictions to parameter uncertainty. The robustness of the optimal control trajectory to the model uncertainty can be obtained for a given model parameter uncertainty distribution using robust control techniques (Nagy and Braatz (2003) and Ma et al. (1999)). In this study, the expectation value of the of the objective function, $E(J)$, and it's variance, σ_J^2 are estimated. Incorporation of uncertainty at the outset of formulation of an OCT problem for PCR can help resolve some of the problems with lack of robustness for problems like minimal time optimal control - since the optimization approach will locate minimum time temperature protocols that are inherently more robust to enzyme binding rate constant uncertainty.

6.2.3 Real-time filtering

In real time PCR machines, which are often used to conduct PCR reactions, it is possible to determine the DNA concentration on-line after every cycle through fluorescence measurements. With these measurements and the state space models which we have presented above, online filtering - a state estimation technique - can be used to estimate other state vector components as well as kinetic model parameters. In this way, the model parameters can be refined based on the real time reaction conditions and the control strategy can be modified on-line in response to state estimates.

Bibliography

Biegler, L. T. *Nonlinear programming: concepts, algorithms, and applications to chemical processes*; SIAM, 2010.

Chakrabarti, R.; Schutt, C. *Gene* **2001**, *274*, 2377–2381.

Chakrabarti, R.; Schutt, C. *Nucleic Acids Research* **2001**, *62*, 383–401.

Skladny, H.; Buchheidt, D.; Baust, C.; Krieg-Schneider, F.; Seifarth, W.; Leib-Mösch, C.; Hehlmann, R. *Journal of Clinical Microbiology* **1999**, *37*, 3865–3871.

Stengel, R. *Optimal Control and Estimation.*; Dover: New York, 1994.

Rychlik, W.; Spencer, W.; Rhoads, R. *Nucleic Acids Research* **1990**, *18*, 6409–6412.

Stolovitzky, G.; Cecchi, G. *Proceedings of National Academy of Sciences, USA* **1996**, *93*, 12947–12952.

Velikanov, M.; Kapral, R. *Journal of Theoretical Biology* **1999**, *201*, 239–249.

Yang, I.; Kim, Y.; Byun, Y., J; Park, S. *Anal. Biochem.* **2005**, *338*, 192–200.

Mehra, S.; Hu, W. *Biotechnology and Bioengineering* **2005**, *91*, 848–860.

Hsu, J.; Das, S.; Mohapatra, S. *Biotechnology and Bioengineering* **1997**, *55*, 359–366.

Gevertz, J.; Dunn, S.; Roth, C. *Biotechnology and Bioengineering* **2006**, *92*, 346–366.

Lee, J.; Lim, H.; Yoo, S.; Zhang, B.; Park, T. *Biochemical Engineering Journal* **2006**, *29*, 109–118.

Griep, M.; Whitney, S.; Nelson, M.; Viljoen, H. *AIChE Journal* **2006**, *52*, 384:392.

Data, K.; LiCata, V. *Nucleic Acids Research* **2003**, *31*, 5590–5597.

Hung, M.; Arnheim, N.; Goodman, M. *Nucleic Acids Research* **1992**, *20*, 4567–4573.

Innis, M. A.; Myambo, K.; Gelfand, D.; Brow, M. *Proceedings of National Academy of Sciences, USA* **1988**, *85*, 9436–9440.

Fish, D. J.; Horne, M. T.; Brewood, G. P.; Goodarzi, J. P.; Alemayehu, S.; Bhandiwad, A.; Searles, R. P.; Benight, A. S. *Nucleic Acids Research* **2007**, *35*, 7197–7208.

Horne, M. T.; Fish, D. J.; Benight, A. S. *Biophysical Journal* **2006**, *91*, 4133–4153.

Hadiwikarta, W. W.; Walter, J.-C.; Hooyberghs, J.; Carlon, E. *Nucleic acids research* **2012**, *40*, e138–e138.

Chalikian, T. V.; Völker, J.; Plum, G. E.; Breslauer, K. J. *Proceedings of the National Academy of Sciences* **1999**, *96*, 7853–7858.

SantaLucia Jr, J.; Hicks, D. *Annu. Rev. Biophys. Biomol. Struct.* **2004**, *33*, 415–440.

SantaLucia, J. *Proceedings of the National Academy of Sciences* **1998**, *95*, 1460–1465.

Breslauer, K. J. *Protocols for oligonucleotide conjugates*; Springer, 1994; pp 347–372.

Breslauer, K. J.; Frank, R.; Blöcker, H.; Marky, L. A. *Proceedings of the National Academy of Sciences* **1986**, *83*, 3746–3750.

Garel, T.; Orland, H. *Biopolymers* **2004**, *75*, 453–467.

Craig, M. E.; Crothers, D. M.; Doty, P. *Journal of molecular biology* **1971**, *62*, 383–401.

Breslauer, K. J.; Bina-Stein, M. *Biophysical chemistry* **1977**, *7*, 211–216.

Pörschke, D.; Eigen, M. *Journal of molecular biology* **1971**, *62*, 361–381.

Wang, J.-Y.; Drlica, K. *Mathematical biosciences* **2003**, *183*, 37–47.

Pfaffl, M. W. *Nucleic acids research* **2001**, *29*, e45–e45.

Gao, Y.; Wolf, L. K.; Georgiadis, R. M. *Nucleic acids research* **2006**, *34*, 3370–3377.

Bloomfield, V. A.; Crothers, D. M.; Tinoco, I. *Nucleic acids: structures, properties, and functions*; University science books, 2000.

Chen, C.; Wang, W.; Wang, Z.; Wei, F.; Zhao, X. S. *Nucleic acids research* **2007**, *35*, 2875–2884.

Bonnet, G.; Krichevsky, O.; Libchaber, A. *Proceedings of the National Academy of Sciences* **1998**, *95*, 8602–8606.

Wetmur, J. G.; Davidson, N. *Journal of molecular biology* **1968**, *31*, 349–370.

Applequist, J.; Damle, V. *Journal of the American Chemical Society* **1965**, *87*, 1450–1458.

Schwarz Jr, M.; Poland, D. *The Journal of Chemical Physics* **1975**, *63*, 557.

Schwarz Jr, M.; Poland, D. *The Journal of Chemical Physics* **1976**, *65*, 2620.

Anshelevich, V.; Vologodskii, A.; Lukashin, A.; Frank-Kamenetskii, M. *Biopolymers* **1984**, *23*, 39–58.

Espenson, J. H. *Chemical kinetics and reaction mechanisms*; McGraw-Hill New York, 1995.

Koehler, R.; Peyret, N. *Bioinformatics* **2005**, *21*, 3333–3339.

Smith, J.; Van ness, H.; Abbott, M. *Introduction to Chemical Engineering Thermodynamic*; McGraw-Hill: New York, 2005.

Jost, D.; Everaers, R. *Biophysical journal* **2009**, *96*, 1056–1067.

Owczarzy, R.; Moreira, B. G.; You, Y.; Behlke, M. A.; Walder, J. A. *Biochemistry* **2008**, *47*, 5336–5353.

Owczarzy, R.; You, Y.; Moreira, B. G.; Manthey, J. A.; Huang, L.; Behlke, M. A.; Walder, J. A. *Biochemistry* **2004**, *43*, 3537–3554.

Rouzina, I.; Bloomfield, V. A. *Biophysical journal* **1999**, *77*, 3242–3251.

Suyama, A.; Wada, A. *Biopolymers* **1984**, *23*, 409–433.

Manyanga, F.; Horne, M. T.; Brewood, G. P.; Fish, D. J.; Dickman, R.; Benight, A. S. *The Journal of Physical Chemistry B* **2009**, *113*, 2556–2563.

Xia, T.; SantaLucia, J.; Burkard, M. E.; Kierzek, R.; Schroeder, S. J.; Jiao, X.; Cox, C.; Turner, D. H. *Biochemistry* **1998**, *37*, 14719–14735.

Baldwin, R. L. In *Molecular Associations in Biology*; Pullman, B., Ed.; Academic Press: New York, 1968.

Gardiner, C. W. *Handbook of Stochastic Methods*; Springer: New York, 2004.

Williams, A. P.; Longfellow, C. E.; Freier, S. M.; Kierzek, R.; Turner, D. H. *Biochemistry* **1989**, *28*, 4283–4291.

Allawi, H. T.; SantaLucia, J. *Biochemistry* **1998**, *37*, 9435–9444.

Peyret, N.; Seneviratne, P. A.; Allawi, H. T.; SantaLucia, J. *Biochemistry* **1999**, *38*, 3468–3477.

Allawi, H. T.; SantaLucia, J. *Biochemistry* **1998**, *37*, 2170–2179.

Bourdélat-Parks, B. N.; Wartell, R. M. *Biochemistry* **2004**, *43*, 9918–9925.

Fish, D. J.; Horne, M. T.; Brewood, G. P.; Goodarzi, J. P.; Alemayehu, S.; Bhandiwad, A.; Searles, R. P.; Benight, A. S. *Nucleic Acids Research* **2007**, *35*, 7197–7208.

Richard, C.; Guttmann, A. J. *Journal of statistical physics* **2004**, *115*, 925–947.

Li, J.; Wang, L.; Mamon, H.; Kulke, M. H.; Berbeco, R.; Makrigiorgos, G. M. *Nature medicine* **2008**, *14*, 579–584.

Blake, R.; Delcourt, S. G. *Nucleic acids research* **1998**, *26*, 3323–3332.

Blake, R.; Bizzaro, J.; J.D., B.; Day, G.; Delcourt, S.; Knowles, J.; Marx, K.; Santalucia, J. *Bioinformatics* **1999**, *15*, 370–375.

Dwight, Z.; Palais, R.; Wittwer, C. T. *Bioinformatics* **2011**, *27*, 1019–1020.

Azbel, M. Y. *Proceedings of the National Academy of Sciences* **1979**, *76*, 101–105.

Patel, S. S.; Wong, I.; Johnson, K. A. *Biochemistry* **1991**, *30*, 511–525.

Boosalis, M.; Petruska, J.; Goodman, M. *The Journal of Biological Chemistry* **1987**, *262*, 14689 –14696.

Mendelman, L. V.; Petruska, J.; Goodman, M. F. *Journal of Biological Chemistry* **1990**, *265*, 2338–2346.

Huang, M.-M.; Arnheim, N.; Goodman, M. F. *Nucleic acids research* **1992**, *20*, 4567–4573.

Brown, J. A.; Suo, Z. *Biochemistry* **2009**, *48*, 7502–7511.

Kuchta, R.; Mizrahi, V.; Benkovic, P.; Johnson, K.; Benkovic, S. *Biochemistry* **1987**, *26*, 8410–8417.

Capson, T. L.; Peliska, J. A.; Kaboord, B. F.; Frey, M. W.; Lively, C.; Dahlberg, M.; Benkovic, S. J. *Biochemistry* **1992**, *31*, 10984–10994.

Fiala, K. A.; Suo, Z. *Biochemistry* **2004**, *43*, 2116–2125.

Wang, Y.; Prosen, D. E.; Mei, L.; Sullivan, J. C.; Finney, M.; Vander Horn, P. B. *Nucleic acids research* **2004**, *32*, 1197–1207.

Davidson, J. F.; Fox, R.; Harris, D. D.; Lyons-Abbott, S.; Loeb, L. A. *Nucleic acids research* **2003**, *31*, 4702–4709.

Invitrogen, *http://probes.invitrogen.com/media/pis/mp22062.pdf* **2006**,

MATLAB, *version 8.0 (R2012a)*; The MathWorks Inc.: Natick, Massachusetts, 2012.

Wittwer, C. T.; Herrmann, M. G. *Rapid thermal cycling and PCR kinetics*; Academic Press, San Diego, CA, USA, 1999.

Wittwer, C. T.; Garling, D. J. *BioTechniques* **1991**, *10*, 76–83.

Bryson, A. E.; Ho, Y.-C. *Applied optimal control: optimization, estimation, and control*; Taylor & Francis, 1975.

D'Alessandro, D. *Introduction to Quantum Control and Dynamics*; Chapman and Hall/CRC: Florida, 2008.

Nagy, Z. K.; Braatz, R. D. *Control Systems Technology, IEEE Transactions on* **2003**, *11*, 694–704.

Ma, D. L.; Chung, S. H.; Braatz, R. D. *AIChE journal* **1999**, *45*, 1469–1476.

Appendix A

Derivation for the relaxation time under steady state assumption

The net forward rate of the i^{th} step

$$r = k_i[\delta_i] - k_{-i}[\delta_{i+1}] \qquad \forall 0 \leq i \leq N-1 \tag{A.1}$$

Note that equation A.1 has been written by applying the equilibrium condition for the formation of the first base pair. Defining the stability constant for the i^{th} base pair formation as

$$s_i = \frac{k_i}{k_{-i}} \tag{A.2}$$

Now equation A.1 can be re-written as

$$\frac{r}{k_i} = [\delta_i] - \frac{1}{s_i}[\delta_{i+1}] \tag{A.3}$$

s_i can be obtained from the free energy formation of the respective base pair. For $i = 0$, equation A.3 can be written as

$$\frac{r}{k_0} = \sigma[\delta_0] - \frac{1}{s_0}[\delta_1] \tag{A.4}$$

For all other i

$$\frac{r}{k_i} = [\delta_i] - \frac{1}{s_i}[\delta_{i+1}] \qquad \forall 1 \leq i \leq N-1 \tag{A.5}$$

At a fixed temperature the forward rate constant (k_i) for the formation of each base pair can be assumed to be a constant Wetmur and Davidson (1968) and it is represented as k_1. Using this assumption and summing equation A.5 through the index i and equation A.4, the equation A.6 is obtained.

$$N\frac{r}{k_f} = \sigma[\delta_0] - \frac{1}{s_0}[\delta_1] + \sum_{i=0}^{N-1}[\delta_i] - \sum_{i=0}^{N-1}\frac{1}{s_i}[\delta_{i+1}] \tag{A.6}$$

As the perturbation parameter δ_i is very small and the stability constant s_i is > 1, the last term in L.H.S of equation A.6 can be approximated as follows

$$\sum_{i=0}^{N-1}\frac{1}{s_i}[\delta_{i+1}] = \frac{1}{s}\sum_{i=0}^{N-1}[\delta_{i+1}]$$

Since stability constant of all possible base pair formation is in the same order of magnitude and the perturbation parameter is a small quantity, above approximation is valid. Hence,

$$N\frac{r}{k_1} = \sigma[\delta_0] - \frac{1}{s_0}[\delta_1] + \sum_{i=0}^{N-1}[\delta_i] - \frac{1}{s}\sum_{i=0}^{N-1}[\delta_{i+1}] \tag{A.7}$$

Further simplifying the equation A.7

$$N\frac{r}{k_f} = \sigma[\delta_0] - \frac{1}{s_0}[\delta_1] + \sum_{i=0}^{N-1}[\delta_i] - \frac{1}{s}\sum_{i=0}^{N-1}[\delta_i] - \frac{1}{s}\delta_N + \frac{1}{s}\delta_1 \tag{A.8}$$

Since

$$\delta_1\left[\frac{1}{s} - \frac{1}{s_0}\right] \approx 0$$

$$N\frac{r}{k_1} = \sigma[\delta_0] + \sum_{i=1}^{N-1}[\delta_i]\left[1 - \frac{1}{s}\right] - \frac{1}{s}\delta_N \tag{A.9}$$

$$A.4 + \frac{1}{s_0} \times A.5(i=1) \implies \frac{r}{k_1} = \left[1 + \frac{1}{s_0}\right]\sigma[\delta_0] - \frac{1}{s_0 s_1}[\delta_2] \tag{A.10}$$

$$A.10 + \frac{1}{s_0 s_1} \times A.5(i=2) \implies$$
$$\frac{r}{k_1} = \left[1 + \frac{1}{s_0} + \frac{1}{s_0 s_1}\right]\sigma[\delta_0] - \frac{1}{s_0 s_1 s_2}[\delta_3] \tag{A.11}$$

Similarly repeating the above steps for all i

$$\frac{r}{k_1}\left[1 + \sum_{i=0}^{N-2}\prod_{j=0}^{i}\frac{1}{s_j}\right] = \sigma[\delta_0] - \delta_N\prod_{i=0}^{n-1}\frac{1}{s_i} \tag{A.12}$$

The mass balance for the perturbation parameter can be written as

$$\delta_0 + \delta_N = -\sum_{i=1}^{N-1} \delta_i \tag{A.13}$$

Substituting the equation A.13 in equation A.9,

$$N\frac{r}{k_1} = \sigma[\delta_0] - ([\delta_0] + [\delta_N])\left[1 - \frac{1}{s}\right] - \frac{1}{s}\delta_N \tag{A.14}$$

$$\implies \frac{r}{k_1} = \frac{\delta_0}{N}\left[\sigma + \frac{1}{s} - 1\right] - \frac{\delta_N}{N} \tag{A.15}$$

Let

$$\left[1 + \sum_{i=0}^{N-2}\prod_{j=0}^{i}\frac{1}{s_j}\right] = C_1$$

Comparing equation A.12 and A.15, an expression for δ_N is obtained as follows

$$\delta_N = \delta_0\left[\frac{\frac{\left[\sigma+\frac{1}{s}-1\right]}{N} - \frac{\sigma}{C_1}}{\frac{1}{N} - \frac{\prod_{i=0}^{N-1}\frac{1}{s_i}}{C_1}}\right] \tag{A.16}$$

Substituting δ_N in equation A.15

$$\frac{r}{k_f} = \frac{\delta_0}{N}\left(\left[\sigma + \frac{1}{s} - 1\right] - \left[\frac{C_1\left(\sigma + \frac{1}{s} - 1\right) - N\sigma}{C_1 - N\prod_{i=0}^{N-1}\frac{1}{s_i}}\right]\right) \tag{A.17}$$

Relaxation time is defined as

$$\tau^{-1} = -\frac{1}{\delta_0}\frac{d\delta_0}{dt} \tag{A.18}$$

Substituting r from equation A.16 for $\frac{d\delta_0}{dt}$ in equation A.17, we obtain the following expression for the relaxation time

$$\tau^{-1} = \frac{k_f}{N}\left(\left[\sigma + \frac{1}{s} - 1\right] - \left[\frac{C_1\left(\sigma + \frac{1}{s} - 1\right) - N\sigma}{C_1 - N\prod_{i=0}^{N-1}\frac{1}{s_i}}\right]\right) \tag{A.19}$$

Appendix B

Nearest Neighbor Parameters for Watson-Crick DNA sequences

Consider a oligonucleotide sequence $GCTAGCTGTAACTG$ for which ΔG needs to be calculated at a temperature, T and a salt concentration $1M$ (Nacl).

$$\Delta G\left(T\right) = \Delta H - T\Delta S \tag{B.1}$$

Assuming that ΔH and ΔS are independent of temperature, above equation can be used to calculate ΔG at a given temperature. ΔH of the above sequence

Table B.1: Nearest-neighbor thermodynamic parameters for DNA Watson-Crick pairs in 1 M NACl

Propagation Sequence	ΔH $(kcal/mol)$	ΔS $(cal/K/mol)$	ΔG_{37^0C} $(kcal/mol)$
AA/TT	-7.6	-21.3	-1.00
AT/TA	-7.2	-20.4	-0.88
TA/AT	-7.2	-21.3	-0.58
CA/GT	-8.5	-22.7	-1.45
GT/CA	-8.4	-22.4	-1.44
CT/GA	-7.8	-21.0	-1.28
GA/CT	-8.2	-22.2	-1.30
CG/GC	-10.6	-27.2	-2.17
GC/CG	-9.8	-24.4	-2.24
GG/CC	-8.0	-19.9	-1.84
Initiation	+0.2	-5.7	1.96
Terminal AT penalty	+2.2	+6.9	+0.05
Symmetry correction	0.0	-1.4	+0.43

is calculated as follows

$$\Delta H\begin{pmatrix} GCTAGCTGTAACTG \\ CGATCGACATTGAC \end{pmatrix} = \Delta H\left(GC_{ini}\right) + \Delta H\begin{pmatrix} GC \\ CG \end{pmatrix}$$
$$+\Delta H\begin{pmatrix} CT \\ GA \end{pmatrix} + \Delta H\begin{pmatrix} TA \\ AT \end{pmatrix} + \Delta H\begin{pmatrix} AA \\ TT \end{pmatrix} + \Delta H\begin{pmatrix} AG \\ TC \end{pmatrix} + \Delta H\begin{pmatrix} GC \\ CG \end{pmatrix}$$
$$+\Delta H\begin{pmatrix} CT \\ GA \end{pmatrix} + \Delta H\begin{pmatrix} TA \\ AT \end{pmatrix} + \Delta H\begin{pmatrix} AG \\ TC \end{pmatrix} + \Delta H\begin{pmatrix} GC \\ CG \end{pmatrix} + \Delta H\begin{pmatrix} CT \\ GA \end{pmatrix}$$

Table B.1 provides the values for the terms in L.H.S of the above equation. Similarly ΔS is also calculated using the entropy values provided in Table B.1. Once ΔH and ΔS for a whole sequence is calculated, Eq B.1 can be used to calculate ΔG.

Appendix C

Nearest Neighbor Parameters for an internal mismatch

Nearest Neighbor Parameters for an internal single base pair mismatch.

MM	ΔH $(kcal/mol)$	ΔS $(cal/K/mol)$	MM	ΔH $(kcal/mol)$	ΔS $(cal/K/mol)$
GC/GG	-6	-15.8	CC/GT	-0.8	-4.5
CT/GT	-5	-15.8	AC/CG	-0.7	-3.8
CG/GG	-4.9	-15.3	AG/TA	-0.7	-2.3
GC/TG	-4.4	-12.3	CA/GG	-0.7	-2.3
GC/GT	-4.1	-11.7	AA/TG	-0.6	-2.3
AG/GC	-4.0	-13.2	GA/CG	-0.6	1.0
AG/TG	-3.1	-9.5	TA/GT	-0.1	-1.7
AC/AG	-2.9	-9.8	AC/TC	0	-4.4

CT/GG	-2.8	-8.0	TA/TT	0.2	-1.5
AT/TT	-2.7	-10.8	AC/GG	0.5	3.2
AT/TG	-2.5	-8.3	AG/CC	0.6	-0.6
GT/CT	-2.2	-8.4	AC/TT	0.7	0.2
CC/GC	-1.5	-7.2	GA/AT	0.7	0.7
CG/TC	-1.5	-6.1	AG/TT	1	0.9
TT/AG	-1.3	-5.3	TT/AC	1	0.7
AT/TC	-1.2	-6.2	AA/TA	1.2	1.7
AG/AC	-9.0	-4.2	TA/CT	1.2	0.7
GA/GT	1.6	3.6	CA/GC	1.9	3.7
AA/TC	2.3	4.6	GC/CT	2.3	5.4
AA/GT	3.0	7.4	GG/CT	3.3	10.4
CA/AT	3.4	8.0	CC/CG	3.6	8.9
GT/TG	4.1	9.5	AA/AT	4.7	12.9
CC/AG	5.2	14.2	CC/TG	5.2	13.5
GA/CC	5.2	14.2	AC/TA	5.3	14.6
GG/TT	5.8	16.3	CA/CT	6.1	16.4
AA/CT	7.6	-20.2			

C.1 Effect of Na^+ concentration on NN parameters

$$\Delta G\left(T\right) = \Delta H - T\Delta S\left(Na^+\right) \tag{C.1}$$

$$\Delta S\left(Na^+\right) = \Delta S\left(1MNa^+\right) + C \tag{C.2}$$

$$C = (4.29 f_{GC} - 3.95) \times 10^{-5} ln[Na^+] + 9.40 \times 10^{-6}\left(ln^2[Na^+]\right) \tag{C.3}$$

f_{GC} is a fraction of GC content of a given oligonucleotide sequence. $[Na^+]$ is a concentration of either $[Na^+]$ or $[K^+]$ salts.

Appendix D

Effect of Mg^{2+} Concentration on NN parameters

$$\Delta G\left(T\right) = \Delta H - T\Delta S\left(Mg^{2+}\right) \tag{D.1}$$

$$\Delta S\left(Mg^{2+}\right) = \Delta S\left(1MNa^{+}\right) + C \tag{D.2}$$

$$C = \left(4.29 f_{GC} - 3.95\right) \times 10^{-5} ln[Mon^{+}] + 9.40 \times 10^{-6} \left(ln[Mon^{+}]\right)^{2} \tag{D.3}$$

$$C = \Delta H^{0} \left[a + bln[Mg^{2+}] + f_{GC}\left(c + dln[Mg^{2+}]\right) + \frac{6 + fln[Mg^{2+}] + g\left(ln[Mg^{2+}]\right)^{2}}{2\left(N_{bp} - 1\right)}\right] \tag{D.4}$$

Table D.1: Parameter for Mg^{2+} Correction factor equations

Parameter	Value (K^{-1})	Standard Error (K^{-1})
a	3.92×10^{-5}	0.2×10^{-5}
b	-9.11×10^{-5}	0.5×10^{-5}
c	6.26×10^{-5}	0.4×10^{-5}
d	1.42×10^{-5}	0.08×10^{-5}
e	-4.82×10^{-5}	0.7×10^{-5}
f	5.25×10^{-5}	0.2×10^{-5}
g	8.31×10^{-5}	0.2×10^{-5}

$$a = 3.92 \times 10^{-5} \left(0.843 - 0.352[Mon^{+}]^{0}.5 \times ln[Mon^{+}]\right) \tag{D.5}$$

$$d = 4.12 \times 10^{-5} \left(1.279 - 4.03 \times 10^{-3} ln[Mon^{+}] - 8.03 \times 10^{-3} \left(ln[Mon^{+}]^{2}\right)\right) \tag{D.6}$$

$$g = 8.31 \times 10^{-5} \left(0.486 - 0.258 ln[Mon^{+}] + 5.25 \times 10^{-3} \left(ln[Mon^{+}]^{3}\right)\right) \tag{D.7}$$

f_{GC} is a fraction of GC content of a given oligonucleotide sequence. $[Mon^{+}]$ is a concentration of mono-valent salt concentration. $[Mg^{2+}]$ is a concentration of Magnesium salt concentration. Table D.1 present the above model parameters. Figure D.1 explains the algorithm that needs to be followed to calculate the correction factor for Mg^{2+} concentration.

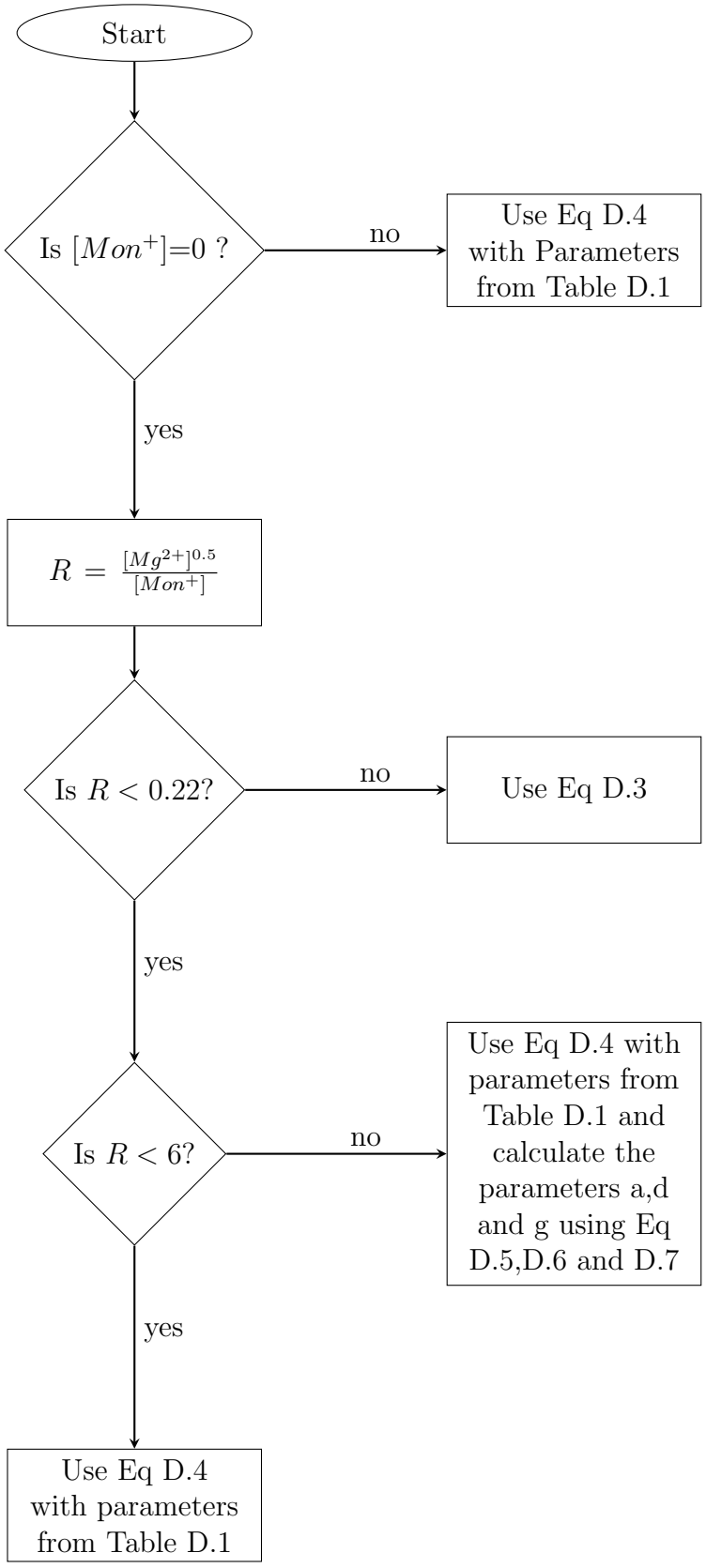


Figure D.1: Calculation of Entropy correction factor for a given $[Mg^{2+}]$

Appendix E

Calculation of the Equilibrium Concentration

E.0.1 Self Complementary Sequences

—

$$2S \underset{k_r}{\overset{k_f}{\rightleftharpoons}} D \tag{R_1}$$

$$K_{annealing} = \frac{[D]}{[S^2]} \tag{E.1}$$

(E.2)

Let x be the conversion of reaction R_1

$$K_{annealing} = \frac{[D_0] + [S_0]x/2}{[S_0]^2(1-x)^2} \tag{E.3}$$

Assuming $[D_0] = 0$.

$$K_{annealing} = \frac{x}{2[S_0](1-x)^2} \tag{E.4}$$

From the above equation, the following equation for the conversion can be written

$$x^2 - x\left(2 + \frac{1}{2[S_0]}K_{annealing}\right) + 1 = 0 \tag{E.5}$$

E.0.2 Non Self Complementary Sequences

—

$$S_1 + S_2 \underset{k_r}{\overset{k_f}{\rightleftharpoons}} D \tag{R_1}$$

$$K_{annealing} = \frac{[D]}{[S_1][S_2]} \tag{E.6}$$

(E.7)

Let x be the conversion of reaction R_1 and assume that initial concentration of S_1 and S_2 is equal

$$K_{annealing} = \frac{[D_0] + [S_{10}]x}{[S_{10}]^2(1-x)^2} \tag{E.8}$$

Assuming $[D_0] = 0$.

$$K_{annealing} = \frac{x}{[S_0](1-x)^2} \tag{E.9}$$

From the above equation, the following equation for the conversion can be written

$$x^2 - x\left(2 + \frac{1}{[S_0]}K_{annealing}\right) + 1 = 0 \tag{E.10}$$

If the concentration of S_1 and S_2 is not equal and S_1 is the limiting reactant

$$K_{annealing} = \frac{[D_0] + [S_{10}]x}{[S_{10}](1-x)\left([S_{20}] - [S_{10}]x\right)} \tag{E.11}$$

Assuming $[D_0] = 0$.

$$K_{annealing} = \frac{x}{[S_{10}](1-x)(M-x)} \tag{E.12}$$

Here $M = [S_{20}]/[S_{10}]$. From the above equation, the following equation for the conversion can be written

$$x^2 - x\left(M + 1 + \frac{1}{[S_{10}]}K_{annealing}\right) + M = 0 \tag{E.13}$$

Appendix F

Calculation of the Relaxation time for Self Complementary Sequences

$$\frac{d[D_{eq}+\delta]}{dt} = k_f[S_{eq}-2\delta]^2 - k_r[D_{eq}+\delta] \tag{F.1}$$

Expanding both sides of the above equation and neglecting the $[\delta^2]$ term,

$$\frac{d[\delta]}{dt} = -\delta(4k_f[S_{eq}]+k_r) \tag{F.2}$$

$$\tau = \frac{1}{(4k_f[S_{eq}]+k_r)} \tag{F.3}$$

Appendix G

Stability Constant Of AU Polymer

Craig et al. (1971) estimated the stability constants of AU base pairs at various temperatures. Figure G.1 presents the Arrhenius relationship of those stability constants and using this relationship it is possible to estimate stability constant at any temperature.

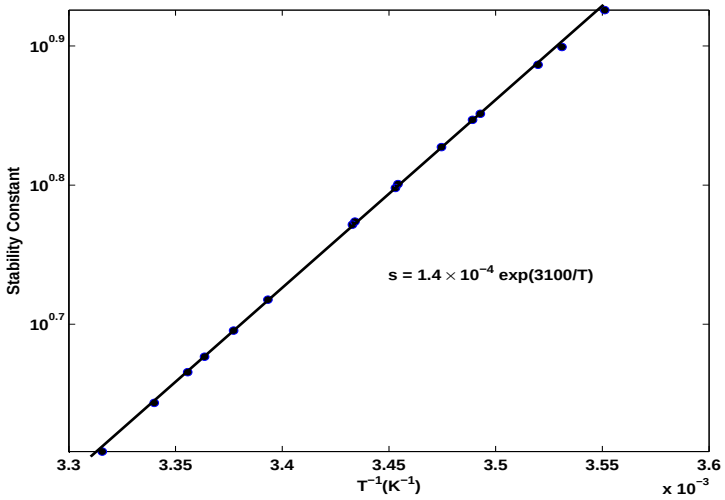


Figure G.1: Stability Constant of AU polymer