Theory of biopolymer templating mechanisms

by

Tam Thi Minh Phan

M.S., Ho Chi Minh City University of Science, Vietnam, 2012

AN ABSTRACT OF A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Abstract

Biopolymer templating is the process in which two or more flexible biopolymers identify specific zipper-like binding state encoded by the sequence of sidechains. The challenge in studying biopolymer templating is that experiments only give macroscopic results from which microscopic processes must be inferred. Here we build a theory of biopolymer templating mechanisms via interactions of the polymers in mis-aligned and aligned states using a random walk model to understand the thermodynamics and kinetics of various biomolecules. We consider two types of biopolymer templating, protein aggregation and DNA hybridization. Protein aggregation is associated with numerous neurodegenerative diseases such as Huntington's and Alzheimer's while DNA hybridization plays important roles in many fields including nanotechnology, biotechnology. Surprisingly, we find that although protein and DNA systems share many similarities, they have different results: the mis-aligned states slow down and hinder the protein aggregation while nonspecific binding helps DNA to perform the alignment search during hybridization, which accelerates the hybridization rate. The findings can contribute to a better understanding of the nature of biopolymer templating in many systems.

Theory of biopolymer templating mechanisms

by

Tam Thi Minh Phan

M.S., Ho Chi Minh City University of Science, Vietnam, 2012

A DISSERTATION

submitted in partial fulfillment of the
requirements for the degree

DOCTOR OF PHILOSOPHY

Department of Physics
College of Arts and Sciences

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2021

Approved by:

Major Professor
Jeremy D. Schmit

Copyright

Abstract

Biopolymer templating is the process in which two or more flexible biopolymers identify specific zipper-like binding state encoded by the sequence of sidechains. The challenge in studying biopolymer templating is that experiments only give macroscopic results from which microscopic processes must be inferred. Here we build a theory of biopolymer templating mechanisms via interactions of the polymers in mis-aligned and aligned states using a random walk model to understand the thermodynamics and kinetics of various biomolecules. We consider two types of biopolymer templating, protein aggregation and DNA hybridization. Protein aggregation is associated with numerous neurodegenerative diseases such as Huntington's and Alzheimer's while DNA hybridization plays important roles in many fields including nanotechnology, biotechnology. Surprisingly, we find that although protein and DNA systems share many similarities, they have different results: the mis-aligned states slow down and hinder the protein aggregation while nonspecific binding helps DNA to perform the alignment search during hybridization, which accelerates the hybridization rate. The findings can contribute to a better understanding of the nature of biopolymer templating in many systems.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First, I would like to send my sincere thanks to my major professor, Dr. Jeremy Schmit. I first met him when I was his TA in the first year of my PhD. I wanted to be one of his students right away because not only he had interesting research, but he was also very professional and fair in work. I did not know he had a great sense of humor, and he was super friendly until I joined the group. I never forget that we always have group meetings full of laughs. He creates an excellent atmosphere for the group in which everybody feels comfortable and happy. He brings positive energy and an optimistic attitude to the group. He is very enthusiastic in teaching and training me every single step on the way to be a PhD. He is always patient to answer all my questions, give me good advice and continued support when I am in need. He is really caring, nice and cheerful to everybody. I truly consider Schmit research group as my second family and Dr. Schmit is like a big brother there. He is the best advisor in my life.

I would like to express my gratitude to professors Yong-Cheng Shi, Christopher Sorensen, Matthew Berg, and John Tomich for serving on my dissertation committee. Thank you so much for spending time on reading and giving me precious advice to improve this dissertation.

I am very grateful to Dr. Christopher Sorensen for all his kind encouragement for all students in the CMS group. I will always remember the moments he slowly came to the students who almost cried and wanted to pass out after their talks in the seminars and nicely said "You did a good job!". He may never notice how meaningful his nice words are, but I do. He is like a father in the CMS group with a really warmed heart.

I wish to give a special thank you to Dr. John Tomich and Susan Whitaker in Tomich's lab in the Department of Biochemistry and Molecular Biophysics for giving me a chance to work on an experimental project. It was difficult for a student in a theoretical group to get familiar with experiments. I started from zero and Susan taught me every single step such as how to hold a pipette properly, how to use many machines in the lab. That is such a

precious experience in my PhD life. Even though the experimental project is not shown in this dissertation, I am grateful that I learned many valuable lessons from the time working in the lab.

My acknowledgment would not be complete without thanking all members of the Schmit research group, my friends, and the Vietnamese community for many beautiful memories in KSU and Manhattan. I would especially like to thank Tra Huynh, my best friend in KSU. We share many things about life, food, music, and shopping. She is always by my side and makes time to help me when I am down. She is a strong girl with a positive attitude and emotional soul. My PhD life would be boring without her.

Last but not least, I would like to thank my parents, my brother's family and a very special thank you goes out to my dear husband. I would not have been able to overcome hardships on my journey without their support and unconditional love.

Dedication

To my parents and my beloved husband.

Chapter 1

Introduction

1.1 Introduction of biotemplating problem

Biomolecular assembly is a major contributor to biophysics because it is associated with most processes of life such as catalysis, signaling, etc. The key to reaching a specific state in biomolecular assembly is a variety of inter-molecular interactions which are patterned to encode the desired state[1]. When studying biomolecular interactions, researchers try to answer many questions such as how binding affinity affects molecule association, at what rate that process occurs, how we can control the process, which factors change the molecule's structure, and what the laws govern their behaviors[2]. Single molecules can interact with each other in various ways resulting in simple to complex configurations. They can form oligomers, micelles, fibrils, crystals, etc. Here we focus on a type of biomolecule interaction that we call "biopolymer templating".

Biopolymer templating is the process in which a flexible biopolymer interacts like a zipper with an existing template by forming and breaking hydrogen bonds. The template can be one or many biomolecules. This process completes when the incoming molecules have the same conformation and alignment as the template[3].

Many methods are used in studying biotemplating problems such as experiments, simulations, and bioinformatic approaches[1,4]. However, there are some challenges. Protein

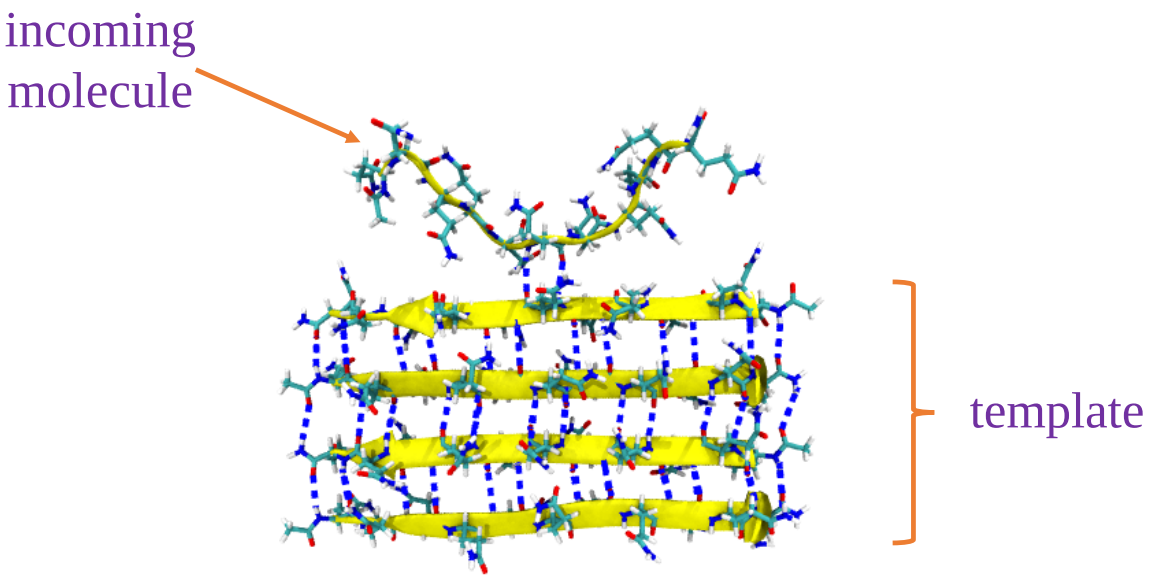


Figure 1.1: *Biotemplating system of polyglutamine. The incoming molecule interacts with the existing template by forming and breaking hydrogen bonds.*

aggregates such as amyloid fibrils take too long a time as ten years to form in numerous neurodegenerative diseases. This is too slow compared to the time scale in vitro aggregation (10^{-2} to 10^{2} s) and the time scale in molecular simulation (10^{-10} to 10^{-3} s)[4]. Moreover, experiments only provide macroscopic observables from which microscopic processes must be inferred. Therefore, developing theoretical models is necessary to create a unified picture between the microscopic and macroscopic world.

In this study, we use the theory of biopolymer templating to investigate the assembly process in two different systems, protein aggregation and DNA hybridization. Native proteins normally fold into secondary structure in a submicrosecond time scale, but we observe that amyloid aggregates form over much longer times scales[3,5]. The explanation for this is that amyloid proteins spend most of the time searching for the correct alignment[3–6]. This search is slow because hydrogen bonds are interchangeable in amyloid aggregates so hydrogen bonds easily form in incorrect alignments. The incoming molecule needs time to break those hydrogen bonds and start searching again before the perfect alignment can be reached[3,5]. DNA hybridization also follows a similar mechanism. The DNA molecules adopt many in-

termediate states before finding some in-register bonds at which point they rapidly zip into the fully bound state[7,8]. Therefore, the templating process involves both mis-aligned and aligned states[5,6]. We model the kinetics of the search process as random walks in a Markov State model, which we can solve either analytically or computationally[5,6].

Here we study two types of biopolymer templating, protein amyloid aggregation and DNA hybridization. Protein amyloid aggregation is associated with many neurodegenerative diseases such as Huntington's, Alzheimer's, and prion[9–11] while DNA hybridization plays important roles in many fields including nanotechnology and biotechnology[7,12]. For more details, I introduce two specific problems of biopolymer templating.

a) Thermodynamics of Huntingtin aggregation

b) Nonspecific binding assists DNA to perform the alignment search during hybridization

1.2 Organization of dissertation

This dissertation is divided into five chapters:

Chapter 1 introduces the templating problem in biomolecular research. It is necessary to build theoretical models because there are some challenges existing in the templating problem using experiments and simulations. This problem is applied on two different objects, protein aggregation and DNA hybridization.

Chapter 2 covers concepts which are used in the two problems. We start with the properties of protein and DNA molecules: what they are, their elementary building blocks, how they fold and misfold, how they interact with each other, etc. Then we introduce on templating systems of protein aggregation and DNA hybridization. We also introduce some models used in the theory.

In chapter 3 we dicuss the thermodynamics of Huntingtin aggregation. Numerous diseases are caused by the aggregation of proteins into amyloids. The similarities between aggregates formed by widely varying proteins raises a question as to the extent that sequence details are important for driving assembly into pathological states. An interesting test case is huntingtin, the aggregating protein in Huntington's disease, which has a remarkably low

complexity sequence featuring a polyglutamine core. This chapter models huntingtin as a triblock copolymer and shows that the aggregation behavior follows directly from generic polymer properties with only minor perturbations from the sequence.

In chapter 4 we apply our templating models to DNA hybridization. DNA hybridization is an important process in biology and nanotechnology. In this chapter, we build a kinetic theory of interactions between DNA molecules in the DNA hybridization process. Although protein and DNA systems share many similarities, our models suggest there are important differences in the kinetic processes.

Chapter 5 provides a summary and a comparison in the results of two problems in chapters 3 and 4, the new knowledge revealed by my theories and the limitations to our models. Possible future work on this topic is recommended as well.

References

[1] C. J. Wilson, A. S. Bommarius, J. A. Champion, Y. O. Chernoff, D. G. Lynn, A. K. Paravastu, C. Liang, M.-C. Hsieh, and J. M. Heemstra, Chemical reviews **118**, 11519 (2018).

[2] D. Goldfarb, *Biophysics demystified* (McGraw-Hill, 2011).

[3] J. D. Schmit, The Journal of chemical physics **138**, 05B611_1 (2013).

[4] R. Nassar, G. L. Dignon, R. M. Razban, and K. A. Dill, Journal of Molecular Biology p. 167126 (2021).

[5] Z. Jia, J. D. Schmit, and J. Chen, Proceedings of the National Academy of Sciences **117**, 10322 (2020).

[6] Z. Jia, A. Beugelsdijk, J. Chen, and J. D. Schmit, The Journal of Physical Chemistry B **121**, 1576 (2017).

[7] T. E. Ouldridge, P. Šulc, F. Romano, J. P. Doye, and A. A. Louis, Nucleic acids research **41**, 8886 (2013).

[8] J. G. Wetmur and N. Davidson, Journal of molecular biology **31**, 349 (1968).

[9] C. A. Ross and M. A. Poirier, Nature medicine **10**, S10 (2004).

[10] A. Aguzzi and T. O'connor, Nature reviews Drug discovery **9**, 237 (2010).

[11] T. T. Phan and J. D. Schmit, Biophysical Journal **118**, 2989 (2020).

[12] Y. Yin and X. S. Zhao, Accounts of chemical research **44**, 1172 (2011).

Chapter 2

Background

Cells are the basic unit of all living things. Human beings, animals, and plants are composed of millions of cells, while other species such as bacteria and yeast are only made up of one cell. The nucleus of each cell contains many chromosomes which each of them contains a very long DNA molecule. A gene is a segment of DNA that holds instructions for a cell to produce a protein[1,2].

In the below sections I will explore the properties of protein and DNA, which are the two main subjects of this dissertation.

2.1 Protein

The world "protein" was derived from the Greek "proteios" which means "holding first place"[3]. Proteins play an essential role in life. For example, proteins help build and maintain tissues, muscles, and organs, they coordinate metabolism, they provide energy for our body, and are important for both physical and mental health[3,4]. A protein is a polymer which is composed of amino acids[5,6]. Each amino acid consists of a carbon atom, an amino group (NH_2), a carboxyl group ($COOH$), a hydrogen atom (H) and a side chain group (R). At neutral pH, the amino acid group receives a proton (H^+) to form (NH_3^+). In contrast, the carboxylic gives up a proton (H^+) and becomes a carboxyl group (COO^-). Therefore,

amino acids are hybrid ions called Zwitterions[7,8]. The side chain is the part which makes each amino acid unique and determines its characteristics.

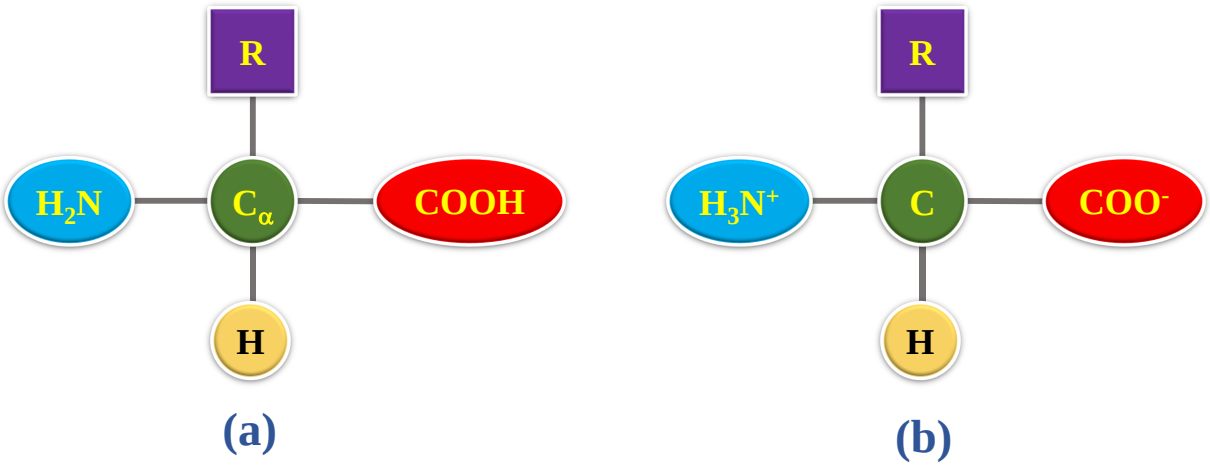


Figure 2.1: *(a) General structure of an amino acid. (b) Dipolar amino acid at neutral pH.*

There are 20 major amino acids. Nine of them are essential amino acids including lysine, valine, phenylalanine, methionine, threonine, isoleucine, histidine, leucine, and tryptophan which cannot be synthesized by human metabolism and must be extracted by consuming food. The other amino acids are non-essential which can be produced by the human body[6]. We can divide 20 types of amino acid into three different classes considering their solubility in water: non-polar, charged polar, and uncharged polar[9,10].

Non-polar amino acids (or hydrophobic amino acids) tend to avoid the watery environment of the cell[7]. They are alanine (Ala – A), glycine (Gly – G), valine (Val – V), leucine (Leu – L), isoleucine (Ile – I), proline (Pro – P), phenylalanine (Phe – F), methionine (Met – M), tryptophan (Trp – W), cysteine (Cys – C). Charged polar amino acids include aspartic acid (Asp – D), glutamic acid (Glu – E) as negatively charged polar (or acidic), and arginine (Arg – R), lysine (Lys – K), histidine (His – H) as positively charged polar (or basic). The last type is uncharged polar amino acids (or hydrophilic) including asparagine (Asn – N), glutamine (Gln – Q), serine (Ser – S), threonine (Thr – T), tyrosine (Tyr – Y)[1,2,7].

From these 20 amino acids, organisms can build many thousands and thousands of different proteins. Amino acids are connected with each other to form a protein polymer which

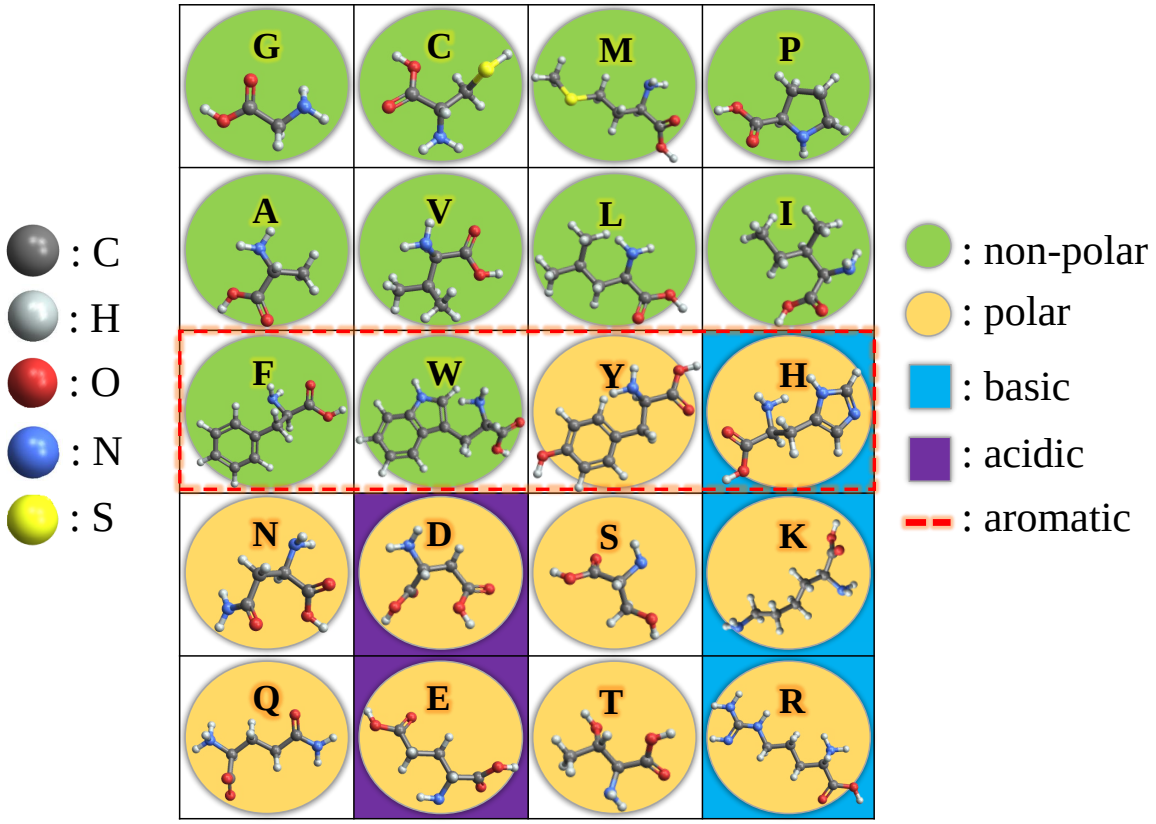


Figure 2.2: *Structure and properties of 20 major amino acids.*

has one amino terminus (N-terminus) and a carboxyl terminus (C-terminus) by forming peptide bonds. When amino acids are linked to form a polypeptide chain, the carboxyl group of one amino acid creates a covalent bond with the amine group of the next amino acid, and a water molecule is released[10].

A peptide of two monomers is called *dipeptide*. A peptide with the number of monomers less than 10 is called *oligopeptide* ("oligo" means "a few"). A peptides having more than 10 monomers is called *polypeptide* ("poly" means "many")[8]. Although proteins range in size from 30 to 10 000 amino acids[2], most proteins are large polypeptides containing from about 200 to 1000 amino acids[10].

Each protein performs its function by folding into a specific structure. The structure of protein is divided into four levels: primary, secondary, tertiary, and quaternary. Primary

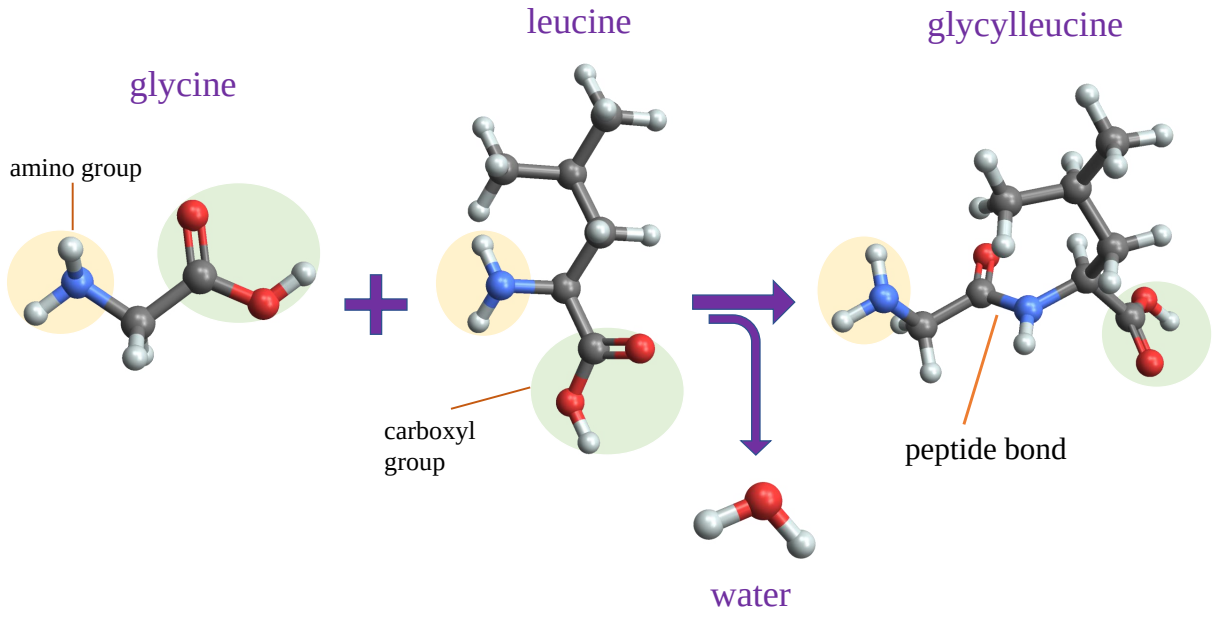


Figure 2.3: *A glycine amino acid connects with a leucine amino acid to form a glycylleucine by forming a peptide bond and releasing a water molecule.*

protein structure presents the order of amino acids within a polypeptide chain. Secondary protein structure occurs when the amino acids backbones interact with each other to form alpha helix or beta-pleated sheet conformations that maximize backbone hydrogen bonds. Tertiary protein structure is a complete folding pattern which produces a three-dimensional structure on a larger scale. At this level, non-polar amino acids tend to be in the interior while the polar ones are on the exterior to maximize the electrostatic interactions they can make in solution. At the final level of quaternary structure, multiple proteins combine together to form a larger structure[10,11].

2.2 Protein aggregation

Normally, newly synthesized proteins can fold into correct shapes very fast to become native proteins so they are able to perform their biological function. However, changing pH or temperature in the cellular environment, genetic mutation, translational errors and modifi-

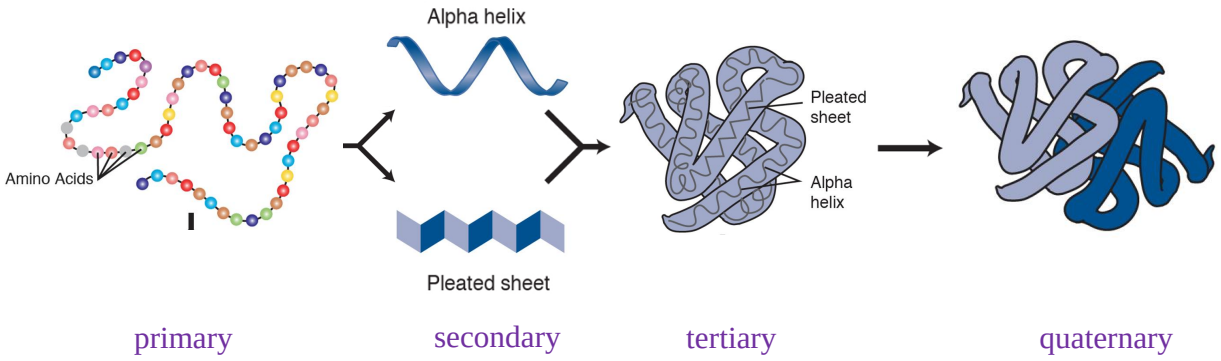


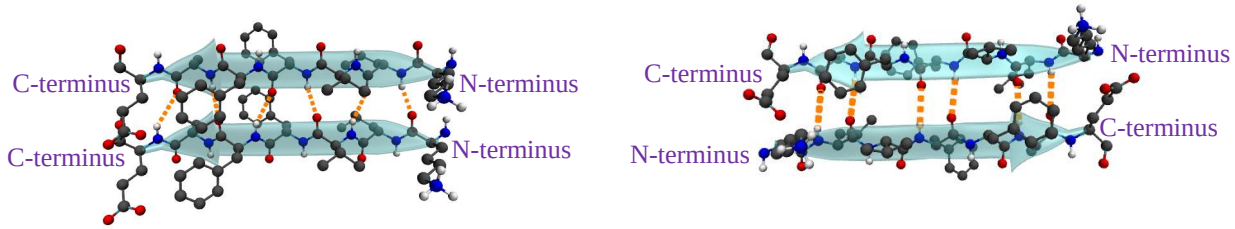
Figure 2.4: *Four levels of protein structure: primary, secondary, tertiary, and quaternary. Image from OpenStax Biology's modification of work by the National Human Genome Research Institute.*

cations in protein may cause misfolded and mutant proteins[12,13]. These proteins can cause aggregation. The aggregation can be disordered or ordered. Many diseases are now associated with protein aggregation and particularly with a form of ordered aggregate called the amyloid fibril. These diseases include Huntington's, Alzheimer's, Parkinson's, and prion[2,5].

The difference between those diseases is the types of protein. For instance, Huntington's disease is caused by mutant huntingtin proteins. They join together to make fibrils inside the nuclei of neuron cells, which make the cells die from the inside. Otherwhile, one of the main causes of Alzheimer's disease is $A\beta$ amyloid fragments from the amyloid precursor protein. $A\beta$ amyloid plaques form outside neuron cells which lead to cell death[14,15].

Two or more misfolded or mutant proteins may connect together to form insoluble clusters as aggregates such as micelle and amyloid fibrils. A micelle is a spherical oligomer which has a hydrophobic core and a hydrophilic exterior[16,17]. An amyloid fibril is a combination of many protein monomers which usually have β-sheet structure[14]. The fibril can be either a parallel β-sheet or antiparallel β-sheet. The molecules in parallel β-sheets have the same direction while the molecules in antiparallel ones have opposite directions which make the hydrogen bonds linear so they are more stable than those in parallel β-sheets (Fig. 2.5).

Fibril formation is divided into three stages including the lag phase, growth phase, and saturation phases which are described by a sigmoid curve. During the lag phase, monomers self-aggregate into oligomers and nuclei. During the growth phase, each nucleus is extended



(a) Parallel beta-sheet (b) Antiparallel beta-sheet

Figure 2.5: *Parallel (a) and antiparallel (b) β-sheets of $A\beta_{16-21}$. The molecules in parallel β-sheets have the same direction while the molecules in antiparallel ones have opposite directions. The hydrogen bonds are slanted in parallel β-sheets and linear in antiparallel β-sheets.*

rapidly and grows into fibrils. At the last stage, aggregation slows because the supply of free protein is depleted. Then many fibrils stick together to form fibers and fibril plaques[15,18].

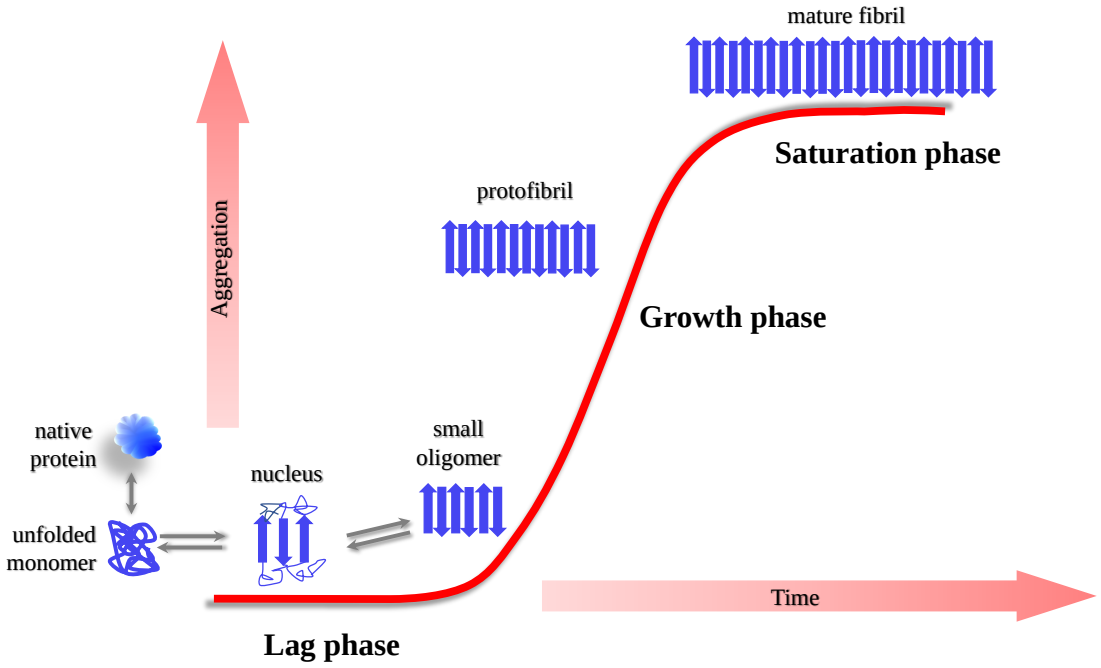


Figure 2.6: *A sigmoid curve describes the fibrillogenesis process. Adapted from Ref. Sgarbossa*[15].

**Huntington's disease is caused by the aggregation of huntingtin protein. Huntingtin aggregates can be both micelles and beta-sheet fibrils which contribute Huntington's progression*[19,20]

Huntington's disease (HD) was first discovered by Dr. George Huntington in 1872. Huntington's disease is very dangerous because it is genetic and causes many problems of movement, thinking, and emotion[5]. Many people struggling with this disease including $5-10$ individuals per 100 000 of the population in North America[21].

Huntington's disease is caused by mutant huntingtin protein. The structure of the huntingtin (Htt) molecule has three parts. The central part is polyglutamine Qn. The second part is the N17 tail which has a sequence MATLEKLMKAFESLKSF, and the third part is C38 which includes numerous prolines (P11-QLPQPPPQAQPLLPQPQ-P10). Huntington's disease is the result of a CAG repeat expansion in the gene which codes the polyglutamine part in huntingtin protein. Many previous studies show that the length of the polyglutamine core is central to HD[22,23]. Normal Htt has less than 36 glutamines (mainly 15-25 glutamines). However, abnormal Htt (or mutant Htt – mHtt) contains more than 36 glutamines which can form amyloid aggregation[21]. A paper by Crick et al. shows that polyQ causes aggregation but N17 and C38 modify it. Specifically, N17 accelerates fibril formation and C38 reduces fibril formation[19].

The thermodynamics of huntingtin aggregation is discussed in Chapter 3.

2.3 DNA

DNA (deoxyribonucleic acid) is one of the most essential molecules in biology and is a central ingredient in nanotechnology and biotechnology[25,26]. DNA contains instructions for all processes that happen in living organisms and carries code to produce proteins.

DNA is a very long molecule which consists of two complementary chains of nucleotides (or bases) called DNA strands[1]. Each nucleotide has three parts including a nitrogenous base (or nucleotide base), a sugar (or ribose or deoxyribose), and a phosphate group. There are four types of nucleotides: adenine (A), cytosine (C), thymine (T), guanine (G).

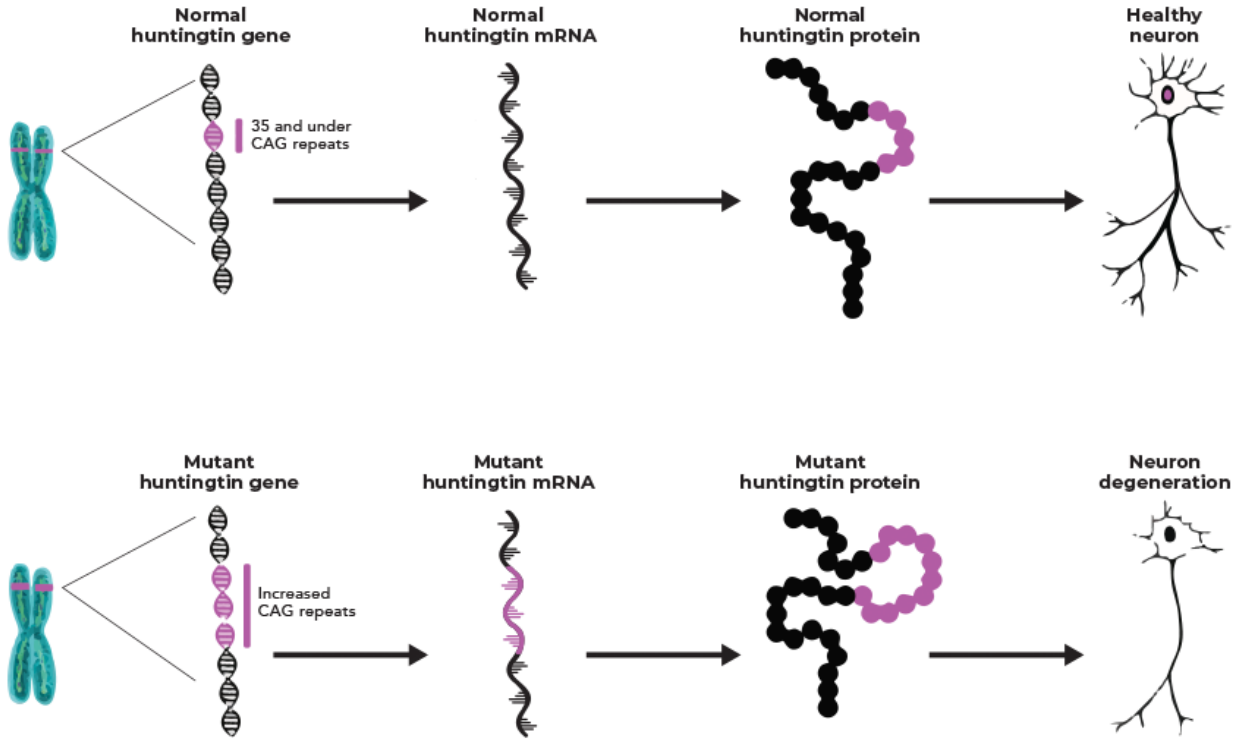


Figure 2.7: *The expansion of CAG repeats in the gene resulting in mutant huntingtin protein. Image from Genentech*[24].

The nucleotides in each DNA strand are linked by a backbone of covalent bonds between sugars and phosphates[1]. Two strands connect together when pairs of nucleotides between them form hydrogen bonds. This makes the DNA molecule look like a twisted ladder with the photsphate-sugar chains as rails and the base pairs forming the rungs[9]. This helical structure is the result of base-stacking interactions[10]. In double stranded DNA, the two DNA strands are antiparallel, which means in one strand the top is the phosphate terminus (5' end) and the bottom is the hydroxyl terminus (3' end) while the other strand is vice versa[27]. Adenine (A) always pairs with thymine (T) while guanine (G) always joins with cytosine (C) by forming hydrogen bonds. This is known as the Watson-Crick-Franklin base pairing rule, which is mentioned many times in Chapter 4. The A-T base pair has two hydrogen bonds but the G-C base pair forms up to three hydrogen bonds. That is one of the reasons why the G-C base pair is more stable than the A-T base pair[7].

The structure of the DNA molecule was investigated by Maurice Wilkins and Rosalind

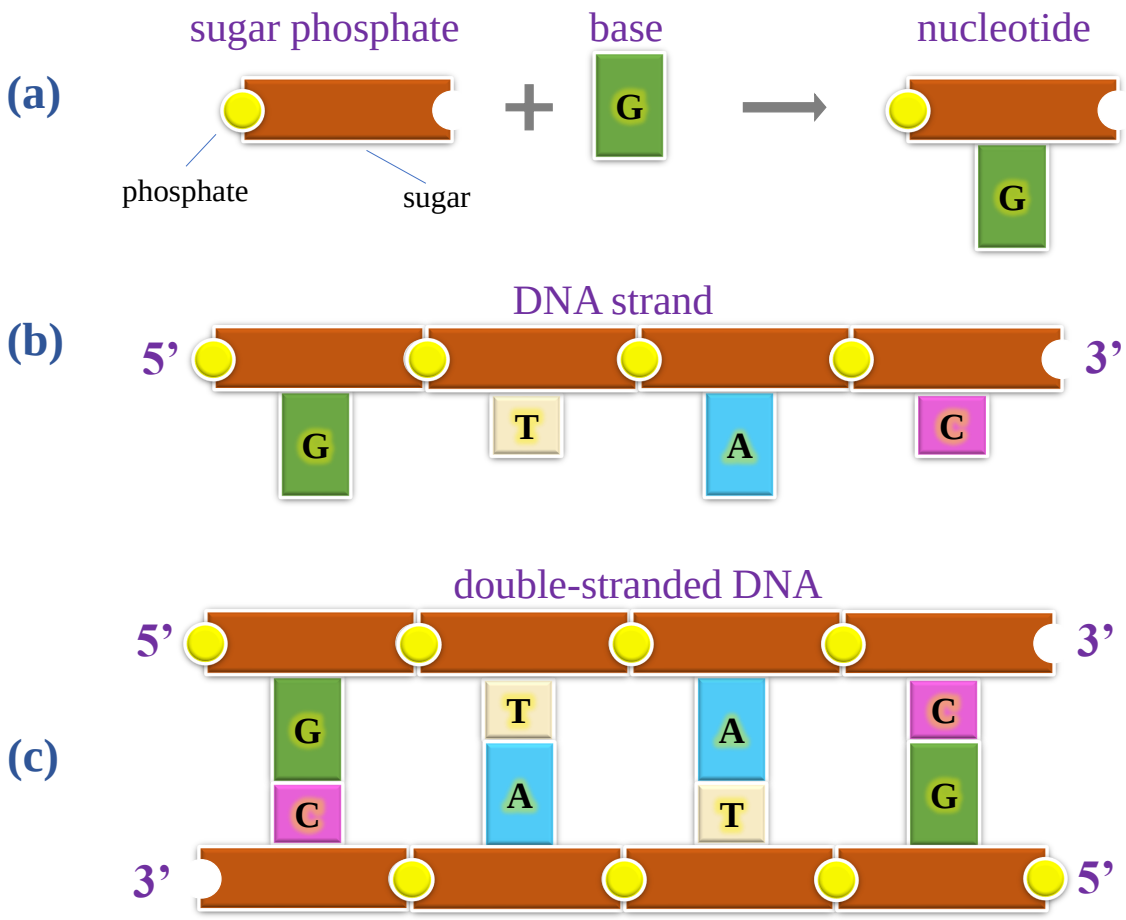


Figure 2.8: *(a) Structure of a nucleotide, (b) Structure of a DNA strand, (c) Two complementary DNA strands are held together by base pairs. Adapted from Ref. Alberts, Johnson, Lewis, Raff, Roberts, Walter*[1].

Franklin using X-ray diffraction in the 1950s. In 1953 Watson and Crick used Franklin' X-ray diffraction data to discover the double helix structure of DNA. Similar to protein and other biopolymers, the primary structure of DNA is the sequence of nucleotide bases, and the secondary DNA is a double helix.

Single-stranded DNA or double-stranded DNA can contain palindrome sequences which means the nucleotide sequence contains its own complement when read backward. The palindrome regions allow DNA to form intra-molecular structures such as hairpins, stem and loop structures and cruciforms depending on if the palindrome region is contiguous or non-contiguous (see Fig. 2.10)[10].

There are many algorithms to predict the structure of DNA molecules. In this study I

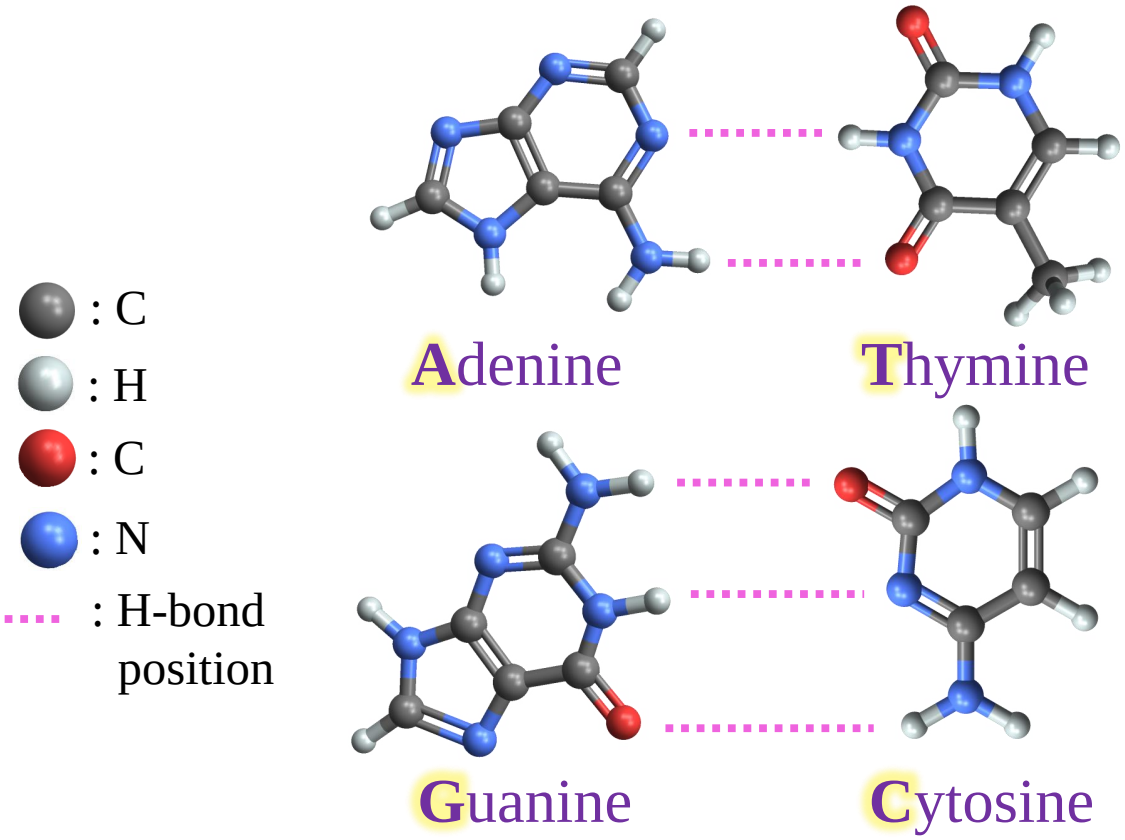


Figure 2.9: *Structure of four types of nucleotides. Adenine (A) always pairs with thymine (T) by forming two hydrogen bonds, while guanine (G) always joins with cytosine (C) and forms three hydrogen bonds. The pink dashed lines indicate the position of hydrogen bonds.*

used the NUPACK software to determine the structure of single-stranded DNA.

2.4 DNA hybridization

There are a few types of hybridization such as DNA-DNA hybridization, DNA-RNA hybridization, RNA-RNA hybridization. In this dissertation, I consider DNA-DNA hybridization, or DNA hybridization for short.

DNA hybridization is a combination of two single-stranded DNA molecules joined by base pairs to form a double stranded DNA molecule. Prior to hybridization it is necessary to obtain single stranded molecules. This can be done by denaturation in high temperature. The hybridization process begins when one DNA strand anneals with a complementary

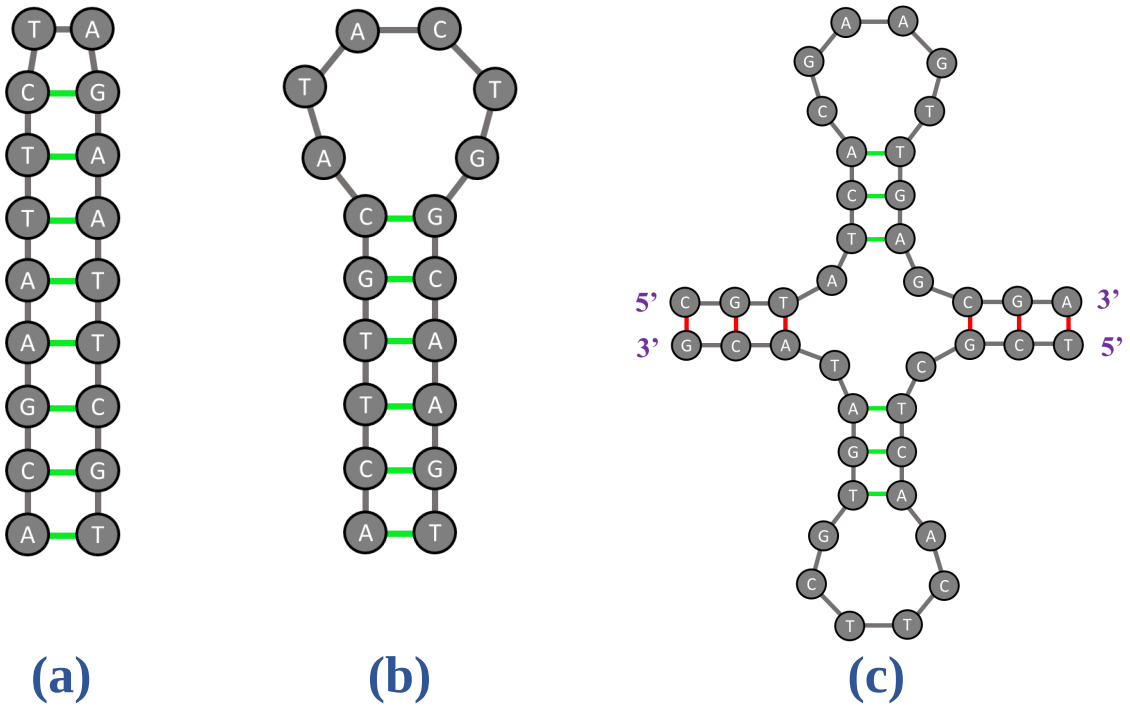


Figure 2.10: *(a) A contiguous palindrome forms a hairpin structure. (b) Non-contiguous palindrome forms loop and stem structure. (c) Non-contiguous palindromes form the cruciform. Adapted from Ref. Goldfarb*[10].

single DNA strand. DNA hybridization can be affected by many factors such as pH, salt concentration, temperature, solvent properties, G-C content, DNA length, and probe density.

DNA hybridization has numerous applications in fields like nanotechnology[25] and biotechnology[26]. Applications include HPV tests, HIV tests[28], cancer diagnosis[29], PCR tests, genotyping and other genomic diagnostics[30–32], DNA origami[31,33–37], etc.

**The stability of DNA helices can be computed with the nearest-neighbor model.*

Mentioned in section 2.3, the G-C base pair is more stable than A-T. The free energy required to break a G-C base pair can be 2 to 3 times that of an A-T pair. However, this free energy of one base pair is not only affected by the number of hydrogen bonds but also by the neighboring base pairs. A common approach to compute the free energy of base pairing is the nearest-neighbor approximation, which assumes that the stability of one base pair is determined by the identity and orientation of the nearest neighbor one[10,38]. In this study

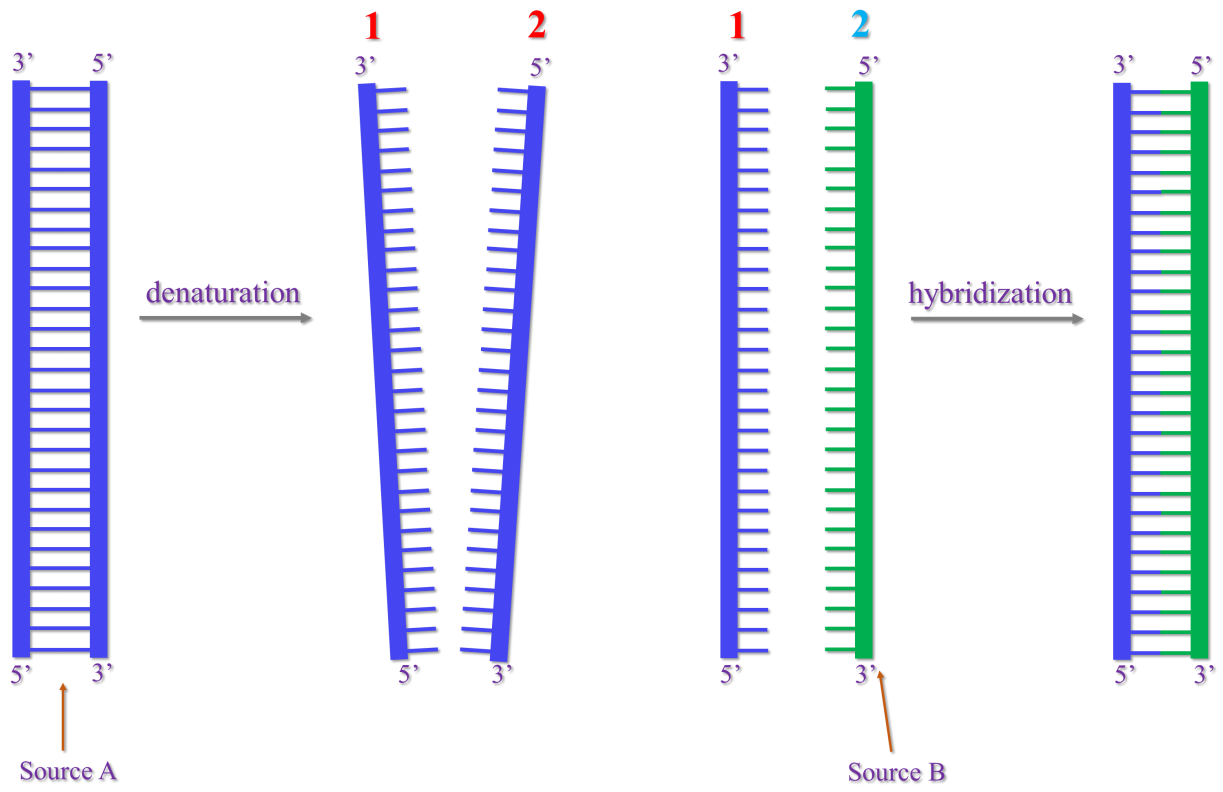


Figure 2.11: *DNA denaturation and DNA hybridization processes.*

we use the nearest neighbor model for Watson-Crick-Franklin (WCF) base pairs, which is unified from databases provided by many labs[30]. A series of articles provide parameters of entropy (ΔH), enthalpy (ΔS) and free energy (ΔG) in 1 M NaCl at 37^oC for WCF pairs and mismatched pairs.[30,38–43]. These parameters can be modified to account for different NaCl concentrations and different temperatures as follows

$$\Delta S[Na^+] = \Delta S[1MNaCl] + 0.368 \cdot \frac{N}{2} \cdot ln[Na^+] \tag{2.1}$$

$$\Delta G_T[Na^+] = \Delta H - T \cdot \Delta S[Na^+] \tag{2.2}$$

where N is the total number of phosphates in the duplex, $[Na^+]$ is the concentration of NaCl in units of mol, T is the temperature in Kelvin, ΔH expressed in units of $cal.mol^{-}1$, and ΔS is in units $K^{-1}mol^{-1}$ (or e.u)[30].

2.5 Connecting bonds between molecules

While the peptide bonds (or backbones) in protein and DNA molecules are covalent bonds, the monomers on the incoming molecule and templating molecules link together by hydrogen bonds (H-bonds). In this section, I introduce many types of bonds which link molecules together including covalent bonds and noncovalent bonds, especially hydrogen bonds which are the main contribution in the biopolymer templating problem.

A covalent bond is formed when two atoms share electrons[1,3]. Peptide bonds in protein polymers and the sugar-phosphate bonds in DNA molecules are two examples of covalent bonds. The strength of the covalent bond in water is about 90 kcal/mol[1]. In contrast, noncovalent bonds have minimal electron sharing and are primarily electrostatic. We can divide noncovalent bonds into three main types: ionic, hydrogen bond, and van der Waals attraction.

Ionic bonds are a bond formed between two oppositely charged ions. Their strength in water is about 3 kcal/mol[1].

Hydrogen bond is a type of dipole-dipole interaction occurring when a hydrogen atom links an electronegative atom. Oxygen and nitrogen are more electronegative than a hydrogen atom so in the O-H interaction and N-H interaction, oxygen and nitrogen atoms pull the electron of hydrogen atom toward them which result in dipoles. The hydrogen bond is created when a dipole is formed. The strength of the hydrogen bond is about 1 kcal/mol in water[1].

Van der Waals attractions occur when electrons of a nonpolar atom fluctuate and result in a dipole. This dipole electron cloud induces nearly atoms to polarize as well. The two atoms produce a very weak attraction (the strength is about 0.1 kcal/mole[1]) between them called the van der Waals force.

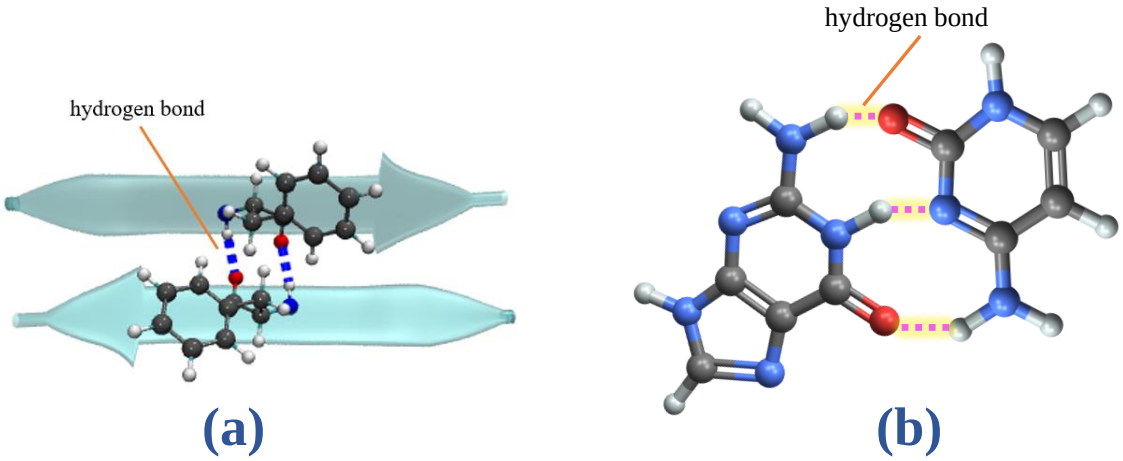


Figure 2.12: *(a) Two amino acids of phenylalnine in two opposite chains in a β-sheet connect together by forming H-bonds. (b) Nucleotide G connects with nucleotide C via three H-bonds.*

2.6 Random walk model

The random-walk problem was first introduced by Karl Pearson (1857-1936) in 1905[44]. The random walk model describes the diffusive motion of a "walker" which can be in one dimension, two dimensions or three dimensions (or random-flight model). In this dissertation, I use the term "random-walk model" to describe a one-dimensional random-walk.

The "walker" can be a particle, a molecule, a linear flexible polymer chain, or number of H-bonds. Each step of the walk takes a unit of time. The direction of each step can randomly be either "forward" (to the right) or "backward" (to the left). Each step is independent of each other. That means the walker does not remember previous steps.[44,45].

In the templating problem, when an incoming molecule interacts with the existing template, they will form and break hydrogen bonds. The 1D random walk model is used to

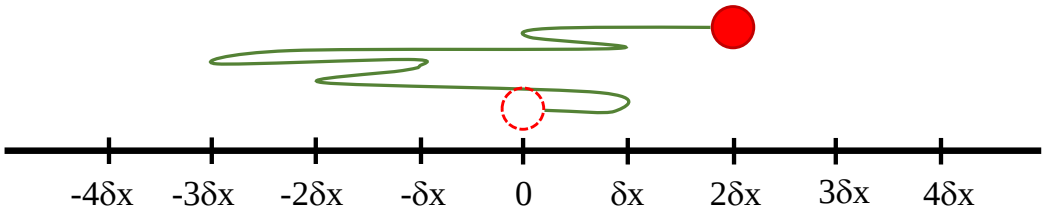


Figure 2.13: *The walker (red ball) moves randomly backward and forward from the origin in one dimension space. Each step consumes the same amount of time δx*

represent the number of hydrogen bonds.

In Fig. 2.14, the number x of H-bonds between two molecules is in the range $[0, N]$ (here $N = 8$), where N is the number of monomers in the incoming molecule. When $x = 0$, there are no H-bond between two molecules. If $x = N$, the two molecules are in perfect alignment with all bonds forming.

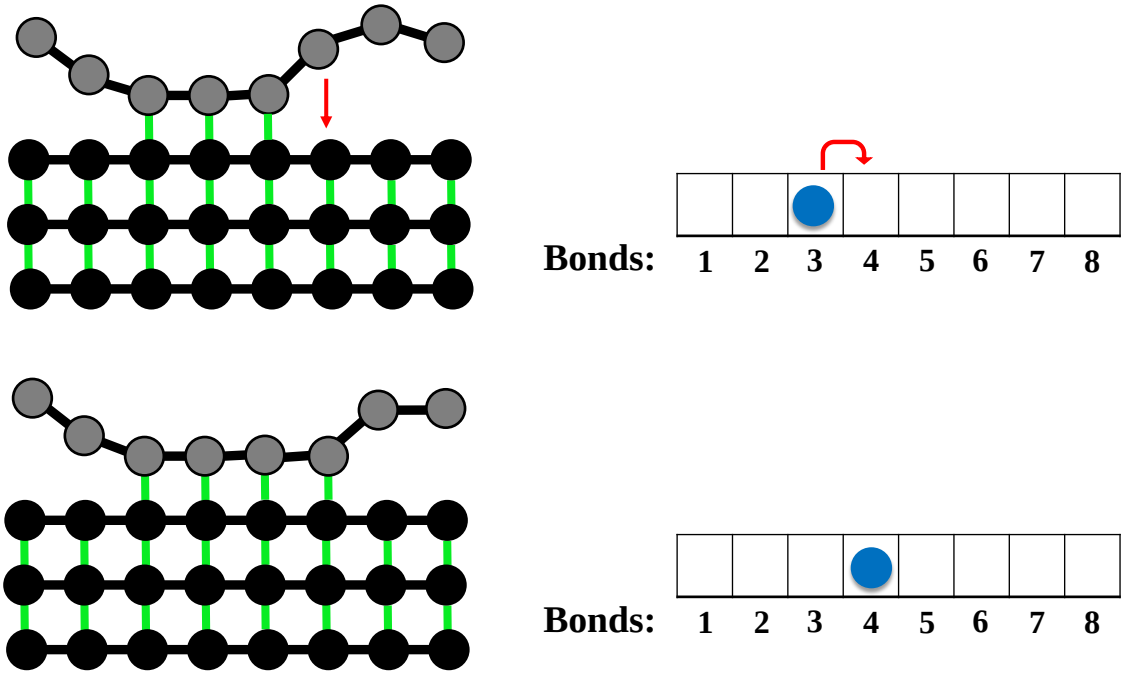


Figure 2.14: *The incoming molecule can form and break H-bonds with the existing template (left). The number of H-bonds (right) describes the kinetics of the system. Adapted from Ref. Schmit*[46].

References

[1] B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter (2007).

[2] B. Alberts, D. Bray, K. Hopkin, A. D. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology* (Garland Science, 2015).

[3] S. Walker and D. M. McMahon, *Biochemistry Demystified* (McGraw-Hill, 2008).

[4] R. Nassar, G. L. Dignon, R. M. Razban, and K. A. Dill, Journal of Molecular Biology p. 167126 (2021).

[5] J. Dodds, *Biology at a Glance* (CRC Press, 2019).

[6] M. J. Lopez and S. S. Mohiuddin, StatPearls [Internet] (2021).

[7] J. G. Salway, *Medical biochemistry at a glance* (John Wiley & Sons, 2012).

[8] D. Voet, J. G. Voet, and C. W. Pratt, *Fundamentals of biochemistry: life at the molecular level* (John Wiley & Sons, 2016).

[9] C. I. Branden and J. Tooze, *Introduction to protein structure* (Garland Science, 2012).

[10] D. Goldfarb, *Biophysics demystified* (McGraw-Hill, 2011).

[11] R. B. Phillips, J. Kondev, J. Theriot, H. Garcia, B. Chasan, et al. (2009).

[12] F. Chiti and C. M. Dobson, Annu. Rev. Biochem. **75**, 333 (2006).

[13] L. M. Fox and A. Yamamoto, in *Autophagy: Cancer, Other Pathologies, Inflammation, Immunity, Infection, and Aging*, edited by M. Hayat (Academic Press, Amsterdam, 2015), pp. 117–137, ISBN 978-0-12-801043-3, URL https://www.sciencedirect.com/science/article/pii/B9780128010433000078.

[14] R. N. Rambaran and L. C. Serpell, Prion **2**, 112 (2008).

[15] A. Sgarbossa, International journal of molecular sciences **13**, 17121 (2012).

[16] Z. P. Aguilar, in *Nanomaterials for Medical Applications*, edited by Z. P. Aguilar (Elsevier, 2013), pp. 33–82, ISBN 978-0-12-385089-8, URL https://www.sciencedirect.com/science/article/pii/B9780123850898000029.

[17] K. T. Oh, T. K. Bronich, and A. V. Kabanov, Journal of Controlled Release **94**, 411 (2004), ISSN 0168-3659, URL https://www.sciencedirect.com/science/article/pii/S0168365903005030.

[18] C.-C. Lee, A. Nayak, A. Sethuraman, G. Belfort, and G. J. McRae, Biophysical journal **92**, 3448 (2007).

[19] S. L. Crick, K. M. Ruff, K. Garai, C. Frieden, and R. V. Pappu, Proceedings of the National Academy of Sciences **110**, 20075 (2013).

[20] A. E. Posey, K. M. Ruff, T. S. Harmon, S. L. Crick, A. Li, M. I. Diamond, and R. V. Pappu, Journal of Biological Chemistry **293**, 3734 (2018).

[21] M. P. Parsons and L. A. Raymond, in *Neurobiology of Brain Disorders*, edited by M. J. Zigmond, L. P. Rowland, and J. T. Coyle (Academic Press, San Diego, 2015), pp. 303–320, ISBN 978-0-12-398270-4, URL https://www.sciencedirect.com/science/article/pii/B9780123982704000203.

[22] D. R. Langbehn, M. R. Hayden, J. S. Paulsen, and P.-H. I. of the Huntington Study Group, American Journal of Medical Genetics Part B: Neuropsychiatric Genetics **153**, 397 (2010).

[23] F. O. Walker, The Lancet **369**, 218 (2007).

[24] Genentech (2019), URL https://www.huntingtonsdiseasehcp.com/mhtt-the-fundamental-cause.html.

[25] T. E. Ouldridge, P. Šulc, F. Romano, J. P. Doye, and A. A. Louis, Nucleic acids research **41**, 8886 (2013).

[26] Y. Yin and X. S. Zhao, Accounts of chemical research **44**, 1172 (2011).

[27] J. L. Nadeau, *Introduction to experimental biophysics: biological methods for physical scientists* (CRC Press, 2017).

[28] G. J. Netto, A. N. Tawil, J. T. Newman, and D. A. Savino, in *Baylor University Medical Center Proceedings* (Taylor & Francis, 1990), vol. 3, pp. 45–52.

[29] Z. Jehan, S. Uddin, and K. S Al-Kuraya, Current medicinal chemistry **19**, 3730 (2012).

[30] J. SantaLucia Jr and D. Hicks, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).

[31] J. M. Huguet, M. Ribezzi-Crivellari, C. V. Bizarro, and F. Ritort, Nucleic acids research **45**, 12921 (2017).

[32] C. T. Wittwer, Human mutation **30**, 857 (2009).

[33] K. E. Bujold, A. Lacroix, and H. F. Sleiman, Chem **4**, 495 (2018).

[34] P. W. Rothemund, Nature **440**, 297 (2006).

[35] S. M. Douglas, H. Dietz, T. Liedl, B. Högberg, F. Graf, and W. M. Shih, Nature **459**, 414 (2009).

[36] Y. Dorsett and T. Tuschl, Nature reviews Drug discovery **3**, 318 (2004).

[37] K. E. Dunn, F. Dannenberg, T. E. Ouldridge, M. Kwiatkowska, A. J. Turberfield, and J. Bath, Nature **525**, 82 (2015).

[38] J. SantaLucia, Proceedings of the National Academy of Sciences **95**, 1460 (1998).

[39] N. Peyret, P. A. Seneviratne, H. T. Allawi, and J. SantaLucia, Biochemistry **38**, 3468 (1999).

[40] H. T. Allawi and J. SantaLucia, Biochemistry **37**, 9435 (1998).

[41] H. T. Allawi and J. SantaLucia, Biochemistry **37**, 2170 (1998).

[42] H. T. Allawi and J. SantaLucia, Biochemistry **36**, 10581 (1997).

[43] H. T. Allawi and J. SantaLucia Jr, Nucleic acids research **26**, 2694 (1998).

[44] K. A. Dill, S. Bromberg, and D. Stigter, *Molecular driving forces: statistical thermodynamics in biology, chemistry, physics, and nanoscience* (Garland Science, 2010).

[45] I. Teraoka (2002).

[46] J. D. Schmit, The Journal of chemical physics **138**, 05B611_1 (2013).

Chapter 3

Thermodynamics of Huntingtin Aggregation

* *We have published this chapter in* **Biophysical Journal 118, 2989 (2020)**.

Amyloid aggregates are found in many neurodegenerative diseases including Huntington's, Alzheimer's, and prion diseases. The precise role of the aggregates in disease progression has been difficult to elucidate due to the diversity of aggregated states they can adopt. Here we study the formation of fibrils and oligomers by exon 1 of huntingtin protein. We show that the oligomer states are consistent with polymer micelles that are limited in size by the stretching entropy of the polyglutamine region. The model shows how the sequences flanking the amyloid core modulate aggregation behavior. The N17 region promotes aggregation through weakly attractive interactions, while the C38 tail opposes aggregation via steric repulsion. We also show that the energetics of cross-β stacking by polyglutamine would produce fibrils with many alignment defects, but minor perturbations from the flanking sequences are sufficient to reduce the defects to the level observed in experiment. We conclude with a discussion of the implications of this model for other amyloid forming molecules.

3.1 Introduction

Protein aggregates are implicated as the causative factor in numerous diseases, including neurodegenerative diseases like Alzheimer's and Huntington's[1,2]. The most conspicuous of these assemblies are insoluble fibrils consisting of molecules stacked in a cross-β motif. However, the predominant evidence is that disease progression is actually driven by smaller, soluble oligomers[3]. These states are more difficult to study than fibrils because they tend to be transient and heterogeneous. In most cases it is believed that the oligomers are metastable with respect to the fibril, but favored kinetically due to the fact that they lack the large nucleation barrier associated with fibril formation[4–10]. Confounding the issue is the fact that in *vitro* conditions inevitably differ from those *in vivo*, raising the question of whether the oligomers observed in the lab are the same as those occurring naturally. This question would be more readily answered with an understanding of the nature and stability of the various states.

The common features of amyloid diseases give rise to another question. To what extent is aggregation and toxicity dependent on the specific sequence and structural states of the proteins? An interesting case study for this question is exon 1 of huntingtin protein, which contains a polyglutamine (polyQ) core flanked by short, unstructured sequences at the N- and C-terminal ends. The aggregation behavior is driven by the polyQ core, with increasing polyQ lengths correlating with earlier disease onset[11,12]. However, the terminal sequences modulate the aggregation propensity with the N-terminus promoting aggregation and the C-terminus promoting higher solubility[13]. The behavior of the latter sequence is not surprising as the C-terminal fragment is composed primarily of proline residues. However, the aggregation promoting property of the N-terminus is more difficult to understand as this segment has a high solubility in isolation [Rohit Pappu, personal communication].

Huntingtin (Htt) shows qualitatively similar aggregation behavior to other amyloid proteins with distinct fibril and oligomer states. The low sequence complexity of huntingtin suggests that these states are not due to sequence-specific interactions, but arise more generally from the polymer nature of the molecule. Here we show that the stability of these

states can be modeled by treating huntingtin as a triblock copolymer.

For a simple polymer, we expect two limiting behaviors; either the swollen random walk of a polymer in good solvent, or the collapsed state typical of a polymer in poor solvent. Recent experiments and simulations have shown that monomeric huntingtin adopts conformations consistent with the poor solvent case[14,15]. Accordingly, we allow the collapsed globules in our model to coalesce further to form copolymer micelles, which we associate with the oligomer state. To account for the fibril state we add a second form of intermolecular interactions in which the backbones and sidechains pack more efficiently at the cost of conformational entropy. Surprisingly, experiments have shown that huntingtin fibrils are highly ordered despite the discrete translational symmetry of the polyglutamine core[16]. We show that this alignment specificity arises naturally from the energetics of the binding ensemble, and it is further assisted by the N- and C- terminal regions.

3.2 Model

3.2.1 Monomer and oligomer are modeled as collapsed globule.

To construct our free energy for Htt, we take the reference state to be a well-solvated Flory coil. In this state, contacts between amino acids are negligible and the random walk entropy is maximized. This state is purely hypothetical because experiments and simulation have shown that Htt adopts configurations consistent with a polymer in poor solvent[15,17–19]. This means that favorable interactions between amino acids are sufficient to pay the entropy cost to collapse the random coil into a globule. These same interactions can also drive the condensation of Htt molecules into oligomers. Since monomer collapse and oligomer formation are driven by the same desolvation reaction, we describe them both by the free energy

$$F_{\text{glob}}(N, \ell_Q, \ell_N) = F_{\text{cont}} + F_{\text{ent}} + F_{\text{C38}} \tag{3.1}$$

where the terms represent the amino acid contact energy, the change in conformational entropy, and the contribution from the C38 tail. N is the size of oligomer in monomer units

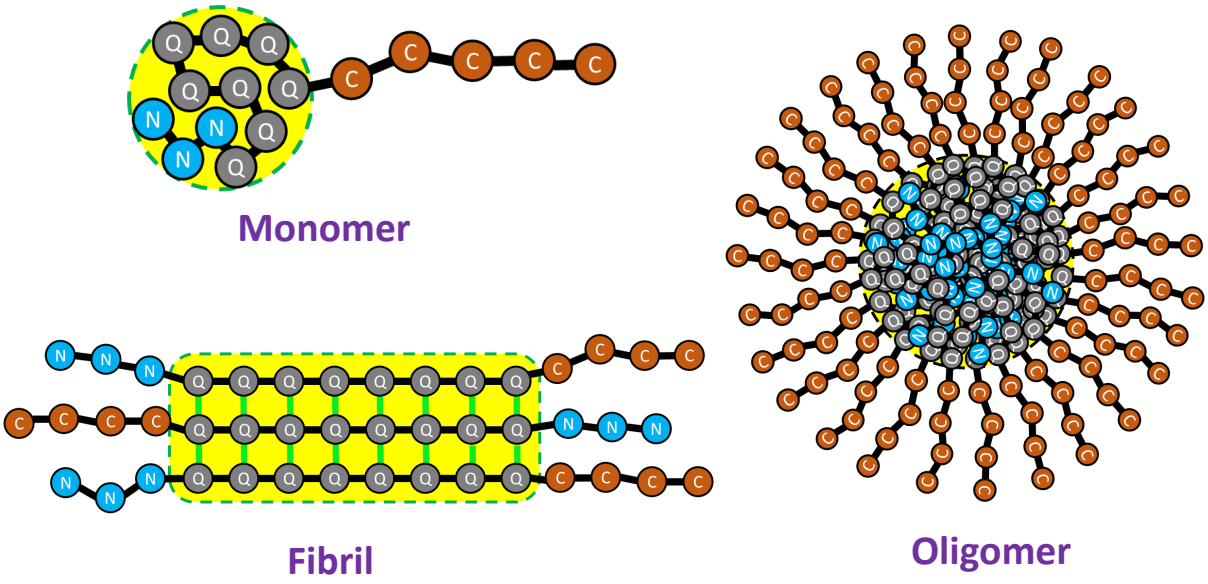


Figure 3.1: *Cartoon representation of the three states of Htt. In the monomer state, the peptide collapses into a globule containing both polyQ and N-terminal regions. The oligomer state is a micelle-like assembly of a few thousand monomers with a spherical core containing the polyQ and N-terminal regions. The fibril state is a cross-β amyloid core of polyglutamine flanked by disordered tails on both sides.*

and $\ell_{Q/N}$ are the number of amino acids in the polyQ and N-terminal (if present) regions. The contact energy has a bulk and surface term

$$\frac{F_{\rm cont}}{k_BT} = LN\varepsilon_G + A_\gamma \varepsilon_G (NL)^{2/3} \tag{3.2}$$

Simulations have shown that the collapsed globule contains both the polyQ region and the N-terminal region[14,17]. Therefore the bulk term is proportional to the total length of these regions $L = \ell_Q + \ell_N$ where $\ell_N = 17$ for molecules containing the N-terminal segment and $\ell_N = 0$ for molecules without the tail. For the burial energy we take a weighted average for the desolvation of glutamine and N17 amino acids.

$$\varepsilon_G = \frac{\ell_Q \varepsilon_Q + \ell_N \varepsilon_N}{\ell_Q + \ell_N} \tag{3.3}$$

The bulk term, $LN\varepsilon_G$, over counts the driving force for collapse because residues at the

surface of the globule are incompletely desolvated. This is corrected by the surface term $A_\gamma \varepsilon_G (NL)^{2/3}$. Simple geometrical considerations give a value $A_\gamma = -2.4$ for the constant (see Appendix).

The entropic contribution to the free energy takes the form

$$\frac{F_{\text{ent}}}{k_B T} = \left(\frac{9}{16\pi^2 L}\right)^{1/3} N^{5/3} + N\frac{L}{g} \tag{3.4}$$

The two terms account for polymer stretching in the oligomer state and the compression of the coils in the collapsed monomer state, respectively. In practice, only one of these terms is significant at any time, which allows for the additive approximation in Eq. 3.4.

The stretching term provides a repulsive energy that arrests oligomer growth. This term arises in polymer micelles because the molecules must extend from the surface to fill the interior of the aggregate when its radius grows longer than the radius of gyration of a random walk polymer[20]. This stretching energy will arrest oligomer growth as long as one end of the molecule remains solvated at the surface of the oligomer. In the full length exon 1 this role will be filled by the C38 region. For constructs lacking C38, surface pinning is likely due to the two lysine residues placed at the C-terminus[13]. It is also possible that the N17 and polyQ regions demix in the oligomer core. Without surface pinning of the C-terminus, this demixing would result in an inverted structure with the more soluble N17 at the surface and polyQ in the interior.

To compute the stretching term we note that each monomer in the oligomer has a stretching energy of $k_B T R^2/R_g^2$ where R is the oligomer radius (calculated in the Appendix) and $R_g = a\sqrt{L}$ is the radius of gyration for a monomer. Here we have taken the Flory exponent to be $\nu = 1/2$ since the excluded volume swelling of the polymer will be screened by excluded volume interactions with neighboring molecules. The total stretching energy for the whole oligomer is then

$$\frac{F_{\text{str}}}{k_B T} = N\frac{\left(\frac{3NL}{4\pi}\right)^{2/3} a^2}{a^2 L} \tag{3.5}$$

which simplifies to the first term in Eq. 3.4.

In the monomer state, the molecules face the opposite problem where the entropic loss is due to compression of the random coil. For a polymer under confinement, the free energy change can be estimated by the blob model[21]. In this model, the polymer can be subdivided into statistically independent segments that are each small enough that the effects of confinement are not felt. Confinement effects arise at the interface between these statistical blobs, where it exerts a perturbation on the order of k_BT. Therefore the free energy of confinement is $\sim L/(g\ell_k)$ where g is the number of statistically independent segments per blob, and ℓ_k is the Kuhn length. The number of segments per blob can be found by requiring that the segment density per blob $g/(ag^\nu)^3$ is equal to the density of the entire system, L/V, where V is the confinement volume. Therefore $g^{3\nu-1} = V/a^3L$. In this case, we are interested in a collapsed globule where the confinement volume is equal to the total volume of the chain $V = La^3$. This gives $g = 1$. Therefore the compression free energy is just k_BT times the number of Kuhn lengths. To estimate this, we note that the persistence length of polyQ is about 1.3 nm[22] and that the Kuhn length is twice the persistence length[23]. Therefore, the statistical correlation along a polymer extends over $\ell_k \simeq 2.6$ nm/ 0.3 nm $\simeq 8.7$ amino acids, where we have taken 0.3 nm as the contour length per amino acid.

The final contribution to the globule free energy comes from the C-terminal tail. This region has the sequence P_{11}-QLPQPPPQAQPLLPQPQ-P_{10}. Given the limited flexibility of proline and the propensity to form polyproline helices, this tail will be more rigid, although largely disordered[24]. We assume that the tails interact primarily by excluded volume interactions. Due to the non-uniform flexibility of the tails, it is difficult to apply the blob model to compute the confinement effect due to neighboring tails. Still, inspection of the sequence suggests that 1-3 blobs per tail is reasonable. In fact, our results are insensitive to values in this range. Here we report results for $f_{\mathrm{C38}} = 2\ k_BT$.

3.2.2 The fibril state is a cross-β core with disordered tails.

Htt fibrils consist of a cross-β core that spans the polyQ region but does not include N17 or C38[16,24]. Evidence suggests that the β-sheet core is most likely anti-parallel as shown in Fig.

3.1[25], although parallel cores may also occur[26]. The specifics of parallel or anti-parallel do not enter our model since the parameters give the average interaction experienced by each sequence block.

We write the fibril free energy as a β-sheet term that scales linearly with the length of the polyQ core, modified by perturbations from the terminal segments.

$$\frac{F_{\text{fibril}}}{k_B T} = \ell_Q \varepsilon_\beta + f_{\text{C38}} + f_{\text{N17}} \tag{3.6}$$

Here ε_β is the free energy gain to move one glutamine residue from the solvated random coil state into the cross-β core. The second term accounts for the interaction between C-terminal segments which we expect to be the same as in the globule state. Finally, f_{N17} accounts for the interaction of N17 tails. These segments are soluble, but not purely repulsive like the proline-rich C38[16]. To account for the possibility of sequence specific interactions between N17 tails, we obtain this parameter by fitting.

3.2.3 Critical concentrations are computed from the change in free energy.

Our next task is to compute the concentration dependence of fibril and oligomer formation. At low concentration the attractive interactions cannot overcome the translational entropy cost of condensing the molecules. At higher concentrations equilibrium is established when the monomer pool is depleted to the point where the translational entropy cost balances the attractive free energy. In the micellization literature this point is called the critical micelle concentration (CMC). This terminology has been adopted to define a critical oligomer concentration (COC) and critical fibril concentration (CFC) in amyloid systems[27,28]. These critical concentrations should not be confused with the critical point of a phase transition and are more similar to the saturation concentration at an arbitrary point along the coexistence (binodal) line. However, the finite size of oligomers results in a more gradual transition than the sharp solubility limit of a macroscopic phase transition[28]. This introduces some

ambiguity in the definition of the critical concentration, although in practice, the transition is sufficiently sharp that this is not experimentally significant.

Following reference[28], we start by writing down the equilibrium constant for N-fold oligomerization.

$$K_a = \frac{C_N C_0^{N-1}}{{C_1}^N} = \exp\left(-\frac{\Delta F_{\mathrm{MO}}}{k_B T}\right) \tag{3.7}$$

where

$$\Delta F_{\mathrm{MO}} = F_{\mathrm{oligomer}} - N F_{\mathrm{monomer}} \tag{3.8}$$

where C_0 is a reference concentration. We identify the critical concentration for oligomerization as the point where there is an equal amount of protein in the monomer and oligomer states $C_1 = NC_N$, which can be combined with Eq. 3.7 to yield a relationship between the critical concentration and the free energy of the oligomer state

$$C_1^{(COC)} = \quad C_0 \left[\frac{1}{N}\exp\left(\frac{\Delta F_{\mathrm{MO}}}{k_B T}\right)\right]^{\frac{1}{N-1}} \tag{3.9}$$

Eq. 3.9 requires the size of the oligomer, N, which we obtain by minimizing Eq. 3.8

$$\begin{aligned}\frac{\partial \Delta F_{\mathrm{MO}}}{\partial N} = L\varepsilon_G + \frac{2}{3}A_\gamma \varepsilon_G \left(\frac{L^2}{N}\right)^{1/3} + \frac{5}{3}\left(\frac{9}{16\pi^2 L}\right)^{1/3} N^{2/3} + \frac{L}{g} + f_{\mathrm{C38}} \\ - \left[L\varepsilon_G + A_\gamma \varepsilon_G L^{2/3} + \left(\frac{9}{16\pi^2 L}\right)^{1/3} + \frac{L}{g}\right] = 0\end{aligned} \tag{3.10}$$

The formation of fibrils can also be associated with a critical concentration. However, unlike the soft transition seen in oligomers, the critical concentration for fibril formation is very sharp[28], analogous to the solubility limit in a phase transition. The critical concentration for fibril formation is[28]

$$C_1^{(CFC)} = C_0 \exp\left(\frac{\Delta F_{\mathrm{MF}}}{k_B T}\right) \tag{3.11}$$

where

$$\Delta F_{\mathrm{MF}} = F_{\mathrm{fibril}} - F_{\mathrm{monomer}} \tag{3.12}$$

is the free energy for transferring a molecule from the monomer state to the fibril state.

3.2.4 Fibril alignment defects incur a free energy penalty.

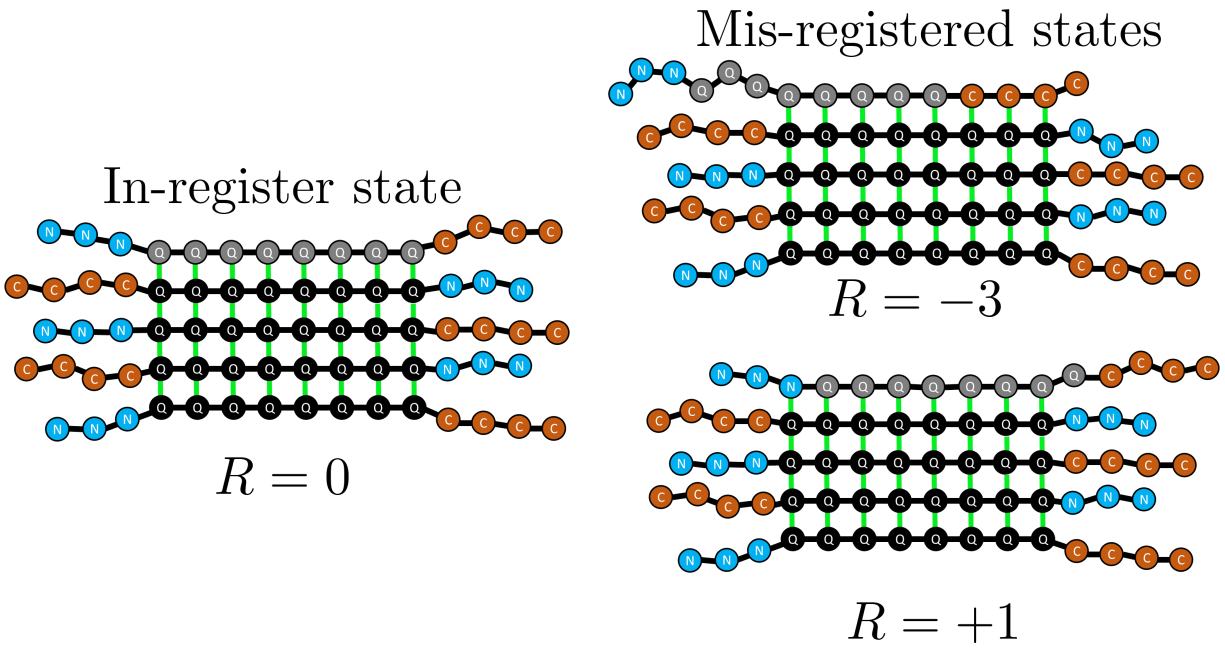


Figure 3.2: *Cartoon representation of the in-register state and mis-registered states. The registry variable, R, defines the alignment of an incoming molecule with the existing fibril. $R = 0$ indicates perfect alignment of the polyQ region, while negative and positive values indicate N-terminal and C-terminal shifts, respectively.*

Atomic resolution models of amyloid fibrils show striking order in the alignment of molecules[29]. However, it is not known whether this order is generally present or if it is an artifact of structural methods that are limited to systems that possess such order. PolyQ aggregates represent an extreme test of the alignment tendency of amyloids due to the discrete translational symmetry they possess.

Here we introduce an equilibrium model to compute the frequency of alignment defects in polyQ fibrils. Following previous work[30], we quantify the alignment using the registry

variable R, which can take the values $-\ell_Q < R < +\ell_Q$, where ℓ_Q is the number of glutamine residues in each molecule. The value $R = 0$ denotes the in-register state, positive values of R indicate that the incoming molecule is shifted toward its C-term, while negative values of R indicate a shift toward the N-term (see Fig. 3.2).

For mis-aligned states with $R < 0$, there will be H-bonds between glutamines of the existing fibril and amino acids in C38 of the incoming molecule. Conversely, if $R > 0$, there will be H-bonds between glutamines of the existing fibril and N17 of the incoming molecule.

We compute the probability for a given alignment by

$$P(R) = \frac{e^{-(\ell_Q - |R|)\varepsilon_\beta - |R|\varepsilon_M}}{\sum_R e^{-(\ell_Q - |R|)\varepsilon_\beta - |R|\varepsilon_M}} \tag{3.13}$$

where the denominator is the partition function for the alignment states and ε_M is equal to $\varepsilon_{\mathrm{NQ}}$ or $\varepsilon_{\mathrm{CQ}}$ to account for interaction between the polyQ core and the N- or C-terminal tail of mis-aligned molecules. The mis-alignment energy is not symmetric because we assume that residues lying outside the β-core are too disordered to have a significant interaction energy.

3.3 Results

3.3.1 Polyglutamine desolvation competes with polymer entropy.

To obtain values for the energies appearing in the model we fit the calculated critical concentration to the experiments of Crick et al[13] and Posey et al[31]. There are four free parameters: ε_Q, ε_N, ε_β, and f_{N17}. To fix the reference concentration we adopt a lattice approximation in which the lattice constant is set by the size of a water molecule $C_0 = 55.5$ M. Other choices for the reference concentration would result in a constant shift to the free energy which would not affect the results in a meaningful way. The measured and fitted free energy are compared in Fig. 3.3. The agreement is good with discrepancies ranging from 0.1-1.9 $k_B T$.

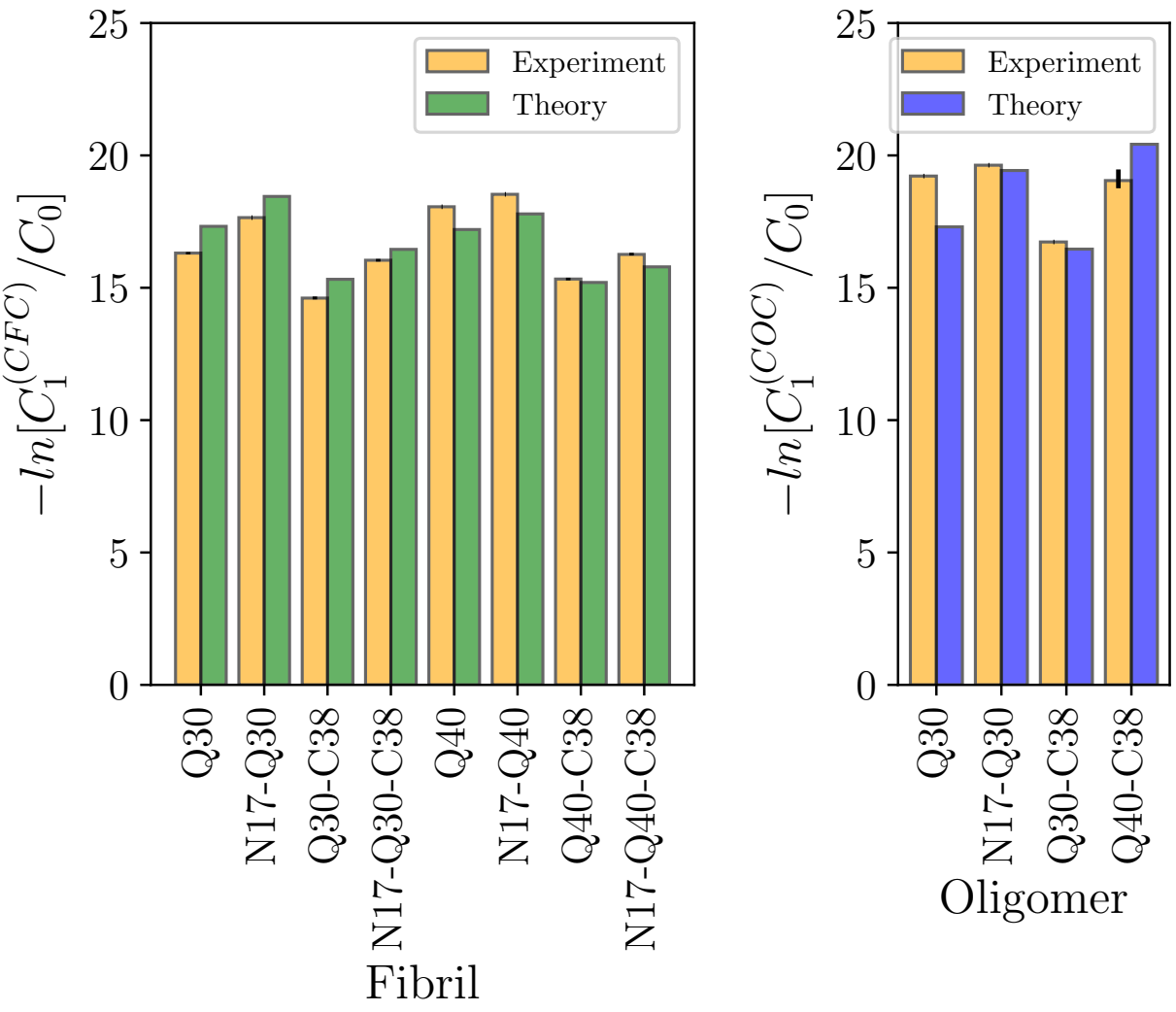


Figure 3.3: *Comparison between the theoretical model and experimentally measured critical concentrations. The model captures the effects of N17 and increasing polyQ length in promoting aggregation and the effect of C38 in inhibiting it.*

The parameter values, shown in Table 4.1, help to clarify the driving forces for aggregation. The free energy of β-sheet formation is almost 1 k_BT per amino acid, which is stronger than the ~ 0.5 k_BT found for Aβ and other amyloid forming molecules[28,32]. This is likely due to fact that polyQ is a homopolymer where all amino acids contribute equally, while other molecules have sequence heterogeneity as well as portions of the molecule in hairpins and disordered fragments that do not contribute to the stability. Interestingly, we find that the free energy of glutamine burial in the oligomer state, ε_Q, is even stronger than that

of β-sheet formation. This reflect the fact that Htt is one of the few molecules where the oligomer state has a lower critical concentration than the fibril state[13,15]. However, it should also be noted that the entropic penalty for elongating the peptide into a β-strand is included in ε_β, while the conformational entropy contributions to globule formation are separately calculated in Eq. 3.4.

The model allows us to understand several features of the aggregation behavior. From our results, the oligomer sizes are in the range of 3300-4900 monomers. Using a density of 1.3 g/cm^3 we estimate an oligomer diameter range of 30-50 nm, consistent with the 10-50 nm spheres measured by EM[13,31].

Fig. 3.4 shows the free energy of monomer collapse as a function of the polyQ length. In the absence of the N17 tail our model predicts that the free energy is zero for $\ell_Q = 17$, meaning that peptides with fewer glutamines will be found primarily in the expanded state while longer polyQ regions will favor the collapsed state. In the presence of N17 the crossover point is at $\ell_Q = 3$. While this is fewer glutamine residues than molecules without N17, the total peptide length is longer (20 vs. 17 amino acids) reflecting the fact that it takes more N17 residues to achieve the same desolvation energy of the glutamines.

Fig. 3.5 shows the oligomer free energy as a function of polyQ length in the presence and absence of the flanking regions. Increasing the peptide length, either by adding glutamines or N17 residues, results in larger oligomers because the longer molecules can more easily stretch to fill the interior of the oligomer. However the C38 region adds a repulsive energy that favors smaller oligomers.

Table 3.1: *Parameters obtained by model fitting*

Parameter	Value (k_BT)
ε_Q	-2.05
ε_N	-1.04
ε_β	-0.91
f_{N17}	-11.70

While the polyQ length has a roughly linear effect on the oligomer size, it has a much more dramatic effect on the critical concentration. Fig. 3.5B shows that the critical concentration

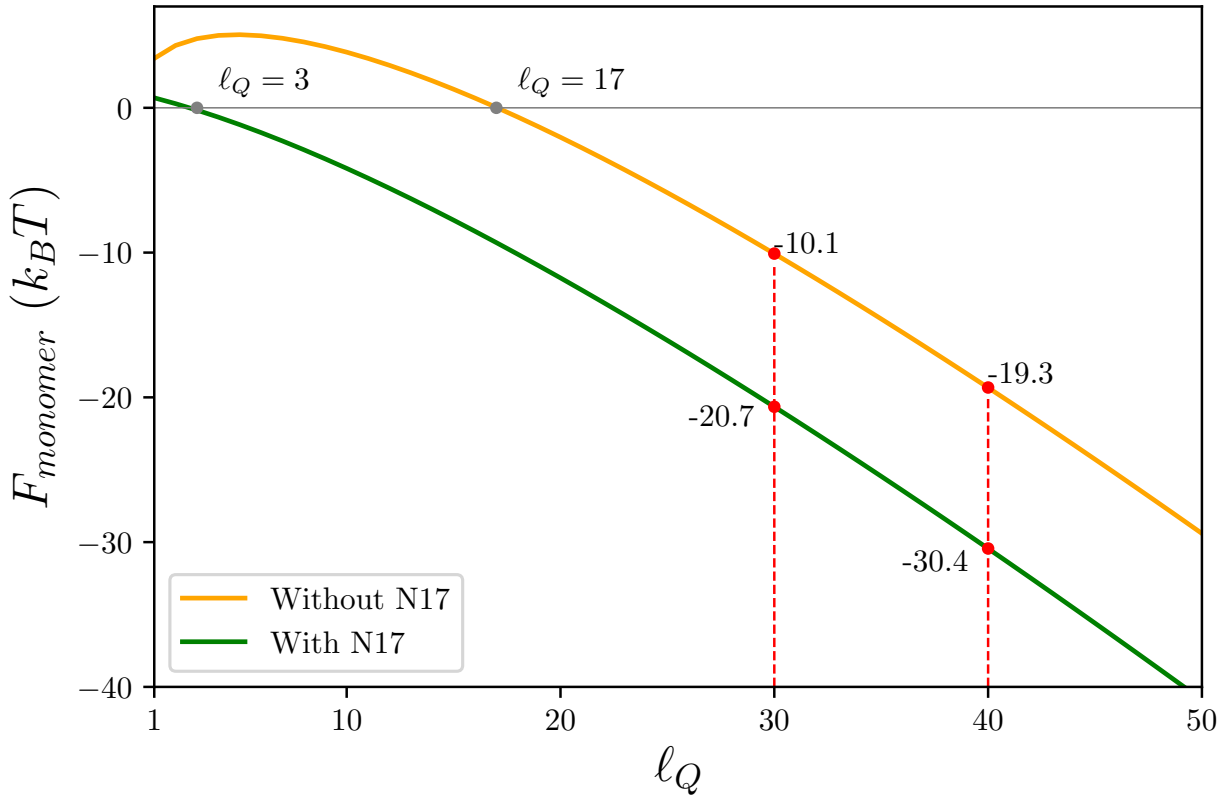


Figure 3.4: *Predicted free energy of monomer collapse for peptides with and without the N-terminal tail as a function of ℓ_Q. The results show that peptides with fewer glutamines will prefer the expanded state while longer glutamine peptides will favor the collapsed states. The presence of the N17 tail contributes to the collapse free energy, but less strongly than glutamine residues.*

scales exponentially with the polyQ length. When both flanking sequences are present the critical concentration varies from 11 nM for $\ell_Q = 40$ to 27000 nM for $\ell_Q = 20$. While this calculation does not account for important cellular factors like crowding, it is easy to speculate that this 10^3 factor could make a difference in the presence of toxic oligomers when the polyQ length increases above the threshold associated with disease.

3.3.2 Flanking sequences prevent large alignment errors.

The parameter ε_Q, obtained by fitting the fibril solubilities, can also be used to compute the frequency of registry errors in polyQ fibrils. NMR experiments have shown that single amino acids shifts occur at frequency of 25% ($R = +1$) and 15% ($R = -1$), with larger shifts

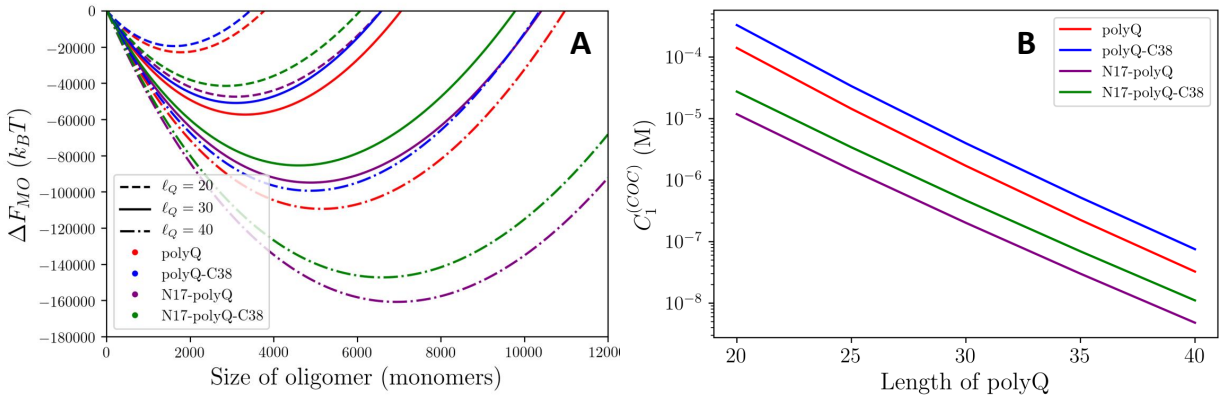


Figure 3.5: *(A) Predicted free energy of oligomer formation for ℓ_Q = 20, 30, and 40 in the presence and absence of N- and C-terminal tails. Increasing the length of the polyglutamine region or adding the N17 tail results in larger oligomers because the extra length more easily stretches to fill the oligomer core. However, adding the C38 tail adds a repulsive energy that favors smaller oligomers. (B) Changing the polyQ length has an exponential effect on the critical concentration for oligomer formation. The critical concentration drops by more than a factor of 10^3 upon changing the ℓ_Q from 20 to 40.*

occurring below the detection limit[16]. In comparison, a simple version of our model that does not account for the N- and C-terminal tails (Eq. 3.13 with $\varepsilon_{\mathrm{NQ}} = \varepsilon_{\mathrm{CQ}} = 0$) yields registry errors of 17% for $R = \pm 1$ and 7% for $R = \pm 2$ (blue line, Fig. 3.6). From this we make two observations. First, even the weak $e^{-\varepsilon_\beta / k_B T}$ penalty for registry shifts is sufficient to prevent registry errors for most molecules. Secondly, the presence of the N- and C-terminal tails have the dual effect of suppressing shifts of $|R| > 1$ and breaking the symmetry between the shift directions. An inspection of the sequence readily suggests mechanisms by which this may occur. The C-terminus of the polyQ region is connected to a stretch of 28 prolines. Prolines will be poorly tolerated in the cross-β core due to their lack of a backbone H-bond donor and their inability to adopt the extended β-sheet conformation. To account for this we add a free energy penalty for negative registry shifts. Fig. 3.6 shows that $\varepsilon_{\mathrm{CQ}}$ values between 0.25 and 1.0 k_BT have the expected effect of shifting the alignment distribution closer to the experimental observation. However, they also raise the probability of $R = +2$ shifts near the 10% level that is experimentally observable. This discrepancy is easily resolved by an inspection of the N-terminal tail, which has a sequence MATLEKLMKAFESLKSF, with the serine and phenylalanine incorporated in the cross-β core[16]. This means that positive

registry shifts would move the lysine into the cross-β core. While the long sidechain could presumably allow for partial solvation of the charge for a $R = +1$ shift, larger registry shifts would require a total desolvation of the charge, thereby incurring a large free energy penalty. If we exclude registry shifts larger than $R = +1$, the predicted distribution of registries is in nearly perfect agreement with experiment (Fig. 3.6, inset).

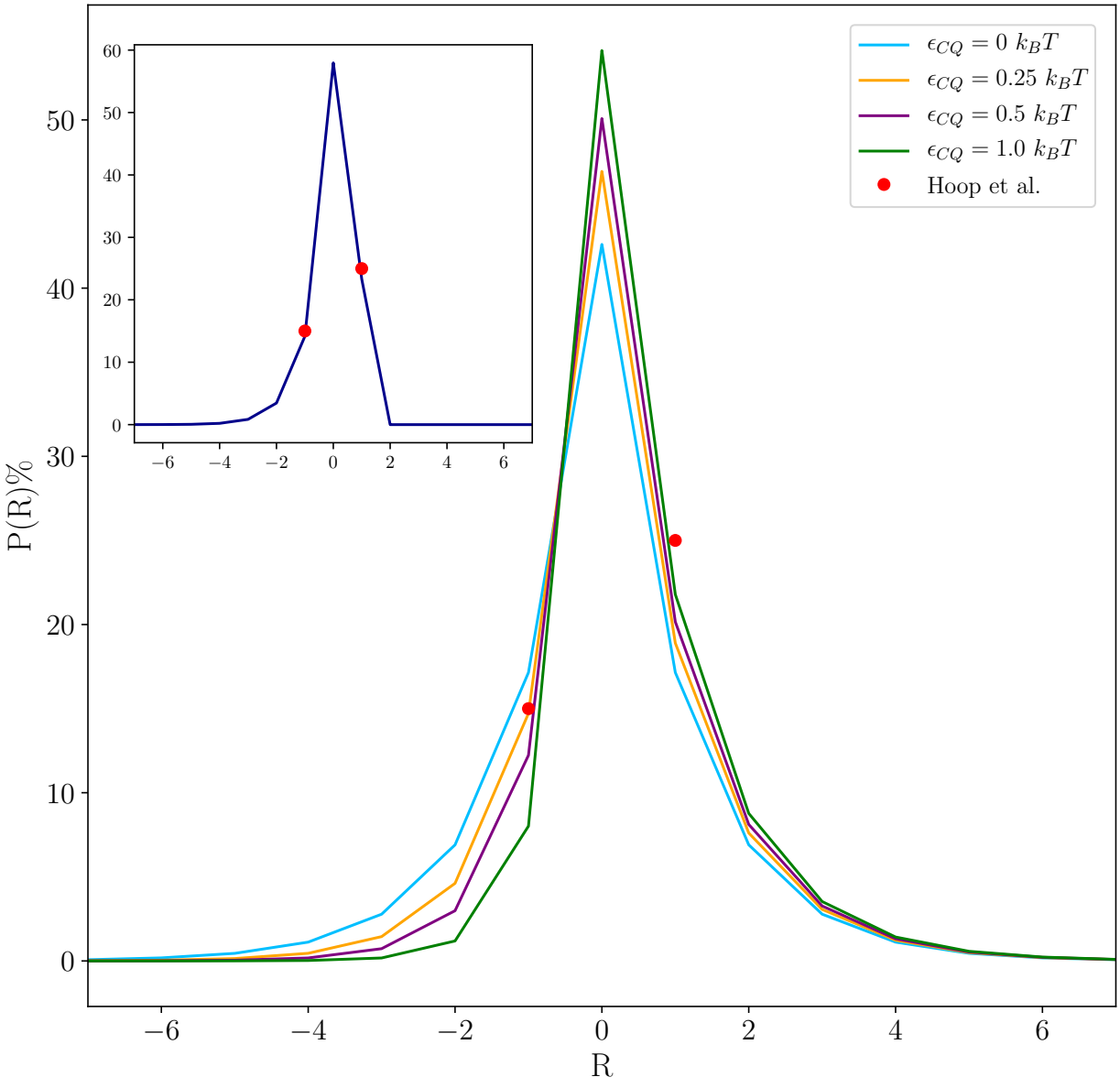


Figure 3.6: *Probabilities of mis-aligned molecules within an Htt fibril as a function of the alignment registry R and* $\varepsilon_{\mathrm{CQ}}$ *(for* $\varepsilon_\beta = -0.91\ k_BT$*). The inset shows alignment probabilities for* $\varepsilon_{\mathrm{CQ}} = 0.5\ k_BT$ *with an additional constraint preventing states with* $R > 1$*, since this would lead to the burial of the lysine charge.*

3.4 Discussion

The micelle-like oligomers described by our theory contrast with the highly ordered β-barrel oligomers that have been reported for other amyloid forming molecules[33,34]. It is difficult to imagine a low complexity sequence like huntingtin adopting such an ordered state. But it is worth asking whether the micellar structure of huntingtin might also be formed by other molecules. Supporting this view is the fact that the A11 antibody, which specifically recognizes amyloid oligomers, was developed by forcing Aβ to form a micelle-like structure[35]. In addition, hydrophobicity correlates strongly with aggregation propensity[36,37], suggesting that most amyloid-forming molecules will contain a stretch of hydrophobic amino acids sufficiently long to form a polymer micelle. This implies that the amyloid phase diagram often contains both ordered and disordered oligomers in addition to the fibril state. The ordered oligomer could be added to our model with a free energy of the form of Eq. 3.9. Notably, due to the smaller size of ordered oligomers (on the order of 4-20 molecules, compared to 10^3 for disordered oligomers), the ordered species will show a softer, power law concentration dependent onset compared to the steep, phase transition-like onset seen with large oligomers and fibrils[28].

Our model also provides insights into the mechanism of fibril formation. Specifically, there is the question of whether the highly ordered fibrils reported from NMR or X-ray studies are typical or an artifact of structural methods that work best with ordered systems. Our results show that even for a homopolymer, the binding energy is sufficient to align almost half of the molecules. Also, consistent with previous work, only minor perturbations from a uniform sequence are necessary to generate highly ordered fibrils[30]. In the equilibrium analysis employed in this work, the fraction of alignment defects is independent of peptide length. However, the kinetic search over alignments scales exponentially with the peptide length, meaning longer peptides will be more easily trapped in non-equilibrium states under conditions of rapid aggregation[28,30].

In conclusion, we have shown that block copolymer model is able to explain many features of oligomer and fibril formation in huntingtin. These findings may also have broader

implications for other amyloid forming systems.

3.5 APPENDIX

Calculation of surface constant

The bulk energy term of Eq. 3.1 accounts for the desolvation of every amino acid in the globule. However, amino acids on the surface of the globule will only be partially desolvated. To estimate the surface correction to the desolvation energy, we assume that amino acids on the surface only get half the desolvation energy. This gives

$$\frac{F_{\text{surface}}}{k_B T} = \left(-\frac{1}{2}\varepsilon_G\right) N_{\text{surface}}$$

where N_{surface} is the number of amino acids on the surface of the globule.

To calculate the N_{surface}, we relate the radius of the globule to the number of molecules

$$V = \frac{4}{3}\pi R^3 = NLa^3 \tag{3.14}$$

where a^3 is the volume of an amino acid. The number of residues on the surface is

$$N_{\text{surface}} = \frac{4\pi R^2}{a^2} = 4\pi \left(\frac{3}{4\pi}\right)^{2/3} (NL)^{2/3}$$

The surface term is therefore

$$\frac{F_{\text{surface}}}{k_B T} = -\left(\frac{1}{2}\varepsilon_G\right) 4\pi \left(\frac{3}{4\pi}\right)^{2/3} N_{\text{total}}^{2/3} \tag{3.15}$$

$$\frac{F_{\text{surface}}}{k_B T} = -2.4\varepsilon_G N_{\text{total}}^{2/3} = A_\gamma \varepsilon_G N_{\text{total}}^{2/3} \tag{3.16}$$

References

[1] C. A. Ross and M. A. Poirier, Nature medicine **10**, S10 (2004).

[2] A. Aguzzi and T. O'connor, Nature reviews Drug discovery **9**, 237 (2010).

[3] Y. E. Kim, F. Hosp, F. Frottin, H. Ge, M. Mann, M. Hayer-Hartl, and F. U. Hartl, Molecular cell **63**, 951 (2016).

[4] L. Zhang and J. D. Schmit, Israel journal of chemistry **57**, 738 (2017).

[5] L. Zhang and J. D. Schmit, Physical Review E **93**, 060401 (2016).

[6] T. Miti, M. Mulaj, J. D. Schmit, and M. Muschol, Biomacromolecules **16**, 326 (2014).

[7] T. P. Knowles, D. A. White, A. R. Abate, J. J. Agresti, S. I. Cohen, R. A. Sperling, E. J. De Genst, C. M. Dobson, and D. A. Weitz, Proceedings of the National Academy of Sciences **108**, 14746 (2011).

[8] S. Auer, P. Ricchiuto, and D. Kashchiev, Journal of molecular biology **422**, 723 (2012).

[9] S. Auer, C. M. Dobson, and M. Vendruscolo, HFSP journal **1**, 137 (2007).

[10] S. Auer, C. M. Dobson, M. Vendruscolo, and A. Maritan, Physical review letters **101**, 258101 (2008).

[11] D. R. Langbehn, M. R. Hayden, J. S. Paulsen, and P.-H. I. of the Huntington Study Group, American Journal of Medical Genetics Part B: Neuropsychiatric Genetics **153**, 397 (2010).

[12] F. O. Walker, The Lancet **369**, 218 (2007).

[13] S. L. Crick, K. M. Ruff, K. Garai, C. Frieden, and R. V. Pappu, Proceedings of the National Academy of Sciences **110**, 20075 (2013).

[14] J. B. Warner IV, K. M. Ruff, P. S. Tan, E. A. Lemke, R. V. Pappu, and H. A. Lashuel, Journal of the American Chemical Society **139**, 14456 (2017).

[15] S. L. Crick, M. Jayaraman, C. Frieden, R. Wetzel, and R. V. Pappu, Proceedings of the National Academy of Sciences **103**, 16764 (2006).

[16] C. L. Hoop, H.-K. Lin, K. Kar, Z. Hou, M. A. Poirier, R. Wetzel, and P. C. Van Der Wel, Biochemistry **53**, 6653 (2014).

[17] E. A. Newcombe, K. M. Ruff, A. Sethi, A. R. Ormsby, Y. M. Ramdzan, A. Fox, A. W. Purcell, P. R. Gooley, R. V. Pappu, and D. M. Hatters, Journal of molecular biology **430**, 1442 (2018).

[18] A. Vitalis, X. Wang, and R. V. Pappu, Biophysical journal **93**, 1923 (2007).

[19] K. M. Ruff, T. S. Harmon, and R. V. Pappu, The Journal of chemical physics **143**, 12B607_1 (2015).

[20] L. Leibler, H. Orland, and J. C. Wheeler, The Journal of chemical physics **79**, 3550 (1983).

[21] S. M. Hoseinpoor, N. Nikoofard, and M. Zahedifar, Journal of Statistical Physics **163**, 593 (2016).

[22] V. R. Singh and L. J. Lapidus, The Journal of Physical Chemistry B **112**, 13172 (2008).

[23] M. Rubinstein, R. H. Colby, et al., *Polymer physics*, vol. 23 (Oxford university press New York, 2003).

[24] M. Chen and P. G. Wolynes, Proceedings of the National Academy of Sciences **114**, 4406 (2017).

[25] D. Punihaole, R. J. Workman, Z. Hong, J. D. Madura, and S. A. Asher, The Journal of Physical Chemistry B **120**, 3012 (2016).

[26] G. Hoffner and P. Djian, Brain sciences **4**, 91 (2014).

[27] C. F. Lee et al., Physical Review E **80**, 031922 (2009).

[28] J. D. Schmit, K. Ghosh, and K. Dill, Biophysical journal **100**, 450 (2011).

[29] R. Tycko, Neuron **86**, 632 (2015).

[30] C. Huang, E. Ghanati, and J. D. Schmit, The Journal of Physical Chemistry B **122**, 5567 (2018).

[31] A. E. Posey, K. M. Ruff, T. S. Harmon, S. L. Crick, A. Li, M. I. Diamond, and R. V. Pappu, Journal of Biological Chemistry **293**, 3734 (2018).

[32] A. J. Baldwin, T. P. Knowles, G. G. Tartaglia, A. W. Fitzpatrick, G. L. Devlin, S. L. Shammas, C. A. Waudby, M. F. Mossuto, S. Meehan, S. L. Gras, et al., Journal of the American Chemical Society **133**, 14160 (2011).

[33] A. Laganowsky, C. Liu, M. R. Sawaya, J. P. Whitelegge, J. Park, M. Zhao, A. Pensalfini, A. B. Soriaga, M. Landau, P. K. Teng, et al., Science **335**, 1228 (2012).

[34] M. Serra-Batiste, M. Ninot-Pedrosa, M. Bayoumi, M. Gairí, G. Maglia, and N. Carulla, Proceedings of the National Academy of Sciences **113**, 10866 (2016).

[35] R. Kayed, E. Head, J. L. Thompson, T. M. McIntire, S. C. Milton, C. W. Cotman, and C. G. Glabe, Science **300**, 486 (2003).

[36] F. Chiti, M. Stefani, N. Taddei, G. Ramponi, and C. M. Dobson, Nature **424**, 805 (2003).

[37] S. Yagi, S. Akanuma, and A. Yamagishi, Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics **1844**, 553 (2014).

Chapter 4

Nonspecific Binding Assists DNA to Perform the Alignment Search During Hybridization

DNA hybridization is the fundamental key of life. It does not only play an important role in propagation of information through generations, but is also a very popular tool in many fields such as biology, biotechnology, nanotechnology. The kinetics of DNA hybridization have been difficult to understand because it is hard for experiments and simulations to reconcile high resolution and low resolution results. In this study we developed a theory of medium resolution which describes the DNA hybridization kinetics via three stages including diffusion, residence and zipping. The model reveals that nonspecific binding in the residence stage helps DNA strands search the alignment state and accelerates the hybridization rate. We also show that the DNA hybridization rate is affected by factors including sequence dependence, intra-molecular structure, temperature, and configurations.

4.1 Introduction

The capability of cells to store, retrieve and transfer genetic information plays a crucial role in making and maintaining a living organism. Cells need DNA as the carrier to precisely propagate genetic instructions through DNA replication. DNA hybridization is essential to this process; it enables cells to pass genes on to their descendants[1].

DNA hybridization also plays a fundamental role in biology, biotechnology, and nanotechnology[2–4]. The ability of DNA to hybridize is essential for techniques like PCR[5], DNA nanostructures (DNA origami)[6], and for the diagnosis of diseases like HPV, HIV[7], cancer[8], genotyping and other genomic diagnostics[5]. Therefore it is important to understand the contributions of both thermodynamics and kinetics of this phenomenon. The thermodynamics of DNA hybridization has been studied extensively, however many interesting and important questions about the kinetics of this process remain unanswered.

Hybridization hinges on the ability of base pairs to recognize specific sequences. This comes from the fact that there are two Watson-Crick-Franklin (WCF) base pairs: A-T and G-C, while other pairs attract only weakly. However, this also means that there is a 1 in 4 chance that a random base pair will be a match. Therefore, there is a non-negligible probability that random portions of DNA will be able to hybridize. This means that in addition to the native, in-register base pairs, DNA is more likely to form non-native base paired structures. We divide these non-native interactions into two categories. The first is intra-molecular base pairs which result in "loop and stem structures". These structures are functionally important in RNA enzymes, but not for DNA. The second is inter-molecular but non-native base pairs, where the molecules find out-of-register alignment that form regions of WCF base pairs. One big question here is how these types of interactions affect hybridization rates.

Precise control of the kinetics of DNA hybridization at the molecular level is vital to the processing of gene replication and regulation. Hybridization kinetics has been observed in recent experiments. Cisse et al. used single molecule fluorescence to precisely quantify the reaction rates of melting (k_{off}) and annealing (k_{on}) between two short DNA strands as

a function of mismatch position. They suggest that seven contiguous pairs are needed for rapid hybridization of DNA[9]. However, the microscopic mechanism was not resolved and the findings are also limited to short sequences. There are many other experimental studies of DNA hybridization in different lengths of DNA. Yazawa et al. used total internal reflection fluorescence microscopy (TIRFM) and a quartz-crystal microbalance (QCM) to observe the DNA hybridization of 8 and 12 nucleotides in DNA sequences. DNA hybridization depends on the length and sequence which results in multiple behaviors. While the 8 base-pair DNA has a single binding mode, the 12 base-pair DNA has at least two different binding modes. The 12 homogeneous base-pair DNA even shows multiple binding modes[10]. Zhang et al. carried out 210 fluorescence kinetics experiments of 100 pairs of DNA strands which have the medium length of 36 nucleotides at a few different temperatures. Then they developed an algorithm to predict the hybridization rate of new DNA sequences[11]. Wetmur et al. studied the kilobase-pair DNA molecules[12]. The studies show that the DNA hybridization rate of the medium DNA length is sequence dependent while the long DNA molecules show less sequence dependence. Moreover, experimental results do not give microscopic mechanisms about the DNA hybridization. It is still very hard to understand the behaviors of DNA molecules based on the experiments.

To solve that problem, molecular simulations have been used to explore the nature of DNA hybridization as well. A coarse-grained model shows that two DNA strands go through a complicated set of intermediate states in which they can form misaligned bonds, then search for alignment bonds via "inchworm" or "pseudoknot" pathways before reaching fully zipped state. "Inchworm" is a pathway in which the DNA strands in a mis-aligned state fluctuate and form an in-register bond resulting in a bulge loop, while "pseudoknot" is a pathway in which the DNA strands have a few mis-aligned bonds forming at their two opposite ends, and their tails can fluctuate and bind to form a few aligned base pairs, resulting in a pseudoknot. Then more $R = 0$ base pairs form, more $R \neq 0$ base pairs break until the system obtains the fully zipped state. The pathways such as "inchworm", "pseudoknot" help to accelerate the hybridization rate. DNA hybridization depends on the zippering, internal displacement and sequence itself[2]. Two DNA strands interact with each other with the assistance of

the nucleation step; a few base pairs form in the in-register state before fully hybridizing. Structured strands such as hairpin have to unfold themselves so they overcome more energy barriers than unstructured ones[3]. Molecular simulations give a good view about the kinetic mechanisms of DNA hybridization by showing multiple mechanisms in DNA hybridization but it is not clear about how they apply to short sequences because if the length is too short, it cannot show the intra-molecular structure in DNA. Therefore, we need a theory which focuses on the assembly process in DNA hybridization.

A theoretical model by Wetmur et al. is developed to reveal the DNA hybridization kinetics. They propose a nucleation and zipping model which assumes that the reaction mechanism of two strands is related to nucleation sites. The theory reveals that in DNA hybridization, after forming one or a few in-register base pairs along two strands, the zipping must occur very fast with the assistance of a large number of nucleation sites[12]. However, the model cannot show clearly what "nucleation" is and how internal structure fits in.

In this study, we develop a theory of DNA hybridization kinetics based on the biopolymer templating problem. Similarly to protein aggregation theory in previous work[13–16] we consider two DNA strands interact with each other by forming and breaking H-bonds. The number of H-bonds made between the incoming molecule (probe) and the templating molecule (target) is described by using a 1D random-walk model[14]. Two DNA strands spend most of the time searching randomly over mis-alignment states before ending up in the perfect alignment state. We find that the nonspecific binding can assist DNA to perform the alignment search during hybridization. This result is opposite to protein aggregation which shows that mis-alignment states can slow down the aggregation rate because incoming molecules in mis-alignment states need to break all H-bonds and start over to contact the existing fibril. This process repeats many times until the incoming molecules have at least a perfect alignment bond[13]. Our theory can be considered as an intermediate resolution model for medium length sequences. DNA hybridization kinetics depend on the sequence and account for internal folding. The model plays a role as a bridge connecting high resolution kinetics shown in simulation to low resolution ensemble measurements which account for large-scale structures. We also propose a mechanism applicable to all scales that explains things like

nucleation.

4.2 Model

4.2.1 Random search affects the DNA hybridization rate

We use the registry variable, R, to describe the alignment of two DNA strands[13,15]. R is the number of base pairs that the top strand is shifted relative to the bottom strand so that $R = 0$ indicates a perfect alignment and $R \neq 0$ expresses a mis-alignment (Fig. 4.1).

Base pairing has many effects on hybridization. There are three types of base pairing:

i) *In-register states* ($R = 0$) allow all bases to form Watson-Crick-Franklin (WCF) pairs[17]. These base pairs are very stable and lock the DNA molecules into place. However, these states contribute minimally to the kinetics of hybridization because most of the hybridization time is spent in non-native traps.

ii) *Mis-registered states* ($R \neq 0$) are less stable than in-register states because most bases will be unable to form WCF pairs. However, because there are only four bases, most alignments will allow for the formation of a few WCF base pairs by random chance. These result in kinetic traps which hold the DNA molecules together (Fig. 4.1). These states are very important to hybridization kinetics since most collisions between DNA molecules will be out of register.

iii) *Intra-molecular contacts* occur when the DNA molecule folds to form WCF base pairs with itself. Intra-molecular base-pairing is usually imperfect resulting in single stranded loops separating base-paired regions. These contacts affect the DNA hybridization rate because self-bonds must be broken before hybridizing.

To disentangle the effects of mis-registered and intra-molecular interactions, we consider DNA molecules that are either "unstructured" state in which the DNA strand does not self-hybridize or has a single self-complementary region resulting in a single loop and single stem (Fig. 4.2). We identify intra-molecular structure using the NUPACK software[18] and classify a sequence as unstructured if looped states have a repulsive free energy.

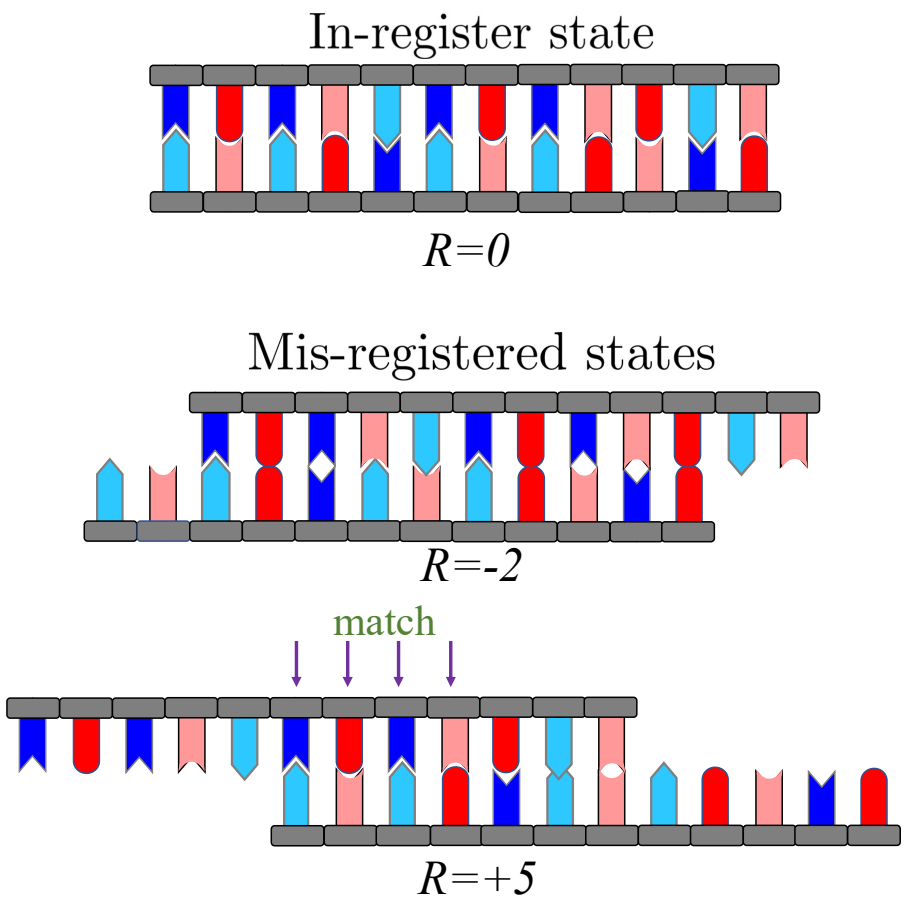


Figure 4.1: *Cartoon representation of the in-register and mis-registered states. In the in-register state, all base pairs follow the Watson-Crick-Franklin (WCF) rule in which A (dark blue) always pairs with T (light blue), G (red) always joins C (light red). In contrast, in the mis-registered states, most base pairs are mismatches. However, because there are only four bases, many alignments will allow for the formation of a few WCF base pairs by random chance, resulting in a kinetic trap. The mis-alignment state at $R = +5$ shows a kinetic trap of four WCF base pairs.*

4.2.2 Kinetics of DNA hybridization is modeled in three stages

We consider DNA hybridization to occur in three stages: the diffusion stage, the residence stage, and the zipping stage (Fig. 4.3). The *diffusion stage* is defined as the time for two DNA strands to diffuse close enough to make the first H-bond. Since WCF H-bonds are short range, the large majority of the time this first interaction will be out-of-register. Regardless of the registry, the diffusion stage ends after the first contact and the strands begin the

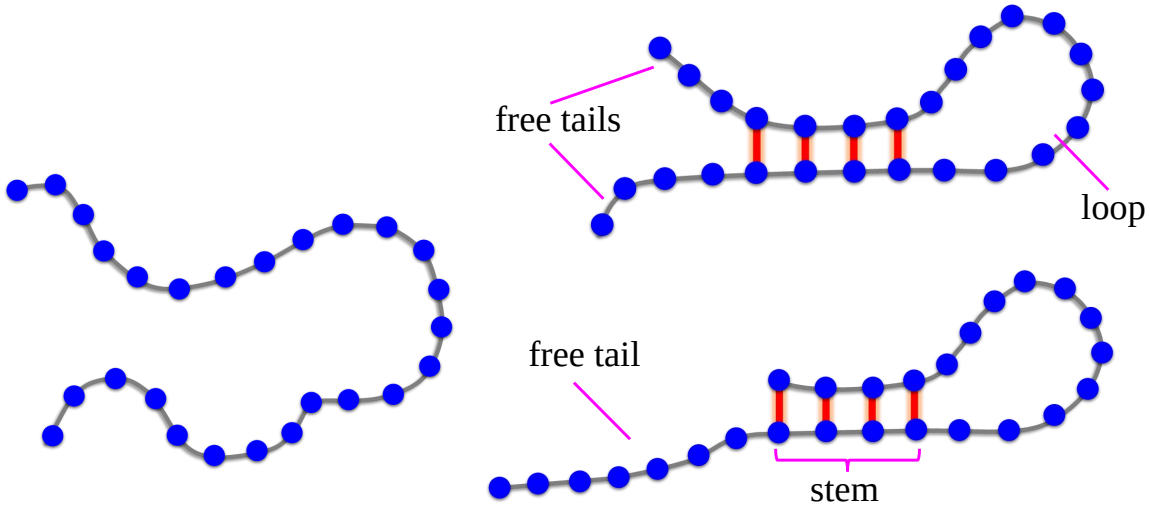


Figure 4.2: *Cartoon representation of the structure of single DNA strands. (a) Unstructured strand is a free molecule which does not self-hybridize resulting in a loop and stem regions. (b) Stem-loop structures occur when a sequence has a single self-complementary region which results in the formation of a double stranded "stem" separating a single stranded loop and one or two free tails.*

residence stage. In the residence stage, the DNA strands search for the in-register state ($R = 0$). In the meantime, they may form or break bonds around the initial contact unless these initial bonds contain WCF pairs, these interactions are usually unstable and short lived. If the interactions at the initial contact point are broken before the in-register state is found, the molecule is considered to return to the diffusion state.

During the residence stage, we assume that unstructured regions of the molecules can interact in search of in-register interactions. If two DNA strands find and form an in-register bond, the *zipping stage* starts. The two DNA strands can form or break in-register bonds in this stage. However, forming in-register bonds is more favorable than breaking because at $R = 0$, all bonds result in WCF pairs. The zipping stage ends when either the molecules are fully hybridized, or all in-register bonds are broken.

In the rare event that two DNA strands make initial contact at $R = 0$, they skip the

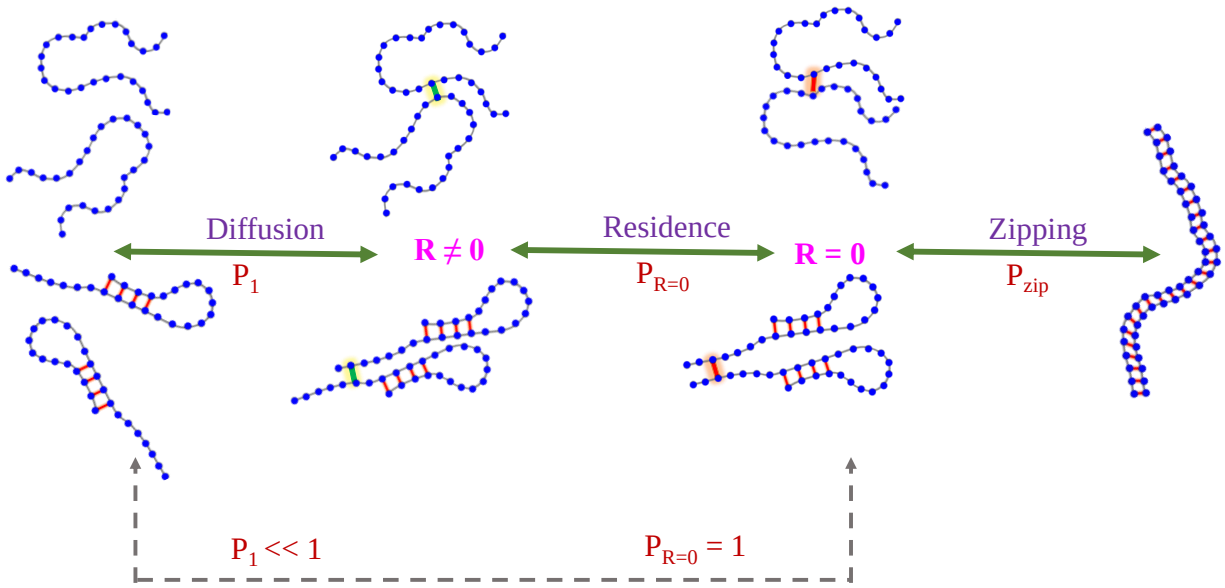


Figure 4.3: *Diagram of the three stages of the DNA hybridization kinetics shown for both unstructured (top) and single loop (middle) molecules. DNA strands go through the diffusion stage which usually results in an initial H-bond at $R \neq 0$, the residence stage which results in the first $R = 0$ H-bond, and the zipping state to obtain a full zipping state. In rare cases the DNA may form an initial bond at $R = 0$ after the diffusion time. In these events the residence stage is skipped (bottom path) and the DNA strands go to the zipping stage immediately after the diffusion stage.*

residence stage and start the zipping stage right after the diffusion stage.

4.2.3 Methods

To account for the combined effect of the three stages we write down the time required for N inter-molecular collisions.

$$\begin{aligned} T_{tot}(N) &= \sum_R [N \cdot (1 - P_1(R)) \cdot \tau_d \\ &+ N \cdot P_1(R) \cdot (1 - P_{R=0}(R)) \cdot (\tau_d + \tau_r) \\ &+ N \cdot P_1(R) \cdot P_{R=0}(R) \cdot (1 - P_{zip}) \cdot (\tau_d + \tau_r + \tau_z) \\ &+ N \cdot P_1(R) \cdot P_{R=0}(R) \cdot P_{zip} \cdot (\tau_d + \tau_r + \tau_z)] \end{aligned} \tag{4.1}$$

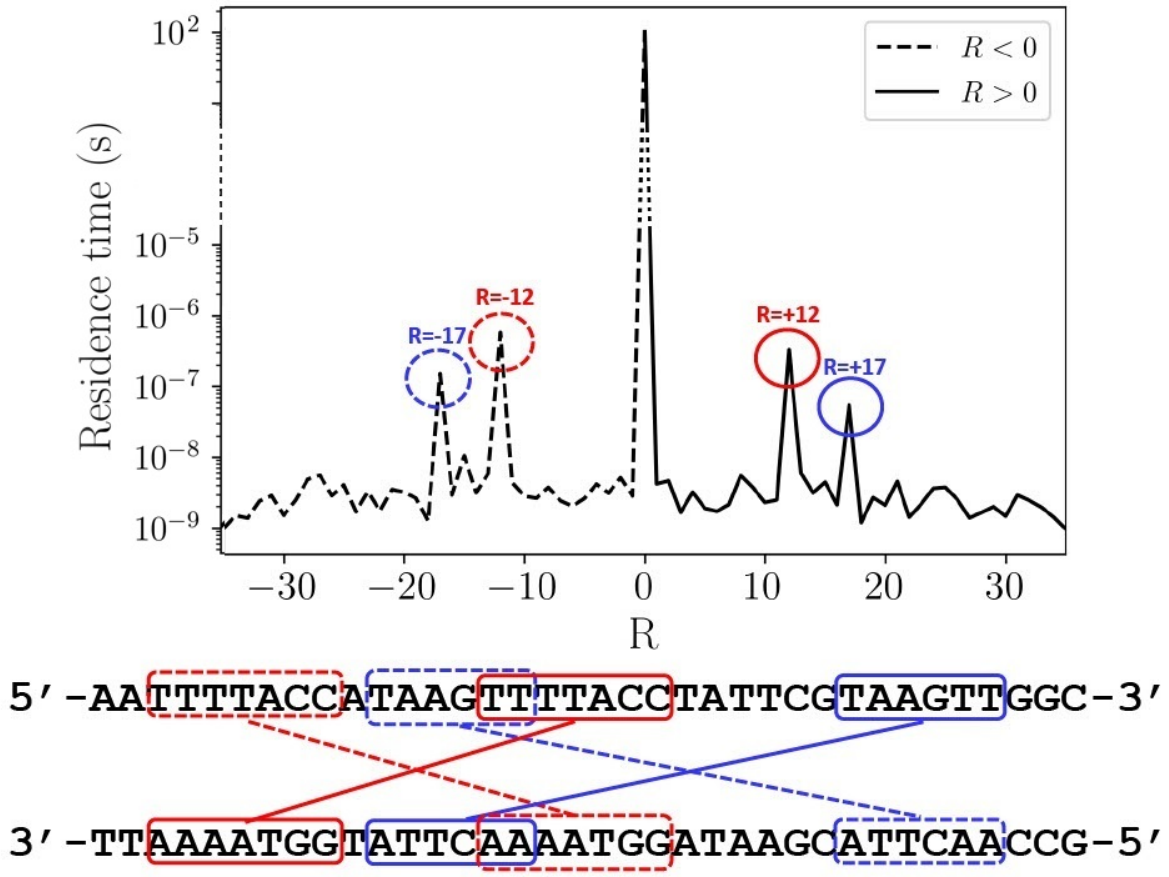


Figure 4.4: *Comparison of the residence times of each registry R of an unstructured sequence (S40) at 55°C. In the mis-registered states, there are four peaks at registries $R = -17$ (blue dash), $R = -12$ (red dash), $R = +17$ (blue solid), and $R = +12$ (red solid). Those peaks represent kinetic traps which arise due to mis-registered WCF base pairs shown in the blocks below the plot with the same color codes.*

The first term represents failed collisions in which two strands diffusing very closely are unable to bring bases into contact. Here $P_1(R)$ is the probability of forming a first bond at registry R after an inter-molecular collision. The second term accounts for molecules that form inter-molecular bonds but do not form any in-register bonds. $P_{R=0}(R)$ is the probability of forming a $R = 0$ H-bond after forming at least one H-bond in the registry R. The third term describes events in which in-register bonds form but fail to reach the fully zipped state. P_{zip} is the probability for fully zipping after forming at least one $R = 0$ base pair. The last term is successful collisions where two strands are able to arrive at the fully zipped state.

τ_d, τ_r, τ_z are the diffusion time, residence time and zipping time, respectively.

The time for N inter-molecular collisions can be rewritten as

$$\begin{aligned} T_{tot}(N) &= N\sum_R[(1-P_1(R))\cdot\tau_d \\ &+ P_1(R)\cdot(1-P_{R=0}(R))\cdot(\tau_d+\tau_r) \\ &+ P_1(R)\cdot P_{R=0}(R)\cdot(1-P_{zip})\cdot(\tau_d+\tau_r+\tau_z) \\ &+ P_1(R)\cdot P_{R=0}(R)\cdot P_{zip}\cdot(\tau_d+\tau_r+\tau_z)] \\ &= N\cdot T_{tot}(1) \end{aligned} \tag{4.2}$$

where $T_{tot}(1)$ is the time required for one inter-molecular collision.

After N collisions there will be $\sum_R[N\cdot P_1(R)\cdot P_{R=0}(R)\cdot P_{zip}]$ successful events, so the average time per hybridization event is

$$t_{hyb} = \frac{T_{tot}(N)}{\sum_R[N\cdot P_1(R)\cdot P_{R=0}(R)\cdot P_{zip}]} \tag{4.3}$$

Therefore the hybridization rate is

$$rate = \frac{\sum_R[P_1(R)\cdot P_{R=0}(R)\cdot P_{zip}]}{T_{tot}(1)} \tag{4.4}$$

We now turn to calculating the times and three probabilities in this expression.

Time

The next step is to calculate the times and probabilities appearing in Eq. 4.4.

Diffusion time is defined as $\tau_d = 1/k_d$ where k_d is approximated by the Smoluchowski

formula for absorbing sphere:

$$\begin{aligned} \frac{k_d}{c} &= 4\pi a D \\ &= 4\pi a \frac{k_B T}{6\pi\eta a} \\ &= \frac{2k_B T}{3\eta} \end{aligned} \tag{4.5}$$

where D is the diffusion constant of the strands, c is concentration far from the surface, η is the viscosity of the solvent (viscosity of water at 55^oC is $5 \times 10^{-4}\mathrm{N}\cdot\mathrm{s/m}^2$ and at 37^oC is $7 \times 10^{-4}\mathrm{N}\cdot\mathrm{s/m}^2$)[19]. Notably, the lengths cancel because the diffusing particle and the target particle are identical. The units of k_d/c are $\mathrm{M}^{-1}\mathrm{s}^{-1}$. Using this formula, the diffusion rate at 55^oC is estimated as $3.6 \times 10^9\mathrm{M}^{-1}\mathrm{s}^{-1}$, and the diffusion rate at 37^oC is estimated as $2.5 \times 10^9\mathrm{M}^{-1}\mathrm{s}^{-1}$.

The *residence time* τ_r is the time two molecules remain in contact. This is computed by the first passage time for the system to go from a state with one base pair to zero base pairs. We compute those first passage times using Gillespie simulations of the bond formation and breakage dynamics[15] (See APPENDIX). We define k_+, k_- as the rate of forming and breaking bonds respectively. We assume that bond formation is independent of sequence and set $k_+ = 10^9 s^{-1}$. In contrast, bond breakage is limited by the bond breakage energy so that $k_- = k_+ e^{\Delta G}$, where ΔG is the base pair binding energy. This bond energy is computed with the nearest neighbor free energies of Santa Lucia et al.[5,20–25].

The *zipping time* is simulated using the Gillespie algorithm which is similar to the residence time simulation. Unlike the residence time simulations, the zipping simulation only considers in-register states. The zipping time is counted from the formation of the first base pair until the molecules reach the fully bonded state. In the zipping simulation the two strands need to satisfy the boundary condition that the bonds can form and break randomly but the number of base pairs cannot reach zero before obtaining the fully zipped state.

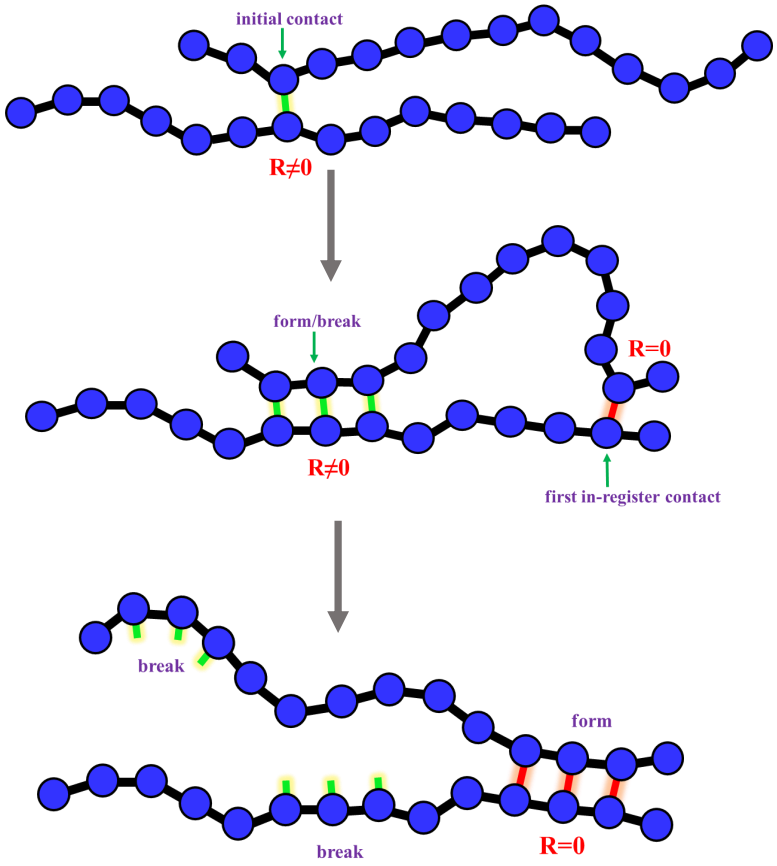


Figure 4.5: *Cartoon representation of alignment searches of DNA strands. After forming the initial contact at $R \neq 0$, the DNA strands may form or break bonds around the initial contact. In the meantime, the free regions around the initial contact fluctuate to search for $R = 0$ positions and have the first in-register contact. They may form or break WCF bonds around that in-register bond, but formation is more favorable. In contrast, mis-aligned bonds are less stable and have a relatively short lifetime.*

Sticking probability P_1

The sticking probability, $P_1(R)$, is the probability that a random collision between DNA molecules results in the formation of a base pair. For this to occur we require that the bases at the site of the collision are not previously engaged in intra-molecular base pairs. For stem-loop molecules this means the contact must be between free tails (we neglect binding in the loops). This limits registries that are possible because registries $|R|$ greater than the length of the tail cannot form base pairs. Therefore,

$$P_1(R) = C_1 \frac{\ell - |R|}{L^2} \tag{4.6}$$

where $\ell = L$ for unstructured sequences, $\ell = \ell_1$ or ℓ_2 for stem-loop sequences with ℓ_1 and ℓ_2 are the length of free tails. Here L^2 is the number of possible random collisions in all registries, and $\ell - |R|$ is the number of possible random collisions of each registry R. C_1 is a sequence independent geometric factor that accounts for collision in orientation unfavorable for base pair formation (i.e, between phosphate backbones). Molecules that have extensive intra-molecular bonding have low values of $P_1(R)$ because the two strands can only interact at the free tails.

Probability $P_{R=0}$

If the molecules are held together by non-native base pairs in the residence stage, the unbound positions of the molecules can search for in-register base pairs. While the molecules are held together by mis-registered base pairs, the free tails are free to fluctuate. We expect that the tails will come into contact on a time scale comparable to the Zimm time, τ_{Zimm}, which describes the dynamic of the DNA strands in solvent accounting for the hydrodynamic interactions[26,27]. Each of these contacts provides an opportunity for the molecule to find in-register base pairs. The probability of success depends on both the amount of time the molecules are held together, which determine the number of attempts, and the length of the free tails, which determine the probability a given attempt is successful. The number of attempts over the residence time is $\tau_r(R)/\tau_{Zimm}$. The probability that a single attempt is successful is

$$P_{succeed}(R) = \frac{(\ell - 1) - |R|}{(\ell - 1)^2} \tag{4.7}$$

where $(\ell-1)-|R|)$ is the number of possible in-register base pairs and $(\ell-1)^2$ is the total possible base pairs which the molecules can search.

If $\tau_r(R)/\tau_{Zimm}$ is small $P_{R=0}(R)$ can be approximated by $\frac{\tau_r(R)}{\tau_{Zimm}}\frac{(\ell-1)-|R|}{(\ell-1)^2}$. However, if $\tau_r(R)$ is large this expression can exceed unity. In this case we cannot neglect the probability of multiple successful attempts in $\tau_r(R)$. The desired quantity, $P_{R=0}(R)$ is the probability that at least one attempt is successful in finding the in-register states. This is equivalent to $P_{R=0}(R) = 1 - P_{fail}(R)$, where $P_{fail}(R)$ is the probability that all attempts fail. The probability of a single attempt fail is $1 - \frac{(\ell-1)-|R|}{(\ell-1)^2}$, so

$$\begin{aligned} P_{R=0}(R) &= 1 - P_{fail}(R) \\ &= 1 - (1 - P_{succeed}(R))^{\tau_r(R)k_{Zimm}} \\ &= 1 - \left[1 - \frac{(\ell-1)-|R|}{(\ell-1)^2}\right]^{\tau_r(R)k_{Zimm}} \end{aligned} \tag{4.8}$$

with the condition $\ell > |R|$ because registries $|R|$ greater than the length of the tail cannot form base pairs and make $P_{R=0}(R)$ be equal 0.

From[26], the Zimm time is given by

$$\tau_{Zimm} \propto \frac{\eta}{kT}R^3 \tag{4.9}$$

This can be rewritten as

$$\begin{aligned} \tau_{Zimm} &= C_2\frac{\eta}{kT}b^3n^{3\nu} \\ &= C_2\frac{\eta b^3}{kT}n^{1.76} \end{aligned} \tag{4.10}$$

where C_2 is an unknown constant, $R = bn^\nu$ is the Flory radius, $b \approx 0.3nm$, $\nu = 0.586$, and n is the length of the free tail that is fluctuating.

The residence stage can be skipped in the rare events where the initial collision results

in the formation of in-register bonds. In these events $P_{R=0}(0)$=1.

Probability $P_{zip}(x)$

The final quantity in Eq. 4.4 is P_{zip}, which is the probability that the molecules successfully hybridize after finding the first in-register base pair. To solve this we introduce the quantity $P_{zip}(x)$, which is the probability that a pair of molecules with x base pairs forms the Lth base pair before falling apart. In Eq. 4.1, we require the probability that a random walk starting at a single base pair, reaches L base pairs before striking an absorbing boundary at $x = 0$. Therefore, the quantity P_{zip} in this dissertation, which describes the probability of the full zipping after forming the first base pair is given by $P_{zip} = P_{zip}(1)$.

a) $P_{zip}(x)$ is the first passage probability.

$P_{zip}(x)$ is the probability for the strands after forming the x bonds of the perfect alignment can form full bonds in zipping time. Specifically we are looking for the probability of zipping all bonds without reaching the unbound state first.

We can prove that $P_{zip}(x)$ is a solution of an ODE (see APPENDIX) which depends on two parameters.

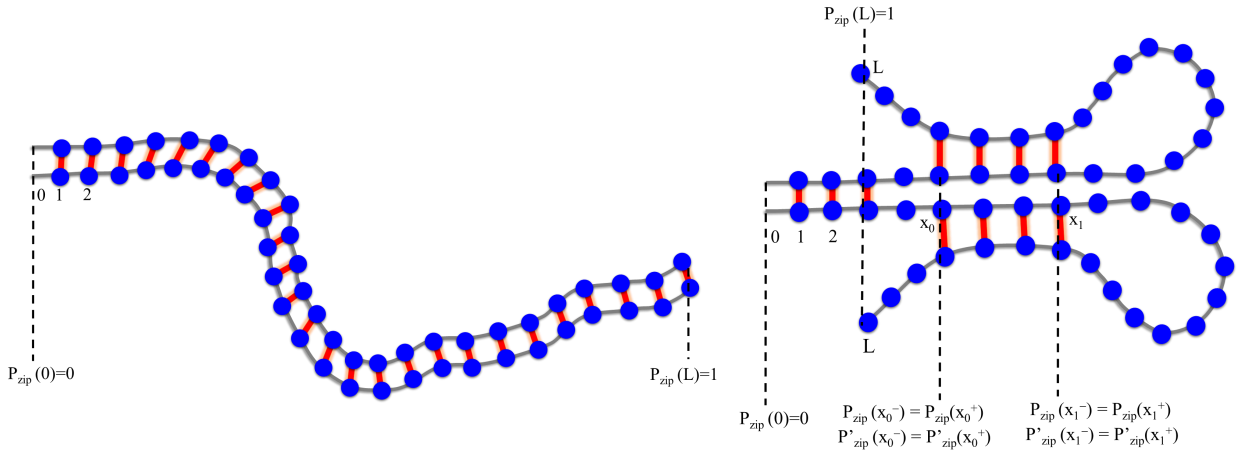
$$v = k_+ - k_- \tag{4.11}$$

$$D_x = \frac{k_+ + k_-}{2} \tag{4.12}$$

where the velocity v is a tendency for x base pairs to move toward the zipped state, and D_x is a diffusion parameter describing the random walk of base pairs, which sets the timescale for changing the bonding state. v and D_x depend on the sequence and presence of intra-molecule structure. When $v > 0$ the structure is unstructured with very low energy. $v < 0$ means the current structure contains a stem with attractive energy which needs to be broken.

b) $P_{zip}(x)$ of unstructured sequences.

For unstructured sequences it is always favorable to form new bonds, so v is always positive. We use boundary conditions (see Fig. 4.6)



(a) Unstructured DNA strands **(b) Stem-loop DNA strands**

Figure 4.6: *Cartoon representation of the boundary conditions of (a) unstructured and (b) stem-loop sequences in solving P_{zip}. The case of unstructured molecules (a) is simple with two boundary conditions while the stem-loop molecule (b) is divided into three regions including the region from 0 to x_0 which is a flexible tail, the region from x_0 to x_1 which is the stem, and the region from x_1 to L which can be considered a free piece after all H-bonds on the stem from x_0 to x_1 is broken.*

$$P_{zip}(0) = 0 \tag{4.13}$$

$$P_{zip}(L) = 1 \tag{4.14}$$

c) $P_{zip}(x)$ of stem-loop sequences.

We divide stem-loop sequence's structure into three regions including the region from 0 to x_0 which is a flexible tail which is always favorable to form new WCF bonds, so v is always positive, the region from x_0 to x_1 which is the stem with strongly attractive energy so $v < 0$, and the region from x_1 to L which can be considered as a free piece after the stem

from x_0 to x_1 is broken with $v > 0$ (see Fig. 4.6). We have a series of three equations

$$P_{zip}(x) = A_1 + B_1 e^{-v_1 x/D_1} \quad (0 \leq x \leq x_0) \tag{4.15}$$

$$P_{zip}(x) = A_2 + B_2 e^{-v_2 x/D_2} \quad (x_0 \leq x \leq x_1) \tag{4.16}$$

$$P_{zip}(x) = A_3 + B_3 e^{-v_3 x/D_3} \quad (x_1 \leq x \leq L) \tag{4.17}$$

To find A_1, B_1, A_2, B_2, A_3 and B_3 for stem-loop sequences, we use boundary conditions (see Fig. 4.6).

$$P_{zip}(0) = 0 \tag{4.18}$$

$$P_{zip}(x_0^-) = P_{zip}(x_0^+) \tag{4.19}$$

$$P'_{zip}(x_0^-) = P'_{zip}(x_0^+) \tag{4.20}$$

$$P_{zip}(x_1^-) = P_{zip}(x_1^+) \tag{4.21}$$

$$P'_{zip}(x_1^-) = P'_{zip}(x_1^+) \tag{4.22}$$

$$P_{zip}(L) = 1 \tag{4.23}$$

In this theory, we calculated $P_{zip} = P_{zip}(1)$ which is the probability that the strands reach the fully bound state after forming the first in-register.

Approximation of DNA hybridization rate

Fig. 4.7 displays the comparison of the diffusion time, the residence time, and the zipping time as a function of the concentration of DNA molecules with two sequences at 55^oC. S12 is a stem-loop sequence which has the longest zipping time and a very stable stem, while S73 is an unstructured sequence which has the shortest zipping time. The red dots show the concentration required for the diffusion time to equal the residence time and the zipping time which are much larger than the concentrations used in the experiments of Zhang et al. (50pM)[11]. Therefore, we can consider that the residence time and the zipping time are very

small in comparison to the diffusion time. This means that we can approximate the times

$$\begin{aligned} \tau_d + \tau_r + \tau_z &\approx \tau_d \\ \tau_d + \tau_r &\approx \tau_d \end{aligned} \tag{4.24}$$

Then we can rewrite the expression of hybridization rate:

$$rate \approx \frac{\sum_R P_1(R) \cdot P_{R=0}(R) \cdot P_{zip}}{(2L-1)\tau_d} \tag{4.25}$$

where L is the length of a DNA molecule and $(2L-1)$ is the number of registries, R.

In this calculation, the free parameters are $C_1 = 0.8$ which is the sequence independent geometric factor in Eq. 4.6, and $C_2 = 1900$ which is the Zimm constant in Eq. 4.11. We determine those parameters using a least squares regression.

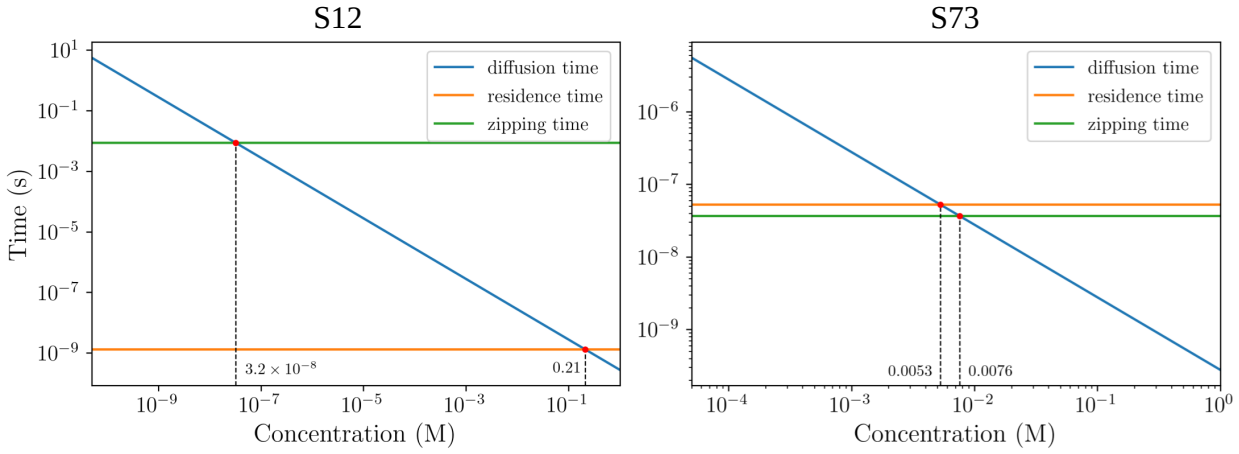


Figure 4.7: *Comparison between the diffusion time (blue line), the residence time (orange line) and the zipping time (green line) as a function of concentration of DNA molecules at 55°C. The stem-loop S12, which has a very stable stem, represents the sequence having the longest zipping time. The unstructured S73 represents the sequence having the shortest zipping time. The red dots show the concentrations required for the diffusion time equals the residence time and the zipping time. These concentrations are much larger than the concentration used in the experiments of Zhang et al. (50 pM).*

4.3 Results

4.3.1 Comparison to experiment

Our model is in good agreement with experiments[11] as shown in Figs. 4.8. We find that the residence stage plays an important role in assisting DNA strands to find in-register states and helps increase the hybridization rate. This is a new finding and it is the opposite of some models for protein aggregation such as Huntingtin[13], A_β[16,28]. In these models sequences in mis-registered states need to break all the mis-alignment bonds to fall off, then start the cycle over again until they find an in-register position. This process may take time before those sequences can end up in an in-register state. In other words, the nonspecific binding in protein aggregation reduces the aggregation rate while nonspecific binding in DNA hybridization increases the hybridization rate.

4.3.2 Discussion

Now we turn to explore the properties of DNA hybridization.

a) Unstructured sequences hybridize more easily than stem-loop sequences.

Fig. 4.8 shows that in general unstructured sequences (red dots) have a higher hybridization rate than stem-loop ones (blue dots). The reason is that unstructured sequences more readily interact at the first contact. Moreover, once they start the zipping stage, the free strands are more likely to complete the hybridization process because they do not have to break existing bonds.

In contrast, stem-loop sequences have several constraints to overcome during hybridization. First, they can only form bonds between their tails, since interactions between intra-molecular stems do not lead to hybridization. Second, once they start the zipping stage, they require additional time and energy to break the intra-molecular bonds, and this makes the probability P_{zip} is lower than the unstructured case. For example, Fig. 4.9 shows that the P_{zip} of the unstructured sequence S19 (red line) is always higher than the stem-loop sequence S14 (blue and orange lines) because the stem hinders the zipping process. Figs. 4.10

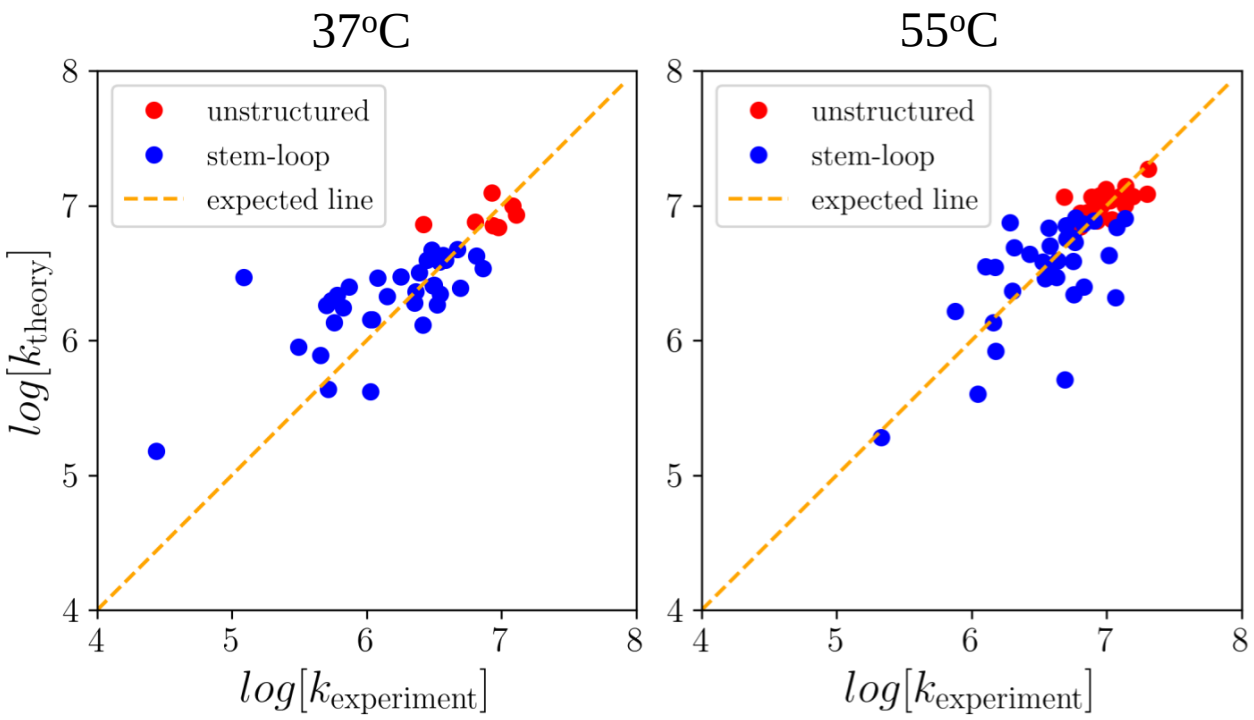


Figure 4.8: *Comparison between the theoretical model and experimentally measured DNA hybridization rate at 37°C and 55°C. The hybridization rates $k_{experiment}$ and k_{theory} are in units of $M^{-1}s^{-1}$. The red dots represent unstructured sequences, the blue dots indicate stem-loop sequences, and the orange-dash line is the expected line in which theoretical and experimental data are the same. In both cases, the theory fits well with the experiment. The hybridization rate at 55°C tends to be higher than 37°C but more scattering.*

and 4.11 show that the unstructured sequences have the shortest zipping time compared to the stem-loop sequences because they have no intra-molecular base pairs to break. Fig 4.14 indicates that unstructured sequences also have the highest probability P_{zip}. That means the zipping time and P_{zip} of the sequences depend on their conformations and have a big effect on hybridization rates between them. The effects of intra-molecule structure on DNA hybridization will be discussed specifically in section (c).

b) Longer residence time helps to increase $P_{R=0}$

Fig. 4.12 shows that a longer residence time correlates with a higher $P_{R=0}$. However, the unstructured molecules are offset from the stem-loop molecules. This is because unstructured sequences have more registries to test and, therefore, need longer residence time to find $R = 0$. Again, this trend expresses the role of residence time to help the sequence find in-register

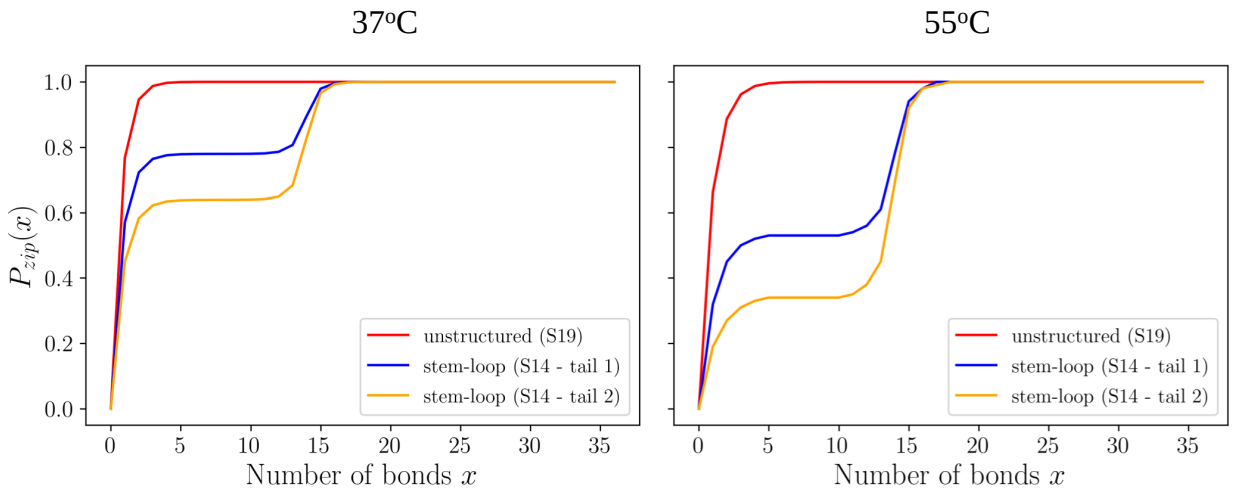


Figure 4.9: *Predicted probability of reaching the full zipped state as a function of the number of available WCF base pairs (x) at 37°C and 55°C. The red line represents the unstructured sequence S19. The blue line shows the stem-loop S14 with the first in-register bond on one free tail, the orange line represents the stem-loop S14 with the first in-register bond on the other free tail. The sequence S14 has the length of the tails as $\ell_1 = \ell_2 = 7$ nucleotides with tail 1 is AATTAGC, tail 2 is TAATCTC. The stem length is 7 base pairs.*

states. Fig. 4.13 shows similar results.

c) Intra-molecular structure and tail length affect the DNA hybridization

Fig. 4.10 reveals that the zipping time depends on the length of the stem in the intra-molecular region. The longer the stem is, the longer the zipping time. It is more useful to compare the zipping time with the sequence's free energies, which are obtained by the NUPACK software (see Fig.4.11). We can see the zipping time increases when the stability of the intra-molecular structure increases. The reason is that the energy is contributed by the base pairs in the stem region. The unstructured sequences which have the shortest zipping time always have no free energy because they have no self-base pairs. The stem-loop sequences have a very wide range of zipping time proportional to the values of their free energies. The more stable the structure is, the more time is required to break H-bonds in the stem.

However, the intra-molecular structure is not the only factor which affects DNA hybridization. The stem length and the tail length have an important relationship. In Fig

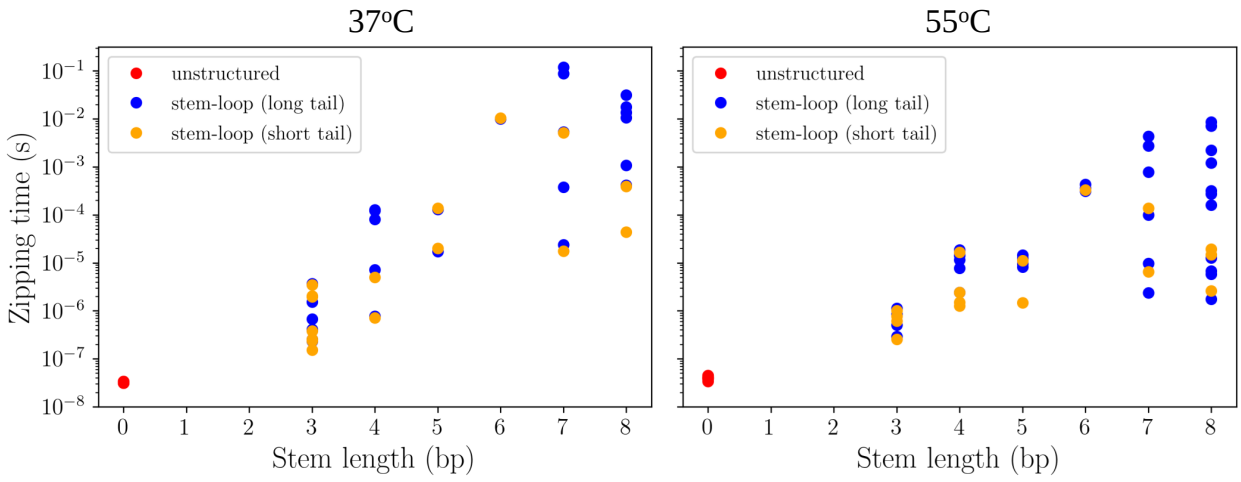


Figure 4.10: *Comparison between the zipping time versus the stem length at* $37°C$ *and* $55°C$*. The red dots represent unstructured sequences, the blue and orange dots indicate stem-loop sequences in which the first in-register bond is on the first tail (usually is the longer tail) and on the second tail (usually is the shorter tail), respectively.*

4.14, the relationship between the stem and the free tails shows a consistent trend at short stem length, but it is scattered at long length. This is explained in Fig. 4.15 which demonstrates the relation between P_{zip} and the ratio of stem length and the free tail length. If the ratio is less than about 0.75, the probability does not change much. This is because there is a competition between the lifetime of the stem and the lifetime of in-register bonds bonded by the tails. If the stem is shorter than the tails it is overwhelmingly likely that the stems unfolds before the tails separate. However, if the ratio is more than 0.75, it is likely the molecules detach before the stem unfold to allow zipping to proceed.

d) Temperature affects the zipping stage

Temperature is one of the factors that has a strong effect on DNA hybridization. Increasing the temperature makes the DNA molecules fluctuate because the WCF and mismatched base pairs are less stable. Based on the nearest-neighbor model, Eq. 2.2 in chapter 2 describes base pair energy as the function of the temperature.

In Fig. 4.8, we see that the hybridization rate at $55°C$ is higher than at $37°C$. This is because the self-bonds and mis-registered bonds are weaker so the DNA molecules can break easier and rapidly obtain the fully zipped state. However, the stem-loop sequences

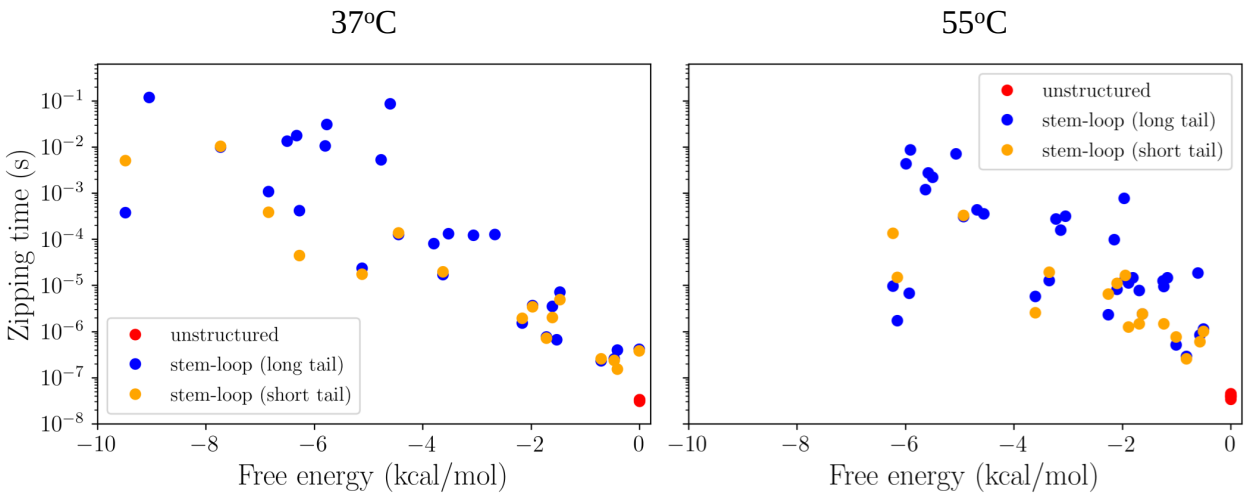


Figure 4.11: *Comparison between the zipping time versus the free energy (from NUPACK software) at 37°C and 55°C. The red dots represent unstructured sequences, the blue and orange dots indicate stem-loop sequences in which the first in-register bond is on the first tail (usually is the longer tail) and on the second tail (usually is the shorter tail), respectively.*

show more scatter around the expected line at 55°C than 37°C. This is because at higher temperatures, the DNA molecules are able to fluctuate more. In Fig. 4.10, the relationship between the zipping time and the stem length has a greater slope at 37°C than at 55°C so that the zipping time at 37°C is generally longer than at 55°C. The reason is that at 37°C the base pairs in the stem are more stable than at 55°C which requires more time to break. This trend is even more pronounced in Fig. 4.11, which plots the zipping time versus the stability of the stem-loop structure.

We compare the zipping time and P_{zip} of unstructured sequences at 37°C and 55°C in Fig. 4.16. The result shows that at lower temperature, the sequences hybridize faster with higher P_{zip}. This trend is reasonable because free energy of H-bonds will be weaker at higher temperatures and makes the in-register bonds less stable than ones at low temperature. Therefore, elevated temperatures increase the probability the molecules fall apart before fully zipping.

e) DNA hybridization depends on the order of nucleotides

One special case is the sequence S14 which is a stem-loop sequence with two free tails. The length of the stem is 7 base pairs, and the lengths of the tails are $\ell_1 = \ell_2 = 7$ nucleotides.

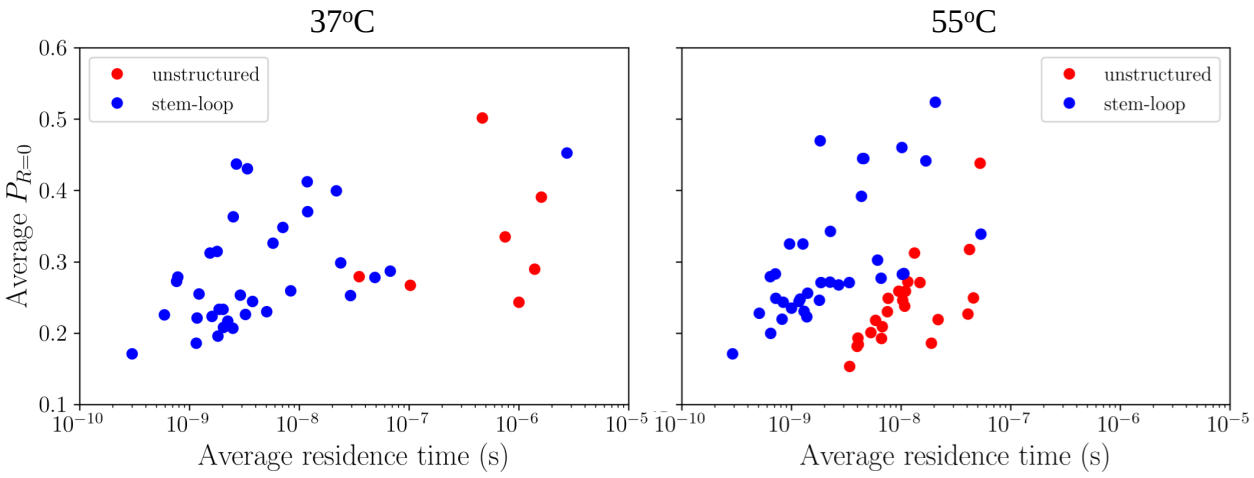


Figure 4.12: *The average probability of $P_{R=0}$ is calculated from the theoretical model as a function of the average of the residence time for all registries R at 37°C and 55°C. The red dots and the blue dots indicate the unstructured and stem-loop, respectively.*

Structure	37°C	55°C
Tail 1 5'-AATTAGC-3' 3'-TTAATCG-5'	-1.00	-1.57
Tail 2 5'-TAATCTC-3' 3'-ATTAGAG-5'	-0.86	-1.41

Table 4.1: *Average energies of free tails of S14 (k_BT/base pair)*

Tail 1 is 5'-AATTAGC-3' and tail 2 is 5'-TAATCTC-3', which both have five A/T nucleotides and two G/C nucleotides. However, P_{zip} for tail 1 is always higher than P_{zip} for tail 2. Even though the two tails look similar, the difference is the sequence of nucleotides which makes the nearest-neighbor energies different. Using the nearest-neighbor model, we calculate the average bonding energy ΔG of tail 1 and tail 2 in Table 4.1. The tail 1 always has higher energy ΔG. This increased stability is responsible for the larger P_{zip}.

f) Effect on hybridization as a function of the free tail length and base pair energy.

In the previous parts, we can see the length of free tails and base pair energies affect the DNA hybridization via affecting the zipping time and P_{zip} . Now we check how they directly affect the hybridization rate. We apply our model to unstructured sequences at 55°C (see

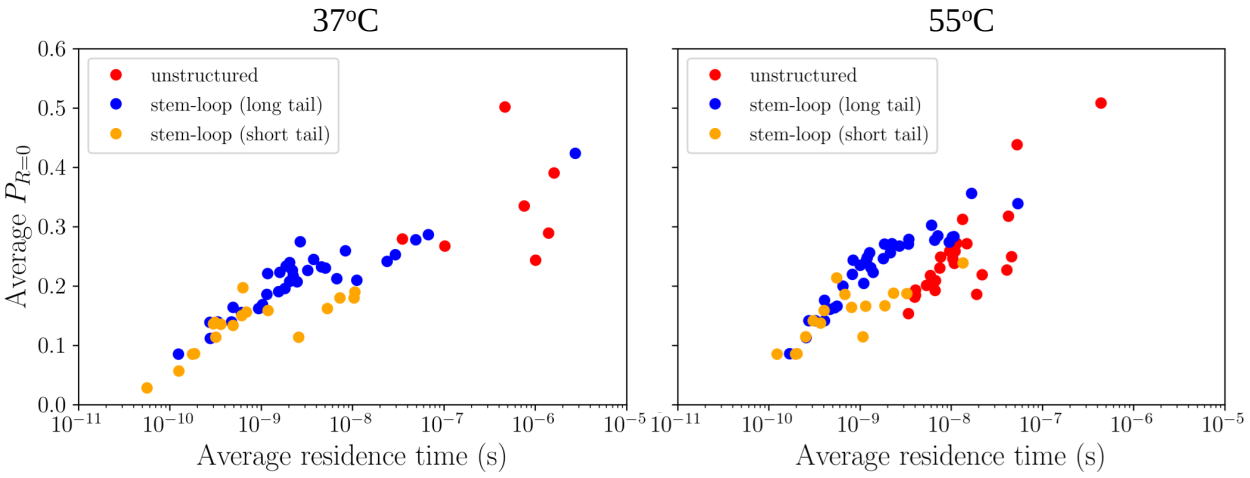


Figure 4.13: *The average of the probability $P_{R=0}$ is calculated from the theory as a function of the average of the residence time for all registries R at 37°C and 55°C. The red dots indicate the unstructured sequences, the blue and orange dots represent the relationship between the residence time and $P_R = 0$ at only the first tail (usually the longer tail) and only the second tail (usually the shorter tail), respectively. The more free bases allows longer residence times but increases the number of registries to be searched.*

Fig. 4.17). In part (a), we reduce the energies of the WCF base pairs in half. This results in a small, but consistent, reduction in the hybridization rate. In Fig. 4.17 (b) we reduce the length of the sequences in half. This makes the hybridization rate increase rapidly by nearly a factor of 6. This can be explained that in the same conditions, shorter sequences have fewer registries R so the probability $P_1(R)$ distributed for each registry R is greater than the longer ones. $P_{R=0}(R)$ is higher as well because as we explain in Fig. 4.12, the shorter tail can find the in-register position easier than the longer one.

4.4 Conclusion

Our model provides insights into the DNA hybridization mechanism. We have shown that DNA hybridization goes through three stages including the diffusion stage, residence stage and zipping stage. In the residence stage, the DNA molecules first stick together at mis-registered bonds. This binding does not hinder the process but accelerates the hybridization rate by assisting DNA molecules to search the in-register alignments. This point is opposite

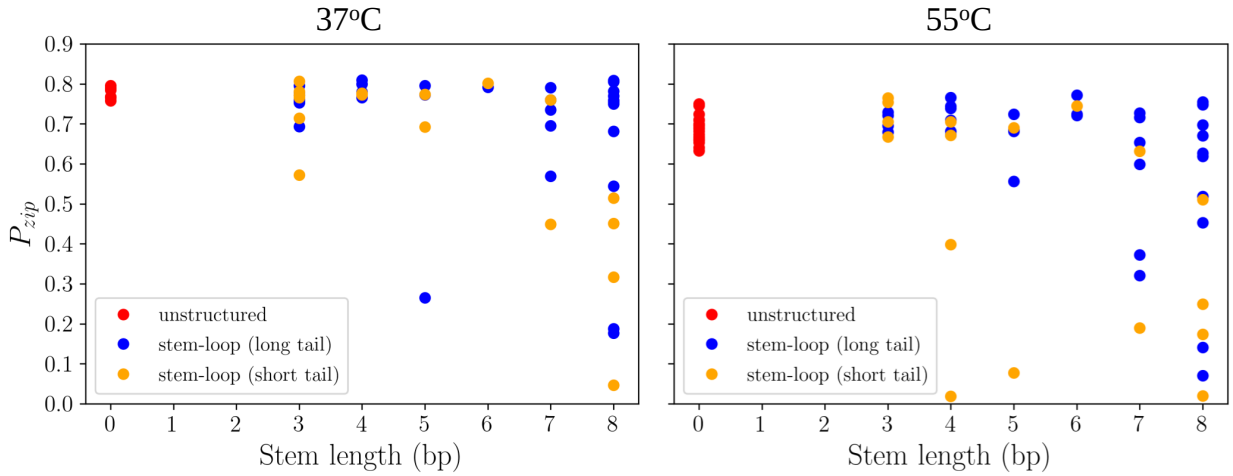


Figure 4.14: *The zipping probability is predicted from the theoretical model as a function of stem length at different temperatures* 37^oC *and* 55^oC*. The red dots represent unstructured sequences, the blue and orange dots indicate stem-loop sequences in which the first in-register bond is on the first tail (usually is the longer tail) and on the second tail (usually is the shorter tail), respectively.*

to amyloid aggregation, in which the non-specific binding slows down the aggregation rate. However, it is unclear why they have this different behavior. Perhaps protein backbones are stickier than DNA mismatch so they cannot slither as easily as DNA.

Our model also reveals that the DNA hybridization rate depends on the sequence conformations, intra-molecular structure, temperature, order of nucleotides and the length of the sequences.

We propose that nucleation sites which are the factors that help to accelerate the DNA hybridization are the sites without internal structure. This explains the theory by Wetmur et al. which reveals that the zipping stage in DNA hybridization occurs very fast with the help of a large number of nucleation sites but the model cannot show what "nucleation" is[12]. Our model indicates that the nucleation sites are the free tails which are very flexible in searching the in-register alignments.

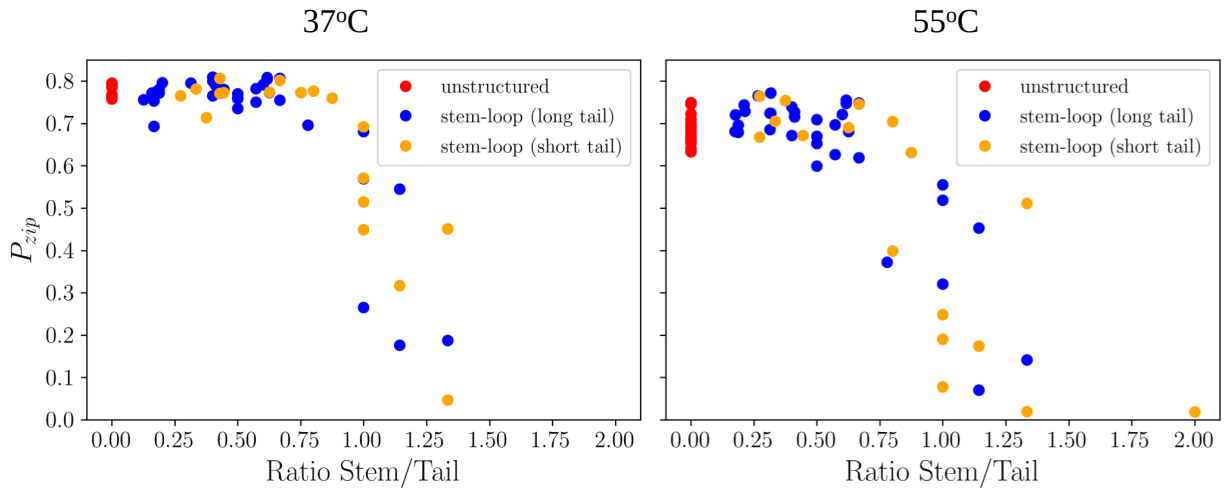


Figure 4.15: *The zipping probability is predicted from the theoretical model as a function of* $\frac{stem\,length}{free\,tail\,length}$ *at different temperatures* $37°C$ *and* $55°C$*. The red dots represent unstructured sequences, the blue and orange dots indicate stem-loop sequences in which the first in-register bond is on the first tail (usually is the longer tail) and on the second tail (usually is the shorter tail), respectively.*

4.5 APPENDIX

4.5.1 Gillespie simulation

At each step of the Gillespie algorithm, two random numbers K_1, K_2 are generated from the interval [0,1]. We call k_1, k_2,..., k_n the rates of all possible transitions from a current state i. In this case the allowed transitions are the formation on breakage of a base pair at either end of the existing base pairs (see Fig. 4.18). The transition $i+1$ is chosen when

$$\frac{\sum_{j=1}^{i-1} k_j}{k_{tot}} < K_1 < \frac{\sum_{j=1}^{i} k_j}{k_{tot}}$$

where $k_{tot} = \sum_{i=1}^{n} k_i$.

The time which elapses before the transition $i+1$ is determined by K_2

$$t = -\frac{1}{k_{tot}} ln(1 - K_2) \tag{4.26}$$

15,28

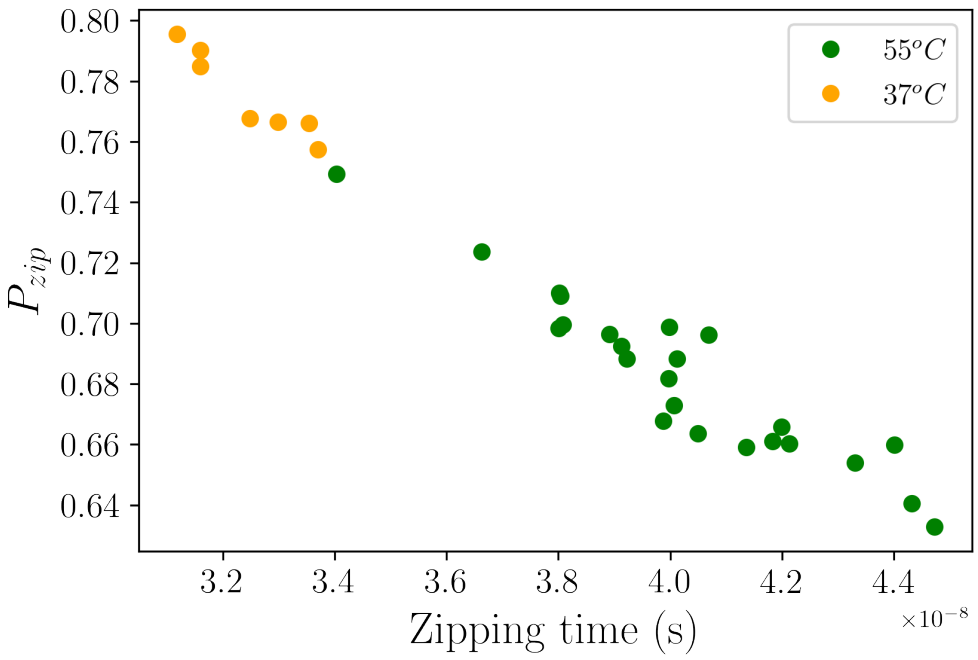


Figure 4.16: *Comparison between the zipping probability P_{zip} of unstructured sequences is predicted from the theoretical model and the zipping time at different temperatures 37^oC (orange dots) and 55^oC (green dots).*

4.5.2 Solving ODE to obtain $P_{zip}(x)$

$P_{zip}(x)$ is defined as[29]

$$P_{zip}(x) = p_+ P_{zip}(x + \delta x) + p_- P_{zip}(x - \delta x) \tag{4.27}$$

where p_+ is the probability to form one more bond, p_- is the probability to break one bond. This means the probability for two DNA strands which have x available base pairs to finally obtain the fully zipped base pairs before breaking all bonds. The two RHS terms indicate the events of forming and breaking one bond after one step.

Taylor expanding

$$P_{zip}(x + \delta x) \quad = \quad P_{zip}(x) + \delta x P'_{zip}(x) + \frac{1}{2}\delta x^2 P''_{zip}(x) \tag{4.28}$$

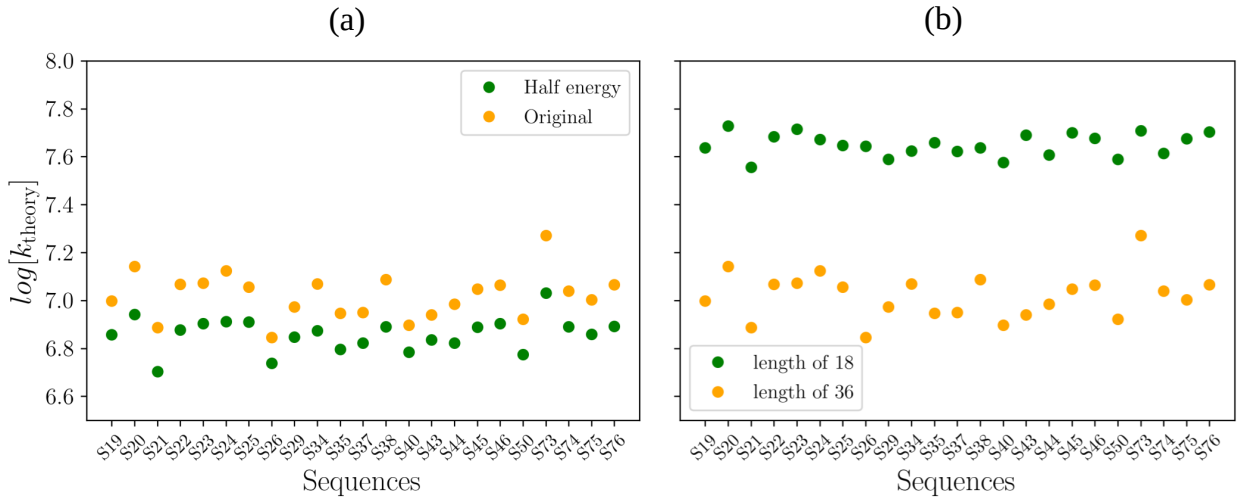


Figure 4.17: *(a) Predicted hybridization rate of unstructured sequences at 55°C as a function of (a) base pair energy, and (b) length of the sequences. In (a), the green dots show the rates as the base pair energies are reduced as a half of the original (orange dots). In (b),the green dots indicate the rates as the length of the sequences is half of the original (orange dots).*

$$P_{zip}(x-\delta x) \quad = \quad P_{zip}(x) - \delta x P'_{zip}(x) + \frac{1}{2}\delta x^2 P''_{zip}(x) \tag{4.29}$$

From Eq. 4.28 and 4.29:

$$\begin{aligned} P_{zip}(x) &= p_+[P_{zip}(x) + \delta x P'_{zip}(x) + \frac{1}{2}\delta x^2 P''_{zip}(x)] \\ &+ p_-[P_{zip}(x) - \delta x P'_{zip}(x) + \frac{1}{2}\delta x^2 P''_{zip}(x)] \\ &= (p_+ + p_-)P_{zip}(x) + (p_+ - p_-)P'_{zip}(x)\delta x + \frac{1}{2}(p_+ + p_-)\delta x^2 P''_{zip}(x) \end{aligned} \tag{4.30}$$

Notice that $p_+ + p_- = 1$, $p_+ - p_- = \frac{k_+ - k_-}{k_+ + k_-}$, $k_- = k_+ e^{\triangle G}$ with $\triangle$G is the binding energy, $\delta x = 1$ in our case. The Eq. 4.30 now becomes

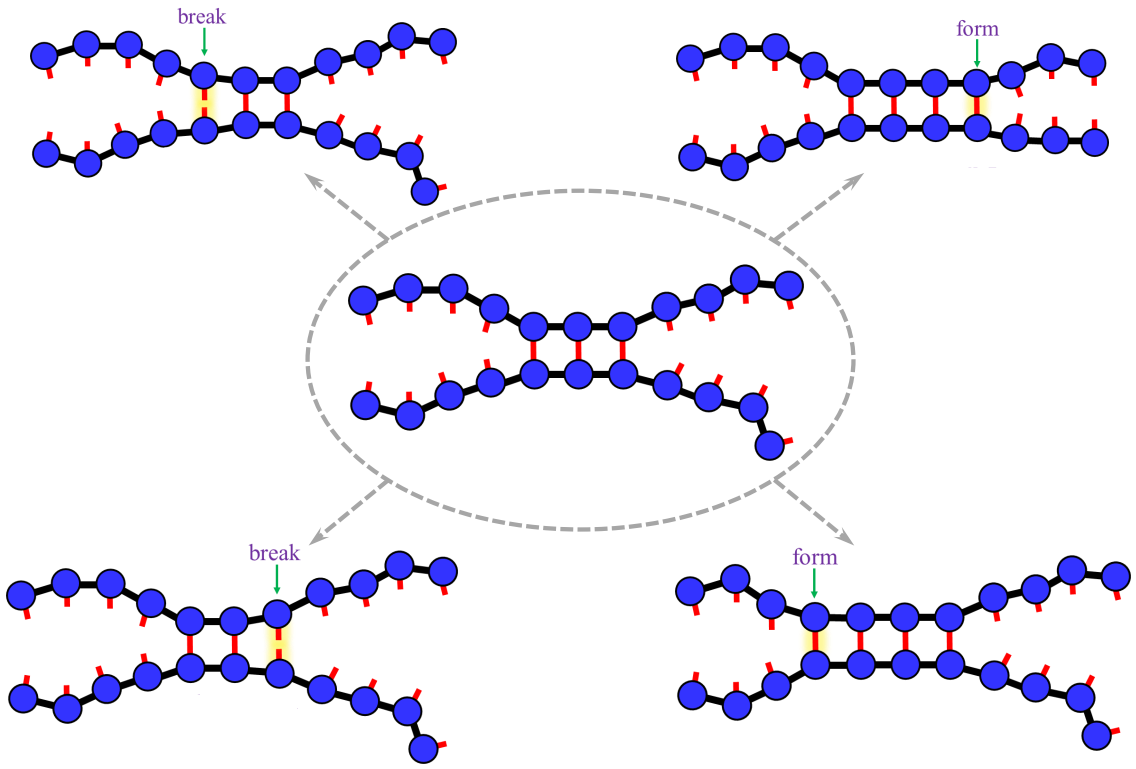


Figure 4.18: *Cartoon representation of all possible transitions from a current state in the Gillespie simulation. The DNA strands can form more or break one base pair at either end of the existing base pairs.*

$$\begin{aligned} P_{zip}(x) &= P_{zip}(x) + (p_+ + p_-)P'_{zip}(x) + \frac{1}{2}P''_{zip}(x) \\ \Leftrightarrow 0 &= (p_+ + p_-)P'_{zip}(x) + \frac{1}{2}P''_{zip}(x) \\ \Leftrightarrow 0 &= \frac{k_+ - k_-}{k_+ + k_-}P'_{zip}(x) + \frac{1}{2}P''_{zip}(x) \end{aligned} \tag{4.31}$$

Let $v = k_+ - k_-$ and $D = \frac{k_+ + k_-}{2}$, then we have

$$DP''_{zip}(x) + vP'_{zip}(x) = 0 \tag{4.32}$$

Eq. 4.32 has a general solution

$$P_{zip}(x) = A + Be^{-vx/D} \tag{4.33}$$

where A and B are constants.

4.5.3 List of single DNA strands

37^oC	ID	Original DNA Strand $(5' - 3')$
Unstructured	S19	ACATTTAGAGTAGTCCTTGGAGATTTTATGGAGATG
	S20	AAGTTGCGGTTGTGGTGATTTTGGCTTAATGTGTTC
	S21	TCACAAGACTAAAGATAATTAAAAAGAAAACCACAG
	S22	GAAACCCCATCTCTACCAAAAATATAAAAACTAGCT
	S73	CCCTGTACTTTCCACTGCCCTACCTAGATGTCCCTG
	S74	GAGATTTTGTCCCTTCATCCACCGGCTTCTAGATTA
	S75	GGACTTGACATTTTAGGGTTTTTAGGTGATTATTCT
Stem-loop	S1	AAGATGGTGAGTGCCATCTTAAAACTTACTGGAGAT
	S2	TTTTCACAAAGATGGTGAGTGCCATCTTAAAACTTA
	S3	TGTTCAACTTTTCACAAAGATGGTGAGTGCCATCTT
	S10	CAGGCGTGAGCCACCACGCCTGGCCAATTATGTAAT
	S11	GGGATTACAGGCGTGAGCCACCACGCCTGGCCAATT
	S12	AAGTGCTGGGATTACAGGCGTGAGCCACCACGCCTG
	S13	ACATAAAAATTAGCCAGGTGTGGTGGTGGGCACCTG
	S14	AATTAGCCAGGTGTGGTGGTGGGCACCTGTAATCTC
	S15	CAGGTGTGGTGGTGGGCACCTGTAATCTCAGCTACT
	S28	AAGATTAAATGGTTAGGTCTTTTTAAAAGTTGCGGT
	S30	TTTTATTTTTATTTTTTTGAGATAATTTCACTCTTG
	S37	ATTCATTTCTCAAAGAGTAAAAGTGCAGGTTGTATG
	S39	TATTCAGGGACAGTGTAGCAAGTAGCTTACAAGGGG
	S44	TACATTATATTGCCCTTCAGAATAGATTCCAGTTCC
	S48	AGGAGGACTGCTTGTGCCCAGAAGTTCGAGGCTGCA
	S51	CTGGGGCTGTTCTCATACTGGGGCTTTCTGCCCCAG

	S52	GTTCTCATACTGGGGCTTTCTGCCCCAGGACCACAC
	S54	GCTCCAGTGCACCCCAGGCTTCGTGGCCAGCCTGGG
	S55	GTGCACCCCAGGCTTCGTGGCCAGCCTGGGAAACTG
	S56	CCCAGGCTTCGTGGCCAGCCTGGGAAACTGTCTCTA
	S58	CCCTCCCAGGCCAGCAGAGGGCTGGCTGTAGCTCCC
	S60	GTGTCAGGAGCCCCTCTCTCCCTCTCTTGGAGAGAG
	S61	GAGCCCCTCTCTCCCTCTCTTGGAGAGAGTCCTGAG
	S62	CTCTCTCCCTCTCTTGGAGAGAGTCCTGAGTGCCCC
	S63	CCCTGTCACCCCGCTTATTTTCATTTCTCTCTGCGG
	S65	CCGCTTATTTTCATTTCTCTCTGCGGAGAAGATCCA
	S66	GCCATCCAATCGAGACCCTGGTGGACATCTTCCAGG
	S67	ATCGAGACCCTGGTGGACATCTTCCAGGAGTACCCT
	S85	TAAGTGAAGTCAAGTTGTTCAGGGGGCTAAGCCCAT
	S89	CTATAAATCCATGAGCAGAAAAATACATAAAATGTG
	S90	TCCCTGTACCTCCTATAAAATCAGCATGGAGCCTGG
	S93	TACCTTTGTGAGCCCCGGGCATCTGTACCTCTTTCC
	S97	ACCCCTTGCCCAGGCCAGACCTTCCTGCTATCCCCT
	S99	CTTATGGCAGCCTCTCCCTGCACTCTCTGCCCGTCT

Table 4.2: List of single DNA strands at 37^oC

55^oC	ID	Original DNA Strand $(5' - 3')$
Unstructured	S19	ACATTTAGAGTAGTCCTTGGAGATTTTATGGAGATG
	S20	AAGTTGCGGTTGTGGTGATTTTGGCTTAATGTGTTC
	S21	TCACAAGACTAAAGATAATTAAAAAGAAAACCACAG
	S22	GAAACCCCATCTCTACCAAAAATATAAAAACTAGCT
	S23	CTTAGTTGGAGTTTGGGGTATTTGAAAACGTCATGC

	S24	TCTGGTGGGGAATTTAAAAATGCATCCTGGAAATCC
	S25	CTTGGAGATTTTATGGAGATGGTGAGCACAAGGTAA
	S26	GCACTTCTCTTGAATTCCTTTATAGATGTACAGTTT
	S29	AAATATTCATTCATGAGCTCTTTTGGCAATCCGTCA
	S34	AGGTTATCTTAGTTGGAGTTTGGGGTATTTGAAAAC
	S35	GCTATCATTTCCCTCAGAAAGCTAAGTAAATTTACT
	S37	ATTCATTTCTCAAAGAGTAAAAGTGCAGGTTGTATG
	S38	CCAGGTTATCTTAGTTGGAGTTTGGGGTATTTGAAA
	S40	AATTTTACCATAAGTTTTACCTATTCGTAAGTTGGC
	S43	GCTCTTTTGGCAATCCGTCATCAGTATATTCTGAAA
	S44	TACATTATATTGCCCTTCAGAATAGATTCCAGTTCC
	S45	TGGAGTTTGGGGTATTTGAAAACGTCATGCCTTCAG
	S46	GCCCAGCTTATTTTGTGTTTTTAGTAGAGACAGGGT
	S50	TTTTAAAAAGGACATTTCTATCAGGGATATATACCT
	S73	CCCTGTACTTTCCACTGCCCTACCTAGATGTCCCTG
	S74	GAGATTTTGTCCCTTCATCCACCGGCTTCTAGATTA
	S75	GGACTTGACATTTTAGGGTTTTTAGGTGATTATTCT
	S76	ACACACTGAAGGAGCTGTAGCATCCAAGAATACTAG
Stem-loop	S2	TTTTCACAAAGATGGTGAGTGCCATCTTAAAACTTA
	S3	TGTTCAACTTTTCACAAAGATGGTGAGTGCCATCTT
	S10	CAGGCGTGAGCCACCACGCCTGGCCAATTATGTAAT
	S11	GGGATTACAGGCGTGAGCCACCACGCCTGGCCAATT
	S12	AAGTGCTGGGATTACAGGCGTGAGCCACCACGCCTG
	S13	ACATAAAAATTAGCCAGGTGTGGTGGTGGGCACCTG
	S14	AATTAGCCAGGTGTGGTGGTGGGCACCTGTAATCTC
	S15	CAGGTGTGGTGGTGGGCACCTGTAATCTCAGCTACT
	S31	GGTCGCCCCAGGAGATCACAGGTAGGGGAGTTGGGA

	S39	TATTCAGGGACAGTGTAGCAAGTAGCTTACAAGGGG
	S51	CTGGGGCTGTTCTCATACTGGGGCTTTCTGCCCCAG
	S52	GTTCTCATACTGGGGCTTTCTGCCCCAGGACCACAC
	S53	CTGGGGCTTTCTGCCCCAGGACCACACCTTCCTGTC
	S54	GCTCCAGTGCACCCCAGGCTTCGTGGCCAGCCTGGG
	S55	GTGCACCCCAGGCTTCGTGGCCAGCCTGGGAAACTG
	S56	CCCAGGCTTCGTGGCCAGCCTGGGAAACTGTCTCTA
	S57	CTGTGAACTTCCCTCCCAGGCCAGCAGAGGGCTGGC
	S58	CCCTCCCAGGCCAGCAGAGGGCTGGCTGTAGCTCCC
	S60	GTGTCAGGAGCCCCTCTCTCCCTCTCTTGGAGAGAG
	S61	GAGCCCCTCTCTCCCTCTCTTGGAGAGAGTCCTGAG
	S62	CTCTCTCCCTCTCTTGGAGAGAGTCCTGAGTGCCCC
	S63	CCCTGTCACCCCGCTTATTTTCATTTCTCTCTGCGG
	S65	CCGCTTATTTTCATTTCTCTCTGCGGAGAAGATCCA
	S66	GCCATCCAATCGAGACCCTGGTGGACATCTTCCAGG
	S67	ATCGAGACCCTGGTGGACATCTTCCAGGAGTACCCT
	S68	CCTGGTGGACATCTTCCAGGAGTACCCTGATGAGAT
	S77	TGTCAACAAAGCACAGATGCTCTCGCTGGGGCCTTG
	S80	AGCTGCCTCCCCCTTTGGGTTTTGCCAGACTCCACA
	S85	TAAGTGAAGTCAAGTTGTTCAGGGGGCTAAGCCCAT
	S93	TACCTTTGTGAGCCCCGGGCATCTGTACCTCTTTCC
	S96	AGTTTGCCCTCTTGGGCGGGGTTATCAGTGGCTGGC
	S97	ACCCCTTGCCCAGGCCAGACCTTCCTGCTATCCCCT
	S99	CTTATGGCAGCCTCTCCCTGCACTCTCTGCCCGTCT

Table 4.3: List of single DNA strands at $55^{o}C$

References

[1] B. Alberts, D. Bray, K. Hopkin, A. D. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter, *Essential cell biology* (Garland Science, 2015).

[2] T. E. Ouldridge, P. Šulc, F. Romano, J. P. Doye, and A. A. Louis, Nucleic acids research **41**, 8886 (2013).

[3] Y. Yin and X. S. Zhao, Accounts of chemical research **44**, 1172 (2011).

[4] M. G. Bajaj, Ph.D. thesis, Massachusetts Institute of Technology (2005).

[5] J. SantaLucia Jr and D. Hicks, Annu. Rev. Biophys. Biomol. Struct. **33**, 415 (2004).

[6] K. E. Bujold, A. Lacroix, and H. F. Sleiman, Chem **4**, 495 (2018).

[7] G. J. Netto, A. N. Tawil, J. T. Newman, and D. A. Savino, in *Baylor University Medical Center Proceedings* (Taylor & Francis, 1990), vol. 3, pp. 45–52.

[8] Z. Jehan, S. Uddin, and K. S Al-Kuraya, Current medicinal chemistry **19**, 3730 (2012).

[9] I. I. Cisse, H. Kim, and T. Ha, Nature structural & molecular biology **19**, 623 (2012).

[10] K. Yazawa and H. Furusawa, ACS omega **3**, 2084 (2018).

[11] J. X. Zhang, J. Z. Fang, W. Duan, L. R. Wu, A. W. Zhang, N. Dalchau, B. Yordanov, R. Petersen, A. Phillips, and D. Y. Zhang, Nature chemistry **10**, 91 (2018).

[12] J. G. Wetmur and N. Davidson, Journal of molecular biology **31**, 349 (1968).

[13] T. T. Phan and J. D. Schmit, Biophysical Journal **118**, 2989 (2020).

[14] J. D. Schmit, The Journal of chemical physics **138**, 05B611_1 (2013).

[15] C. Huang, E. Ghanati, and J. D. Schmit, The Journal of Physical Chemistry B **122**, 5567 (2018).

[16] Z. Jia, J. D. Schmit, and J. Chen, Proceedings of the National Academy of Sciences **117**, 10322 (2020).

[17] J. G. Salway, *Medical biochemistry at a glance* (John Wiley & Sons, 2012).

[18] J. N. Zadeh, C. Steenberg, J. Bois, B. Wolfe, M. Pierce, A. Khan, R. Dirks, and N. Pierce, Journal of Computational Chemistry **28**, 73 (2009).

[19] E. ToolBox (2004), URL https://www.engineeringtoolbox.com/water-dynamic-kinematic-viscosity-d_596.html.

[20] J. SantaLucia, Proceedings of the National Academy of Sciences **95**, 1460 (1998).

[21] N. Peyret, P. A. Seneviratne, H. T. Allawi, and J. SantaLucia, Biochemistry **38**, 3468 (1999).

[22] H. T. Allawi and J. SantaLucia, Biochemistry **37**, 9435 (1998).

[23] H. T. Allawi and J. SantaLucia, Biochemistry **37**, 2170 (1998).

[24] H. T. Allawi and J. SantaLucia, Biochemistry **36**, 10581 (1997).

[25] H. T. Allawi and J. SantaLucia Jr, Nucleic acids research **26**, 2694 (1998).

[26] M. Rubinstein, R. H. Colby, et al., *Polymer physics*, vol. 23 (Oxford university press New York, 2003).

[27] M. Doi and S. F. Edwards, *The theory of polymer dynamics*, vol. 73 (oxford university press, 1988).

[28] Z. Jia, A. Beugelsdijk, J. Chen, and J. D. Schmit, The Journal of Physical Chemistry B **121**, 1576 (2017).

[29] S. Redner, *A guide to first-passage processes* (Cambridge university press, 2001).

Chapter 5

Conclusion and Future Directions

In this dissertation, I develop the theory of biopolymer templating to investigate the assembly process in different systems including Huntingtin aggregation and DNA hybridization. The theory of biopolymer templating mechanisms is based on the interactions of the polymers in mis-aligned and aligned states using a random walk model to understand the thermodynamics and kinetics of some different biomolecules such as Huntingtin and DNA. The models of both Huntingtin aggregation and DNA hybridization are simple but turn out many interesting results.

1) In the Htt system, the incoming molecule is a mutant Huntingtin (mHtt) protein, the template is a Htt fibril composed of many mHtt molecules. The incoming molecule interacts with the existing fibril by forming and breaking H-bonds via mis-aligned and aligned states. The thermodynamics of Htt aggregation has three stages: monomer, oligomer, and fibril. Our model investigates that (a) the mis-aligned states slow down and hinder protein aggregation, (b) flanking segments help to accelerate the fibril formation (N17) or reduce the fibril formation (C38), (c) in fibril aggregation, only minor perturbations from uniform sequences are necessary to generate highly ordered fibrils.

2) In the system of two DNA strands, an incoming molecule is one DNA molecule, and the template is the other one. The two DNA strands can exchange their roles with each other. One is the incoming molecule, one is the template and vice versa. Similarly to the Htt

molecules, two DNA strands interact with each other by forming and breaking H-bonds via mis-aligned and aligned states. The kinetics of DNA hybridization includes three stages: the diffusion stage, the residence stage, and the zipping stage. Our model shows that: (a) The mis-aligned states speed up and increase DNA hybridization rate, (b) DNA hybridization rate depends on the sequence conformations, intra-molecular structure, temperature, order of nucleotides, and the length of the sequences.

Surprisingly, although protein and DNA systems share many similarities, they have different results. The mis-aligned states slow down and hinder the protein aggregation while nonspecific binding helps DNA to perform the alignment search during hybridization, which accelerates the hybridization rate. This can be explained:

1) Perhaps protein backbones are stickier than DNA mismatch so they cannot slither flexibly as DNA.

2) A system of DNA hybridization only contains two DNA single strands so the tails of both strands are very flexible and have space to fluctuate and search the in-register state. In contrast, amyloid fibril contains many monomers which make it stiff and bulky. Only the incoming monomer is flexible but in reality, there are not one-dimensional fibrils but multiple layers. They limit the space of interaction between the incoming molecule and the template.

3) Because the system contains only two DNA molecules, it requires the last state to be an in-register fully bound state. On the contrary, the protein system is more complicated. It contains many monomers so it cannot avoid minor perturbations in the fibril.

At this moment, the predictions above can open a new direction for future research where we can continue to improve our medium-resolution model to high resolution model to find the reasons leading to the different behaviors between protein and DNA.

1) We can develop the protein aggregation problem into 2-dimensional or 3-dimensional problems.

2) We can consider other factors that can affect protein aggregation such as sequence length, temperature.

	Htt aggregation	**DNA hybridization**
Incoming molecule	A Htt molecule	A DNA strand
Template	A stiff fibril with many Htt molecules	The other DNA strand (2 DNA strands can exchange roles with each other. One is the incoming molecule, one is the template and vice versa.)
Model	Thermodynamics (equilibrium)	Kinetics
Methods	Paper and pen	- Paper and pen - Gillespie simulation
Stages	- Monomer - Oligomer - Fibril	- Diffusion - Residence - Zipping
Assumption	The incoming molecule interacts with the template by forming and breaking H-bonds via mis-aligned and aligned states.	The incoming molecule interacts with the template by forming and breaking H-bonds via mis-aligned and aligned states.
Results	- The mis-aligned states slow down and hinder protein aggregation. - Flanking segments help to accelerate the fibril formation (N17) or reduce the fibril formation (C38).	- The mis-aligned states speed up and increase DNA hybridization rate. - DNA hybridization rate depends on the sequence conformations, intramolecular structure, temperature, nearest-neighbors of base pairs, and the length of the sequences.

	- In fibril aggregation, only minor perturbations from uniform sequences are necessary to generate highly ordered fibrils.	

Table 5.1: Summary of two biopolymer templating problems including Htt aggregation and DNA hybridization.