Physical mechanisms for self-assembly, chaperoning, and phase separation in biological systems

by

Terry J. Salava Jr.

B.S., Emporia State University, Kansas, 2021

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

2026

Abstract

Highly ordered complexes are necessary for many processes within the cell must assemble before they can function. Some systems are mobile enough that they can self-assemble while others require the help of other molecules to assemble correctly. Understanding the assembly process is critical for reducing the complexity of the broader system so that mechanisms can be studied in isolation.

We discuss three projects with the first being the kinetics of amyloid self-assembly. The second is how RNA navigates a frustrated free energy landscape with the mediation of molecular chaperones. Last is liquid-liquid phase separation and how small changes in chemical structures can result in large changes in solubility. Each of these projects involves the emergence of assemblies within the cell from simple processes. By studying these systems from a thermodynamic lens, we can gain intuition into how certain properties emerge from simple rules.

Our preliminary results show that amyloids are self poisoning systems that have long periods of arrested growth. We also show that entropic buffering of the unfolded RNA states leads to a faster folding rate. Lastly, we show how dilute phase complexes in the standard solubility product formalism for precipitation can lead to re-entrant behavior as well as magic number effects. By understanding the inner mechanisms of the cell, we can more reliably predict emergent behaviors which can be targets for drug development.

Physical mechanisms for self-assembly, chaperoning, and phase separation in biological systems

by

Terry J. Salava Jr.

B.S., Emporia State University, Kansas, 2021

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

2026

Approved by:

Major Professor
Dr. Jeremy D. Schmit

Abstract

Highly ordered complexes are necessary for many processes within the cell must assemble before they can function. Some systems are mobile enough that they can self-assemble while others require the help of other molecules to assemble correctly. Understanding the assembly process is critical for reducing the complexity of the broader system so that mechanisms can be studied in isolation.

We discuss three projects with the first being the kinetics of amyloid self-assembly. The second is how RNA navigates a frustrated free energy landscape with the mediation of molecular chaperones. Last is liquid-liquid phase separation and how small changes in chemical structures can result in large changes in solubility. Each of these projects involves the emergence of assemblies within the cell from simple processes. By studying these systems from a thermodynamic lens, we can gain intuition into how certain properties emerge from simple rules.

Our preliminary results show that amyloids are self poisoning systems that have long periods of arrested growth. We also show that entropic buffering of the unfolded RNA states leads to a faster folding rate. Lastly, we show how dilute phase complexes in the standard solubility product formalism for precipitation can lead to re-entrant behavior as well as magic number effects. By understanding the inner mechanisms of the cell, we can more reliably predict emergent behaviors which can be targets for drug development.

Table of Contents

Table of Contents v

List of Figures ix

1 Theoretical Background 1

1.1 Statistical Mechanics . . . 1

1.1.1 Introduction . . . 1

1.1.2 Thermodynamics . . . 2

1.1.3 Thermodynamic Potentials . . . 3

1.1.4 Ensembles . . . 6

1.2 Phase Transitions . . . 7

1.2.1 Overview . . . 7

1.2.2 Flory-Huggins Theory . . . 8

1.3 First Passage Problems . . . 9

1.3.1 Continuous Time Random Walk . . . 9

1.3.2 Kinetic Monte Carlo Method . . . 10

1.3.3 First Passage Processes . . . 11

1.3.4 Analytical First Passage Examples . . . 12

1.4 Other Math Methods . . . 15

1.4.1 Method of Transfer Matrices . . . 15

2 Amyloid Aggregation Kinetics are Self Poisoning 17

2.1 Introduction . . . 17

2.1.1 Overview . . . 17
2.1.2 Characteristic Regions . . . 18
2.2 Model . . . 20
2.2.1 Definitions . . . 20
2.2.2 Monomer Binding/Unbinding . . . 21
2.2.3 Disordered Cap Partition Function . . . 22
2.2.4 Ordered/Disordered Rates . . . 24
2.3 Results . . . 25
2.3.1 Variable Reduction . . . 25
2.3.2 Simulations . . . 27
2.3.3 Experimental Comparison . . . 29
2.3.4 Composition Estimates . . . 32
2.4 Discussion . . . 33
2.4.1 Fibril Stability . . . 33
2.4.2 Secondary Nucleation . . . 36
2.4.3 N State Models . . . 36

3 Nonspecific Binding with RNA Increases Folding Rate 38

3.1 Introduction . . . 38
3.1.1 The Iterative Annealing Model . . . 39
3.2 Model . . . 40
3.2.1 RNA as a Zipper . . . 40
3.2.2 Modeling Chaperones . . . 41
3.2.3 Chaperone Coordinate Grouping . . . 43
3.2.4 1D Random Walk First Passage Time . . . 45
3.3 Simulations and Experiments . . . 46

3.3.1 Simulations . . . 46
3.3.2 RNA as a Chaperone Experiment . . . 48
3.3.3 Protein Chaperone Spy Experiment . . . 49
3.4 Discussion . . . 51
3.4.1 Dominant Timescales . . . 51
3.4.2 OBM Explains Chaperone Condensates . . . 52
3.4.3 Advantages Over IAM . . . 53

4 The Magic Number Effect of Phase Separated Systems 54
4.1 Introduction . . . 54
4.2 Re-Entrant Phase Diagrams in Binary Systems . . . 55
4.2.1 Solubility Product and Ion Product of Two Component Systems . . . 55
4.2.2 Dilute Phase Modeling of Oligomers . . . 56
4.2.3 Stable Dilute Oligomers Compete with the Dense Phase . . . 58
4.2.4 The Solubility Product Formalism can Model Re-Entrant Behavior of Phase Diagrams . . . 59
4.3 Magic Numbers in Binary Systems . . . 63
4.3.1 Discussion . . . 63
4.3.2 Structural Perturbations of Monomers . . . 63
4.3.3 Ion Product for Phase Diagrams . . . 65
4.3.4 Discrete Binding Site Model Generates the Magic Number Effect . . . 66

5 Conclusion 69
5.1 Predictions . . . 69
5.2 Future Work . . . 70

Bibliography 71

A Derivation of Reversible Mean First Passage Time for Linear State Spaces 80

A.1 Linear First Passage Process 80

A.2 First Passage Probability 81

A.3 First Passage Time 84

A.4 Total Time to Reach N 86

B Derivation of the 1D Fibril Partition Function and Rates 88

List of Figures

1.1 The left figure shows the mean first passage time of a general 3 state system. The right hand figure shows a linear 3 state system which is equivalent to the general 3 state system when $k_{13} = k_{31} = 0$ 12

1.2 This figure shows the linear N state model. Notice that this model starts at state 0 and ends at N unlike the previous 3 state model. If we set $N = 2$, we will recover the 3 state linear model result identically 14

2.1 Monomer addition onto a fibril (A) can bind correctly (blue) or incorrectly (red). Fibrils can undergo degradation (B) where monomers detach themselves from the end of the fibril one at a time. Unbinding rates are determined by the free energy of the last two monomers giving 4 possible rates. Figures A and B account for all six microscopic events that can occur in our model. We can categorize the macroscopic growth of the fibril in terms of ordered (C) and disordered (D) growth. Both ordered and disordered growth end in a blue-blue bond, the growth is irreversible. Ordered growth does not capture any defect states and grows one blue monomer at a time. Disordered growth does capture defects and grows multiple monomers at a time per growth event. 19

2.2 This plot compares the Kinetic Monte Carlo simulations to our reduced theory. All simulations had the default values $p = 0.5, k_0 = 1, c_0 = 2, \epsilon_{bb} = -10, \epsilon_{br} = -2.0, \epsilon_{rb} = 4.0, \epsilon_{rr} = 4.0$. In plot A, we modify ϵ_1 which redefines $\epsilon_{br} = \epsilon_1$. Due to our choice of p, there are no added entropic terms $\log(q/p)$ for states ending in a red. Similarly plot B redefines $\epsilon_{rb}, \epsilon_{rr} = \epsilon_2$. The total time it took for a fibril to grow to 10000 monomers was used to calculate the average growth rate 28

2.3 Here are the fitted plots for several experiments involving 1D fibril growth for Amyloid Beta and Alpha Synuclein proteins[1,2,3,4]. The data was compiled and normalized by the Knowles Group where the $A\beta$ data was collected through labeled and unlabeled optical microscopy and αSyn was collected through ThT assays. Our fitting suggests that the disordered growth of these systems is negligible since ϵ_2 is very large effectively disallowing red-red and red-blue binding states. 30

2.4 The normalized growth rate of insulin fibrils measured by Quartz Crystal Microbalance[5]. A shows the line of best fit in the ordered growth limit $\epsilon_2 \gg 1$ which is identical to the registry search model[6]. The incorporation of disordered growth in figure B can now account for previously outlying points suggesting the importance of disordered growth microscopically. 32

2.5 The defect rate of a simulated fibril grown to 10000 monomers as a function of concentration. The simulations were conducted with $\epsilon_0 = -10, \epsilon_1 = -2, k_0 = 1, c_0 = 2, p = 0.5$ which are the exact same conditions found in figure 2.2B. The theory curves are calculated as $1 - P_{\text{ordered}}$ and is only valid for low concentrations. When the defect rates are below 5%, which experiments show is a plausible value[7], the theory curve perfectly matches the simulation. In the high concentration limit, the fibril proportions converge to a fixed value. 34

2.6 A gives a toy model for the geometry of a monomer, fibril, and a fibril defect. B shows how this toy model maps onto our red blue model. Red monomers represent defects with respect to the previous monomer. C gives a geometric example of how new fibrils might be nucleated. The misaligned registry of a defect monomer can serve as a template for the nucleation of new fibrils. . . 35

3.1 Figure A shows the Iterative Annealing Model (IAM) for RNA annealing. There are various forms of IAM models, but they broadly operate by unfolding the RNA repeatedly destabilizing both native and misfolded states. This is often a faster process than the RNA unfolding naturally by itself since the rate of attempts increases. Figure B shows our orthogonal binding model where the chaperone binding and the RNA folding are treated as separate independent processes composed into a single state space. This model can accommodate an arbitrary number of chaperones at the cost of more complexity. 39

3.2 Figure A shows what the state space of a single RNA fold (misfold or native) appears like in the absence of a chaperone. Figure B is a toy model of an RNA fold in the presence of a molecule that binds to RNA, but does not compete for ssRNA bonds. The RNA binding reactions are independent of the chaperone binding reactions, giving the same first passage time from all left states to all right states as figure A. Figure C is the state space of an RNA fold where the chaperone competes for ssRNA bonds. This results in some states becoming impossible to reach since dsRNA bonds and RNA-chaperone bonds are mutually exclusive 41

3.3 These plots show the folding rates of 3 different RNA strands of 30 base pairs forming helices of the same length but varying loop size and free tail length. The plots vary only in chaperone size with the left plot having a footprint of 5 base pairs and the right having a footprint of 10 base pairs. 46

3.4 These plots show how varying the chaperone properties affects the first passage times. By changing the size of the chaperone, we can change the properties of the graph as seen in the left hand graph. In the right hand graph, we change the chaperone binding rate. As the chaperone slows, the probability a chaperone binds at all becomes arbitrarily small leading to the total graph converging to the RNA only solution. It appears like we can recover the chaperone solution by interpolating between the fast chaperone and the RNA only curves. 47

3.5 This figure plots the folding rate of an RNA with a synthetically developed chaperone which is just a truncated form of the original RNA. The RNA was fret labeled so the native state will fluoresce and the misfolded state will not. By varying the chaperone concentration and measuring the florescence as a function of time, we can fit time dependent data to an exponential curve giving us the reaction rate. The concentration of the chaperone is measured as a proportion of RNA concentration. 48

3.6 Plots of the extracted data from Stull 2015[8]. The measured folding rate of a protein when in the presence of a non-specific chaperone Spy. There is an optimal concentration of spy somewhere between 10μM and 100μM. Initial addition of a nonspecific chaperone will increase the folding rate with diminishing returns. Excessive chaperones will hinder the folding of proteins. 50

4.1 Plots of 3 systems discussed in section 4.2.4. Z_I is a system with no oligomers, Z_{II} is a system with one oligomer that matches the dense phase stoichiometry, and Z_{III} is a system with stable oligomers with a different stoichiometry than the dense phase. This plot uses constants $K_{sp}, K_{N,N}, K_{2,1}, K_{1,2}, N, r = 1$. The shaded regions represent concentrations where phase separation spontaneously occurs. Notice that system III is the only system that shows any re-entrant behavior. 60

4.2 The dense phase stoichiometry r can affect how a system will phase transition. The 5 phase diagrams at the top are for a fixed dense phase ratio r with intersections at $A_t = 20$ with a dilute solution that supports oligomers up to a 3:3 ratio. The phase diagram at the bottom is built from slices of the above phase diagrams as a function of the dense phase composition. 64

4.3 Phase diagrams for systems with the free energy outlined in equation 4.39. Red areas are where dilute and dense phases coexists while blue areas only support the dilute phase. Both figures have the same constants $\log(K_{sp}) = -30.5$, $\epsilon = -3k_BT$, $v_A = 10$, $A_t = 1$, $r = v_A/v_B$. The only difference between these plots are what oligomers are present in the dilute phase. Stable oligomers delay the entry to the dense phase and cause an early exit out of the dense phase. A key feature of both diagrams is that no dense phase is nucleated when the dense phase ratio is 1:1 67

Chapter 1

Theoretical Background

1.1 Statistical Mechanics

1.1.1 Introduction

In my opinion, thermodynamics is the most important theory of physics, since it is directly responsible for powering modern every day life. It governs the laws that make all modern amenities possible by explaining how energy, heat, work, and other macroscopic thermodynamic quantities are related. The macroscopic properties of a thermal system must emerge from the microscopic behavior of the same system. Statistical mechanics is the theory that bridges the gap between macroscopic thermodynamics and microscopic statistics. This bridge is concise and general, offering an extremely powerful method of predicting the macroscopic thermal behavior of a system through microscopic modeling of the same system.

1.1.2 Thermodynamics

Suppose that we can determine the entropy of a system as a function of three extensive variables X_1, X_2, X_3. The defining feature of extensive variables is that they scale with the size of the system and not something intrinsic to the system. The three most common extensive variables are volume, internal energy, and particle number. For non-interacting systems, extensive variables must be additive meaning that they obey the equation

$$\lambda S(X_1, X_2, X_3) = S(\lambda X_1, \lambda X_2, \lambda X_3) \tag{1.1}$$

If we double each of the extensive variables in our system with $\lambda = 2$, the total entropy of our system will also double. In this process of doubling, the intensive variables that are an intrinsic property of the system remain unchanged, such as temperature, pressure, and chemical potential. The application of the chain rule for the entropy then results in the thermodynamic identity.

$$dS(X_1, X_2, X_3) = \frac{\partial S}{\partial X_1} dX_1 + \frac{\partial S}{\partial X_2} dX_2 + \frac{\partial S}{\partial X_3} dX_3 \tag{1.2}$$

This gives us the relation for the derivative of S. We could also take the derivative of equation 1.1 with respect to λ. This gives a different relationship

$$\begin{aligned} S(X_1, X_2, X_3) &= \frac{d}{d\lambda} \lambda S(X_1, X_2, X_3) && (1.3)\\ &= \frac{d}{d\lambda} S(\lambda X_1, \lambda X_2, \lambda X_3) && (1.4)\\ &= \frac{\partial S}{\partial \lambda X_1} \frac{d\lambda X_1}{d\lambda} + \frac{\partial S}{\partial \lambda X_2} \frac{d\lambda X_2}{d\lambda} + \frac{\partial S}{\partial \lambda X_3} \frac{d\lambda X_3}{d\lambda} && (1.5)\\ &= \frac{\partial S}{\partial \lambda X_1} X_1 + \frac{\partial S}{\partial \lambda X_2} X_2 + \frac{\partial S}{\partial \lambda X_3} X_3 && (1.6) \end{aligned}$$

Evaluating the above expression at $\lambda = 1$ then gives the Euler relation

$$S(X_1, X_2, X_3) = \frac{\partial S}{\partial X_1} X_1 + \frac{\partial S}{\partial X_2} X_2 + \frac{\partial S}{\partial X_3} X_3 \tag{1.7}$$

This method is not limited to 3 variables and it should be trivial to see how equation 1.1 could be extended to accommodate more extensive variables. Other more exotic extensive variables can include the total magnetic moment of a system or conservative external forces. We can replace the extensive variables found in the generalized Euler relation 1.7 with the more standard energy, volume, and particle number, given by U, V, N respectively. The conjugate variables given by $\frac{\partial S}{\partial U}, \frac{\partial S}{\partial V}, \frac{\partial S}{\partial N}$ now have special definitions, namely $\frac{1}{T}, \frac{p}{T}, \frac{-\mu}{T}$. Combining everything gives the more common Euler relation as

$$S(U, V, N) = \frac{1}{T} U + \frac{p}{T} V - \frac{\mu}{T} N \tag{1.8}$$

Note that this derivation relied on the assumption of a non-interacting additive system with extensive variables of energy, volume, and particle number. This equation is also widely applied to systems with short range interactions since they can be considered approximately additive in the finite case, and exactly additive in the infinite limit. This equation can be rearranged to solve for the internal energy of the gas giving

$$U(S, V, N) = TS - pV + \mu N \tag{1.9}$$

1.1.3 Thermodynamic Potentials

Thermodynamic potentials are a measure of how much energy a system and reservoir pair has stored. Energy can be stored in the form of a temperature, pressure, or particle gradient. A system in thermal contact with a reservoir will freely exchange temperature. A system

in mechanical contact with a reservoir will freely exchange volume. A system in chemical contact with a reservoir will freely exchange particles and temperature. The free exchange of temperature, volume, and particle number will eventually result in a thermal, mechanical, or chemical equilibrium with the reservoir. A system that has reached equilibrium with a reservoir with respect to any of its extrinsic variables can no longer store any energy in that form since the gradient will be zero (unless the reservoir is changed). Therefore, any thermodynamic potential energy that is stored must arise from the parts that are not freely exchangeable between the system and reservoir. By letting each extrinsic variable exchange freely or not at all, we would get eight possible thermodynamic potentials. Since particles can not be exchanged without exchanging temperature, this number is reduced to six. Lastly, a system that is free to transfer all extrinsic variables has no stored potential giving only the five core thermodynamic potentials. These thermodynamic potentials can be derived by taking the internal energy of the system and subtracting the energy terms that will always have a zero gradient determined by what kind of contact the system has with the reservoir. This gives the thermodynamic potentials as

$$U(S,V,N) = U = TS - pV + \mu N \tag{1.10}$$

$$F(T,V,N) = U - TS = -pV + \mu N \tag{1.11}$$

$$\Phi(T,V,\mu) = U - TS - \mu N = -pV \tag{1.12}$$

$$H(S,p,N) = U + pV = TS + \mu N \tag{1.13}$$

$$G(T,p,N) = U - TS + pV = \mu N \tag{1.14}$$

The natural variables for all of the thermodynamic potentials have changed because this process is equivalent to a Legendre transform. Taking the derivative of both sides of the free energies above will give us the thermodynamic relation for each free energy. If we let

our U be defined by equation 1.9, then we get the following relations

$$dU = TdS - pdV + \mu dN \tag{1.15}$$

$$dF = -SdT - pdV + \mu dN \tag{1.16}$$

$$d\Phi = -SdT - pdV - Nd\mu \tag{1.17}$$

$$dH = TdS + Vdp + \mu dN \tag{1.18}$$

$$dG = -SdT + Vdp + \mu dN \tag{1.19}$$

Any system open to the reservoir thermally, mechanically, and chemically can not have stored potential energy. This is equivalent to taking the internal energy of the system and subtracting off all energy terms arising from temperature, pressure, and volume i.e. $0 = U - TS + pV - N\mu$. Taking the derivative gives

$$0 = d(U - ST + Vp - N\mu) \tag{1.20}$$

$$= dU - TdS - SdT + pdV + Vdp - \mu dN - Nd\mu \tag{1.21}$$

$$= (TdS - pdV + \mu dN) - TdS - SdT + pdV + Vdp - \mu dN - Nd\mu \tag{1.22}$$

$$= -SdT + Vdp - Nd\mu \tag{1.23}$$

This equation is the Gibbs-Duhem relation and is the thermodynamic relation for an unconstrained system. The interesting part of this equation is that the right hand side can contain individual non-zero terms. This is because U has natural variables S, V, N and taking the derivative of U only gives dS, dV, dN terms.

1.1.4 Ensembles

Statistical Mechanics builds on top of thermodynamics by relating the macroscopic thermal properties and microscopic statistical properties of a system. Many important quantities that we are interested in such as the heat capacity, free energy, average occupancy, etc. can be calculated using ensembles and their associated partition functions. The partition functions for the five thermodynamic potentials listed previously are

$$\Omega = \sum_{i \in \text{States}} \delta_{E,E_i} \tag{1.24}$$

$$Z = \sum_{i \in \text{States}} e^{\frac{-E_i}{k_B T}} \tag{1.25}$$

$$\mathcal{Z} = \sum_{i \in \text{States}} e^{\frac{-E_i + \mu N}{k_B T}} \tag{1.26}$$

$$\Gamma = \sum_{i \in \text{States}} \delta_{H,E_i + pV} \tag{1.27}$$

$$\Delta = \sum_{i \in \text{States}} e^{\frac{-E_i - pV}{k_B T}} \tag{1.28}$$

The first ensemble is the micro canonical ensemble and includes all states of the system with fixed volume and particle number that have exactly energy E. Ω is also known as the degeneracy of the system. The second ensemble is the canonical ensemble that measures all states with fixed volume and particle number, but allows the system energy E to vary. The third is the grand partition function which sums over all states with a fixed volume. The fourth is not widely used since the degeneracy is easier to calculate but is called the isoenthalpic-isobaric ensemble only allowing for volume exchange with the reservoir. The last ensemble is known as the isothermal-isobaric ensemble and keeps the number of particles in the system fixed but allows the volume and temperature to freely exchange. All of these

systems are closely related to the free energies listed previously and are given by

$$TS = \quad k_B T \log(\Omega) \tag{1.29}$$

$$F = -k_B T \log(Z) \tag{1.30}$$

$$\Phi = -k_B T \log(\mathcal{Z}) \tag{1.31}$$

$$TS = \quad k_B T \log(\Gamma) \tag{1.32}$$

$$G = -k_B T \log(\Delta) \tag{1.33}$$

By breaking down our system into a sum of individual states, it make it easy to calculate properties of the system like average occupancy in the grand canonical ensemble, or average energy in the canonical ensemble. These are tools that we will need for all of our projects

1.2 Phase Transitions

1.2.1 Overview

Phase transitions are a reoccurring feature in biological systems. These transitions can be beneficial or detrimental to cellular functions such as membraneless organelles or viroplasms. A phase transition is defined as a rapid change in the extrinsic properties of a system given a small continuous change in its intrinsic properties. This is marked by any non-analytic behavior of the thermodynamic potential's derivatives. In the case of boiling water, there is some temperature and pressure at which it is more energetically favorable for the system to form a gas instead of a liquid. This is because the system gains more energy in the form of entropy than the system doe's in the form of hydrogen bonds. Over a small change in free energy at the boiling point dG, there is a discontinuous change in the volume of the system $V_{\text{gas}} - V_{\text{liquid}}$. A thermal system always attempts to minimize the corresponding

thermodynamic potential, but the optimal extrinsic variables that actually minimize it do not need to change smoothly.

1.2.2 Flory-Huggins Theory

One specific category of phase transitions that is relevant in biological physics is two component phase transitions. These phase transitions can be plotted as a function of the concentrations of each of the individual components. One common method for determining whether such a system will phase separate is Flory-Huggins theory. It is a mean field lattice theory that can model the mixing of polymers. It can account for polymer mixing and polymer interactions. We are interested in Flory Huggins since biological systems contain many kinds of polymers like DNA, Proteins, Amyloids, etc. The 2 component Flory Huggins free energy is

$$\frac{F}{V} = \frac{\phi_1}{N_1}\log(\phi_1) + \frac{\phi_2}{N_2}\log(\phi_2) + \chi\phi_1\phi_2 \tag{1.34}$$

Where ϕ_1, ϕ_2 are the volume fractions of the components that obey the equation $\phi_1 + \phi_2 = 1$. N_1, N_2 are the number of lattice sites that a single polymer of each component occupies. χ is the interaction parameter that accounts for the energy difference between the pure and mixed states. The components of Flory-Huggins can represent rigid bodies, polymers, solvents, or single lattice site gasses. Flory-Huggins can also model systems with more than two components. The N molecule Flory-Huggins free energy is

$$\frac{F}{V} = \sum_i^N \frac{\phi_i}{N_i}\log(\phi_i) + \frac{1}{2}\sum_i^N\sum_j^N \chi_{ij}\phi_i\phi_j \tag{1.35}$$

where ϕ_i is the volume fraction of molecule i, N_i is the number of lattice sites a single molecule i occupies. χ_{ij} is the interaction parameter between molecules i and j. The volume fractions must sum to 1 giving the equation $\sum_i^N \phi_i = 1$ since every lattice site must

be occupied. The first term is known as the entropic term since this accounts for the entropy of mixing. The second term is known as the enthalpic term since it accounts for all energetic interactions.

1.3 First Passage Problems

1.3.1 Continuous Time Random Walk

Stochastic processes, such as Brownian motion, are a common problem encountered in biological physics. One method for solving problems like this is the first passage method. By measuring the time it takes to go from an initial state to some final state on average, we can estimate reaction rates even if we don't have an exact solution. The type of random walk that we are most interested in is the continuous time random walk. Since the spatial dimensions of our random walk are discrete, we can encode all the states of our system into a vector v. v can be written as

$$v(t) = \begin{bmatrix} a_1(t) \\ a_2(t) \\ \vdots \\ a_{n-1}(t) \\ a_n(t) \end{bmatrix} \tag{1.36}$$

Where n is the dimensionality of the system. Each dimension $a_i(t)$ corresponds to a population or occupancy of an individual state at time t. Now we define some set of reactions K for our system which transfers the occupancy from one state to another. By tracking all outgoing and incoming rates from every state individually, we can measure how the populations of our individual states changes with time $\frac{dv}{dt}$. This leads to the master equation

$$\frac{dv(t)}{dt} = Kv(t) \tag{1.37}$$

where the matrix K contains all reaction rates k_{ij} from any state i to j. For small systems, this matrix can be diagonalized with an eigen system where we can inspect the largest/smallest eigenvalues to determine the rate limiting step of our system. Other systems are so highly dimensional that diagonalization is numerically difficult and prone to errors. For these high dimensional systems, it is sometimes easier to simulate trajectories through the state space rather than diagonalizing a high dimensional matrix.

1.3.2 Kinetic Monte Carlo Method

The Kinetic Monte Carlo Method (KMC) is the method we use to simulate the continuous time random walk listed in the previous section. This simulation is guaranteed to give a statistically correct trajectory through state space. This can be used to collect data on systems that are too complicated to numerically solve due to the high dimensionality of states. This method is particularly useful because we can simulate a difficult original system and compare it to and approximated reduced system to measure when the approximation breaks down. This method acts as if it is following a single molecule in the ensemble on its trajectory through state space. Unlike the previous section where all populations are modified at the same time, this one traces a single particle through state space. An individual simulated trajectory can be completed with a series of steps that results in a single first passage time. By sampling many of these trajectories, we can calculate the mean first passage time (MFPT) to go from some initial state to some final state. This can give us more intuition into how a system operates compared to observing the evolution of the entire distribution. If we know the particle is in state i, KMC must obey the master equation

$$\frac{dP_i}{dt} = -\sum_j k_{ij} P_i \tag{1.38}$$

where P_i is the probability that our system remains in state i and k_{ij} is the transfer rate from state i to j. The interpretation of this equation is that the longer the state remains in state i, the more likely it is to react into one of its connected neighbors. This is a first order linear ODE and can easily be solved giving the equation

$$P_i(t) = \exp\left(-t \sum_j k_{ij}\right) \tag{1.39}$$

The normalization constant is set to 1, since at time 0, we know that we must still be in state i with probability 1. Rearranging the terms of this equation, we can solve for the time, which gives the equation

$$t = \frac{\log(1/P_i)}{\sum_j k_{ij}} \tag{1.40}$$

This is the time it takes for the system to leave state i and enter one of its neighboring states. This lets us treat our system as a Markov Chain where every step adds the time 1.40 where the probability P_i is generated by a pseudo random number generator. This allows us to generate a possible trajectory from the beginning to the end state as well as its time. By averaging many trajectories, we can calculate the mean first passage time.

1.3.3 First Passage Processes

The first passage time for a continuous time random walk is the average time it takes for some initial state to reach some final set of states. Given the exit condition of reaching the state j, the first passage time from any other state i to the final state j is fixed given by T_{ij}. This mean first passage time must obey the recurrence relations

$$T_{jj} = 0 \tag{1.41}$$

$$T_{ij} = \sum_k p_{ik}(r_i + T_{kj}) \tag{1.42}$$

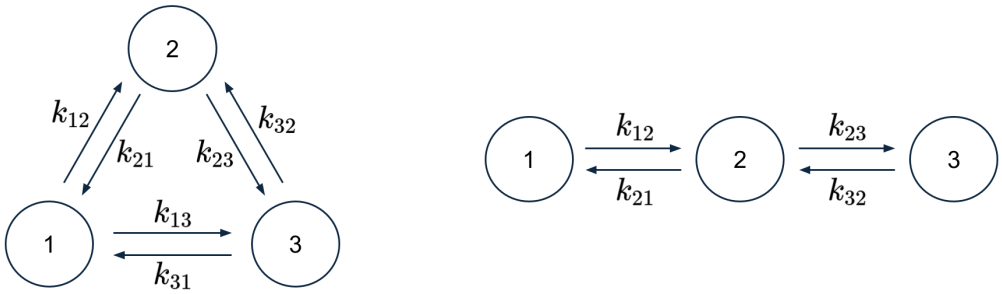


Figure 1.1: *The left figure shows the mean first passage time of a general 3 state system. The right hand figure shows a linear 3 state system which is equivalent to the general 3 state system when* $k_{13} = k_{31} = 0$

where r_i the average time spent in state i, or the residence time. Both the probabilities p_{ij} and the residence times r_i can be calculated from the individual reaction rates k_{ij} given by the continuous time random walk. Rewriting the formula gives the equation

$$r_i = \frac{1}{\sum_j k_{ij}} \tag{1.43}$$

$$p_{ij} = k_{ij} r_i \tag{1.44}$$

Oftentimes it is more natural to express the first passage time in terms of probabilities and times. It is important to know that these terms can be calculated entirely from the rate constants given by the differential equation 1.37. For small systems, such as the examples shown in figure 1.1, these first passage times can be calculated exactly and analytically. For other systems, matrix denationalization is too difficult, and we must make use of simulations or approximations.

1.3.4 Analytical First Passage Examples

3 State First Passage Times The first graph that we will solve the first passage problem for is a 3 state model where each state is bidirectionally connected to every other state. This means that we will have 6 kinetic rates as seen in figure 1.1. We can then calculate the

transition probabilities and residence times using equations 1.43 and 1.44. We can find the first passage time from the initial state 1 to the final state 3 by exhaustively summing over all paths.

$$T_{13} = p_{13}r_1 + p_{12}p_{23}(r_1 + r_2) + p_{12}p_{21}p_{13}(r_1 + r_2 + r_1) + ... \tag{1.45}$$

$$= \sum_{n=0}^{\infty} p_{13}(p_{12}p_{21})^n(r_1 + n(r_1 + r_2)) + p_{12}p_{23}(p_{12}p_{21})^n(r_1 + r_2 + n(r_1 + r_2)) \tag{1.46}$$

Although this method will give you the correct answer, it does not generalize to higher order systems. The same result can be achieved using the recurrence relationships outlined in equations 1.41 and 1.42. After two applications of these equations, we get

$$T_{13} = p_{13}r_1 + p_{12}p_{23}(r_1 + r_2) + p_{12}p_{21}(r_1 + r_2 + T_{13}) \tag{1.47}$$

We can rewrite this equation in terms of rate constants $k_{m,n}$ giving the equation

$$\begin{aligned} T_{13} = \frac{k_{12}}{k_{12}+k_{13}}\frac{1}{k_{12}+k_{13}} + \frac{k_{12}}{k_{12}+k_{13}}\frac{k_{23}}{k_{23}+k_{21}}\left(\frac{1}{k_{12}+k_{13}} + \frac{1}{k_{23}+k_{21}}\right) \\ + \frac{k_{12}}{k_{12}+k_{13}}\frac{k_{21}}{k_{23}+k_{21}}\left(\frac{1}{k_{12}+k_{13}} + \frac{1}{k_{23}+k_{21}} + T_{13}\right) \end{aligned} \tag{1.48}$$

Notice that the linear 3 state system can be recovered by setting $k_{13} = 0$ giving the linear 3 state model the mean first passage time as

$$T_{13}^L = \frac{1}{k_{12}} + \frac{k_{23}}{k_{23}+k_{21}}\left(\frac{1}{k_{12}} + \frac{1}{k_{23}+k_{21}}\right) + \frac{k_{21}}{k_{23}+k_{21}}\left(\frac{1}{k_{12}} + \frac{1}{k_{23}+k_{21}} + T_{13}^L\right) \tag{1.49}$$

N State Linear First Passage Times The linear N state model can be represented as a non-uniform random walk where the probability of going left or right depends on its current location. A visual representation can be seen with figure 1.2. This can be uniquely determined by individual rates $k_{n,n+1}, k_{n,n-1}$ for all n. We can represent the rate matrix K

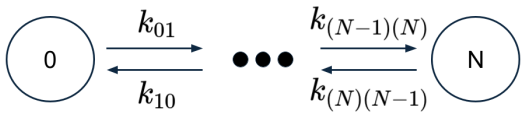


Figure 1.2: *This figure shows the linear N state model. Notice that this model starts at state 0 and ends at N unlike the previous 3 state model. If we set $N = 2$, we will recover the 3 state linear model result identically*

as

$$K = \begin{bmatrix} 0 & k_{0,1} & 0 & \dots & 0 & 0 \\ k_{1,0} & 0 & k_{1,2} & \dots & 0 & 0 \\ 0 & k_{2,1} & 0 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & 0 & k_{N-1,N} \\ 0 & 0 & 0 & \dots & k_{N,N-1} & 0 \end{bmatrix} \tag{1.50}$$

The astonishing part is that we can still find the first passage time from state 0 to state N for this state space using a single closed form equation. Its derivation is long and cumbersome and can be found in appendix A. The mean first passage time of this system is given by

$$T_{0,N} = r_0 \sum_{i=1}^{N-1} \prod_{j=1}^{i-1} \frac{q_j}{p_j} + \sum_{i=1}^{N-1} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j} \tag{1.51}$$

where p_i, r_i, q_i are given by

$$p_i = \frac{k_{i,i+1}}{k_{i,i-1} + k_{i,i+1}} \tag{1.52}$$

$$q_i = \frac{k_{i,i-1}}{k_{i,i-1} + k_{i,i+1}} \tag{1.53}$$

$$r_i = \frac{1}{k_{i,i-1} + k_{i,i+1}} \tag{1.54}$$

1.4 Other Math Methods

1.4.1 Method of Transfer Matrices

Transfer matrices are a method used for solving one dimensional problems. Transfer matrices can be used to solve N'th order linear difference equations, partition function calculations, or sequentially constructed mathematical systems. If we assume the finite difference approximation, the method of transfer matrices can be used to solve some first order ordinary partial differential equations. Practically, the interaction distance should be kept small since the dimensionality of the transfer matrix can grow exponentially with increasing interaction distance. In general, regardless of the system, this lets us define a recurrence relation

$$v_{n+1} = Av_n \tag{1.55}$$

Where the dimensionality D of the vector v_n is determined by the local interactions of the system. The vector v_n encodes all information of the system up to the index n and the transfer matrix A encodes all local interactions of the system. If we know the initial boundary condition v_0, then this allows us to calculate any state v_n by repeatedly applying

the transfer matrix A to the base state v_0 giving the equation

$$v_n = A^n v_0 \tag{1.56}$$

For most problems involving transfer matrices, we are only interested in calculating the final vector v_n. This problem is made simple if the matrix A is diagonalizable. If there exists a matrix P that diagonalizes A, then there exists a diagonal matrix D such that $A = PDP^{-1}$. Expanding A^n and substituting A gives

$$v_n = PD^n P^{-1} v_0 \tag{1.57}$$

Although both equations 1.56 and 1.57 give a correct expression for v_n, equation 1.57 tends to be much easier to calculate in practice. The most difficult part of this expression is generating the matrix P that diagonalizes A. Since D is diagonal, it is trivial to calculate the value D^n. Repeated multiplication of a diagonal matrix is trivial compared to trivial multiplication of a non-diagonal matrix. This allows us to calculate the value of v_n in a reasonable amount of time.

Chapter 2

Amyloid Aggregation Kinetics are Self Poisoning

2.1 Introduction

2.1.1 Overview

1D filaments are an important feature of many systems in biophysics. These systems range from beneficial cytoskeleton structures to toxic amyloid structures. Many of these toxic filament structures cause neuro-degenerative diseases such as Parkinson's and Huntington's disease since they form 1D filaments in the brain. Understanding how an isolated 1D filament can grow might be useful in understanding how drugs affect the progression of the disease. When the growth of the filaments is measured against the monomer concentration, several trends appear. The first is the existence of a linear growth regime where increasing the concentration of monomers increases the rate of growth. This region is then followed by a plateau where increasing the concentration no longer increases the growth rate of the fibril[9]. This can be explained using the registry search model, which has thermally stable states that are incompatible with further growth[6,10]. The second story is about secondary nucleation as

experiments show exponential growth after a fibril is nucleated[9,7]. This suggests some sort of secondary nucleation mechanism in which previously formed fibrils support the nucleation of new fibrils. We claim that the defect states in the red-blue binding model for fibril growth are sufficient to explain both the self poisoning plateau regime as well as the secondary nucleation of new fibrils.

2.1.2 Characteristic Regions

Many fibrils in the cell self-assemble through the addition of monomers[9,11,12]. In this paper, we propose a model for the self-assembly of 1D filaments that can resolve the issues faced by previous models[13]. The most notable issue with the previous model is that it does not generalize to all fibril systems. Some fibril systems such as Aβ and αSyn have a very long plateau region that some theories could predict[6]. Other fibril systems seemingly have no plateau at that other theories could model[13]. Our theory offers a generalization of the previous red blue model that can smoothly interpolate between the two types of fibril systems with and without plateaus. The growth rate as a function of concentration can be categorized into three regimes with our theory. The first is linear growth in the low concentration regime where the growth rate is primarily determined by the monomer hitting rate with the fibril. On this time scale, fibril growth is almost entirely ordered since any incorrect conformation has enough time to be removed before becoming locked in. The second region is the plateau region. This region is characterized by self poisoning behavior of thermally stable defect states that are incompatible with further growth. Amyloid Beta 42 is an example of a system that exhibits this property[1]. This leads to a flat plateau region of the rate vs time graphs seen in figure 2.3 The composition of the fibril in the plateau region is still mostly ordered although defects become more common as the concentration increases as seen in figure 2.5. The last region is the disordered aggregation regime where the concentration of monomers becomes so high that any attempt at rejecting any monomer

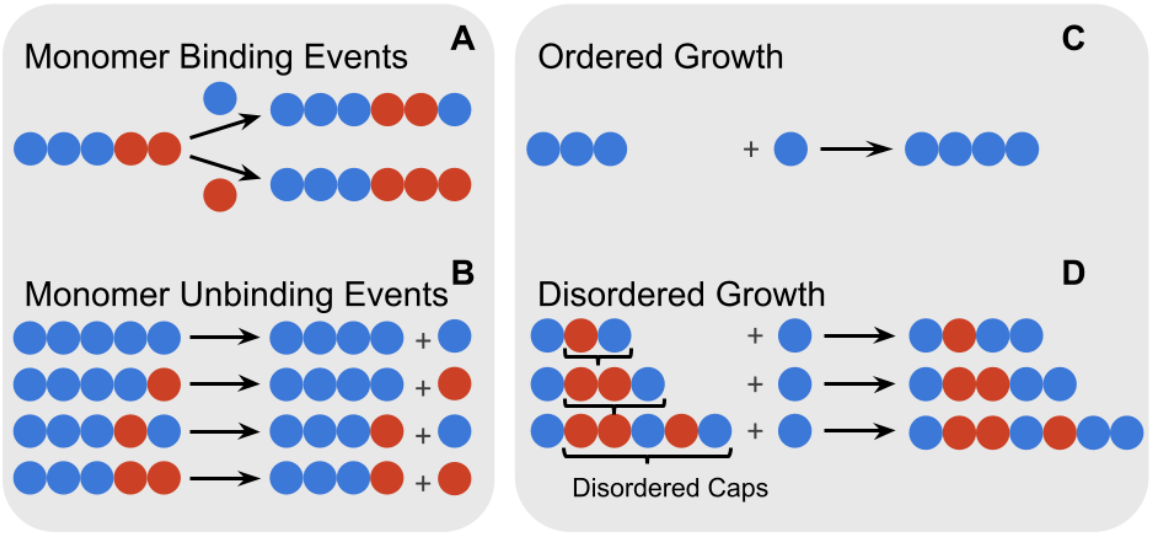


Figure 2.1: *Monomer addition onto a fibril (A) can bind correctly (blue) or incorrectly (red). Fibrils can undergo degradation (B) where monomers detach themselves from the end of the fibril one at a time. Unbinding rates are determined by the free energy of the last two monomers giving 4 possible rates. Figures A and B account for all six microscopic events that can occur in our model. We can categorize the macroscopic growth of the fibril in terms of ordered (C) and disordered (D) growth. Both ordered and disordered growth end in a blue-blue bond, the growth is irreversible. Ordered growth does not capture any defect states and grows one blue monomer at a time. Disordered growth does capture defects and grows multiple monomers at a time per growth event.*

is negligible. The composition of the fibril in this regime also plateaus out and converges to a constant that is dependent on the initial probability of binding correctly. Although our theory breaks down in this regime, the effective growth of the fibril will be linear with concentration and the composition of the fibril will also become constant. This region can be observed in figure 2.4

2.2 Model

2.2.1 Definitions

Our model describes a solution of monomers that self-assemble into 1D fibrils. Fibrils grow when monomers attach to the ends of a fibril in a correct (blue) or incorrect (red) conformation. Blue binding represents the "correct" conformation and is the most energetically stable binding state. Red binding represents the ensemble of incorrectly bound monomer conformations and is less energetically stable. We assume a nearest neighbor model where the free energy of a monomer binding to a fibril depends only on the monomer at the end of the fibril and the incoming monomer to be added. Because of this, we refer to the bond types as red-red, red-blue, blue-red, and blue-blue where the first term represents the monomer at the end of the fibril and the second is the incoming monomer. The fibril grows with successive binding of blue or red monomers. Red defect states appear in both an equilibrium and kinetic capacity. Following previous work, fibril growth can be categorized into ordered growth and disordered growth[13]. Starting from an ordered fibril (ending in a blue-blue bond), ordered growth is defined by immediate binding of another blue. A fibril entirely grown through ordered growth does not contain defects. Disordered growth occurs when a blue-blue bond locks in at least one defect. This introduces numerous possible ways to grow the fibril so we introduce the concept of a disordered cap. The disordered cap includes all monomers that have attached to the fibril since the most recent blue-blue bond. This, by definition, excludes blue-blue bonds from the disordered cap. The disordered cap can become permanent by locking it in with a stable blue-blue bond and resetting the disordered cap size to zero. The locking in of disordered caps is what we call disordered growth.

2.2.2 Monomer Binding/Unbinding

To model monomer binding, we assume that the rate at which a monomer in solution binds to a fibril is proportional to the monomer concentration. Any given monomer has the potential to bind correctly or incorrectly and it does so with probability p and q respectively. These probabilities are independent of the fibrils composition. This gives the monomer binding rates as

$$k_{b+} = k_0 pc \tag{2.1}$$

$$k_{r+} = k_0 qc \tag{2.2}$$

where k_0 is the reference rate determined by the monomer hitting rate and reaction probability, c is the unitless concentration of monomers, and $q = 1 - p$. Here, k_{b+} is the binding rate of blue monomers and k_{r+} is the binding rate of red monomers. The monomer unbinding rate is determined by the free energy of the bond between the outter most monomer and its nearest neighbor. This gives four binding energies which leads to the four binding rates

$$k_{bb-} = k_0 e^{\epsilon_{bb}} \tag{2.3}$$

$$k_{br-} = k_0 e^{\epsilon_{br}} \tag{2.4}$$

$$k_{rb-} = k_0 e^{\epsilon_{rb}} \tag{2.5}$$

$$k_{rr-} = k_0 e^{\epsilon_{rr}} \tag{2.6}$$

k_0 is the same rate constant as before due to a detailed balance condition. Here k_{ij-} represents the breaking of an ij bond where i is the new end of the fibril and j is the monomer breaking off into solution. Each rate is determined by the free energy of binding ϵ_{ij}. We neglect fragmentation of the fibril where it breaks in the middle. Since blue is the native state of the 1D fibril, it should be the most energetically favorable requiring

$\epsilon_{bb} << \epsilon_{br}, \epsilon_{rb}, \epsilon_{rr}$. The consequence of this assumption is that all blue-blue bonds are nearly permanent and irreversible. Any blue-blue bond will result in nearly permanent macroscopic growth of the fibril.

2.2.3 Disordered Cap Partition Function

Continuing with previous work, categorize the permanent macroscopic growth into two categories, ordered growth and disordered growth[13]. Both ordered and disordered growth are permanent and must end in a blue-blue bond. Ordered growth occurs when a blue monomer is added directly onto an already ordered fibril ending in a blue-blue bond. Ordered growth is free of defects and grows the fibril one monomer at a time. Disordered growth occurs when there is at least one red monomer on the fibril that is locked in with a blue-blue bond further along. Disordered growth allows for the capture of defects in the fibril that previous models did not allow[6]. A key assumption of this model is that all disordered caps are sampled from a quasi-equilibrium distribution that takes a negligible amount of time to reach relative to the macroscopic growth events. This allows us to generate a distribution of disordered caps in the form of a partition function that we use to calculate the ordered and disordered growth rates. Here we define some partition functions

$$Z = Z_r + Z_b \tag{2.7}$$

$$Z_r = \sum_{l}^{\infty} Z_{r,\ell} \tag{2.8}$$

$$Z_b = \sum_{l}^{\infty} Z_{b,\ell} \tag{2.9}$$

Here, $Z_{r,\ell}(Z_{b,\ell})$ is the partition function of all states starting from a blue-blue bond ending in a red(blue) monomer where the disordered cap is of length ℓ. By summing over all lengths, we get the partition function of all states ending in a red(blue) as $Z_r(Z_b)$. Adding these

partition functions, we get the partition function of all disordered caps Z. We can generate these partition functions using the following transfer matrix

$$\begin{bmatrix} Z_{b,\ell+1} \\ Z_{r,\ell+1} \end{bmatrix} = \begin{bmatrix} 0 & pce^{-\epsilon_{rb}} \\ qce^{-\epsilon_{br}} & qce^{-\epsilon_{rr}} \end{bmatrix} \begin{bmatrix} Z_{b,\ell} \\ Z_{r,\ell} \end{bmatrix} \tag{2.10}$$

The top left entry of the matrix is 0 to prevent any blue-blue bonds from being included in the calculation of the disordered caps. Any inclusion of a blue-blue bond would lock in the disordered cap resetting the disordered cap size to 0. Since both disordered growth and ordered growth end in a blue-blue bond, we can assume that all fibrils without a disordered cap end in a blue monomer. This leads to the boundary conditions of $Z_{b,0} = 1, Z_{r,0} = 0$. Combining the recurrence relation with the base case, we can generate all $Z_{b,\ell}, Z_{r,\ell}$ which can be summed to get the total partition function as (see appendix B for details)

$$Z = \frac{1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}}{1 - qce^{-\epsilon_{rr}} - qpc^2e^{-\epsilon_{rb}-\epsilon_{br}}} \tag{2.11}$$

This partition function can be used with the quasi-equilibrium assumption to calculate the probability distribution of the disordered caps and their sizes. This is a generalization of the partition function from former models which treats the energies independently instead of imposing the condition $\epsilon_{rb} = \epsilon_{br} = \epsilon_{rr}$ [13]. If we set the energies to be equal, then the partition function from the previous work can be recovered exactly. It should be noted that the derivation assumes the summation of an infinite geometric series which is only valid for concentrations less than the divergence concentration c_{div}. This is the concentration where our model breaks down and is defined by the equation

$$(1 - qc_{\text{div}}e^{-\epsilon_{rr}} - qpc_{\text{div}}^2e^{-\epsilon_{br}-\epsilon_{rb}}) = 0 \tag{2.12}$$

This causes the denominator of the partition function to go to 0 allowing arbitrarily large captures of monomers in very short amounts of time. Physically, this is the concentration where the free energy of adding a red monomer is negative and thermally stable. This will result in amorphous aggregation likely breaking the 1D fibril assumption. This means that our theory is valid for concentrations in the range of $0 < c < c_{\text{div}}$.

2.2.4 Ordered/Disordered Rates

As seen in figure 2.1C, ordered growth requires that there is no disordered cap at the time of a blue binding event. Using the partition function previously defined, we know the probability that the fibril does not have a disordered cap is $Z_{b,0}/Z$. If the fibril exists in this state, and binds a blue monomer with rate k_{b+}, then an ordered growth event has occurred. We also group the fibril degradation term with the ordered growth term since both are proportional to the probability that there is no disordered cap $Z_{b,0}/Z$. This gives us the ordered growth rate as

$$R_{\text{ord}} = (k_{b+} - k_{bb-})\frac{Z_{b,0}}{Z} \tag{2.13}$$

$$= \frac{(k_0 pc - k_0 e^{\epsilon_{bb}})(1 - qce^{-\epsilon_{rr}} - qpc^2 e^{-\epsilon_{rb}-\epsilon_{br}})}{1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}} \tag{2.14}$$

The degradation term assumes that only a single monomer is removed from the fibril from the removal of the outer monomer. This may not be the case in reality since the broken blue-blue bond could be locking in a defect leading to the rapid degradation of several more monomers. It should then be noted that equation 2.13 rate is only valid for fibrils mostly composed of blue monomers that do not fragment. This is a reasonable assumption for amyloid systems since measured defect rates are rare, ranging from 1/99 to 1/834 for some systems[7]. To calculate the disordered growth rate as seen in figure 2.1D, we must sum over all possible defect captures. Every defect capture event must have a disordered cap

ending in a blue state since a blue-blue bond is required for a permanent growth event. The probability that the fibril has length ℓ and ends in a blue is given by $Z_{b,\ell}/Z$. Any blue addition will then capture this defect adding $\ell+1$ monomers to the fibril, ℓ from the defect and one from the last blue monomer. The disordered growth rate can then be calculated as (see appendix B for details)

$$R_{dis} = \sum_{\ell=1}^{\infty} k_{b+}(\ell+1)\frac{Z_{b,\ell}}{Z} \tag{2.15}$$

$$= \frac{k_0 q p^2 c^3 e^{-\epsilon_{rb}-\epsilon_{br}}(3-2qce^{-\epsilon_{rr}}-qpc^2e^{-\epsilon_{br}-\epsilon_{rb}})}{(1-qce^{-\epsilon_{rr}}+qce^{-\epsilon_{br}})(1-qce^{-\epsilon_{rr}}-qpc^2e^{-\epsilon_{br}-\epsilon_{rb}})} \tag{2.16}$$

Similarly to the partition function, both $R_{\text{dis}}, R_{\text{ord}}$ are only valid for concentrations smaller than the divergence concentration c_{div}.

2.3 Results

2.3.1 Variable Reduction

This model has seven free parameters $\epsilon_{bb}, \epsilon_{br}, \epsilon_{rb}, \epsilon_{rr}, p, c_0, k_0$, which is too many for experimental comparison. The first reduction we make is by noticing that the only instance of ϵ_{bb} is in the ordered growth term 2.13. Notice that the rest of R_{ord} can be factored out leaving $(pc - e^{\epsilon_{bb}})$. Since the other terms are always positive, this is the only term that determines the x intercept. In many amyloid systems, this x intercept is indistinguishable from 0 making it difficult to accurately determine ϵ_{bb}. For this reason, we set $\epsilon_{bb} = -\infty$ which sets the x intercept to 0. Next, we can eliminate p and q by noticing $q = \frac{qp}{p} = pe^{\ln q/p}$.

This allows us to transform our system using the equations

$$\epsilon_{bb} \to \epsilon_0 \tag{2.17}$$

$$\epsilon_{br} - \ln(q/p) \to \epsilon_1 \tag{2.18}$$

$$\epsilon_{rb} \to \epsilon_2 \tag{2.19}$$

$$\epsilon_{rr} - \ln(q/p) \to \epsilon_2 \tag{2.20}$$

$$pc \to c' \tag{2.21}$$

We categorize these energies based on where they appear relative to the end of an ordered fibril. Since the blue-blue bond only appears at the end of an ordered fibril and never in the disordered cap, we label this with ϵ_0. Since our disordered cap must start with a red monomer since blue-blue bonds are forbidden, our first bond is guaranteed to be a blue-red bond hence ϵ_1. Since our disordered cap after one bond is guaranteed to be red, our second bond can be either red-blue or red-red so we group them together hence ϵ_2. The term $-\ln(q/p)$ added onto energies $\epsilon_{br}, \epsilon_{rr}$ represents the extra entropy provided by the ensemble of red states but is mostly for mathematical convenience. The justification for this grouping of ϵ_{rb} and ϵ_{rr} is that these are weakly bound states resulting in very short binding lifetimes so grouping should not introduce significant errors. The last replacement is the pc grouping as c' which is the average hitting rate of a monomer that will bind as a blue. Replacing all of these variables in equations 2.13 and 2.15 gives

$$R_{\text{ord}}^{\text{reduced}} = \frac{(k_0 pc - k_0 e^{\epsilon_{bb}})(1 - qce^{-\epsilon_{rr}} - qpc^2 e^{-\epsilon_{rb}-\epsilon_{br}})}{1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}} \tag{2.22}$$

$$= \frac{k_0(pc - e^{\epsilon_{bb}})(1 - pce^{-\epsilon_{rr}+\log\frac{q}{p}} - p^2c^2e^{-\epsilon_{rb}-\epsilon_{br}+\log\frac{q}{p}})}{1 - pce^{-\epsilon_{rr}+\log\frac{q}{p}} + pce^{-\epsilon_{br}+\log\frac{q}{p}}} \tag{2.23}$$

$$= \frac{k_0(c' - e^{\epsilon_0})(1 - c'e^{-\epsilon_2} - c'^2e^{-\epsilon_1-\epsilon_2})}{(1 - c'e^{-\epsilon_2} + c'e^{-\epsilon_1})} \tag{2.24}$$

$$R_{\text{dis}}^{\text{reduced}} = \frac{k_0 q p^2 c^3 e^{-\epsilon_{rb}-\epsilon_{br}}(3 - 2qce^{-\epsilon_{rr}} - qpc^2 e^{-\epsilon_{br}-\epsilon_{rb}})}{(1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}})(1 - qce^{-\epsilon_{rr}} - qpc^2 e^{-\epsilon_{br}-\epsilon_{rb}})} \tag{2.25}$$

$$= \frac{k_0 p^3 c^3 e^{-\epsilon_{rb}-\epsilon_{br}+\log\frac{q}{p}}(3 - 2pce^{-\epsilon_{rr}+\log\frac{q}{p}} - p^2 c^2 e^{-\epsilon_{br}+\log\frac{q}{p}-\epsilon_{rb}})}{(1 - pce^{-\epsilon_{rr}+\log\frac{q}{p}} + pce^{-\epsilon_{br}+\log\frac{q}{p}})(1 - pce^{-\epsilon_{rr}+\log\frac{q}{p}} - p^2 c^2 e^{-\epsilon_{br}+\log\frac{q}{p}-\epsilon_{rb}})} \tag{2.26}$$

$$= \frac{k_0 c'^3 e^{-\epsilon_2-\epsilon_1}(3 - 2c'e^{-\epsilon_2} - c'^2 e^{-\epsilon_1-\epsilon_2})}{(1 - c'e^{-\epsilon_2} + c'e^{-\epsilon_1})(1 - c'e^{-\epsilon_2} - c'^2 e^{-\epsilon_1-\epsilon_2})} \tag{2.27}$$

ϵ_1 now represents the strength of the first defect on the fibril and is responsible for determining the height of the plateau. If we assume that the system is in a low concentration regime with $c'^2 = 0$, and limit the size of our disordered cap to a single monomer with $\epsilon_2 = \infty$, then our disordered growth rate disappears and our ordered growth rate reduces to

$$R_{\text{ord}}^{\text{low c'}} = \frac{k_0(c' - e^{\epsilon_0})}{(1 + c'e^{-\epsilon_1})} \tag{2.28}$$

This equation blends between linear fibril growth that does not have any blocking red states ($\epsilon_1 = \infty$) and a system that has a defined plateau.

2.3.2 Simulations

To test the validity of our quasi-equilibrium assumption and variable reduction, we compare the theory to the Kinetic Monte Carlo simulations of our original system as seen in figure 2.2. There are some notable differences between the theory curves and the simulations that we want to address here. The most obvious difference is the divergence in the theory curve at c_{div}. This is because the quasi-equilibrium assumption breaks down and the system starts incorporating infinitely large defects in arbitrarily short periods of time. The simulations show the growth will continue to rise linearly in the high concentration regime because at these concentrations, the hitting rate is faster than the rejection rate of both red and blue monomers. Of our four free parameters, k_0, c_0 are just trivial scaling factors responsible for the stretching of the x and y axes. Analysis of ϵ_1, ϵ_2 shows more interesting behavior that

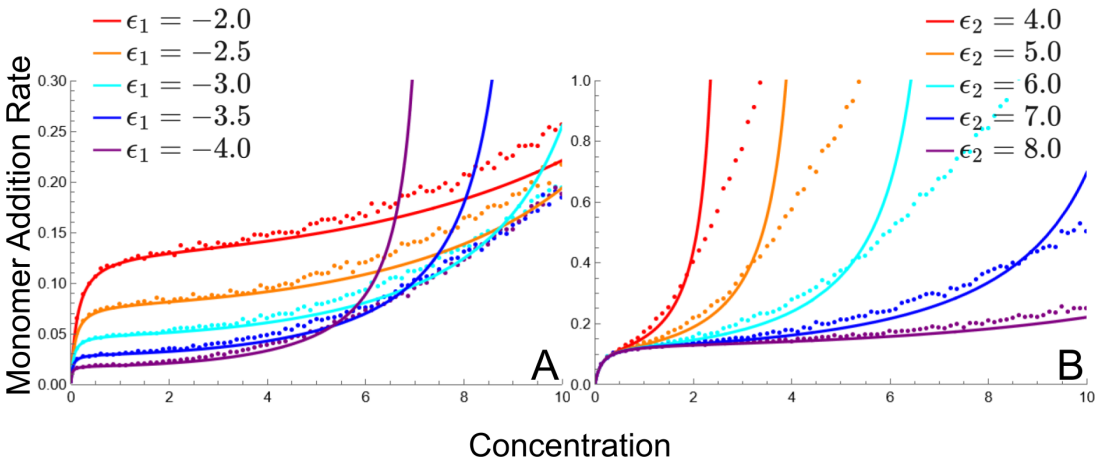


Figure 2.2: *This plot compares the Kinetic Monte Carlo simulations to our reduced theory. All simulations had the default values* $p = 0.5, k_0 = 1, c_0 = 2, \epsilon_{bb} = -10, \epsilon_{br} = -2.0, \epsilon_{rb} = 4.0, \epsilon_{rr} = 4.0$. *In plot A, we modify* ϵ_1 *which redefines* $\epsilon_{br} = \epsilon_1$. *Due to our choice of* p, *there are no added entropic terms* $\log(q/p)$ *for states ending in a red. Similarly plot B redefines* $\epsilon_{rb}, \epsilon_{rr} = \epsilon_2$. *The total time it took for a fibril to grow to 10000 monomers was used to calculate the average growth rate*

is found in figure 2.2. The role of ϵ_1 in our theory is to smoothly interpolate between pure linear growth and the existence of a plateau as observed in equation 2.28. ϵ_1 is primarily responsible for determining the height of the plateau while ϵ_2 is responsible for determining the width of the plateau. The reason why a plateau even exists is because the growth of the system is limited by how fast the first defect given by ϵ_1 is rejected. In addition to the plateau height, ϵ_1 also affects the defect capture rate since it can represent third and higher order defects on the disordered cap. The divergence occurs earlier due to the red-blue bonds increasing the partition function of the disordered cap. Analysis of the effect of ϵ_2 shows that it only affects the disordered growth rate. This is because red-blue and red-red defects can only appear as a second or higher order defect leaving the plateau height unaffected. Taking the limit as $\epsilon_2 \to \infty$, the growth rate exactly reduces to those found in previous work[6]. Making ϵ_2 more energetically favorable increases the partition function of the disordered cap which increases the disordered growth rate.

2.3.3 Experimental Comparison

There are several experiments that measure the elongation of fibril systems. Some examples include Aβ, αSyn, and Insulin. We classify these systems into the categories of more and less accepting of defects. Systems that are less accepting of defects have a strong observed plateau such as the Aβ and αSyn systems observed in figure 2.3. Other systems are more accepting of defects and will still form fibrils even if the monomers are not aligned geometrically. Our simulations predict that defect tolerant systems of this form will not show a significant plateau but will continue to increase linearly in the high concentration limit. One system that we claim has evidence of this property are the fibrils formed by insulin as observed in figure 2.4. Our theory can smoothly interpolate between both cases. There are four prominent features on a generalized growth curve that can be used for curve fitting including the x intercept, the plateau height, the divergence concentration, and the roundness

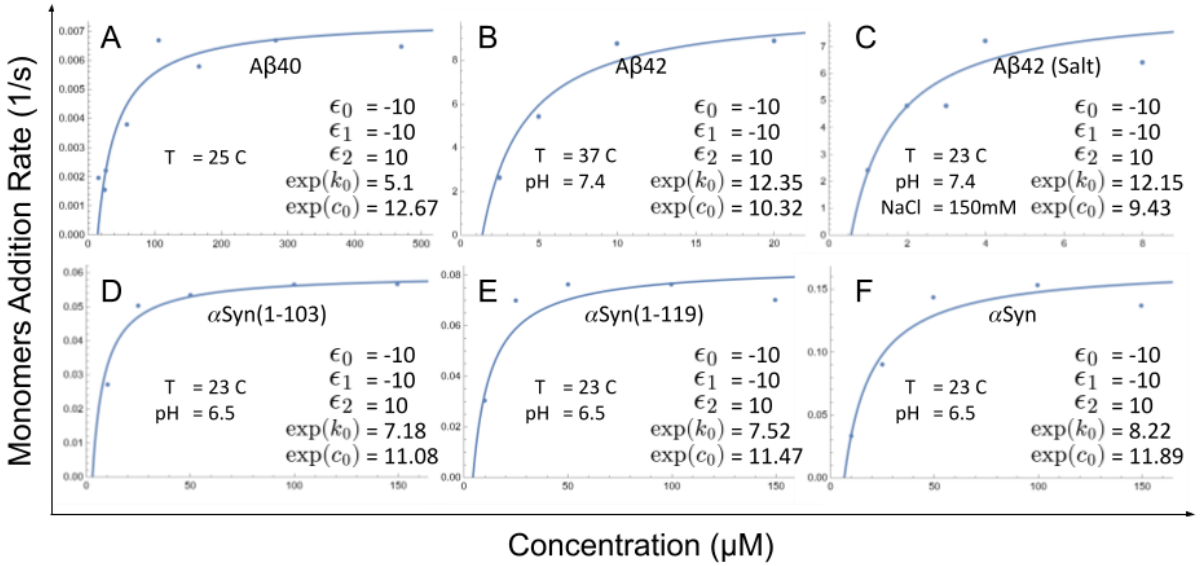


Figure 2.3: *Here are the fitted plots for several experiments involving 1D fibril growth for Amyloid Beta and Alpha Synuclein proteins*[1,2,3,4]*. The data was compiled and normalized by the Knowles Group where the Aβ data was collected through labeled and unlabeled optical microscopy and αSyn was collected through ThT assays. Our fitting suggests that the disordered growth of these systems is negligible since* ϵ_2 *is very large effectively disallowing red-red and red-blue binding states.*

of the initial linear to plateau transition. We can fit each of these regions independently of the other regions since each only some variables affect some features of the graph. The x intercept is determined mostly by ϵ_0 found in the $c' - e^{\epsilon_0}$ term of the ordered growth rate. The plateau height is determined by either ϵ_1 or k_0 since ϵ_1 shifts the plateau up or down and k_0 stretches the entire graph along the y axis. The divergence concentration is determined by either ϵ_2 or c_0 since ϵ_2 causes the disordered growth to diverge earlier but c_0 stretches everything along the x axis. Each of these three fitting parameters can be fitted mostly independently of one another. The roundness of the linear to plateau region is the final fitting parameter which can be achieved by varying ϵ_1 and k_0 simultaneously such that the plateau remains unchanged but the curvature changes. Another way of fitting the roundness is by varying ϵ_2 and c_0 simultaneously in a manner that keeps the divergence concentration c_{div} fixed. We can fit all of the relevant features of both defect tolerant and intolerant systems.

The variable fitting provides some intuition into the microscopic behavior of the experimental systems described in figures 2.3 and 2.4. If disordered growth is highly suppressed with $\epsilon_2 \gg 1$, and the first defect on an ordered fibril is stable $\epsilon_1 < 0$, then there is a significant plateau observed as in figure 2.3. Such a system can still have defects but in very low quantities that will not meaningfully affect the growth curves for experimentally relevant concentrations. As a system becomes more tolerant of defect states $\epsilon_2 \approx 1$, the plateau is replaced by a second linear growth regime as seen in 2.4B. We believe that the macroscopic observed differences between figures 2.3 and 2.4 is primarily due to the amount of defects that are incorporated into the fibril in each system. The experimental defect rates measured in Aβ fibrils range from 1/99 to 1/834 which we believe qualifies as a low defect rate[7]. In our simulations, a defect rate of $5 - 10\%$ is required to observe the transition away from a plateau into a new linear growth regime. Defect rates are not commonly measured in these systems and should be investigated further. One reason why this phenomena is not observe

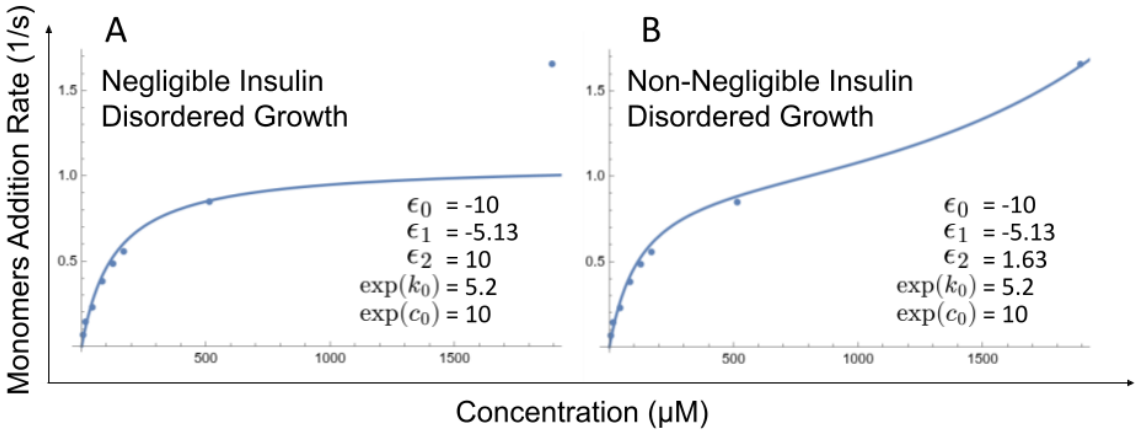


Figure 2.4: *The normalized growth rate of insulin fibrils measured by Quartz Crystal Microbalance*[5]*. A shows the line of best fit in the ordered growth limit* $\epsilon_2 \gg 1$ *which is identical to the registry search model*[6]*. The incorporation of disordered growth in figure B can now account for previously outlying points suggesting the importance of disordered growth microscopically.*

many real systems is that other mechanisms of aggregation may be more favorable than a strict 1D structure in the high concentration limit. In this case, no second linear growth regime would ever be observed since our model requires the assumption of a 1D fibril. Every fibril system will have defects as a matter of statistical mechanics, but our theory predicts that fibril systems without a strongly observable plateau are more tolerant of defects.

2.3.4 Composition Estimates

The composition of the fibril can play an important role in secondary nucleation. Defects in a fibril can serve as a template for fibril nucleation increasing the number of fibrils in solution. Former models have predicted defect rates before but not in a manner that can also accommodate the disordered growth observed in insulin[6]. To estimate the fibril composition,

we assume that our system is in the low concentration limit. This allows us to assume that all defects added are of the form red-blue-blue as well as assume that ordered growth is much much greater than disordered growth. This lets us calculate the fibril composition as $R_{\text{ord}}^{\text{reduced}}$ and $R_{\text{dis}}^{\text{reduced}}$. The portion of the fibril that is ordered is given by the equation

$$P_{\text{ordered}} = \frac{R_{\text{ord}}^{\text{reduced}} + \frac{2}{3}R_{\text{dis}}^{\text{reduced}}}{R_{\text{ord}}^{\text{reduced}} + R_{\text{dis}}^{\text{reduced}}} \tag{2.29}$$

$$= 1 - \frac{\frac{1}{3}R_{\text{dis}}^{\text{reduced}}}{R_{\text{ord}}^{\text{reduced}} + R_{\text{dis}}^{\text{reduced}}} \tag{2.30}$$

$$\approx 1 - \frac{R_{\text{dis}}^{\text{reduced}}}{3R_{\text{ord}}^{\text{reduced}}} \tag{2.31}$$

$$= 1 - \frac{c'^2 e^{-\epsilon_2-\epsilon_1}(3 - 2c'e^{-\epsilon_2} - c'^2 e^{-\epsilon_1-\epsilon_2})}{3(1 - c'e^{-\epsilon_2} - c'^2 e^{-\epsilon_2-\epsilon_1})^2} \tag{2.32}$$

This result is verified by the simulations found in figure 2.5. The coefficient of 2/3 in the disordered growth term comes from counting the number of blue monomers in a red-blue-blue defect capture event. We hope to use this result for the estimation of secondary nucleation rates

2.4 Discussion

2.4.1 Fibril Stability

In some cases, our models can be used to measure the stability of a fibril. If there is a data set with a measurable x intercept in the rate vs concentration graph, then the blue-blue binding affinity can be determined. We define a saturation concentration c_{sat} where the fibril neither grows nor shrinks on average which is given by the x intercept. This can be approximately calculated through the inspection of R_{ord} since for very low concentrations

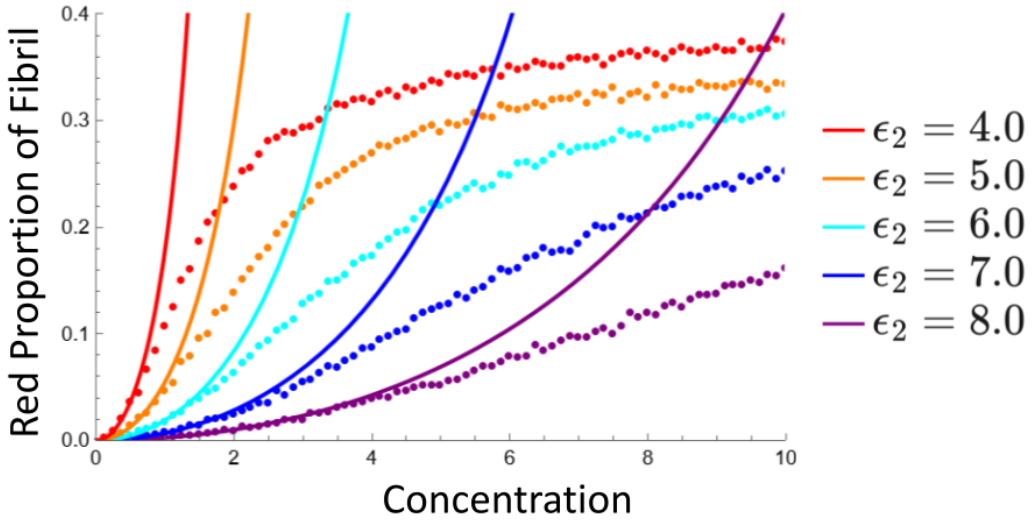


Figure 2.5: *The defect rate of a simulated fibril grown to 10000 monomers as a function of concentration. The simulations were conducted with* $\epsilon_0 = -10, \epsilon_1 = -2, k_0 = 1, c_0 = 2, p = 0.5$ *which are the exact same conditions found in figure 2.2B. The theory curves are calculated as* $1 - P_{ordered}$ *and is only valid for low concentrations. When the defect rates are below 5%, which experiments show is a plausible value*[7]*, the theory curve perfectly matches the simulation. In the high concentration limit, the fibril proportions converge to a fixed value.*

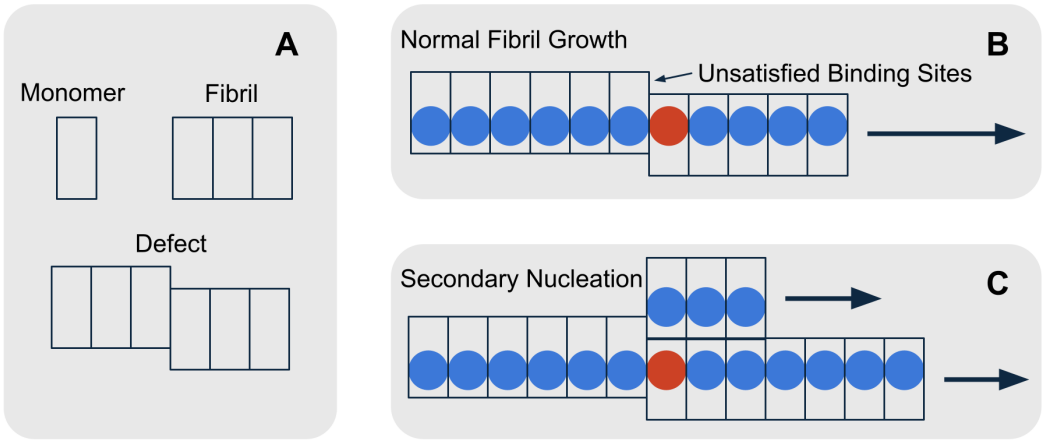


Figure 2.6: *A gives a toy model for the geometry of a monomer, fibril, and a fibril defect. B shows how this toy model maps onto our red blue model. Red monomers represent defects with respect to the previous monomer. C gives a geometric example of how new fibrils might be nucleated. The misaligned registry of a defect monomer can serve as a template for the nucleation of new fibrils.*

R_{dis} is nearly 0. The equation governing c_{sat} is

$$(k_0 p c_{\text{sat}} - k_0 e^{\epsilon_{bb}}) = 0 \tag{2.33}$$

solving for c_{sat} gives

$$c_{\text{sat}} = \frac{e^{\epsilon_{bb}}}{p} \tag{2.34}$$

If c_{sat} is measured, then ϵ_{bb} can be solved. One issue with this approach is that the error increases as the intercept approaches 0. The closer to 0 the intercept is, the more uncertain we are in the binding affinity. Most of the systems that we analyzed have an x intercept that is near 0 which is why we have set the fitting parameters to -10 in figures 2.3 and 2.4. With the data sets currently available to us, we can not extract binding affinities with any degree of certainty.

2.4.2 Secondary Nucleation

An important concept related to fibril growth is secondary nucleation. Recent evidence shows that defect sites in amyloid systems serve as secondary nucleation sites[7]. In our model, this manifests itself in our model as the boundary between ordered and disordered growth. The analysis of figure 2.6 offers a geometric argument for how secondary nucleation might occur. If this model were true, red monomers which represent fibril defects can serve as a site for secondary nucleation. If this is the case then the nucleation rate of new fibrils is proportional to the number of defects present on already existing fibrils

$$R_{\text{nuc}} \propto 1 - P_{\text{ordered}} \tag{2.35}$$

The reality of secondary nucleation is likely more complex. We assume that all red monomers can cause secondary nucleation but there could be arguments made that only red monomers near a long chain of blue monomers can nucleate. Even if this condition is imposed, the equation listed previously would still be valid in the low concentration regime. Some experiments measure the defect rates of the fibrils between 1 in 99 to 1 in 834[7].

2.4.3 N State Models

The math we developed to generalize the fibril growth problem can be used to solve any N state model where N is the number of distinct states (colors) a monomer can take on the fibril. An N state model will have N^2 free variables for the energies, $N-1$ free variables for the probabilities of each color, and 2 free variables for the reference rate and reference concentration scaling factors. For a 3 state model, the maximum number of free parameters required is 13 which is far too many for an accurate fit. Still though, we analyzed a 3 state model including Native, Misaligned, and Nucleating states where nucleating is a special defect that can foster the growth of secondary fibrils. This model had issues where it

diverges much faster than the 2 state model and had a much more inconsistent plateau with worse fits. Because of this, we decided to stick with the 2 state model.

Chapter 3

Nonspecific Binding with RNA Increases Folding Rate

3.1 Introduction

Unlike the self-assembly of the previous section, there are many systems within a cell that require help in the form of chaperones to assemble. One class of systems that requires chaperones for assembly is structured RNA[14,15,16,17,18]. The role of structured RNA within the cell is to fold into a useful conformation that can carry out important processes such as protein production. Examples include transfer RNA or ribosomal RNA. The free energy landscape an RNA must traverse to reach its native conformation is highly frustrated with many kinetic traps that require the help of chaperones[19,20,21,22,23]. These chaperones operate by destabilizing secondary structures forme by the RNA including both misfolded and native structures, or by locking in native conformations. Some chaperones require external energy in the form of ATP known as active chaperones. Other chaperones do not require ATP and are known as passive chaperones[24,25,26,27,28]. Passive chaperones operate by binding many chaperones to the RNA at once, competing for bonds that would otherwise be occupied

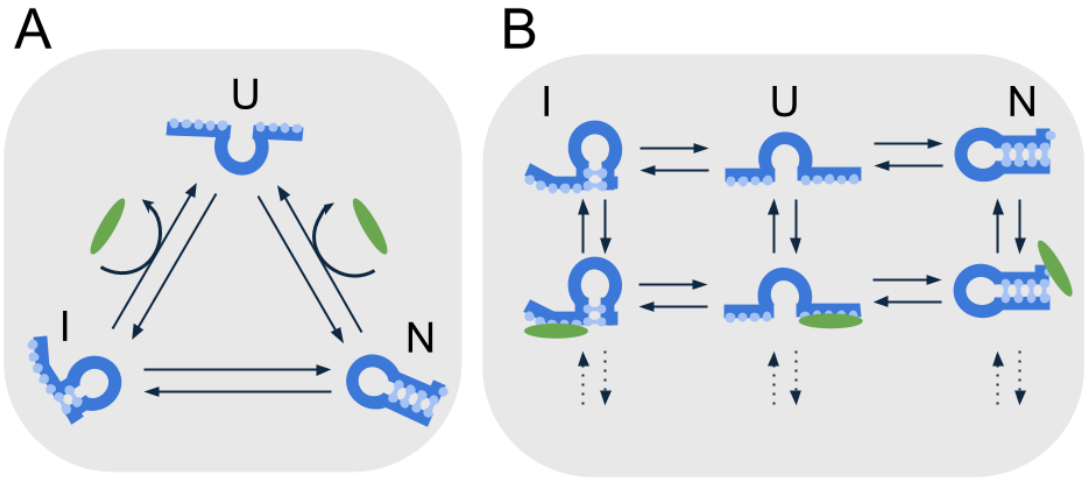


Figure 3.1: *Figure A shows the Iterative Annealing Model (IAM) for RNA annealing. There are various forms of IAM models, but they broadly operate by unfolding the RNA repeatedly destabilizing both native and misfolded states. This is often a faster process than the RNA unfolding naturally by itself since the rate of attempts increases. Figure B shows our orthogonal binding model where the chaperone binding and the RNA folding are treated as separate independent processes composed into a single state space. This model can accommodate an arbitrary number of chaperones at the cost of more complexity.*

with normal double stranded RNA bonds[29,30,31,32]. This entropically buffers the unfolded intermediate states, lowering the effective free energy reaction barrier. We offer a mechanistic model for entropic buffering by a chaperone that destabilizes secondary structures and allows polymers such as RNA to anneal without the use of ATP.

3.1.1 The Iterative Annealing Model

The gold standard for RNA annealing is the iterative annealing model (IAM)[33,22,19]. A cartoon of the IAM can be found in figure 3.1A. The IAM models a 3-state system with misfolded, native, and unfolded/intermediate states. Chaperones can interact with these states only by bringing misfolded/native states to the unfolded state. This is a reduced

view of the IAM but captures the most important features. The best feature of this model is that it has very few fitting parameters and is able to capture how chaperone concentration affects the RNA folding rate. The primary issue with this model for our purposes is that it does not allow RNA to fold while bound to the chaperone. This model also implicitly assumes that only one chaperone binds at a time. Our proposed solution for this issue is to introduce a new orthogonal binding model that can accommodate many bound chaperones simultaneously.

3.2 Model

3.2.1 RNA as a Zipper

One method for representing the RNA is as a zipper[34]. This model treats base pairs as discrete steps that encompass their own local free-energy minima. This model is easy to work with and sufficiently detailed to model the complex dynamics of RNA. Our zipper model works by breaking bonds in order from the free tails to the hairpin structure. Other zipper models allow for RNA bonds to break from both the hairpin and the free tail directions. We choose to work with the simpler model since it is already complex enough to explain how chaperones operate. This model assumes that all RNA bonds up to state i have some free energy $F_r(i)$ in units of k_BT where i is the order parameter that represents how far the zipping has progressed. This allows for individual base pairs to have varying energies depending on it's sequence. We let the order parameter i represents how many base pairs have formed in the zipper where $F_r(i)$ is the total free energy of all i bonds. We assume the rate of forming an RNA bond is uniformly equal to k_r regardless of the folding state or RNA composition. To calculate the bond breaking rates, we assume a detailed balance condition requiring a thermal fluctuation greater than the difference in free energies. This

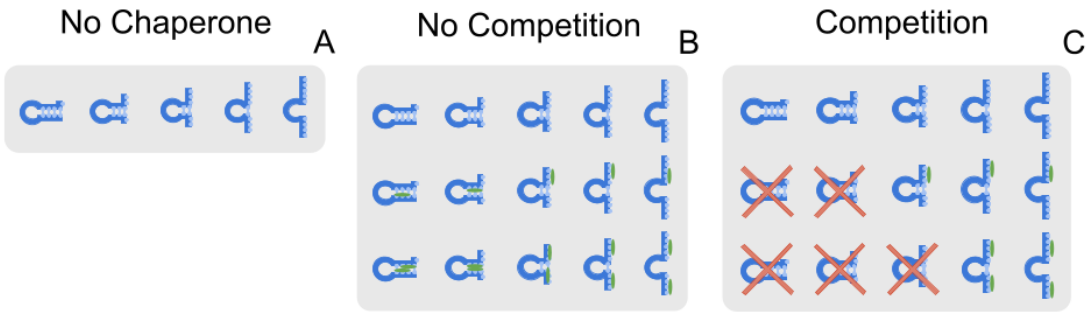


Figure 3.2: *Figure A shows what the state space of a single RNA fold (misfold or native) appears like in the absence of a chaperone. Figure B is a toy model of an RNA fold in the presence of a molecule that binds to RNA, but does not compete for ssRNA bonds. The RNA binding reactions are independent of the chaperone binding reactions, giving the same first passage time from all left states to all right states as figure A. Figure C is the state space of an RNA fold where the chaperone competes for ssRNA bonds. This results in some states becoming impossible to reach since dsRNA bonds and RNA-chaperone bonds are mutually exclusive*

fluctuation must have at least energy $\Delta F_r(i)$. This gives us the equations

$$k_{r+}(i) = k_r \tag{3.1}$$

$$k_{r-}(i) = k_r e^{-\Delta F_r(i)} \tag{3.2}$$

Here k_{r+} is the rate of forming an RNA bond $i \to i+1$ and k_{r-} is the rate of breaking an RNA bond $i \to i-1$

3.2.2 Modeling Chaperones

We define passive multiple-binding chaperones as any molecule that increases the folding rate of RNA without the use of ATP by binding many at once. Chaperones of this category operate by entropically buffering the unfolded state, making the unfolded state more thermally stable. This lowers the reaction free energy barrier allowing thermal fluctuations to

more easily kick the RNA out of kinetic traps. Chaperones that have this behavior include HFQ, Csp, StpA, NSP2, and more[29]. For a chaperone to buffer the unfolded state, it must bind to single stranded RNA more favorably than double stranded RNA. This allows the chaperone to compete for hydrogen bonds from exposed nucleic acids. A visual proof of this mechanism is given by figure 3.2B showing that chaperones that bind to single and double stranded RNA equally do not create a bias towards the unfolded state. An increase in folding rate only happens if there is a biased entropic buffering towards the unfolded state. We introduce the chaperone into our model by assuming that our system consists of a single RNA in a bath of chaperones. This is sometimes referred to as the semi-canonical ensemble, since the chaperone population is allowed to vary but the RNA population is not. Since we exist in a bath of chaperones, multiple chaperones can bind to the RNA at once. Depending on the size of a chaperone, this puts an upper limit on how many chaperones can bind at once. The maximum number of chaperones that can bind to the RNA n depends on the folding index i and the size of the chaperone. This is because the RNA folding index determines how many base pairs are currently single stranded. If the RNA is unfolded, there is a maximum number of chaperones that can bind to the RNA given by N. We can construct a free energy for chaperone binding $F_c(i,n)$ by combining the chemical potential μ_c, binding energy ϵ_c, and the degeneracy of the binding configurations $\Omega(i,n)$. It should be noted that the free energy $F_c(i,n)$ is the free energy of only the RNA bonds and the degeneracy. For the kinetics, we assume that the chaperone binding rate is proportional to the chaperone concentration in solution e^{μ_c} with some reference rate k_c. Applying the detailed balance condition again, we can construct the binding rates as follows

$$k_{c+}(n) = k_c e^{\mu_c} \tag{3.3}$$

$$k_{c-}(n) = k_c e^{-\Delta F_c(i,n)} \tag{3.4}$$

k_{c+} is the rate at which a chaperone leaves the bath and enters the system, binding to the RNA and increasing the number of chaperones bound $n \to n+1$. k_{c-} is the rate at which a thermal fluctuation breaks an RNA-chaperone bond and the chaperone leaves the system, decreasing the number of chaperones bound $n \to n-1$. Combining this chaperone system with the RNA system allows for 4 possible kinetic events, RNA binding/unbinding k_{r+}, k_{r-}, and chaperone binding/unbinding k_{c+}, k_{c-}.

3.2.3 Chaperone Coordinate Grouping

The key assumption of our model is that the rate at which chaperones bind/unbind k_{c+}, k_{c-} happens on a much shorter timescale than RNA binding/unbinding k_{r+}, k_{r-}. We can then make a quasi-equilibrium assumption where we assume that all states with RNA folding index i obey an equilibrium distribution in the semi-canonical ensemble $\mathcal{Z}_i$ with ground state $e^{-F_r(i)}$ given by

$$\mathcal{Z}_i = \sum_{n=0}^{N} e^{(-F_c(i,n)+n\mu_c)/k_B T} \tag{3.5}$$

This statistical ensemble is a grand partition function with respect to the chaperone number. We can also calculate the partition function of the entire system as

$$\Xi = \sum_{i=0}^{I} \sum_{n=0}^{N} e^{(-F_r(i)-F_c(i,n)+n\mu_c)/k_B T} \tag{3.6}$$

$$= \sum_{i=0}^{I} e^{-F_r(i)} \mathcal{Z}_i \tag{3.7}$$

To truly convert our 2D system to a 1D system, we must renormalize the kinetic rates into new binding and unbinding rates given by k_+, k_-. Compared to a system without chaperones, we can not simply assume that the folding rate of RNA is k_r because this ignores cases when a chaperone blocks the folding of an RNA base pair. We can remedy this by considering all possible RNA binding events starting with folding index i weighted by

their probability, and counting how many folding events towards index $i+1$ are successful. This process gives the following equation which can be interpreted as the probability that an RNA folding event is physically blocked by a chaperone already bound to one of the base pairs attempting to form a bond.

$$\sum_{n=0}^{N} \underbrace{\frac{e^{(-F_c(i,n)+n\mu_c)/k_BT}}{\mathcal{Z}_i}}_{\text{Weighted State}} \underbrace{\frac{\Omega(i+1,n)}{\Omega(i,n)}}_{\text{Success Probability}} \tag{3.8}$$

$$= \sum_{n=0}^{N} \frac{e^{(-F_c(i,n)+n\mu_c)/k_BT}}{\mathcal{Z}_i} \frac{e^{\log(\Omega(i+1,n))} e^{(-n\epsilon_c+n\mu_c)/k_BT}}{e^{\log(\Omega(i,n))} e^{(-n\epsilon_c+n\mu_c)/k_BT}} \tag{3.9}$$

$$= \sum_{n=0}^{N} \frac{e^{(-F_c(i,n)+n\mu_c)/k_BT}}{\mathcal{Z}_i} \frac{e^{(-F_c(i+1,n)+n\mu_c)/k_BT}}{e^{(-F_c(i,n)+n\mu_c)/k_BT}} \tag{3.10}$$

$$= \sum_{n=0}^{N} \frac{e^{(-F_c(i+1,n)+n\mu_c)/k_BT}}{\mathcal{Z}_i} \tag{3.11}$$

$$= \frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} \tag{3.12}$$

The probability of adding a base pair bond is determined by the degeneracy of the state the system is currently in $\Omega(i,n)$, as well as the state it is entering $\Omega(i+1,n)$. In other words, if our current state at (i,n) has $\Omega(i,n)$ possible states and the system is attempting to move to state $(i+1,n)$ with $\Omega(i+1,n)$ possible states, exactly $\Omega(i+1,n)$ attempts succeed. Since the amount of single stranded RNA for state $i+1$ is less than that of i, the ratio of grand partition functions is guaranteed to be less than 1 so we can interpret it as the probability of successfully forming an double stranded RNA bond. The opposite process of RNA unbinding is trivial since our model assumes that the chaperone can not block an unbinding event. This leaves the RNA unbinding rate unchanged as given by equation 3.2.

The kinetic rates of the reduced model is then

$$k_+(i) = k_r \frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} \tag{3.13}$$

$$k_-(i) = k_r e^{-\Delta F_r(i)} \tag{3.14}$$

This system is almost identical to the RNA only system with the only difference being that the RNA binding term is slowed by a probability term. This probability term represents the probability that the chaperone is physically preventing the formation of an RNA-RNA bond. The result shows that the chaperone does not increase the unfolding rate, rather, it slows the folding rate to both the misfolded and native states. By slowing the folding rate, it slows the rate at which the system regresses back to a misfolded state after a thermal fluctuation has made progress towards the native start. This leads to the apparent lowering of the free energy reaction barrier.

3.2.4 1D Random Walk First Passage Time

The 1D zipper is the same as a discrete 1D random walk over continuous time. Random walks are usually thought of with probabilities and residence times so we must first convert our rate equations accordingly. The zipper probabilities and times can be calculated from the reduced rates given by equations 3.13 and 3.14 giving

$$p_i = \frac{k_+(i)}{k_+(i) + k_-(i)} \tag{3.15}$$

$$q_i = \frac{k_-(i)}{k_+(i) + k_-(i)} \tag{3.16}$$

$$r_i = \frac{1}{k_+(i) + k_-(i)} \tag{3.17}$$

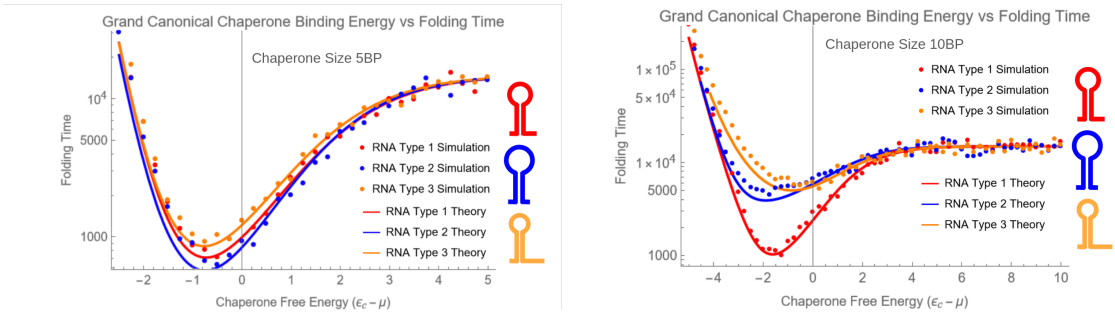


Figure 3.3: *These plots show the folding rates of 3 different RNA strands of 30 base pairs forming helices of the same length but varying loop size and free tail length. The plots vary only in chaperone size with the left plot having a footprint of 5 base pairs and the right having a footprint of 10 base pairs.*

The exact mean first passage time of a 1D system of any length is known and its derivation can be found in appendix A. The mean first passage time is given by

$$T_{\text{MFPT}} = \underbrace{r_0 \sum_{i=0}^{N} \prod_{j=1}^{i} \frac{q_j}{p_j}}_{\text{Time Spent at the Start } T_s} + \underbrace{\sum_{i=1}^{N} \sum_{j=1}^{i} \left(\prod_{k=j+1}^{i} \frac{q_k}{p_k} \right) \frac{r_j}{p_j}}_{\text{Time Spent in the Middle } T_m} \tag{3.18}$$

Any theory curves plotted in figures are derived from this equation.

3.3 Simulations and Experiments

3.3.1 Simulations

To test the validity of our approximations, we have generated a Kinetic Monte Carlo simulation of the 2D system to compare with the 1D theory. Our simulation takes in constants such as RNA size/structure/binding energy/rate, and chaperone size/chemical potential/binding energy/rate to apply to the equations 3.1-3.4. We then simulate thousands of trajectories from the misfolded state to the native state to get a mean first passage time. By varying parameters such as chaperone size or RNA structure, we can discover interesting emergent

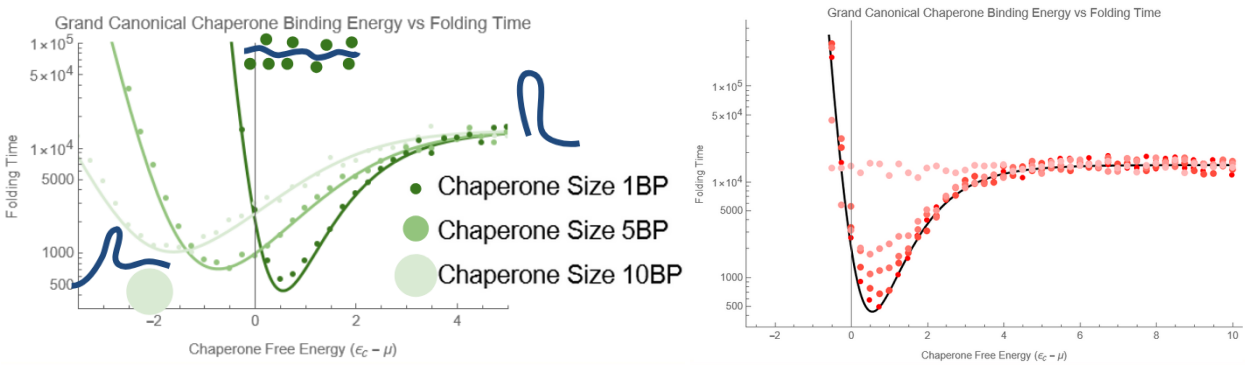


Figure 3.4: *These plots show how varying the chaperone properties affects the first passage times. By changing the size of the chaperone, we can change the properties of the graph as seen in the left hand graph. In the right hand graph, we change the chaperone binding rate. As the chaperone slows, the probability a chaperone binds at all becomes arbitrarily small leading to the total graph converging to the RNA only solution. It appears like we can recover the chaperone solution by interpolating between the fast chaperone and the RNA only curves.*

properties of the system. When we simultaneously vary the structure of the RNA and the size of the chaperone as seen in figure 3.3, we can see that specificity starts to arise from purely non-specific interactions. What is happening on the right hand side figure is that larger chaperones have a harder time finding a contiguous chunk of RNA to bind to. This leads to RNA with longer chunks available for the chaperone having a larger degeneracy. However, a large degeneracy does not mean that the chaperone is working better, in fact, the red RNA shows the best chaperoning behavior. This is because it must unfold halfway before the chaperone can even bind. This is a better situation compared to the yellow RNA where the chaperone can bind almost immediately after the simulation starts and not save any progress. Another interesting feature of these diagrams is that small non-specific chaperones do not have the emergent specificity like the large ones do. This is because they do not need to wait for a large contiguous chunk of RNA to bind to so they can almost always bind.

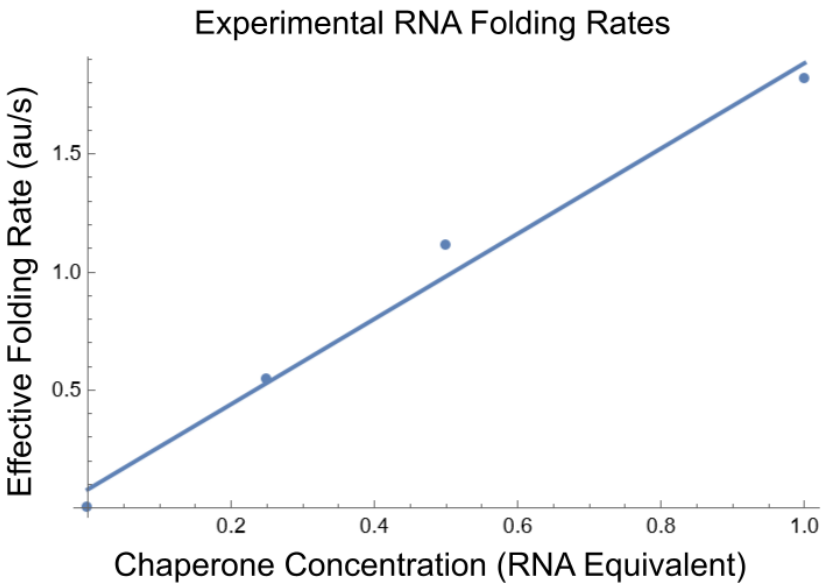


Figure 3.5: *This figure plots the folding rate of an RNA with a synthetically developed chaperone which is just a truncated form of the original RNA. The RNA was fret labeled so the native state will fluoresce and the misfolded state will not. By varying the chaperone concentration and measuring the florescence as a function of time, we can fit time dependent data to an exponential curve giving us the reaction rate. The concentration of the chaperone is measured as a proportion of RNA concentration.*

3.3.2 RNA as a Chaperone Experiment

We have had a difficult time finding RNA folding experiments with many different chaperone concentrations. To prove our theory, we need data that plots the RNA folding rate against the chaperone concentration for many different concentrations. The data that we have found usually contains too few data points to reliably test our model so we only compare to two experiments here. The first experiment shows a remarkable similarity to our simulations even though it only contains 4 data points[35]. This system consists of a synthetic RNA engineered to have 2 free energy wells representing the misfolded and native states. The RNA is fret labeled so the solution will fluoresce when it reaches the native state. The chaperone is also a synthetic RNA designed so it can partially unfold the target RNA through entropic buffering. The misfolded state contains two small hairpins and the native state contains one

large hairpin in a very similar manner to our simulations. The chaperone binds to one of the two smaller hairpins leaving it completely unfolded until the second hairpin can unfold naturally. This system perfectly maps onto our model since this process does not require ATP and we know that there can only be 1 chaperone bound to the RNA at a time due to their construction. We extract the data in figure 4 of this paper and fit the curves to a first order reaction curve $A - B\exp(-k[c])$. After fitting, we extract the rate constants k leading to our figure 3.5. If we assume that the chaperone RNA binds with the same affinity as the RNA hairpin by itself, then our model predicts a slope of 1. Our actual calculated slope is 1.805. This difference could be attributed to our assumption that the chaperone binding affinity is stronger than the RNA by itself due to the extra conformational entropy gained by the newly free tail.

3.3.3 Protein Chaperone Spy Experiment

Although our theory has been centered around RNA, it should also be applicable to proteins. The only assumption we have made is that there is some order parameter i that represents how far folding has progressed. It should be possible to construct an order parameter for a protein that models how much folding has progressed. Proteins also have non-specific chaperones with one example being the Spy protein. Spy is considered to be a non-specific passive protein chaperone. This chaperone can be bound to the protein while folding[8] just as in our model. The folding rate of the protein was measured with various concentrations of Spy present shown in figure 3.6. Just like our theory predicts for RNA, there is an optimal concentration for chaperoning. In the low concentration regime the chaperone increases the folding rate of proteins by destabilizing misfolded structures. In the high concentration regime, the chaperone decreases folding rates by interfering with the formation of native bonds.

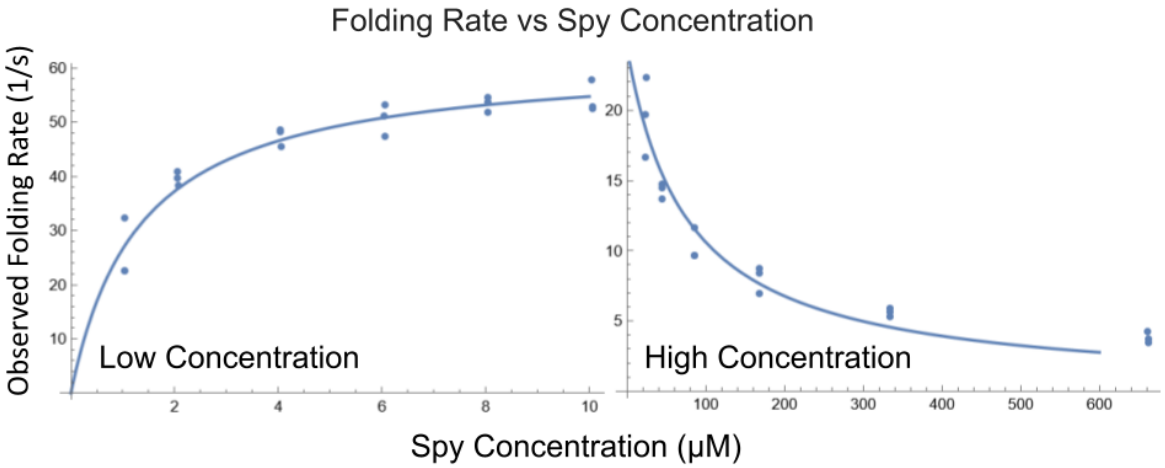


Figure 3.6: *Plots of the extracted data from Stull 2015*[8]*. The measured folding rate of a protein when in the presence of a non-specific chaperone Spy. There is an optimal concentration of spy somewhere between 10μM and 100μM. Initial addition of a nonspecific chaperone will increase the folding rate with diminishing returns. Excessive chaperones will hinder the folding of proteins.*

3.4 Discussion

3.4.1 Dominant Timescales

Multiple-binding chaperones have two dominant timescales. The first timescale is determined by how deep the free energy basin of the misfolded state and is the rate limiting step for chaperones of low concentration. Conversely, the rate limiting step when there are an excess of chaperones is being able to remove the chaperones. This can be seen in equations 3.13 and 3.14, where the folding probability can become so slow that the unfolding rate of the RNA is higher than the folding rate. This stage is not biologically relevant since this is equivalent to a new global free energy minima but could be relevant in laboratory settings. It turns out that the fastest rate of folding is when the free energy landscape is the flattest. If the free energy landscape rises approximately linearly, then the fastest folding rate happens when the on and off rate are equal because the free energy landscape is completely flat here. This condition is given by

$$\frac{\mathcal{Z}_{i+1}}{\mathcal{Z}_i} = e^{-\Delta F_c(i)} \tag{3.19}$$

Since we can approximate an RNA as a zipper whose free energy rises linearly as it unzips, this should be a valid approximation. If we also make the assumption that the RNA has a discrete number of non-overlapping binding sites, then we can calculate the grand partition function on a per binding site basis. So for chaperone length L, we get a binding site partition function of $(1 + e^{-(\epsilon_c - \mu_c)})$ with available binding sites $(i_0 - 2i)/L$. Here i_0 is the total number of single stranded bases on the unfolded RNA and i is the length of the current helix in base pairs. Combining everything gives

$$\mathcal{Z}_i = (1 + e^{-\epsilon_c + \mu_c})^{(i_0 - 2i)/L} \tag{3.20}$$

Applying this relation to equation 3.19, we get

$$(1 + e^{-\epsilon_c + \mu_c})^{(-2/L)} = e^{\epsilon_r} \tag{3.21}$$

This equation tells us how strongly the optimal chaperone with an RNA binding footprint of L should bind to the RNA for the fastest annealing. Any chaperone more attractive than this will create a new global minima for the system, making the native state thermodynamically unfavorable. It should be noted that ϵ_c is the energy of the entire chaperone binding, not per base pair. In reality though, it is likely that the non-specific binding sites do overlap leading to the grand partition function making this only an approximation. This is in agreement with figure 3.4 where the absolute minima changes with chaperone length. In the case of the right hand graph. This predicts that the best chaperones should bind strongly if they are large, and weakly if they are small.

3.4.2 OBM Explains Chaperone Condensates

RNA chaperones that bind simultaneously tend to form condensates[36,32,37]. Some experiments even show that RNA folds better in condensates[32]. This is is in agreement with equation 3.19 where the rate of annealing increases modulate the chaperone concentration to a near optimal concentration for RNA folding. We propose that these condensates could be acting like reaction vessels where they localize many chaperones, increasing the effective concentration, so that RNA can be annealed. Once the RNA secondary structure is formed, there would be fewer single stranded bonds for the chaperones to bind to allowing the RNA to leave the condensate easily too. The OBM qualitatively matches what is observed with multiple-binding chaperones that form condensates.

3.4.3 Advantages Over IAM

The orthogonal binding model has two major advantages over the IAM model. The first advantage is that it can model multiple-binding chaperones. The second is that can predict behavior that the IAM can not, namely the optimal binding energy for a chaperone. Analysis of the IAM shows that the unfolding rate from M to U happens with a rate proportional to the concentration of RNA in the state M times the concentration of the chaperone. This gives an overall rate proportional to $[M][C]$. This rate is monotonic increasing with chaperone concentration implying that chaperones can only help. Both the experiments from Mayer 2007 and our Monte Carlo simulations show that after a certain point, chaperones start slowing the folding of RNA. The orthogonal binding model predicts that anything that preferentially binds to single stranded RNA will chaperone it if its concentration is modulated correctly.

Chapter 4

The Magic Number Effect of Phase Separated Systems

4.1 Introduction

Liquid liquid phase separation (LLPS) has been studied extensively in the field of biophysics. LLPS occurs when differing chemicals become/are immiscible and naturally separate from each other. LLPS condensates have many potential roles such as concentration buffering, reaction vessels, self organization, molecular storage, and many more[38]. LLPS RNA chaperone examples include HFQ and NSP2 condensates as well as the Nucleolus which all behave like a reaction vessels for RNA folding[32,39,36,40]. Another example is the chromatin structure in the nucleolus that organizes DNA in an organized phase separated structure to prevent the DNA from tangling. LLPS structures are ubiquitous in the cell and lead to higher order structure within the cell often without the use of ATP. In this chapter, we discuss the Magic Number Effect and how it applies to the phase separation of H1 and DNA related to the chromatin structure in the cell.

4.2 Re-Entrant Phase Diagrams in Binary Systems

4.2.1 Solubility Product and Ion Product of Two Component Systems

Chemists describe the solubility and precipitation of salts using the ion product and the solubility product. The solubility product is a constant that defines how saturated a solution of dissolved ions is and whether that solution will undergo spontaneous precipitation. Similarly to the precipitation of salt, if biological molecules reach a high enough concentration, they will undergo liquid-liquid phase separation. The solubility product for predicting the liquid-liquid phase separation of charged polymers too with some modification. The solubility product of a solution is defined as

$$K_{sp} = [A][B]^r = AB^r \tag{4.1}$$

where [A] or A represents the concentration of A monomers at the phase boundary and [B] or B represents the concentration of B monomers at the phase boundary. r is the optimal ratio of A's to B's in the dense phase and does not need to be an integer. Although the concentrations of A and B refer to the dilute phase, the solubility product can be thought of as a property of the dense phase and is constant for any dilute phase composition. The ion product is a property of the dilute phase and is given by

$$Q(A, B) = [A][B]^r = AB^r \tag{4.2}$$

The expression for Q is identical to K_{sp} but without the restriction of existing at the phase boundary. Since Q is non-constant, we note that it is a function of the freely varying concentrations A and B. This lets us determine whether a solution is sub-saturated, sat-

urated, or super-saturated when compared to the solubility product through the relations $Q(A,B) < K_{sp}, Q(A,B) = K_{sp}$, or $Q(A,B) > K_{sp}$. The immediate consequence of this is that for a precipitate or dense phase to form, the ion product must be greater than or equal to the solubility product. No dense phase can form if the ion product is less than that of the solubility product.

4.2.2 Dilute Phase Modeling of Oligomers

Biological molecules tend to interact with each other and form small assemblies, which we call oligomers. The population of these oligomers in the dilute phase can be represented using the equilibrium constants which we express through the disassociation constants $K_{m,n}$

$$[A_m B_n] = K_{m,n}[A]^m[B]^n = K_{m,n}A^m B^n \tag{4.3}$$

This lets us calculate the concentration of these oligomeric complexes through the monomer concentrations and the equilibrium constant. We define a partition function Z that represents the entire dilute phase at equilibrium as a sum of concentrations

$$Z = \underbrace{\sum_{m=0}^{\infty}\sum_{n=0}^{\infty}[A_m B_n]}_{(m,n)\neq(0,0)} = \sum_{m,n} K_{m,n}A^m B^n \tag{4.4}$$

m and n are integers representing the number of monomers of A and B per oligomer. An oligomer with no A's or B's can not exist so we disallow the state of $(m,n) = (0,0)$. Note that Z is a grand partition function where the concentration terms are related to the chemical potentials and $K_{m,n}$ is related to the interaction energy. We can calculate the total concentrations of A's(B's) as $A_t(B_t)$ by summing all possible oligomers weighted by

the concentration of A and B in each oligomer

$$A_t(A,B) = \sum_{m,n} mK_{m,n}A^m B^n = \frac{dZ}{d\log A} \tag{4.5}$$

$$B_t(A,B) = \sum_{m,n} nK_{m,n}A^m B^n = \frac{dZ}{d\log B} \tag{4.6}$$

It is helpful to rewrite these equations in terms of $\log A = \mu_A$ and $\log B = \mu_B$ both for ease of notation. The logarithm will also disallow negative concentrations which helps with the optimization of numeric routines. Eliminating negative concentrations prevents numerical root finding algorithms from discovering complex solutions. Redefining our equations in terms of these chemical potentials gives

$$Z = \sum_{m,n} e^{\log K_{m,n} + m\log A + n\log B} = \sum_{m,n} e^{\Delta G_{m,n} + m\mu_A + n\mu_B} \tag{4.7}$$

$$A_t = \frac{dZ}{d\mu_A} = \sum_{m,n} me^{\log K_{m,n} + m\log A + n\log B} = \sum_{m,n} me^{\Delta G_{m,n} + m\mu_A + n\mu_B} \tag{4.8}$$

$$B_t = \frac{dZ}{d\mu_B} = \sum_{m,n} ne^{\log K_{m,n} + m\log A + n\log B} = \sum_{m,n} ne^{\Delta G_{m,n} + m\mu_A + n\mu_B} \tag{4.9}$$

We also take $\Delta G_{m,n} = \log K_{m,n}$ which relates the Gibbs free energy to the dissociation constant. Both the Gibbs free energy $\Delta G_{m,n}$ and the chemical potentials μ_A, μ_B are measured in units of $k_B T$ making them unitless. Often in titration experiments, we will know the total concentrations A_t and B_t at all times throughout the experiment. We can then numerically solve for the monomer concentrations A and B to calculate an ion product. We must use numerical methods in general because polynomials of degree greater than 4 can not generally be inverted. The comparison of the solubility product to the ion product, we can predict whether the system phase separates or not.

4.2.3 Stable Dilute Oligomers Compete with the Dense Phase

Dilute phase oligomers have been shown to suppress the onset of a dense phase experimentally[41,42]. Others have described this behavior using the solubility product before but only for oligomers of a 1:1 and 1:2 stoichiometry[41]. We offer a generalization that works for any oligomers that might exist in solution. Consider two systems with equal A_t, B_t and K_{sp} that we label as system I and II. The only difference between the two systems is that system I does not contain any stable oligomers, while system II does. We can then express the dilute phase partition function of each system as

$$Z_I = A_I + B_I \tag{4.10}$$

$$Z_{II} = A_{II} + B_{II} + \underbrace{\sum_{m=0}^{\infty}\sum_{n=0}^{\infty} K_{m,n}A_{II}^m B_{II}^n}_{(m,n)\neq(0,0),(1,0),(0,1)} \tag{4.11}$$

$$= A_{II} + B_{II} + \sum_{\text{Oligos}} K_{m,n}A_{II}^m B_{II}^n \tag{4.12}$$

Where $Z_I(Z_{II})$ is the partition function for system $I(II)$ respectively. Applying our initial assumption that both systems have equal total concentrations of A's and B's gives the following relations

$$A_t = A_I = A_{\text{II}} + \sum_{\text{Oligos}} mK_{m,n}A_{II}^m B_{II}^n \tag{4.13}$$

$$B_t = B_I = B_{\text{II}} + \sum_{\text{Oligos}} nK_{m,n}A_{II}^m B_{II}^n \tag{4.14}$$

Since the equilibrium constant $K_{m,n}$ is always positive and the concentrations A, B are always positive, we know that the summation over the oligomers is positive. This gives us

the general inequalities

$$A_I > A_{II} \tag{4.15}$$

$$B_I > B_{II} \tag{4.16}$$

Because the monomer concentrations are positive, raising the concentrations to any positive power will preserve the ordering. Therefore, for any positive value r, we have

$$A_I B_I^r > A_{II} B_{II}^r \tag{4.17}$$

$$Q_I > Q_{II} \tag{4.18}$$

The ion product of a system that does not form stable oligomers Q_I will always be greater than a system that forms stable oligomers Q_{II}. Since the ion product of system I is greater than system II for all concentrations A_t and B_t, system I will phase separate at lower concentrations given that both systems have an equal solubility product K_{sp} and ratio r.

4.2.4 The Solubility Product Formalism can Model Re-Entrant Behavior of Phase Diagrams

Here we analyze 3 systems to prove the existence of re-entrant behavior and under what conditions it occurs. The first system (I) is one that does not form oligomers in the dilute phase such as a simple monovalent salt. The second system (II) is one that produces stable oligomers with an equal stoichiometry as the dense phase but nothing else. The last system (III) is one in which some stable oligomers have a stoichiometry different than the dense

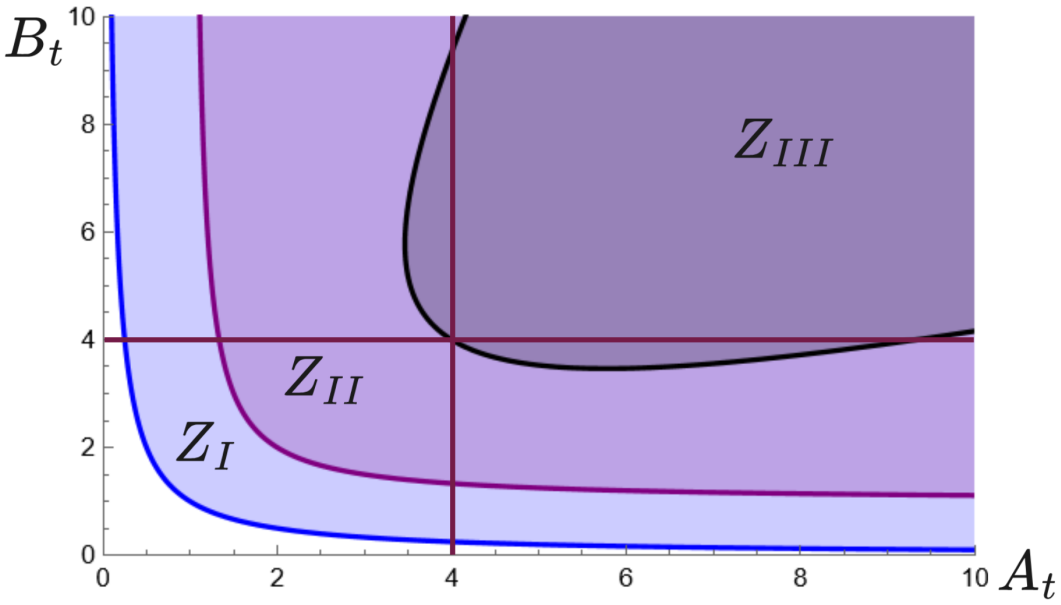


Figure 4.1: *Plots of 3 systems discussed in section 4.2.4. Z_I is a system with no oligomers, Z_{II} is a system with one oligomer that matches the dense phase stoichiometry, and Z_{III} is a system with stable oligomers with a different stoichiometry than the dense phase. This plot uses constants $K_{sp}, K_{N,N}, K_{2,1}, K_{1,2}, N, r = 1$. The shaded regions represent concentrations where phase separation spontaneously occurs. Notice that system III is the only system that shows any re-entrant behavior.*

phase. Defining these systems gives us

$$Z_I = A + B \tag{4.19}$$

$$Z_{II} = A + B + K_{N,N}(AB^r)^N \tag{4.20}$$

$$Z_{III} = A + B + K_{2,1}A^2B + K_{1,2}AB^2 \tag{4.21}$$

$$K_{sp} = AB^r \tag{4.22}$$

We want to generate phase diagrams for these systems with respect to the total concentrations of A_t and B_t. Since re-entrant behavior can't be expressed using single valued functions, we resort to plotting our phase boundary parametrically. A_t and B_t are both functions of the monomer concentrations A and B. If we impose the condition that we only want A_t and B_t that exist on the phase boundary, then we can substitute $A = \frac{K_{sp}}{B^r}$. This allows us to eliminate the concentration A and parametrically plot our phase diagram with respect to B.

$$A_T(B) = A_t(\frac{K_{sp}}{B^r}, B) \tag{4.23}$$

$$B_T(B) = B_t(\frac{K_{sp}}{B^r}, B) \tag{4.24}$$

The functions $A_T(B)$ and $B_T(B)$ represent the total concentration of A's and B's specifically at the phase boundary. The subscript has been capitalized to express that this function only

applies to the phase boundary. Expansion of the right hand side yields

$$A_T(B)_I = \frac{K_{sp}}{B^r} \tag{4.25}$$

$$B_T(B)_I = B \tag{4.26}$$

$$A_T(B)_{II} = \frac{K_{sp}}{B^r} + K_{N,N}(\frac{K_{sp}}{B^r})^N B^{rN} \tag{4.27}$$

$$B_T(B)_{II} = B + K_{N,N}(\frac{K_{sp}}{B^r})^N B^{rN} \tag{4.28}$$

$$A_T(B)_{III} = \frac{K_{sp}}{B^r} + 2K_{2,1}(\frac{K_{sp}}{B^r})^2 B + K_{1,2}(\frac{K_{sp}}{B^r})B^2 \tag{4.29}$$

$$B_T(B)_{III} = B + K_{2,1}(\frac{K_{sp}}{B^r})^2 B + 2K_{1,2}(\frac{K_{sp}}{B^r})B^2 \tag{4.30}$$

Plotting these curves parametrically results in figure 4.1. Notice that system I gives the typical result for a monovalent salt with no re-entrant behavior. Inspection of the functional forms of system II reveals that it also lacks re-entrant since A_T is monotonically decreasing and B_T is monotonically increasing. Notice that A_T and B_T for system II are shifted by the constant value $K_{N,N}K_{sp}^N$ and no matter what value we choose for N, we will not get any re-entrant behavior. The only system with re-entrant behavior is system III and its because the dense phase is competing with oligomers of differing stoichiometry. For large A_t, the A^2B oligomers out-compete the dense phase. Similarly, for large B_t, the B^2A oligomers out-compete the dense phase too. This results in the dilute phase monomers dissolving the dense phase by ripping the complementary monomers away from the dense phase. For a system to have re-entrant behavior, the dilute phase must support stable oligomers with an oligomer stoichiometric ratio different than the dense phase stoichiometric ratio.

4.3 Magic Numbers in Binary Systems

4.3.1 Discussion

Some systems show a "magic number effect" where small perturbations to the structure of a molecule results in very large change of the stability of the molecule. Experimental examples include Rubisco-EPYC1 and polySUMO-polySIM systems while theoretical examples include attractive systems with discrete binding sites[43,44]. In these systems, condensation is highly dependent on the number of binding sites on the monomer designated as the valence. When the valence of one of these molecules is an integer multiple of the other, the dense phase is suppressed in favor of small dilute phase oligomers[43]. This effect only occurs when the energy lost from a single valence mismatch is very high with some systems requiring $14k_BT$ difference to show this effect[44]. The qualitative observation that the dense phase is suppressed by stable oligomers is in complete agreement with our findings from the solubility product formalism laid out in section 4.2.3. What our previous formalism doesn't explain yet is how perturbations to either the A or B molecule leads to the magic number effect.

4.3.2 Structural Perturbations of Monomers

Here we introduce valence into our solubility product formalism where v_A and v_B represents the valence of A and B respectively. The energy these valences provide comes from the pairwise binding of the A and B binding sites. Any unsatisfied binding site results in a large energetic penalty. Applying this to the solubility product formalism will modify the dense phase reference state and the equilibrium constants. This allows us to rewrite the equilibrium constants $K_{sp}, K_{m,n}$, and dense phase ratio r as a function of the valences giving

$$Z = A + B + \sum_{\text{Oligos}} K_{m,n}(v_A, v_B) A^m B^n \tag{4.31}$$

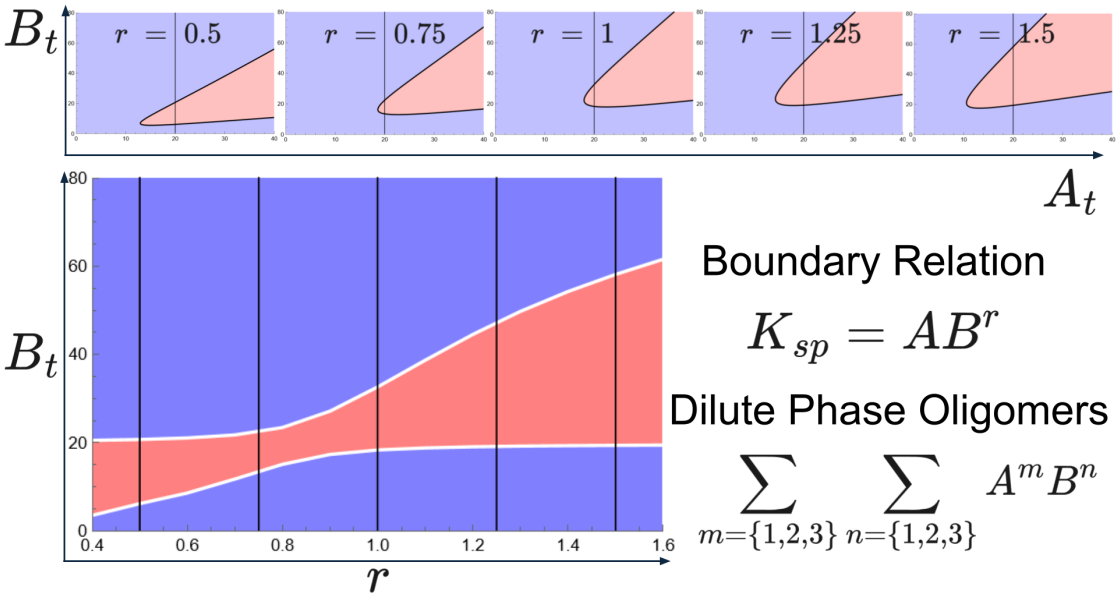


Figure 4.2: *The dense phase stoichiometry r can affect how a system will phase transition. The 5 phase diagrams at the top are for a fixed dense phase ratio r with intersections at $A_t = 20$ with a dilute solution that supports oligomers up to a 3:3 ratio. The phase diagram at the bottom is built from slices of the above phase diagrams as a function of the dense phase composition.*

$$K_{sp}(v_A, v_B) = [A][B]^r = AB^{r(v_A,v_B)} \tag{4.32}$$

The monomer concentrations are independent variables and are not functions of the valences. From this, we can generate phase diagrams as functions of any two variables A_t, B_t, v_A, v_B so long as the other two are held constant. Since we can isolate the concentration A in the second equation, we can plot the phase diagrams as outlined in the previous section. The parametric plots are generated using the equations

$$A_T(B) = A_t(\frac{K_{sp}(v_A, v_B)}{B^{r(v_A,v_B)}}, B) \tag{4.33}$$

$$B_T(B) = B_t(\frac{K_{sp}(v_A, v_B)}{B^{r(v_A,v_B)}}, B) \tag{4.34}$$

In principle, we could parametrically plot this system with fixed total concentration A_T and solve for its intersections B_T. Although we get a valid phase diagram, it is usually slower than explicitly calculating the ion product.

4.3.3 Ion Product for Phase Diagrams

It is faster to generate phase diagrams using alternative numeric methods. This method will generate a phase diagram by calculating the ion product and comparing its value to the solubility product. The ion product for systems where the monomers can be perturbed can be expressed as follows

$$Q(A_t, B_t, v_A, v_B) = A(A_t, B_t, v_A, v_B)B(A_t, B_t, v_A, v_B)^{r(v_A,v_B)} \tag{4.35}$$

Because the ion product is expressed in terms of the monomer concentrations but we are only given the total concentrations, we must numerically solve for the monomer concentrations. This usually ends up being faster than the previous method for generating phase diagrams since there is exactly one solution A, B for any given A_t, B_t, v_A, v_B. The numeric routines

from the parametric plotting have up to two solutions, requiring the routine to spend extra time verifying that there are no more solutions. By representing the concentrations in their logarithmic forms μ_A, μ_B, we can find the total concentrations A_t, B_t using the following equations

$$A_t = \sum_{m,n} m e^{\Delta G_{m,n}(v_A, v_B) + m\mu_A + n\mu_B} \tag{4.36}$$

$$B_t = \sum_{m,n} n e^{\Delta G_{m,n}(v_A, v_B) + m\mu_A + n\mu_B} \tag{4.37}$$

Simultaneously satisfying both of these equations numerically allows us to extract μ_A, μ_B allowing us to calculate the ion product as

$$\log(Q) = \mu_A + \mu_B r(v_A, v_B) \tag{4.38}$$

Since logarithms preserve ordering, it is easier to compare $\log(Q)$ to $\log(K_{sp})$ since the magic number effect appears with very high energies. By numerically solving for $\log(Q)$, a phase diagram can be constructed.

4.3.4 Discrete Binding Site Model Generates the Magic Number Effect

The magic number effect arises from stable oligomers in the dilute solution that suppress phase separation. An example of a system that exhibits this behavior is a model where A has v_A binding sites, and B has v_B binding sites. This model assumes pairwise binding of A and B binding sites ignoring any A-A and B-B binding. Ignoring entropic and pressure effects, the free energy is then proportional to the number of binding sites fulfilled

$$\Delta G_{m,n} = \epsilon \min(m v_A, n v_B) \tag{4.39}$$

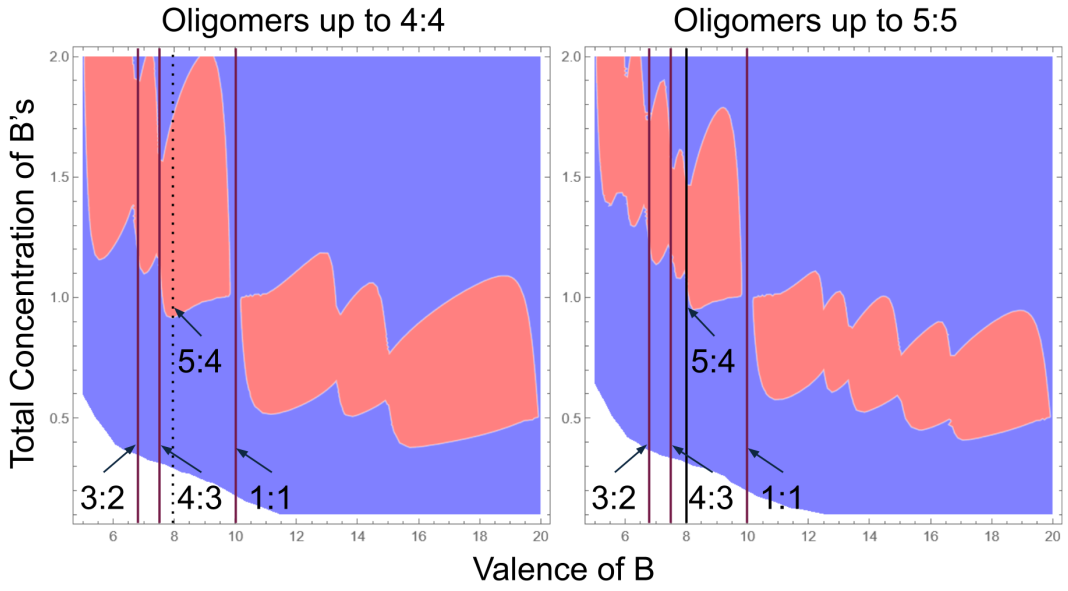


Figure 4.3: *Phase diagrams for systems with the free energy outlined in equation 4.39. Red areas are where dilute and dense phases coexists while blue areas only support the dilute phase. Both figures have the same constants* $\log(K_{sp}) = -30.5$, $\epsilon = -3k_BT$, $v_A = 10$, $A_t = 1$, $r = v_A/v_B$. *The only difference between these plots are what oligomers are present in the dilute phase. Stable oligomers delay the entry to the dense phase and cause an early exit out of the dense phase. A key feature of both diagrams is that no dense phase is nucleated when the dense phase ratio is 1:1*

Here ϵ is the affinity per binding site. Although this is an extremely crude model that ignores PV by assuming that the volume term is negligible. It exhibits magic number features. As we sweep through many possible ratios r for the dense phase as seen in figure 4.3, we can see the dense phase is less likely to form when the dense phase ratio matches that of one of the oligomers in solution. In the left hand phase diagram of 4.3, the regions with stable dilute complexes tend not to form dense phases for very long as seen with the 4:3, 3:2, and 1:1 complexes. Since this phase diagram did not include oligomers up to 5:4, there is no narrowing of the phase separating region. In the right hand side of the graph where the 5:4 oligomer does exist, there is a significant narrowing of the dense phase in phase space. This is because the dense phase is now competing with a stable dilute phase oligomer. The regions of phase space where the dense phase lasts the longest are places where there are fewer stable oligomers. This works only for small oligomers because if the oligomers get too large, the chemical potential severely penalizes larger oligomers even if they are more energetically favorable. After oligomers of up to 10:10, only very minimal changes can be seen.

Chapter 5

Conclusion

5.1 Predictions

Most of this work has been about bottom up modeling of bio-molecular structures. By comparing the results of the modeling to experiments, we can compare the features and make new predictions. The first prediction that we make with the 1D fibril growth model is the potential for negative slopes. The math shows negative slopes can arise in the general model, but not the reduced model. This has also not been observed experimentally with any degree of significance to our knowledge. Another prediction we will make is that RNA chaperone condensates help the RNA unfold through entropic buffering. Although some experiments have shown that HFQ condensates show a stronger chaperoning effect compared to HFQ in dilute solution, there are several other RNA chaperones on which this has not been tested[32]. A more interesting prediction coming out of the chaperone model is that any molecule that preferentially binds to single stranded RNA has an optimal concentration and can chaperone RNA.

5.2 Future Work

We plan on applying the Magic Number Effect of the solubility product to the condensation of a DNA and H1 system by testing more free energy functions. We have developed a theoretical framework that can explain both the re-entrant behavior, as well as the magic number effect expressed by the DNA H1 system through the solubility product. Our choice of free energy function was too simple and did not give a good match for the experiments. There were several free energy models tried and not listed for the DNA H1 system including Poisson Boltzmann Electrostatics, Flory Huggins Polymer Mixing, as well as the Polymer Blob Model for electrostatic confinement. Each one of these models had their own issue with complexity or inability to model general oligomers. Some models came close and showed enough promise to merit a continuation of the project. The issue is that there are infinite possible oligomers so we must choose the right free energy function for robust results. Even then, results could be thrown off by dense phase induced fluctuations caused by the finite size of the system.

Bibliography

[1] Laurence J. Young, Gabriele S. Kaminski Schierle, and Clemens F. Kaminski. Imaging a(1–42) fibril elongation reveals strongly polarised growth and growth incompetent states. *Phys. Chem. Chem. Phys.*, 19(41):27987–27996, 2017. ISSN 1463-9084. doi: 10.1039/c7cp03412a. URL http://dx.doi.org/10.1039/c7cp03412a.

[2] A Lomakin, D S Chung, G B Benedek, D A Kirschner, and D B Teplow. On the nucleation and growth of amyloid beta-protein fibrils: detection of nuclei and quantitation of rate constants. *Proceedings of the National Academy of Sciences*, 93(3): 1125–1129, February 1996. ISSN 1091-6490. doi: 10.1073/pnas.93.3.1125. URL http://dx.doi.org/10.1073/pnas.93.3.1125.

[3] Manuela R. Zimmermann, Subhas C. Bera, Georg Meisl, Sourav Dasadhikari, Shamasree Ghosh, Sara Linse, Kanchan Garai, and Tuomas P. J. Knowles. Mechanism of secondary nucleation at the single fibril level from direct observations of a42 aggregation. *Journal of the American Chemical Society*, 143(40):16621–16629, 2021. ISSN 1520-5126. doi: 10.1021/jacs.1c07228. URL http://dx.doi.org/10.1021/jacs.1c07228.

[4] Ingrid M. van der Wateren, Tuomas P. J. Knowles, Alexander K. Buell, Christopher M. Dobson, and Céline Galvagnion. C-terminal truncation of -synuclein promotes amyloid fibril amplification at physiological ph. *Chemical Science*, 9(25):5506–5516, 2018. ISSN 2041-6539. doi: 10.1039/c8sc01109e. URL http://dx.doi.org/10.1039/c8sc01109e.

[5] Tuomas P. J. Knowles, Wenmiao Shu, Glyn L. Devlin, Sarah Meehan, Stefan Auer, Christopher M. Dobson, and Mark E. Welland. Kinetics and thermodynamics of amy-

loid formation from direct measurements of fluctuations in fibril mass. *Proceedings of the National Academy of Sciences*, 104(24):10016–10021, 2007. ISSN 1091-6490. doi: 10.1073/pnas.0610659104. URL http://dx.doi.org/10.1073/pnas.0610659104.

[6] Jeremy D Schmit. Kinetic theory of amyloid fibril templating. *The Journal of Chemical Physics*, 138(18), May 2013. ISSN 1089-7690. doi: 10.1063/1.4803658. URL http://dx.doi.org/10.1063/1.4803658.

[7] Dev Thacker Emil Axell Elin Stemme Urszula Łapińska Stefan Wennmalm Georg Meisl Samo Curk Maria Andreasen Michele Vendruscolo Paolo Arosio Anela Šarić Jeremy D. Schmit Tuomas P. J. Knowles Emma Sparr Sara Linse Thomas C. T. Michaels Alexander J. Dear Jing Hu, Tom Scheidt. Structural defects in amyloid- fibrils drive secondary nucleation. *Nature Communications*, 17(1), February 2026. ISSN 2041-1723. doi: 10.1038/s41467-026-69377-1. URL http://dx.doi.org/10.1038/s41467-026-69377-1.

[8] Frederick Stull, Philipp Koldewey, Julia R Humes, Sheena E Radford, and James C A Bardwell. Substrate protein folds while it is bound to the atp-independent chaperone spy. *Nature Structural amp; Molecular Biology*, 23(1):53–58, November 2015. ISSN 1545-9985. doi: 10.1038/nsmb.3133. URL http://dx.doi.org/10.1038/nsmb.3133.

[9] Sean R Collins, Adam Douglass, Ronald D Vale, and Jonathan S Weissman. Mechanism of prion propagation: Amyloid growth occurs by monomer addition. *PLoS Biology*, 2 (10):e321, 2004. ISSN 1545-7885. doi: 10.1371/journal.pbio.0020321. URL http://dx.doi.org/10.1371/journal.pbio.0020321.

[10] Caleb Huang, Elaheh Ghanati, and Jeremy D. Schmit. Theory of sequence effects in amyloid aggregation. *The Journal of Physical Chemistry B*, 122(21):5567–5578, February 2018. ISSN 1520-5207. doi: 10.1021/acs.jpcb.7b11830. URL http://dx.doi.org/10.1021/acs.jpcb.7b11830.

[11] Stephen Whitelam and Robert L. Jack. The statistical mechanics of dynamic pathways to self-assembly. *Annual Review of Physical Chemistry*, 66(1):143–163, April 2015. ISSN 1545-1593. doi: 10.1146/annurev-physchem-040214-121215. URL http://dx.doi.org/10.1146/annurev-physchem-040214-121215.

[12] Filip Hasecke, Tatiana Miti, Carlos Perez, Jeremy Barton, Daniel Schölzel, Lothar Gremer, Clara S. R. Grüning, Garrett Matthews, Georg Meisl, Tuomas P. J. Knowles, Dieter Willbold, Philipp Neudecker, Henrike Heise, Ghanim Ullah, Wolfgang Hoyer, and Martin Muschol. Origin of metastable oligomers and their effects on amyloid fibril self-assembly. *Chemical Science*, 9(27):5937–5948, 2018. ISSN 2041-6539. doi: 10.1039/c8sc01479e. URL http://dx.doi.org/10.1039/c8sc01479e.

[13] Stephen Whitelam 2 Georg Meisl 3 Tuomas P.J. Knowles 3 4 Jeremy D. Schmit 1 Sk Ashif Akram 1, Tyler Brown 1. Competing addition processes give distinct growth regimes in the assembly of 1d filaments. *Biophysical Journal*, 124(5):778–788, March 2025. ISSN 0006-3495. doi: 10.1016/j.bpj.2025.01.018. URL http://dx.doi.org/10.1016/j.bpj.2025.01.018.

[14] P B Sigler. An analysis of the structure of trna. *Annual Review of Biophysics and Bioengineering*, 4(1):477–527, June 1975. ISSN 0084-6589. doi: 10.1146/annurev.bb.04.060175.002401. URL http://dx.doi.org/10.1146/annurev.bb.04.060175.002401.

[15] D. Thirumalai and S. A. Woodson. Kinetics of folding of proteins and rna. *Accounts of Chemical Research*, 29(9):433–439, September 1996. ISSN 1520-4898. doi: 10.1021/ar9500933. URL http://dx.doi.org/10.1021/ar9500933.

[16] Daniel Herschlag. Rna chaperones and the rna folding problem. *Journal of Biological Chemistry*, 270(36):20871–20874, September 1995. ISSN 0021-9258. doi: 10.1074/jbc.270.36.20871. URL http://dx.doi.org/10.1074/jbc.270.36.20871.

[17] DH Turner, N Sugimoto, and SM Freier. 3.6. 2 methods: 3.6 thermodynamics and kinetics of base-pairing. *Spectroscopic and Kinetic Data. Physical Data I*, pages 201–202, 1990.

[18] Shi-Jie Chen. Rna folding: Conformational statistics, folding kinetics, and ion electrostatics. *Annual Review of Biophysics*, 37(1):197–214, June 2008. ISSN 1936-1238. doi: 10.1146/annurev.biophys.37.032807.125957. URL http://dx.doi.org/10.1146/annurev.biophys.37.032807.125957.

[19] D. Thirumalai, George H. Lorimer, and Changbong Hyeon. Iterative annealing mechanism explains the functions of the groel and rna chaperones. *Protein Science*, 29(2):360–377, December 2019. ISSN 1469-896X. doi: 10.1002/pro.3795. URL http://dx.doi.org/10.1002/pro.3795.

[20] Bettina G. Keller, Andrei Kobitski, Andres Jäschke, G. Ulrich Nienhaus, and Frank Noé. Complex rna folding kinetics revealed by single-molecule fret and hidden markov models. *Journal of the American Chemical Society*, 136(12):4534–4543, March 2014. ISSN 1520-5126. doi: 10.1021/ja4098719. URL http://dx.doi.org/10.1021/ja4098719.

[21] Roy Nassar, Gregory L. Dignon, Rostam M. Razban, and Ken A. Dill. The protein folding problem: The role of theory. *Journal of Molecular Biology*, 433(20):167126, October 2021. ISSN 0022-2836. doi: 10.1016/j.jmb.2021.167126. URL http://dx.doi.org/10.1016/j.jmb.2021.167126.

[22] Shaon Chakrabarti, Changbong Hyeon, Xiang Ye, George H. Lorimer, and D. Thirumalai. Molecular chaperones maximize the native state yield on biological times by driving substrates out of equilibrium. *Proceedings of the National Academy of Sci-*

ences, 114(51), December 2017. ISSN 1091-6490. doi: 10.1073/pnas.1712962114. URL http://dx.doi.org/10.1073/pnas.1712962114.

[23] Wenbing Zhang and Shi-Jie Chen. Exploring the complex folding kinetics of rna hairpins: I. general folding kinetics analysis. *Biophysical Journal*, 90(3):765–777, February 2006. ISSN 0006-3495. doi: 10.1529/biophysj.105.062935. URL http://dx.doi.org/10.1529/biophysj.105.062935.

[24] Philipp Wenter, Boris Fürtig, Alexandre Hainard, Harald Schwalbe, and Stefan Pitsch. Kinetics of photoinduced rna refolding by real-time nmr spectroscopy. *Angewandte Chemie International Edition*, 44(17):2600–2603, April 2005. ISSN 1521-3773. doi: 10.1002/anie.200462724. URL http://dx.doi.org/10.1002/anie.200462724.

[25] Hermann Hämmerle, Mads Beich-Frandsen, Branislav Večerek, Lukas Rajkowitsch, Oliviero Carugo, Kristina Djinović-Carugo, and Udo Bläsi. Structural and biochemical studies on atp binding and hydrolysis by the escherichia coli rna chaperone hfq. *PLoS ONE*, 7(11):e50892, November 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0050892. URL http://dx.doi.org/10.1371/journal.pone.0050892.

[26] Todd M. Link, Poul Valentin-Hansen, and Richard G. Brennan. Structure of escherichia coli hfq bound to polyriboadenylate rna. *Proceedings of the National Academy of Sciences*, 106(46):19292–19297, November 2009. ISSN 1091-6490. doi: 10.1073/pnas.0908744106. URL http://dx.doi.org/10.1073/pnas.0908744106.

[27] Jon R. Lorsch. Rna chaperones exist and dead box proteins get a life. *Cell*, 109 (7):797–800, June 2002. ISSN 0092-8674. doi: 10.1016/s0092-8674(02)00804-8. URL http://dx.doi.org/10.1016/s0092-8674(02)00804-8.

[28] Zenobia F. Taraporewala and John T. Patton. Identification and characterization of the helix-destabilizing activity of rotavirus nonstructural protein nsp2. *Journal of Virology*,

75(10):4519–4527, May 2001. ISSN 1098-5514. doi: 10.1128/jvi.75.10.4519-4527.2001. URL http://dx.doi.org/10.1128/jvi.75.10.4519-4527.2001.

[29] Christina Lorenz Robert Konrat Renée Schroeder Oliver Mayer, Lukas Rajkowitsch. Rna chaperone activity and rna-binding properties of the e. coli protein stpa. *Nucleic Acids Research*, 35(4):1257–1269, January 2007. ISSN 1362-4962. doi: 10.1093/nar/gkl1143. URL http://dx.doi.org/10.1093/nar/gkl1143.

[30] Weining Jiang, Yan Hou, and Masayori Inouye. Cspa, the major cold-shock protein of escherichia coli, is an rna chaperone. *Journal of Biological Chemistry*, 272(1):196–202, January 1997. ISSN 0021-9258. doi: 10.1074/jbc.272.1.196. URL http://dx.doi.org/10.1074/jbc.272.1.196.

[31] Katharina F Hohmann, Anja Blümler, Alexander Heckel, and Boris Fürtig. The rna chaperone stpa enables fast rna refolding by destabilization of mutually exclusive base pairs within competing secondary structure elements. *Nucleic Acids Research*, 49(19):11337–11349, October 2021. ISSN 1362-4962. doi: 10.1093/nar/gkab876. URL http://dx.doi.org/10.1093/nar/gkab876.

[32] Omer Goldberger, Tamar Szoke, Anat Nussbaum-Shochat, and Orna Amster-Choder. Heterotypic phase separation of hfq is linked to its roles as an rna chaperone. *Cell Reports*, 41(13):111881, December 2022. ISSN 2211-1247. doi: 10.1016/j.celrep.2022.111881. URL http://dx.doi.org/10.1016/j.celrep.2022.111881.

[33] M J Todd, G H Lorimer, and D Thirumalai. Chaperonin-facilitated protein folding: optimization of rate and yield by an iterative annealing mechanism. *Proceedings of the National Academy of Sciences*, 93(9):4030–4035, April 1996. ISSN 1091-6490. doi: 10.1073/pnas.93.9.4030. URL http://dx.doi.org/10.1073/pnas.93.9.4030.

[34] Serguei V. Kuznetsov and Anjum Ansari. A kinetic zipper model with intrachain interactions applied to nucleic acid hairpin folding kinetics. *Biophysical Journal*, 102 (1):101–111, January 2012. ISSN 0006-3495. doi: 10.1016/j.bpj.2011.11.4017. URL http://dx.doi.org/10.1016/j.bpj.2011.11.4017.

[35] Xinchang Wang Guoliang Ke Xiaoyu Cao Chunhai Fan Chaoyong James Yang Haojun Liang Zhong-Qun Tian Zhi-Chao Lei, Xue Yin. Rna can function as molecular chaperone for rna folding. *Giant*, 1:100008, March 2020. ISSN 2666-5425. doi: 10.1016/j.giant.2020.100008. URL http://dx.doi.org/10.1016/j.giant.2020.100008.

[36] Yuan Lin, David S.W. Protter, Michael K. Rosen, and Roy Parker. Formation and maturation of phase-separated liquid droplets by rna-binding proteins. *Molecular Cell*, 60(2):208–219, October 2015. ISSN 1097-2765. doi: 10.1016/j.molcel.2015.08.018. URL http://dx.doi.org/10.1016/j.molcel.2015.08.018.

[37] Amber J LaPeruta, Jelena Micic, and John L Woolford Jr. Additional principles that govern the release of pre-ribosomes from the nucleolus into the nucleoplasm in yeast. *Nucleic Acids Research*, 51(20):10867–10883, 2022. ISSN 1362-4962. doi: 10.1093/nar/gkac430. URL http://dx.doi.org/10.1093/nar/gkac430.

[38] Dikesh Kumar Thakur, Sonal Padole, Tapati Sarkar, Somasundaram Arumugam, and Shiladitya Chattopadhyay. Liquid–liquid phase separation: Mechanisms, roles, and implications in cellular function and disease. *FASEB BioAdvances*, 7(11), November 2025. ISSN 2573-9832. doi: 10.1096/fba.2025-00140. URL http://dx.doi.org/10.1096/fba.2025-00140.

[39] Sarah L. Nichols, Emil M. Nilsson, Heather Brown-Harding, Leslie E. W. LaConte, Julia Acker, Alexander Borodavka, and Sarah McDonald Esstman. Flexibility of the rotavirus nsp2 c-terminal region supports factory formation via liquid-liquid phase separation.

Journal of Virology, 97(2), February 2023. ISSN 1098-5514. doi: 10.1128/jvi.00039-23. URL http://dx.doi.org/10.1128/jvi.00039-23.

[40] Mrityunjoy Kar, Furqan Dar, Timothy J. Welsh, Laura T. Vogel, Ralf Kühnemuth, Anupa Majumdar, Georg Krainer, Titus M. Franzmann, Simon Alberti, Claus A. M. Seidel, Tuomas P. J. Knowles, Anthony A. Hyman, and Rohit V. Pappu. Phase-separating rna-binding proteins form heterogeneous distributions of clusters in subsaturated solutions. *Proceedings of the National Academy of Sciences*, 119(28), July 2022. ISSN 1091-6490. doi: 10.1073/pnas.2202222119. URL http://dx.doi.org/10.1073/pnas.2202222119.

[41] Sarah J. Nehm, Barbara Rodríguez-Spong, and Naír Rodríguez-Hornedo. Phase solubility diagrams of cocrystals are explained by solubility product and solution complexation. *Crystal Growth amp; Design*, 6(2):592–600, December 2005. ISSN 1528-7505. doi: 10.1021/cg0503346. URL http://dx.doi.org/10.1021/cg0503346.

[42] Ian Seim, Ammon E. Posey, Wilton T. Snead, Benjamin M. Stormo, Daphne Klotsa, Rohit V. Pappu, and Amy S. Gladfelter. Dilute phase oligomerization can oppose phase separation and modulate material properties of a ribonucleoprotein condensate. *Proceedings of the National Academy of Sciences*, 119(13), March 2022. ISSN 1091-6490. doi: 10.1073/pnas.2120799119. URL http://dx.doi.org/10.1073/pnas.2120799119.

[43] Bin Xu, Guanhua He, Benjamin G. Weiner, Pierre Ronceray, Yigal Meir, Martin C. Jonikas, and Ned S. Wingreen. Rigidity enhances a magic-number effect in polymer phase separation. *Nature Communications*, 11(1), March 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-15395-6. URL http://dx.doi.org/10.1038/s41467-020-15395-6.

[44] Yaojun Zhang, Bin Xu, Benjamin G Weiner, Yigal Meir, and Ned S Wingreen. Decoding the physical principles of two-component biomolecular phase separation. *eLife*, 10, March 2021. ISSN 2050-084X. doi: 10.7554/elife.62403. URL http://dx.doi.org/10.7554/eLife.62403.

Appendix A

Derivation of Reversible Mean First Passage Time for Linear State Spaces

A.1 Linear First Passage Process

Consider a stochastic system of discrete space and continuous time. Now impose the condition that the discrete space is a straight line with absorbing boundaries. Then the system can be modeled like a random walk on an integer line starting at index 0 and n since we imposed the condition of discrete space. Furthermore, we can define properties of an ensemble of particles as follows

$$C(x, t + \Delta t) = (1 - k_{x+} - k_{x-})C(x, t) + k_{(x+1)-}C(x + 1, t)\Delta t + k_{(x-1)-}C(x - 1, t)\Delta t$$

Where $C(x, t)$ is the concentration of the random walkers at index x and time t. Expanding the left hand side of the above equation then reducing gives

$$\frac{\partial C(x, t)}{\partial t} = -(k_{x-} + k_{x+})C(x, t) + k_{(x+1)-}C(x_1, t) + k_{(x-1)-}C(x, t)$$

Note that for the edge cases at 0 and n, the probabilities of the indices -1 and $n+1$ are 0 as well as the rates going to and leaving from these states. Lastly, if we consider the equations from the viewpoint of a random walker, then we we know we are in state i with probability 1 at time t. Then, the probability we remain in state x is given by the equation

$$\frac{\partial P(i,t)}{\partial t} = -(k_{i+} + k_{i-})P(i,t)$$

This partial differential equation then has the trivial solution

$$P(i,t) = e^{-}(k_{i+} + k_{i-})t$$

Taking the average of this distribution, we can find the average time spent in state i, also known as the residence time, as

$$r_i = \frac{1}{k_{i+} + k_{i-}}$$

We also are then able to find the probability moving to the right p_i and the probability of moving left $q_i = 1 - p_i$ as

$$p_i = \frac{k_{i+}}{k_{i+} + k_{i-}}$$
$$q_i = \frac{k_{i-}}{k_{i+} + k_{i-}}$$

A.2 First Passage Probability

Here, we want to solve for the first passage probability. Consider the walk starting at index 1 and we want to know the probability it reaches 0 or n first. Define g_i as the probability

of reaching 0 before reaching n starting from index i, then we are left with the system of equations

$$g_0 = 1$$

$$g_n = 0$$

$$g_i = q_i g_{i-1} + p_i g_{i+1}$$

From our definition, if we start our random walk at index 1, then the probability of reaching index n before reaching index 0 (P_{success}) is given by $1 - g_1$. Now define a new function $u_i = g_{i-1} - g_i$. Now we get a new recurrence relation

$$u_{i+1} = \frac{q_i}{p_i} u_i = \left(\prod_{j=1}^{i} \frac{q_j}{p_j} \right) u_1$$

Calculating the sum of u_i gives

$$\begin{aligned} \sum_{i=1}^{n} u_i &= u_1 + u_2 + ... + u_n \\ &= (g_0 - g_1) + (g_1 - g_2) + ... + (g_{n-1} - g_n) \\ &= g_0 - g_n \\ &= 1 - 0 \end{aligned}$$

Substituting u_i from the recurrence relation gives

$$\begin{aligned}
1 &= \sum_{i=1}^{n} u_i \\
&= \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{q_j}{p_j} \right) u_1 \\
&= u_1 \sum_{i=1}^{n} \prod_{j=1}^{i-1} \frac{q_j}{p_j} \\
&= (1 - g_1) \sum_{i=1}^{n} \prod_{j=1}^{i-1} \frac{q_j}{p_j} \\
&= P_{\text{success}} \sum_{i=1}^{n} \prod_{j=1}^{i-1} \frac{q_j}{p_j}
\end{aligned}$$

Thus we are left with the probability of reaching index n before index 0 starting at index 1 as

$$P_{\text{success}} = u_1 = \frac{1}{\sum_{i=1}^{n} \prod_{j=1}^{i-1} \frac{q_j}{p_j}} = \frac{1}{1 + \frac{q_1}{p_1} + \frac{q_1 q_2}{p_1 p_2} + ... + \frac{q_1 q_2 ... q_{n-1}}{p_1 p_2 ... p_{n-1}}}$$

We can extrapolate this result to any starting index. First we must notice that the probability of reaching index n before reaching index 0 starting at index k is equal to

$$\begin{aligned}
1 - g_k &= \sum_{i=1}^{k} u_i \\
&= \sum_{i=1}^{k} \left(\prod_{j=1}^{i-1} \frac{q_j}{p_j} \right) u_1 \\
&= u_1 \sum_{i=1}^{k} \prod_{j=1}^{i-1} \frac{q_j}{p_j} \\
&= \frac{\sum_{i=1}^{k} \prod_{j=1}^{i-1} \frac{q_j}{p_j}}{\sum_{i=1}^{n} \prod_{j=1}^{i-1} \frac{q_j}{p_j}}
\end{aligned}$$

If we were to start a random walk at index 0, since we have a boundary and can't move to index -1 we must move forward to index 1 with probability 1 After reaching index 1, we

then have probability u_1 of reaching index n before reaching 0 again.

A.3 First Passage Time

Similarly to the previous problem, we can model the first passage time as a system of equations. Define the function t_i as the average time it takes to hit index 0 or index n. This can be modeled using a new inhomogenious system of equations

$$t_0 = 0$$

$$t_n = 0$$

$$t_i = q_i t_{i-1} + p_i t_{i+1} + r_i$$

t_i is the average time it takes to reach either index 0 or index n starting at index i. This means that t_1 is the average time it takes to be absorbed by a boundary when starting at index 1. Again, defining a new function $v_i = t_{i-1} - t_i$, we can modify our previous recurrence relation as

$$v_{i+1} = \frac{q_i}{p_i} v_i + \frac{r_i}{p_i}$$

The equation can further be expanded as

$$
\begin{aligned}
v_{i+1} &= \frac{q_i}{p_i} v_i + \frac{r_i}{p_i} \\
&= \frac{q_i}{p_i} \left(\frac{q_{i-1}}{p_{i-1}} v_{i-1} + \frac{r_{i-1}}{p_{i-1}} \right) + \frac{r_i}{p_i} \\
&= \frac{q_i q_{i-1}}{p_i p_{i-1}} v_{i-1} + \frac{q_i}{p_i} \frac{r_{i-1}}{p_{i-1}} + \frac{r_i}{p_i} \\
&= \frac{q_i q_{i-1}}{p_i p_{i-1}} (...) + \frac{q_i}{p_i} \frac{r_{i-1}}{p_{i-1}} + \frac{r_i}{p_i} \\
&= \frac{q_i q_{i-1} ... q_1}{p_i p_{i-1} ... p_1} v_1 + \frac{q_i q_{i-1} ... q_2}{p_i p_{i-1} ... p_2} \frac{r_1}{p_1} + \frac{q_i q_{i-1} ... q_3}{p_i p_{i-1} ... p_3} \frac{r_2}{p_2} + ... + \frac{p_i}{q_i} \frac{r_{i-1}}{p_{i-1}} + \frac{r_i}{p_i} \\
&= \left(\prod_{j=1}^{i} \frac{q_j}{p_j} \right) v_1 + \sum_{j=1}^{i} \left(\prod_{k=j+1}^{i} \frac{q_k}{p_k} \right) \frac{r_j}{p_j}
\end{aligned}
$$

Calculating the sum of v_i gives

$$
\begin{aligned}
\sum_{i=1}^{n} v_i &= v_1 + v_2 + ... v_n \\
&= (t_0 - t_1) + (t_1 - t_2) + ... + (t_{n-1} - t_n) \\
&= t_0 - t_n \\
&= 0 - 0
\end{aligned}
$$

Lastly, we can form the equation

$$
\begin{aligned}
0 &= \sum_{i=1}^{n} v_i \\
&= \sum_{i=1}^{n} \left(\left(\prod_{j=1}^{i-1} \frac{q_j}{p_j} \right) v_1 + \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j} \right) \\
&= \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{q_j}{p_j} \right) v_1 + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j} \\
&= v_1 \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{q_j}{p_j} \right) + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j} \\
&= (t_0 - t_1) \sum_{i=1}^{n} \left(\prod_{j=1}^{i-1} \frac{q_j}{p_j} \right) + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j} \\
&= \frac{0 - t_1}{P_{\text{success}}} + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j}
\end{aligned}
$$

Bringing the t_1 to the other side gives the equation its final form

$$
\frac{t_1}{P_{\text{success}}} = \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j}
$$

A.4 Total Time to Reach N

Thus far we have calculated the probability that any attempt reaches n as P_{success} and the average time per attempt t_1 starting at index 1. We still need to calculate the total time T that it takes for the system to reach state N. For this, consider a system that starts at index 0. After some time r_0 the system moves to state 1 with probability 1. After that, the system spends time t_1 somewhere in the middle until it returns to index 0 or finishes at index N. We expect the total number of attempts to be $1/P_{\text{success}}$. Therefore, we can

calculate the total average time it takes to reach index n as

$$T = (\text{Average Time Per Attempt}) * (\text{Expected Number of Attempts}) = (r_0 + t_1) * \frac{1}{P_{\text{success}}}$$

We can solve this analytically in terms of the residence times and probabilities as follows.

$$T = r_0 \sum_{i=1}^{n} \prod_{j=1}^{i-1} \frac{q_j}{p_j} + \sum_{i=1}^{n} \sum_{j=1}^{i-1} \left(\prod_{k=j+1}^{i-1} \frac{q_k}{p_k} \right) \frac{r_j}{p_j}$$

Appendix B

Derivation of the 1D Fibril Partition Function and Rates

This appendix covers the derivation of the partition function found in equation 2.11. Consider the transfer matrix found in equation 2.10 given by A.

$$A = \begin{bmatrix} 0 & pce^{-\epsilon_{rb}} \\ qce^{-\epsilon_{br}} & qce^{-\epsilon_{rr}} \end{bmatrix} \tag{B.1}$$

This matrix can be diagonalized by the matrix P where $P^{-1}AP = D$. The individual matrices are given as follows

$$P = \begin{bmatrix} -\frac{\left(\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}+\sqrt{q}e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}}\right)}{2\sqrt{q}e^{-\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}+\epsilon_{rr}}} & -\frac{\left(\sqrt{q}e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}}-\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}\right)}{2\sqrt{q}e^{-\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}+\epsilon_{rr}}} \\ 1 & 1 \end{bmatrix} \tag{B.2}$$

$$P^{-1} = \begin{bmatrix} -\frac{\sqrt{q}e^{-\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}+\epsilon_{rr}}}{\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}} & -\frac{\sqrt{q}e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}}-\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}}{2\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}} \\ \frac{\sqrt{q}e^{-\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}+\epsilon_{rr}}}{\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}} & \frac{\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}+\sqrt{q}e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}}}{2\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}p}} \end{bmatrix} \tag{B.3}$$

$$D = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} \tag{B.4}$$

$$= \begin{bmatrix} -\frac{1}{2}c\sqrt{q}\frac{\left(\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}}p-\sqrt{q}e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}}\right)}{e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}+\epsilon_{rr}}} & 0 \\ 0 & \frac{1}{2}c\sqrt{q}\frac{\left(\sqrt{qe^{\epsilon_{br}+\epsilon_{rb}}+4e^{2\epsilon_{rr}}}p+\sqrt{q}e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}}\right)}{e^{\frac{\epsilon_{br}}{2}+\frac{\epsilon_{rb}}{2}+\epsilon_{rr}}} \end{bmatrix} \tag{B.5}$$

These equations were generated using Mathematica. The multiplication of the matrices PDP^{-1} can be used to recover our original matrix A. The partition sum over the disordered cap states is given by

$$\begin{bmatrix} Z_b \\ Z_r \end{bmatrix} = \sum_{\ell=0}^{\infty} \begin{bmatrix} Z_{b,\ell} \\ Z_{r,\ell} \end{bmatrix} \tag{B.6}$$

$$= \sum_{\ell=0}^{\infty} A^{\ell} \begin{bmatrix} Z_{b,0} \\ Z_{r,0} \end{bmatrix} \tag{B.7}$$

$$= \sum_{\ell=0}^{\infty} (PDP^{-1})^{\ell} \begin{bmatrix} Z_{b,0} \\ Z_{r,0} \end{bmatrix} \tag{B.8}$$

$$= \sum_{\ell=0}^{\infty} PD^{\ell}P^{-1} \begin{bmatrix} Z_{b,0} \\ Z_{r,0} \end{bmatrix} \tag{B.9}$$

$$= P\left(\sum_{\ell=0}^{\infty} D^{\ell}\right) P^{-1} \begin{bmatrix} Z_{b,0} \\ Z_{r,0} \end{bmatrix} \tag{B.10}$$

$$= P\bar{D}P^{-1} \begin{bmatrix} Z_{b,0} \\ Z_{r,0} \end{bmatrix} \tag{B.11}$$

Here we introduce the new notation of $\bar{D}$. This notation represents the infinite geometric series of the diagonal matrix D with elements λ_i. Our matrix happens to have eigenvalues between -1 and 1 prior to the first divergence which is the region we are interested in. This

allows us to sum the geometric series as follows

$$\bar{D} = \sum_{\ell=0}^{\infty} D^{\ell} \tag{B.12}$$

$$= \sum_{\ell=0}^{\infty} \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}^{\ell} \tag{B.13}$$

$$= \begin{bmatrix} \sum_{\ell=0}^{\infty} \lambda_1^{\ell} & 0 \\ 0 & \sum_{\ell=0}^{\infty} \lambda_2^{\ell} \end{bmatrix} \tag{B.14}$$

$$= \begin{bmatrix} \frac{1}{1-\lambda_1} & 0 \\ 0 & \frac{1}{1-\lambda_2} \end{bmatrix} \tag{B.15}$$

Since this is the partition function of a disordered cap on an ordered fibril, the fibril is guaranteed to end in a blue monomer giving the boundary conditions of $Z_{b,0} = 1, Z_{r,0} = 0$. We then generate the partition functions by evaluating

$$\begin{bmatrix} Z_b \\ Z_r \end{bmatrix} = P\bar{D}P^{-1} \begin{bmatrix} 1 \\ 0 \end{bmatrix} \tag{B.16}$$

$$= \begin{bmatrix} \frac{e^{\epsilon_{br}+\epsilon_{rb}}(e^{\epsilon_{rr}}-cq)}{c^2(-e^{\epsilon_{rr}})pq-cqe^{\epsilon_{br}+\epsilon_{rb}}+e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}} \\ \frac{cqe^{\epsilon_{rb}+\epsilon_{rr}}}{c^2(-e^{\epsilon_{rr}})pq-cqe^{\epsilon_{br}+\epsilon_{rb}}+e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}} \end{bmatrix} \tag{B.17}$$

Finally, after some miraculous cancellation, the partition function Z can be calculated by summing both terms of the vector resulting in

$$Z = Z_b + Z_r \tag{B.18}$$

$$= \frac{e^{\epsilon_{br}+\epsilon_{rb}}\left(e^{\epsilon_{rr}} - cq\right)}{c^2\left(-e^{\epsilon_{rr}}\right)pq - cqe^{\epsilon_{br}+\epsilon_{rb}} + e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}} + \frac{cqe^{\epsilon_{rb}+\epsilon_{rr}}}{c^2\left(-e^{\epsilon_{rr}}\right)pq - cqe^{\epsilon_{br}+\epsilon_{rb}} + e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}} \tag{B.19}$$

$$= \frac{e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\left(1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}\right)}{c^2\left(-e^{\epsilon_{rr}}\right)pq - cqe^{\epsilon_{br}+\epsilon_{rb}} + e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}} \tag{B.20}$$

$$= \frac{1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}}{1 - qce^{-\epsilon_{rr}} - pqc^2e^{-\epsilon_{br}-\epsilon_{rb}}} \tag{B.21}$$

We use this partition function in the calculation of the ordered and disordered growth rates found in equations eq 2.13 and eq 2.15. The ordered rate is given by

$$R_{\text{ord}} = (k_{b+} - k_{bb-})\frac{Z_{b,0}}{Z} \tag{B.22}$$

$$= (k_0pc - k_0e^{\epsilon_{bb}})\frac{1}{Z} \tag{B.23}$$

$$= (k_0pc - k_0e^{\epsilon_{bb}})\frac{1 - qce^{-\epsilon_{rr}} - pqc^2e^{-\epsilon_{br}-\epsilon_{rb}}}{1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}} \tag{B.24}$$

The disordered rate is then given by

$$R_{dis} = \sum_{\ell=1}^{\infty}(\ell+1)\frac{Z_{b,\ell}}{Z} \tag{B.25}$$

$$= \frac{k_{b+}}{Z}\sum_{\ell=1}^{\infty}(\ell+1)Z_{b,\ell} \tag{B.26}$$

$$= \frac{k_{b+}}{Z}\sum_{\ell=1}^{\infty}(\ell+1)\begin{bmatrix}1 & 0\end{bmatrix}\begin{bmatrix}Z_{b,\ell}\\ Z_{r,\ell}\end{bmatrix} \tag{B.27}$$

$$= \frac{k_{b+}}{Z}\sum_{\ell=1}^{\infty}(\ell+1)\begin{bmatrix}1 & 0\end{bmatrix}A^{\ell}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.28}$$

$$= \frac{k_{b+}}{Z}\sum_{\ell=1}^{\infty}(\ell+1)\begin{bmatrix}1 & 0\end{bmatrix}PD^{\ell}P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.29}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix}P\left(\sum_{\ell=1}^{\infty}(\ell+1)D^{\ell}\right)P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.30}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix}P\begin{bmatrix}\sum_{\ell=1}^{\infty}(\ell+1)\lambda_1^{\ell} & 0\\ 0 & \sum_{\ell=1}^{\infty}(\ell+1)\lambda_2^{\ell}\end{bmatrix}P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.31}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix}P\begin{bmatrix}\frac{d}{d\lambda_1}\sum_{\ell=2}^{\infty}\lambda_1^{\ell} & 0\\ 0 & \frac{d}{d\lambda_2}\sum_{\ell=2}^{\infty}\lambda_2^{\ell}\end{bmatrix}P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.32}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix}P\begin{bmatrix}\frac{d}{d\lambda_1}(\frac{1}{1-\lambda_1}-\lambda_1-1) & 0\\ 0 & \frac{d}{d\lambda_2}(\frac{1}{1-\lambda_2}-\lambda_2-1)\end{bmatrix}P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.33}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix}P\begin{bmatrix}\frac{1}{(1-\lambda_1)^2}-1 & 0\\ 0 & \frac{1}{(1-\lambda_2)^2}-1\end{bmatrix}P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.34}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix}P(\bar{D}\bar{D}-I)P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.35}$$

Resolving the matrix multiplication $P(\bar{D}\bar{D} - I)P^{-1}$ in Mathematica gives the following result

$$B = P(\bar{D}\bar{D} - I)P^{-1} \tag{B.36}$$

$$= \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \tag{B.37}$$

Where the individual elements of B are given by

$$b_{11} = -\frac{c^2 e^{\epsilon_{rr}} pq \left(c^2 e^{\epsilon_{rr}} pq + 2cqe^{\epsilon_{br}+\epsilon_{rb}} - 3e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)}{\left(c^2 e^{\epsilon_{rr}} pq + cqe^{\epsilon_{br}+\epsilon_{rb}} - e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)^2} \tag{B.38}$$

$$b_{12} = \frac{cp\left(2e^{\epsilon_{rr}} - cq\right) e^{2\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}}{\left(c^2 e^{\epsilon_{rr}} pq + cqe^{\epsilon_{br}+\epsilon_{rb}} - e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)^2} \tag{B.39}$$

$$b_{21} = \frac{cq\left(2e^{\epsilon_{rr}} - cq\right) e^{\epsilon_{br}+2\epsilon_{rb}+\epsilon_{rr}}}{\left(c^2 e^{\epsilon_{rr}} pq + cqe^{\epsilon_{br}+\epsilon_{rb}} - e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)^2} \tag{B.40}$$

$$b_{22} = -\frac{cq\left(c^3 e^{2\epsilon_{rr}} p^2 q + 2c^2 pqe^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}} - 3cpe^{\epsilon_{br}+\epsilon_{rb}+2\epsilon_{rr}} + cqe^{2(\epsilon_{br}+\epsilon_{rb})} - 2e^{2\epsilon_{br}+2\epsilon_{rb}+\epsilon_{rr}}\right)}{\left(c^2 e^{\epsilon_{rr}} pq + cqe^{\epsilon_{br}+\epsilon_{rb}} - e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)^2} \tag{B.41}$$

This lets us calculate the disordered rate as

$$R_{\text{dis}} = \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix} P(\bar{D}\bar{D} - I)P^{-1}\begin{bmatrix}Z_{b,0}\\ Z_{r,0}\end{bmatrix} \tag{B.42}$$

$$= \frac{k_{b+}}{Z}\begin{bmatrix}1 & 0\end{bmatrix} B\begin{bmatrix}1\\ 0\end{bmatrix} \tag{B.43}$$

$$= \frac{k_{b+}}{Z} b_{11} \tag{B.44}$$

$$= \frac{k_{b+}}{Z}\frac{c^2 e^{\epsilon_{rr}} pq\left(-c^2 e^{\epsilon_{rr}} pq - 2cqe^{\epsilon_{br}+\epsilon_{rb}} + 3e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)}{\left(c^2 e^{\epsilon_{rr}} pq + cqe^{\epsilon_{br}+\epsilon_{rb}} - e^{\epsilon_{br}+\epsilon_{rb}+\epsilon_{rr}}\right)^2} \tag{B.45}$$

$$= \frac{k_{b+}}{Z}\frac{pqc^2 e^{-\epsilon_{br}-\epsilon_{rb}}\left(-pqc^2 e^{-\epsilon_{br}-\epsilon_{rb}} - 2qce^{-\epsilon_{rr}} + 3\right)}{\left(c^2 e^{-\epsilon_{br}-\epsilon_{rb}} pq + cqe^{-\epsilon_{rr}} - 1\right)^2} \tag{B.46}$$

$$= \frac{k_0 pc(1 - qce^{-\epsilon_{rr}} - pqc^2 e^{-\epsilon_{br}-\epsilon_{rb}})}{1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}}}\frac{pqc^2 e^{-\epsilon_{br}-\epsilon_{rb}}\left(-pqc^2 e^{-\epsilon_{br}-\epsilon_{rb}} - 2qce^{-\epsilon_{rr}} + 3\right)}{\left(-c^2 e^{-\epsilon_{br}-\epsilon_{rb}} pq - cqe^{-\epsilon_{rr}} + 1\right)^2} \tag{B.47}$$

$$= \frac{k_0 qp^2 c^3 e^{-\epsilon_{rb}-\epsilon_{br}}(3 - 2qce^{-\epsilon_{rr}} - qpc^2 e^{-\epsilon_{br}-\epsilon_{rb}})}{(1 - qce^{-\epsilon_{rr}} + qce^{-\epsilon_{br}})(1 - qce^{-\epsilon_{rr}} - qpc^2 e^{-\epsilon_{br}-\epsilon_{rb}})} \tag{B.48}$$

Although we only show the solution for the 2 state red-blue model, we also solved it for a 3 state red-blue-purple model. The linear algebra is exactly the same although the 3 state case will yield different matrices for A, P, P^{-1}, D, and $\bar{D}$. We determined that the 3 state model was worse than the 2 state model since it would have an earlier divergence giving the 2 state model more predictive capability. In general, the N state model still obeys the equations

$$Z = \sum_{\ell=0}^{\infty} Z_\ell = \sum_{\ell=0}^{\infty} A^\ell Z_0 = P\bar{D}P^{-1}Z_0 \tag{B.49}$$

$$R_{\text{ord}} = \frac{(k_{0+} - k_{00-})}{Z} \tag{B.50}$$

$$R_{\text{dis}} = \frac{k_{0+}}{Z}\begin{bmatrix}1 & 0 & 0 & \cdots\end{bmatrix} P(\bar{D}\bar{D} - I)P^{-1}Z_0 \tag{B.51}$$

where k_{0+} and k_{00-} are the on and off rates of the color that is most stable and Z_0 is the base case vector