1 Model Development

1.1 Freely Jointed Chain

We create a generic model of a disordered protein using a simplified θ-solvent freely-jointed chain (FJC) from polymer physics. This model requires only specifying a number of rods (N) and a length per rod (Kuhn length, δ). The FJC consists of N rigid rods of length δ which are allowed to perform a random walk where the only constraints are the chain length and connections. In our simulation, the FJC is allowed to explore its configuration space through randomized movements. The FJC model tracks where each joint is located, making steric interactions with the ligand easy to simulate compared to using the continuous WLC model. The ligand is simulated as an idealized sphere which may interact with the FJC. We compute quasi-equilibrium statistics of the chain and its bound or unbound ligands using a Monte Carlo (Metropolis) Algorithm.

We can relate the change in dissociation constants between a floppy, unstructured protein ($K_{D_F} = K_F$) and a rigid, structured protein ($K_{D_R} = K_R$) to the probability a ligand may bind to the protein in an unstructured state. From detailed balance, which is what and we can use it why? the dissociation constants $K_D \equiv k_{\text{off}}/k_{\text{on}}$ of the floppy (native) state and the rigid (partially phosphorylated) state are

$$\frac{k_{on}}{k_{off}} = \exp\left(\frac{-\Delta G}{k_B T}\right) \tag{1}$$

$$\frac{K_{D_F}}{K_{D_R}} = \frac{K_F}{K_R} = \frac{\frac{1}{\exp\left(\frac{-\Delta G_F}{k_B T}\right)}}{\frac{1}{\exp\left(\frac{-\Delta G_R}{k_B T}\right)}} \tag{2}$$

$$\frac{K_F}{K_R} = \frac{\exp\left(\frac{-\Delta G_R}{k_B T}\right)}{\exp\left(\frac{-\Delta G_F}{k_B T}\right)} \tag{3}$$

$$= \exp\left(\frac{\Delta G_F - \Delta G_R}{k_B T}\right) \tag{4}$$

$$= \exp\left(\frac{(E_F - TS_F) - (E_R - TS_R)}{k_B T}\right) \tag{5}$$

$$= \exp\left(\frac{S_R - S_F}{k_B}\right) \tag{6}$$

$$= \text{ex}\left(\frac{(S^R_{on} - S^R_{off}) - (S^F_{on} - S^F_{off})}{k_B}\right) \tag{7}$$

$$= \exp\left(\frac{(k_B \ln(\frac{\Omega_R P_R}{\Omega_R})) - (k_B \ln(\frac{\Omega_F P_F}{\Omega_F})}{k_B}\right) \tag{8}$$

$$= \frac{P_R}{P_F} \tag{9}$$

$$= \frac{1}{P_F} \qquad \text{assuming } P_R = 1 \tag{10}$$

where $G_j = E_j - TS_j$ is the free energy of binding in the free or rigid state, $S_j = k_B \ln W$ is the entropy of binding, where W is the number of microstates and P_j is the probability in the canonical ensemble that the configuration allows for binding. We let Ω be the total number of microstates, with $\Omega_F * P_F$ being the microstates available when the ligand is bound and $\Omega_R * P_R$ is the number of microstates available when the polymer is rigid. We also note $E_R = E_F$ since we are working at equilibrium. Since the perfectly rigid state always allows binding, $P_R = 1$, it is sufficient to compute P_F. We define P_{occ} as the probability that the region of space needed by the kinase domain is occupied by some of the polymer. Thus, $P_{occ} = 1 - P_F$.

Then

$$\frac{K_F}{K_R} = \frac{1}{(1 - P_{occ})}. \tag{11}$$

Our problem of interest is now reduced to computing the occlusion probability.

We calculate how often a ligand is able to bind to an oriented sphere tangentially attached to the polymer, where 'able to bind' refers to the specified sphere being empty of both other polymer segments and half-space barriers weird wording?. To determine if a site is occluded in a given conformation, we check if any of the segment end points are located within the sphere of interest. We also check if the sphere crosses below the half-space surface designated at $z = 0$. Given that the binding sphere is large compared to the Kuhn length, we assume the probability of tangential occlusion (where a segment has end points outside the sphere but part of the segment lies within) is negligible compared to end point occlusion. Should we produce a plot for this someday to show it doesn't matter? Or create a diagram to make it more obvious what all the types of occlusion are?

1.2 Code Validation

There are theoretical solutions for many aspects of the freely-jointed chain. This provides a basis with which to verify our code.

1.2.1 Root-mean-square end-to-end distance

First, we look at the average end-to-end distance of the polymer (RMS). In our simulations, we normalize by the Kuhn length, so all simulations assume $\delta = 1$ and record all other parameters in units of Kuhn lengths. We know from polymer physics that the RMS should increase as $\delta\sqrt{N}$, which given $\delta = 1$, is just $\sqrt{N}$.

Average end-to-end distance (Root-mean-square end-to-end distance): [**Reeves2011**]

$$\sqrt{\langle r_{ee}^2 \rangle} = \sqrt{N\delta^2} = \delta\sqrt{N}$$

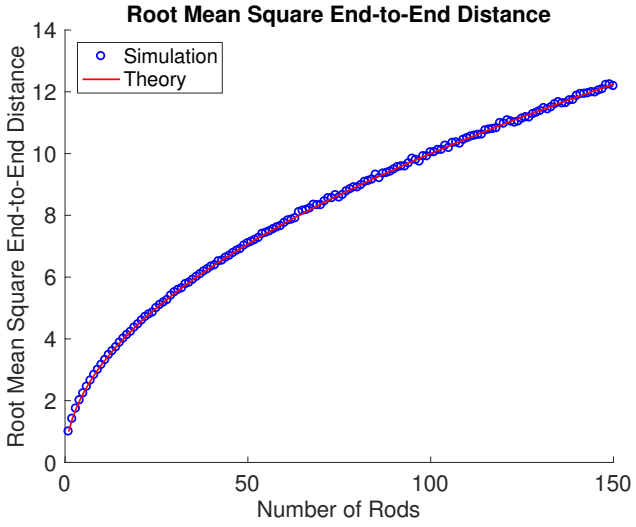


Figure 1: Theoretical root mean square end-to-end distance (red line) against simulated values (blue circles).

1.2.2 R_{ee} Distribution

End-to-end distribution: [**VanValen2009**], [**Reeves2011**]

$$P(r_{ee}) = 4\pi r^2 \left(\frac{3}{2\pi N\delta^2}\right)^{\frac{3}{2}} exp\left(\frac{-3r^2}{2N\delta^2}\right)$$

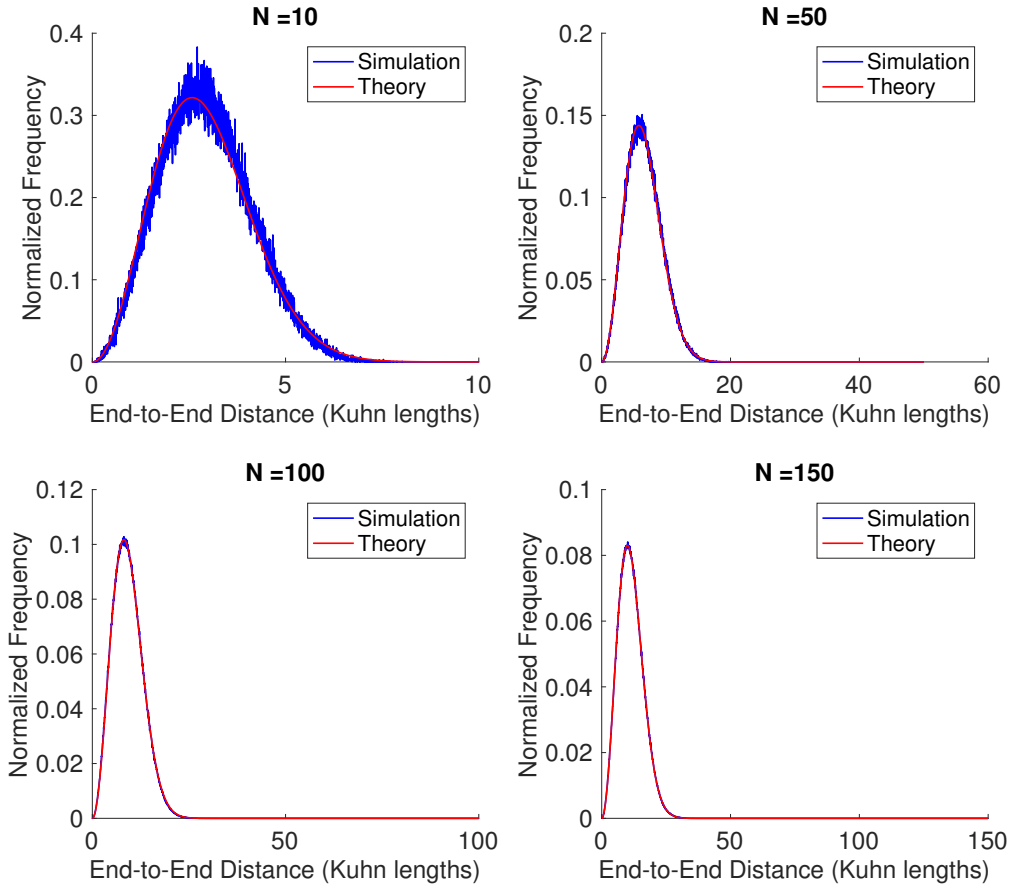


Figure 2: Simulated end-to-end distance distribution (blue) against theoretical distribution (red) for multiple polymer lengths (N).

and show derivation (?) of distribution formula?

1.2.3 Occlusion Probability of End Site VS Analytical Result

We consider analytical solutions to the simplest case, where the ligand is attempting to bind at the end of the freely-jointed chain in free-space. The probability that a random walk over time 0 to $\delta * N$, beginning at the edge of a sphere of radius, R, will cross into the sphere at any time is analagous to the probability the ligand is occluded by the freely jointed chain. This is a confusing description In order to solve this analytically, we must assume the random walk begins ϵ away from the sphere or it will always count as a failure. Then we may formulate as follows for the probability of survival, $p(\vec{x}, t)$ when the random walk begins at a starting point, $r = R + \epsilon$:

I got confused.

$$\begin{cases} \frac{\partial p}{\partial t} = D\nabla^2 p, & p = 0 \quad \text{at} \quad r = R \\ p(\vec{x}, 0) = \delta(\vec{x} - (R + \epsilon)) \end{cases} \tag{12}$$

In order to get a finite estimate to compare, we consider the solution when $N \to \infty$.

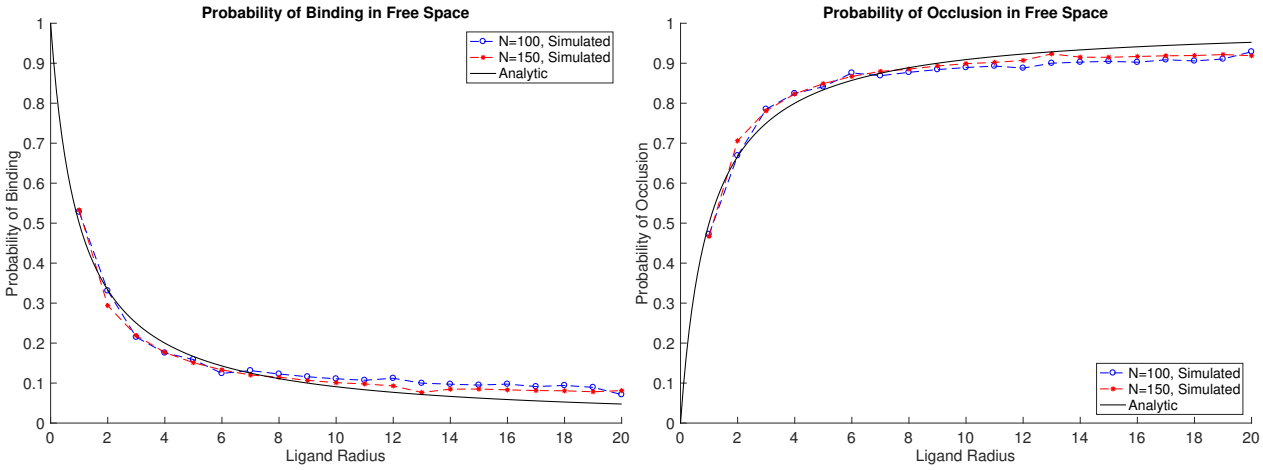
$$\begin{cases} q(r) = \mathbb{P}(\text{not hit sphere} | x(0) = r) \\ 0 = \frac{2}{r} q'(r) + q''(r) \\ q(R) = 0 \\ q(r) \to 1 \quad \text{as} \quad r \to \infty \end{cases} \tag{13}$$

We assume $\epsilon = 1$ Kuhn length and R measured in Kuhn lengths:

$$q(r) = 1 - \frac{R}{r} q(r) = 1 - \frac{R}{R+1} \tag{14}$$

We see that when we compare our simulated binding probability to the analytic solution, we get good agreement. This tells us both our code is working as desired and that N=100 is approximately equivalent to $N \to \infty$.

pick one of these - don't need both



(a) (b)

Figure 3: Analytic Solution

1.3 Case Study: T Cell Receptor Zeta Chain

For the following study, we will focus on the mouse TCR CD3 ζ chain. The TCR CD3 ζ chain is a subunit of CD3 consisting of 164 amino acids. Of these, 21 are included in the signal peptide region of the protein. The remaining 143 amino acids make up the extracellular, transmembrane and cytoplasmic regions of the CD3 ζ chain. The cytoplasmic tail is an intrinsically disordered chain of 113 amino acids containing multiple phosphorylation sites, called ITAMs (immunoreceptor tyrosine-based activation motif). There are three ITAMs on the ζ chain, each containing two tyrosines. The tyrosine kinase Lck phosphorylates each tyrosine

in the ζ chain and it is thought that phosphorylation of one site influences the binding of the kinase to other sites.

In mouse CD3 ζ, the cytoplasmic tail spans residues 52-164 and the tyrosines are located at residues 72, 83, 111, 123, 142, 153. Therefore, if we were to renumber to begin at the beginning of the cytoplasmic tail, the region would be $N = 113$ amino acids long, with tyrosines located at $i = 21, 32, 60, 72, 91, 102$. (UniProt, entry P24161) Given an assumption of 0.3 nmper Kuhn length (i.e. one Kuhn length is equivalent to one amino acid), then the tyrosines are similarly located along the 113 segments of the FJC.

Cite UniProt? Or find paper?

Mouse Lck is composed of an SH3, SH2 and protein kinase domain connected by small loops. The domains are 61, 98, and 254 amino acids respectively. (UniProt entry P06240) Using a protein molecular mass calculator (http://www.bioinformatics.org/sms/prot_mw.html) , we calculate that the kinase domain is 29.08 kDa. If we assume a protein density of 1.41 g / cm^3 then we can estimate the volume of the kinase domain [**Fischer2004**]:

Cite UniProt? Or find paper?

Cite molecular mass calculator?

$$(29 \times 1000\text{Da}) * (1.66 \times 10^{-27}\text{kg / Da}) * (1000\text{g / kg})/(1.41\text{g / cm}^3) = 34\,\text{nm}^3.$$

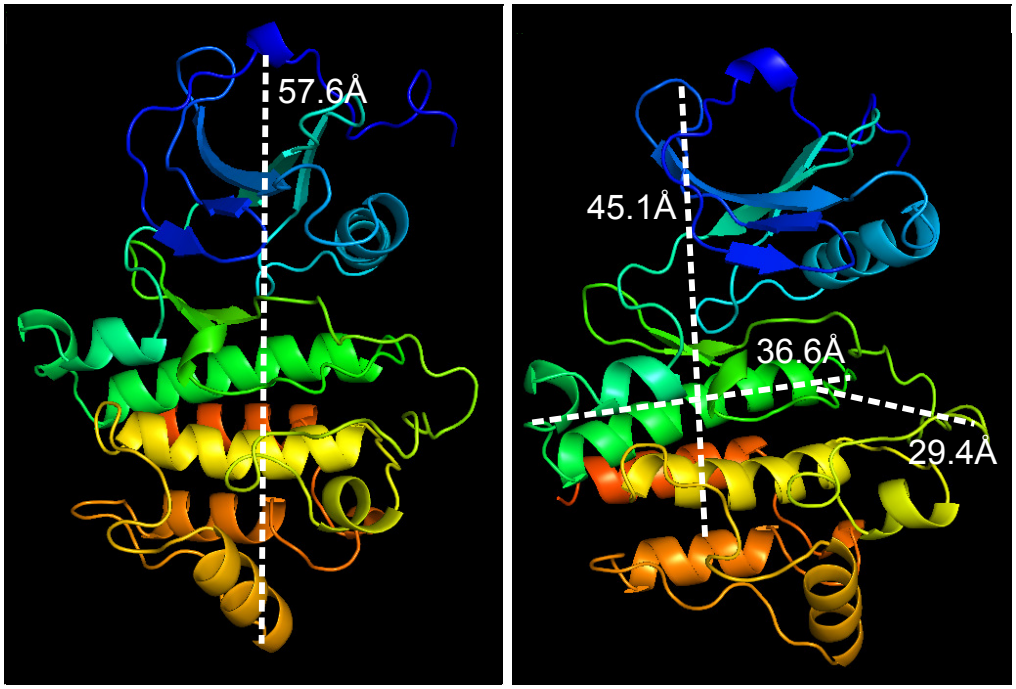
If we then approximate the kinase domain as a sphere, then we can estimate a radius:

$$\begin{aligned} V &= \frac{4}{3}\pi r^3 \\ 34\,\text{nm}^3 &= \frac{4}{3}\pi r^3 \\ r &\approx 2\,\text{nm} \end{aligned}$$

Measurements from the crystal structure of Lck suggest a volume of 45 nm^3 for the kinase domain.Where did this come from? This estimate is on the same order of magnitude as our previous estimate. Note also, if we recalculate the kinase radius using a volume of 45 nm^3, we estimate a radius of 2.2 nm.

Cite or show PyMol for Lck

We measure the maximal length of the kinase domain in PyMol from PDB 3LCK. Rounding up for error, we have a maximal distance of 58Å. This gives a maximal spherical estimate with a radius of 2.9 nm, or about ten Kuhn lengths. (Fig. 4a) If we instead measure rough length, width, and height for the kinase domain, we have measurements of 36.6Å, 29.4Å, and 45.1Årespectively. (Fig. 4b) From these we can estimate a sphere with volume corresponding to the volume of the rectangular prism with those dimensions. This estimates a sphere with radius 2.3 nm, or about eight Kuhn lengths. Based on all of these estimates, we choose to represent Lck with a radius of seven Kuhn lengths.



(a) (b)

Figure 4: PDB 3LCK - kinase domain of Lck

Introduce estimates of ZAP-70 SH2 domain

1.4 Possibly unnecessary stuff

where does this stuff go?

In this model, the FJC is allowed to pass through itself. However, this turns out to be unimportant. Although the disordered protein would have a nonzero bond width, this size would be small in comparison to the kinase volume and volume of exploration space. If you consider two thin strings in 3-dimensional space, they will almost never intersect. Therefore, we model our polymer with infinitely thin width. Since it is infinitely thin, the time when it would take on a conformation where it overlaps itself is negligible compared to the number of legal conformations it assumes.

Additionally, we are only interested in the ensemble of configurations, not the dynamics. Therefore, the way the polymer achieves its conformations is unimportant to the results and any instances of self-intersection are negligible compared to the larger ensemble. Allowing the FJC to pass through itself, while unphysical, does not change the validity of our simulation.