1 Model Development

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 model the polymer in the canonical ensemble, i.e. equilibrium. Steric occlusion of binding sites by the rest of the chain therefore gives rise to a change in K_D. We can compute the difference in K_D between two states, e.g. fully phosphorylated compared to unphosphorylated. From detailed balance, we have $K_D \equiv k_{\text{off}}/k_{\text{on}}$, which allows us to compute the change in K_D as follows:

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

$$\frac{K_{D_1}}{K_{D_2}} = \frac{\frac{1}{\exp\left(\frac{-\Delta G_1}{k_B T}\right)}}{\frac{1}{\exp\left(\frac{-\Delta G_2}{k_B T}\right)}} \tag{2}$$

$$= \frac{\exp\left(\frac{-\Delta G_2}{k_B T}\right)}{\exp\left(\frac{-\Delta G_1}{k_B T}\right)} \tag{3}$$

$$= \exp\left(\frac{\Delta G_1 - \Delta G_2}{k_B T}\right) \tag{4}$$

$$= \exp\left(\frac{(E_1 - TS_1) - (E_2 - TS_2)}{k_B T}\right) \tag{5}$$

$$= \exp\left(\frac{S_2 - S_1}{k_B}\right) \tag{6}$$

$$= \exp\left(\frac{(S^2_{on} - S^2_{off}) - (S^1_{on} - S^1_{off})}{k_B}\right) \tag{7}$$

$$= \exp\left(\frac{(k_B \ln(\frac{\Omega_2 P_2}{\Omega_2})) - (k_B \ln(\frac{\Omega_1 P_1}{\Omega_1})}{k_B}\right) \tag{8}$$

$$\frac{K_{D_1}}{K_{D_2}} = \frac{P_2}{P_1} \tag{9}$$

$$\tag{10}$$

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 $\Omega_1 P_1$ being the microstates available when the ligand is bound and $\Omega_2 P_2$ is the number of microstates available when the polymer is rigid. We also note $\Delta E_1 = \Delta E_2$ since the change in energy due to ligand binding will be the same, regardless of conformation.

For example, we can consider the change in K_D between a fully rigid polymer, where all binding sites are available and $K_D = K_{D_R}$, compared to a floppy, unstructured polymer, where binding site availability

depends on the conformation of the polymer and $K_D = K_{D_F}$. Then we calculate the change in K_D as above, but since the rigid state always allows binding, we may write $P_R = 1$.

$$\frac{K_{D_F}}{K_{D_R}} = \frac{\frac{1}{\exp\left(\frac{-\Delta G_F}{k_B T}\right)}}{\frac{1}{\exp\left(\frac{-\Delta G_R}{k_B T}\right)}} \tag{11}$$

$$= \frac{P_R}{P_F} \tag{12}$$

$$\frac{K_{D_F}}{K_{D_R}} = \frac{1}{P_F} \qquad \text{assuming } P_R = 1 \tag{13}$$

$$\tag{14}$$

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, $P_{occ} = 1 - P_F$. Our problem of interest is now reduced to computing the occlusion probability.

$$\frac{K_{D_F}}{K_{D_R}} = \frac{1}{(1 - P_{occ})}. \tag{15}$$

We have already used this comparison as a possibility to account for the change in phosphorylation rates of the TCR ζ chain [**?**]. We consider the generalized version for the rest of our work.

Steric occlusion by the polymer will impact the ability of a ligand to localize to the binding site. Although entropic forces could also impact unbinding of the polymer, we assume this influence to be negligible compared to the change in k_{on}. Therefore, for the body of this work, we assume that the change in K_D stems from a change in k_{on}.

We calculate the occlusion probability by computing how often a ligand is able to bind to an oriented sphere tangentially attached to the polymer. We define 'able to bind' as when the specified sphere is empty of both other polymer segments and any surface constraints. 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. If a surface is present, we also check if the ligand 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.

1.1 Code Validation

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

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 $\sqrt{N}$. We also consider the distribution of end-to-end distances, which is known from previous work [**? ?**].

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

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

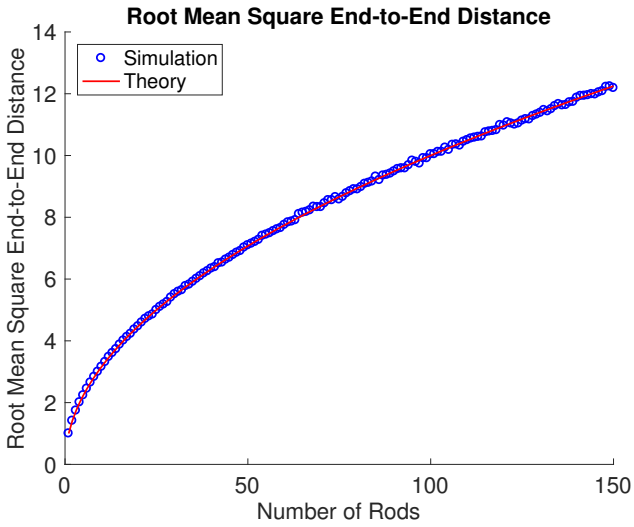


Figure 1: Theoretical root mean square end-to-end distance (red line) against simulated values (blue circles).

End-to-end distribution [**? ?**]:

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

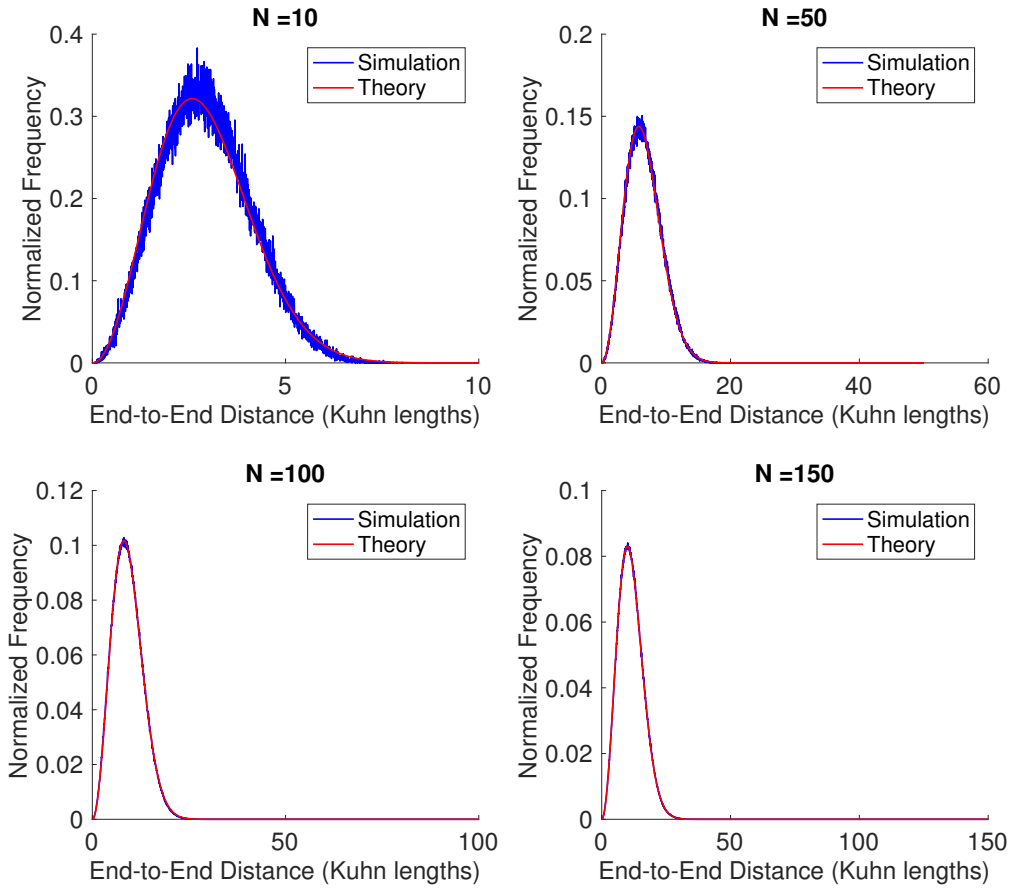


Figure 2: Simulated end-to-end distance distribution (blue) against theoretical distribution (red) for multiple polymer lengths (N).

1.2 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.

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 0.3nm per 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.

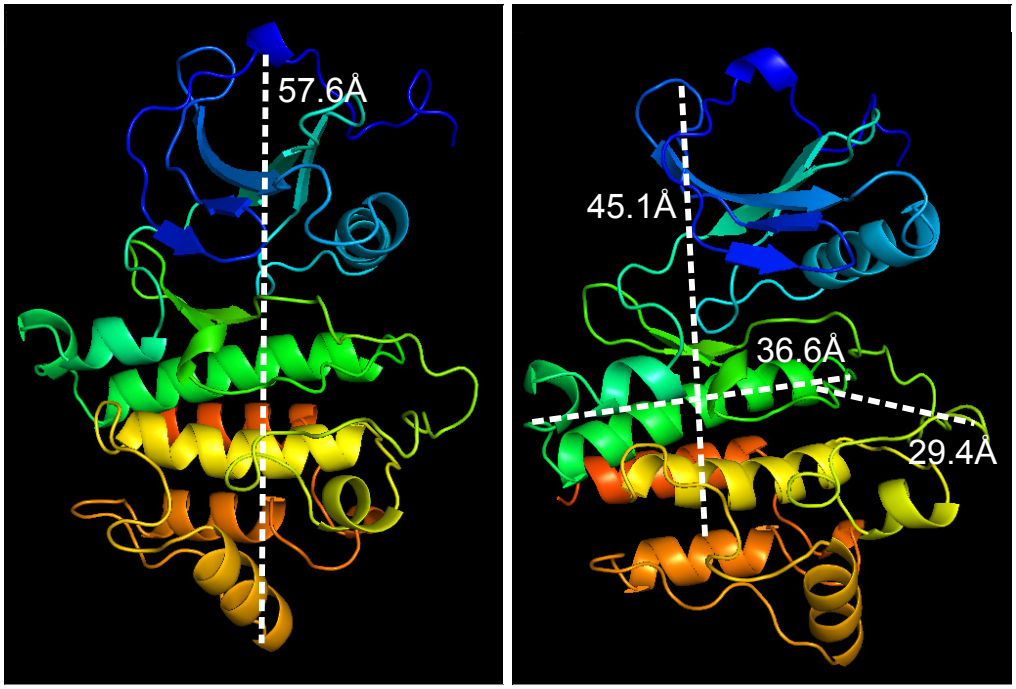
Mouse Lck is composed of an SH3, SH2 and protein kinase domain connected by small loops. The domains are 61, 98, and 254 amino acids respectively (UniProt entry P06240). Using a protein molecular mass calculator, we calculate that the kinase domain is 29.08 kDa. If we assume a protein density of 1.41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

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

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

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

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. 3a). 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. 3b). 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 3: PDB 3LCK - kinase domain of Lck