PHYS180 Project 6
Protein Folding

Karoli Clever, Rey Cervantes, Owen Morehead

May 21, 2021

1. Protein Folding: Foldit Program, all

Owen: Completed 12 of of the puzzles. They were fun and a good way to learn about protein structure and design. Also tried my best at a handful of science puzzles. I moved my way up to 7th place in the Docking Design science puzzle (protein in the figure). The goal of this puzzle involves designing a small protein section to best bond to a specific spot on the larger protein. I kept the protein design as a helix as other structures were scoring lower. Overall I felt like I should have been able to better bond the two sections but this was the best I could do without spending too much time!

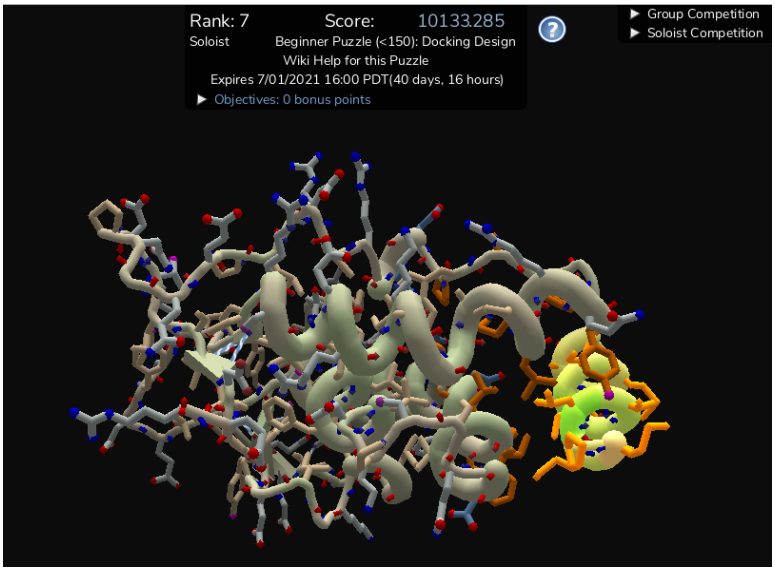


Figure 1: My best attempt at the Docking Design science puzzle in the Foldit program. My high score was 10133 which put me in 7th place. Not as great as I wanted but not bad!

Rey: Completed all of the introduction and beginner's puzzles. The science puzzle that I took on was the Coronavirus one. I made it to 33rd place in this puzzle, and my protein can be seen in the figure below. I think I could have done a lot better if I dedicated more time to it, but I think I had a really nice increase in ranking. I think the Foldit program was very fun to explore with, and I will be trying out other puzzles with my little sister!

Karoli: Did the intro and the beginners puzzles, tried intermediate ones, but they are quite the time commitment. Did alright on one of them, The Protein Design Sandbox, and went up from rank 1800+ to rank 426 score started at (-25000+) and I got to 1800, so a gain in points of nearly 27000. The second best was a beginners puzzle, Thioredoxin, where I ranked 120 of 800+ and got a score starting at 0 to 10770.427. Third highest score of my trials was Influenza HA, from rank 700+ to rank 22, score also starts at 0 rounded but is actually at -2195.070, so to get my score of 2319.354 I actually gained nearly 5500 points. It was fun,

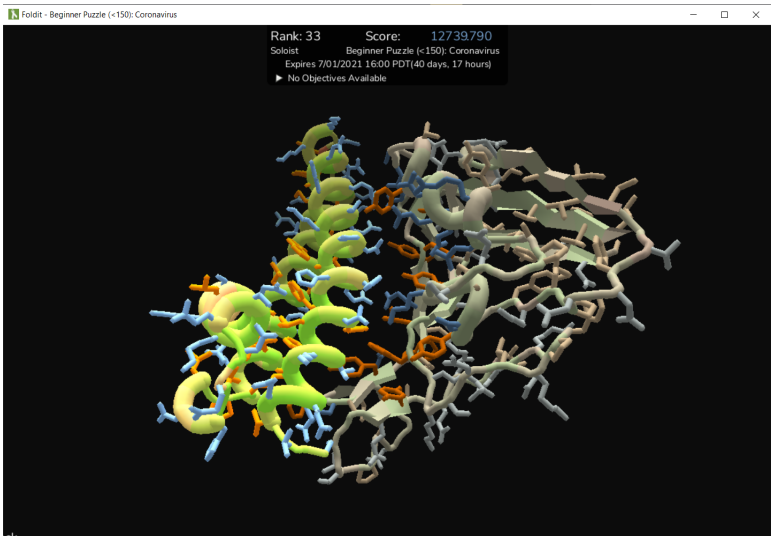


Figure 2: This shows my attempt at the Coronavirus science puzzle. My ranking was 33 and my high score was 12739.790.

but again, takes a lot of time to actually get to the point where side chains can be added or mutated, etc. Nice to know the program though! Will send images in a pdf with descriptions in my personal assignment submission.

2. Coil Globual Transition

Pt 1.

Here we explore a simulation using reptation dynamics for a lattice model describing a polymer chain with attractive interactions with its neighbors but is also self avoiding (can't occupy lattice site twice). We run the code coil_globule.py for different parameters. We run the simulation for different chain lengths (N = 32, 64), and also for different values of attractive coupling, J. J = 0 means no attraction between nearest neighbors, and so in this phase we expect the polymer to be self avoiding. The average radius of gyration, R, as a function of chain length, N, should obey:

$$\mathbf{R} \propto N^{v} \tag{1}$$

where v is a constant, and in three dimensions is about 0.6.
If J is increased, say $J = 0.5$, there is attractive coupling between nearest neighbors in the chain and so the dynamics represent a collapsed (globule) phase. In this phase the chain should be of constant density, that is the ratio in equation 1, N/R^3, or $R/N^{1/3}$ should be constant so that $v = 1/3$. In order to test these values we can calculate from the simulation the average radii of gyration for J = 0 (N = 32 and 64), and for J = 0.5 (N = 32 and 64). From this we can obtain a numerical estimate for the two different values of v. Because we don't know the proportionality factor between R and N, let us estimate v by taking ratios of the simulations. We can define these ratios from equation 1 but with the added unknown scaling constant, C.

$$C\mathbf{R} \propto N^{v} \rightarrow C = N^{v}/\mathbf{R} \tag{2}$$

So for J = 0 or J = 0.5, consider the ratio where N = 32 and 64:

$$C = const = 32^{v}/\mathbf{R}_1 = 64^{v}/\mathbf{R}_2 \rightarrow (\frac{32}{64})^{v} = \frac{\mathbf{R}_1}{\mathbf{R}_2}. \tag{3}$$

The simulation gives us values $\mathbf{R}_1 = 3.143$ and $\mathbf{R}_2 = 4.78$. Therefore, for the self avoiding coil phase (J = 0):

$$(\frac{32}{64})^v = \frac{3.142}{4.782} \rightarrow v = \frac{ln(3.142/4.782)}{ln(32/64)} \rightarrow \boxed{\text{v = 0.606}} \tag{4}$$

And for the collapsed globule phase (J = 0.5), we estimate a value of v:

$$(\frac{32}{64})^v = \frac{2.394}{3.072} \rightarrow \frac{ln(2.389/3.018)}{ln(32/64)} \rightarrow \boxed{\text{v = 0.337}} \tag{5}$$

The simulation was run for approximately 2000 measurements and after taking the average gyration radius from all of the measurements we see that our simulation results, $v = 0.606$ and $v = 0.337$ agree well with the theoretical values, $v = 0.6$ and $v = 0.333$. Running these simulations for different chain lengths is nice in that it lets us measure parameters such as v by simply taking ratios of the two simulation results. This eliminates the need to know any proportionality factor between $\mathbf{R}$ and N because they are eliminated when taking ratios. This nicely illustrates the power of scaling in fractal systems.

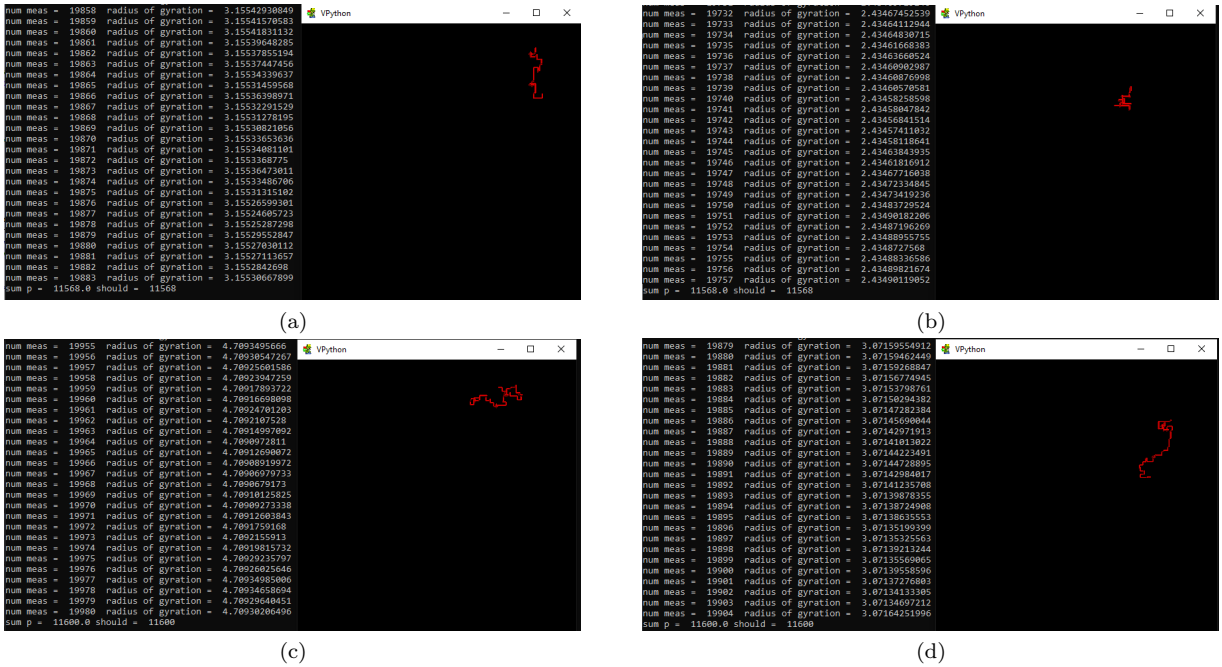


Figure 3: Running the python code for (a) N=32, J=0, (b) N=32, J=0.5, (c) N=64, J=0, and (d) N=64, J=0.5

Pt 2. Applications to Biology - Method of Inducing Coil Globule Transition

In their paper, Zhu et al. (2017) discuss the process of gene therapy and the purpose of carrying recombinant DNA inside the cell in order to achieve gene expression, as well as to correct or compensate for genetic defects and diseases caused by abnormal conditions. Due to gene conformational and electronic reasons, transport of DNA across the cell membrane proves difficult. Agents of interest which are acting as gene vectors include cationic surfactants because of their unique structures. Furthermore, they can reduce the charge repulsion via binding to DNA backbones, which leads to compaction of the DNA, helping it cross cell membranes. When a photoactive function group is introduced into the ionic liquid surfactants, new ionic liquid surfactant-DNA complexes could possibly be produced. Zhu et al. (2017) point out that research in the applications in the DNA photocleavage and DNA condensing agents is of high importance. In their experiment a photoactive isoquinoline ring was introduced into the ionic liquids to synthesize the photoactive ionic liquid surfactants. The binding characteristic of the photoactive ionic liquid surfactants with DNA were investigated by UV–vis spectroscopy, steady-state and time-resolved fluorescence spectroscopy, dynamic light scattering (DLS), cryogenic transmission electron microscopy (cryo-TEM), circular dichroism (CD) spectroscopy, FT-IR spectroscopy, isothermal titration microcalorimetry (ITC), 1H NMR and 2D-NOESY. Zhu et al. (2017) state that the binding characteristic of the ionic liquid surfactants on DNA, such as the all-or-none type transition, and the effect of the chain length of the ionic liquid surfactants were obtained. Furthermore, the binding mode of the ionic liquid surfactants on DNA was revealed by investigating the binding sites. The unique binding behavior of the photoactive ionic liquid surfactants on DNA would provide an important insight in the design of a DNA photosensitizer for photodynamic therapy or DNA compensation for gene transportation across cell membrane (Zhu et al., 2017).

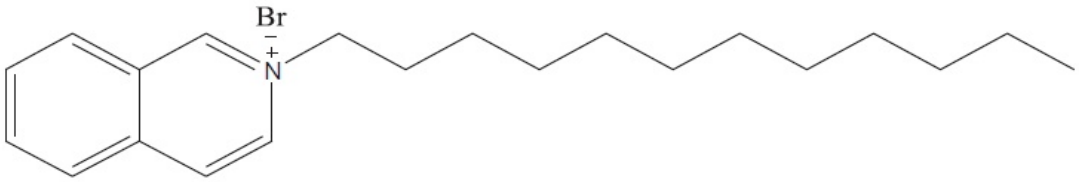


Figure 4: Visible the chemical structure of [C_{12}iQuin]Br (Zhu et al., 2017).

Methods and Steps – direct citation as listed in the paper to avoid misinterpretation through paraphrasing, all credits go to Zhu et al. (2017)

1) The ionic liquid surfactant, lauryl isoquinolinium bromide([C_{12}iQuin]Br), was synthesized with purity of 98 %.
2) The absorption spectra of [C_{12}iQuin]Br/DNA were recorded in the wavelength range of 200–400 nm on a Perkin-Elmer Lambda-650S spectrophotometer at 25 $\pm$0.1°C, using a 1 cm path length quartz cell.
3) Steady-state fluorescence spectra were performed with a Hitachi F-7000 fluorescence spectrophotometry.
4) The fluorescence spectroscopy of the intercalated probe EB and the minor groove bound probe DAPI were conducted, and their intensities were recorded by keeping the stoichiometric ratio of DNA (0.05 mM) and dye constant.
5) The dynamic light scattering experiments were performed with an ALV (CGS-8F) laser light scattering spectrometer working in a pseudo cross-correlation mode.
6) The cryo-TEM samples were prepared as follows: a small droplet of each sample solution was placed on a carbon-coated TEM copper grid. The excess solution was blotted by filter paper, and the sample grids were then quickly plunged into liquid ethane and kept there before imaging.
7) A JASCO J-810 spectropolarimeter was used to perform CD spectra. The measurements were performed under a constant nitrogen flow, which was used to purge the ozone generated by the light source of the instrument.
8) The FT-IR spectra of DNA, [C_{12}iQuin]Br, and DNA/[C_{12}iQuin]Br complexes in water were recorded on a Bruker VERTEX 70v IR spectrometer by a coaddition of 312 interferograms collected at a 4 cm^{-1} resolution.

9) The binding mechanism of [C_{12}iQuin]Br on DNA was further investigated by 1H NMR and 2D-NOESY experiments.
10) Zeta potential measurements were performed by laser Doppler electrophoresis using a Zetasizer Nano ZS90 spectrometer in a standard DTS 1060 zeta cell.
11) ITC measurements were conducted in order to determine the binding enthalpy between [C_{12}iQuin]Br and DNA. The experiments were performed using a VP-ITC calorimeter at 25 °C.

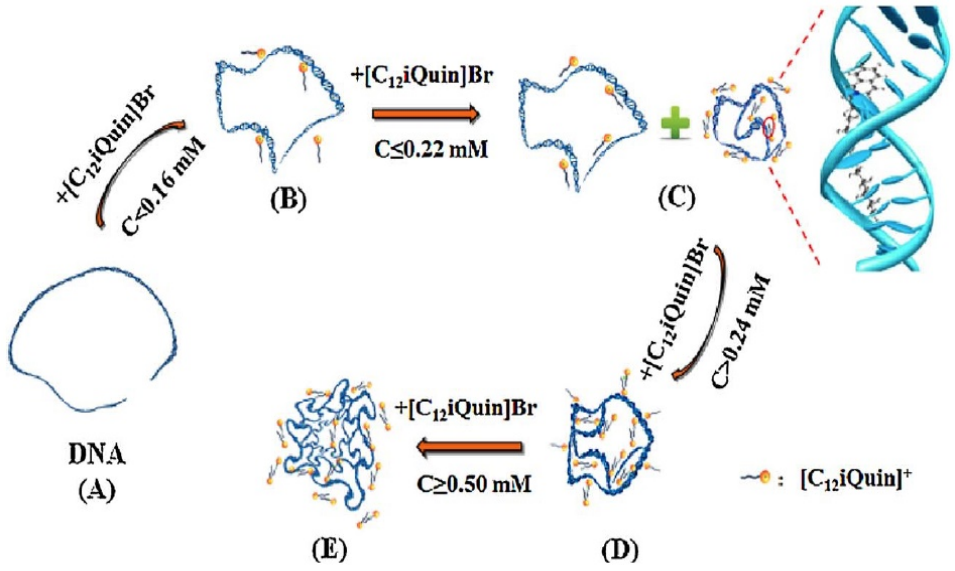
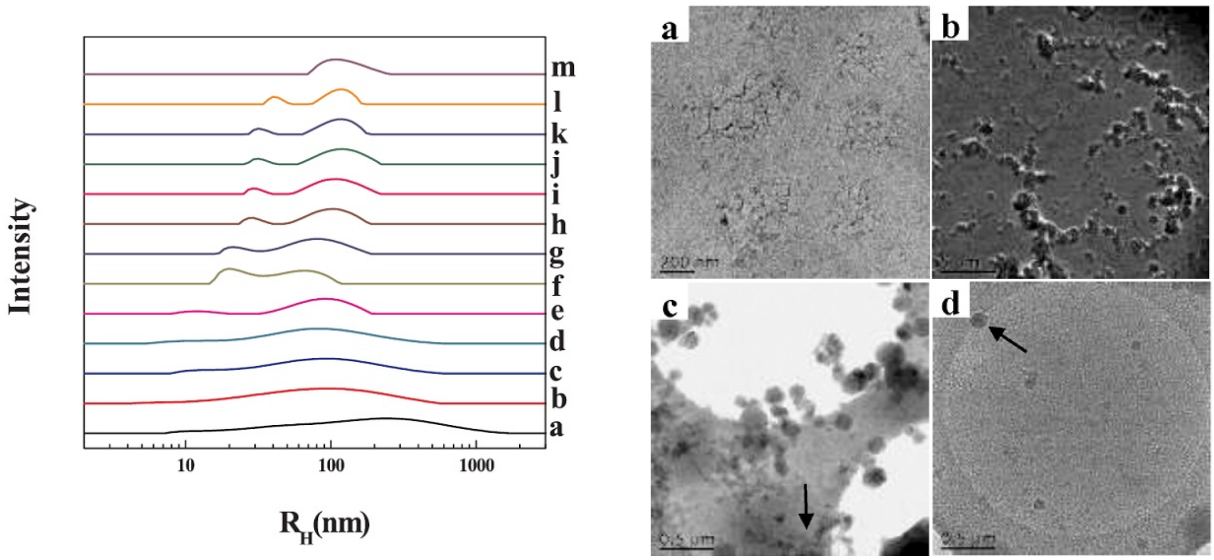


Figure 5: Mechanism showing photoactive ionic liquid surfactant [C_{12}iQuin]Br inducing coil globule structure transition and binding mechanism of DNA (Zhu et al., 2017).

Results – key points, directly cited as listed in the paper to avoid misinterpretation through paraphrasing, all credits go to Zhu et al. (2017)

1) The binding behavior of the ionic liquid surfactant [C_{12}iQuin]Br with DNA in PBS buffer was monitored by UV–vis spectroscopy.
2) Fluorescence spectroscopy was used to reveal the binding extent of [C12iQuin]Br on DNA.
3) Ethidium bromide (EB), as an intercalation fluorescence probe, can intercalate into the base pairs of double-stranded DNA and stretches the double helix of DNA. The increase of the fluorescence intensity of the probe arises from the hydrophobic micro-environment between the base pairs, which protects the probe from the quenching of water molecules and molecular oxygen.
4) In order to investigate the micro structure changes of DNA induced by [C_{12}iQuin]Br, the hydrodynamic radii were determined by DLS in the presence of different concentrations of [C_{12}iQuin]Br. Furthermore, the structural change of DNA is ascribed to the electrostatic binding of [C_{12}iQuin]Br to DNA and the strong hydrophobic association between the hydrocarbon chains of [C_{12}iQuin]Br and hydrophobic bases of DNA. When [C_{12}iQuin]Br is added to the DNA system, it is bound to DNA, which is driven by the electrostatic and hydrophobic interactions between [C_{12}iQuin]Br and DNA. Such cooperative interactions can promote the DNA compaction and make the DNA-[C_{12}iQuin]Br complexes to undergo a significant coil-globule transition of DNA. The resulting compacted DNA structures show mostly globule forms and observed by cryo-TEM.
5) To verify the micro structure of DNA/[C_{12}iQuin]Br complexes, cryo-TEM was performed to examine the morphology change of the complexes.
6) To reveal the change of the secondary structure of DNA during the binding process of [C_{12}iQuin]Br to DNA, the circular dichroism spectra in absence or presence of [C_{12}iQuin]Br were recorded.
7) In order to check the effects of the alkyl chain lengths of the ionic liquid surfactants on the binding behavior, the relative fluorescence intensity of DNA-bound EB, and the position band intensity of CD of

DNA were recorded upon addition of [C_niQuin]Br with different chain lengths.
8) For gaining insights into the interaction mechanisms of [C_{12}iQuin] Br and DNA, FT-IR spectroscopy was used to investigate the interaction sites of [C_{12}iQuin]Br on DNA.
9) To further detect the specific binding sites of [C_{12}iQuin]Br on DNA, a similar oligonucleotide to the fragment of calf thymus DNA was used in the 1H NMR and 2D-NOESY measurements.
10) The zeta potential can reveal the specific binding state of [C_{12}iQuin]Br on DNA surface.
11) The thermodynamic behavior of [C_{12}iQuin]Br-DNA binding associated with the interaction between them was investigated by isothermal titration microcalorimetry (ITC).
12) Based on the above experimental evidences the binding mode of [C_{12}iQuin]Br on DNA was proposed. When the [C_{12}iQuin]Br concentration is below 0.16 mM, the DNA molecules exist in a free stretch mode in nature state, and the individual [C_{12}iQuin]Br molecules are bound on the DNA to form [C_{12}iQuin]Br-DNA complexes. The electrostatic attraction between the cationic isoquinolinium ring and the negative phosphate groups, and the hydrophobic interaction between the hydrocarbon chains of [C_{12}iQuin]Br and the hydrophobic bases of DNA make the [C_{12}iQuin]Br-DNA complexes to be compacted, accompanied by the rearrangement of DNA helix at the same time. Upon addition of [C_{12}iQuin]Br at the concentration between 0.16 mM and 0.22 mM, the electrostatic and hydrophobic forces make the DNA structure transition from coil to globule forms, as observed by DLS and cryo-TEM results. When the [C_{12}iQuin] Br concentration reaches 0.24 mM, the compression of DNA induced by[C_{12}iQuin]Br is almost finished. Thereafter, both the coiled and globular DNA structures grow in the sizes with addition of [C_{12}iQuin]Br, which can be attributed to the hydrophobic binding of the added [C_{12}iQuin]Br molecules to those bound on the DNA surface. Finally, the [C_{12}iQuin]Br molecules bound on the coiled and globular DNA surfaces bind with each other via the association of their hydrocarbons, resulting in the fusion of the two structures and thus the formation of uniform globular [C_{12}iQuin]Br-DNA complexes of about 100 nm. During the binding process of [C_{12}iQuin]Br to DNA, the hydrophobic interactions are the dominant driving force (Zhu et al., 2017).



(a) (b)

Figure 6: (a) Hydrodynamic radius of DNA solution in the presence of [C_{12}iQuin]Br in PBS solution at 25 °C. The concentration (mM): a-0; b-0.04; c-0.08; d-0.12; e-0.16; f-0.20; g-0.22; h-0.24; i-0.26; j-0.28; k-0.30; l-0.34; m-0.50. DNA concentration: 0.05 mM. (b) Cryo-TEM images of DNA structures in the presence of [C_{12}iQuin]Br in PBS solution at 25 °C. [C12iQuin]Br concentration (mM): a-0; b-0.12; c-0.26; d-0.50. (a, b: figures and captions by Zhu et al., 2017).

Zhu et al. (2017) conclude that the photoactive [C_{12}iQuin]Br can be effectively bound to DNA by the electrostatic attraction between the cationic isoquinolinium ring and the negative phosphate groups of DNA, and the hydrophobic interaction between the hydrocarbon chains of [C_{12}iQuin]Br and the DNA bases. The obtained binding constant calculated suggests that the binding is with multiple sites. The intercalated fluorescence probe EB and the minor-groove-bound probe DAPI can be extruded from DNA in the presence of [C_{12}iQuin]Br, confirming the strong binding of [C_{12}iQuin]Br on DNA (Zhu et al., 2017). The [C_{12}iQuin]Br can induce a complete coil-to-globe transition of DNA, confirmed by cryo-TEM images, showing that the DNA structures, such as the binding compactness, binding sites, and DNA conformation can be adjusted by changing the molar ratio of [C_{12}iQuin]Br to DNA, as well as the chain length of [C_niQuin]Br, and the base packing and helical structure of DNA varied largely but maintained the B-form state (Zhu et al., 2017). Zhu et al. (2017) state that according to the FT-IR, 1H NMR and 2D NOESY experimental evidence, the binding mode at a molecular level were proposed. At a lower [C_{12}iQuin]Br concentration, the [C_{12}iQuin]Br molecules lie down at the AT sites in minor groove of DNA, the isoquinolinium ring would be arranged facing to the plane of the minor groove of DNA, and the hydrophobic tails of [C_{12}iQuin]Br would be arranged parallel to the minor groove. Zhu et al. (2017) further concluding that the H7 of the isoquinolinium ring is localized within several angstroms of the DNA phosphates on one strand of DNA, and the H4 of the ring is localized near the sugar 1 hydrogen protons on the other strand of DNA. When the [C_{12}iQuin]Br concentration is larger than 0.24 mM, the binding of [C_{12}iQuin]Br molecules in the minor groove of DNA has reached saturation. The added [C_{12}iQuin]Br molecules would be bound on the DNA surface with the cationic isoquinolinium ring near DNA phosphates, and the hydrocarbon chains attached to the DNA surface at a certain angle. During the binding process the hydrophobic interaction between the hydrocarbon chains of [C_{12}iQuin]Br and DNA bases provides the dominate driving force. The unique binding mode of [C_{12}iQuin]Br on DNA causes the formation of [C_{12}iQuin]Br-DNA complexes with different structures. Furthermore, the introduction of the photoactive group makes the formed [C_{12}iQuin]Br-DNA complexes to show high sensitivity in the optical detection. Thus, [C_{12}iQuin]Br could be a potential photosensitizer for DNA cleavage (Zhu et al., 2017).
The authors (Zhu et al., 2017) argue that results may promise important applications of isoquinoline-based photoactive ionic liquid surfactants in the design of DNA photosensitizer for photodynamic therapy or DNA compensation for gene transportation across cell membrane.

Comparison with transition of a protein from denatured to folded

To get a better idea of how our model of coil-globule transition compares to protein folding and other models, three papers were chosen that report experiments to this topic. The general idea of each research is summarized below, and after going through them, it is a consensus that our model works quite well when compared to protein folding or denaturing. Of course this is a very well-covered topic, and overall quite complex depending on the specific interest of each research group, but models are getting better and new technologies and research techniques have made it possible to continue models that allow the best understanding of the processes to date.

Luo et al. (2014) performed an experiment pertaining to an end-grafted hydrophobic-polar (HP) model protein chain with alternating H and P monomers to examine interactions between the critical adsorption transition due to surface attraction and the collapse transition further due to pairwise attractive HH interactions. Critical adsorption temperature TCAP is influenced by the attractive HH interactions in some cases. Should collapse temperature Tc be lower than TCAP, the critical adsorption of the HP chain is similar to that of a homopolymer without intrachain attractions and TCAP remains unchanged. Yet the collapse transition is suppressed by the adsorption and for cases where Tc is close to or higher than TCAP, TCAP of the HP chain is increased, indicating that a collapsed chain is more easily adsorbed on the surface. The strength of the HH attraction also influences the statistical size and shape of the polymer, with strong HH attractions resulting in adsorbed and collapsed chains adopting two-dimensional, circular conformations. Luo et al. (2014) first determine the collapse transition of the alternating HP model chain in a dilute solution, and the chain is annealed from a high to a low temperature, next simulation of the adsorption of the end-grafted HP chain with EHH = -1 by annealing the chain with the head H monomer grafted on a flat surface and estimate the collapse transition and the critical adsorption transition temperatures. The coilglobule transition temperature for an end-grafted chain was estimated from the temperature dependence

of the mean square end-to-end distance. Observations were made between the critical adsorption of a lattice HP protein with alternating H and P monomers and the coilglobule transition with the dynamical Monte Carlo method. Simulations are carried out in the simple cubic lattice, finding that the critical adsorption temperature TCAP is influenced by the presence of intrachain attractions responsible for the collapse transition of the polymer. There is more difficulty in a surface to be absorbed HP polymer chain to go through the coilglobule collapse than one that is free in solution, yet if the intrinsic coilglobule transition temperature Tc 0 is higher than TCAP, a collapsed chain can be more easily adsorbed, and the conformational properties of the end-grafted HP chain are strongly influenced by the pairwise HH attraction. The research team state in their overall conclusion about the mutual impact on the coilglobule transition and the critical adsorption transition is likely still valid (Luo et al., 2014).

The Maffie et al. (2012) research team presents a Vibrating Interacting Self-Avoiding Walk (VISAW) model that is a natural generalization of the Interacting Self-Avoiding Walk (ISAW), stemming from basic considerations about the structure of a realistic potential compatible with the ISAW approximations. There were inclusions of fluctuations around inherent structures in the calculation of the partition function. This is coherent with modern developments in condensed matter physics and provides a clearer picture of the CGT. Maffie et al. (2012) say that the change in the order of the transition is brought about by the force constant k of non-bonded contacts, due to the depleting effect of low-frequency modes for globular conformations, being crucial mention that it is not the absolute flexibility of the polymers that determines the transition order change, but rather the softness difference between globule and coil conformations (Maffie et al., 2012).

Dai et al. (2015) are interested in why single polymer chains undergo a phase transition from coiled conformations to globular conformations as the effective attraction between monomers becomes strong enough. The authors investigated the coil-globule transition of a semiflexible chain confined between two parallel plates, a slit, using the lattice model and Pruned-enriched Rosenbluth method (PERM) algorithm. Dai et al. (2015) found that as the slit height decreases, the critical attraction for the coil-globule transition changed non-monotonically due to the competition of the confinement free energies of the coiled and globular states. In wide/narrow slits, the coiled state experiences more/less confinement free energy, therefore the transition becomes easier/more difficult. The authors conclude that their simulations reveal that with decreasing slit height the critical attraction for the coil-globule transition first decreases and then increases, due to the competing of the confinement free energies of the coiled state and the globular states. This suggests that the critical slit height corresponding to the minimum critical attraction mainly depends on the smallest dimension of the globule transition, and that the coil-globule transition becomes less sharp with the decreasing slit height (Dai et al., 2015).

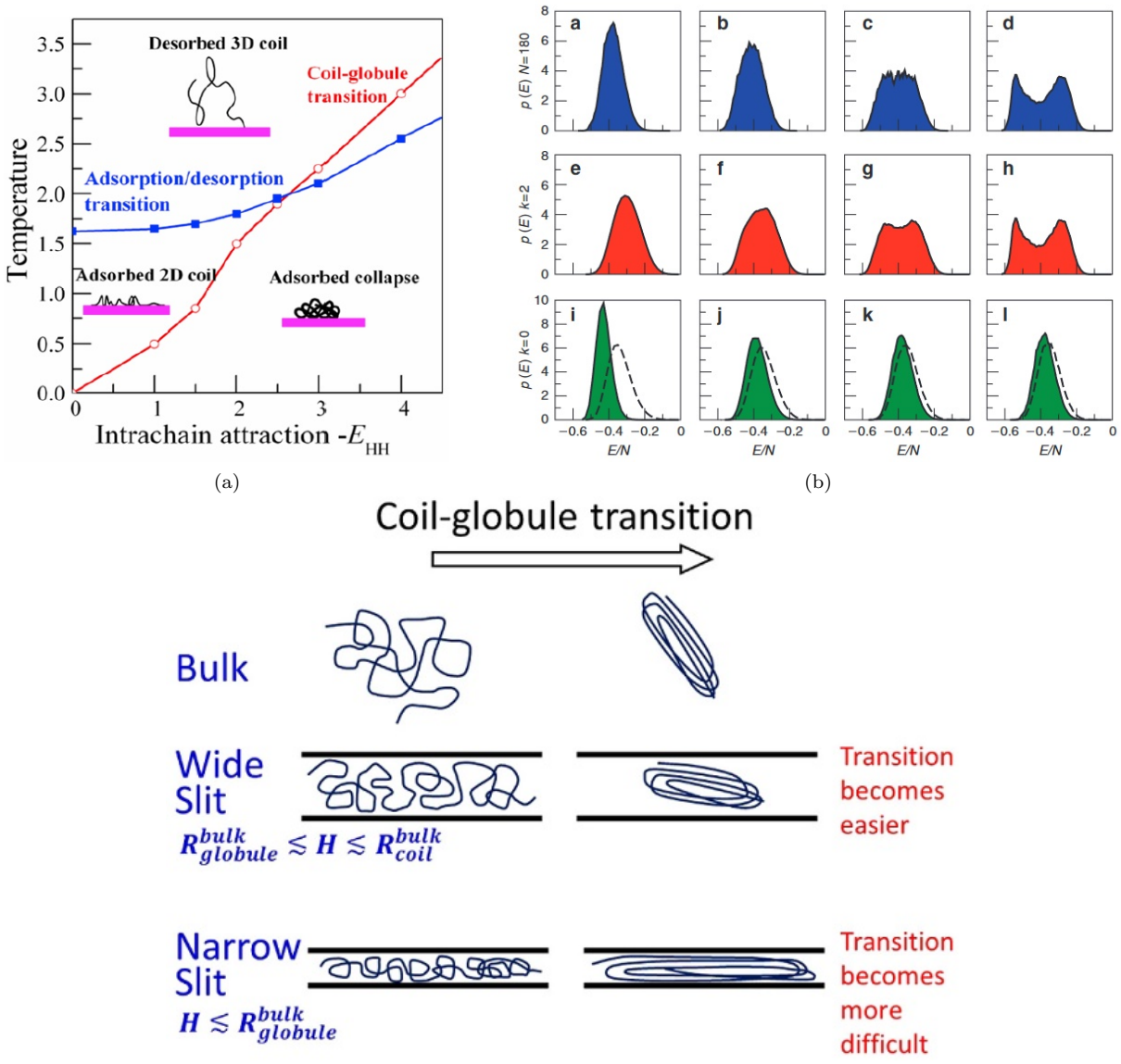
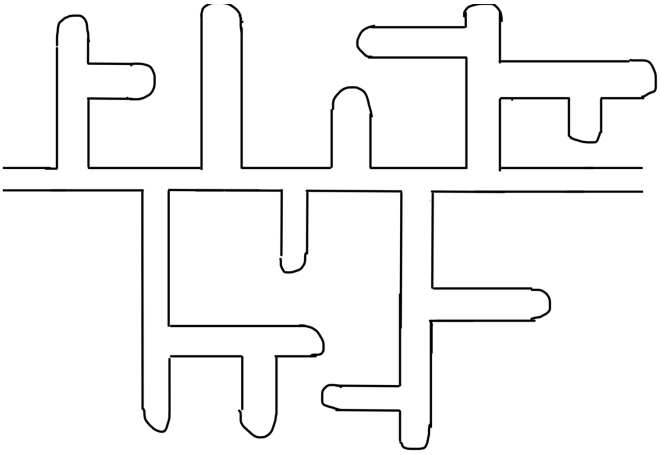


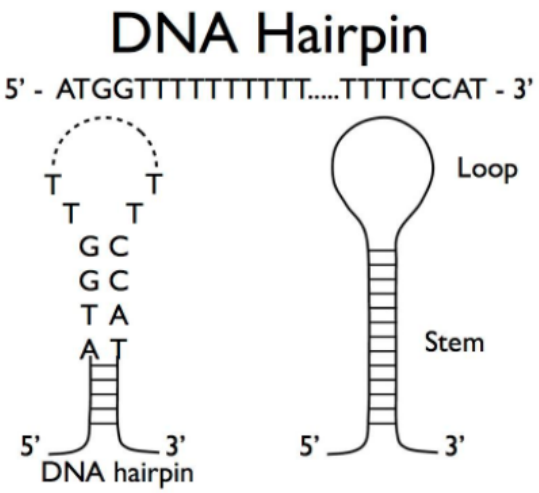
Figure 7: (a) Coil-globule transition graph of temperature vs. Intrachain attraction, including adsorption/desorption transition of 2D, 3D and collapsed coils (Luo et al., 2014). (b) Evolution of the energy probability distribution of the Vibrating Interacting Self-Avoiding Walk (VISAW). Row 1: evolution of the energy per monomer probability distribution for a fixed polymer length, N = 180, and varying values of k; distribution progressively broadens until two peaks become clearly discernible, signalling phase coexistence. Row 2: evolution of the distribution for a fixed value of k (k = 2); progressively longer polymers eventually characterized by two peaks. VISAW model for k = 0 and increasing lengths (Maffi et al., 2012). (c) Coil-globule transitions in bulk and slits (Dai et al., 2015).

Pt 3. Applications to Biology - Condensation of Long SS DNA Molecule

Here we are presented with the hypothetical of having a single-stranded DNA molecule with the sequence of ATAT...AT—the nucleotides of adenine and thymine only. When we consider the high-temperature denatured phase, it is one where the DNA solution has been heated up to the point where it breaks the hydrogen bonds preset in the initial double-stranded DNA configuration and turns it into that of a single strand. This will not look like a coil-globule transition, as this is expected at low values of temperature, where attraction is dominant. At very low temperatures, the expected result is one single long hairpin (with the pairings of adenine and thymine), but this is not a possibility in our experiments due to water freezing. Below is a sketch of a typical configuration that can be expected considering these elements. In this rough sketch,



(a)



(b)

Figure 8: (a) Sketch of a typical configuration of the Single Strand DNA molecule as it condenses from a high temperature denatured phase. For simplicity, the turns are drawn at 90 degree turns, but this does not have to be the case. The straight segments of the molecule have adenine and thymine paired together, with the turns or loops at the ends of the individual hairpins having unpaired nucleotides. (b) Simple picture of a DNA hairpin for reference to use in the sketch made. Credit goes to Taekjip Ha at the University of Illinois at Urbana-Champaign.

the hairpins can be easily shown, along with the hairpin branches that come from them. This configuration produces a much higher entropy compared to that of the one long hairpin at higher temperatures and is thus

the favored configuration. The branches of hairpins are products of the randomness that the configuration experiences. The aim is to minimize the free energy (such as nature does!), given as

$$F = E - T * S$$

With this model, the entropy is maximized with having more configurations while also expending a lower amount of energy.

4. 2D HP Model

Pt 1.

The HP lattice model is a simple yet accurate model that captures much of the physics behind protein folding. Proteins live on a lattice, normally taken to be square in two dimensions. Polymers can't intersect themselves but interact by nearest neighbor interactions. Instead of 20 kinds of amino acids, this is simplified down to two: hydrophobic, "H", and polar, "P". An picture of the ground state of such a model is shown in figure 8.

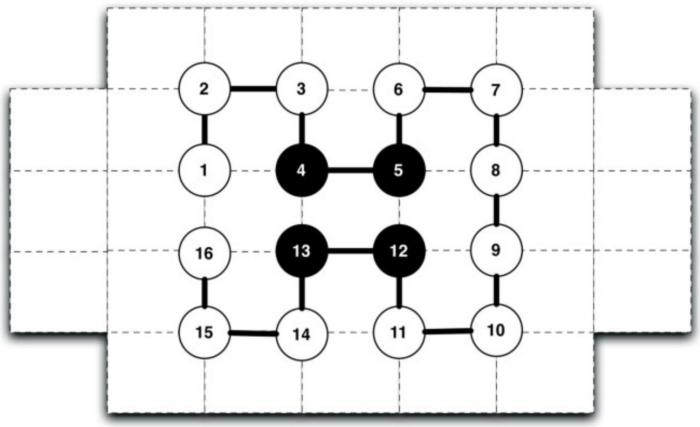


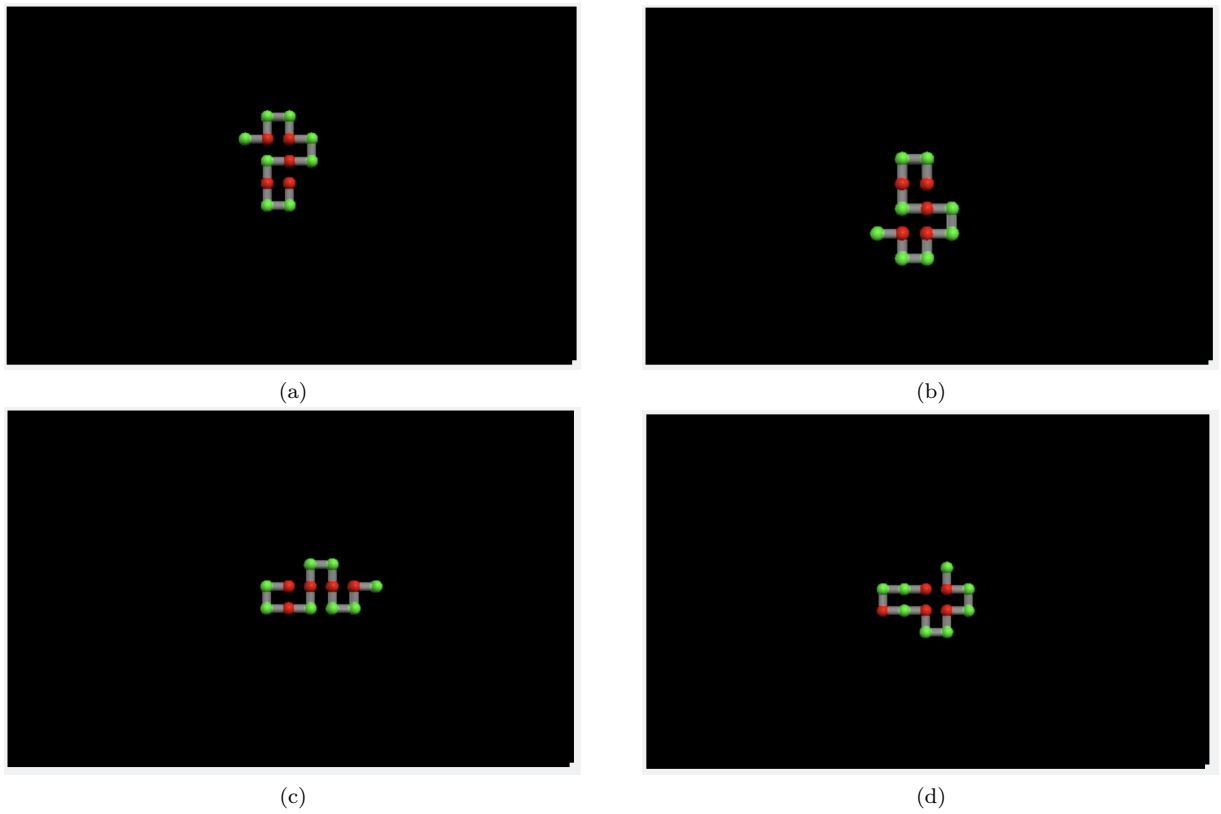
Figure 9: A ground state conformation in the 2D HP model. The grid represents the 2D square lattice which this conformation, or amino acids that makeup the protein molecule, are fixed on. The black circles represent the hydrophobic, "H", residues while the unfilled white circles represent the polar, "P", residues. The ground state energy of such a conformation is calculated as the number of H-H contacts between neighbors. We see that the ground state energy of this conformation is -2. With an energy of -1 being contributed from the neighboring residues 4 and 13, and 5 and 12 (Thachuk et al., 2007).

This HP model can be used to understand the different intermediate states of a protein as it folds. We are essentially probing the number of states at specific energies by running this simulation multiple times. One study of this HP model was done by Ken Dill and Hue Chan, in which they consider a 13 length chain and analyze its folding kinetics (Dill Chan, 1997). The ground state energy is -5, which is unique for this specific sequence. In addition, we can study how unique the higher energy states are for this system.
The code in lat_prot_py_2d.py is a simulation of this 2D HP model and lets us visualize how it folds as the protein cools down. The program is set to stop if the minimum energy state is -4 or below. The ground state energy is -5. We can also change the minimum energy value to say -3 such that the simulation will stop if the minimum energy is -3 or below. Probing these minimum energy states can allow us to see the corresponding protein configuration, and compare multiple simulations to determine if the configuration is unique. We define a conformation to have energy $E(c_i)$, where C_s is the set of all valid self-avoiding walks

on some lattice, L, for sequences s:

$$E(c_i) = \sum_{j=1}^{n-1} \sum_{k=j+1}^{n} N_{jk} \qquad \text{with} \qquad N_{jk} = \begin{cases} -1 & \text{j and k are both H residues and topological neighbors} \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

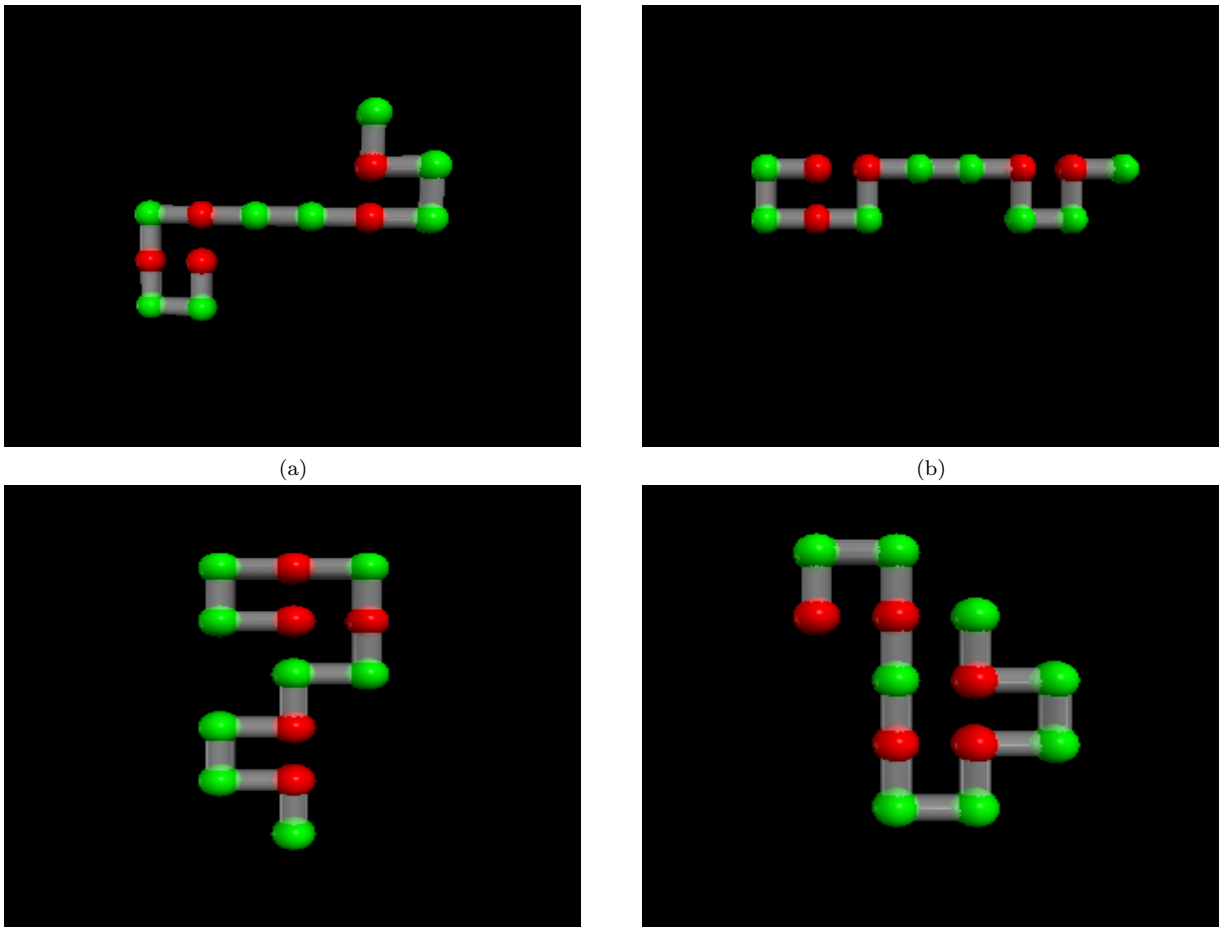
A conformation C^* that minimizes $E(c_i)$ is considered a solution and is also called a ground state conformation of the given protein sequence.
We run the model simulation multiple times and below show screenshots of the simulated protein configurations at certain energy states.



(a) (b) (c) (d)

Figure 10: Resulting protein configurations after running the 2D HP model simulation. Shown are 4 images of of the higher energy state, $E = -4$, with $E_{min} = -4$. Notice that each of these configurations is different although they all have 4 connections of neighboring H amino acids.

We see from the above figure that the configurations we find at different energies higher than the ground energy are, in addition to rotation and reflection differences, *not unique*. This suggests that there are a number of intermediate states one has when a protein folds. Many instances of the HP protein folding problem exhibit solution degeneracy. This means there is more than one minimum-energy conformation. Although the ground state conformation of our simulation is unique, the intermediate states when a protein folds are not unique as there are many different configurations the protein can end up in for the same intermediate energy state. This density of states is important in determining the thermodynamic stability of proteins (specifically larger ones).



(a) (b) (c) (d)

Figure 11: Resulting protein configurations after running the 2D HP model simulation. Shown are 4 images of of the higher energy state with $E = -3$. Notice that each of these configurations is different (in addition to rotation and reflection differences) although they all have 3 connections of neighboring H amino acids.

Pt 2. Interactions which should be Included in an Accurate Model

To understand whether a sequence will fold into a unique and stable structure, a demonstration is necessary showing that the sequence prefers the folded state to both the unfolded state, folding intermediates, and alternative folded states. Sets of representative conformations must be used to approximate folding intermediates and possibly other stable or meta-stable states. Without explicitly considering decoy structures, the inter-residue contacts from different conformations can be randomly sampled. Furthermore, a scoring function is needed to quantify differences between the possible states, where strategies include those derived statistically from existing structures and those from physical first principles. Both types of scoring function assume that the native sequence lies close to a thermodynamic optimum, as well as the presence of a gap in energy between the native state and the closest non-native state (Grahnen et al., 2011). Protein biological function depends on proper folding as well as on the ability to bind the target ligands, hence binding must be evaluated and from a physical perspective the only states to consider are the bound and the unbound states. It has been considered that selective pressures on proteins to avoid non-specific binding are also an important aspect of protein fitness. The binding decoy characteristics also affect the level of selective constraint on the binding interface (Grahnen et al., 2011).

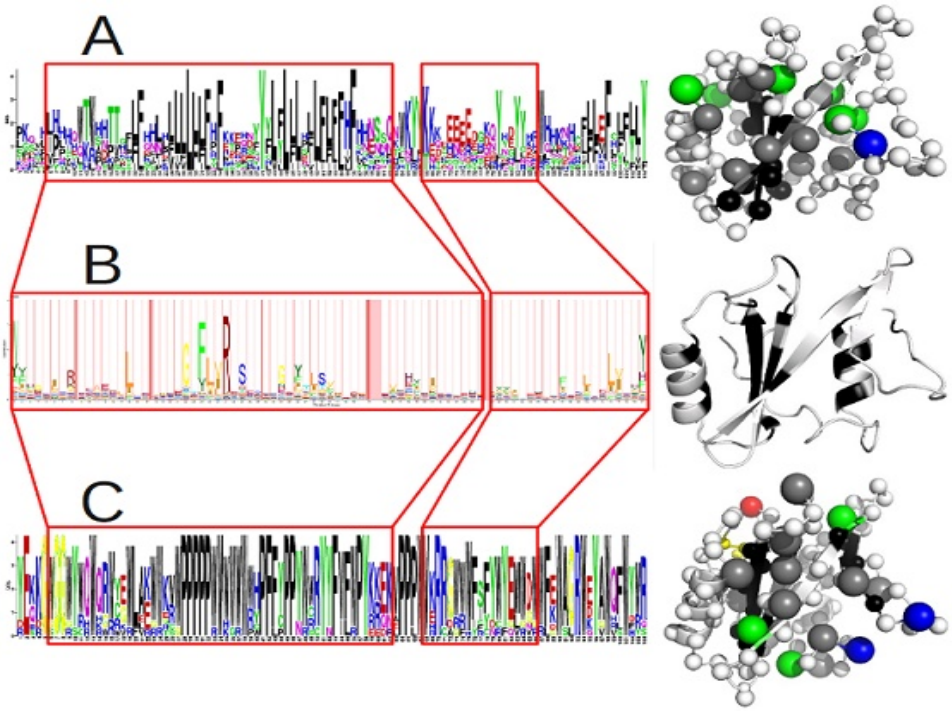


Figure 12: A comparison of simulated sequences and structures with those of known SH2 proteins. Here A) and C) depict sequences which were simulated under the informational and physics-based models. B) depicts the SH2 family from the Pfam database. Residue frequency distributions are shown on the left as sequence logos, A) and C), and as HMM emission probabilities for B). Red boxes indicate matching positions between the sequences. (Grahnen et al., 2011).

Grahnen et al. (2011) state that site and function independent models do not capture critical elements of protein evolution. For structure-based models, a good scoring function must produce sequences with similar properties to real proteins. This includes a hydrophobic core that evolves slower than the hydrophilic surface. Furthermore, while the gamma function for rate heterogeneity was adopted for model fit rather than mechanistic purposes, it is one of the most common parameters in standard evolutionary models and accounts for heterogeneity in the substitution rate driven by structural signals as well as other signals in evolutionary sequence data (Grahnen et al., 2011). In addition to rate heterogeneity, sequences must also have an energy gap between the native and alternative conformations to ensure rapid and stable folding.

The model should place the native sequence near an optimum to not provide a signal of directional selection when function is not changing, and therefore most mutations should be deleterious or nearly neutral rather than adaptive. Lastly, proteins must retain their binding function (Grahnen et al., 2011).
Population size will dictate what fitness changes are neutral as well as what mutations become substitutions. To enable use in forward and backward studies of evolution, the model should be coarse grained at a level that permits sufficient computational speed. Protein folding and function cause interdependence between sites in the protein sequence. For instance, a mutation that removes a cysteine involved in a disulfide bridge is likely to be deleterious, whereas mutation of other cysteines which are not involved in disulfide bridge formation may be more neutral. There are two ways of calculating such scores and are statistically motivated informational methods and the first principles physics-based methods. Informational models score the likelihood of observing specific types of contacts in the folded protein based on those seen in previously known structures, whereas the Physics-based approaches evaluate structures by measuring the fit of residues to geometrical properties such as backbone torsion and residue-residue distances. In either cases the fit to the native and the many possible non-native conformations must be measured to ensure specificity (Grahnen et al., 2011).

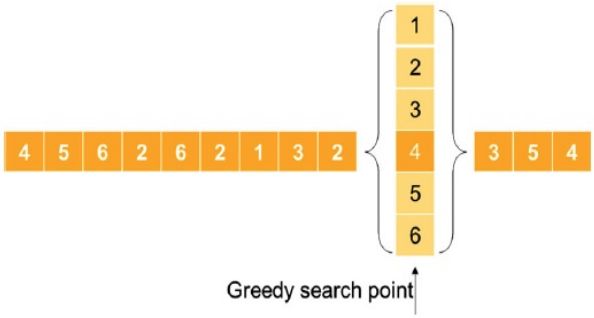


Figure 13: IMO and local search by a greedy algorithm, utilized for randomly selecting the position point, as well as the best fitness of six neighbors (Yang et al., 2018).

Yang et al. (2018) claim that the self-avoiding walk for protein folding is one of the harder NP problems. This difficulty has led to several heuristic and meta-heuristic algorithms proposed to find best protein structure predictions, including genetic algorithm, branch and bound, replica exchange Monte Carlo, Evolutionary Monte Carlo, and greedy-like-algorithm. The authors use an ions motion optimization (IMO) algorithm as a heuristic algorithm newly combined with a greedy algorithm for local search within the 2D-triangle-lattice-model. This should optimize protein folding predictions of high stability, high efficiency, and outstanding performance. The greedy algorithm makes a current local optimal decision at each stage for global optimization, is easy to implement and works efficiently depending on the problems, hence plays a useful role as a optimization method according to its characteristics. Greedy algorithms are used in bioinformatics tools as in DNA sequence alignment, co-phylogeny reconstruction problems, detection of transient calcium signaling, and resolving the structure and dynamics of biological networks. Yang et al. (2018) experimental results show that the proposed IMOG method can reliably find the best solution for short sequences, but also obtain satisfying results with longer sequences. The authors convey that the hybrid algorithm, combining the IMO algorithm with a greedy algorithm provides a novice and useful tool for protein folding predictions (Yang et al., 2018).
Any additional interactions will definitely affect intermediate states. Depending on the interactions itself and how many of these can occur in the folding process, it will give us a diversity of intermediate states with different energy levels. Although we are generally limited 20 amino acids, and a predictable range of interactions are still predictable with a mathematical model, there will still be a plethora of conformations.

1 Sources and References

Axelrod, D., Koppel, D. E., Schlessinger, J., Elson, E., and Webb, W. W. (1976). Mobility measurement by analysis of fluorescence photobleaching recovery kinetics. Biophysical journal, 16(9), 1055–1069. https://doi.org/10.1016/S0006-3495(76)85755-4

Dai, L., Renner, C., Yan, J. et al. Coil-globule transition of a single semiflexible chain in slitlike confinement. Sci Rep 5, 18438 (2016). https://doi.org/10.1038/srep18438

Dill, A. Ken., Chan S. Hue. (1997). From Levinthal to Pathways to Funnels. Nature Structural Biology, vol4, 1.
Grahnen, Johan A., et al. "Biophysical and structural considerations for protein sequence evolution." BMC evolutionary biology 11.1 (2011): 1-18.

HP gound state pic:
https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-342

Luo, Meng-Bo et al. "Interplay of Coil-Globule Transition and Surface Adsorption of a Lattice HP Protein Model." The journal of physical chemistry. B vol. 118,51 (2014): 14913-21. doi:10.1021/jp506126d

Maffi, C., Baiesi, M., Casetti, L. et al. First-order coil-globule transition driven by vibrational entropy. Nat Commun 3, 1065 (2012). https://doi.org/10.1038/ncomms2055

Yang, Cheng-Hong, et al. "Protein folding prediction in the HP model using ions motion optimization with a greedy algorithm." BioData mining 11.1 (2018): 1-14.

Zhu, Panpan, Yuanhua Ding, and Rong Guo. "Coil-globule structure transition and binding characteristics of DNA molecules induced by isoquinoline-based photoactive ionic liquid surfactant." Colloids and Surfaces A: Physicochemical and Engineering Aspects 531 (2017): 150-163.