Protein Folding

TFY4235 - Assignment 3

Sami Laubo

April 8, 2024

Abstract*—*Simulating protein folding with two- and three dimensional polymers gave insight into how to most effectively navigate the energy landscape of polymer configurations. High temperatures allows the polymer to quickly surpass local minimas, but have higher oscillations than for lower temperatures. Too low temperatures gave configurations that get stuck in local minimas and never get close to the optimal energy configuration. Simulated annealing gave the most effective search for minimal energy configurations, where it can quickly surpass metastable configurations and find equilibriums in lower energy configurations.

1. Introduction

Understanding the folding dynamics of polymers is crucial in many scientific and industrial fields, ranging from drug design to material science. Polymers are large molecules composed of amino acids, and their folding behavior governs their mechanical properties, biological functions, and interactions with their surroundings. Despite the significance of polymer folding, the prediction of its intricate folding pathways remains a challenging task due to the complexity of molecular interactions.

Polymer folding describes the process in which a primary structure of a polymer (often stretched out) folds onto itself into its tertiary structure, which is the structure it has when it is biologically active. The folding and final structures depend on the interactions between the amino acids, temperature, pressure and chemical composition of the environment. The aim of this paper is to model polymer folding with Monte Carlo simulations and to explore the energy landscapes, phase transitions, and other interesting properties that occur. A simplified model is used, but it still gives insight into the fascinating properties of polymer folding which can be transferred to more complex simulations. [2]

2. Methodology

In this paper, we model a polymer as a chain of amino acids, also called monomers, where two consecutive amino acids have rigid covalent bonds of fixed length. The polymer is modeled on a grid in two- and three dimensions, which only allows the angles between the covalent bonds to be 0 or 90 degrees and a fixed bond length 1.

2.1. Initiating polymers

The initial structure of a polymer is created monomer by monomer. A direction $\vec{r}$ has the direction to the next monomer, and this direction is changed randomly to the other possible directions with a probability p. When $p = 0$, the initial polymer will be in a straight line, while a higher p gives more flexibility to the polymer. The direction of the polymer can change to two possible directions, 90 degrees clockwise or counter clockwise in 2D, as long as the positions are not occupied by another monomer. In 3D, there are four possible directions. Thus, taking N steps with this method creates a N monomer long polymer, where the "straightness" is controlled by a probability parameter.

2.2. Energy

Each amino acid has an interaction energy with the other amino acids in the order of k_b. The interactions are between monomers that do not have covalent bonds (previous or next in the chain), but still are nearest neighbours. This means that the L_1-distance (Manhattan distance) between the two monomers is 1. This energy is denoted by $J_{[A(i),A(j)]}$, where $A(i)$ is the amino acid type of monomer i, out of 20 possible amino acid types. Each amino acid type thus has a specific interaction energy wwith each other amino acid type. These interaction-energies are picked randomly (but remain fixed for a simulation) on the interval $-4k_b$ to $-2k_b$ and can be represented in a symmetric matrix, where one instance is presented in Figure 1.

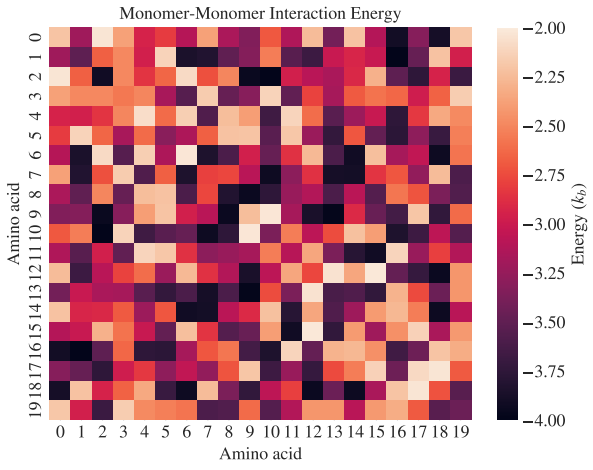

Figure 1. *One instance of random interaction energies between amino acid types.*

The total energy of the system, excluding covalent bonds, is

$$E = \sum_{(i,j)} \delta_{(i,j)} J_{[A(i),A(j)]}, \tag{1}$$

where $\delta_{(i,j)}$ indicates that monomer i and j are nearest neighbours. When proteins fold they will minimize their energy and thus finding the global minima of Eq. (1) over possible tertiary structures gives the final structure of the protein.

2.3. Radius of gyration

The radius of gyration (RoG) of a protein is defined as the radius at which the moment of inertia equals that of the protein, if all its mass were concentrated at this radius [2]. For a discrete polymer where all the monomers have the same mass,

the equation reads

$$RoG = \sqrt{\frac{\sum_{i=1}^{N}(\mathbf{x}_i - \overline{\mathbf{x}})}{N}}, \tag{2}$$

where $\overline{\mathbf{x}}$ is the average position of the monomers and N is the number of monomers. This gives a measure of how compact the polymer is packed.

2.4. Metropolis Monte Carlo

When a protein folds it is preferred to minimize the energy of the system, in our case Eq. (1), and the tertiary configuration of the protein will have the minimal energy out of all the possible configurations. The amount of configurations can quickly grow very large (detailed more in Appendix A) and an exhaustive search through all the configuration is computationally infeasible. The Metropolis Monte Carlo (MMC) algorithm is one way to search through the configuration space more effectively.

The search is performed by starting with an initial configuration $\mathbf{x}^{(i)}$ and finding the possible transitions for a random monomer in the polymer. In the two-dimensional case, the end monomers can have a maximum of two possible transitions and an inner monomer can have one maximum possible transition. An inner monomer only has possible transition when it together with the neighboring monomers create a 90 degree angle, thus it can "flip" to the other side and maintain a bond length of 1. In the three-dimensional case, the end monomers have a maximum of four possible transitions and an inner monomer still has a maximum of one possible transition. A transition is defined as a jump to a lattice site which keeps the covalent bonds to the length of 1. Numerically, this is done by checking if the surrounding lattice sites of the monomer is empty and if the bond length to both the previous and next monomer would remain one if the jump is performed.

After the possible transitions for one monomer is found, a random transition is chosen and the energy of the new configuration is calculated. The transition is accepted based on a probability using the MMC rule. If the energy of the new configuration E_j is lower than the energy of the current configuration E_i, the transition is accepted. However, if $E_j > E_i$, the new configuration is accepted with probability $\exp(-\beta(E_j - E_i))$. This is obtained from the Boltzmann distribution, where $\beta = 1/k_bT$ [3]. This is called a Monte Carlo draw, whereas N such draws accounts to one Monte Carlo step.

This model has a finite probability to reach all possible states in the system, but will quickly go in the direction of lower energies.

2.5. Simulated annealing

Using the Boltzmann measure to calculate the probability of accepting a transition to a new configuration has the characteristic that the temperature greatly affects the probability of accepting a jump to a higher energy configuration. This means that a higher temperature will give the system more randomness and thus less probability to get stuck in local minimas. Starting with a high temperature makes the system explore a larger part of the configuration space and then slowly lowering the temperature to more precisely reach a global minimum. This concept is called simulated annealing.

3. Results

3.1. Initial tertiary structures

An example of a polymer consisting of 15 monomers is presented in Figure 2. This polymer was initiated with the maximum probability of changing direction as the polymer grows. The total internal energy for 10000 such polymers is plotted in a histogram in Figure 3.

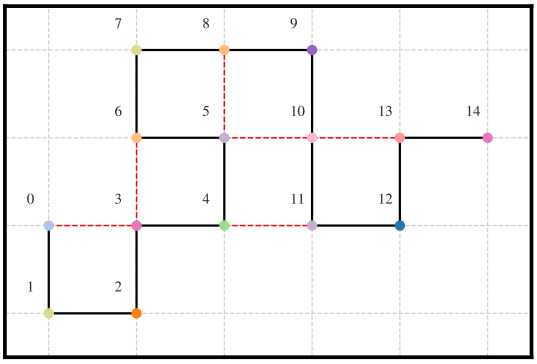

Figure 2. *Example initial polymer with maximal probability of changing direction as it grows. $N = 15$ monomers indexed in ascending order, where the colors indicate its amino acid type. Black lines are covalent bonds and red dashed lines are nearest neighbours.*

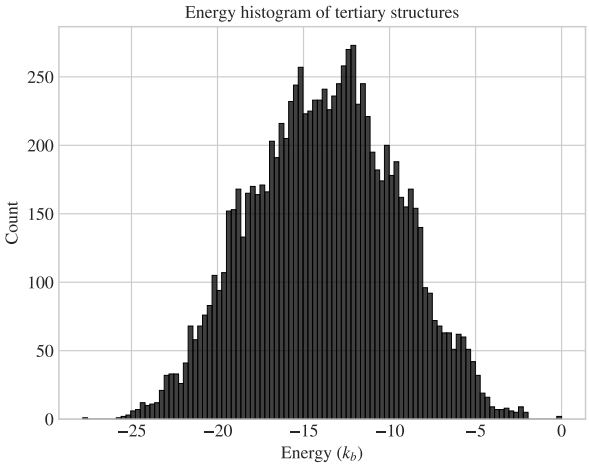

Figure 3. *Energy histogram of total internal energy in* 10000 *polymers made up of* 15 *monomers, including only nearest neighbour interactions.*

3.2. Monte Carlo simulations

At temperature $T = 10$ and $N = 15$ monomers, the system after 1, 10 and 100 Monte Carlo steps is shown in Figure 4. Additionally, the observables energy, radius of gyration and end-to-end distance is plotted in Figure 5 for the Monte Carlo steps. The same experiment is carried out at temperature $T = 1$, and the obsevables are plotted in Figure 6. Furthermore, for $T = 1$ and a longer chain of 50 monomers, the observables are plotted in Figure 7. The energy, end-to-end distance and radius of gyration is averaged over total steps/20 to smooth the

graphs. This means that the first and last total steps/40 steps are excluded from the plot. All the following presentations of the observables are presented in this way.

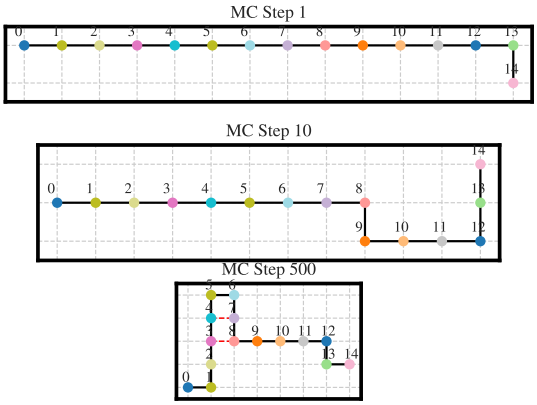

Figure 4. *A polymer of* 15 *monomers plotted after Monte Carlo (MC) sweep* 1, 10 *and* 100, *at* $T = 10$.

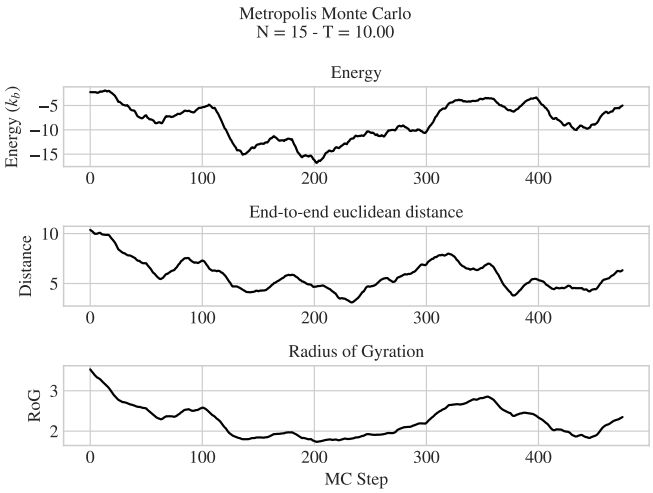

Figure 5. *The energy, radius of gyration and end-to-end distance (E, RoG and Distance respectively) of the polymer of* 15 *monomers plotted over Monte Carlo (MC) sweeps with an averaging of* 10 *sweeps, at* $T = 10$.

3.3. Phase diagrams

To find the phase diagrams $E(T)$ and $RoG(T)$, we first start with finding the equilibrium at different temperatures. To find the equilibrium the average over every 100 energy datapoints is continously taken. When the total sum of the R^2 distance of the last 3 such points away from the average of the same 3 datapoints are less than the threshold 0.4, the simulation is stopped and the average of the last 100 points are used as the equilibrium temperature and radius of gyration. The Monte Carlo simulation of a monomer of length 15 is presented in Figure 8 for multiple temperatures. A supplementation of monomers of lengths 25, 35, 45 and 55 is given in Appendix B.

The phase diagram uses the average of the last 100 datapoints of the Monte Carlo simulation as the equilibrium energy and radius of gyration. These phase diagrams are presented in

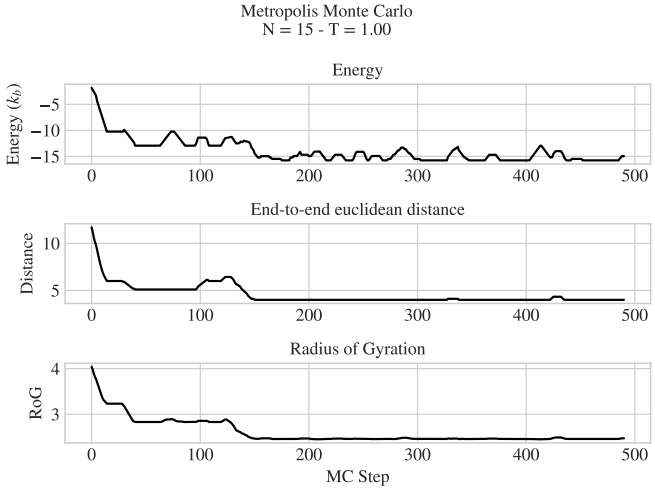

Figure 6. *The energy, radius of gyration and end-to-end distance (E, RoG and Distance respectively) of the polymer of* 15 *monomers plotted over Monte Carlo (MC) sweeps with an averaging of* 10 *sweeps, at* $T = 1$.

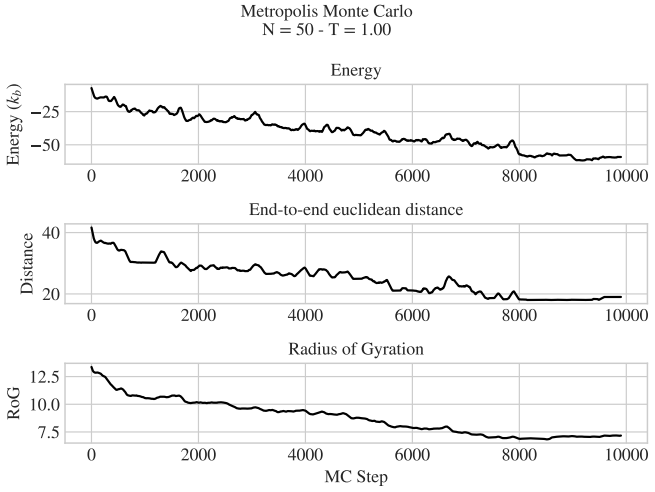

Figure 7. *The energy, radius of gyration and end-to-end distance (E, RoG and Distance respectively) of the polymer of* 50 *monomers plotted over Monte Carlo (MC) sweeps with an averaging of* 10 *sweeps, at* $T = 1$.

Figure 9 and 10 for the energy and radius of gyration respectively.

3.4. Simulated annealing

Starting from the same initial unfolded polymer of length 30, Figure 11 and 12 shows two possible tertiary structures after 10000 MC steps at $T = 1$. For the same inital configuration simulated annealing where applied with temperatures starting from $T = 4$ to $T = 1$ over 2000 MC steps. The result from the simulated annealing experiment is presented in Figure 13 .

Changing the sign of some random monomer-monomer interaction energies results in the matrix in Figure 14 and a simulated annealing Monte Carlo simulation of an unfolded protein is presented in Figure 15.

3.5. Three dimensions

Similar to Figure 4, the Monte Carlo evolution of a 15 monomer chain is plotted after MC steps 1, 10 and 100 at temperature $T = 10$ in Figure 16.

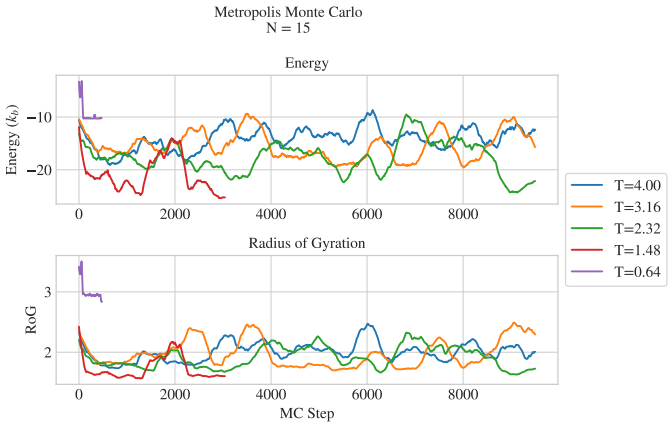

Figure 8. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

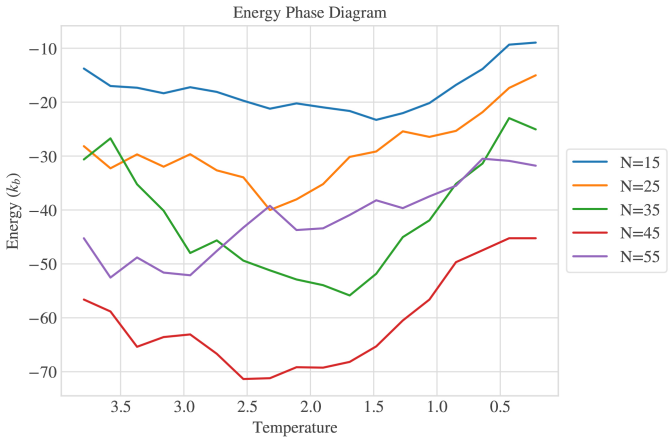

Figure 9. *Phase diagram of the equilibrium energies, $E(T)$, for different polymer lengths N.*

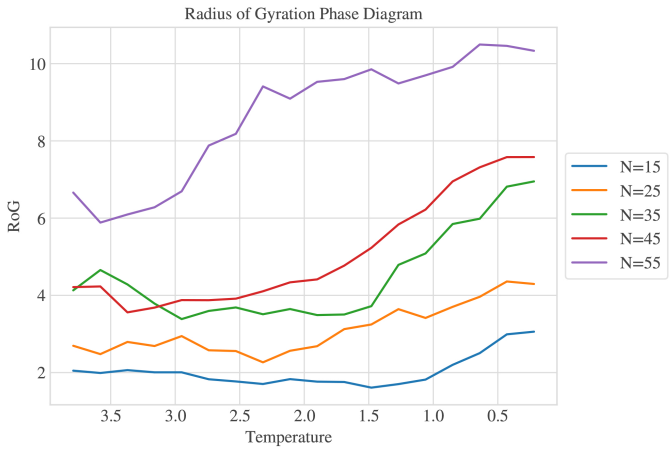

Figure 10. *Phase diagram of the equilibrium radius of gyration, $RoG(T)$, for different polymer lengths N.*

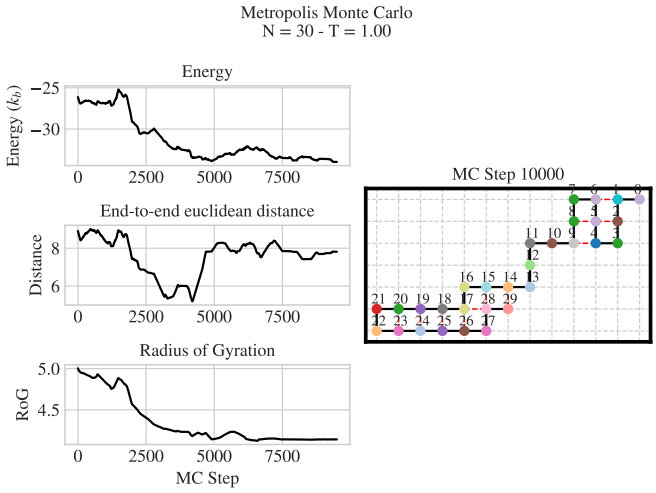

Figure 11. *One possible tertiary structure for a polymer of* 30 *monomers at* $T = 1$.

The phase diagrams for energy and radius of gyration are also presented in the three dimensional case in Figure 17 and 18 respectively.

4. Discussion

Initiating 10000 folded tertiary structures of 15 monomers such as the example presented in Figure 2 gives the energy distribution in Figure 3. The distribution looks like a normal distribution centered around approximately $-14k_b$. Given that the average monomer-monomer interaction energy is $-3k_b$, this amounts to an average of almost 5 nearest neighbour interactions per polymer. The negative interaction energy imply that the polymer wants to be as folded as possible for as many of these interactions to happen such that the energy in the system becomes as low as possible.

A straight polymer of 15 monomers can only move its end monomers the first Monte Carlo step. This we can see in Figure 4 and as the polymer evolves it becomes more and more folded. We can see that the energy of the system presented in Figure 5 does not go very low, only oscillates about $-10k_b$. This is probably due to the high temperature which makes the probability of a jump to a higher energy state relatively likely compared to a jump to a lower energy state. However, if the temperature of the system is decreased to 1, we can see that the energy in Figure6 stabilises at a lower energy with much less fluctuations. This is likely due to the not having a high enough temperature to jump out of metastable phases and gets stuck in one local minima. For $T = 1$ it takes around 200 iterations to reach a steady state. For a longer chain of 50 monomers the energy decreases more slowly as we see in Figure 7, and it takes almost 9000 steps to reach equilibrium. On average each end monomer should move once per Monte Carlo step, which means longer chains will take longer to curl up than shorter chains. A straight line initiation of the polymer is therefore not preferable.

Figure 8 shows the Monte Carlo evolution of a polymer when simulated in different system temperatures. We can clearly see that the lowest temperature does not decrease much before reaching an equilibrium and thus is stuck in a local minimum. For higher temperatures the polymer reaches lower total energy, but as the energy increases the fluctuations are even larger. The phase diagrams for both energy and radius of gyration, in Figure 9 and 10 respectively, includes more temperature simulations and the graphs more clearly is decreasing before increasing again around $T = 1.5$. For higher temperatures, the temperature is too high for the polymer to stabilise in one low energy configuration, which explains the initial steady decrease in energy. However, at a certain point

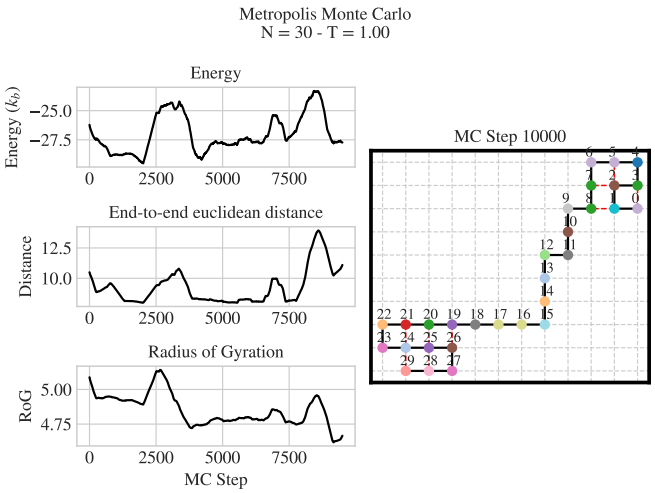

Figure 12. *One possible tertiary structure for a polymer of* 30 *monomers at* $T = 1$.

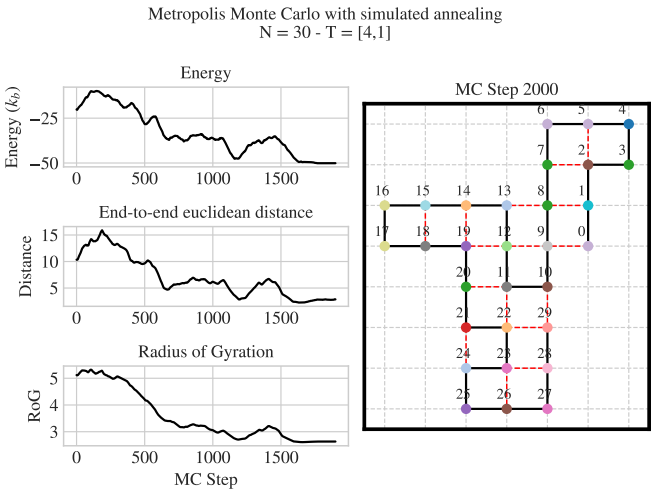

Figure 13. *One possible tertiary structure for a polymer of* 30 *monomers using simulated annealing from* $T = 4$ *to* $T = 1$.

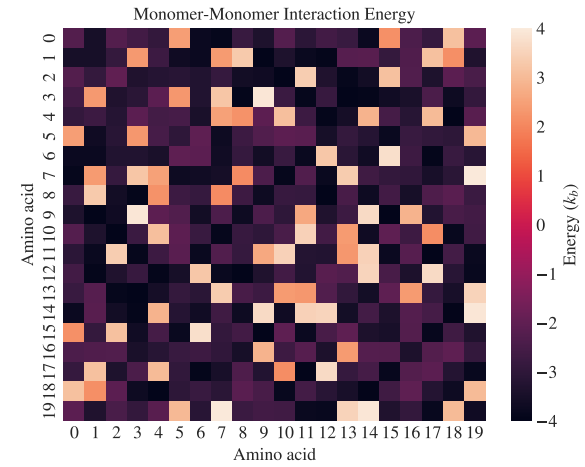

Figure 14. *Instance of random interaction energies between amino acid types.*

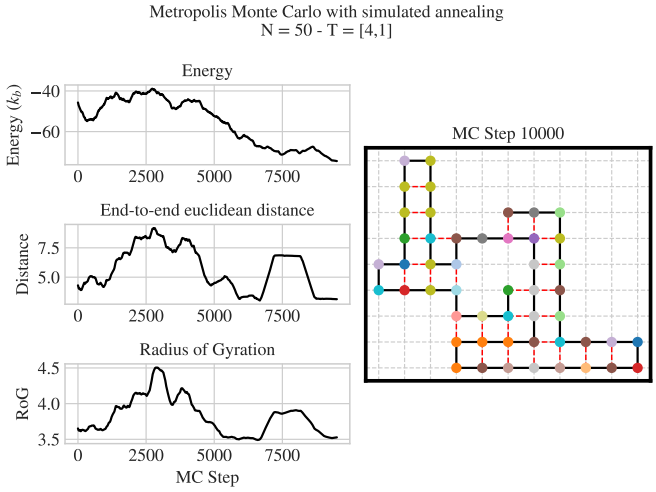

Figure 15. *Simulated annealing of unfolded protein with some positive monomer-monomer interaction energies.*

the temperature becomes too low for the polymer to explore a wide variety of configurations and gets stuck in local minimas before getting near the global mimimum. This effect is more and more prominant for lower temperatures and explains the increase in equilibrium energy and radius of gyration after $T = 1$. The same effects can be seen in the three dimensional case in Figures 17 and 18. The increase in energy for lower temperatures seams to happen earlier for shorter polymers, and thus longer polymers are more effected by lower temperatures.

Two possible tertiary structures for a 30 monomer long polymer is presented in Figure 11 and 12. The two tertiary structures are different and they have different energies. Additionally one can see that the first simulation goes faster down in energy than the second with less fluctuations. The path in the energy landscape that the polymer takes is thus very different from simulation to simulation, and it has a chance of getting stuck in local energy minimas. However, when using simulated annealing on the same initial polymer, the energy goes a lot faster and further down than for the constant temperature case, which we can see in Figure 13. It takes around 1600 Monte Carlo steps with simulated annealing against 10000 steps without, which is a massive increase in efficiency. This happens because the polymer in the start can easily navigate the energy landscape without getting stuck in local minimas because of the high temperatures. After time when the energy lowers, the polymer will find a closer configuration to the global energy minimum than without the simulated annealing. It is thus preferable to use simulated annealing as it is less likely to get stuck in smaller minimas and navigates better and faster towards a global minima. As it is infeasible in most cases to explore all possible configurations of a polymer, simulated annealing gives an effective way to surpass uninteresting configurations and quickly go towards energy minimums.

When changing the sign of some interaction terms, the optimal configuration is not necessarily the configuration with most nearest neighbour anymore. One such instance of monomer-monomer interaction energies is presented in Figure 14 and the following polymer in Figure 15 shows a simulation of a 50 monomer long polymer. We clearly see that there are a hole in the polymer, where the monomers around are indeed in the positive interaction range, and it is thus not energetically favorable for the monomers to have interactions.

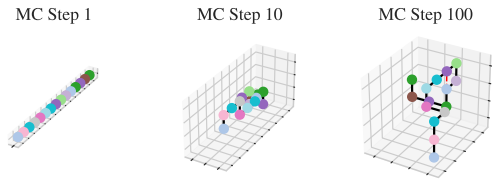

Figure 16. *A polymer of* 15 *monomers plotted after Monte Carlo (MC) sweep* $1, 10$ *and* 100*, at* $T = 10$*.*

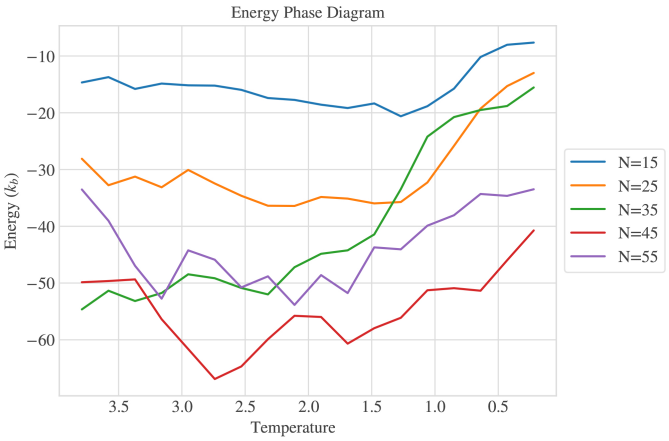

Figure 17. *Phase diagram of the equilibrium energy,* $E(T)$*, for different polymer lengths* N*.*

5. Conclusion

Using Monte Carlo simulations for both two- and three dimensional polymers gave insight into the problem of protein folding. It clearly comes forward that there is difficulties for the polymer to navigate to the lowest energy configuration, and it is highly affected by the temperature in the system. A high temperature allows the polymer to quickly move past local energy minimas and closer to the global minimum, whereas lower temperatures gets stuck in local energy minimas. Simulated annealing starts the simulation at high temperatures such that the system moves fast in the configuration space, and then slowly decreases the temperature to stabilise the system in an equilibrium. Using simulated annealing proved a faster decrease in energy and also lower total energy than for running the simulation in fixed system temperatures. It thus provides a much more efficient search for the optimal folded configuration of the protein.

Systems where some monomer-monomer interaction energies are positive gave rise to different folding configurations. Since some interactions are not energetically advantageous, some polymers have holes inside the folding where monomers avoid each other.

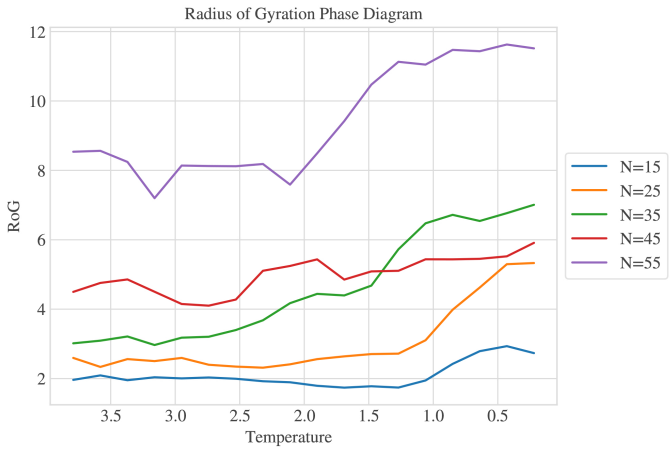

Figure 18. *Phase diagram of the equilibrium radius of gyration,* $RoG(T)$*, for different polymer lengths* N*.*

References

[1] *64-bit computing*, Accessed 15.03.24. [Online]. Available: https://en.wikipedia.org/wiki/64-bit_computing.

[2] J. Frafjord and R. Cabriolu, "Assignment 3: Protein folding", Spring 2024.

[3] M. Hjorth-Jensen, "Computational physics lecture notes fall 2015", August 2015.

[4] I. Simonsen and R. Cabriolu, "Computational physics lectures", Spring 2024.

A. Preliminary questions

This section answers the preliminary questions of the assignment.

If a primary structure has 300 monomers with 299 covalent bonds and we assume that each covalent bond can take 4 directions, there's 4^{299} possible tertiary structures that can form. If a computer then needs 10^{-12} s to check each configure, a sequential simulation would need around 10^{168} s to finish the simulation, which is relatively high. Relatively high compared to everything, ever.

In a typical 64-bit architecture following the IEEE 754 standard, a computer can store unsigned integer numbers up to $2^{64} - 1$[1] and floating point numbers up to $1.8 \cdot 10^{308}$[4]. One also has to be considerate when adding or multiplying single-precision numbers with double-precision numbers. Usually, the single-precision number is converted to double-precision before doing the arithmetic operation, however if the magnitude of the double-precision number is much larger than the single-precision number, the less significant digits will be lost in addition. This is not as likely to occur with multiplication.

The smallest possible number one can add to 1.0 to get a number different than 1.0 is 2^{-23} for single precision and 2^{-52} for double precision floating point numbers.

B. Supplementary data

Figure 19, 20, 21, 22 is a supplement to Figure 8 in the Results section. It presents the averaged energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.

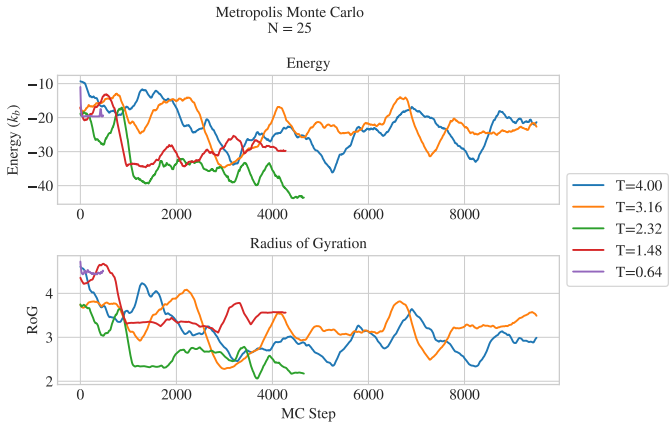

Figure 19. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

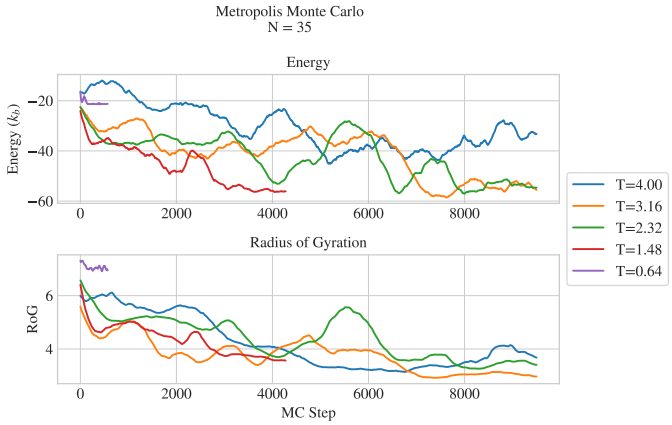

Figure 20. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

The same experiments are presented for the three dimensional case in Figures 23, 24, 25 and 26.

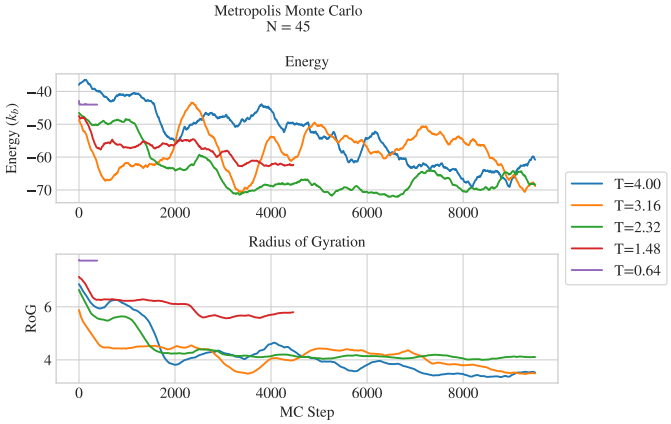

Figure 21. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

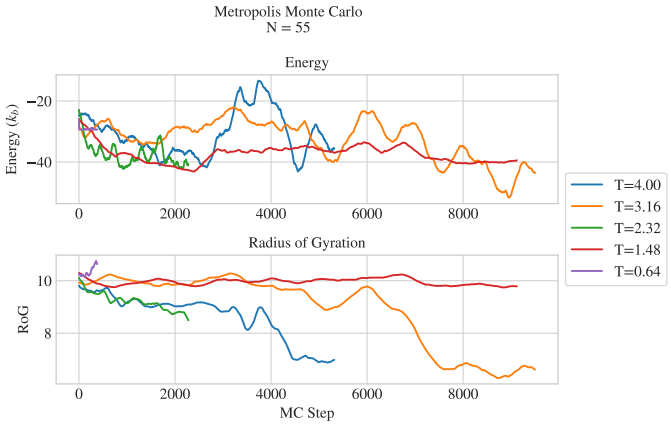

Figure 22. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

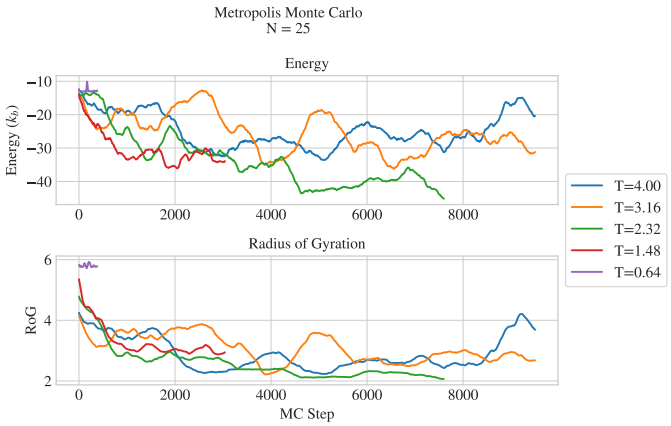

Figure 23. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

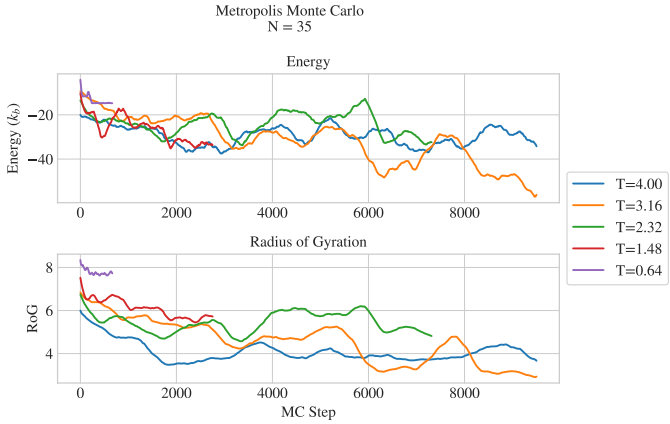

Figure 24. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

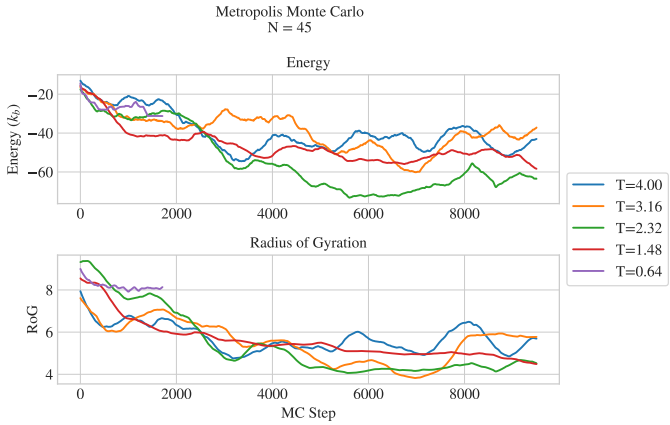

Figure 25. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*

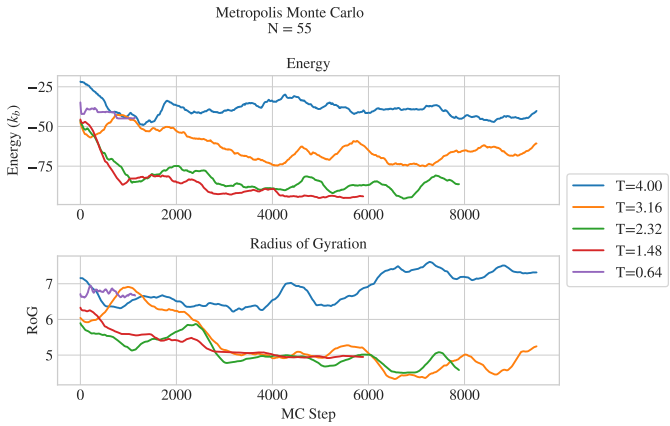

Figure 26. *Energy and radius of gyration over a Monte Carlo simulation for multiple different temperatures.*