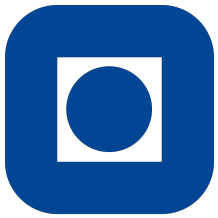
Contact information:
Jonas Frafjord (jonas.frafjord@ntnu.no)
Raffaela Cabriolu (raffaela.cabriolu@ntnu.no)



NTNU
Department of Physics

TFY4235/FY8904: Computational Physics (spring 2024)
Assignment 3: Protein folding
Soft deadline: 14 April 2024

Allowed help: Alternative **A**
This problem set consists of 6 pages.

Introduction
In this assignment, we will study the protein folding problem by running Monte Carlo simulations. The system will be investigated in both 2- and 3-dimensions, and different potentials are discussed. Different aspects of the physics around the equilibration should be carefully explored and it is expected that the simulations contain a sufficient number of sampling points to be representative of the current state of the system. Phase transformation will be investigated and a critical temperature for this transformation will be calculated. The critical temperature is often difficult to calculate accurately due to the unstable nature of the transitional phases. Efficient code will be important to solve this assignment.

Report
The report must be structured as a scientific report. It must contain an abstract, an introduction, a methodology, a result and discussion section, a conclusion and references. The length of the report can not exceed 6 pages, however supplementary figures can be added to an appendix. The supplementary figures are not the main result, but can underline critical points in the discussion. The report should explain what you have been doing, your results, and how you interpret these results. Details should be included to the extent that we as the graders can follow your way of reasoning. Remember that if you have written an original and/or clever code for solving the problem, but are not able to explain it well in the report, it is hard to give you full credit.

Relevant fields: Biophysics, material physics, statistical physics

Mathematical and numerical methods: Monte Carlo simulations, unit evaluation, phase transformations, equilibration, ensembles

1. Introduction

Many problems in nature have an interdisciplinary character. The lines between physics, biology, atmospheric science, chemistry and so on are not clearly defined and solutions to problems can often be relevant across different fields of study. Protein structure is one example which has its strongest roots in biology, but with interesting properties from a physics point of view.

A protein is a type of polymer which is crucial for many biological processes. They are involved in energy conversion and storage, as structural components in bone and tendons or even handle the communication between cells. A polymer is a class of large molecules that are composed of multiple simpler units called monomers that are repeated in the structure. Polymers are known in many different research fields, e.g. in civil engineering where it is used as a structural part in the form of plastic polymers.

Proteins are polymers since they are links of amino acids chained together from start to end. Each link in the chain is one of the 20 unique amino acids that exist in nature, where one amino acid is referred to as a monomer. Different combinations and numbers of these monomers construct all the known proteins in nature, where each monomer is a relatively short molecule containing 10s of atoms. The sequence of monomers is known as the primary structure of the specific protein. This primary structure determines all the properties and functions related to the protein. The shortest proteins are chains of around 50 monomers, while the longer proteins can be several thousand monomers long.

The links between the monomers are flexible, making it possible for the protein to assume different shapes. The proteins in living cells are found in a solution that is mostly water, which facilitates rearrangements of the monomers and changes of the protein shape. Examples of different assumed shapes are illustrated in Figure 1. In one case, the protein is in a folded state, while the other illustrates an unfolded structure. The folded shape can be a sheet in some cases or sphere-like in others. The point is that many different shapes can be formed by using the same chain of monomers. The equilibrium shape depends on various factors such as the temperature, pressure and/or the chemistry of the solution.

When a protein is in its biologically active[1] state it is said to be in its tertiary structure. Understanding the tertiary structure will give insight into how the protein works and what other molecules a protein can bind to, where it fits etc. The tertiary structure is generally some sort of folded state.

The folding of the chains follows a set of rules. The links are not completely flexible, and the bonds are of covalent nature. This means that they share electrons and form electron pairs which are directional and strong. The angles between the monomers are strongly affected by their covalent bonds (see Figure 2). The angle, θ, in the x-y plane and the angle, ϕ, in the x-z plane can in most cases be reversed, yielding 4 combinations of angles of relative orientation that result in the same energy. It is possible to unfold proteins by changing the environment. It has been demonstrated experimentally that when a protein is reintroduced to its biologically natural environment it will reinstate its tertiary structure. The ability to reform into the tertiary structure after an unfold is remarkable when thinking of all the different structures it could assume. We will get a better idea of the vast configurational space in the preliminary questions below.

This seemingly impossible task of proteins to reinstate their tertiary structure is known as Levinthal's paradox. However, the proteins do in fact find back to the tertiary structure without searching through every structure. This feature is not fully understood and the problem is known as "the protein-folding problem".

If the energy of a given primary structure of a protein was solely determined by the covalent bonds, then there would be a multitude of degenerate tertiary structures. However, when the proteins are in a folded state the monomers interact with the other parts of the chain. This interaction is much weaker than the covalent bonds, but they are crucial for differentiating the energy of the structures. One of these interactions is the well-known van-der-Waals interaction. Figure 3 illustrates the interactions and bonds in a folded protein. The simplified illustration, Figure 3a), depicts the flexibility of the folded structure, while the toy model, Figure 3b), illustrates how to easier visualise the weak interactions and the covalent bonds where the circles indicate different amino acids. The toy model limits the angular configurations of the covalent bonds to only allow 0° and 90° orientation relationships.

Another interaction force is the hydrogen bonds. Proteins that are strongly attracted to water molecules can maximise the number of interactions with the solvent by unfolding or positioning the hydrophilic amino acids to the outer surface of the tertiary structure, e.g. the large-blue amino acids in Figure 3. Then there will be a competition between the forces that want to keep the protein folded and the forces that want to unfold the protein. The competition leads to a structure that best minimises the total energy, or there is an arrangement where there are fluctuations between the phases. In this context, a phase refer to the state of being unfolded or folded.

The interaction forces between parts of the monomer chain are on the order of $\mathrm{k_b}T$ per protein, where $\mathrm{k_b}$ is the Boltzman constant and T is around room temperature. This means that there is also a fine balance between the disordering effect of temperature and the interaction forces.

The first task will outline the protein folding problem with a simplified two-dimensional lattice. The primary structure consisting of N amino acids is denoted as,

$$\mathrm{A}(1), \mathrm{A}(2) \ldots \mathrm{A}(N), \tag{1}$$

where $\mathrm{A}(n)$ is a number between 1 and 20 to represent the possible amino acids. Each monomer is connected to its closest neighbours by two covalent bonds (of course, the chain ends have only one connection). The interaction with the neighbour-

[1] Biologically active means that the protein is primed to execute the designed task, e.g. ready to affect a biological process.

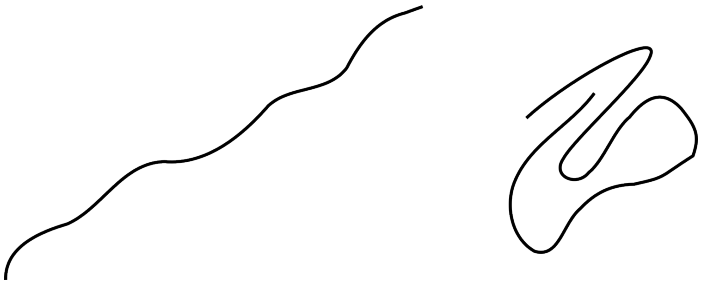


Figure 1: Illustration of an unfolded and a folded protein.

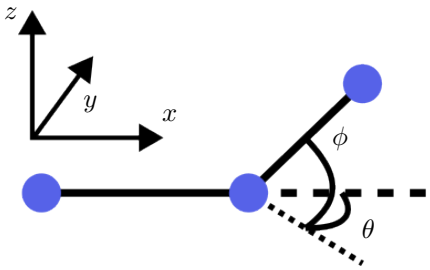


Figure 2: The angles, θ and ϕ, of the covalent bonds, solid lines, can be inverted without energy cost. This yield 4 combinations of degenerate states.

ing monomers that are not covalently bonded is denoted by the energy $J_{[A(i),A(j)]}$. For example, in Figure 4 the interaction $J_{[A(5),A(8)]} = J_{[A(8),A(5)]}$ is the non-zero interaction between amino acids 5 and 8 and its determined by the type of amino acids in question. Note the symmetry for this pair-wise interaction. We assume that only the first (non-covalently bonded) nearest neighbours interact via these J coefficients. The energy of the system, taking only non-covalent interaction into account, is given by,

$$E = \sum \delta_{(i,j)} J_{[A(i),A(j)]}, \tag{2}$$

where the sum is over all pairs of monomers. The variable $\delta_{(i,j)} = 1$ if $A(i)$ and $A(j)$ are nearest-neighbors and not covalently bonded. The interaction energies $J_{[A(i),A(j)]}$ will be given random numbers and they will not change during the simulations for each monomer pair interaction. We will specify the corresponding interaction energy in the task.

Figure 4 illustrates a valid jump. The covalent bonds are for simplicity modelled as rigid. Thus, amino acid 5 can only move diagonally as shown in the figure to keep the length of the covalent bonds equal. A transition is defined as the movement of one amino acid from one position to another.

1.1. Useful terminology

- Amino acid/monomer: one link in the chain that makes the polymer/protein
- Primary structure: The order of N specific types of amino acids in a protein.
- Tertiary structure: The distinct spatial arrangement of the protein. Typically the shape of its folded state.
- End-to-end distance (e2e): vector-length between the first and last monomer.
- Radius of gyration (RoG): This is often used in mechanical physics. It is the radius which yields the same moment of inertia as the structure as if the mass was concentrated at this radius.
- Monte Carlo (MC) draw/trial: One random selection of a monomer in a chain to be evaluated.
- MC step/sweep: N MC draws account to *one* MC step/sweep. On average, all monomers should be drawn within one sweep.

1.2. Preliminary task

The following preliminary tasks will demonstrate how quickly the number of perturbations in this type of system explodes. It should also make you aware of how numbers are represented in computers and the errors in numerical precision. The preliminary task should be answered in an appendix of the report.

- How many tertiary structures can arise from a given primary structure of 300 monomers if we assume that each covalent bond can take 4 directions? See Figure 2.
- How much time would it take in practice to try all the combinations if one spends 1E-12 s for each structure?
- What are the highest integer and floating point numbers you can store in a computer? Specify any assumptions you use.
- Do you lose precision when you a) add or b) multiply a single precision float to a double precision float?
- What is the smallest number you can add to 1.0, for single- and double-precision floating point numbers, to achieve a different number on a computer?

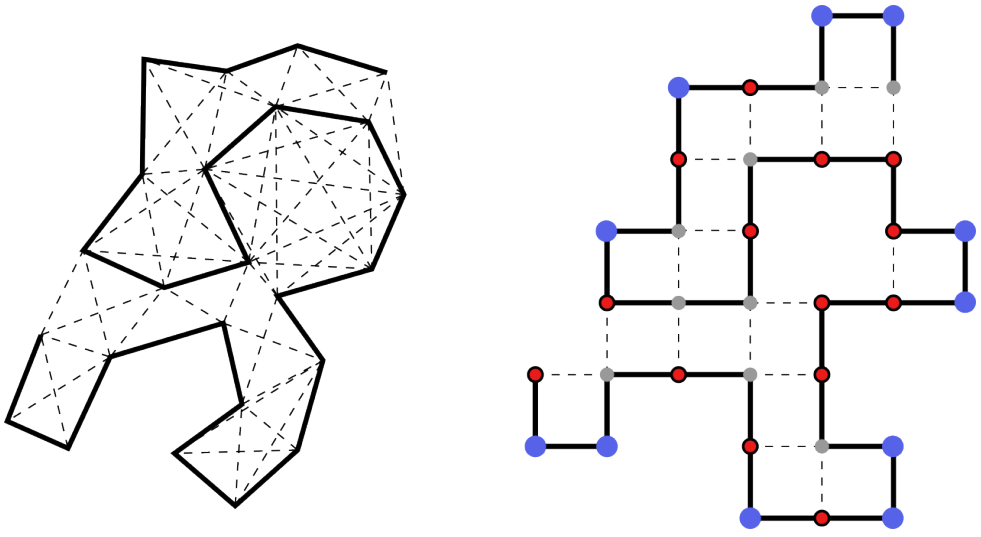


Figure 3: Figure a) illustrates the complexity of a folded protein, where the dotted lines represent the interaction forces between amino acids that are not covalently bonded. The simplified illustration, b), illustrates only nearest-neighbour interactions with stippled lines. The angles of all covalent bonds in b) are 0 or 90°. The large-blue amino acids have no nearest neighbour interactions other than the covalent bonds, the framed-red ones have 1 amino acid within NN interaction, while the small-grey have 2 amino acids with interaction.

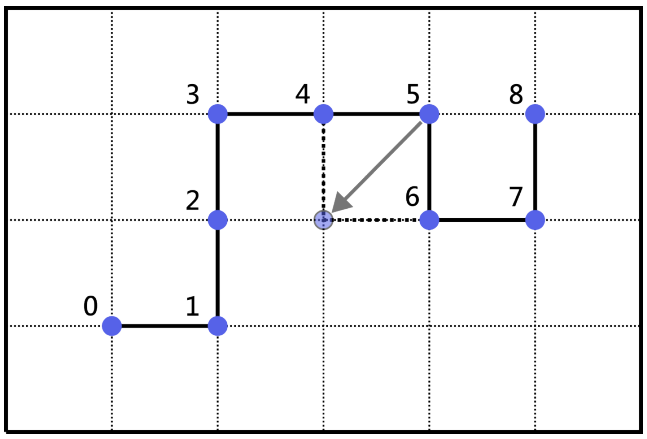


Figure 4: Illustration of a chain of amino acids, blue circles, where the number represent the index in the chain. The solid lines represent the covalent bonds between the amino acids. The arrow indicates a possible transition in the system, where amino acid 5 changes position ('corner-switch'). Note that the chain ends can make movements as well.

2. Tasks

The tasks are written in a specific order to guide you along the problem. Find a way to present your data in a way that illustrates the physical behaviour of the system and shows your understanding of it. You may also use additional simulations with other parameters to support your discussion. Discuss your choices and assumptions. Keep in mind that your code should be well-structured and clearly commented on.

The primary tasks are divided into three parts. The first part is a 2D investigation of protein folding and unfolding. The second part is a 3D extension of the model with relevant discussion. The third part is a bonus task which employs an adaptive model for the monomer-monomer interaction and investigates periodic boundary conditions.

The bonus part is written as an open-ended task of this assignment, and you should prioritise the first two parts and the report before starting the bonus task.

2.1. Task 1

1. Create a primary structure of $N = 15$ monomers in a chain on a 2D grid as in Figure 4. For simplicity, we recommend the vector between the monomers to be a unit vector. Give each amino acid a type (ranging between 1-20).
2. Create a nearest-neighbor list for each monomer, where only the first (non-covalent) nearest neighbour interaction is included, see Figure 3b). Build a general 20×20 matrix of the monomer-monomer interaction energies, where the values in the matrix should be uniformly distributed between -2 $\mathrm{k_b}$ and -4 $\mathrm{k_b}$. Notice that the unit for the energy is $\mathrm{k_b}$, which effectively makes the temperature and $J_{(i,j)}$ unitless. Document one example of the interaction matrix you have used (by writing it out or plotting it) in your report. Think about which properties the matrix should have. The interaction energy matrix should be constant during simulations. It defines the physical interaction between amino acids and should not be changed during simulations once it has been determined.
3. Investigate different tertiary structures by initiating folded structures and calculating their energies. What does the negative value of $J_{(i,j)}$ imply physically? We will later see how some positive numbers can affect the system.

4. Implement a classical Metropolis Monte Carlo method to search through possible tertiary structures. A classical MC method is detailed in the lecture notes, see Chapter 12 therein. In order to set up the general MC procedure, you need the following steps:
   (a) Draw one of the amino acids randomly and find possible transitions.
   (b) If a move is possible, calculate the energy for the new configuration.
   (c) Accept or decline the new configuration according to the Metropolis Monte Carlo method.
   (d) Repeat the previous steps N times for an MC step, also known as a "sweep". At the end of each sweep, one should save relevant system variables.

   It is often beneficial to create a logger instance which contains information about the system as it evolves. A logger is often used when working with systems of many parameters to keep track of the observables as a function of time (or MC sweeps, in our case). Implement your own logger to suit your system.
5. Now start with a fully unfolded protein ($N = 15$) and show the configuration of the protein after the first MC steps, e.g the system after $X = 1, 10, 100$ sweeps. Use $T = 10$ as a starting temperature, which should be sufficiently high to allow most transitions to occur. Plot the energy, end-to-end distance and the RoG as a function of MC steps and comment on the tertiary structures that appear. How long does it take to reach a steady state (that is, the energy fluctuates around a constant value)? Tip: Use a running average to smooth out the energy fluctuations. By 'end-to-end', we mean the first and last amino acid of the chain, respectively.
6. At $T = 10$, the system should go through a variety of structures. Change the temperature to $T = 1$. Comment on the effect of temperature. How long does it take to reach a steady state? Increase N to $N \approx 100$. Comment on challenges of the straight-line initiation of a long monomer chain. Does it reach a steady state?
7. Next we will find a phase diagram for the transition from an unfolded state to a folded structure by annealing the system from a high to a low temperature. From this point onward, an initialisation routine of the structure can be beneficial to reduce the time to reach a steady state. High and low temperatures refer to the temperature where the system is definitely in an unfolded or folded state, respectively. You should use (at least 3) different N, e.g. $N = 15, 50, 100$.
   (a) Plot E and RoG over MC sweeps as you decrease the temperature. How long does it take to reach equilibrium at different T?
   (b) In order to find the phase diagram E(T) and RoG(T), plot the time-averaged values for a selected set of (decreasing) T over a representative temperature range. Describe how you choose the starting point for taking the time averages.
   (c) What is the critical temperature, i.e. the point where the phase transition occurs? Are the phase diagrams similar in all three (or more) cases?
8. Initiate a primary structure consisting of $N = 30$ amino acids in an unfolded state.
   (a) From this initial condition, find at least two different tertiary structures from a constant temperature simulation at $T = 1$. Compare the two results with respect to the energy and fluctuations of the energy in the final structure.
   (b) Now, use the concept of simulated annealing (SA) to find a tertiary structure for the same primary structure. SA is a method to quickly surpass many metastable phases in the search for the global energy minimum. Start above the critical temperature and then gradually go past it as you decrease the temperature continuously. Did you end up with the same structure as before? Did you reach a stable configuration faster due to the annealing process?
   (c) Discuss the existence of metastable phases and the complexity of the energy landscape in the context of the Levinthal's paradox.
9. Finally, change the sign of some monomer-monomer pair interactions in the 20×20 interaction matrix (make sure that the corresponding monomers are included in your primary structure!) and investigate the tertiary structures that arise for N = 50 by using the SA strategy. Document the new interaction energy matrix.

2.2. Task 2

1. Extend the system to 3 dimensions on a simple cubic lattice. Keep the restriction of rigid covalent bonds with a length that equals the lattice constant. Considering this restriction, how many possible lattice sites can an amino acid at the end of the chain jump? And how many possible jumps does an internal amino acid have in the 3D case?
2. Plot the evolution of a 3D protein, initialised fully unfolded, with $N = 15$ for $X = 1, 10, 100$ MC sweeps at a temperature of $T = 10$.
3. Repeat point 7 of task 1 (finding phase diagrams) in 3D and compare to the result of the 2D system. Describe and discuss differences.
4. (Bonus) Only start this implementation if you have finished all the other tasks. The Lennard-Jones potential (LJP) is one of the simplest, but efficient, mathematical representations of an atomic interaction model. This potential is an approximation to the van-der-Waals interaction. Implement this potential as a replacement for the simple nearest neighbour interaction $J_{[A(i),A(j)]}$. Choose how long the interaction energy should be (not limited to NN-interactions), and make sure that the LJP is smooth by multiplying it with a cut-off function. The use of a cut-off function is crucial to remove discontinuities that will often ruin a good model. The LJP should give the same result as the previous tasks when only taking the 1st nearest neighbour-interaction into account. Repeat point 3 (in this task), and describe and discuss the differences. Include how you implemented the potential in the report.

5. (Bonus - open ended) Define a bounding box that contains the protein. Add periodic boundary conditions and allow the protein to interact with the periodic image of itself. Initiate a folded protein and let the system evolve at high temperatures. Lower the temperature step-wise and investigate the interaction with the periodic image. This is an open-ended task which is meant to motivate further investigation into the field of protein folding and entanglement.