CSE 3504: Project - Part 1

Joseph Sweeney and Soumya Kundu

April 1, 2017

Practice 1:

1. The path length is 76

2. With 100 runs of the simulation:

    (a) The average path length is 69.23

    (b) The standard deviation of the path length is 45.02

3. With 100 runs of the simulation using a $1000X1000$ lattice:

    (a) The average path length is 74.08

    (b) The standard deviation of the path length is 51.77

Practice 2:

1. The distance between the start point and the end point is 24

2. The lattice size is 208

3. With 10 runs of the simulation:

    (a) Distance:

        i. Mean is 14.6

        ii. Median is 6

        iii. Variance is 251.04

        iv. Standard Deviation is 15.84

    (b) Lattice Size:

        i. Mean is 245.1

        ii. Median is 66

        iii. Variance is 140614.89

        iv. Standard Deviation is 374.99

4. With 100 runs of the simulation:

   (a) Distance:

        i. Mean is 15.63
        ii. Median is 13
        iii. Variance is 131.49
        iv. Standard Deviation is 11.47

   (b) Lattice Size:

        i. Mean is 223.29
        ii. Median is 130
        iii. Variance is 73434.01
        iv. Standard Deviation is 270.99

5. With 1000 runs of the simulation:

   (a) Distance:

        i. Mean is 15.20
        ii. Median is 12
        iii. Variance is 134.67
        iv. Standard Deviation is 11.60

   (b) Lattice Size:

        i. Mean is 222.22
        ii. Median is 135
        iii. Variance is 66287.33
        iv. Standard Deviation is 257.46

6. Between multiple runs of the script, with each run containing 10, 100, and 1000 runs of the simulation, the mean, median, variance, and standard deviation of both the distance and the lattice size have less variation when the number of runs of the simulation is higher. Therefore, those numbers vary much more between different runs of the script for the case of 10 runs of the simulation than they do for 1000 runs of the simulation.

Question 3:

Self-avoiding random walks pose interesting and often difficult problems in the fields of mathematics, probability theory, and enumerative combinatorics. However, their applications are also prevalent in numerous fields of scientific research. One such application is its use for protein structure prediction. The problem of protein structure prediction involves making inferences about the shape of a protein given its sequence of amino acids. Since the structure of a protein influences its function, accurately predicting protein structures can provide great insights into their functions, which can be extremely useful when investigating the role of such proteins in certain diseases. For low resolution backbone structure prediction of a protein, a two or three dimensional hydrophobic-hydrophilic square lattice model is used, where hydrophobic interactions are supposed to dominate protein folding. This happens as the protein core is formed by hydrophobic amino acids, while the hydrophilic amino acids tend to move to the outer surface due to their affinity with the solvent.

Using this model, the protein structure prediction problem can be reduced to an optimization problem with a self-avoiding walk (SAW) on a two or three dimensional lattice, depending on the resolution. The goal is to find a SAW such that the free energy of the protein, which depends on the topological neighboring contacts between hydrophobic amino acids that are not contiguous in the primary structure, is minimal. For finding the structure of a protein with an amino acid sequence of length n, we need to find the most optimal n-step SAW. However, given all the n-step self-avoiding walks and the sequence of hydrophobicity of a protein, finding a minimal conformation of the protein is very difficult. In fact, it has been proven that this problem is NP-hard. As a result, instead of trying to develop exact algorithms to solve this problem, researchers have developed heuristics to predict the most probable conformation of the protein using self-avoiding random walks. These heuristics simulate protein folding in the real biological world by folding SAWs in order to minimize the free energy of the associated conformation.

The application of self-avoiding random walks for protein structure prediction varies a lot from the simulations that were done in the previous two practices. First, the SAWs used for protein structure prediction must minimize a given cost function in order to accurately predict the true structure of the protein. In our simulations, no such optimization problem was considered. Next, the elements of the SAWs in the application discussed above are not homogeneous, as some of them are labeled as hydrophobic while others are labeled as hydrophilic, according to the given protein, to represent the two types of amino acids and their properties that are relevant to protein structure prediction. On the other hand, in our simulations, all elements of the SAWs were homogeneous. Finally, for the protein structure prediction problem, SAWs are also used on three dimensional lattices. However, our simulations were restricted to only two dimensional lattices.

3

References

1. Bahi, J.M., Guyeux, C., Nicod, J., & Philippe, L (2013). Protein structure prediction software generate two different sets of conformations. Or the study of unfolded self-avoiding walks.

2. Bhansali, R., & Tangerman, F. (2014). Are there Unfoldable Proteins in Dimension Three?