Protein Folding: An Algorithmic, Graph-Theoretic Approach

Tara Basu Trivedi

Protein folding is the process by which a sequence of amino acids (the building blocks of proteins) achieves its 3-dimensional shape. Understanding how proteins fold is of huge biological importance because, in most cases, the structure of a protein is essential to its function. Failure to fold correctly results in inactive or malfunctional proteins. Many allergies and diseases are caused by errors in the folding process of proteins, including cystic fibrosis, Alzheimer's disease, and Huntington's disease.

For all we know about the link between protein structure and function, the process of protein folding remains mostly a mystery to scientists. Three main areas of research in protein folding are:

(1) the folding code: how the specific interatomic interactions acting on amino acid sequences influence protein folding;
(2) the prediction of protein structure: how to predict the native 3-dimensional configuration of a sequence of amino acids; and
(3) the kinetics of folding: understanding how proteins can fold quickly and replicably. A famous thought experiment called Levinathal's paradox observes that a protein cannot fold by trying all possible conformations, since it would take an astronomical amount of time to do so. Proteins must therefore fold via a sequence of increasingly stable conformations.

My Capstone work focuses on the second question of structure prediction. In order to develop a computational method that runs in a reasonable amount of time, I made the following assumptions in my protein model:

- discretization of space onto a 3-dimensional lattice such that amino acids rest on the vertices of the lattice and protein folds are represented by self-avoiding paths on the lattice;
- an HP model, where amino acids are categorized either as hydrophillic (P, i.e. water-loving) or hydrophobic (H, i.e. water-hating) since proteins tend to form densely-packed cores of hydrophobic molecules surrounded by hydrophillic molecules; and
- an optimization function that counts the number of H-H amino acid contacts. The protein folding prediction algorithm under consideration seeks to find a conformation that maximizes this optimization function.

More formally, the structure prediction problem I seek to solve is defined as finding the fold that maximizes the optimization function for a given sequence of amino acids. Even under the assumptions above, finding the optimal answer to the structure prediction problem is provably NP-complete.[1] I implemented a gradient descent optimization algorithm that optimized the above optimization function for an amino acid sequence of length 100. The biggest challenge was coming up with proposal functions for subsequent locations of amino acids in the lattice that resulted in valid (self-avoiding) folds. I compared the number of H-H contacts in folds that my algorithm predicted with the theoretical upper bound of number of H-H contacts in a 3-dimensional lattice for a given sequence of amino acids: [2] $4 \min\{\#\text{ H's at odd indices in sequence}, \#\text{ H's at even indices in sequence}\} + 2$.

My algorithm's performance was spotty and highly dependent on the parameters with which it was run. No conclusions can be made based on the algorithm's output despite my considering only a simplified model of proteins. Much work still needs to be done in the development of approximation algorithms that reflect biologically-sound, rigorous, and efficient.

[1] B. Berger and T. Leighton. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *Journal of Computational Biology*, 5:27–40, 1998.

[2] W.E. Hart and S. Istrail. Lattice and off-lattice side chain models of protein folding: Linear time structure prediction better than 86% of optimal. *Journal of Computational Biology*, 4:241–259, 1997.