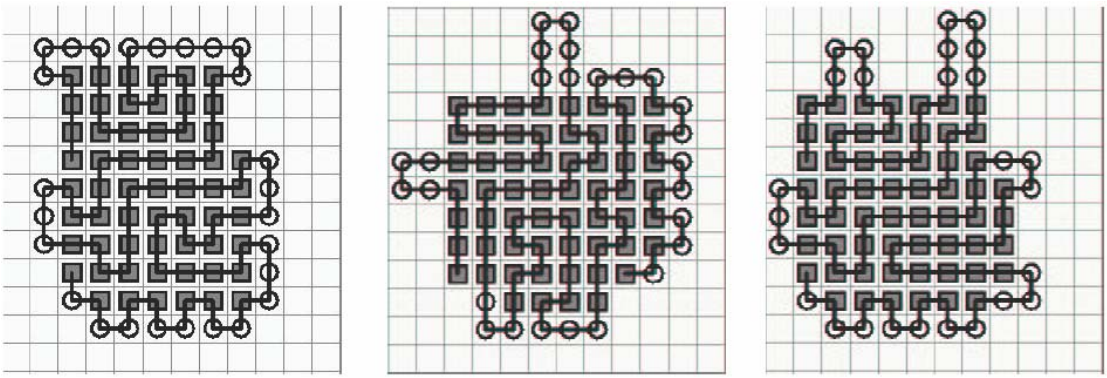
Project description: Simplified 2D HP protein folding

Background

Protein folding prediction is one of the hardest problems within the bioinformatics research fields. Such complex problems are often reduced to more easily solved instances ("more easily" does not always mean easy). First, we reduce the problem many-fold by only considering a two dimensional protein structure. Second, if we do not consider the actual amino acids in the protein other constraints and influences diminish. The reduced problem, although non-realistic, captures some important problems in protein folding.

The simplified two dimensional hydrophobic-hydrophilic (2D HP) model abstracts the hydrophobic interaction in protein folding by labeling the amino acids in a given polymer as hydrophobic (H for non-polar) or hydrophilic (P for polar). The chain must be placed on a two-dimensional grid without overlapping, so that adjacent amino acids (vertices) in the chain remain horizontally or vertically adjacent in the grid. A fold is assigned an energy score that is computed by summing certain adjacency occurrences. The energy score of a configuration on the grid is determined by counting pairs of vertices that are adjacent in the grid and both labeled H; each such pair decreases the score by 1. Pairs of vertices that are adjacent in the chain are not counted, since they must occur in all valid configurations. The goal is to minimize the energy, which corresponds to maximizing the number of adjacent hydrophobic pairs in the polymer. An unfolded amino acid chain has 0 energy.

The following example (N. Lesh et al, A Complete and Effective Move Set for Simplified Protein Folding, RECOMB'03, April 10–13, 2003, Berlin, Germany) shows three different folds for the amino acid sequence: (4H 4P 12H 6P 12H 3P 12H 3P 12H 3P 1H 2P 2H 2P 2H 2P 1H 1P 1H), with energy -53. 4H denotes four hydrophobic vertices (H H H H) etc.



Assignment

Your task is to:

1. Find out a convenient representation for amino acid chains with HP labels and coordinates.

2. Implement procedures for altering configurations by for example rotation. A simple procedure would rotate the end tip of a chain around a given vertex. Other “shortcut” procedures may also be useful.

3. Implement an energy calculation procedure given the above informal definition.

4. Think of a search strategy for solving the problem. The problem has been proven NP-complete, and hence unlikely to be solved in polynomial time. This means that a good heuristic is essential. Think of how one can avoid getting stuck in a local minimum.

5. Implement a search algorithm for solving the problem, i.e. minimizing the energy by folding the protein in two dimensions. Use some stop criterion (energy threshold, number of states visited etc.) for searching different configurations as exhaustive search will require too much computational effort.

(define programming-language ‘scheme)