

Spring 2015: CSCI 6635 Programming Assignment #2

DUE: Sunday, ~~March/01/2015~~ March/08/2015 (11pm)

Instructions

- ❑ All work must be your own (other than the instructor provided codes and hints to be used). You are not to work in teams on this assignment.
- ❑ Format: Your solution must be typed. Submit as a single compressed file (via moodle) **containing all the related files in it**. Name it as PA2_<Your_name_id>.
- ❑ The top/cover page of the report should have the title, "Spring 2015: CSCI 6635 Programming Assignment #2". Then your, "Name: _____ and ID: _____"

Description

A protein sequence can be expressed as a 2 type of amino acids only in a simplified presentation: hydrophobic (H) and hydrophilic or polar (P). For example, the sequence of Figure 1(a) can be expressed as hphpphhphpphphpph ('1' is indicating the first residue). And for the possible folds generated from the sequence, it can also be placed on a 2D (square) HP-model as showed in Figure 1.

Protein structure prediction (PSP) using hydrophobic (H) and hydrophilic (P or Polar) or HP lattice model was introduced by Dill. It uses a simplified version of amino acid sequence having only two types of monomers, namely 'H' and 'P', and the chain is placed as a self-avoiding-walk (SAW) on this lattice path. Search using this model looks for the valid conformation (i.e. SAW) which has the maximum number of topological neighboring (TN) (Figure 1) of H-H contacts, where the Hs are not already covalent bonded (or sequentially connected) within the amino acid chain or sequence.

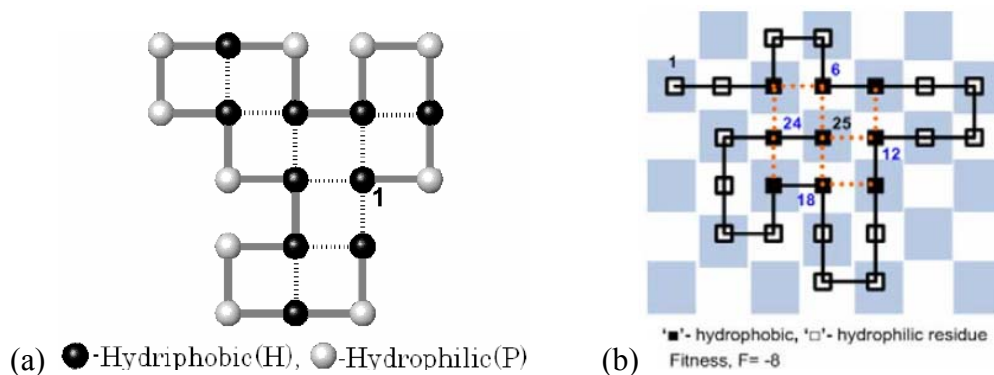


Figure-1: Conformation in 2D HP Model shown by solid line. Dotted line indicates TN.
(a) Fitness = -(TN Count) = -9. (b) Fitness = -(TN Count) = -8.

For the given HP-model we will follow a fitness function to find the best fold on a 2D HP model, which can be expressed by the following simplified energy matrix:

	H	P
H	-1	0
P	0	0

Assuming amino-acid sequence can be given as $S = S_1, S_2, S_3, \dots, S_m$. and a conformation (structure) c out of that is searched such that $c^* \in C(S)$, energy $E_{\min} = \min \{E(c) \mid c \in C\}$. Here, m is the total number of amino-acid or residue in a sequence and $C(S)$ is the set of all valid (i.e., SAW) conformations of S . If the number of TNs in a conformation c is k then the value if $E(c) = -k$ which is regarded as fitness function and expressed as $F = -k$.

To Do:

You need to develop a Genetic Algorithm (GA) based structural search algorithms. The algorithm for a given sequences will search for the best conformation or will go for minimum Energy conformation and it will visually show or draw the best conformation found in each generations along with the computed fitness value.

- Submit program code and data such a way so that it can be run to check and verify the result visually. Thus, you will need to describe, 'How to run your code', in your report.
- Well commented programming code is expected.
- In a 'readme.txt/.doc' file you must describe how to run your program. Avoid asking to install programming package to run your program, rather provide executables.

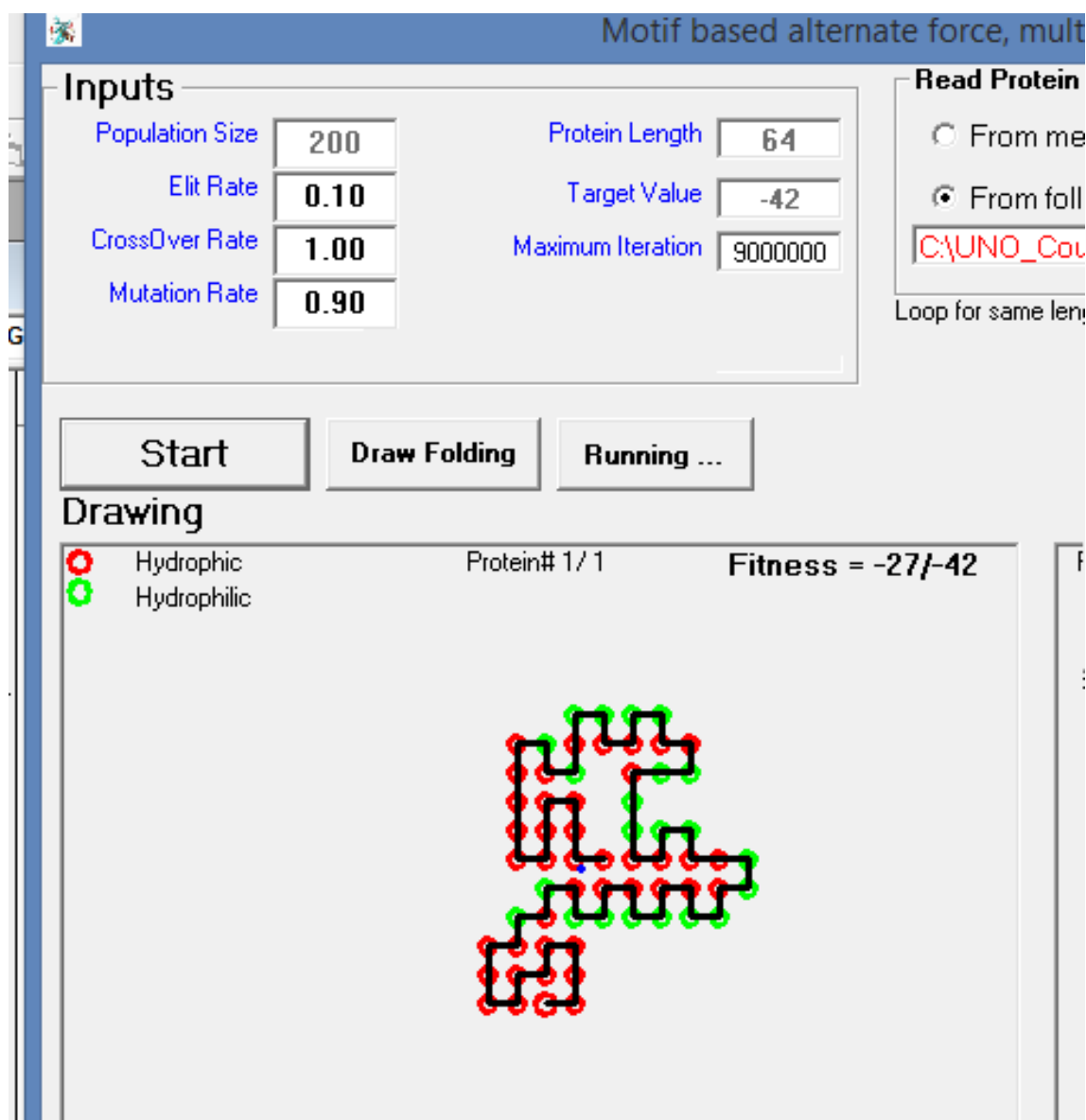
Note: A sample input file (input.txt) is provided for the problem sequences along with their best fitness found.

Declaration

Inside the PA2_<Your_name_id>.zip, you must submit a single paged document declaring that the submitted work is **completely yours and you have not either partially or fully copies someone else's assignments directly or indirectly.**

Without the declaration being submitted, the assignment will NOT be graded and the default score will be '0' (zero).

Progressing Sample (GUI) Output



--- X ---