Spr 2018: CSCI 6990, ML-II, Programming Assignment #1

DUE: Monday, Feb/19/2018 (**Softcopy** @10am; **Hardcopy**@2pm/in class)

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 including the following report**. Name it as PA1_<Your_name_id>.

- **Provide hard copy of the report in class. Your report should contain the well-commented code and snap-shots of the outputs.**

- The top/cover page of the report should have the title, "Spr 2018: CSCI 6990, ML-II, Programming Assignment #1". Then your, "Name:________________________ and ID:_______________________

Part 1 [Marks 20]

#1. Write a *Hill-Climbing* algorithm to find the maximum-value of a function f, where $f = |12 \bullet \mathit{one}(v) - 160|$. Here, v is the input binary variable of 40 bits and the *one* counts the number of '1's in v. Set MAX =100, and thus *reset* algorithm 100 times for the global maximum and print the found maximum-value for each *reset* separated by comma.

Part 2 [Marks 80]

#2. A protein sequence can be expresses as if it is made of only 2 types of amino-acids a simplified presentation: hydrophobic (H) and hydrophilic or polar (P). For example, the sequence of Figure 1(a) can be expressed as hphpphhphpphphhpphph ('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.

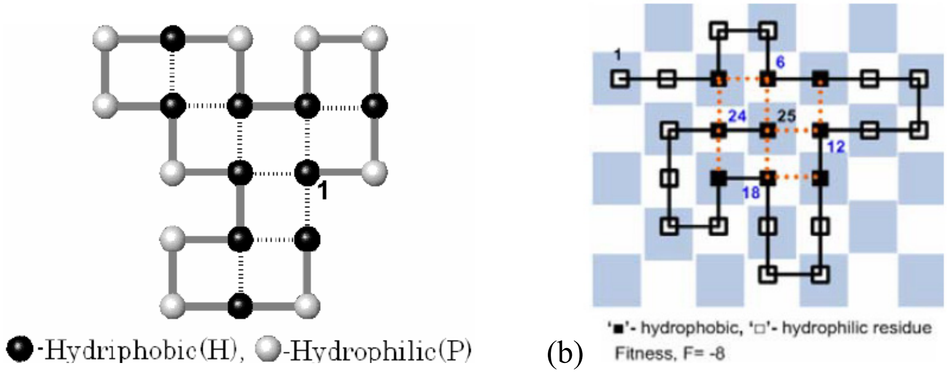

1
●-Hydriphobic(H), ○-Hydrophilic(P)
(b)
1
6
24
25
12
18
'■'- hydrophobic, '□'- hydrophilic residue
Fitness, F= -8

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 $\boldsymbol{S} = S_1, S_2, S_3, \ldots, S_m$. and a conformation (structure) ***c*** out of that is searched such that $\boldsymbol{c}^* \in C(S)$, 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 *run_readme.txt* file. Please, avoid asking to install programming package to run your program, rather provide executable(s).

- Well commented programming code will score high.

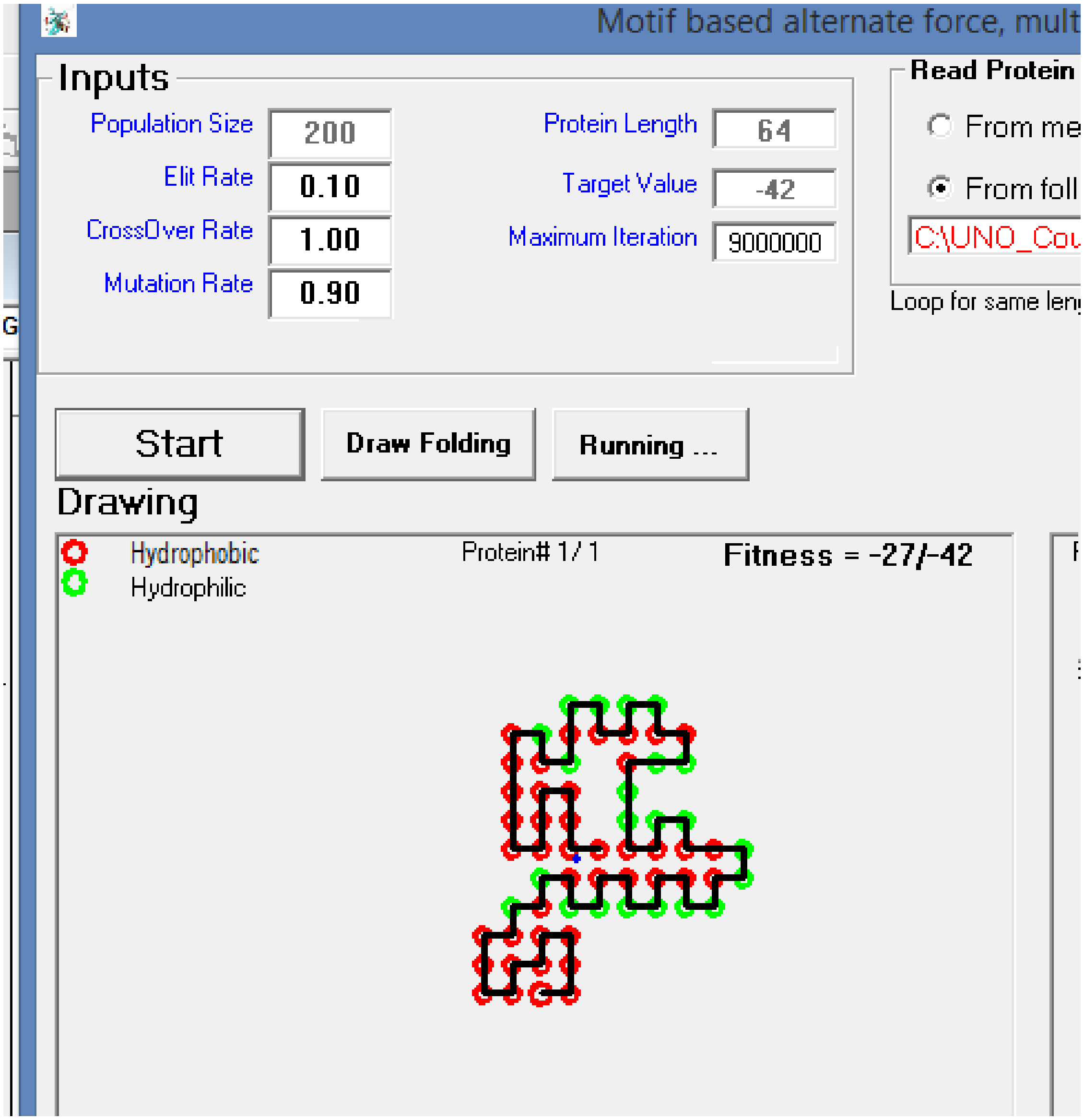
Note: A sample input file (input.txt) is provided for the problem sequences along with their best fitness found.

Declaration

Inside the PA1_<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).

Sample (GUI) Output for Part#2



Motif based alternate force, mult

Inputs

Population Size 200

Elit Rate 0.10

CrossOver Rate 1.00

Mutation Rate 0.90

Protein Length 64

Target Value -42

Maximum Iteration 9000000

Read Protein

○ From me

● From foll

C:\UNO_Cou

Loop for same len

Drawing

🔴 Hydrophobic

🟢 Hydrophilic

Protein# 1/ 1

Fitness = -27/-42

--- x ---