

# Fall 2020: CSCI 4588/5588 Prog. Assignment #1

## [Bonus => 11%]

**DUE: Monday, Sep/14/2020 (Softcopy @3 PM via Moodle)**

### Instruction

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.

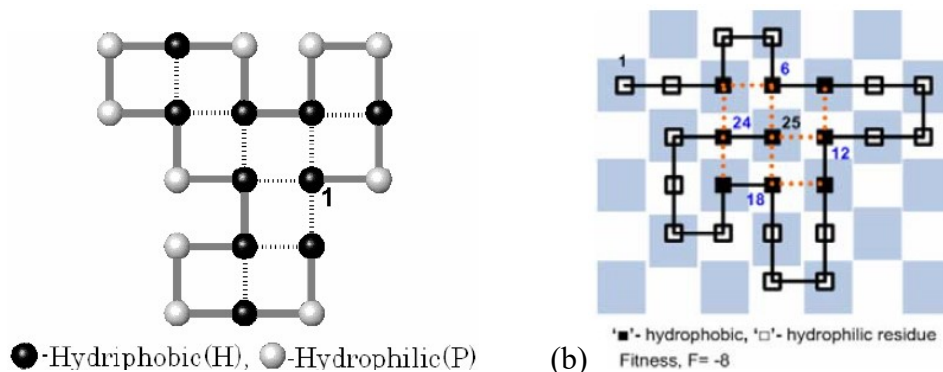
Format: Your solution must be typed. Submit as a single compressed file (via Moodle) **containing all the related files in it including the following report**. Name it as PA\_1B\_<Your\_name\_id>.

Your report should contain the well-commented code and some snapshots of the outputs.

The top/cover page of the report should have the title, "Fall 2020: CSCI 4588/5588 Programming Assignment #1". Then your, "Name: \_\_\_\_\_ and ID: \_\_\_\_\_"

A protein sequence can be expressed as a simplified presentation, which is made of only 2 types of amino-acids, hydrophobic (H) and hydrophilic or polar (P). For example, the sequence of Figure 1(a) can be expressed as hphpphhphpphphpphph ('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 the 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  $\mathcal{S} = S_1, S_2, S_3, \dots, S_m$ . and a conformation (structure)  $c$  out of that is searched such that  $c^* \in C(\mathcal{S})$ , 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  $C(\mathcal{S})$  is the set of all valid (i.e., SAW) conformations of  $\mathcal{S}$ . If the number of TNs in a conformation  $c$  is  $k$  then the value of  $E(c) = -k$  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 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 /structure found in each generation 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.
- Describe, 'How to run your code', in your *run\_readme.txt* file.
- Well commented programming code will score high.
- Please, avoid asking to install (programming) package to run your program, rather provide executable(s).

**Input:** A sample input file (input.txt) is provided for the problem sequences along with their best fitness found. Your program will read one problem sequence at a time, go for searching the corresponding solution structure. Once the termination is meet, the program will select the next problem sequence and search for the corresponding structure, and so on. The program terminates when it is done with all the sequences in the input file.

(see next)

## Sample (GUI) Output for Part#2

Motif based alternate force, multi

### Inputs

Population Size	200	Protein Length	64
Elit Rate	0.10	Target Value	-42
CrossOver Rate	1.00	Maximum Iteration	9000000
Mutation Rate	0.90		

### Read Protein

☐ From me

☒ From foll


C:\UNO\_Cou


Loop for same leng

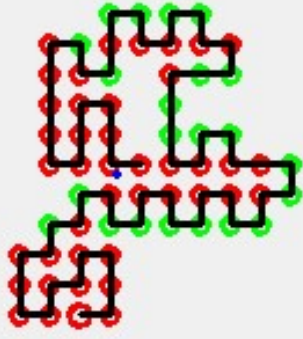
**Start** **Draw Folding** **Running ...**

### Drawing

Protein# 1/1 **Fitness = -27/-42**

 Hydrophobic

 Hydrophilic



--- X ---