Akram Hecini
Septembre 2020 - MASTER 2 BIOIFORMATIQUE

A replica exchange Monte Carlo algorithm for protein folding in the HP model

Introduction :

The technological progress that has been made lately has made it possible to obtain quick and massive data from protein sequencing.

The protein sequencing only enables us to obtain the amino acid sequence that constitutes the protein of interest but it does not have any information about the protein structure wether its secondary, tertiary and quaternary structure.

One of the main goals of researches today is to be able to predict protein structures and folding without having to perform laboratory analyzes such as X-ray crystallography. One of the solutions to do this is to create random folding models in order to try to extrapolate a protein folding as near to the reality as possible.

For several years now, different teams have tried to create models capable of predicting protein folding and among the various works carried out, there is the work of Thachuk and his team which is based on a Monte Carlo algorithm.

What I'm going to explain here in this report is my work which is based on the study carried by Thachuk et al. (2007), in order to implement a simple protein folding model with Monte Carlo algorithms.

Material and methods:

The algorithm created here will be carried out with a "Polar / Hydrophobic" sequence model (Dill 1985). This model classifies the amino acids into two categories: polar amino acids and hydrophobic amino acids.

The protein sequence is therefore modeled by a sequence of H and P amino acids. For this model, a random sequence was used which is available in the data

directory of the git. The Monte Carlo algorithm implemented here consists of taking a random amino acid in the sequence and making it perform a random movement.

In fact, each amino acid can move into one position with one square in any direction provided if the new square is empty. In order to optimize protein folding, the model seeks to find the most energetically stable 2D folding which means the model which has the lowest energy. This conformational energy corresponds to the number of neighboring hydrophobic amino acids in the 2D grid (Lau & Dill 1989).

If the energy following the movement is less than or equal to the initial energy, then the new conformation is kept.

only one parameter must be defined for the execution of this simplified Monte Carlo algorithm which is the number of times the Monte Carlo algorithm will be run: nstep.

Results :

sequence : hhpppphhh
computing the initial energy : -6

```
initial energy is: -6
[['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' 'h' 'h' 'p' 'p' 'p' 'p' 'h' 'h' 'h' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
```

final energy after computing the algorithm : also -6

```
final energy is: -6
[['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' 'p' 'p' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' 'h' 'h' '*' 'p' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' 'h' 'h' 'h' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']]
```

in this case the algorithm was applied on a sequence of 9 amino acids.

It is important to note that sometimes the algorithm fails to succzfully find a new optimal conformation where the initial energy of the conformation is identical to the final energy howver the conformation is not the same which is the case in the previous images above.

In some other cases the algorithm could successfully find a new conformation where the energy is lower than the energy of the initial conformation. (image below where the final energy is -8)

```
final energy is: -8
[['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' 'p' '*' 'h' 'p' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' 'h' '*' '*' '*' 'h' 'h' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' 'p' 'h' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']
 ['*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*' '*']]
```

In some other cases the algorithm fails to find any confirmation, it has to be executed a few more times for it to work.

Discussion :

The Monte Carlo algorithm for protein folding implemented in this project allows a slight folding of the amino acid sequence chain to be performed by adopting a lower energy conformation.
The problem of this implemented algorithm is that it is sometimes unable to predict a folding of a protein chain which could be attributed to several reasons such as the fact that the Monte Carlo algorithm is based on a random choice of amino acids to be moved and also on a random movement, also, the movements implemented in my algorithm are much simpler than in the study of Thachuk and his team. They modeled several types of movements: End moves, Corner moves, Crankshaft moves, Pull moves.. Here the movements are the same for each amino acid and do not take into consideration the conformation of the protein chain which is not the case in the study by Thachuk et al. And finally, my work only concerns a simple Monte Carlo algorithm.