REMC: A protein folding algorithm in the HP model

Author : Oussaren Mohamed

September 14, 2022

Contents

Abstract

Predicting the 2D or 3D structure of a protein from a sequence is a major problem. One of the solutions consists in evaluating the energy function. For this purpose, we have implemented the replica exchange Monte-Carlo (REMC) method based on a HP model representation [2] and the combination of well identified motions. Our results show that our algorithm is able to identify a conformation according to our expectations.

1 Introduction

One of the major problems of ab initio protein folding is the prediction of native folding from a given sequence. Indeed, the resolution of protein structures could provide invaluable information in the biological process, but currently only a few hundred thousand experimentally resolved protein structures

(about 10Therefore, the use of computational tools is necessary to allow the prediction of protein structures. Currently the best known algorithm for protein prediction is AlphaFold2. AlphaFold2 is a tool for 3D protein modelling by prediction. It is a structural prediction algorithm in Machine Learning and has made it possible to extend the coverage of solved proteins with a reliable prediction. The principle is more complex but the idea is that the algorithm will seek to create a distance matrix characterising the position of each amino acid in the target sequence and then shape a 3D structure of our protein [3].

In this project, we are interested in developing a method that is conceptually simple. We have developed the REMC algorithm (*Replica Exchange Monte-Carlo*)[1], a method based on a representation of the residues by HP Dill models [2] and the implementation of a pseudo-random motion system. This method relies on a 2D representation of the protein but remains rather time consuming since its ab initio prediction of protein structure relies on an internal representation very different from that of Alphafold2.

2 Materials et methods

The objective of REMC is to determine a conformation with minimal energy. The method consists of two steps, which we describe: transient change of replicates and evaluation of replicates. First, the sequence of interest is retrieved with which we generate n (a parameter provided at runtime) replicates with pseudorandom conformations. These replicates will undergo transient conformational changes using the Monte Carlo method. Out of all the new candidate conformations, some will undergo an exchange only if the temperature of one replicate is close to another. Furthermore, the objective of this method is to determine the conformation that minimises energy. Thus, out of a set of conformations, it is possible to have several energy minimising conformations. Note that each conformation is represented by the model proposed by Dill [2] allowing to discriminate between polar (P) and hydrophobic (H) residues, and to calculate the energy as the number of non-adjacent hydrophobic residues. Thus if two non-neighbouring residues are hydrophobic then the energy drops by -1.

Our REMC algorithm is a program interpreted on the latest version of Python, namely Python 3, and can be run on any operating system. The program loads a sequence available in a FASTA file corresponding to the protein structure we wish to predict and will allow to visualize in 2D or semi-3D the conformation of the sequence. The python module *matplotlib* is necessary for the visualisation as is the

module *tqdm* which is of slightly less importance. A number of options are available to specify the type of output of the desired program.

3 Results

We present here different results of our algorithm. Firstly, we sought to visualise the 2D structure of the p01013 protein, in particular the hydrophobic and non-hydrophobic regions that theoretically should be embedded inside and outside the protein respectively.

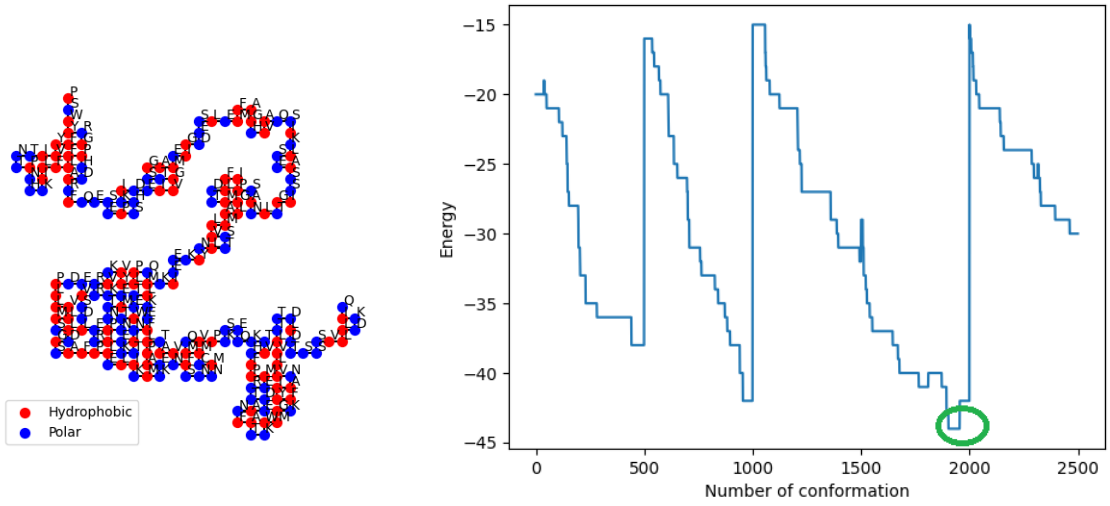

Figure 1: Representation of the 2D structure of p01013 and the energy function obtained after simulation

We provided the following parameters: optimal energy = -30, number of iterations = 500, minimum and maximum temperature at 220 and 250 respectively and number of replicates = 5. According to the figure above, in the set of generated conformations, a global minimum could be detected namely the 2000th conformation approximately generated with an energy of about -42, this is represented in green.

4 Conclusion

Finally, we have developed the REMC algorithm to predict the 2D structure of a given protein. The results presented are encouraging and show that our algorithm is able to identify a conformation whose energy level is minimised. Optimisations can be made to allow calculations to be performed on potentially longer sequences since the algorithm performs calculations over polynomial time.

References

[1] Thachuk C, Shmygelska A, Hoos HH. A replica exchange Monte Carlo algorithm for protein folding in the HP model. BMC Bioinformatics. 2007 Sep17;8:342. PubMed PMID: 17875212; PubMed Central PMCID: PMC2071922.

[2] Dill KA. Theory for the folding and stability of globular proteins. Biochemistry. 1985 Mar 12;24(6):1501-9. doi: 10.1021/bi00327a032. PMID: 3986190.

[3] Jumper, J., Evans, R., Pritzel, A. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). https://doi.org/10.1038/s41586-021-03819-2

[4]] Tunyasuvunakool, K., Adler, J., Wu, Z. et al. Highly accurate protein structure prediction for the human proteome. Nature 596, 590–596 (2021). https://doi.org/10.1038/s41586-021-03828-1