Algorithms in Computational Biology (INFO-F438)
Assignment 1: Optimal Protein Folding in the HP Model

Jean Cardinal (`jcardin@ulb.ac.be`)
Sofia Papadimitriou (`Sofia.Papadimitriou@vub.be`)

February 2020

Protein Folding in the Hydrophobic-Hydrophilic Model

Protein folding is one of the major challenges in computational biology. Here we will consider a well-known simple model that was devised by K. A. Dill around 1985. In this model, the amino acid types are classified into two classes: hydrophobic (H), and hydrophilic (P). The best folding is the one that maximizes the number of adjacent but not covalently bound $H - H$ pairs, thereby allowing for the hydrophobic effect.

Practically, the protein is modeled as a simple string of n symbols in the alphabet $\{H, P\}$. For our purpose here, a folding is a representation of the string on a square grid. The symbols that are adjacent in the string must be adjacent in the grid (in the 4-way square adjacency sense) and no more than one symbol can occupy a square. Hence in particular, the successive symbols of the string must form a non-intersecting walk on the grid. The score of such a folding is the number of pairs of hydrophobic (H) symbols that are adjacent in the grid but not in the string. An example of a folding of the string $HPHHHHPPHPPH$ is given below. The score of this folding is 4 (see red dashed adjacencies).
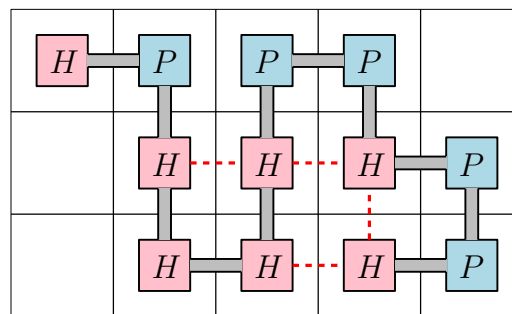
Your Work

Write a program that takes a string in $\{H, P\}^n$ as input and outputs the maximum score of a folding. For this purpose, you can search the space of foldings exhaustively using recursion. You should also think about a simple branch-and-bound technique to quickly discard uninteresting branches of the exploration.

Requirements:

1. The source code of a program performing the above task, preferably in Python.

2. A number of relevant, non-redundant comments and explanations about your code, either in the form of comment lines, or in a separate report

3. The result output by your program on the following input examples:

   - `HHHHHHH`
   - `HPPHPHP`
   - `HHHPHPPHPHPPHHHPH`

Further Readings

Alantha Newman: A new algorithm for protein folding in the HP model. *ACM-SIAM Symposium on Discrete Algorithms* 2002: 876-884.

Evaluation

The evaluation will be based on the following criteria:

1. the general understanding of the instructions,

2. the proper use of the programming language,

3. the efficiency and correctness of the implemented algorithm,

4. the clarity and relevance of the comments and explanations.

Ethics

Plagiarism will be severely sanctioned. Plagiarism cases include reusing someone else's written or drawn material, or any kind of work, without an explicit quote or reference.

Please refer to the following webpages:

`http://www.bib.ulb.ac.be/fr/aide/eviter-le-plagiat/`
`http://www.plagiarism.org/.`

Deadline

March 4, 2020.