Algorithms and Data Structures for Biology
11 March 2019 — Lab Session

Ugo Dal Lago Thomas Leventis

1 The Problem

The problem we will deal with during this lab session is the one of approximate string matching (ASM), that we have already considered in the theory module. We will consider strings over the alphabet $\Sigma = \{\mathtt{A}, \mathtt{T}, \mathtt{C}, \mathtt{G}\}$.

Edit Distance. There are three *editing operations* on strings: *insertion* (for instance st turns into $s\mathtt{A}t$), *deletion* ($s\mathtt{A}t$ turns into st), *substitution* ($s\mathtt{A}t$ turns into $s\mathtt{T}t$, $s\mathtt{C}t$ or $s\mathtt{G}t$). The *edit distance* between two strings s and t is the minimal number of editing operations required to go from s to t.

Approximate String Matching. Given a string s of size $|s|$, a pattern p and a distance d, we want to find all the positions i between 1 and $|s|$ such that some substring of s starting at position i is at distance at most d from the pattern p.

2 A Suggested Algorithm

You need to define an algorithm to solve the ASM problem. A relatively simple algorithm could be built along the following lines:

1. First of all write an algorithm which, given a string p and a natural number d, generates all the strings which are at edit distance at most d from p.
2. Then write an exact string matching algorithm, but generalized to sets of patterns: given a string s and a *set* of strings P, it should produce in output the subset of $\{1, \ldots, |s|\}$ of positions i such that a substring of s starting in i is in P.
3. Finally, a third algorithm which simply (appropriately) calls the two algorithms we have just defined needs to be given.

3 The Program

Implement the algorithm defined above in Python. Please do so without invoking any external module.

4 The Efficiency

Define a function to generate random strings of given length over the alphabet Σ and use it to test your program. To do so you can use the function `choice` from the module `random`.

In particular, generate strings s of length $10, 20, 30, \ldots, 100$ and patterns p of length $2, 4, 6, 8$. How long does it take to solve the ASM with a distance 1? 2? 3? 4?