"Algorithms and Data Structures for Computational Biology"
Module 3 – 1/22/2018

1. The command *diff* of the *Unix* operating system examines two text files and outputs their differences in terms of rows. For instance, given the two input files *P* and *T* shown in the first two colums, *diff(P, T)* gives the output file *U* shown in the third column:

```
Questo è il testo originale
alcune linee non dovrebbero
cambiare mai
altre invece vengono
rimosse
altre vengono aggiunte
```

```
Questo è il testo nuovo
alcune linee non dovrebbero
cambiare mai
altre invece vengono
cancellate
altre vengono aggiunte
come questa
```

```
- Questo è il testo originale
+ Questo è il testo nuovo
  alcune linee non dovrebbero
  cambiare mai
  altre invece vengono
- rimosse
+ cancellate
  altre vengono aggiunte
+ come questa
```

The problem of designing the pseudo-code for the *diff* command can be solved by a *dynamic programming* algorithm. One can assume that *P* and *T* have, respectively, *m* and *n* rows, and that two rows can be compared in O(1) time, since the number of characters in each row is upper bounded by a constant and each character is coded by a constant number of bits. Define first the recurrence relations giving the optimal sub-structure property of such a dynamic programming algorithm, and then write its corresponding pseudo-code and analyze its complexity.

2. Consider the string "b a b a c a r". Write (by hand) its corresponding:

- Suffix trie;
- Suffix tree;
- Suffix array;
- Burrows-Wheeler transform;
- LF mapping.