

Uvod u Bioinformatiku, Junski rok

Industrija 4.0

Školska godina 2022/2023

Uputstvo:

Na Desktop-uu se nalazi folder sa nazivom BIO.JUN u kome se nalaze nepotpuni kodovi rešenja zadataka koje treba dopuniti kako bi traženi algoritmi bili implementirani. Folder preimenovati u format bio_ispitniRok_Prezime_Ime_BrIndeksa (npr. bio_jun_Peric_Pera_4005-2021) i sve kodove čuvati u preimenovanom folderu dopunjujući nepotpune kodove.

1. [12.5p] Napisati kod *Median String* algoritma koji pronalazi kanonski motiv za dati skup sekvenci `dna_sequences` određivanjem niske medijane tj. `k`-grama za koji postoje u svim sekvencama `k`-grami (instance kanonskog motiva) koji u odnosu na njega imaju minimalno ukupno odstupanje.

Primer:

```
dna_sequences = ['ttacggtAAC',
                 'gATAtctgtc',
                 'ACGgcgttcg',
                 'ccctAAAgag',
                 'cgtcAGAggt']
```

`k = 3`

Izlaz: 'AAA'

2. [12.5p] Napisati kod algoritma *Longest Common Subsequence* koji pronalazi najdužu zajedničku podsekvencu za niske `seq1` i `seq2`. Treba vratiti i skor pronađene najduže podsekvence, tj. broj njenih tačnih poklapanja (*matches*) sa odgovarajućim podsekvencama niski `seq1` i `seq2`.

Primer:

```
seq1 = 'ACGTCCATTA'
seq2 = 'ATGTTATA'
```

Izlaz: LCS score: 6

LCS: AGTATA

3. [12.5p] Napisati kod algoritma *Aditivne filogenije* koji konstruiše nekorensko filogenetsko stablo koje odgovara matrici rastojanja `D` dimenzije $n \times n$.

Primer:

```
D = [[0, 13, 21, 22],
      [13, 0, 12, 13],
      [21, 12, 0, 13],
      [22, 13, 13, 0]]
```

`n = 4`

Izlaz: {0 : [('1+0', 11.0)], 1 : [('1+0', 2.0)], '1+0' : [(1, 2.0), (0, 11.0), ('2+1+0', 4.0)],
2 : [('2+1+0', 6.0)], '2+1+0' : [(2, 6.0), ('1+0', 4.0), (3, 7.0)], 3 : [('2+1+0', 7.0)]}

4. [12.5p] Napisati kod koji korišćenjem strukture *prefiksnog stabla* vrši pretraživanje (uparivanje) šablona iz liste `patterns` unutar date niske `sequence`.

Primer:

```
patterns = ['ACCG', 'CTCT', 'GGA', 'TATA', 'ATG']
sequence = 'ATATGCTCTTGCTAGATGTGCTATA'
```

Izlaz: [('ATG', 2), ('CTCT', 5), ('ATG', 15), ('TATA', 21)]