Mini projekt 1. Kje so geni?

Ivan Antešić (63130003)

November 5 2019

1 Odgovori na vprašanja

- **Kolikšna je dolžina genoma bakterije Mycoplasma genitalium?**
  580076 bp.
- **Koliko genov vsebuje?**
  563. Od tega je 509 genov, ki se prevedejo v proteine.
- **Kolikšna je dolžina največjega, najmanjšega gena in mediana dolžina genov (v kodonih)?**
  Če upoštevamo samo 509 genov, ki se preslikajo v proteine in katere smo analizirali je minimalna dolžina 37, maksimalna 1805 in mediana 287 kodonov. Če upoštevamo vse gene je minimalna dolžina 21, maksimalna 1803 in mediana 237 kodonov.
- **Kolikšna je občutljivost (angl. recall) in natančnost (angl. precision) vašega algoritma za iskanje genov pri L=50 in L=125 kodonov?**
  Pri L=50 je občutljivost enaka 0,81, natančnost pa 0,33. Pri L=125 sta občutljivost in natančnost obe enaki 0,72.

2 Rezultati

Na sliki 1 so prikazani rezultati osnovnega algoritma, ki upošteva vsa dovoljena odprta bralna okna (anlg. ORF - open reading frames).Če odgovorimo na zadnje vprašanje z osnovnim algoritmom bi bila pri L=50 občutljivost enaka 0,81, natančnost pa 0,33. Na sliki 2 pa so vidni rezultati algoritma, ki med iskanjem genov zajame le največja odprta bralna okna in morebitna okna, ki so vsebovana znotraj večjih oken, ne upošteva. S tem ne zajamemo vseh možnih genov - namesto približno 4500 jih zajamemo okoli 1800.

Z primerjavo rezulatov lahko opazimo občutno izboljšanje natančnosti - mere, ki nam pove kolikšen delež od pridobljenih genov je previlen. To je predvsem zaradi manjšega števila vseh pridobljenih genov. Zaradi istega razloga pa nam malo upade mera občutljivosti, ki opisuje kolikšen delež vseh pravilnih genov smo zajeli. Iz grafa 2 lahko razberemo, da je optimalna minimalna dolžina L ravno 125 kodonov, ker pri tej vrednosti ujamemo ravnovesje natančnosti in občutljivosti.

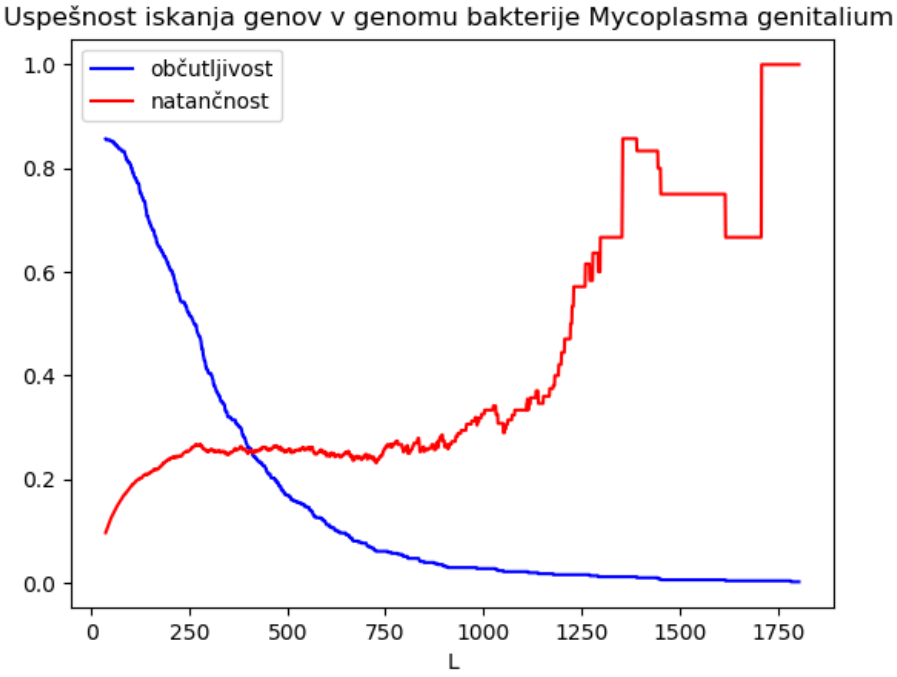

Figure 1: Graf občutljivosti in natančnosti algoritma v odvisnosti od minimalnega števila kodonov v genu (L).

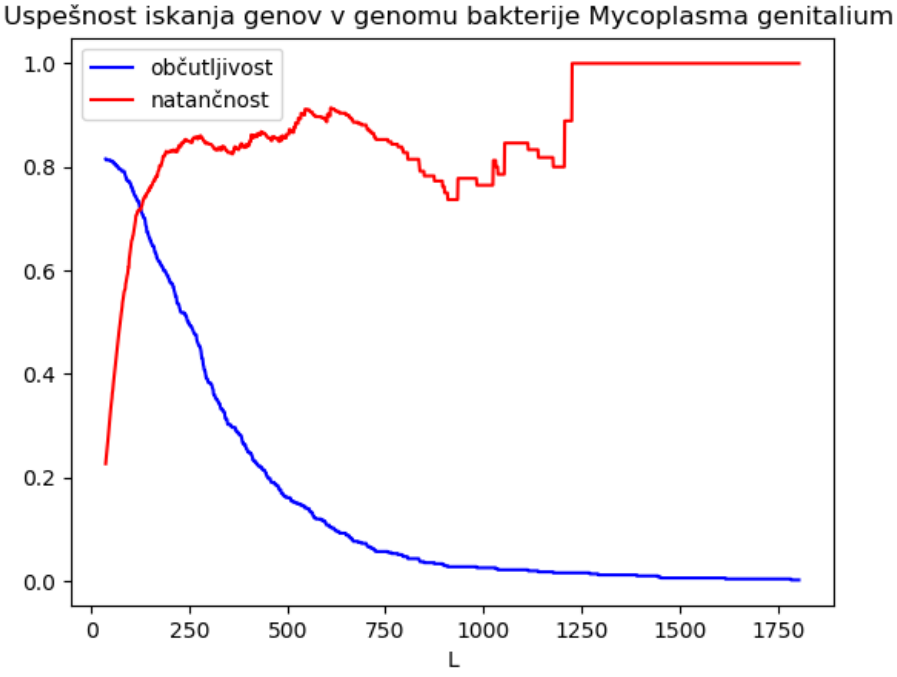

Figure 2: Graf občutljivosti in natančnosti algoritma v odvisnosti od minimalnega števila kodonov v genu (L). Algoritem v tem primeru uporablja opisano izboljšavo.