SVEUČILIŠTE U ZAGREBU

FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

BIOINFORMATIKA 1

Navarrov algoritam za približno uspoređivanje teksta

Fran Medved, Adam Vuković

Voditelj: Krešimir Križanović

Zagreb, svibanj, 2024.

**Sadržaj**

[1. Uvod 1](#_Toc159987575)

[2. Seminarski rad 2](#_Toc159987576)

[3. Zaključak 3](#_Toc159987577)

[4. Literatura 4](#_Toc159987578)

[5. Sažetak 5](#_Toc159987579)

# Uvod

Približno uspoređivanje teksta jedan je od temeljnih zadataka bioinformatike. To je tehnika koja se koristi za pronalaženje podudarnosti između tekstualnih nizova kada dopuštamo određene razlike, kao što su umetanja, brisanja ili zamjene znakova. Klasična implementacija ove tehnike koristila je dinamičko programiranje te je zahtijevala vremensku složenost od O(mn), gdje je n duljina teksta, a m duljina uzorka koji tražimo u tekstu.

Danas se koristi pretraživanje uzoraka u hipertekstu. U takvoj implementaciji, tekst je prikazan kao graf gdje čvorovi sadržavaju dio teksta, a bridovi označavaju iduće dijelove teksta koji mogu slijediti trenutni čvor. Amir i suradnici razvili su algoritam koji ima vremensku složenost od O(m(n log m + e)) i prostornu složenost od O(mn), gdje je e broj bridova.

U sklopu ovog projekta, implementirat ćemo algoritam profesora G. Navarra koji ima vremensku složenost O(m(n + e)) i prostornu složenost O(n). Algoritam ćemo usporediti s bit parallel sequence-to-graph alignment algoritmom na 3 vrste graf topologija (linearni, SNP i twopath graf). Nećemo razmatrati topologije koje imaju cikluse, poput tangle grafa.

# Navarrov algoritam

**Opis algoritma**

Prvo opišimo klasični algoritam. U ovom algoritmu stvaramo matricu *Ci, j* na način da je *Ci, 0 = C0, i = i* za sve *i*, te za sve ostale elemente na način:

*Ci, j = ako (a[i] = b[j]) onda Ci−1, j−1*

*inače 1 + min(Ci−1, j, Ci, j−1,Ci−1, j−1),*

gdje *Ci−1, j* predstavlja brisanje, *Ci, j−1* predstavlja umetanje te *Ci−1, j−1* predstavlja zamjenu. Važno je zamijetiti kako je prostorna složenost ovog algoritma O(mn) što je jako skupo za tekstove koji imaju veliku duljinu n.

Navarrov algoritam rješava taj problem tako što pamtimo samo jedan stupac naše matrice Ci, j tako da je prostorna složenost O(m). Točnije, za računanje *j*-tog stupca treba nam samo *j-1.* stupac.

Primjenjujući algoritam na hipertekst, ...

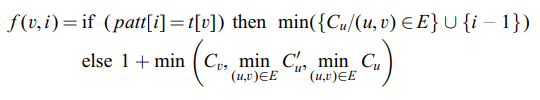
Search (V, E, patt)

1. for all 𝑣∈𝑉, 𝐶𝑣 ← 0.

2. for 𝑖=1 to 𝑚

3. for all 𝑣∈𝑉, 𝐶𝑣′←𝑓(𝑣,𝑖)

4. for all 𝑣∈𝑉, 𝐶𝑣←𝐶𝑣′



**Primjer**

# Rezultati testiranja

# Zaključak

# Literatura

[1] Mikko Rautiainen, Veli Makinen, Tobias Marschall; *Bioinformatics*, Volume 35, Issue 19, 1 October 2019; Bitt-paralleel sequence-to-graph alignment https://academic.oup.com/bioinformatics/article/35/19/3599/5372677

[2] Gonzalo Navarro; *Theoretical Computer Science*, Volume 237, Issues1-2, 28 April 2000; <https://www.sciencedirect.com/science/article/pii/S0304397599003333>