SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4191

REKONSTRUKCIJA FILOGENETSKOG STABLA KORISTEĆI METODU UDRUŽIVANJA SUSJEDA

Lucija Megla

Zagreb, lipanj 2015.

Hvala mojoj obitelji na svemu!

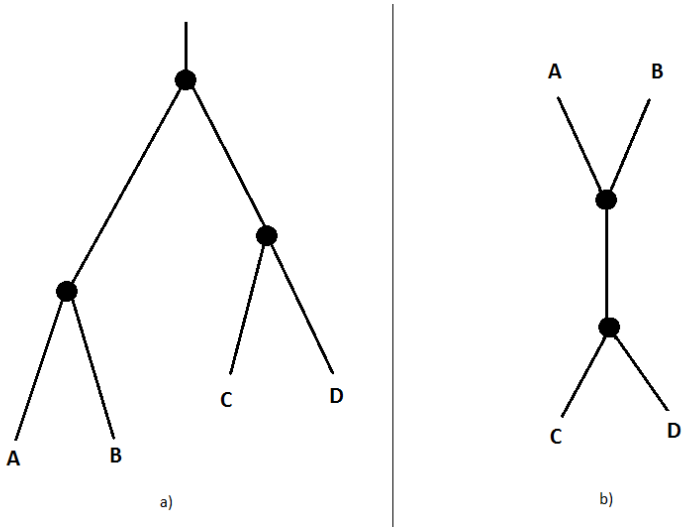
Hvala Ani Bulović na pomoći pri izradi ovog rada i hvala mom mentoru Mili Šikiću!

Sadržaj

1. Uvod

Filogenija ili filogeneza je područje istraživanja koje se bavi pronalaskom bioloških veza između vrsta na temelju njihovih osobina i karakteristika. Glavna je pretpostavka ove grane biologije međusobno srodstvo promatranih vrsta kako bi se na temelju DNK-a ili sekvence proteina odredilo njihovo srodstvo. Središnji je pojam filogenije filogenetsko stablo, pomoću kojega se najčešće prikazuju evolucijski odnosi među organizmima. Listovi filogenetskog stabla označavaju organizme koji mogu biti pripadnici neke taksonomske jedinice, odnosno taksona: vrste, populacije itd.

Razlikujemo dva osnovna tipa filogenetskog stabla: ukorijenjeno i neukorijenjeno stablo.



Slika 1.1 a) ukorijenjeno stablo i b) neukorijenjeno stablo

Ukorijenjeno stablo određuje odnos između pretka i potomka, odnosno određuje smjer evolucijskog procesa (Slika 1.1a).
Neukorijenjeno stablo prikazuje relativne udaljenosti između taksonomskih jedinica,

ali nema uređen odnos između pretka i potomka, odnosno ne uzima u obzir pretpostavljeni tok vremena (Slika 1.1b).

Za razumijevanje tijeka evolucije i otkrivanja povezanosti između taksonomskih jedinica potrebno je rekonstruirati filogenetsko stablo. Ta se rekonstrukcija danas vrlo često obavlja računalom pomoću niza algoritama koji uvijek balansiraju točnost i brzinu izvođenja.

2. Pregled metoda za rekonstrukciju filogenetskih stabala

Metode za rekonstrukciju filogenetskih stabala se mogu podijeliti u dvije glavne skupine:

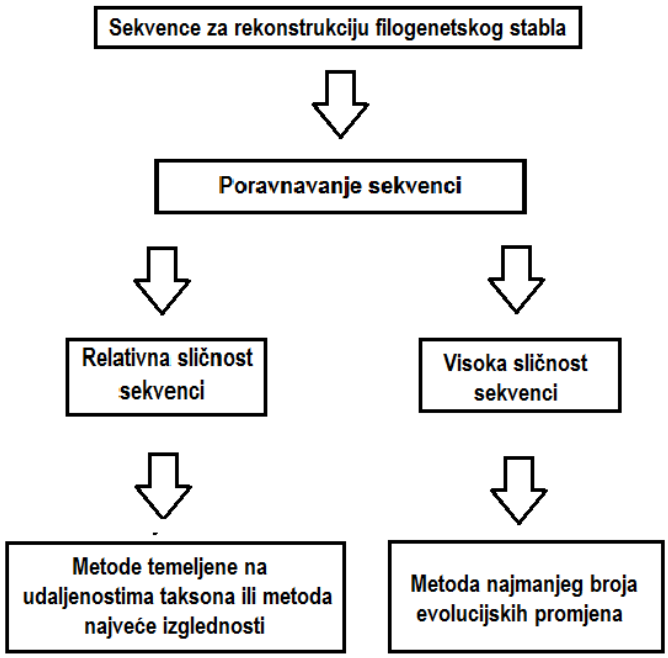
1. Metode temeljene na obilježjima taksona
2. Metode temeljene na udaljenostima između taksona

Metode temeljne na obilježjima taksona se oslanjaju na poravnavanje više sekvenci odjednom (*eng. multiple sequence alignment*). Svaka poravnata pozicija nosi evolucijsku informaciju o organizmima, koja se koristi za izgradnju filogenetskog stabla. Stablo izgrađeno na ovakav način se naziva kladogram. Metode koje spadaju u ovu skupinu su: metoda najmanjeg broja evolucijskih promjena i metoda najveće izglednosti.

Metode temeljene na udaljenostima između taksona koriste matricu udaljenosti na temelju koje određuju izgled filogenetskog stabla. Matrica udaljenosti se određuje računajući udaljenosti između DNK-sekvenci ili proteina. Svaka razlika u sljedovima DNK ili proteina povećava udaljenost, dok sličnost smanjuje ukupnu udaljenost. Udaljenost se računa pod pretpostavkom da su srodnosti i sličnosti organizama u korespondenciji, odnosno da su evolucijske udaljenosti i razlike među organizmima ekvivalentne. Metode koje spadaju u ovu skupinu su: UPGMA i metoda udruživanja susjeda.

Metode temeljene na obilježjima taksona su sporije i memorijski zahtjevnije od metoda temeljenih na udaljenostima među taksonima, ali daju bolje i kvalitetnije rezultate kod rekonstrukcije filogenetskih stabala, i za DNK slijedove i za proteinske slijedove ([2], [3]). Metode temeljene na udaljenostima su zbog svoje brzine pogodne za analizu većeg broja sekvenci čiji organizmi nisu u velikoj srodnosti, no sažimanjem evolucijskih podataka sadržanih u sekvencama u jedan broj, na temelju kojeg se radi rekonstrukcija, daje prostora pogreškama ([2], [3]).

Nažalost, niti jedna od ovih metoda ne garantira da će rekonstruirano filogenetsko stablo uistinu prikazati prave evolucijske odnose među danim sekvencama, ali pametnim biranjem metoda možemo dobiti najtočniji rezultat (Slika 2.1).



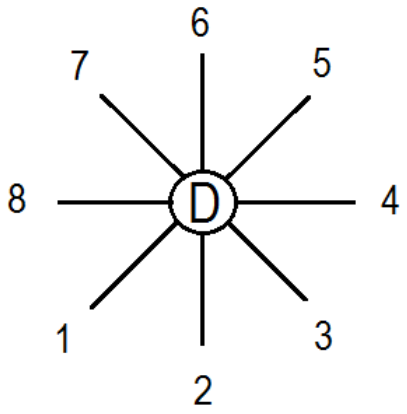
Slika 2.1 – odabir pogodnih metoda za rekonstrukciju filogenetskog stabla

U daljnjim poglavljima ćemo se orijentirati na metodu udruživanja susjeda kao jednu od metoda rekonstrukcije filogenetskih stabala.

3. Metoda udruživanja susjeda

Kao što je prije navedeno, metoda udruživanja susjeda spada u skupinu metoda temeljenih na udaljenostima taksona, odnosno ova metoda rekonstruira filogenetsko stablo pomoću matrice udaljenosti. Osim što određuje topologiju filogenetskog stabla, ova metoda određuje i duljinu grana filogenetskog stabla, odnosno udaljenosti između svakog čvora stabla. Grana filogenetskog stabla predstavlja evolucijsku udaljenost između bioloških sljedova iz kojih se radi rekonstrukcija filogenetskog stabla.

U idućim ćemo poglavljima opisati metodu udruživanja susjeda, dati njezin pseudokod i pokazati rad algoritma na jednom primjeru.



Slika 3.1.1: Zvjezdoliko stablo

3 .1 Opis metode

Cijeli se postupak metode udruživanja susjeda temelji na obnavljanju topologije stabla spajanjem dvaju taksona u svakom koraku tako da zbroj grana stabla u tom koraku bude minimalan. Kako je prije spomenuto, za spajanje taksona ova metoda koristi matricu udaljenosti koja sadrži sve udaljenosti između taksona. Udaljenost između taksona *i* i *j* ćemo označavati oznakom D_{ij}, što je ujedno i *ij*-ti element

matrice udaljenosti. Kako bi udaljenost D_{ij} mogli proglasiti mjerom udaljenosti između taksona, za funkciju D: N x N → R moraju vrijediti sljedeći izrazi (1.1):

1. $D_{ij} \geq 0$;
2. $D_{ij} = 0$ ako i samo ako $i = j$;
3. $D_{ij} = D_{ji}$ (simetričnost); (1.1)
4. $D_{ij} + D_{jk} \geq D_{ik}$ (nejednakost trokuta);

   za svaki i, j, k, definirano na skupu taksona N.

Metoda na početku uzima svih *n* poravnatih taksona sa izračunatom matricom udaljenosti i od njih stvara zvjezdoliko stablo (Slika 3.1.1).

Kako bi došli do sume zvjezdolikog stabla, uvest ćemo oznaku D_{iD} koja predstavlja duljinu grane od taksona *i* do čvora D. Vidljivo je sa slike 3.1.1 da ukupnu sumu grana stabla možemo izraziti kao sumu svih duljina između taksona i čvora D:

$$S_0 = \sum_{i=1}^{N} D_{iD} \tag{1.2}$$

Međutim, duljine između taksona i središnjeg čvora D nisu poznate, no ono s čime raspolažemo je matrica udaljenosti i duljine D_{ij}.

Kada raspišemo udaljenosti D_{ij} između taksona *i* i *j* dobijemo sljedeći izraz:

$$D_{ij} = D_{iD} + D_{Dj} \tag{1.3}$$

što prema svojstvu 3 za mjeru udaljenosti (izraz 1.1) možemo zapisati kao:

$$D_{ij} = D_{iD} + D_{jD} \tag{1.4}$$

Ako probamo raspisati sumu svih udaljenosti između taksona dolazimo do izraza (1.5):

$$\sum_{i<j}^{N} D_{ij} = D_{13} + \ldots + D_{1n} + D_{23} + \ldots + D_{2n} + \ldots + D_{(n-2)(n-1)} + D_{(n-2)n} + D_{(n-1)n} \quad (1.5)$$

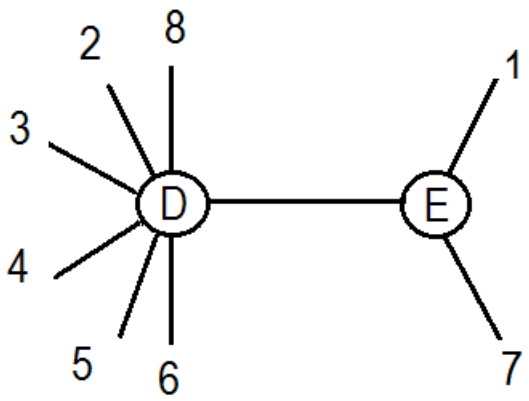
Kada svaki član izraza (1.5) zamjenimo izrazom (1.4) i grupiramo iste članove dolazimo do formule:

$$\sum_{i<j}^{N} D_{ij} = (n-1)D_{1D} + (n-1)D_{2D} + \ldots + (n-1)D_{(n-1)D} + (n-1)D_{nD} \quad (1.6)$$

Ako izraz (1.6) podjelimo sa $(n-1)$, možemo uočiti da je desna strana jednaka formuli za računanje sume S_0. Tako konačno dobivamo izraz za sumu zvjezdolikog stabla (slika 3.1.1):

$$S_0 = \sum_{i=1}^{N} D_{iD} = \frac{1}{N-1} \sum_{i<j} D_{ij} \quad (1.7)$$

Idući je korak nakon stvaranja zvjezdolikog stabla pronalaženje prva dva taksona koja će činiti prvi par susjeda. Njih možemo izabrati na $N(N-2)/2$ načina, pri čemu je N broj taksona. No, mi ne želimo izabrati bilo koja dva taksona, već želimo, kako je prije rečeno, odabrati dva taksona čijim će spajanjem zbroj duljina grana nastalog stabla biti minimalan.



Slika 3.1.2: Izgled stabla nakon udruživanja taksona 1 i 7

Spajanjem dva taksona dodajemo novi čvor u stablo, čiji izgled možemo vidjeti na slici 3.1.2., u slučaju da su takson 1 i takson 7 odabrani kao par.

Kako bi odredili točan par taksona koje ćemo spojiti novim čvorom, moramo za svaki mogući par *i* i *j* odrediti sumu grana stabla koje će nastati sparivanjem ta dva taksona. Tako ćemo zbroj grana u stablu sparivanjem taksona *i* i *j* označiti s S_{ij}. Izraz pomoću kojeg se računa ukupan zbroj grana u stablu glasi:

$$S_{ij} = D_{xy} + D_{ix} + D_{jx} + \sum_{k!=i,k!=j}^{N} D_{ky} \tag{1.8}$$

Označimo udaljenost D_{xy} iz izraza (1.8) kao udaljenost između dva unutarnja čvora stabla, D_{ix} i D_{jx} kao udaljenosti taksona *i* odnosno *j* od čvora X s kojim su spojeni, a $\sum_{k!=i,k!=j}^{N} D_{ky}$ kao sumu svih udaljenosti između taksona koji nisu odabrani za udruživanje i čvora Y na kojeg su spojeni. Izraz (1.8) možemo još dodatno raspisati. Naime, udaljenost između dva unutrašnja čvora D_{xy} moramo odrediti iz poznatih podataka. Ideja je da zbrojimo sve udaljenosti koje sadrže duljinu između čvorova X i Y te potom oduzmemo sve nebitne grane:

$$D_{xy} = \frac{1}{2(N-2)}\left[\sum_{k!=i,k!=j}^{N}(D_{ik} + D_{2k}) - (N-2)(D_{ix} + D_{jx}) - 2\sum_{k!=i,k!=j}^{N} D_{ky}\right] \tag{1.9}$$

Prvi izraz u zagradi izraza (1.9) je zbroj svih udaljenosti između taksona *i* i *j* i ostalih taksona. Ako bolje promotrimo, uvidjet ćemo da sve te udaljenosti sadržavaju duljinu D_{xy}. Kako bi izbacili irelevantne duljine, oduzimamo $(N-2)(D_{ix} + D_{jx})$ od prvog izraza jer smo upravo te dvije duljine prebrojali $(N-2)$ puta – imamo dva taksona koja uparujemo i računamo duljine prema ostalih $N-2$ taksona. Još moramo oduzeti od prvog izraza sve duljine grana od čvora Y do preostalih $N-2$ taksona. Kako smo svaku tu granu prebrojali 2 puta, prvi put kod računanja duljine od taksona *i,* a drugi put kod računanja od taksona *j,* izraz koji moramo oduzeti je upravo $2\sum_{k!=i,k!=j}^{N} D_{ky}$. Kako bi do kraja izračunali izraz za sumu stabla, moramo raspisati izraz

$\sum_{k!=i,k!=j}^{N} D_{ky}$ preko poznatih vrijednosti iz matrice udaljenosti. Ako se poslužimo istom logikom kao u izrazima (1.5) i (1.6), pri raspisu ćemo uvidjeti da se ovaj izraz može zapisati kao:

$$\sum_{k!=i,k!=j}^{N} D_{ky} = \frac{1}{N-3}\sum_{m,n}^{N} D_{mn} \tag{1.10}$$

odnosno suma svih taksona vezanih za čvor Y je jednaka sumi svih njihovih međusobnih udaljenosti podjeljenoj sa $N-3$. Kada izraze (1.9) i (1.10) uvrstimo u izraz (1.8) dobit ćemo konačnu formulu za izračun sume stabla kada odaberemo dva taksona za udruživanje:

$$S_{ij} = \frac{1}{2(N-2)} \sum_{k!=i,k!=j}^{N} (D_{ik} + D_{jk}) + \frac{1}{2} D_{ij} + \frac{1}{N-2} \sum_{m!=i,n!=j,m<n}^{N} D_{mn} \tag{1.11}$$

Izračunom sume S_{ij} za svaki par taksona *i* i *j* i određivanjem one minimalne, dobit ćemo novi par taksona. Time se broj taksona s kojima provodimo algoritam smanjuje za jedan, matricu udaljenosti moramo preračunati dodajući novi čvor (i, j) te istovremeno moramo ukloniti udaljenosti od taksona *i* i *j* prema ostalim taksonima. Udaljenost između novonastalog čvora (i, j) i ostalih taksona računamo prema izrazu:

$$D_{(i\ -\ j)k} = (D_{ik} + D_{jk}) / 2 \tag{1.12}$$

dok ostale udaljenosti ostaju jednake.

Metoda udruživanja susjeda, osim što stvara toplogiju stabla, računa i duljine grana tako da nakon udruživanja dva para taksona, računamo duljinu grana između taksona *i* i *j* do čvora X. Za to se koriste izrazi metode Fitch-Margoliash (Nei, 1987):

$$D_{iX} = \left(D_{ij} + \frac{1}{N-2} \sum_{k!=i,k!=j}^{N} D_{ik} - \frac{1}{N-2} \sum_{k!=i,k!=j}^{N} D_{jk} \right) / 2 \qquad (1.13)$$

$$D_{jX} = \left(D_{ij} + \frac{1}{N-2} \sum_{k!=i,k!=j}^{N} D_{jk} - \frac{1}{N-2} \sum_{k!=i,k!=j}^{N} D_{ik} \right) / 2 \qquad (1.14)$$

U slučaju da se računa duljina grane između dva unutrašnja čvora, na primjer da u idućem koraku spojimo čvor (i, j) s nekim taksonom *m* u čvor $(i-j, m)$, izraz za računanje udaljenosti između (i, j) i $(i-j, m)$ se pretvara u:

$$D_{(i-j)(i-j,m)} = \left(D_{(i-j)m} + \frac{1}{N-2} \sum_{k!=i,k!=j}^{N} D_{(i-j)k} - \frac{1}{N-2} \sum_{k!=i,k!=j}^{N} D_{mk} \right) / 2 - D_{ij} / 2 \qquad (1.15)$$

Postupak uparivanja taksona ili čvorova stvorenih uparivanjem taksona se nastavlja sve dok broj taksona ne padne na dva, odnosno, gore navedeni postupak se ponavlja sve dok nam ne ostanu dva moguća taksona koja možemo spojiti. Tada izračunamo procjenjenu duljinu grane između njih izrazom (1.15) i metoda udruživanjem susjeda tu završava.

3.2 Pseudokod metode udruživanja susjeda

```
matrica_udaljenosti = učitaj();
taksoni = učitaj();
broj_taksona = taksoni.veličina();
matrica_suma[][] = 0;

dok (broj_taksona > 2) {

     za svaki par taksona (i,j) {
      matrica_suma[i][j] = izračunaj_sumu_grana_stabla(i,j);
     }

     iMin,jMin=odredi_minimalnu_sumu(matrica_suma);

     n = stvori_novi_čvor(iMin,jMin);

     izračunaj_duljinu_grane(iMin, n);
```

```
      izračunaj_duljinu_grane(jMin, n);

      makni_iz_liste_taksona(iMin);
      makni_iz_liste_taksona(jMin);
      dodaj_u_listu_taksona(n);
      broj_taksona--;

      matrica_udaljenosti = preračunaj_udaljenosti();

}

izračunaj_duljinu_grane(taksoni);
```

Nakon učitavanja matrice udaljenosti i taksona slijedi izgradnja stabla. Dok broj čvorova, odnosno taksona, ne padne na 2, za svaka dva para taksona izračunamo sumu grana stabla prema izrazu (1.11). Svaku tu vrijednost spremimo u matricu suma, gdje indeksi *i* i *j* označavaju na koje čvorove se odnosi suma stabla. Kada se sve sume izračunaju, traži se ona najmanja i taksoni koji određuju tu sumu. Oni su predstavljeni varijablama *iMin* i *jMin*. Pomoću njih stvorimo novi čvor koji će biti čvor-roditelj tim taksonima te izračunamo duljinu grana između tog čvora i taksona. Ako su *iMin* i *jMin* izvorni taksoni za izračun duljine koristimo izraz (1.13) ili izraz (1.14), a ako je jedan od *iMin* i *jMin* stvoreni unutarnji čvor, koristimo izraz (1.15). Novostvoreni čvor dodajemo u listu taksona i na njega gledamo kao na novi takson, dok taksone *iMin* i *jMin* mičemo iz liste.

Kada vrijednost liste taksona padne na dva, izlazimo iz petlje i izračunavamo posljednju duljinu grane između dva preostala čvora pomoću izraza (1.13), (1.14) ili (1.15) ovisno o tome jesu li u listi ostali ukomponirani čvorovi ili orginalni taksoni.

3.3 Primjer rekonstrukcije filogenetskog stabla

U ovom odjeljku ćemo rekonstruirati filogenetsko stablo metodom udruživanja susjeda, za N = 5 taksona. Promotrimo matricu udaljenosti danu u tablici 1.
Iz matrice udaljenosti danoj u tablici 1. računamo matricu suma prema izrazu (1.11). Matrica suma je dana u tablici 2.

Tablica 1: Matrica udaljenosti za N=5

taksoni	*1*	*2*	*3*	*4*	*5*
1	*0*	*7*	*10*	*6*	*4*
2	*7*	*0*	*5*	*8*	*11*
3	*10*	*5*	*0*	*4*	*6*
4	*6*	*8*	*4*	*0*	*9*
5	*4*	*11*	*6*	*9*	*0*

Tablica 2: Matrica suma za N = 5

taksoni	*1*	*2*	*3*	*4*	*5*
1	*-*	*17.17*	*19.67*	*17.33*	*15.83*
2	*17.17*	*-*	*16.50*	*17.67*	*18.67*
3	*19.67*	*16.50*	*-*	*16.67*	*17.17*
4	*17.33*	*17.67*	*16.67*	*-*	*18.33*
5	*15.83*	*18.67*	*17.17*	*18.33*	*-*

Iz izračunatog vidimo da se u prvom koraku trebaju spojiti taksoni 1 i 5 pošto je u njihovom slučaju ukupna suma grana stabla minimalna. Kada se ta dva taksona spoje, preračunava se matrica distanci i smanjuje se broj taksona za 1. Novu matricu udaljenosti možemo vidjeti u tablici 3.

Tablica 3: Matrica udaljenosti za N = 4

taksoni	*2*	*3*	*4*	*(1,5)*
2	*-*	*5*	*8*	*9*
3	*5*	*-*	*4*	*5*
4	*8*	*4*	*-*	*7.5*
(1,5)	*9*	*8*	*7.5*	*-*

Primjetimo kako su se promjenile udaljenosti jedino između čvora (1,5) i ostalih taksona. Još nam je preostalo izračunati duljinu grane između taksona 1 i čvora (1,5), kao i između taksona 5 i čvora *(1,5).* Pošto će oba taksona biti listovi stabla, za njih koristimo izraze (1.13) i (1.14) te dobivamo $D_{1(1-5)} = 1.5$ i $D_{5(1-5)} = 2.5$. U idućem koraku ponovno računamo matricu suma iz matrice udaljenosti, određujemo par koji ćemo povezati, preračunavamo matricu udaljenosti i na kraju računamo duljinu grana. Matrica suma je dana u tablici 4, a preračunata matrica udaljenosti u tablici 5. Primjetimo kako je minimalna suma grana stabla jednaka za taksone 2 i 3, te taksone 4 i (1,5). U ovakvom slučaju možemo izabrati bilo koji od ta dva para, pa izabiremo 2 i 3.

Tablica 4: Matrica suma za N = 4

taksoni	***2***	***3***	***4***	***(1,5)***
2	*-*	*13.50*	*14.38*	*13.63*
3	*13.50*	*-*	*13.63*	*14.38*
4	*14.38*	*13.63*	*-*	*13.50*
(1,5)	*13.63*	*14.38*	*13.50*	*-*

Tablica 5: Matrica udaljenosti za N=3

taksoni	***(2,3)***	***4***	***(1,5)***
(2,3)	*-*	*7.5*	*6*
4	*7.5*	*-*	*8.5*
(1,5)	*6*	*8.5*	*-*

Za udaljenosti između taksona 2 i čvora (2,3) i taksona 3 i čvora (2,3) ponovno koristimo izraze (1.13) i (1.14) te dobivamo vrijednosti $D_{2(2-3)} = 3.75$ i $D_{3(2-3)} = 1.25$. U sljedećem koraku ponovno ponavljamo prethodne postupke. Matrica suma je dana u tablici 6, a matrica udaljenosti u tablici 7. Primjetimo ponovno kako su sume svugdje jednake, pa možemo birati bilo koji par taksona, a

u ovom slučaju biramo 4 i (1,5). Duljinu grane između taksona 4 i čvora (4-1,5) računamo prema starom izrazu (1.13), međutim duljinu grane između taksona (1,5) i čvora (4-1,5) računamo prema izrazu (1.15), pošto je takson (1,5) unutrašnji čvor stabla, a ne list stabla. Udaljenosti koje dobijemo su sljedeće: $D_{4(4-1,5)} = 2.5$ i $D_{(1,5)(4-1,5)} = 3$.

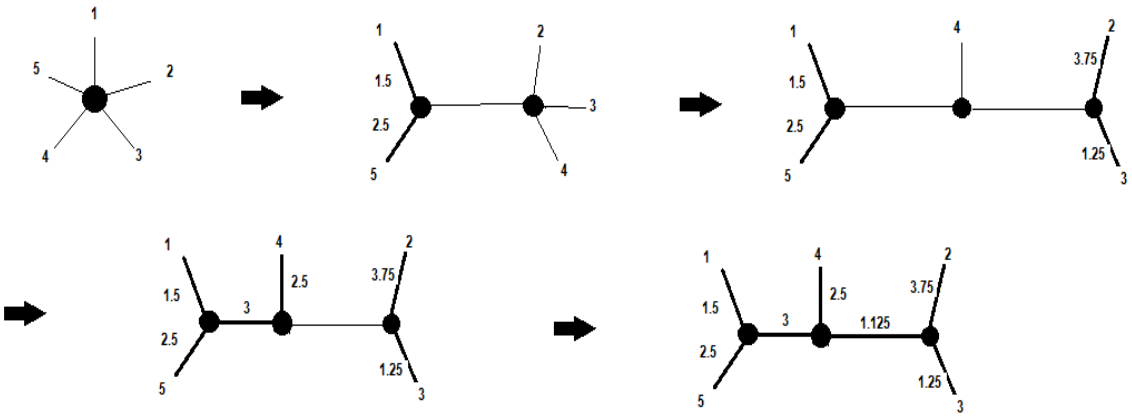
Tablica 6: Matrica suma za N = 3

taksoni	***(2,3)***	***4***	***(1,5)***
(2,3)	-	*11*	*11*
4	*11*	-	*11*
(1,5)	*11*	*11*	-

Tablica 7: Matrica udaljenosti za N = 2

taksoni	***(2,3)***	***(4-1,5)***
(2,3)	-	*7.25*
(4-1,5)	*7.25*	-

Broj taksona je pao na N = 2, te se rekonstrukcija stabla zaustavlja. Još je potrebno izračunati udaljenost između čvorova (2,3) i (4-1,5) pomoću formule (1.15). Udaljenost koju dobijemo je: $D_{(2,3)(4-1,5)} = 1.125$. Potpuni postupak rekonstrukcije filogenetskog stabla možemo vidjeti na Slici 3.3.1.



Slika 3.3.1: Rekonstrukcija filogenetskog stabla

4. Implementacija

U nastavku je opisana implementacija metode udruživanja susjeda prema algoritmu iz poglavlja 3.2.. Opis implementacije se sastoji od rada algoritma, dijagrama razreda, opisa metoda i vremenskog i memorijskog zauzeća

Cijela metoda je napisana u C++ prema C++11 standardu. Implementacija je napravljena po objektno-orijentiranoj paradigmi i izvođena na Intel CORE i5 vPro procesoru. Osim same implementacije, program nudi i poravnavanje zadanih sekvenci u slučaju da iste nisu poravnate. Poravnavanje sekvenci je nužno za određivanje matrice udaljenosti. Za poravnate biološke sljedove koristi se Levenshteinova udaljenost pri određivanju matrice udaljenosti. Levenshteinova udaljenost zadovoljava uvjete iz izraza (1.1), stoga se može koristiti kao mjera udaljenosti (vidjeti Dodatak B za pojašnjenje računanja udaljenosti).

4.1 Algoritam implementacije metode udruživanja susjeda

1. Ulaz u program je jedna FASTA datoteka (vidjeti Dodatak A za pojašnjenje formata).
2. Zadana FASTA datoteka se parsira kako bi se dobile sekvence za obradu.
3. Određuje se matrica udaljenosti pomoću Levenshteinove mjere. U slučaju da biološki sljedovi u FASTA datoteci nisu poravnati, prije samog računanja matrice udaljenosti sljedovi se poravnavaju uz pomoć Smith-Waterman algoritma (vidjeti Dodatak B za opis algoritma).
4. Provodi se metoda udruživanja susjeda.
5. Izlaz iz programa je datoteka u NEWICK formatu (vidjeti Dodatak A za pojašnjenje formata).
6. Pseudokod programa za rekonstrukciju filogenetskog stabla metodom udruživanja susjeda je sljedeći:

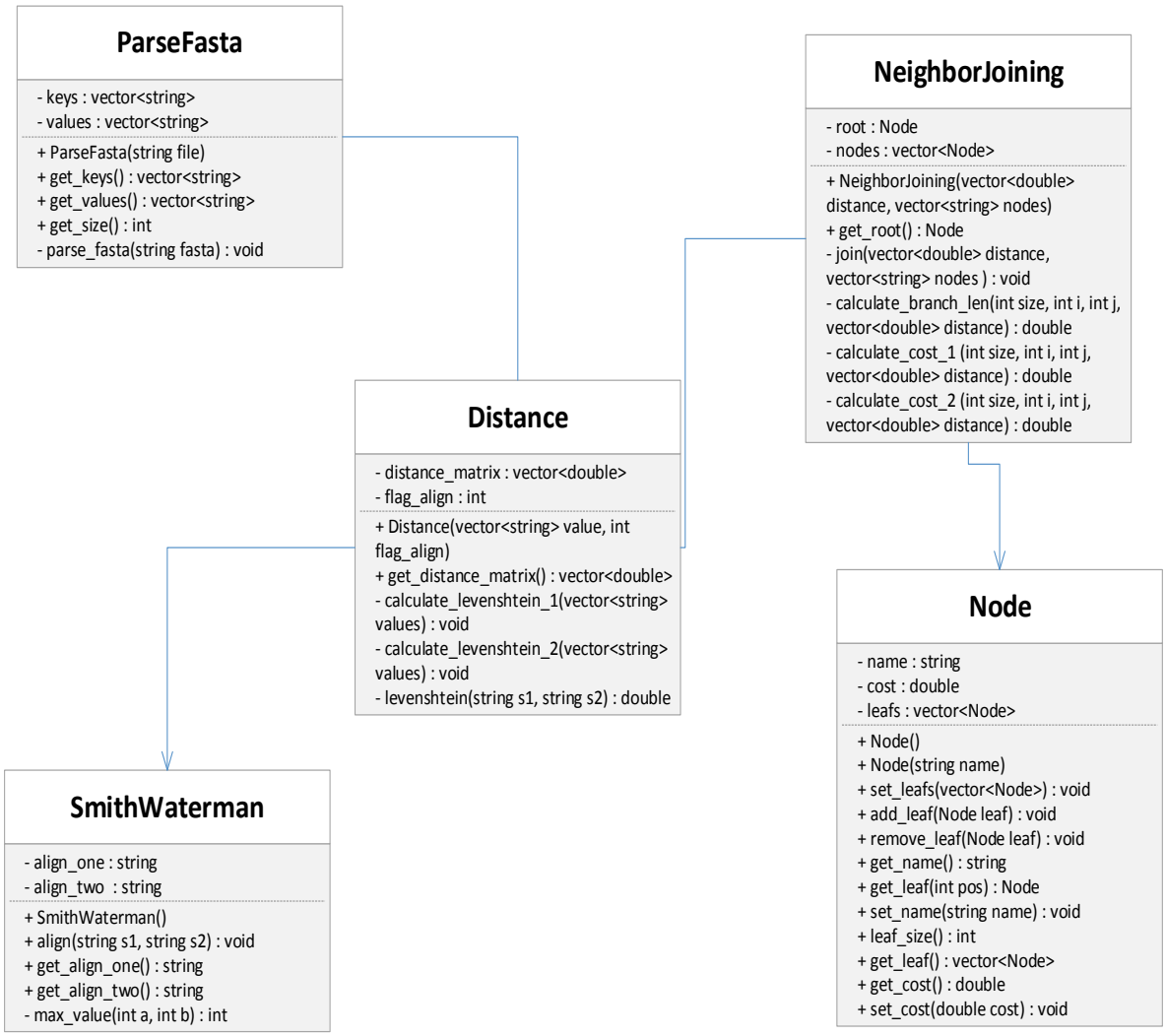
```
file, zastavica = parsiraj_ulazne_podatke();
fatsa_file = učitaj_FASTA _file();
ako (!fasta_file) {
     dojavi_grešku();
     prekini_izvođenje();
}
sekvence = parsiraj(fasta_file);

ako (zastavica) {
     poravnaj_sekvence_i_odredi_matricu_udaljenosti();
} inače {
     odredi_matricu_udaljenosti();
}
rekonstrukcija_stabla_NJM ();
stvori_nwk_file();
izlaz();
```

4.2 Dijagram razreda i opis metoda

Na idućoj strani, na slici 4.2.1 je dan UML dijagram razreda.

Kao što je rečeno u poglavlju 4.1, funkcija main učitava jednu datoteku FASTA formata i poziva konstruktor klase ParseFasta, koji poziva metodu *parse_fasta(string fasta)*. Metoda *parse_fasta(string fasta)* provjerava valjanost dane FASTA datoteke i završava program ukoliko se datoteka ne može otvoriti. Ako je datoteka ispravna, *parse_fasta* mijenja vrijednosti varijabli *keys* i *values.* U varijabli *keys* se čuvaju imena sekvenci iz FASTA datoteke, a u varijabli *values* same sekvence. Nakon parsiranja, metoda *main* prosljeđuje konstruktoru klase *Distance* listu sekvenci za koje treba odrediti matricu udaljenosti i zastavicu *flag_align* koja je znak klasi Distance je li potrebno obaviti poravnanje nad sekvencama ili su one već poravnate. U slučaju da je vrijednost zastavice *flag_align = 1*, poravnavanje se obavlja Smith-Watermanovim algoritmom,



Slika 4.2.1: Dijagram razreda

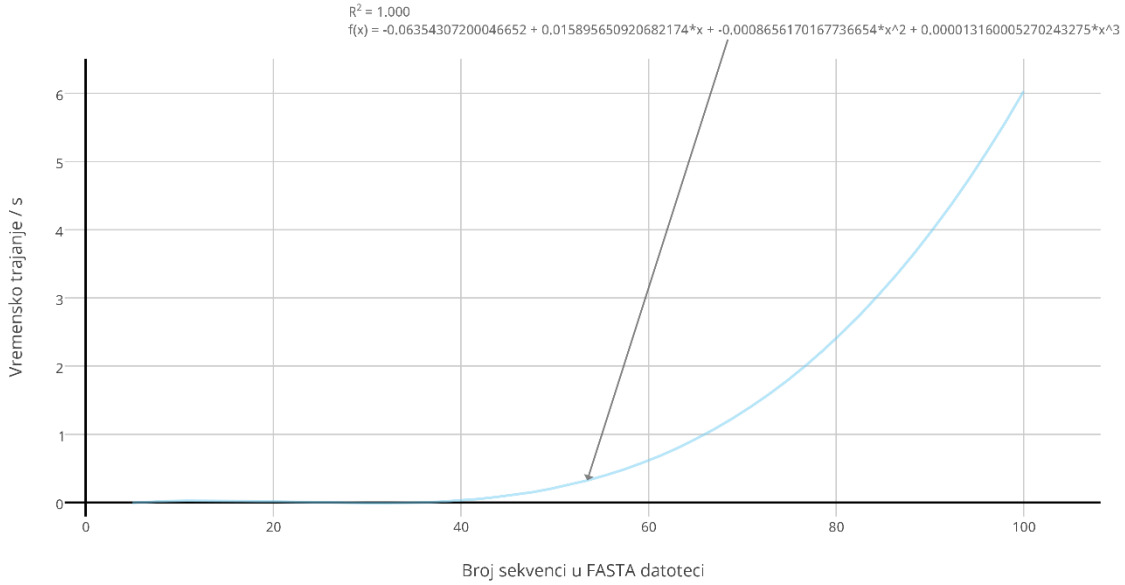
odnosno za računanje udaljenosti između sekvenci poziva se metoda *calculate_levenshtein_2* koja poziva metodu *align* iz SmithWaterman klase. Ako su pak sekvence već poravnate, poziva se metoda *calculate_levenshtein_1* koja samo računa Levenshtein udaljenost između sekvenci.

Nakon izračunavanje matrice udaljenosti, metoda *main* poziva klasu NeighborJoining, čiji konstruktor poziva metodu *join* koja obavlja rekonstrukciju filogenetskog stabla metodom udruživanja susjeda. Metoda *join* poziva tri metode koje predstavljaju prethodno izvedene izraze za računanje sume grana stabla i računanje duljine grana između čvorova. Metoda *calculate_branch_len* računa ukupnu sumu grana stabla prema izrazu (1.11). Računanje duljine grane stabla između listova stabla i unutrašnjih čvorova ili samo unutrašnjih čvorova prema

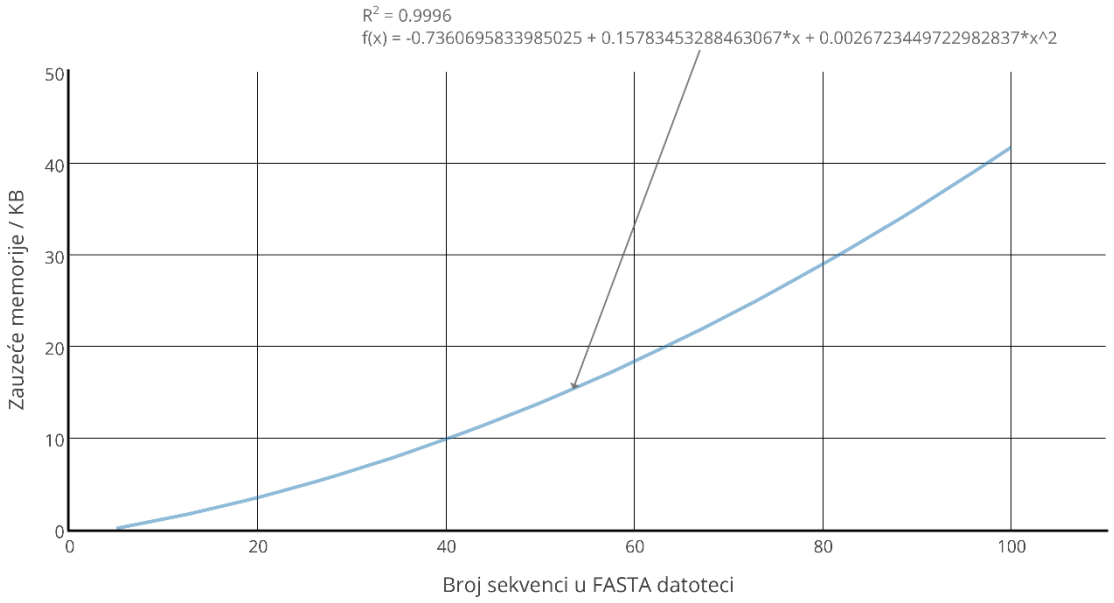
izrazima (1.13), (1.14) i (1.15) implementiraju metode *calculate_cost_1* i *calculate_cost_2.*

Nakon što je filogenetsko stablo rekonstruirano, metoda main ispisuje rezultat u datoteku NEWICK formata i time program završava.

4.3 Vremensko i memorijsko zauzeće



Slika 4.3.1: Prikaz trajanja izvođenja programa u ovisnosti o količini sekvenci u FASTA datoteci



Slika 4.3.2: Prikaz zauzeća memorije u ovisnosti o količini sekvenci u FASTA datoteci

Ako označimo broj bioloških sljedova u datoteci s n, implementacija metode udruživanja susjeda ima vremensku složenost $O(n^3)$, dok joj je memorijska složenost $O(n^2)$. Algoritam je pokrenut na 5 različitih FASTA datoteka. Sve FASTA datoteke sadrže već poravnate sekvence, tako da se testira samo izvođenje metode udruživanja susjeda. Na slikama 4.3.1 i 4.3.2 su prikazane ovisnosti veličine FASTA datoteke i trajanja algoritma, odnosno zauzeća memorije.

Vremenskoj složenosti najviše pridonosi računanje elemenata matrice, čija je vremenska složenost $O(n^2)$. Računanje elemenata matrice se odvija unutar petlje koja se ponavlja $n-2$ puta, iz čega proizlazi eksponencijalna vremenska složenost $O(n^3)$.

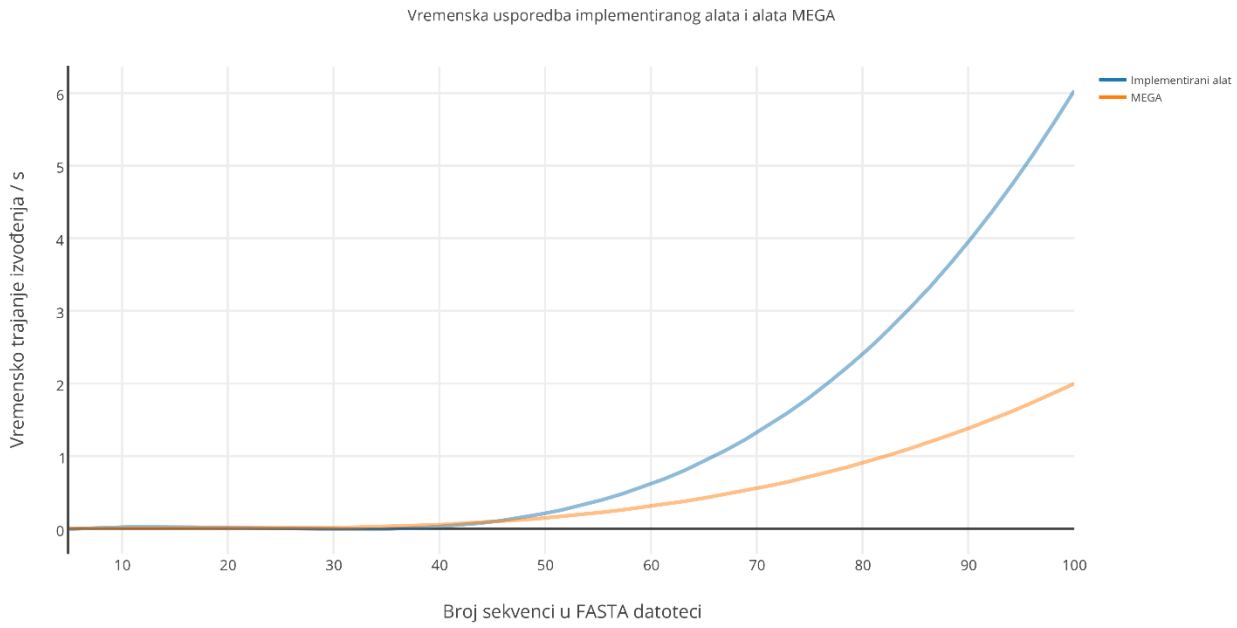
Memorijska složenost implementacije se sastoji u čuvanju elemenata matrice veličine $n \times n$ te čuvanju čvorova stabla čija se veličina povećava dodavanjem unutarnjih čvorova. Sveukupno memorijska složenost iznosi $2n^2$, odnosno $O(n^2)$.

5. Diskusija i usporedba sa drugim alatima

Metoda udruživanja susjeda je jedna od najbržih metoda za rekonstrukciju filogenetskih stabala ([2],[3]). To je čini izuzetno pogodnom za rekonstrukciju filogenetskih stabala iz puno bioloških sljedova. Također, ova metoda računa duljine grana stabla što nam daje podatke o međusobnim udaljenostima organizama koji se promatraju. Međutim, od svih mogućih topologija stabala, metoda udruživanja susjeda razmatra samo jednu toplogiju od nekoliko mogućih, što može dovesti do rekonstrukcije različitih topologija stabala korištenjem različitih implementacija metode udruživanja susjeda na istim podacima. Ta razlika se događa u dijelu biranja čvorova za spajanje novim unutarnjim čvorom. Naime, pri računanju sume grana stabla za svaki par *i* i *j* može se dogoditi da više različitih parova imaju istu ukupnu sumu grana stabla. Ako je ta suma ujedno i minimalna suma grana stabla u tom koraku, slobodni smo izabrati bilo koji od tih parova [1]. Pošto se biranjem parova u svakom koraku algoritma stvara topologija stabla, tako će i različiti odabir ovih parova stvoriti različitu toplogiju konačne rekonstrukcije filogenetskog stabla. Ovdje možemo primjetiti moguće poboljšanje metode udruživanja susjeda. Ukoliko u algoritmu naiđemo na slučaj da se više parova može odabrati za spajanje unutarnjim čvorom, mogle bi se stvoriti dodatne topologije filogenetskog stabla koje će se razlikovati upravo prema odabiru parova koji su u jednom od koraka algoritma imali istu ukupnu sumu grana stabla koja je za taj korak bila minimalna.

Metoda udruživanja susjeda u svakom koraku algoritma minimizira sumu grana stabla, međutim konačna rekonstrukcija topologije filogenetskog stabla ne mora imati minimalnu sumu grana stabla u usporedbi s ostalim mogućim topologijama filogenetskog stabla [1]. Treba napomenuti da filogenetsko stablo s najmanjom sumom grana ne mora nužno biti stablo koje odražava realno stanje u evoluciji [1], stoga ovaj mogući nedostatak ne mora dati nužno lošu toplogiju filogenetskog stabla.

Ovu smo implementaciju metode udruživanja susjeda usporedili s alatom MEGA za molekularnu i genetičku analizu. Alat MEGA ima mogućnost analize FASTA datoteka, poravnavanja sekvenci i rekonstrukcije filogenetskog stabla metodom udruživanja susjeda. Na slici 5.1 vidimo vremensko trajanje rekonstrukcije filogenetskog stabla metodom udruživanja susjeda i implementiranog alata u sklopu ovog rada. Usporedba vremenskog trajanja na ova dva alata je provedena uz pet FASTA datoteka različitih veličina.



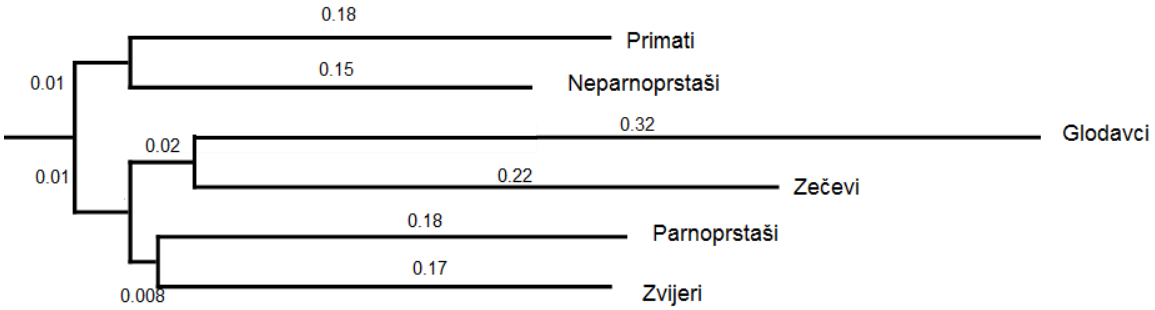
Slika 5.1: Vremenska usporedba implementiranog alata i alata MEGA

Iz slike vidimo da implementirani alat ima vremenski lošije performanse od alata MEGA što se tiče rekonstrukcije filogenetskog stabla metodom udruživanja susjeda. Sama implementacija ima prostora za moguće optimizacije kako bi se ubrzao njezin rad.

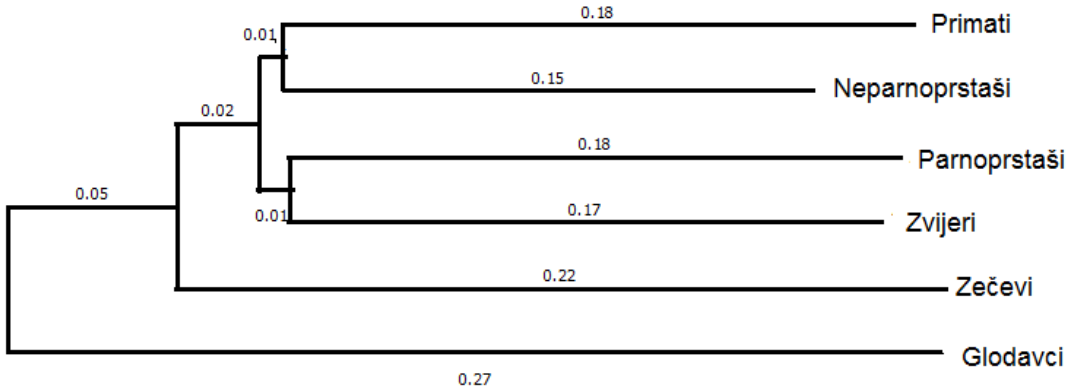
Kako bi testirali biološku ispravnost implementiranog alata usporedili smo ga s alatom MEGA na nekoliko istih skupova podataka. Primjer jedne takve skupine je dan u tablici 9. Tablica 9 sadrži matricu udaljenosti između šest skupina organizama: Glodavci, Primati, Zečevi, Parnoprstaši, Zvijeri i Neparnoprstaši. Rezultati metode udruživanja susjeda su vidljivi na slikama 5.2 i 5.3.

Tablica 8: Matrica udaljenosti iz bioloških podataka

Taksoni:	Glodavci	Primati	Zečevi	Parnoprstaši	Zvijeri	Neparnoprstaši
Glodavci	0	0.514	0.535	0.530	0.521	0.500
Primati	0.514	0	0.436	0.388	0.353	0.331
Zečevi	0.535	0.436	0	0.418	0.417	0.402
Parnoprstaši	0.530	0.388	0.418	0	0.345	0.327
Zvijeri	0.521	0.353	0.417	0.345	0	0.349
Neparnoprstaši	0.500	0.331	0.402	0.327	0.349	0



Slika 5.2: Rekonstrukcija filogenetskog stabla implementirani alatom



Slika 5.3: Rekonstrukcija filogenetskog stabla alatom MEGA

Na slici 5.2 je prikazana rekonstrukcija filogenetskog stabla implementiranim alatom, dok je na slici 5.3 prikazana rekonstrukcija filogenetskog stabla alatom MEGA. Možemo primjetiti kako se toplogije stabla razlikuju u pojedinim grana. Kao što smo prije spomenuli, kada imamo iste ukupne minimalne sume grana stabla za pojedine parove taksona u nekom koraku algoritma, slobodni smo proizvoljno odabrati koji par taksona ćemo spojiti unutarnjim čvorom. Takvo proizvoljno biranje utječe na toplogiju stabla kakvu primjećujemo u ovom primjeru. Valja napomenuti da su obje topologije ispravne, Ono što je bitno jest da evolucijske udaljenosti između organizama budu jednake, što je evidentno iz slika 5.2 i 5.3. Možemo zaključiti kako je implementirani alat ispravno rekonstruirao filogenetsko stablo.

6. Zaključak

Za pronalazak bioloških veza između taksonomskih jedinica filogenija koristi filogenetska stabla. U današnjem su vremenu razvijene metode i algoritmi koji omogućuju rekonstrukciju filogenetskih stabala pomoću računala. U ovom je radu opisana metoda udruživanja susjeda koja pripada u skup metoda rekonstrukcije filogenetskog stabla temeljenih na međusobnoj udaljenosti taksona. Metoda udruživanja susjeda pomoću matrice udaljenosti radi rekonstrukciju filogenetskog stabla te kao rezultat daje jednu moguću topologiju stabla.

U sklopu ovog rada implementiran je jednostavniji alat za rekonstrukciju filogenetskog stabla metodom udruživanja susjeda. Navedeni su i objašnjeni formati ulaznih i izlaznih podataka te rad Smith-Waterman algoritma koji je bio potreban za poravnavanje bioloških sljedova.
Implementirani je alat uspoređen sa postojećim alatom MEGA. Ustanovljeno je da implementirani alat ima slabije performanse. Postoji više mogućnosti optimizacije s kojima bi se poboljšao rad implementiranog alata. Usporedbom s alatom MEGA ustanovljen je ispravan rad implementiranog alata. Dobivene topologije stabala uporabom implementiranog alata odgovaraju biološkoj stvarnosti.

Metode rekonstrukcije filogenetskih stabala su jedno od bitnih područja bioinformatike na kojem se rade mnoga istraživanja i poboljšanja.

7. Literatura

[1] Saitou N. i Nei M., The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees, 1987.

[2] Rastogi S.C., Rastogi P. i Mendiratta N., Bioinformatics methods and applications: Genomics Proteomics and Drug Discovery, 2008.

[3] http://bioinformatics.oxfordjournals.org/content/23/2/e136.full, posjećeno: 23.svibanj 2015.

[4] Janowski T., Distributed Computing and Internet Technology: 6th International Conference, 2010.

[5] Edwards, Stajich i Hansen, Bioinformatics: Tools and Applications, 2009.

[6] Šikić M., Domazet-Lošo M., Bioinformatika, 2013.

8. Sažetak

Za područje filogenije rekonstrukcija filogenetskih stabala je izuzetno bitna. Pomoću njih se određuju biološke veze između taksonomskih jedinica. Za rekonstrukciju filogenetskih stabala pomoću računala su razvijene mnoge metode i algoritmi. U ovom je radu opisana metoda udruživanja susjeda kao jedna od metoda rekonstrukcije filogenetskih stabala temeljenih na međusobnoj udaljenosti taksona te je ostvarena njezina programska implementacija. Usporedbom implementiranog programa s postojećim alatom ustanovljen je ispravan rad implementiranog alata. Topologije stabala dobivene implementiranim alatom odgovaraju evolucijskim odnosima u stvarnosti.

Ključne riječi: filogenetsko stablo, matrica udaljenosti, rekonstrukcija, topologija, metoda udruživanja susjeda

9. Abstract

In phylogeny, reconstruction of phylogenetic trees is extremely important. With them one can determine biological connection between taxa. For computional phylogenetic tree reconstruction there are many methods and algorithms developed. We described neighbor-joining method as one of distance-based methods for phylogenetic reconstruction. We also implemented a tool that uses this method for phylogenetic tree reconstruction. By comparative analysis with already implemented tool we determined soundness of our implementation. Furthermore, we determined that topologies reconstructed with implemented tool reflect evolutionary relationships in reality.

Keywords: phylogenetic tree, distance matrix, reconstruction, topology, neighbor-joining method.

Dodatak A

FASTA format

U bioinformatici FASTA format je jedan od najbitnijih formata za zapisivanje sekvenci nukleotida, slijedova aminokiselina. Uobičajeno se u FASTA datotekama nalaze zapisi proteina, gena ili čak čitavih genoma. Svaki nukleotid je predstavljen jednim slovom abecede-A, T, C, G ili U. Ukupan opis jedne sekvence u FASTA datoteci se sastoji od naziva sekvence koji opisuje što ona predstavlja te samih slijedova nukleotida koji čine sekvencu.
Izgled FASTA datoteke se može vidjeti na Slici A.1. Naziv sekvence uobičajeno počinje znakom '>' te se iza naziva nastavlja niz nukleotida koji predstavlja sekvencu tog naziva.

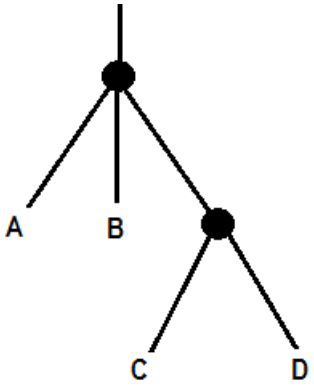
```
>gi|544886608|ref|NZ_AURB01000207.1| Alicyclobacillus acidoterrestris
GTTTTGGGAACGGGTATTGACAGGGAGTTTTGGGTACATGGATTACGCAAGGGCAGAAACACCCAGTTCC
ACATGGACGATATTGCTCTGACCGTCATTTTCGGTTCCTTGCTGGGTCAGGAACGGATTTTCCACTTTGA
GGACATCGAACAAGATCCCCTGTTGAAGCTGAAGTTGGACGTGCCGAAACTGCCTGATACGACTCTGTTG
>gi|220683588|gb|FJ158840.1| Jeotgalicoccus huakuii
AGAGTTTGATCCTGGCTCAGGATGAACGCTGGCGGCGTGCCTAATACATGCAAGTCGAGCGCGAGCGTAA
GGAGCTTGCTCCTTACAATCGAGCGGCGGACGGGTGAGTAACACGTGGGCAACCTACCCTTTAGACTGGG
ATAACTACCGGAAACGGTAGCTAATACCGGATAAGTTGGATTACACAAGTAATCTTAATGAAAGGCGGAT
TTATCTGTCACTAAAGGATGGGCCTGCGGTGCATTAGCTAGTTGGTGAGGTAGTGGCTCACCAAGGCAAC
```

Slika A.1 : primjer FASTA formata

Newick format

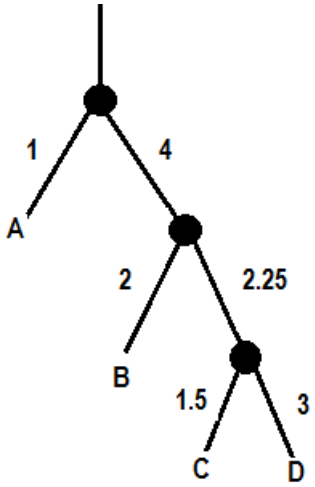
Newickov format je raširenI format za zapis grafova(u smislu mreža) i stabala. Pomoću Newickovog formata je moguće zapisati odnose između čvorova stabala i dodatno definirati udaljenosti od roditeljskih čvorova ako je potrebno. Ovaj format

se tvori pomoću zagrada i zareza, a njegov izgled je prikazan na Slici A.2 i Slici A.3 kao i odgovarajuća stabla.



Zapis u Newick formatu: (A, B,(C,D));

Slika A.2: crtež i zapis stabla u Newick formatu



Zapis u Newick formatu: (A:1,(B:2,(C:1.5,D:3):2.25):4);

Slika A.2: crtež i zapis stabla u Newick formatu s definiranim udaljenostima

Dodatak B

Levenshteinova mjera udaljenosti

Levenshteinova mjera udaljenosti izražava udaljenost između dva biološka sljeda. Ako imamo niz *s1* i niz *s2,* Levenshteinova će udaljenost između njih biti minimalan broj potrebnih modifikacija kako bi se jedan niz pretvorio u drugi. Pod pojmom modifikacije se smatra zamjena, umetanje ili brisanje jednog znaka.

Smith-Waterman algoritam

Smith-Waterman algoritam služi za optimalno lokalno poravnanje sekvenci, odnosno koristi se kada sekvence dijele izolirane regije sličnosti među sobom. Danas se ovaj algoritam vrlo često koristi za uspoređivanje bioloških sljedova te je izuzetno bitan u bioinformatici.

Kako bi poravnao dvije sekvence, SWA prvo stvara matricu veličine $m \times n$ pri čemu je *m* duljina prve sekvence, a *n* duljina druge sekvence. Znakovi prve sekvence će predstavljati redove matrice, dok će znakovi druge sekvence predstavljati stupce matrice. Član matrice $M(i, j)$ se popunjava rezultatom poravnanja sekvenci na području $(0, i)$ za prvu sekvencu, odnosno $(0, j)$ za drugu sekvencu. Matematički se računanje člana matrice može prikazati izrazom (B.1):

$$M(i, j) = \begin{cases} 0 & \mathrm{i} = 0 \vee \mathrm{j} = 0 \\ \max \begin{cases} M(i\text{-}1, j\text{-}1) + s(i,j) \\ M(i-1, j) + p \\ M(i, j\text{-}1) + p \\ 0 \end{cases} & ostalo \end{cases} \qquad \text{(B.1)}$$

U izrazu (B.1) $s(i, j)$ označava funkciju sličnosti između znakova sekvenci na indeksima *i* i *j,* a *p* označava kaznu za umetanje ili brisanje jednog ili više znakova u sekvencama.

Nakon popunjavanja matrice, pronalazi se maksimalna vrijednost u matrici te se rekonstruira put poravnanja praćenjem unatrag dok se ne dosegne polje u matrici čija je vrijednost nula.