

SVEUČILIŠTE U ZAGREBU
FAKULTET ELEKTROTEHNIKE I RAČUNARSTVA

ZAVRŠNI RAD br. 4994

**IZGRADNJA FILOGENETSKOG STABLA
KORIŠTENJEM METODA UDALJENOSTI**

Nikola Gajski

Zagreb, lipanj 2017.

Zagreb, 8. ožujka 2017.

ZAVRŠNI ZADATAK br. 4994

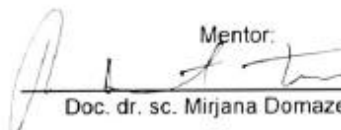
Pristupnik: **Nikola Gajski (0036483903)**
Studij: **Računarstvo**
Modul: **Programsko inženjerstvo i informacijski sustavi**

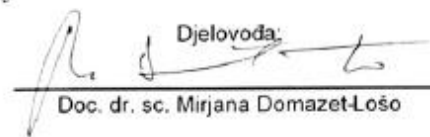
Zadatak: **Izgradnja filogenetskog stabla korištenjem metoda udaljenosti**


Opis zadatka:

U okviru ovoga završnog rada potrebno je proučiti metode za izgradnju filogenetskog stabla temeljem metoda udaljenosti: metodu UPGMA i metodu povezivanja susjeda. Potrebno je napraviti implementaciju tih algoritama i vizualizaciju izgradnje filogenetskog stabla u programskom jeziku Java. Potrebno je analizirati vrijeme izvođenja i memorijske zahtjeve implementacije te usporediti s postojećim programskim rješenjima.

Zadatak uručen pristupniku: 10. ožujka 2017.
Rok za predaju rada: 9. lipnja 2017.

Mentor:

Doc. dr. sc. Mirjana Domazet-Lošo

Djelovoda:

Doc. dr. sc. Mirjana Domazet-Lošo

Predsjednik odbora za
završni rad modula:

Doc. dr. sc. Ivica Botički

Zahvala

Zahvaljujem se mentorici doc. dr. sc. Mirjani Domazet-Lošo na ukazanom vremenu i podršci.

Sadržaj

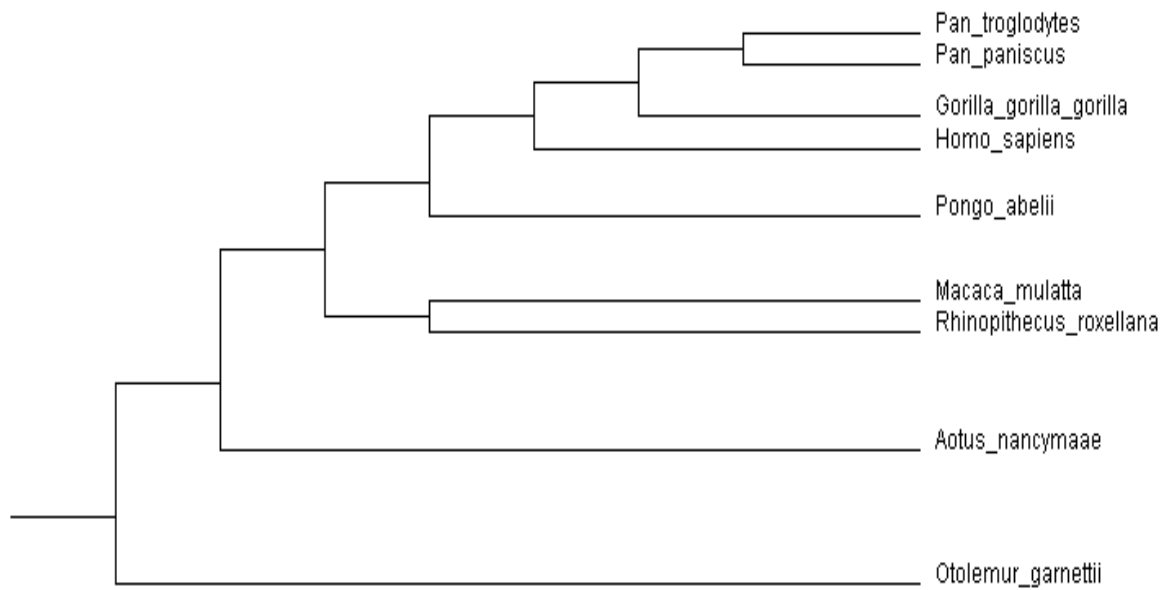
1. Uvod.....	1
2. Povijesni razvoj ideja	3
2.1. <i>Binarna nomenklatura ili dvoimeno nazivlje.....</i>	<i>3</i>
2.2. <i>Evolucijske osnove.....</i>	<i>4</i>
3. Definicija problema	5
3.1. <i>Ukorijenjeno filogenetsko stablo</i>	<i>5</i>
3.2. <i>Neukorijenjeno filogenetsko stablo</i>	<i>6</i>
3.3. <i>Pretvorba neukorijenjenog u ukorijenjeno stablo.....</i>	<i>7</i>
3.3.1. <i>Dodavanje korijena na pola puta između dva najudaljenija lista u stablu</i>	<i>7</i>
3.3.2. <i>Dodavanje najudaljenijeg taksona kao taksona koji se veže direktno na korijen</i>	<i>8</i>
3.4. <i>Matrica udaljenosti</i>	<i>9</i>
3.5. <i>Broj mogućih filogenetskih stabala u ovisnosti o broju taksona</i>	<i>11</i>
3.6. <i>Udaljenost sljedova.....</i>	<i>12</i>
3.6.1. <i>Jukes – Cantorov evolucijski model</i>	<i>12</i>
4. UPGMA	16
4.1. <i>Algoritam.....</i>	<i>16</i>
4.2. <i>Primjer.....</i>	<i>17</i>
5. Metoda povezivanja susjeda	21
5.1. <i>Algoritam.....</i>	<i>21</i>
5.2. <i>Primjer.....</i>	<i>22</i>
6. Analiza programske potpore	26
7. Usporedba programske potpore sa MEGA7.....	32
8. Zaključak.....	42
9. Literatura	43

1. Uvod

Bioinformatika je znanost koja se bavi istraživanjem i procesuiranjem genetskih informacija pomoću računalne tehnologije i statističkih metoda. Ubrzano se razvija zadnjih dvadesetak godina, a velika dostupnost tehnologija sekvenciranja rezultirala je stvaranjem velikih skupova bioloških podataka. Količina tih podataka potaknula je razvoj novih tehnologija i računalnih metoda koje bi omogućile analizu u što je moguće kraćem vremenu i najvećom mogućom točnošću. Dvije od tih metoda biti će opisane u daljnjem radu.

Filogenija (Haeckel E., 1866) je znanost koja proučava proces nastanka određene sistematske kategorije organizama, a bazirana je na sličnostima i razlikama u njihovim morfološkim i molekularnim podacima. U osnovi, filogenija se temelji na pretpostavci o zajedničkom pretku svih organizama na Zemlji. Izvorno je zasnovana na spoznaji da su različite vrste biljaka i životinja potekle od zajedničkih predaka. Međutim za veliku većinu vrsta i dan danas nije odgovoreno na pitanja od kuda i kada su egzistirale, koliko su trajale, kako i kada su izumrle, a relativno mali broj njihovih ostataka sačuvan je u obliku fosila. Iz navedenog razloga veliki dio filogenije počiva na hipotezama, koje se baziraju na indirektnim dokazima. Uvođenjem molekularne genetike na bazi DNK analize i napredovanjem filogenetskih metoda učvršćena je opća suglasnost da je filogenetsko stablo rezultat zajedničkog porijekla organskog svijeta, što će biti i vidljivo u ovom tekstu.

Filogenetsko stablo (slika 1) ili evolucijsko stablo je razgranati dijagram koji prikazuje evolucijske odnose između različitih vrsta ili rjeđe, nekih drugih sistematskih kategorija.



Slika 1. Filogenetsko stablo

Filogeneza je proces biološke evolucije živih bića kroz Zemljinu povijest. Pojam nije ograničen samo na evoluciju životinjskih stabala, nego uključuje i razvoj pojedinih taksonomskih jedinica, na svim razinama sistematike. Taksonomske jedinice su poznate skupine živih bića, a osnovna jedinica rasprostranjenosti je vrsta. Prigodna obrada osigurava prikaz rasprostranjenosti svih podvrsta jedne vrste na istom stablu, ili ako je to potrebno, viših taksonomskih kategorija. Istraživanje filogeneza se provodi:

- vrednovanjem morfoloških i anatomskih osobina fosila
- uspoređivanjem morfoloških, anatomskih i fizioloških osobina živih bića
- analizom DNK, naročito pojedinih segmenata DNK i molekularno filogenetskim metodama

2. Povijesni razvoj ideja

2.1. Binarna nomenklatura ili dvoimeno nazivlje

Za kreaciju sustava botaničke i zoološke nomenklature zaslužan je švedski botaničar i liječnik Carl Linné (1707–1778) koji je u želji da opiše čitav poznati živi svijet, svakoj vrsti, mineralu, biljci ili životinji, pripisao dva latinska ili starogrčka imena, što se zove binarna nomenklatura. Dvojno nazivlje u različitim oblicima postojalo je i prije, međutim tek se pojavom Linnéa raširilo. Danas u biologiji označava službenu metodu imenovanja vrsta:

1. Znanstveno nazivlje se ispisuje u kurzivu, npr. *Homo sapiens* (čovjek).
2. Prva riječ (ime roda) se prema aktualnom hrvatskom pravopisu i međunarodnom dogovoru uvijek piše velikim, dok se druga riječ (ime vrste) uvijek piše malim početnim slovom, npr. *Homo sapiens*
3. Znanstveni naziv se pri prvoj upotrebi u tekstu ili pri popisivanju više vrsta istoga roda treba pisati u punom obliku. U daljnjem tekstu može se kratiti korištenjem početnog slova s točkom umjesto imena roda npr. *Homo sapiens* će biti *H. sapiens*. Iznimke su slučajevi kada je skraćeni naziv prerastao u opću upotrebu; npr. bakterija *Escherichia coli* se najčešće navodi kao *E. coli*, a *Tyrannosaurus rex* je poznatiji po nazivu *T. rex*.
4. Dvostrukom se imenu dodaje i puno prezime znanstvenika koji je biljku ili životinju prvi opisao, te godina kada ju je opisao, npr. *Mus musculus* Linné 1758.
5. Pravo autorstva: Ako vrsta ima više imena, pripada joj prvo i najstarije ime koje joj je dao neki stručnjak.

2.2. Evolucijske osnove

1859. godine Charles Darwin (1809 - 1882) objavljuje čuvenu knjigu „O podrijetlu vrsta“ (puni naziv „O postanku vrsta putem prirodnog odabira ili očuvanje boljih pasmina u borbi za opstanak“), najvažniju knjigu iz područja biologije ikad napisanu. U njoj je iznio teoriju da živa bića nisu novostvorena, već su evoluirala od predaka procesom prirodne selekcije.

U prirodi svaka vrsta stvori više jedinki nego će ih se održati održati na životu. Iz tog razloga dolazi do borbe za opstanak u kojoj preživljavaju jedinke bolje prilagođene uvjetima okoline, te se dalje razmnožavaju. Nasljeđivanjem njihovih gena, novi naraštaji bolje su prilagođeni okolini, te se tako izdvajaju vrste koje se sve više razlikuju od zajedničkog pretka. Ova teorija je u čast Darwinu nazvana darvinizam.

“One general law, leading to the advancement of all organic beings, namely, multiply, vary, let the strongest live and the weakest die.” (Darwin C., The Origin of Species, 1859)

Iako je Charles Darwin bio začetnik teorije evolucije, njemački znanstvenik Ernst Haeckel je prvi uveo pojam filogenije.

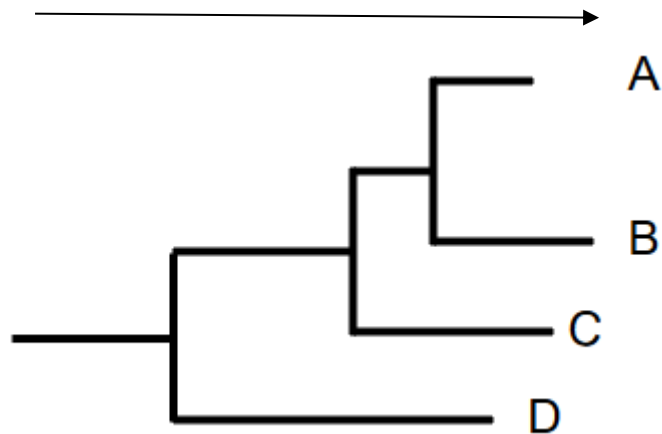
3. Definicija problema

Filogenetsko stablo prikazuje evolucijske odnose između različitih organizama za koje se pretpostavlja da potječu od zajedničkog pretka. Čvorovi u stablu predstavljaju taksonomske jedinice (vrsta, rod, red itd.). Filogenetska stabla danas se grade uglavnom na temelju analize sljedova pojedinih vrsta. Sljedovi mogu biti skupovi gena ili cijeli genom i u tom slučaju se radi o genskim stablima, ili pak skupovi proteina i u tada se radi o proteinskom stablu. Analizom slijeda izračunavaju se dužine pojedinih segmenata istog gena određene vrste te se koriste njihove sličnosti i razlike za izgradnju stabla. Vrste čiji su sljedovi sličniji, na stablu se nalaze međusobno bliže od onih, čiji se sljedovi jako razlikuju. Kako kompleksnost izračuna takvih stabala s brojem i duljinom sljedova eksponencijalno raste, metode za izgradnju istih stalno se unaprjeđuju, te se stvaraju nove, efikasnije i brže.

U današnje vrijeme poznato je da se svi organizmi, pa tako i njihovi geni, nisu razvijali ravnomjerno. Iz tog razloga cilj izgradnje filogenetskog stabla je što detaljnije objašnjenje tijeka evolucije. Primjerice, neki geni koje posjeduje svaki čovjek imaju zajedničkog pretka samo sa čimpanzama, dok se neki drugi pojavljuju kod svih sisavaca. Posljedica toga je da analiza različitih skupova gena iste vrste rezultira različitim filogenetskim stablima, od kojih je svako korektno. Kako bi se utvrdio pravi razvojni slijed kao i grananja u evoluciji pojedinih organizama, neophodno je ispitivanje različitih sljedova gena.

3.1. Ukorijenjeno filogenetsko stablo

Ukorijenjeno filogenetsko stablo predstavlja usmjereni graf koji određuje smjer evolucije, a duljina grana predstavlja relativni broj evolucijskih promjena u jedinici vremena.

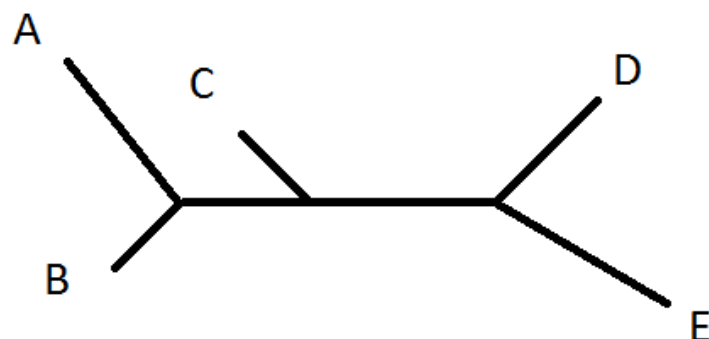


Slika 2. Ukorijenjeno filogenetsko stablo

Na slici 2 prikazano je ukorijenjeno stablo. Stablo sadrži četiri lista koji predstavljaju četiri taksona za sljedove A, B, C i D. Listovi stabla predstavljaju najmlađe potomke, dok čvorovi predstavljaju pretke. Strelica usmjerena od korijena prema listovima prikazuje smjer vremena (vrijeme diferencijacije), odnosno smjer evolucije. Iz stabla se može zaključiti da su sljedovi A i B u bližem srodstvu nego A i C ili B i C. Također se može vidjeti da je slijed C srodniji sa sljedovima A ili B nego što je sa slijedom D. Sva četiri slijeda A, B, C i D imaju zajedničkog pretka, a to je korijen stabla.

3.2. Neukorijenjeno filogenetsko stablo

Neukorijenjeno filogenetsko stablo, za razliku od ukorijenjenog, nema poznati smjer evolucije te ne sadrži korijen.



Slika 3. Neukorijenjeno filogenetsko stablo

Na slici 3 prikazano je neukorijenjeno filogenetsko stablo sa 5 listova koji predstavljaju taksone za sljedove A, B, C, D i E. Iz stabla se može zaključiti da su sljedovi A i B u bližem srodstvu nego A i C ili B i C. Također se može vidjeti da je slijed C srodniji sa sljedovima A ili B nego što je sa slijedom D ili E, jer je dužina grane od C do korijena A i B manja nego dužina grane od C do korijena D i E. Za razliku od ukorijenjenog stabla na slici 2, kod ove vrste stabla nije vidljiv najbliži zajednički predak sljedova A, B, C, D i E jer na slici 3 ne postoji vidljiv korijen. Isto tako nije moguće odrediti smjer evolucije.

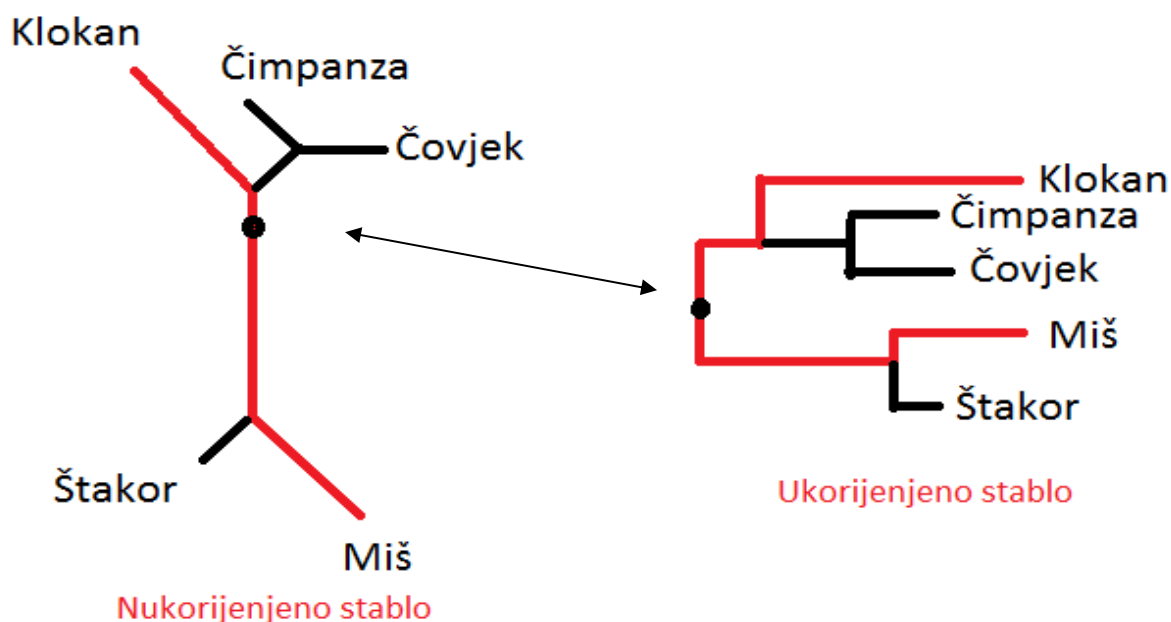
3.3. Pretvorba neukorijenjenog u ukorijenjeno stablo

Iz razloga što neukorijenjeno stablo sadrži manje informacija u odnosu na ukorijenjeno stablo, poželjno je napraviti pretvorbu iz neukorijenjenog u ukorijenjeno stablo. To se može napraviti na jedan od sljedeća dva načina (Šikić M. & Domazet-Lošo M., 2013):

1. Dodavanje korijena na pola puta između dva najudaljenija lista u stablu (*eng. midpoint rooting*)
2. Dodavanje najudaljenijeg taksona kao taksona koji se veže direktno na korijen (*eng. outgroup rooting*)

3.3.1. Dodavanje korijena na pola puta između dva najudaljenija lista u stablu

Kao što ime kaže, kod ove metode traže se dva najudaljenija lista u stablu, te se na polovici te dužine stvori korijen.

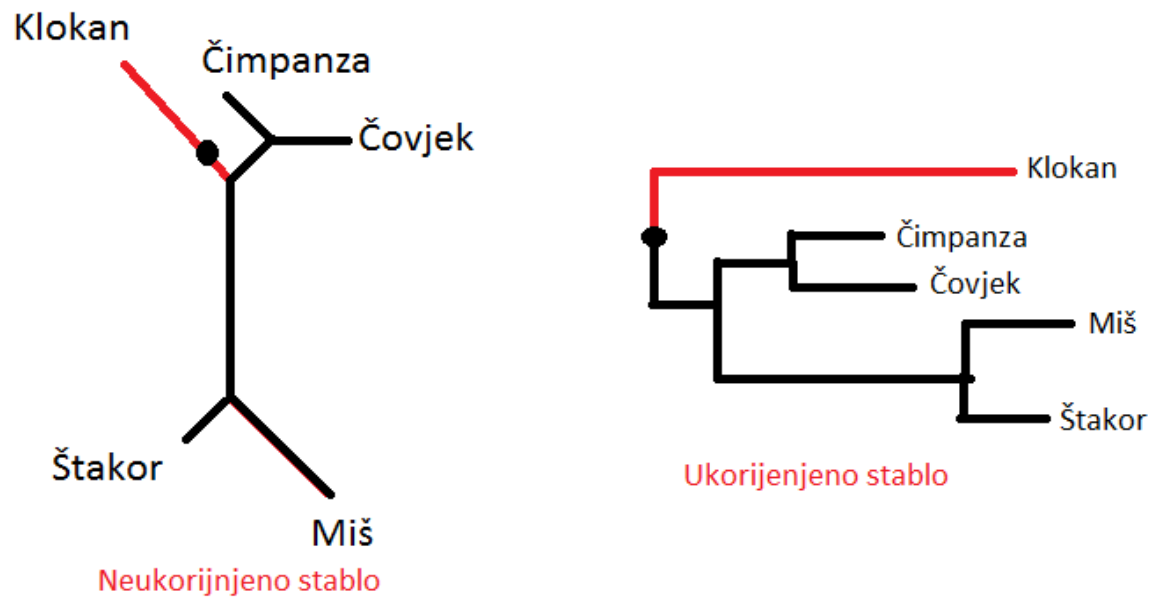


Slika 4. Dodavanje korijena na pola puta između dva najudaljenija lista u stablu

Na slici 4, kod neukorijenjenog stabla najudaljeniji taksoni su klokan i miš. Točno na polovici te dužine napravi se korijen. Ukoliko je evolucija kod svih organizama tekla jednakom brzinom, korijen je napravljen na pravom mjestu. Međutim, u stvarnosti supstitucijska stopa (*eng. substitution rate*) nije jednaka u svim organizmima tijekom vremena, stoga ne vrijedi pretpostavka konstantne frekvencije molekularnog sata (*eng. molecular clock*). Iz navedenog razloga češće se koristi metoda pretvorbe dodavanjem najudaljenijeg taksona kao taksona koji se veže direktno na korijen.

3.3.2. Dodavanje najudaljenijeg taksona kao taksona koji se veže direktno na korijen

Za razliku od prethodne metode kod koje se ukorijenjeno stablo crta na temelju neukorijenjenog, kod ove metode koristi se već postojeće znanje za odabir lokacije korijena. Odabire se najudaljeniji, vanjski (*eng. outgroup*) takson, za koji se zna da je najmanje povezan sa svim ostalim taksonima u stablu, te se on veže direktno na korijen.



Slika 5. Dodavanje najudaljenijeg taksona kao taksona koji se veže direktno na korijen

Na slici 5, kod neukorijenjenog stabla takson koji je najviše udaljen od svih ostalih je klockan. Zna se da su klockani sisavci koji pripadaju podrazredu tobolčara, dok svi ostali sisavci imaju posteljicu. Iz tog razloga ukorijenjeno stablo sa slike 4 ne prikazuje točne evolucijske odnose, dok stablo sa slike 5 prikazuje.

Postoji nekoliko razloga zašto je određivanje točnog korijena bitno. Prvi razlog se svodi na razumijevanje stabla. Sva 3 stabla, neukorijenjeno i 2 ukorijenjena imaju istu topologiju, međutim, ukorijenjeno stablo sa slike 4 prikazuje krive evolucijske odnose. Drugi, važniji razlog, svodi se na smjer evolucije i zaključke o bliskosti taksona. Promatrajući neukorijenjeno stablo i ukorijenjeno stablo sa slike 4 može se zaključiti da je klockan sa čovjekom i čimpanzom u bližem srodstvu nego čimpanza i čovjek sa mišem i štakorom. Gledajući samo razlike u genomu to je točno, međutim evolucijski gledano svi sisavci sa posteljicom su potekli od zajedničkog pretka i jednako su udaljeni od tobolčara.

3.4. Matrica udaljenosti

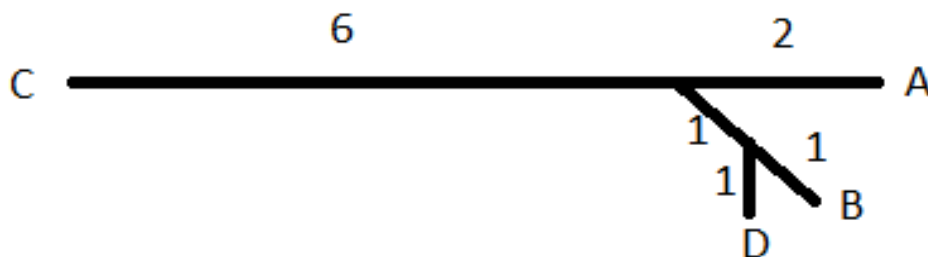
Matrica udaljenosti je kvadratna matrica ($N \times N$) čije su vrijednosti udaljenosti između taksona, te se pomoću nje može izgraditi neukorijenjeno i ukorijenjeno stablo. Matrica zadovoljava sljedeća svojstva:

- udaljenosti na glavnoj dijagonali su 0 ($X_{ij} = 0$ za sve $i=j$, $1 \leq i,j \leq N$)
- sve vrijednosti matrice koje nisu na glavnoj dijagonali su pozitivne ($X_{ij} > 0$ za sve $i \neq j \in 1 \leq i,j \leq N$)
- matrica je simetrična ($X_{ij} = X_{ji}$ za sve $i,j \in 1 \leq i,j \leq N$)
- zadovoljava nejednakost trokuta (za svaki i, j $X_{ij} \leq X_{ik} + X_{kj}$ za sve k)

Tablica 1. Matrica udaljenosti

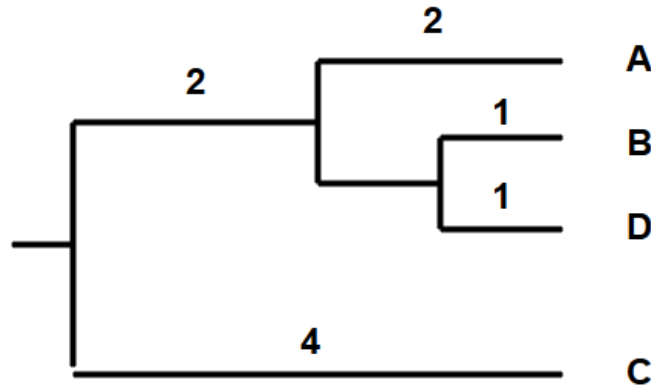
	A	B	C	D
A	0	4	8	4
B	4	0	8	2
C	8	8	0	8
D	4	2	8	0

Iz matrice udaljenosti sa tablice 1 jednostavno se može nacrtati neukorijenjeno stablo. Duljina grana odgovara duljinama zapisanim u matrici udaljenosti.



Slika 6. Neukorijenjeno stablo nastalo iz matrice udaljenosti

Pomoću prethodno opisane metode „Midpoint rooting“ neukorijenjeno stablo sa slike 6 pretvori se u ukorijenjeno stablo na slici 7, tako da se pronađe najveća udaljenost na stablu i u središtu napravi korijen.



Slika 7. Ukorijenjeno stablo nastalo iz neukorijenjenog

3.5. Broj mogućih filogenetskih stabala u ovisnosti o broju taksona

Filogenetska stabla mogu biti binarna, gdje svaki korijen ima najviše 2 podstabla (lijevo i desno), ili korijen može imati više djece. U ovom radu proučavati će se samo binarna filogenetska stabla.

Neka N predstavlja broj taksona za koje se gradi binarno filogenetsko stablo. Ako se radi o neukorijenjenom filogenetskom stablu, broj mogućih stabala je (Cavalli-Sforza L. L., Edwards A. W. F, 1967)

$$(2N - 3)!! = \frac{(2N-3)!}{2^{N-2}(N-2)!}, \text{ za } N \geq 2 \quad (3.1)$$

Ukoliko se radi o ukorijenjenom filogenetskom stablu, broj mogućih stabala je:

$$(2N - 5)!! = \frac{(2N-5)!}{2^{N-3}(N-3)!}, \text{ za } N \geq 3 \quad (3.2)$$

Iz (3.1) i (3.2) je vidljivo da broj mogućih stabala raste u ovisnosti o faktorijelama. Primjerice ako uzmemo da je broj taksona $N = 3$, broj mogućih neukorijenjenih stabala je 3, dok je broj mogućih ukorijenjenih stabala 1. Ukoliko dodamo još 2 taksona u stablo ($N = 5$), broj mogućih neukorijenjenih stabala raste na 105, a ukorijenjenih na 15. Zbog navedenog razloga nije moguće testirati sva stabla za velike N , već se koristi heuristički pristup.

Heuristika obuhvaća metode i tehnike rješavanja problema, učenja i otkrivanja koje su bazirane na iskustvu. Heurističke metode se koriste da ubrzaju proces

pronalaženja dovoljno dobrog (približno točnog) rješenja u situacijama kada provođenje detaljnog istraživanja nije praktično ili je pak nemoguće.

3.6. Udaljenost sljedova

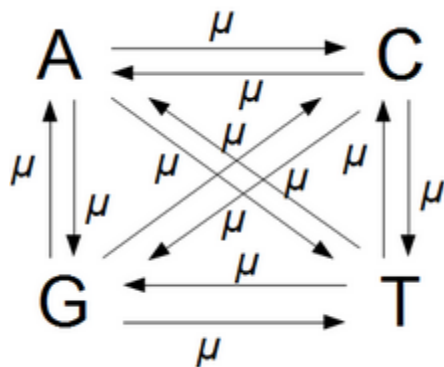
Kako bi izgradili matricu udaljenosti potrebno je znati koliko su sljedovi međusobno udaljeni. Prvo se sljedovi poravnavaju, a onda se broje vidljive razlike između tih sljedova. Međutim broj tih razlika je uvijek manji nego stvarni broj mutacija koje su nastale. Razlog tome je što u jednom slijedu nukleotid A mutira u nukleotid C i zatim ponovno može mutirati u nukleotid A. Time su zapravo nastale 2 mutacije, dok promatrači to vide kao da nije bilo nikakvih mutacija.

U cilju određivanja stvarnog broja mutacija razvijeno je nekoliko evolucijskih DNA modela koji uzimaju u obzir vrijeme koje je preteklo od razdvajanja sljedova i brzinu kojom je svaki od sljedova mutirao. Neki od tih modela su: model je JC69 (Jukes & Cantor 1969), K80 (Kimura 1980), F81 (Felsenstein 1981), HKY85 (Hasegawa, Kishino & Yano 1985), T92 (Tamura 1992), TN93 model (Tamura & Nei 1993), GTR: Generalised time-reversible (Tavaré 1986). U ovom radu detaljnije će biti obrađen prvi evolucijski Jukes - Cantorov model.

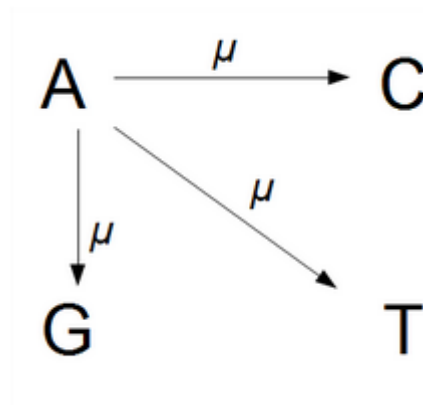
3.6.1. Jukes – Cantorov evolucijski model

Jukes – Cantorov evolucijski model (Jukes & Cantor, 1969) je prvi evolucijski model, a ujedno je i najjednostavniji. Zbog jednostavnosti nije pogodan za određivanje evolucijske udaljenosti jako udaljenih sljedova.

Kreće se od pretpostavke da svaki od četiri nukleotida, A, C, G i T ima jednaku vjerojatnost mutacije μ u bilo koji drugi nukleotid. Primjerice vjerojatnost da nukleotid A mutira u C, G ili T je 3μ .



Slika 8. Sve moguće mutacije



Slika 9. Mutacija nukleotida A

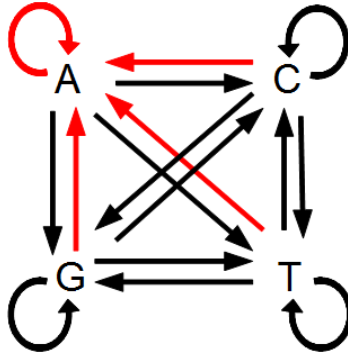
Druga pretpostavka je da je frekvencija svih nukleotida u slijedu jednaka, odnosno vrijedi $f_A = f_C = f_G = f_T = \frac{1}{4}$. To se može vidjeti i na slici 8. Postoje 3 mutacije koje vode prema nukleotidu A, dok 9 mutacija ne vodi: $\frac{3}{9+3} = \frac{1}{4}$. Ako su prema prvoj pretpostavci sve mutacije jednako vjerojatne, uvijek će biti $\frac{1}{4}$ nukleotida A.

Pogledajmo sada Poissonovu raspodjelu (*eng. Poisson distribution*) u kojoj je vjerojatnost pojedinačnog događaja mala, događaji su nezavisni, a brojimo ih tijekom vremena. Jednostavan primjer Poissonove raspodjele je broj ljudi koji kupi auto iz salona tijekom dana. Poissonova raspodjela kaže da je vjerojatnost da se dobije broj k može izračunati ako se zna očekivani broj λ :

$$P(k | \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Primjerice ako se zna da dnevno troje ljudi kupi novi auto, kolika je vjerojatnost da će na današnji dan 5 ljudi kupiti auto?

$$P(5 | 3) = 0,10$$



Slika 10. Juke-Cantorov model

Slika 10 predstavlja Juke-Cantorov model evolucije. U odnosu na sliku 8 dodana je mogućnost da nukleotid mutira u samog sebe, odnosno da nukleotid uopće ne mutira. Ta vjerojantost je opet μ . Iz toga proizlazi da je vjerojatnost mutacije 4μ , a prema Poissonu $\lambda = 4\mu t$ (t predstavlja vrijeme). Vjerojatnost barem jedne mutacije je:

$$1 - P(0 | \lambda) = 1 - e^{-\lambda} = 1 - e^{-4\mu t}$$

Vjerojatnost da jedan nukleotid neće mutirati, primjerice nukleotid A je:

$$P(A|A, \mu, t) = \frac{1}{4}(1 - e^{-4\mu t})$$

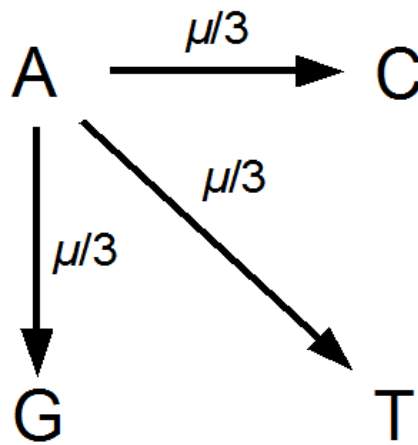
a vjerojatnost da će mutirati u neki drugi nukleotid je:

$$P(\text{neki drugi} | \mu, t) = d = \frac{3}{4}(1 - e^{-4\mu t})$$

Da bi se izračunala udaljenost sljedova kroz vrijeme, potrebno je supstituirati $D = \mu t$:

$$D = -\frac{1}{4} \ln(1 - \frac{4}{3}d)$$

Jedan od pristupa je podijeliti μ sa 3 kako bi vjerojatnost mutacije iz jednog nukleotida bila μ (slika 11).



Slika 11. Vjerijatost mutacije iz jednog nukleotida u druge

Iz toga se dobiva konačna udaljenost dvaju sljedova:

$$D = -\frac{3}{4} \ln\left(1 - \frac{4}{3}d\right)$$

Ova formula je najčešće korištena formula Juke - Cantorovog modela.

4. UPGMA

UPGMA (Unweighted Pair Group Method with Arithmetic Mean) (Michenet & Sokal, 1958) je najjednostavnija metoda crtanja filogenetskog stabla. Velika mana ove metode je što se polazi od pretpostavke molekularnog sata, odnosno pretpostavlja se da je mutacijska stopa svih organizama jednaka. Posljedica toga je da su svi listovi u stablu jednako udaljeni od korijena. Iz tog se razloga UPGMA metoda izgradnje stabla koristi za prikaz udaljenosti blisko srodnih organizama, inače će topologija stabla biti kriva. Još jedna od upotreba metode je za određivanje poravnanja više sljedova odjednom, te za izgradnju inicijalnog stabla (eng. *guide tree*) kao pomoć drugim složenijim i preciznijim metodama (Šikić M. & Domazet-Lošo M., 2013).

4.1. Algoritam

Vremenska složenost UPGMA algoritma je $O(n^2)$, n predstavlja broj taksona.

OZNAKA	OPIS
T_i	i-ti takson
$T_{i,j}$	Takson nastao spajanjem taksona T_i i T_j
$d(T_i, T_j)$	Udaljenost između taksona T_i i T_j
$h(T_i)$	Visina taksona (udaljenost taksona od listova)
$M_{i,j}$	Matrica udaljenosti

1. UPGMA metoda kreće pronalaženjem dva najsrodnija taksona T_i i T_j , taksona čija je udaljenost u matrici udaljenosti $M_{i,j}$ najmanja.
2. Stvoriti novi takson $T_{i,j}$ koji predstavlja najbliži zajednički predak T_i i T_j . Takson T_i spojiti s taksonom $T_{i,j}$. Isto napraviti i sa taksonom T_j .
 - Visina taksona $T_{i,j}$ je:

$$h(T_{i,j}) = d(T_i, T_j) / 2 \quad (4.1)$$

- Udaljenost T_i do najbližeg zajedničkog pretka $T_{i,j}$ je razlika visina $T_{i,j}$ i T_i . Isto vrijedi i za T_j .

$$d(T_i, T_{i,j}) = h(T_{i,j}) - h(T_i) \quad (4.2)$$

$$d(T_j, T_{i,j}) = h(T_{i,j}) - h(T_j) \quad (4.3)$$

3. Ako su T_i i T_j zadnji par taksona algoritam je gotov.
4. Potrebno je promijeniti udaljenosti u novo nastaloj matrici udaljenosti. Za sve $T_l \neq T_k$

$$d(T_k, T_l) = (|T_i| \cdot d(T_i, T_k) + |T_j| \cdot d(T_j, T_k)) / (|T_i| + |T_j|) \quad (4.4)$$
5. Eliminirati taksoni T_i i T_j iz matrice udaljenosti i dodati novi takson $T_{i,j}$, vratiti se na korak 1.

4.2. Primjer

Zadan je skup taksona A, B, C i D te matrica udaljenosti sa Tablice 1.

1. Među svim mogućim parovima taksona, taksoni između kojih je udaljenost najmanja su B i D.

	A	B	C	D
A	0	4	8	4
B	4	0	8	2
C	8	8	0	8
D	4	2	8	0

2. Takson koji nastaje spajanjem taksona B i D je takson BD. Visina taksona BD prema (4.1) je:

$$h(BD) = d(B, D) / 2 = 2 / 2 = 1$$

Udaljenosti od B i D prema (4.2) i (4.3) respektivno su:

$$d(B, BD) = h(BD) - h(B) = 1 - 0 = 1$$

$$d(D, BD) = h(BD) - h(D) = 1 - 0 = 1$$

3. Taksoni B i D nisu posljednji par taksona pa se algoritam nastavlja.

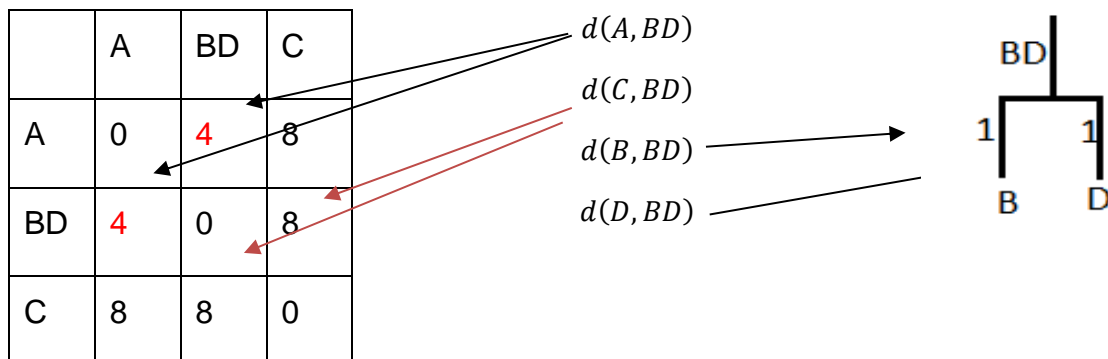
4. Za takson A, BD i C potrebno je izračunati nove udaljenosti:

$$d(A, BD) = \frac{|B| \cdot d(B, A) + |D| \cdot d(D, A)}{|A| + |D|} = \frac{1 \cdot 4 + 1 \cdot 4}{1 + 1} = 4$$

$$d(C, BD) = \frac{|B| \cdot d(B, C) + |D| \cdot d(D, C)}{|A| + |D|} = \frac{1 \cdot 8 + 1 \cdot 8}{1 + 1} = 8$$

$$d(A, C) = 8$$

5. Eliminirati takson B i D i zatim dodati takson BD u matricu udaljenosti (slika 12) :



Slika 12. Matrica i stablo nastali dodavanjem taksona BD

6. Među svim mogućim parovima taksona na slici 12, taksoni između kojih je udaljenost najmanja su A i BD.

7. Takson koji nastaje spajanjem taksona A i BD je takson ABD. Visina taksona ABD prema (4.1) je:

$$h(ABD) = d(A, BD) / 2 = 4 / 2 = 2$$

Udaljenosti od A i BD prema (4.2) i (4.3) su:

$$d(A, ABD) = h(ABD) - h(A) = 2 - 0 = 2$$

$$d(BD, ABD) = h(ABD) - h(BD) = 2 - 1 = 1$$

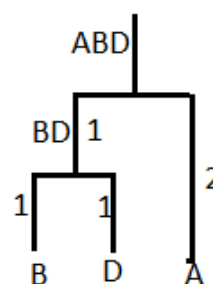
8. Taksoni A i BD nisu posljednji par taksona pa se algoritam nastavlja.

9. Za takson ABD i C potrebno je izračunati nove udaljenosti prema (4.4):

$$d(C, ABD) = \frac{|A| \cdot d(A, C) + |BD| \cdot d(BD, C)}{|A| + |BD|} = \frac{1 \cdot 8 + 2 \cdot 8}{3} = 8$$

10. Eliminirati takson A i BD i zatim dodati takson ABD u matricu udaljenosti:

	ABD	C
ABD	0	8
C	8	0



Slika 13. Matrica i stablo nastalo dodavanjem taksona ABD

11. Među svim mogućim parovima taksona na slici 13, taksoni između kojih je udaljenost najmanja su C i ABD.

12. Takson koji nastaje spajanjem taksona C i ABD je takson ABCD. Visina taksona ABCD prema (4.1) je:

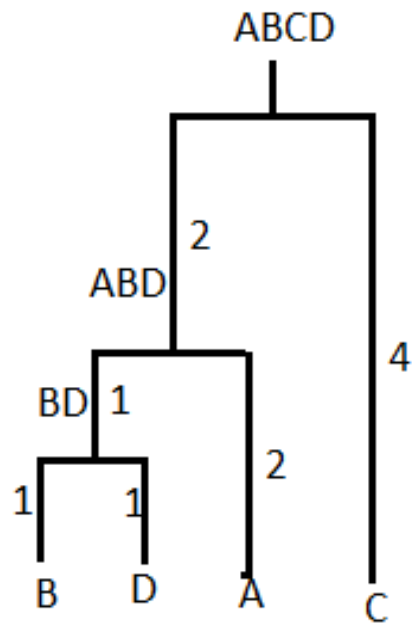
$$h(ABCD) = d(C, ABD) / 2 = 8 / 2 = 4$$

Udaljenosti od C i ABD prema (4.2) su:

$$d(C, ABCD) = h(ABCD) - h(C) = 4 - 0 = 4$$

$$d(ABD, ABCD) = h(ABCD) - h(ABD) = 4 - 2 = 2$$

13. Taksoni C i ABD jesu posljednji par taksona. Taksoni se spajaju, a algoritam završava:



Slika 14. Stablo izgrađeno UPGMA algoritmom

5. Metoda povezivanja susjeda

Metoda povezivanja susjeda (*eng. Neighbour joining*) (Saotou N. & Nei M., 1987) je metoda izgradnje stabla od listova prema korijenu. Za razliku od UPGMA, ova metoda osim topologije filogenetskog stabla određuje i duljinu njegovih grana, odnosno pretpostavka molekularnog sata ne vrijedi.

5.1. Algoritam

Vremenska složenost metode povezivanja susjeda je $O(n^3)$, n predstavlja broj taksona.

OZNAKA	OPIS
--------	------

T_i	i-ti takson
$D_{i,j}$	Udaljenost između taksona T_i i T_j
S_i	Suma udaljenosti taksona T_i od svih ostalih taksona
$M_{i,j}$	Mjera udaljenosti između taksona T_i i T_j

1. Potrebno je pronaći dva taksona T_i i T_j koji su najbliži, odnosno njihova mjera udaljenosti $M_{i,j}$ mora biti najmanja. Mjera udaljenosti M za par taksona računa se prema sljedećoj formuli:

$$S_i = \sum_{j=1}^n D_{i,j} \quad (5.1)$$

$$M_{i,j} = D_{i,j} - \frac{1}{n-2} (S_i + S_j) \quad (5.2)$$

2. U stablo se dodaje novi čvor T_k , koji povezuje čvorove T_i i T_j . Udaljenost T_k od T_i i T_j računa se prema formuli:

$$D_{i,k} = \frac{1}{2} D_{i,j} + \frac{1}{2(n-2)} (S_i - S_j) \quad (5.3)$$

3. Izračunati nove udaljenosti $D_{l,k}$ za sve taksone T_l , $T_l \neq T_i, T_j$ prema formuli:

$$D_{k,l} = \frac{1}{2}(D_{i,l} + D_{j,l} - D_{i,j}) \quad (5.4)$$

Iz matrice udaljenosti maknuti taksone T_i i T_j .

4. Ponavljati prethodna tri koraka dok $n > 2$. Ukoliko $n = 2$, preostala dva taksona potrebno je spojiti. Udaljenost između tih taksona je posljednja preostala vrijednost u matrici.

5.2. Primjer

Zadan je skup taksona A,B, C i D te matrica udaljenosti prema tablici 1.

1. Izračun mjere udaljenosti za parove taksona:

	A	B	C	D
A	0	4	8	4
B	4	0	8	2
C	8	8	0	8
D	4	2	8	0

Prema (5.1):

$$S_A = 0 + 4 + 8 + 4 = 16$$

$$S_B = 4 + 0 + 8 + 2 = 14$$

$$S_C = 8 + 8 + 0 + 8 = 24$$

$$S_D = 4 + 2 + 8 + 0 = 14$$

Mjera udaljenosti za parove takson prema (5.2)

$$M_{A,B} = D_{A,B} - \frac{1}{4-2}(S_A + S_B) = 4 - \frac{1}{2}(16 + 14) = -11$$

$$M_{A,C} = D_{A,C} - \frac{1}{4-2}(S_A + S_C) = 8 - \frac{1}{2}(16 + 24) = -12$$

$$M_{A,D} = -11, \quad M_{B,C} = -11, \quad M_{B,D} = -12, \quad M_{C,D} = -12$$

Parovi taksona (A,C), (B,D), (C,D) imaju najmanju mjeru udaljenosti $M = -12$, stoga možemo odabrati bilo koji par od navedenih kao najbliži.

2. U stablo se dodaje novi čvor AC, te se prema (5.2) računa udaljenost čvorova A i C od dodanog čvora AC:

$$D_{A,AC} = \frac{1}{2} D_{A,C} + \frac{1}{2 * (n - 2)} (S_A - S_C) = \frac{1}{2} * 8 + \frac{1}{2(4 - 2)} (16 - 24) = 2$$

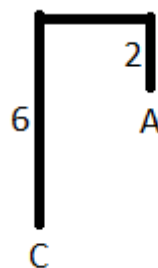
$$D_{C,AC} = \frac{1}{2} D_{A,C} + \frac{1}{2 * (n - 2)} (S_C - S_A) = \frac{1}{2} * 8 + \frac{1}{2(4 - 2)} (24 - 16) = 6$$

3. Taksoni A i C se brišu iz matrice udaljenosti, a dodaje se novi takson AC. Prema (5.4) računaju se nove udaljenosti taksona AC od svih preostalih taksona:

$$D_{AC,B} = \frac{1}{2} (D_{A,B} + D_{C,B} - D_{A,C}) = \frac{1}{2} (4 + 8 - 8) = 2$$

$$D_{AC,D} = \frac{1}{2} (D_{A,D} + D_{C,D} - D_{A,C}) = \frac{1}{2} (4 + 8 - 8) = 2$$

	AC	B	D
AC	0	2	2
B	2	0	2
D	2	2	0



Slika 15. Matrica udaljenosti i stablo nastalo dodavanjem taksona AC

4. Iz razloga što je broj taksona $n = 3$, algoritam se nastavlja od prvog koraka:

$$S_{AC} = 4, S_B = 4, S_D = 4$$

$$M_{AC,B} = D_{AC,B} - \frac{1}{3-2}(S_{AC} + S_B) = 2 - (4 + 4) = -6$$

$$M_{AC,D} = -6, M_{B,D} = -6$$

Svi parovi taksona imaju jednaki mjeru udaljenosti $M = -6$, stoga možemo odabrati bilo koji par od navedenih kao najbliži.

5. U stablo se dodaje novi čvor BD, te se prema (5.2) računa udaljenost čvorova B i D od dodanog čvora BD:

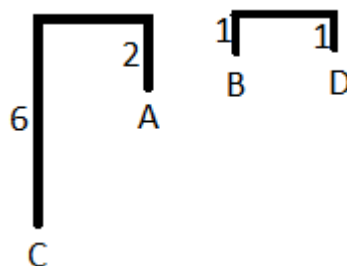
$$D_{B,BD} = \frac{1}{2}D_{B,D} - \frac{1}{2 * (n-2)}(S_B - S_D) = \frac{1}{2} * 2 - \frac{1}{2(3-2)}(4 - 4) = 1$$

$$D_{D,BD} = \frac{1}{2}D_{B,D} - \frac{1}{2 * (n-2)}(S_D - S_B) = \frac{1}{2} * 2 - \frac{1}{2(3-2)}(4 - 4) = 1$$

6. Taksoni B i D se brišu iz matrice udaljenosti, a dodaje se novi takson BD. Prema (5.4) računaju se nove udaljenosti taksona BD od svih preostalih taksona:

$$D_{BD,AC} = \frac{1}{2}(D_{B,AC} + D_{D,AC} - D_{B,D}) = \frac{1}{2}(2 + 2 - 2) = 1$$

	AC	BD
AC	0	1
BD	1	0

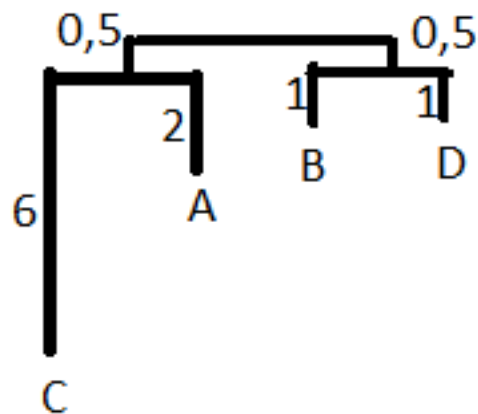


Slika 16. Matrica udaljenosti i stablo nastalo dodavanjem taksona BD

7. Broj taksona je $n = 2$. Posljednja dva preostala taksona potrebno je spojiti.

$$D_{BD,ACBD} = \frac{1}{2}D_{BD,AC} = 0,5$$

$$D_{AC,ACBD} = \frac{1}{2}D_{BD,AC} = 0,5$$



Slika 17. Filogenetsko stablo nastalo spajanjem taksona AC i BD

6. Analiza programske potpore

Programska potpora, u daljnjem tekstu Program, omogućava prikaz matrice udaljenosti i filogenetskog stabla proizvoljnog broja genoma ili proteinskih lanaca međusobno srodnih organizama. Ulaz programa moraju biti poravnati lanci jednakih duljina u FASTA formatu (slika 18). FASTA format predstavlja tekstualnu reprezentaciju genskih ili proteinskih lanaca, gdje je svaki nukleotid ili aminokiselina kodirana jednim slovom abecede. Udaljenost između lanaca računa se pomoću Jukes-Cantorovog modela, prethodno opisanog u poglavlju 3.6.1.

```
>WP_044889774.1 hypothetical protein [Myxococcus hansupus]
MVRVPGRDGGAMTRINSFNGPMIQPAATQAQSRRAASTSFGSLVN----PMAPANRPGDP
TMVSGGAVVASALASVGSSPNNGLNSYLSAVGRGPISEDGSRGPVS---QAPAGSEQAQ
QEQLMELAEMSAATLSNSIIHMGNMKMKVDMERE
>AEI64772.1 hypothetical protein LILAB_14335 [Myxococcus fulvus HW-1]
-----MTRINSFTGPMIQPATTQAQTRAASTSFGALVN----PMAPANRPGDP
LMVSGGAVVASALASVGSSPNNGLNSYLSAVGRGPISEDGTRGPAS---QAPAGSEQAQ
QEQLMELAEMSAATLSNSILHMGNMKMKVDLDRE
```

Slika 18. Poravnati proteinski lanci u FASTA formatu

Svi proteinski lanci koji su analizirani preuzeti su sa NCBI-a (National Center for Biotechnology Information), a poravnati su sa Clustal Omega (2013), programom za poravnanje više genskih ili proteinskih lanaca veličine do 4000 taksona. Poravnate datoteke se nalaze na CD-u priloženom uz rad u direktoriju examples. Datoteke su imenovane na način alignedNxL.fasta, gdje N predstavlja broj taksona, a L duljinu lanaca.

Testiranja su napravljena na računalu čije su karakteristike navedene u tablici 2. Izračunate vrijednosti u tablicama 3, 4, 5, 6 i 7 su aritmetička sredina 10 mjerenja.

Tablica 2. Karakteristike računala

Procesor	Intel(R) Core(TM) i5-4460 CPU @ 3.20GHz, 3201 Mhz, 4 Core(s)
RAM	8GB DDR3, 1333 Mhz
Operacijski sustav	Microsoft Windows 10 Pro 64 bit

Tablica 3. Vrijeme potrebno za izračun matrice udaljenosti iz FASTA datoteke, broj taksona je konstantan

Duljina lanca [broj nukleotida]	Vrijeme [ms]
20000	1.0
64000	3.0
100000	4.0
200000	8.0
500000	21.3
700000	33.0
1000000	41.6

Tablica 4. Vrijeme potrebno za izračun matrice udaljenosti iz FASTA datoteke, duljina lanca je konstanta

Broj taksona	Vrijeme [ms]
100	40.0
200	213.0
400	397.0
800	1944.3
1000	3274.3

Tablica 5. Iskorištenost centralne procesne jedinice (CPJ)

Broj taksona	UPGMA [%]	Metoda povezivanja susjeda [%]
25	1.8	1.8
50	2.8	3.0
100	12.8	26.2
500	26.2	26.8
1000	26.2	26.7

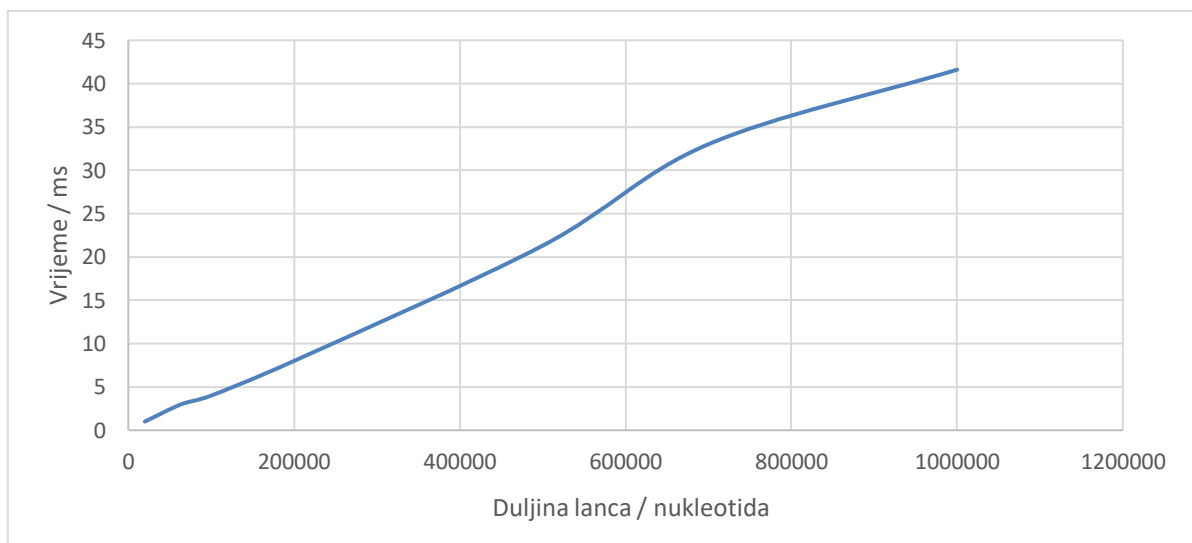
Tablica 6. Iskorištenost memorije (RAM)

Broj taksona	UPGMA [MB]	Metoda povezivanja susjeda [MB]
25	60	77
50	80	175
100	317	725
500	568	729
1000	628	880

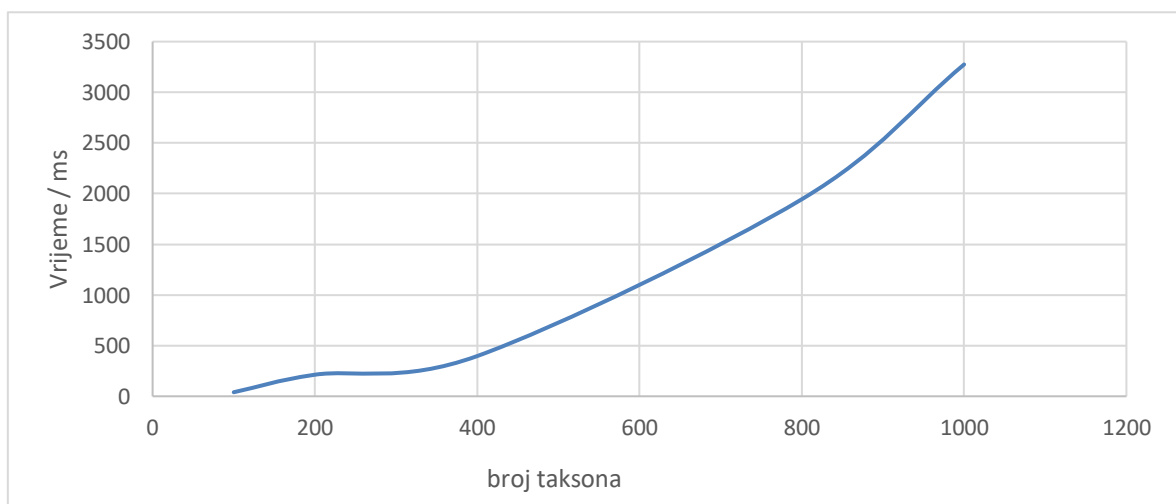
Tablica 7. Vrijeme potrebno za izgradnju stabla iz izračunate matrice udaljenosti

Broj taksona	UPGMA [ms]	Metoda povezivanja susjeda [ms]
25	12.4	24.2
50	44.6	233.8
100	517.8	3334.4
500	46697.2	2434326.4
1000	368707.0	>>2434326.4

Graf na slici 19 prikazuje vrijeme potrebno za izgradnju matrice udaljenosti iz ulazne datoteke u FASTA formatu. Vidljivo je da je se radi o linearnoj složenosti $O(L)$, L predstavlja duljinu lanca. S druge strane graf na slici 20 prikazuje vrijeme potrebno za izgradnju matrice udaljenosti u ovisnosti o broju taksona, dok je duljina lanaca konstantna. Ovaj put se radi o eksponencijalnoj složenosti, točnije $O(n^2)$, a n predstavlja broj taksona. Ukupno vrijeme izgradnje matrice ovisi broju taksona i njihovoj duljini, odnosno ukupna složenost je $O(n^2 * L)$.



Slika 19 . Vrijeme potrebno za izračun matrice udaljenosti u ovisnosti o duljini lanca

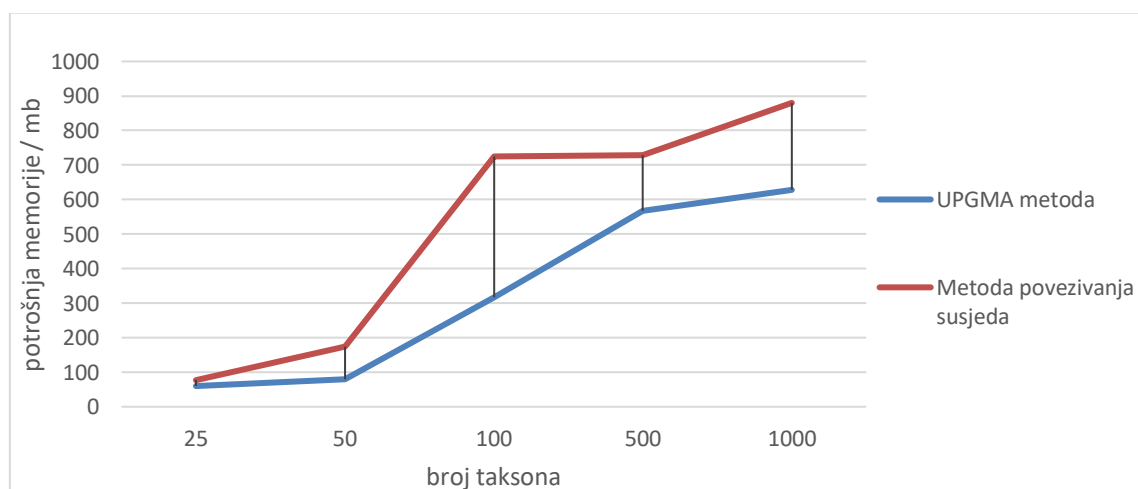


Slika 20. Vrijeme potrebno za izračun matrice udaljenosti u ovisnosti o broju taksona

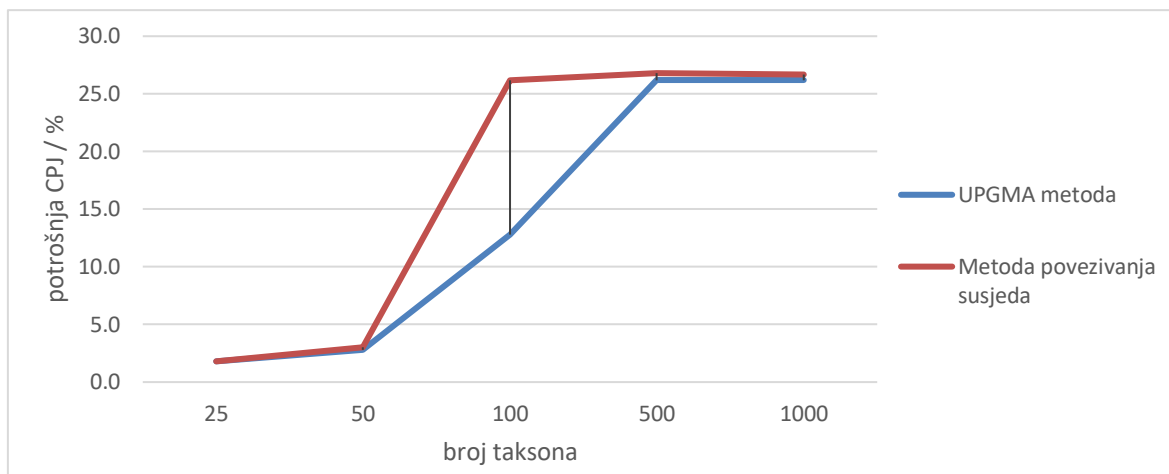
Grafovi na slikama 21, 22 i 23 prikazuju ovisnost potrošnje memorije, potrošnje CPJ i vremena izgradnje filogenetskog stabla o broju taksona. Zbog toga što se vrijeme izgradnje stabla mjerilo od trenutka kada je matrica izračunata, duljina sljedova taksona se ne uzima u razmatranje. Izgradnja stabla ne ovisi o duljini slijeda, već o veličini matrice koju određuje broj taksona.

Vidljivo je da je metoda povezivanja susjeda u sva tri aspekta zahtjevnija, međutim takav rezultat nije neočekivan iz razloga što je složenost UGMA metode $O(n^2)$, n predstavlja broj taksona, dok je složenost metode povezivanja susjeda $O(n^3)$. Ova je razlika izrazito vidljiva kod grafa koji prikazuje vrijeme potrebno za izgradnju stabla (slika 23). Razlika u potrošnji CPJ nije toliko velika, jer procesna jedinica uzme maksimalno resursa koliko može za obavljanje pojedinog zadatka, u ovom slučaju izgradnju filogenetskog stabla. Potrošnja memorije kod metode povezivanja susjeda je nešto veća u odnosu na UGMA metodu zbog toga što algoritam metode povezivanja susjeda generira jednu matricu više za pronalazak najmanje udaljenosti između dva taksona.

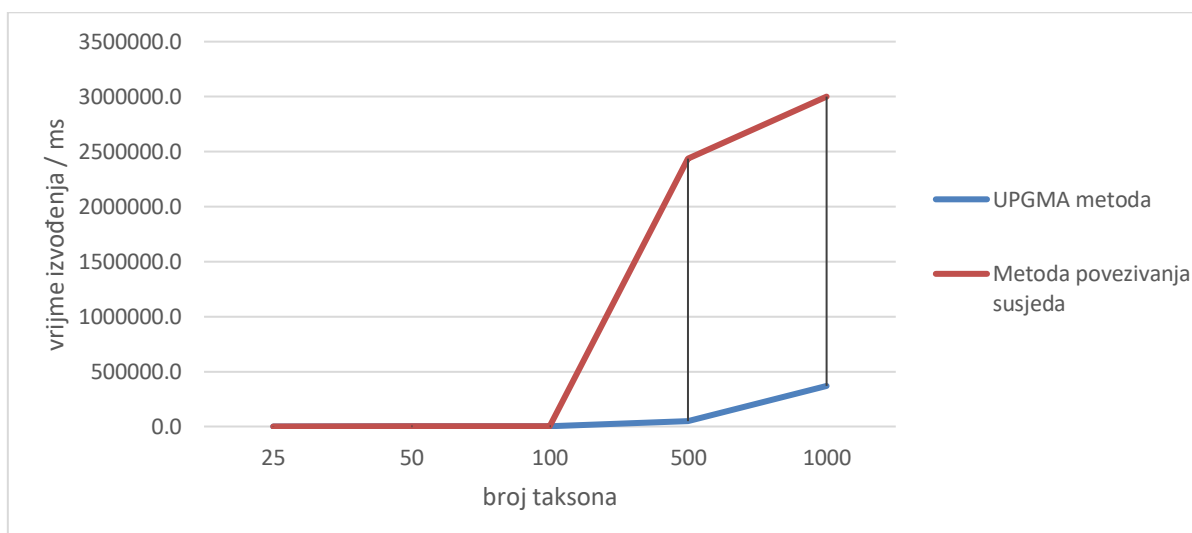
Program je u mogućnosti izgraditi i nacrtati filogenetska stabla veličine do 100 taksona u nekoliko sekundi. Ukoliko se korisnik odluči izgraditi veće stablo, primjerice 500 taksona, za UPGMA metodu morat će pričekati nešto manje od 1 minute, dok će za isto stablo izgrađeno metodom povezivanja susjeda trebati čekati 40-tak minuta (slika 23).



Slika 21. Ovisnost potrošnje memorije o broju taksona



Slika 22. Ovisnost potrošnje CPJ o broju taksona



Slika 23. Vrijeme potrebno za izgradnju filogenetskog stabla iz matrice udaljenosti u ovisnosti o broju taksona

7. Usporedba programske potpore sa MEGA7

MEGA (Molecular Evolutionary Genetics Analysis) verzija 7 (Kumar, Stecher, & Tamura 2015) je program za statističku analizu procesa evolucije i izgradnju filogenetskog stabla. Uključuje mnoge sofisticirane alate koji se koriste u filogenezi, međutim za usporedbu s programskom potporom koja prati ovaj rad koristiti će se sljedeći alati:

- prikaz matrice udaljenosti s mogućnošću odabira modela računanja udaljenosti
- prikaz filogenetskog stabla izgrađenog UPGMA metodom uz mogućnost odabira modela računanja udaljenosti
- prikaz filogenetskog stabla izgrađenog metodom povezivanja susjeda (*eng. neighbour joining*) uz mogućnost odabira modela računanja udaljenosti
- prikaz duljine grana izgrađenog filogenetskog stabla

Neki od modela računanja udaljenosti koje MEGA7 podržava su: Poissonov model, Dayhoffov model (Dayhoff, Schwartz & Orcutt, 1978) i Jones-Taylor-Thorntonov model (Jones, Taylor & Thornton, 1992). Iz razloga što MEGA ne podržava Jukes-Cantorov model, pri uspoređivanju će se uvijek koristiti Poissonov model.

	1	2	3	4	5	6	7	8	9
1: Otolemur_garnettii	0.0000	0.0536	0.0609	0.0577	0.0557	0.0567	0.0577	0.0577	0.0598
2: Aotus_nancymaae	0.0536	0.0000	0.0330	0.0300	0.0279	0.0290	0.0300	0.0300	0.0340
3: Pongo_abelii	0.0609	0.0330	0.0000	0.0149	0.0149	0.0149	0.0159	0.0249	0.0300
4: Homo_sapiens	0.0577	0.0300	0.0149	0.0000	0.0059	0.0059	0.0069	0.0239	0.0290
5: Gorilla_gorilla_gorilla	0.0557	0.0279	0.0149	0.0059	0.0000	0.0049	0.0059	0.0219	0.0269
6: Pan_troglodytes	0.0567	0.0290	0.0149	0.0059	0.0049	0.0000	0.0010	0.0229	0.0279
7: Pan_paniscus	0.0577	0.0300	0.0159	0.0069	0.0059	0.0010	0.0000	0.0239	0.0290
8: Macaca_mulatta	0.0577	0.0300	0.0249	0.0239	0.0219	0.0229	0.0239	0.0000	0.0169
9: Rhinopithecus_roxellana	0.0598	0.0340	0.0300	0.0290	0.0269	0.0279	0.0290	0.0169	0.0000

Slika 24. Program: Matrica udaljenosti nastala iz datoteke aligned9x1032.fasta

	1	2	3	4	5	6	7	8	9
1. Otolemur garnettii PRED		0.0508	0.0603	0.0571	0.0550	0.0561	0.0571	0.0581	0.0603
2. Aotus nancymaae PRED			0.0312	0.0281	0.0261	0.0271	0.0281	0.0302	0.0333
3. Pongo abelii PREDICTE				0.0150	0.0150	0.0150	0.0160	0.0241	0.0302
4. Homo sapiens UPF0577					0.0060	0.0060	0.0070	0.0231	0.0292
5. Gorilla gorilla gorilla PR						0.0050	0.0060	0.0210	0.0271
6. Pan troglodytes PREDIC							0.0010	0.0220	0.0281
7. Pan paniscus PREDICTE								0.0231	0.0292
8. Macaca mulatta PREDIC									0.0160
9. Rhinopithecus roxellana									

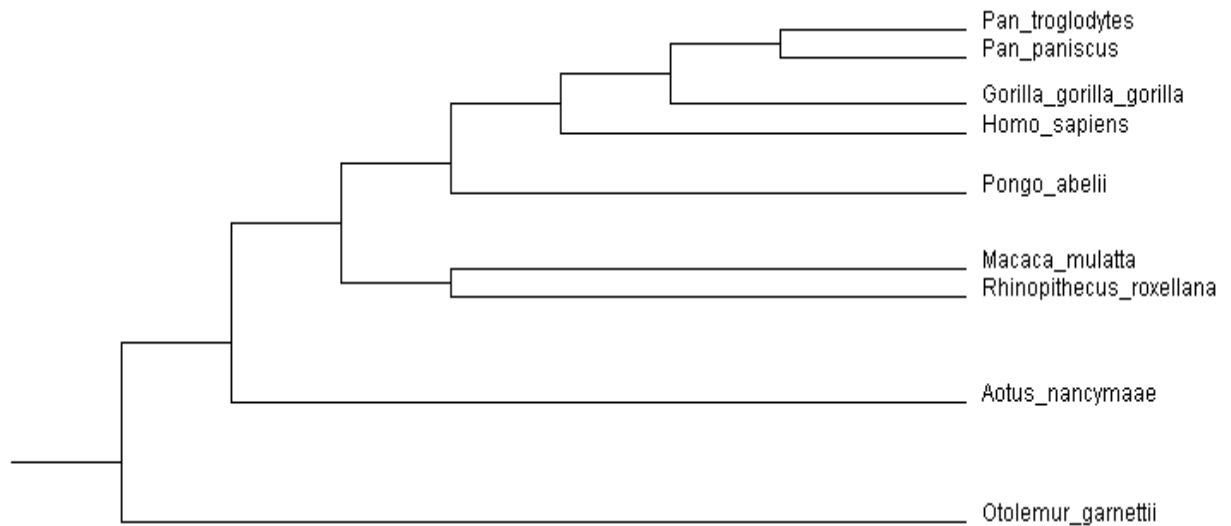
Slika 25. MEGA: Matrica udaljenosti nastala iz datoteke aligned9x1032.fasta

Slike 24 i 25 prikazuju matrice udaljenosti za 9 proteinskih lanaca duljine 1032 aminokiseline, generiranih iz datoteke aligned9x1032.fasta. Podaci su preuzeti sa NCBI-a i poravnati programom Clustal Omega. Udaljenosti prikazane u matricama nisu u potpunosti jednake, a to je posljedica korištenja različitih evolucijskih modela za izračun udaljenosti. Matrica generirana od strane Programa koristi Jukes-Cantorov model, dok MEGA koristi Poissonov model. Bez obzira na činjenicu da se udaljenosti u potpunosti ne poklapaju, odnosno razlikuju se na drugoj ili trećoj znamenki od decimalne točke, matrice se mogu uspoređivati. Obje matrice pokazuju da su najsrodniji organizmi Pan_troglodytes i Pan paniscus, a udaljenost između njih je 0.0010, šesti redak i sedmi stupac obiju matrica (6,7).

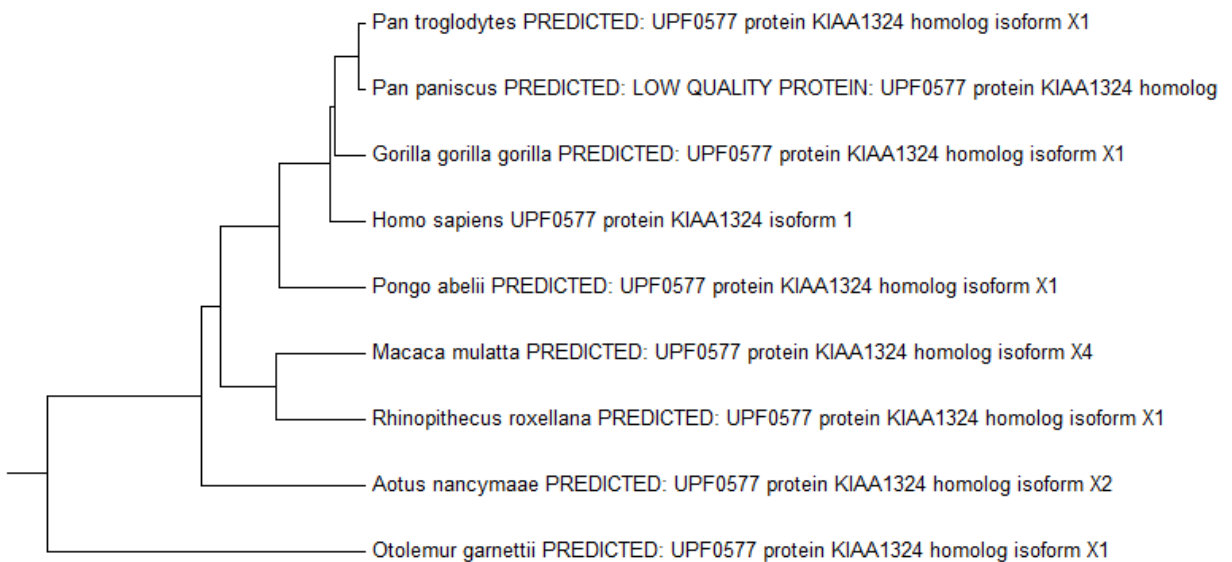
Vidljivo je da su najudaljeniji par taksona Otolemur_garnetti i Pongo_abelii (1,3). I u ovom slučaju oba programa su izračunala isti najudaljeniji par, međutim ovaj put udaljenosti nisu jednake. Program je izračunao udaljenost 0.609, dok je MEGA za istu udaljenost izračunao 0.603. Ta razlika iznosi manje od 1%, a kako bi dokazali

da je zanemariva, filogenetska stabla nastala iz matrica trebala bi biti jednaka.

Stabla nastala iz matrice udaljenosti UPGMA metodom prikazana su na slikama 26 i 27, a stabla nastala metodom povezivanja susjeda prikazana su na slikama 28 i 29.

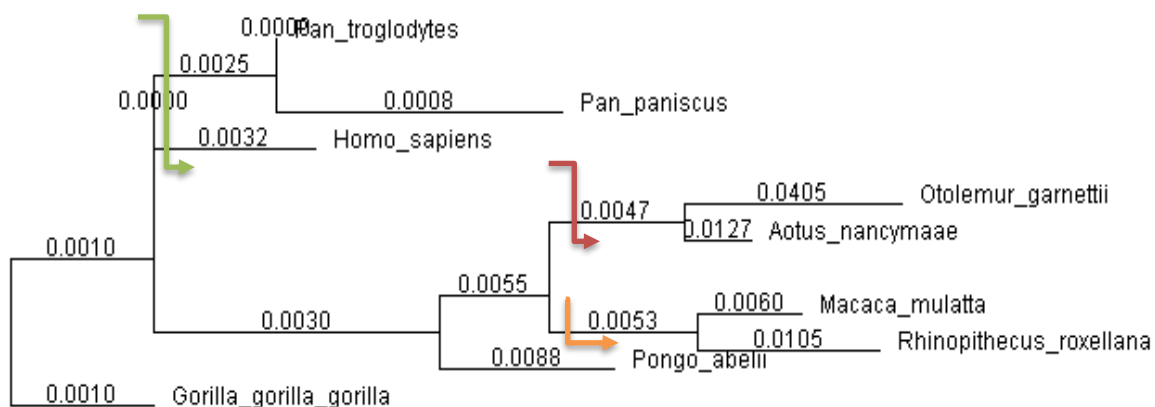


Slika 26. Program: UPGMA stablo nastalo iz matrice sa slike 24

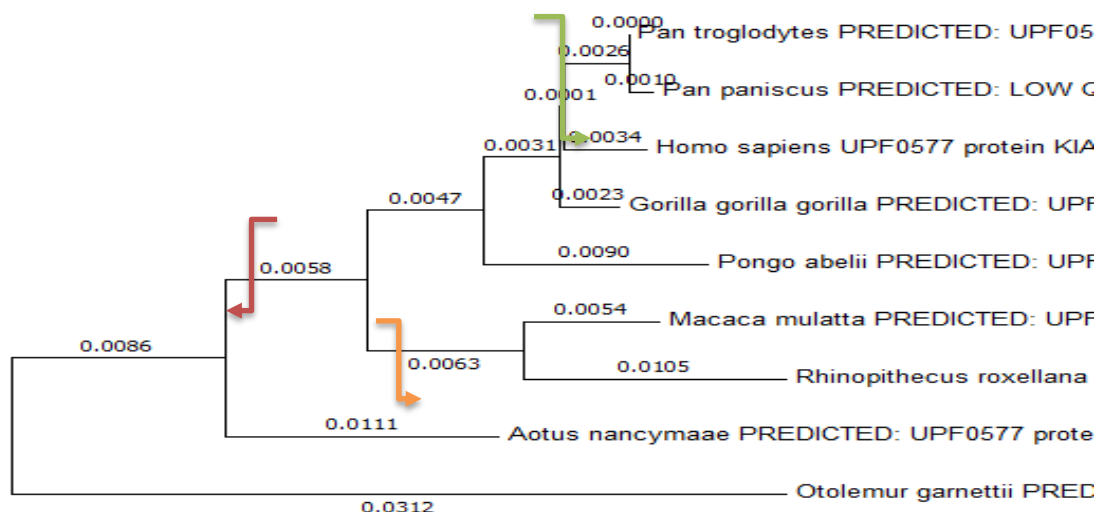


Slika 27. MEGA: UPGMA stablo nastalo iz matrice sa slike 25

Slike 26 i 27 predstavljaju ukorijenjena stabla nastala UPGMA metodom. Vidljivo je da je topologija oba filogenetska stabala ista, a iz razloga što se duljina grana kod ove metode ne uzima u obzir, može se reći da su stabla identična.



Slika 28. Program: stablo nastalo iz matrice sa slike 24 metodom povezivanja susjeda



Slika 29. MEGA: stablo nastalo iz matrice sa slike 25 metodom povezivanja susjeda

Slike 28 i 29 predstavljaju neukorijenjena stabla nastala metodom povezivanja susjeda, koja je složenija UPGMA metode i duljine grana se uzimaju u obzir. Algoritmi crtanja stabala Programa i MEGA-e nisu isti, pa stabla na prvi pogled izgledaju različito. Međutim, ukoliko se bolje pogleda, uočljivo je da je na oba stabla *Otolemur_garnettii* najudaljeniji čvor, a od najsirodnijeg taksona (*Aotus_nancymae*) udaljen je za 0.05. Parovi taksona koji su međusobno najsirodniji su isti na oba filogenetska stabla, odnosno *Pan_troglodytes* i *Pan_paniscus*, i *Macaca_mulatta* i *Rhinopithecus_roxellana*. Jednaka podstabla zbog lakšeg snalaženja označena su strelicama istih boja.

Udaljenost između *Otolemur_garnettii* i *Pongo_abelii* prema stablu na slici 28 iznosi 0,00595, a prema stablu slici 29 iznosi 0,00593. Tablica 8 prikazuje udaljenosti između nasumično odabranih taksona sa slike 28 u usporedbi sa udaljenostima sa slike 29.

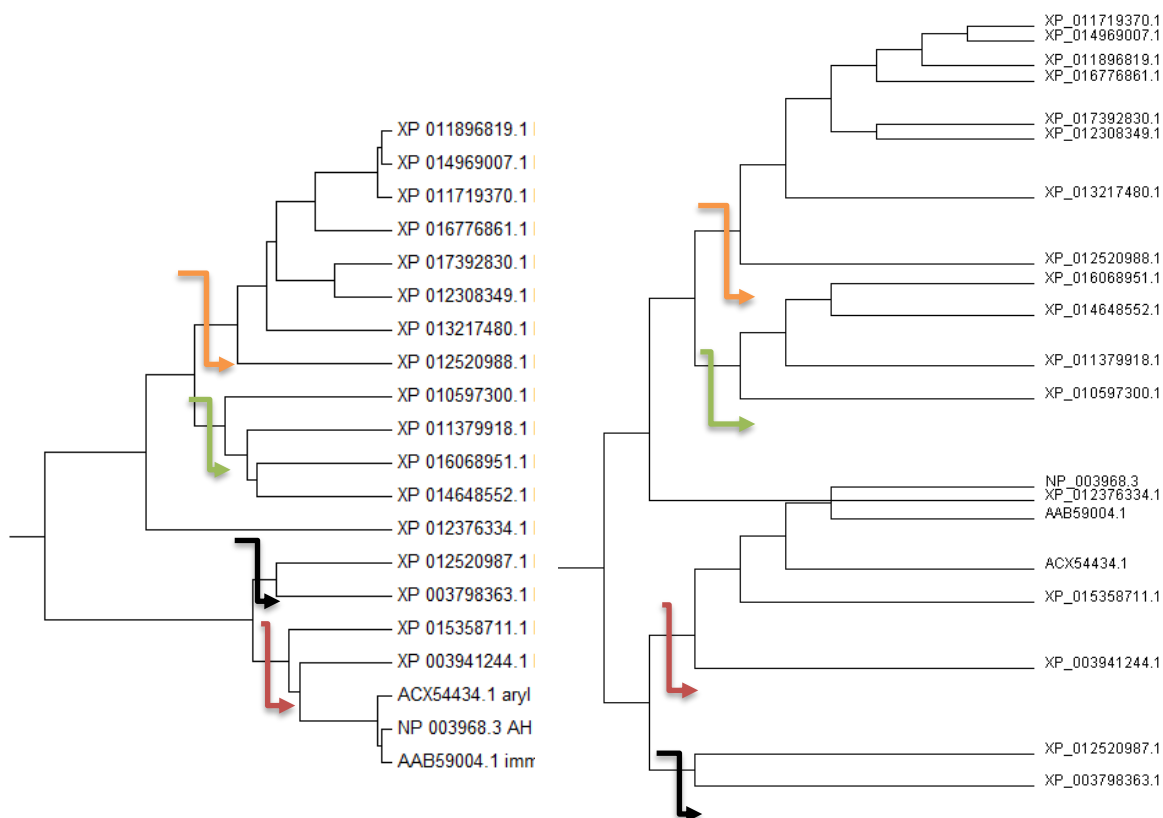
Tablica 8 - Usporedba udaljenosti taksona sa slika 28 i 29

<i>(Takson A,Takson B)</i>	<i>Program (slika 28)</i>	<i>MEGA (slika 29)</i>
<i>(Pan_paniscus, Pongo_abelii)</i>	0,0151	0,0158
<i>(Homo_sapiens, Macaca mulatta)</i>	0,0230	0,0230
<i>(Aotus nancymae, Gorila gorilla gorilla)</i>	0,0279	0,0270
<i>(Gorilla gorilla gorilla, Homo sapiens)</i>	0,0052	0,0058

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1: XP_012520987.1	0.0000	0.0402	0.0434	0.0561	0.0561	0.0529	0.0561	0.1431	0.1223	0.1466	0.1292	0.1258	0.0690	0.1155	0.1258	0.1087	0.1223	0.1155	0.1121	0.1121
2: XP_003798363.1	0.0402	0.0000	0.0339	0.0465	0.0434	0.0402	0.0434	0.1361	0.1155	0.1361	0.1189	0.1155	0.1020	0.1087	0.1189	0.1020	0.1121	0.1087	0.1054	0.1054
3: XP_015358711.1	0.0434	0.0339	0.0000	0.0370	0.0339	0.0308	0.0339	0.1327	0.1121	0.1292	0.1155	0.1189	0.1087	0.1020	0.1054	0.0722	0.1054	0.0986	0.0953	0.0953
4: XP_003941244.1	0.0561	0.0465	0.0370	0.0000	0.0339	0.0308	0.0339	0.1361	0.1087	0.1361	0.1223	0.1155	0.0690	0.0788	0.1020	0.0986	0.1020	0.0986	0.0986	0.0986
5: ACXS44341	0.0561	0.0434	0.0339	0.0339	0.0000	0.0030	0.0061	0.1466	0.1223	0.1327	0.1258	0.1223	0.0986	0.1087	0.1054	0.0755	0.0953	0.0920	0.0920	0.0920
6: NP_003968.3	0.0529	0.0402	0.0308	0.0308	0.0030	0.0000	0.0030	0.1431	0.1189	0.1431	0.1292	0.1189	0.0953	0.1054	0.1020	0.0722	0.0920	0.0887	0.0887	0.0887
7: AABS90041	0.0561	0.0434	0.0339	0.0339	0.0061	0.0030	0.0000	0.1466	0.1223	0.1327	0.1223	0.1223	0.0986	0.1087	0.1054	0.0755	0.0953	0.0920	0.0920	0.0920
8: XP_012376334.1	0.1431	0.1361	0.1327	0.1361	0.1466	0.1431	0.1466	0.0000	0.0625	0.0920	0.0821	0.0625	0.0854	0.0788	0.0854	0.0722	0.0854	0.0788	0.0821	0.0821
9: XP_010597300.1	0.1223	0.1155	0.1121	0.1087	0.1223	0.1189	0.1223	0.0625	0.0000	0.0625	0.0529	0.0465	0.0638	0.0529	0.0561	0.0465	0.0625	0.0638	0.0593	0.0625
10: XP_011379918.1	0.1466	0.1361	0.1292	0.1361	0.1466	0.1431	0.1466	0.0920	0.0625	0.0000	0.0497	0.0434	0.0854	0.0755	0.0788	0.0690	0.0854	0.0788	0.0722	0.0755
11: XP_016068951.1	0.1292	0.1189	0.1155	0.1223	0.1327	0.1292	0.1327	0.0821	0.0529	0.0497	0.0000	0.0434	0.0690	0.0625	0.0529	0.0690	0.0690	0.0625	0.0625	0.0625
12: XP_014648552.1	0.1258	0.1155	0.1189	0.1189	0.1258	0.1223	0.1223	0.0625	0.0465	0.0434	0.0434	0.0000	0.0593	0.0529	0.0625	0.0561	0.0497	0.0434	0.0465	0.0465
13: XP_012520988.1	0.0690	0.1020	0.1087	0.1155	0.1223	0.1189	0.1223	0.0854	0.0638	0.0854	0.0690	0.0593	0.0000	0.0465	0.0625	0.0465	0.0529	0.0497	0.0434	0.0465
14: XP_017392830.1	0.1155	0.1087	0.1020	0.0690	0.0986	0.0953	0.0986	0.0788	0.0529	0.0755	0.0625	0.0529	0.0465	0.0000	0.0183	0.0402	0.0308	0.0370	0.0308	0.0339
15: XP_012308349.1	0.1258	0.1189	0.1054	0.0788	0.1087	0.1054	0.1087	0.0854	0.0561	0.0788	0.0658	0.0625	0.0625	0.0183	0.0000	0.0497	0.0434	0.0434	0.0370	0.0402
16: XP_013217480.1	0.1087	0.1020	0.0722	0.1020	0.1054	0.1020	0.1054	0.0722	0.0465	0.0690	0.0529	0.0561	0.0465	0.0402	0.0497	0.0000	0.0402	0.0339	0.0370	0.0370
17: XP_016776861.1	0.1223	0.1121	0.1054	0.0986	0.0755	0.0722	0.0755	0.0854	0.0625	0.0854	0.0690	0.0561	0.0529	0.0308	0.0434	0.0402	0.0000	0.0277	0.0214	0.0245
18: XP_0111719370.1	0.1155	0.1087	0.0986	0.1020	0.0953	0.0920	0.0953	0.0854	0.0658	0.0788	0.0690	0.0497	0.0370	0.0434	0.0402	0.0277	0.0000	0.0061	0.0000	0.0030
19: XP_011896819.1	0.1121	0.1054	0.0953	0.0986	0.0920	0.0887	0.0920	0.0788	0.0593	0.0722	0.0625	0.0434	0.0308	0.0370	0.0339	0.0214	0.0061	0.0000	0.0030	0.0030
20: XP_014969007.1	0.1121	0.1054	0.0953	0.0986	0.0920	0.0887	0.0920	0.0821	0.0625	0.0755	0.0658	0.0465	0.0339	0.0402	0.0370	0.0245	0.0030	0.0030	0.0000	0.0000

Silka 30. Program: Matrica udaljenost nastala iz datoteke alligned20x335.fasta

Slika 30 prikazuju matrice udaljenosti za 20 proteinskih sljedova duljine 335 aminokiselina, generiranih iz datoteke aligned20x335.fasta. Proteinski sljedovi su preuzeti sa NCBI-a, a poravnati programom Clustal Omega. Redci matrice su označeni parovima (broj, takson), dok su stupci označeni samo brojevima zbog preglednosti. Svaki broj koji numerira stupac označava takson koji mu je pridružen prilikom numeriranja retka.



Slika 31. Stabla nastala UPGMA metodom, lijevo MEGA, desno Program

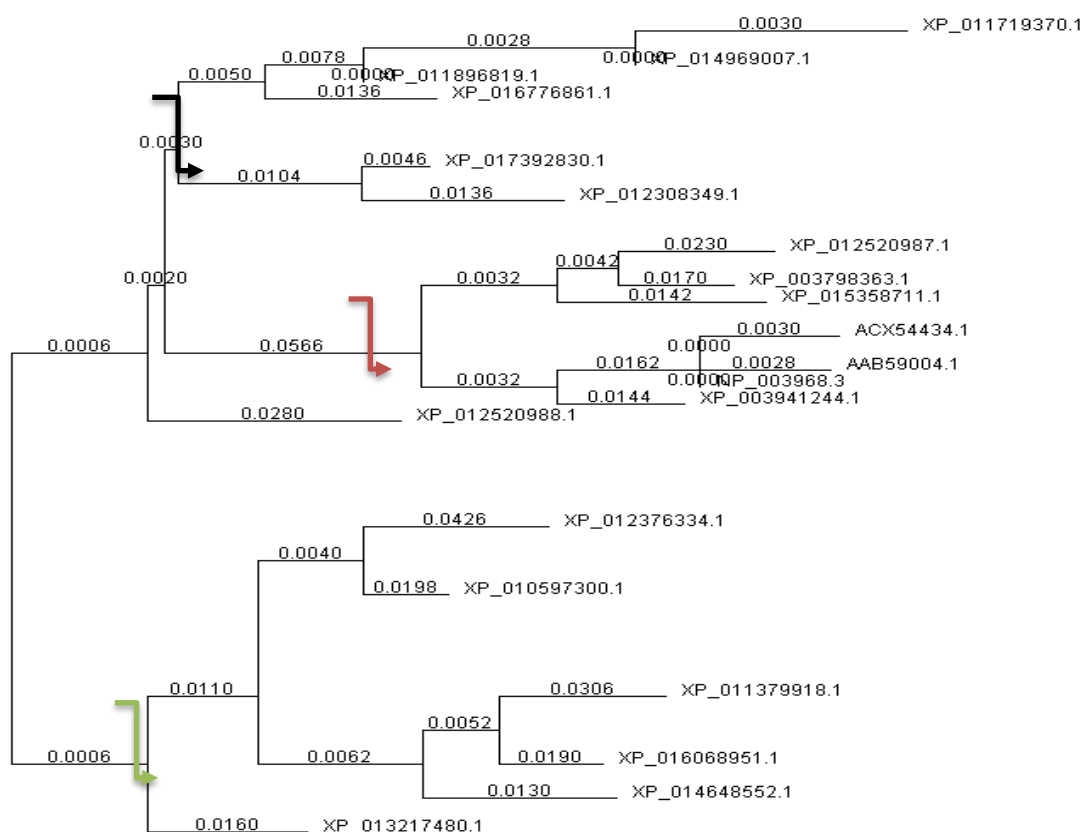
Slika 31 usporedno prikazuje dva filogenetska stabla nastala UPGMA metodom. Stablo na lijevoj strani je izgrađeno uz pomoć MEGA-e, a na desnoj strani uz pomoć Programa.

Taksoni lijevog i desnog stabla nisu poredani istim redoslijedom, međutim parovi najsirođnijih taksona su jednaki. Ti parovi kako ih je nacrtao Program redom su: XP_017392830.1 i XP_012308349.1, XP_01606951.1 i XP_01469552.1, NP_003968.3 i AAB59004.1 i XP_012502987.1 i XP_003798363.1. Jednaka podstabla označena su strelicama istih boja zbog lakšeg snalaženja.

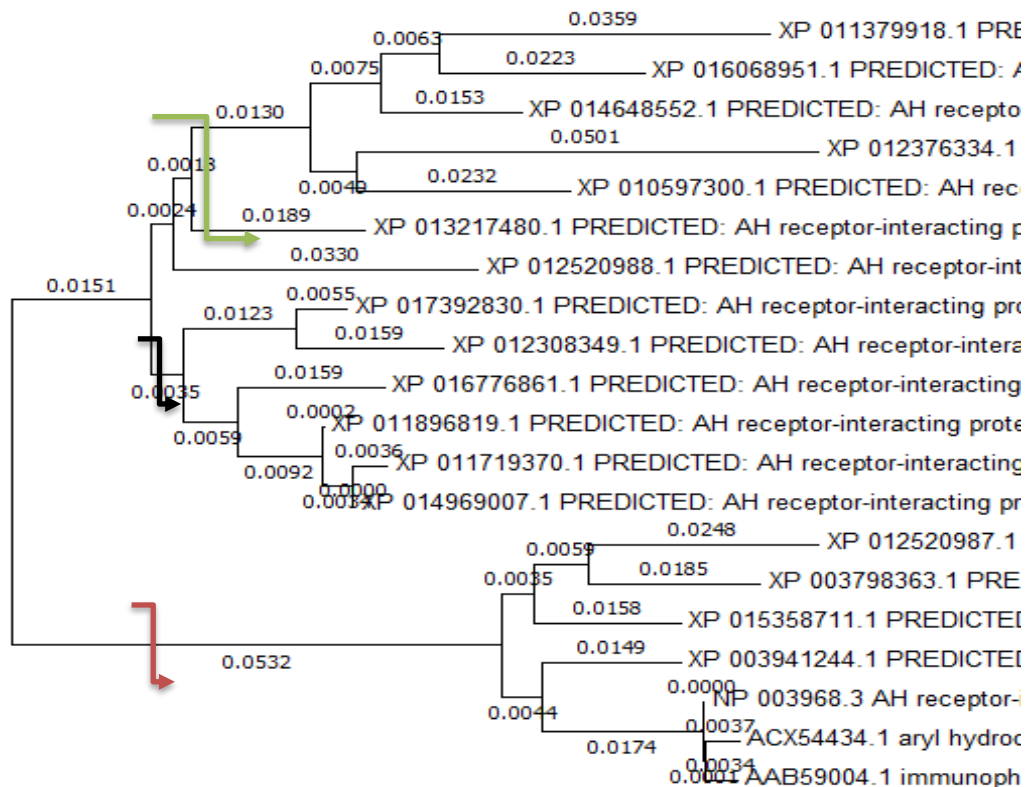
Ukoliko prilikom gradnje filogenetskog stabla postoje više od dva taksona koji imaju jednaku udaljenost Program će kao bliži takson uzeti onaj čije ime abecedno

prethodi. Posljedica toga je da takson XP_014969007.1, koji je jednako udaljen od XP_011719370.1 i XP_011896819.1 (slika 30), spojen sa XP_011719370.1, a zatim je njihov roditelj spojen sa taksonom XP_011896819.1. MEGA koristi drugačiju logiku prilikom izgradnje stabla, pa su prvo spojeni taksoni XP_014969007.1 i XP_011896819.1, a zatim njihov roditelj s XP_011719370.1.

Kod filogenetskih stabala koja imaju veći broj taksona (listova) dolazi do međusobnog presijecanja grana prilikom crtanja. Ovaj problem je izražen samo kod Programa zbog toga što MEGA koristi naprednije algoritme crtanja binarnih stabala. Presijecanje je vidljivo kod taksona NP_003968.3 koji siječe takson XP_012376334.1. Mogućnost pojave presijecanja grana raste s povećanjem dubine stabla.



Slika 32. Program: stablo nastalo metodom povezivanja susjeda



Slika 33. MEGA: stablo nastalo metodom povezivanja susjeda

Slike 32 i 33 predstavljaju neukorijenjena stabla nastala metodom povezivanja susjeda, te se duljine grana uzimaju u razmatranje. Parovi najsirodnijih taksona na filogenetskim stablima su jednaki, a to su prema stablu na slici 32 redom: XP_011719370.1 i XP_014969007.1, XP_017392930.1 i XP_012308349.1, XP_01250987.1 i XP_003798363, ACX54431.1 i AAB59004, XP_012376334.1 i XP_010597300.1 i XP_011379918.1 i XP_016068951.1. Vidljivo je da je topologija stabla jednaka. Jednaka podstabla označena su strelicama istih boja zbog lakšeg snalaženja.

Kod iscrtavanja taksona ACX54431.1 i AAB59004.1 došlo je do preklapanja.

Tablica 7 prikazuje udaljenosti između 10 nasumično odabranih taksona sa slike 32 u usporedbi sa udaljenostima iz matrice sa slike 30.

Tablica 7. Usporedba udaljenosti taksona

(Takson A,Takson B)	Stablo (slika 32)	Matrica udaljenosti (slika 30)	<i>Susjedni taksoni</i>
(XP_010597300.1, XP_012376334.1)	0,0624	0,0625	
(AAB59004.1, ACX54434.1)	0,0058	0,0061	
(XP_011719370.1, XP_016776861.1)	0,0272	0,277	
(XP_012308349.1, XP_016776861.1)	0,0368	0,434	
(XP_012376334.1, XP_014648552.1)	0,0658	0,625	
(XP_012520988.1, XP_010597300.1)	0,0634	0,658	
(XP_010597300.1, XP_013217480.1)	0,0508	0,465	
(XP_013217480.1, XP_011379918.1)	0,0690	0,690	
(XP_014648552.1, XP_011379918.1)	0,0488	0,434	
(XP_012520987.1, XP_003941244.1)	0,0480	0,561	

Taksonima koji su susjedni u stablu, udaljenosti prikazane na stablu i udaljenosti u matrici su skoro jednake, razlikuju se tek na četvrtoj decimali. Ukoliko taksoni nisu susjedni, udaljenosti prikazane na stablu i u matrici se razlikuju. Ta razlika raste što su taksoni udaljeniji, a posljedica je matematičkog algoritma metode povezivanja susjeda. Svakim dodavanjem novog čvora u stablo, računaju se nove udaljenosti koje povećavaju pogrešku.

8. Zaključak

Filogenetsko stablo ima važnu ulogu u razumijevanju evolucije. Kako bi se utvrdio razvojni slijed kao i grananja u evoluciji pojedinih vrsta, neophodno je ispitivanje različitih vrsta organizama. Danas je poznato da se geni nisu razvijali ravnomjerno. Neki geni koji se danas nalaze kod čovjeka, imaju zajedničkog pretka samo sa čimpanzama, dok se neki drugi pojavljuju kod svih sisavaca.

Konačni izgled filogenetskog stabla ovisi o nekoliko faktora: matrici udaljenosti, metodi kojom se taksoni dodaju u stablo, te konačno o algoritmu koji crta izgrađeno stablo. Udaljenosti u matrici udaljenosti se mogu izračunati na više načina, a u ovom radu je odabran Jukes – Cantorov model iz razloga što je jedan od prvih i ujedno najjednostavnijih modela evolucije. Metode koje su odabrane za izgradnju filogenetskog stabla su UPGMA i metoda povezivanja susjeda. UPGMA metoda je odabrana zbog toga što je jedna od prvih metoda za izgradnju stabla, dok je razlog za odabir metode povezivanja susjeda to što je preciznija i suvremenija od UPGMA.

Ukoliko stablo postane previše razgranato prilikom crtanja dolazi do presijecanja grana, te tada topologija stabla prestaje biti jasno vidljiva. U svrhu unaprjeđenja programske potpore treba proširiti model crtanje stabala s velikim brojem taksona. Rješenje prethodno navedenog problema i provođenje dodatnih poboljšanja, poput računanja matrice udaljenosti na više načina, pretvorbe stabla u Newick format (Cayley A., 1857) i dodavanja novih metoda izgradnje stabla, ostavljam za rad u budućnosti.

9. Literatura

- Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. "Phylogenetic Analysis. Models and Estimation Procedures." *American Journal of Human Genetics* 19 (3 Pt 1): 233–57.
- "Clustal Omega < Multiple Sequence Alignment < EMBL-EBI." 2017. Accessed June 7. <http://www.ebi.ac.uk/Tools/msa/clustalo/>.
- Darwin, Charles, and Julian Huxley. 2003. *The Origin of Species: 150th Anniversary Edition*. Rep Anv edition. Signet.
- "FASTA Format." 2017. Accessed May 26. http://www.bioinformatics.nl/tools/crab_fasta.html.
- Felsenstein, Joseph. 1981. "Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach." *Journal of Molecular Evolution* 17 (6): 368–76. doi:10.1007/BF01734359.
- Hasegawa, Masami, Hirohisa Kishino, and Taka-aki Yano. 1985. "Dating of the Human-Ape Splitting by a Molecular Clock of Mitochondrial DNA." *Journal of Molecular Evolution* 22 (2): 160–74. doi:10.1007/BF02101694.
- Information, National Center for Biotechnology, U. S. National Library of Medicine 8600 Rockville Pike, Bethesda MD, and 20894 Usa. 2017. "National Center for Biotechnology Information." Accessed June 7. <https://www.ncbi.nlm.nih.gov/>.
- Kimura, Motoo. 1980. "A Simple Method for Estimating Evolutionary Rates of Base Substitutions through Comparative Studies of Nucleotide Sequences." *Journal of Molecular Evolution* 16 (2): 111–20. doi:10.1007/BF01731581.
- Kishino, Hirohisa, and Masami Hasegawa. 1989. "Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea." *Journal of Molecular Evolution* 29 (2): 170–79. doi:10.1007/BF02100115.
- Kumar, Sudhir, Glen Stecher, and Koichiro Tamura. 2016. "MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets." *Molecular Biology and Evolution* 33 (7): 1870–74. doi:10.1093/molbev/msw054.
- Linné, Carl von, and Lars Salvius. 1753. *Caroli Linnaei ... Species Plantarum :Exhibentes Plantas Rite Cognitas, Ad Genera Relatas, Cum Differentiis Specificis, Nominibus Trivialibus, Synonymis Selectis, Locis Natalibus, Secundum Systema Sexuale Digestas...* Vol. 1. Holmiae : Impensis Laurentii Salvii,. <http://www.biodiversitylibrary.org/item/13829>.
- Morgan, Gregory J. 1998. "Emile Zuckerkandl, Linus Pauling, and the Molecular Evolutionary Clock, 1959–1965." *Journal of the History of Biology* 31 (2):

155–78. doi:10.1023/A:1004394418084.

Pearl, Judea. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Saitou, N., and M. Nei. 1987. "The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees." *Molecular Biology and Evolution* 4 (4): 406–25. doi:10.1093/oxfordjournals.molbev.a040454.

Shin, Chan-Su, Sung Kwon Kim, Sung-Ho Kim, and Kyung-Yong Chwa. 1998. "Algorithms for Drawing Binary Trees in the Plane." *Information Processing Letters* 66 (3): 133–39. doi:10.1016/S0020-0190(98)00049-0.

Sokal, RR, and CD Michener. 1958. "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Scientific Bulletin* 28: 1409–38.

Šikić, M. and M. Domazet-Lošo, "Bioinformatika." 2017. Accessed June 1. https://www.fer.unizg.hr/download/repository/bioinformatika_skripta_v1.2.pdf.

Tamura, K., and M. Nei. 1993. "Estimation of the Number of Nucleotide Substitutions in the Control Region of Mitochondrial DNA in Humans and Chimpanzees." *Molecular Biology and Evolution* 10 (3): 512–26.

"The Newick Tree Format." 2017. Accessed June 8. <http://evolution.genetics.washington.edu/phylip/newicktree.html>.

Whelan, Simon, and Nick Goldman. 2001. "A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach." *Molecular Biology and Evolution* 18 (5): 691–99. doi:10.1093/oxfordjournals.molbev.a003851.

Naslov, sažetak i ključne riječi

Završni rad:

Izgradnja filogenetskog stabla korištenjem metoda udaljenosti

Sažetak

Izgradnja filogenetskog stabla kreće izračunom matrice udaljenosti. Udaljenosti se računaju po Jukes - Cantorovom modelu tako da se kao ulaz programa predaju poravnati genomi ili proteinski lanci u FASTA formatu. Nakon što je iz ulazne datoteke kreirana matrica udaljenosti, na temelju tih udaljenosti programska potpora omogućuje prikaz: matrice udaljenosti, filogenetskog stabla izgrađenog UPGMA metodom ili stabla izgrađenog metodom povezivanja susjeda.

Ključne riječi: filogenetsko stablo , UGPMA, metoda povezivanja susjeda, matrica udaljenosti, genom, protein, takson

Bachelor thesis:

Distance-Based Phylogenetic Tree Construction Methods

Abstract

Phylogenetic tree construction starts with computation of the distance matrix. Distances are based on Juke – Cantor's model of evolution. Input of the program are aligned genomes or protein sequences in FASTA format. When the distance matrix is computed, the program will construct phylogenetic tree using UPGMA algorithm or neighbor joining method.

Key words: phylogenetic tree, UGPMA, neighbor joining method, distance matrix, gene, protein, taxon