Bioinformatika

30. april 2020..

Glava 1

Filogenetska stabla

U ovom poglavlju se bavimo rekonstrukcijom filogenetskih (evolutivnih) stabala. Dva su biološka primera koja ćemo posmatrati. U prvom problemu se bavimo utvrđivanjem kako je virus SARS prešao na čoveka pri čemu razmatramo inicijalni virus SARS mada bi se analogno metodologija mogla primeniti i na aktuelni SARS-CoV2 (inače oba virusa pripadaju grupi takozvanih koronavirusa). Drugi problem koji razmatramo je utvrđivanje od kojih su predaka potekle aktuelne vrste i kako su one povezane u filogenetsko stablo.

1.1 Matrice rastojanja i evolutivna stabla

1.1.1 Konstrukcija matrice rastojanja na osnovu višestrukog poravnanja

Da bi odredili kako je SARS prešao sa životinja na ljude, naučnici su počeli da sekvenciraju genome koronavirusa iz različitih organizama. Ideja je bila da višestrukim poravnanjem uzoraka iz različitih organizama utvrdi između kojih uzoraka je najveća sličnost što bi sugerisalo da je između ta dva organizma došlo do prelaska. Konstrukcija višestrukog poravnanja za cele genome se pokazala teškom jer se geni unutar viralnih genoma često preuređuju i sadrže mnogo insercija i delecija. Zbog toga su se naučnici fokusirali samo na jedan od ukupno 6 gena koji postoje u genomu SARS-a, i to na gen koji kodira takozvani *Spike* protein, zadužen za vezivanje virusa za domaćina.

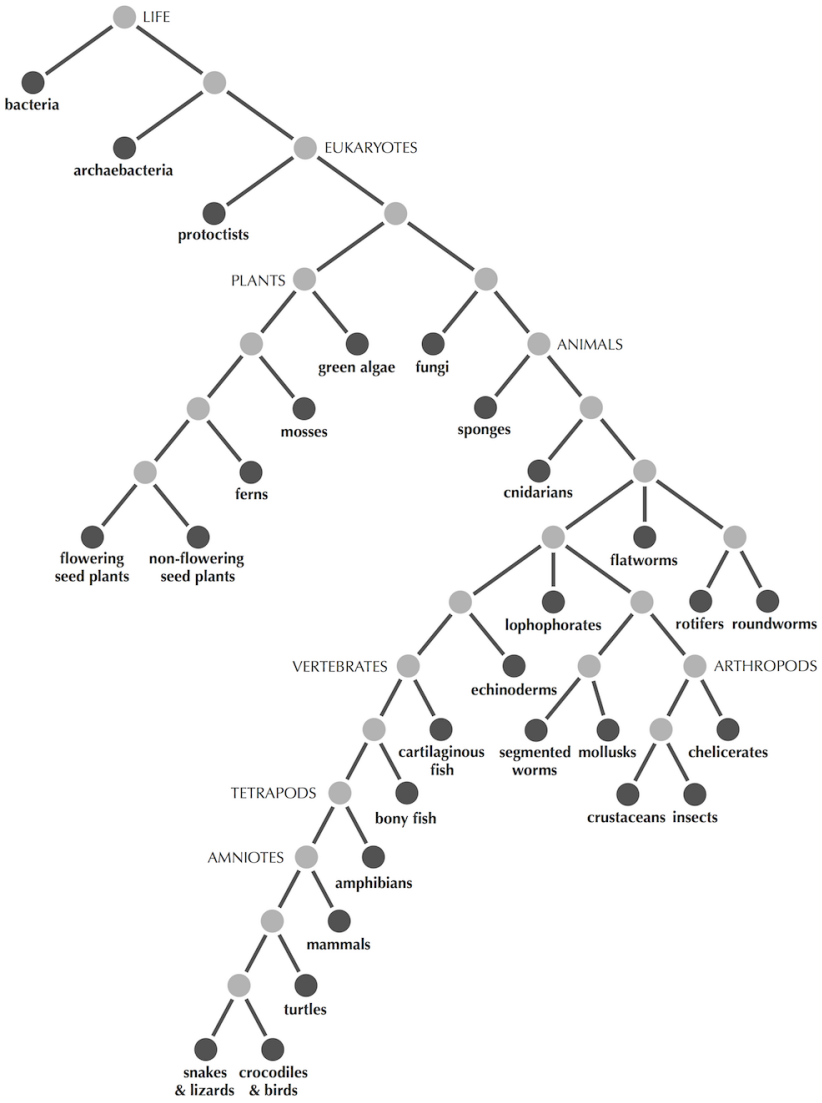
Prisetimo se da poravnanje n sekvenci predstavlja matricu od n vrsta gde se u svakoj koloni nalazi simbol neke sekvence ili praznina (simbol $-$). Na osnovu ovakve matrice poravnanja moguće je konstruisati *matricu rastojanja* dimenzije $n \times n$ u oznaci D. S obzirom da se ovom matricom definiše rastojanje, funkcija $D_{i,j}$ koja određuje vrednosti matrice mora biti metrika (da bude nenegativna, simetrična i da zadovoljava nejednakost trougla). Postoji više primera takvih funkcija, pa tako $D_{i,j}$ može biti broj pozicija u matrici poravnanja na kojima se razlikuju sekvence i i j (takozvano *edit* rastojanje), zatim rastojanje 2-prekida i slično. Izbor funkcije zavisi od primene same matrice rastojanja. Na slici 1.1 je dat primer konstrukcije matrice rastojanja gde je kao mera sličnosti korišćeno edit rastojanje.

1.1.2 Grafovska reprezentacija evolutivnih stabala

Evoluciju modelujemo stablima i takva stabla nazivamo evolutivnim ili filogenetskim. U evolutivnom stablu listovi predstavljaju današnje vrste, dok unutrašnji čvorovi predstavljaju izumrle vrste. Koreni čvor predstavlja najdaljeg zajedničkog predaka. Na slici 1.2 se može videti primer evolutivnog stabla koje prikazuje evoluciju života na Zemlji (takozvano drvo života, *tree of life*).

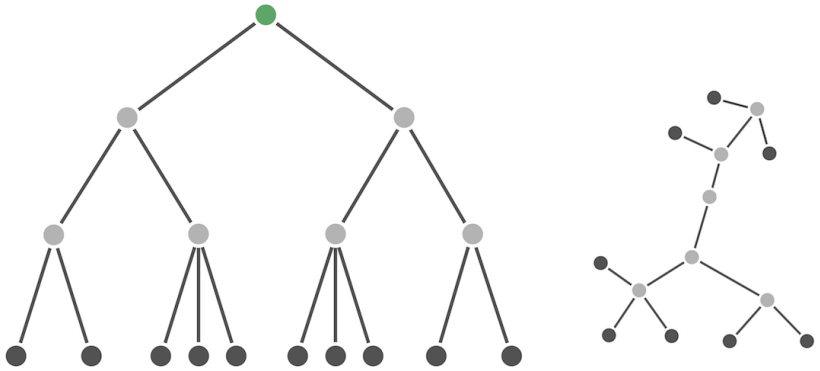
Vrsta	Poravnanje	Matrica rastojanja			
		Šimpanza	Čovek	Foka	Kit
Šimpanza	ACGTAGGCCT	0	3	6	4
Čovek	ATGTAAGACT	3	0	7	5
Foka	TCGAGAGCAC	6	7	0	2
Kit	TCGAAAGCAT	4	5	2	0

Slika 1.1: Prikaz konstrukcije matrice rastojanja



Slika 1.2: Evolutivno stablo živog sveta na Zemlji

Evolutivna stabla se mogu predstavljati kao korena ili kao nekorena stabla (1.3). Kod korenih stabala, grane su implicitno orijentisane *od* korena. Ovakva orijentacija modeluje vreme jer se u korenu nalazi predak svih vrsta a evolucija teče od korena ka listovima. Kod nekorenih stabala nemamo koren pa tako ni pretpostavku o jedinstvenom zajedničkom pretku. U nastavku ćemo se baviti algoritmima za rekonstrukciju i korenih i nekorenih evolutivnih stabala.



Slika 1.3: Korena i nekorena stabla

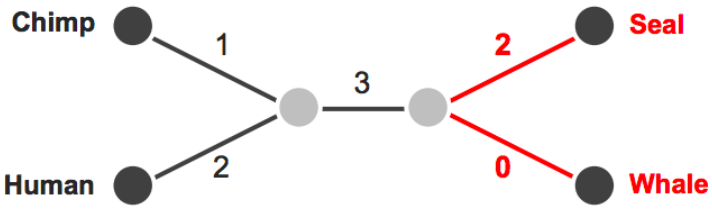
Podsetimo se da stablo predstavlja povezani aciklički graf. Navodimo bez dokaza nekoliko svojstava za ovakve grafove koja će biti od koristi u kasnijim razmatranjima:

- Svako stablo sa bar dva čvora sadrži bar dva lista.
- Svako stablo sa n čvorova sadrži tačno $n - 1$ grana.
- Za proizvoljna dva čvora postoji tačno jedna putanja koja ih povezuje.

1.1.3 Rekonstrukcija filogenetskih stabala na osnovu rastojanja

Prvo ćemo se fokusirati na izvođenje nekorenog stabla iz matrice rastojanja. Neka je data jedna matrica rastojanja D čije su vrste obeležene vrstama organizama i jedno filogenetsko stablo T čiji su listovi označeni istim vrstama organizama kao u matrici rastojanja a granama su dodeljene nenegativne težine. Rastojanje između dva lista u filogenetskom stablu D definišemo kao sumu težina grana na putanji između njih. Kažemo da stablo T *odgovara* matrici D ako za svaki par listova (i, j) važi da je rastojanje između njih jednako $D_{i,j}$ (slika 1.4).

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



Slika 1.4: Prikaz matrice rastojanja i njoj odgovarajućeg evolutivnog stabla

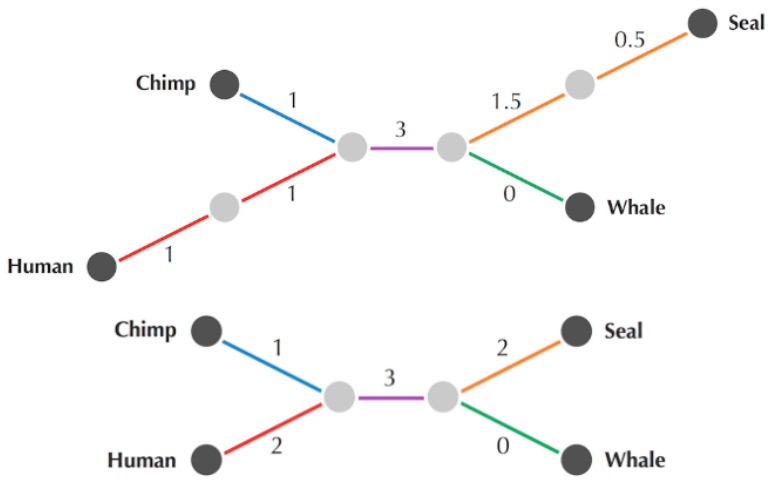
Za dato filogenetsko stablo, lako je rekonstruisati matricu rastojanja kojoj odgovara. U ovom poglavlju ćemo se baviti obrnutim pitanjem: kako na osnovu date matrice rastojanja konstruisati filogenetsko stablo koje odgovara toj matrici. U vezi sa tim, možemo se zapitati sledeće:

1. Da li postoji više filogenetskih stabala koji odgovaraju istoj matrici rastojanja?

2. Da li za svaku matricu rastojanja postoji filogenetsko stablo koje joj odgovara?

Odgovor na prvo pitanje je pozitivan. Na slici 1.5 imamo primer dva filogenetska stabla koja odgovaraju istoj matrici rastojanja. Možemo primetiti da su stabla slične topologije i da bi bilo korisno na neki način definisati kanonsku topologiju filogenetskog stabla koje odgovara datoj matrici rastojanja. Zbog toga uvodimo pojam *prostog* stabla. Prosto stablo je stablo koje ne sadrži čvorove stepena dva. Zbog toga ćemo u nastavku podrazumevati potragu za prostim filogenetskim stablom za datu matricu rastojanja.

Species	Alignment	Distance Matrix			
		Chimp	**Human**	**Seal**	**Whale**
Chimp	ACGTAGGCCT	0	3	6	4
Human	ATGTAAGACT	3	0	7	5
Seal	TCGAGAGCAC	6	7	0	2
Whale	TCGAAAGCAT	4	5	2	0



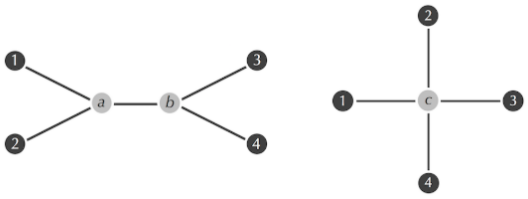
Slika 1.5: Dva različita evolutivna stabla koja odgovaraju istoj matrici rastojanja

Odgovor na drugo pitanje je negativan za matrice dimenzije veće od 3. Primer matrice za koju ne postoji odgovarajuće filogenetsko stablo je prikazan na slici 1.6.

	v_1	v_2	v_3	v_4
v_1	0	3	4	3
v_2	3	0	4	5
v_3	4	4	0	2
v_4	3	5	2	0

Slika 1.6: Neaditivna matrica rastojanja

Naime, prosto stablo sa 4 lista može imati jednu od dve topologije sa slike 1.7 (pri čemu kod levog stablo listovi mogu zameniti mesta i tako dobijamo još dva stabla) i ne možemo dodeliti težine granama stabla tako dobijeno filogenetsko stablo odgovara datoj matrici. Možemo zaključiti da se za neke matrice rastojanja može konstruisati filogenetsko stablo i takve ćemo nazivati *aditivnim*, a za neke ne. Termin aditivnosti potiče od definicije rastojanja između dva lista kod filogenetskih stabala: to je zbir težina grana na putanji između njih koji je jednak odgovarajućem polju u matrici rastojanja.

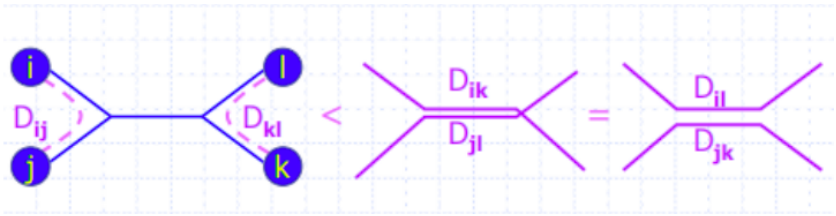


Slika 1.7: Topologije prostih stabala sa 4 lista

Pod kojim uslovima će matrica rastojanja biti aditivna odnosno kada ćemo za datu matricu rastojanja moći da konstruišemo odgovarajuće filogenetsko stablo? Odgovor na ovo pitanje daje naredna teorema koja se često naziva i *uslovom četiri tačke* (slika 1.8). Teoremu navodimo bez dokaza.

Teorema 1.1. *Matrica rastojanja D je aditivna akko za proizvoljna četiri indeksa i, j i k, l u D važi:*

$$D_{ij} + D_{kl} <= D_{ik} + D_{jl} = D_{il} + D_{jk}$$



Slika 1.8: Uslov četiri tačke

Uslov četiri tačke govori da je matrica aditivna ako za njena proizvoljna 4 čvora i 3 njihova međusobna zbira važi da su dva od tri jednaki a treći manji od maksimalna dva.

Aditivnost matrice obezbeđuje jedinstveno prosto filogenetsko stablo koje joj odgovara:

Teorema 1.2. *Postoji tačno jedno prosto stablo koje odgovara aditivnoj matrici.*

Sada možemo formulisati problem filogeneze na osnovu rastojanja.

Problem filogeneze na osnovu rastojanja Konstruisati evolutivno stablo na osnovu aditivne matrice rastojanja.

Ulaz: Aditivna matrica rastojanja.

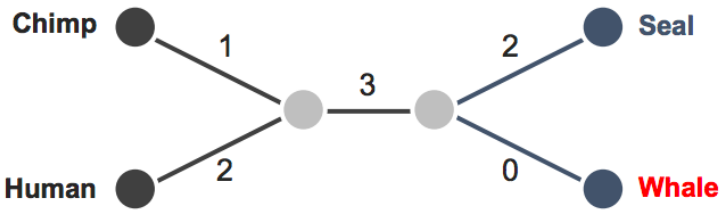
Izlaz: Prosto stablo koje odgovara datoj matrici rastojanja.

1.2 Različiti pristupi rešavanju problema filogeneze na osnovu rastojanja

1.2.1 Pristup preko susednih listova

Na slici 1.9 prikazani su matrica rastojanja i odgovarajuće filogenetsko stablo. U ovom stablu postoje dva para listova koji imaju zajedničkog pretka: $(chimp, human)$ i $(seal, whale)$. Takve listove nazivamo *susednim listovima*.

	Chimp	Human	Seal	Whale
Chimp	0	3	6	4
Human	3	0	7	5
Seal	6	7	0	2
Whale	4	5	2	0



Slika 1.9: Prikaz susednih listova

Primetimo da u ovom primeru minimalna pozitivna vrednost matrice rastojanja odgovara jednom paru susednih listova. Zbog toga ima smisla uvesti ovakvu pretpostavku prilikom konstruišanja filogenetskog stabla na osnovu date matrice rastojanja. Ako uvedemo ovakvu pretpostavku, postavlja se pitanje da li ćemo u prostom stablu uvek imati susedne listove. Odgovor na to daje sledeća teorema.

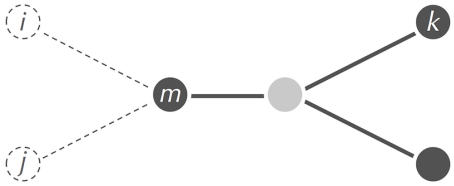
Teorema 1.3. *Za svako prosto stablo sa bar četiri čvora postoji bar jedan par susednih listova.*

Dokaz. Neka je dato stablo T koje sadrži bar četiri čvora i neka je $P = (v_1, \ldots, v_k)$ najduža putanja u tom stablu od svih mogućih putanja. Svako stablo je povezan graf pa k mora biti bar 3. Dalje, v_1 i v_k moraju biti listovi jer bi u suprotnom mogli da produžimo putanju P pa onda ne bi bila najduža. Pošto je T prosto stablo, svaki njegov unutrašnji čvor mora biti stepena bar 3. Stoga, pošto je v_1 list, v_2 mora biti njegov roditelj, unutrašnji čvor stepena bar 3. Čvor v_2 je dakle povezan sa v_1, i mora imati bar još dva susedna čvora. Jedan od njih mora pripadati putanji T i njega označimo sa $v3$ a drugi ne pripada putanji T i njega označimo sa w.

Tvrdimo da je čvor w list čime bi v_1 i w bili susedni listovi jer dele istog roditelja v_2. Pretpostavimo suprotno, da w nije list već unutrašnji čvor. Ako je w unutrašnji čvor prostog stabla T, on mora biti povezan sa bar tri čvora od kojih je jedan v_2 a preostale možemo označiti su u i x. To znači da u stablu T postoji putanja $P' = (u, w, v_2, \ldots, v_k)$ dužine $k + 1$ što dovodi do kontradikcije da polazna pretpostavka da je P putanja maksimalne dužine nije tačna. Time smo dokazali da je w list i da su v_1 i w susedni listovi. ■

Uočimo rekurzivnu prirodu ovog problema. Da bismo smanjili dimenziju matrice, treba ukloniti neke njene redove i kolone. Svaki red u matrici predstavlja jedan list u filogenetskom stablu pa uklanjanje redova/kolona odgovara uklanjanju listova iz stabla. Ako uklonimo sve listove nekog

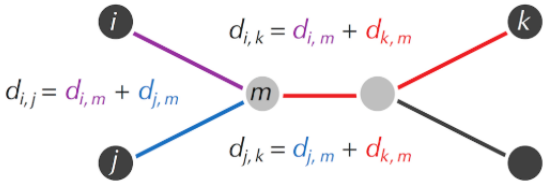
unutrašnjeg cvora, taj unutrašnji cvor ce postati list i rastojanja između njega i ostalih listova bi morala da se nađu u matrici rastojanja (slika 1.10).



Slika 1.10: Brisanje listova iz filogenetskog stabla

Tako bismo iz stabla uklonili dva susedna lista a dodali jedan novi list (njihovog zajedničkog roditelja) a iz matrice rastojanja obrisali redove i kolone koji odgovaraju uklonjenim listovima a dodali jedan novi red i jednu novu kolonu koja odgovara novom listu. Koje će vrednosti imati nova vrsta/kolona u matrici rastojanja? Razmotrimo rastojanja novog lista od listova preostalih u matrici rastojanja. Posmatrajmo stablo od bar četiri čvora, dva susedna lista (i, j) sa zajedničkim pretkom unutrašnjim čvorom m i jedan list k koji im nije susedan. Slika 1.11 prikazuje računanje rastojanja između m i k. Kako su j, k i l listovi, rastojanja između njih su poznata iz matrice rastojanja što znači da možemo izračunati rastojanje između proizvoljnog unutrašnjeg čvora i nesusednog lista.

$$d_{k,m} = \frac{(d_{i,m} + d_{k,m}) + (d_{j,m} + d_{k,m}) - (d_{i,m} + d_{j,m})}{2} = \frac{d_{i,k} + d_{j,k} - d_{i,j}}{2}$$



Slika 1.11: Rastojanja između čvorova u filogenetskom stablu

Prethodna razmatranja se mogu objediniti u sledeći rekurzivni algoritam za rešavanje problema filogeneze na osnovu rastojanja:

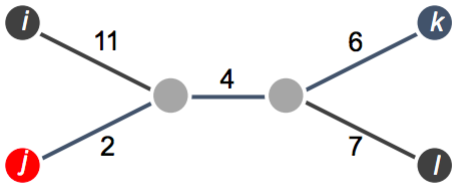
- Naći par susednih listova kojima odgovara minimalno pozitivno rastojanje iz matrice rastojanja
- Ukoliti redove i i j iz matrice i dodati red koji bi odgovarao njihovom zajedničkom roditelju. Rastojanja od ostalih listova izračunati kao na slici 1.11
- Rekurzivno rešiti problem za matricu manje dimenzije
- Dodati čvorove i i j na stablo dobijeno rekurzivnim korakom

1.2.2 Pristup preko spoljnih grana

Ako bismo pokušali pristupom preko susednih listova da konstruišemo filogenetsko stablo za matricu sa slike 1.12, ne bismo u tome uspeli. Ipak, ova matrica je aditivna i ima odgovarajuće

filogenetsko stablo prikazano ispod. Možemo uočiti da u njenom filogenetskom stablu minimalna vrednost matrice rastojanja *ne odgovara* susednim listovima.

	i	*j*	*k*	*l*
i	0	13	21	22
j	13	0	12	13
k	21	12	0	13
l	22	13	13	0

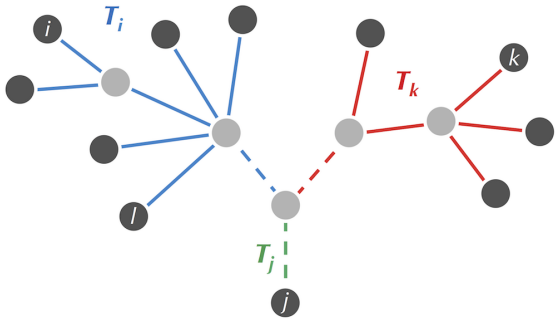


Slika 1.12: Primer matrice za koju ne možemo primeniti prethodni pristup

Ovaj primer pokazuje da je potrebno promeniti polaznu pretpostavku da minimalna nenula vrednost matrice rastojanja odgovara rastojanju između susednih listova. Uz to, treba da nađemo i drugačiji način za smanjivanje dimenzije problema. Umesto uklanjanja susednih listova iz stabla (što bi u matrici rastojanja podrazumevalo uklanjanje vrste i kolone u kojoj se nalazi minimum), drugi pristup podrazumeva uklanjanje jednog lista. Nakon rešavanja problema manje dimenzije, uklonjeni list ćemo vratiti na rekonstruisano stablo i tako dobiti stalo koje odgovara polaznom problemu. Da bismo mogli da dodamo list na stablo, razmotrimo kako da izračunamo dužinu grane koja ga povezuje sa roditeljskim čvorom. Takve grane (na čijem je jednom kraju list) nazivamo *spoljnim granama* (eng. *limb*) dok ostale grane nazivamo unutrašnjim.

Računanje dužine spoljnih grana

Da bismo izračunali dužinu spoljne grane do lista j ($LimbLength(j)$), primetimo sledeće: pošto je stablo $Tree(D)$ koje odgovara matrici rastojanja D prosto, onda je svaki njegov unutrašnji čvor, pa tako i roditeljski čvor datog lista $Parent(j)$, stepena bar tri. Tako $Parent(j)$ particioniše skup listova stabla $Tree(D)$ na bar tri podskupa koja odgovaraju podstablima koja bi nastala kada bi se uklonio $Parent(j)$ (slika 1.13).



Slika 1.13: Prosto stablo u kom su neki od listova i, j, k i l. Listovi i i l pripadaju podstablu T_i, list j podstablu T_j čiji je jedini čvor a list k stablu T_k

Teorema 1.4. ***O dužini spoljnih grana.*** *LimbLength(j) je jednako minimalnoj vrednosti* $(D_{i,j} + D_{j,k} - D_{i,k})/2$ *po svim parovima listovima i i k.*

Dokaz. Par listova može da pripada istom podstablu ili različitim podstablima pa u odnosu na to razlikujemo dva slučaja.

Slučaj 1. Listovi pripadaju različitim podstablima. Označimo ih sa i i k (slika 1.13). S obzirom da je $Parent(j)$ na putanji koja povezuje i i k, važi sledeće:

$$d_{i,j} = d_{i,Parent(j)} + LimbLength(j)$$

$$d_{j,k} = d_{k,Parent(j)} + LimbLength(j)$$

Kada saberemo ove jednačine, dobijamo:

$$d_{i,j} + d_{j,k} = d_{i,Parent(j)} + d_{k,Parent(j)} + 2 \cdot LimbLength(j)$$

S obzirom da je $d_{i,Parent(j)} + d_{k,Parent(j)} = d_{i,k}$, važi sledeće:

$$LimbLength(j) = \frac{d_{i,j} + d_{j,k} - d_{i,k}}{2} = \frac{D_{i,j} + D_{j,k} - D_{i,k}}{2}$$

Slučaj 2. Listovi pripadaju istom podstablu. Označimo ih sa i i l (slika 1.13). U ovom slučaju čvor $Parent(j)$ pa važi sledeća nejednakost:

$$d_{i,Parent(j)} + d_{l,Parent(j)} \geq d_{i,l}$$

Kao i u slučaju 1, važi:

$$d_{i,j} + d_{j,l} = d_{i,Parent(j)} + d_{l,Parent(j)} + 2 \cdot LimbLength(j)$$

Kombinovanjem ova dva izraza dobijamo:

$$LimbLength(j) = \frac{d_{i,j} + d_{j,l} - (d_{i,Parent(j)} + d_{l,Parent(j)})}{2} \leq \frac{d_{i,j} + d_{j,l} - d_{i,l}}{2} = \frac{D_{i,j} + D_{j,l} - D_{i,l}}{2}$$

Pokazali smo da $LimbLength(j)$ mora biti manje ili jednako od $(D_{i,j} + D_{j,k} - D_{i,k})/2$ za svaki par listova (i, k) što je ekvivalentno sa tvrđenjem da je $LimbLength(j)$ jednako minimalnoj vrednosti $(D_{i,j} + D_{j,k} - D_{i,k})/2$ za svaki par listova (i, k). ■

Potkresivanje stabla

Pristup preko spoljnih grana zasnovan je na tome da se dimenzija problema smanjuje uklanjanjem jednog lista filogenetskog stabla, a time i grane koja taj list povezuje sa ostatkom stabla. S obzirom da u trenutku rekonstrukcije stabla ne znamo kako stablo izgleda, već imamo samo matricu rastojanja, potrebno je da efekat uklanjanja jednog lista modelujemo u okviru matrice rastojanja. Skica ovog postupka prikazana je na slici 1.14.

Najpre, pretpostavimo da je poznato $Tree(D)$ i odaberimo proizvoljno jedan njegov list j. Odseći ćemo granu kojom je j zakačen tako što ćemo smanjiti njenu težinu za $LimbLength(j)$ i tako je svesti na nulu. S obzirom da $Tree(D)$ nije poznato, ovo odsecanje ćemo predstaviti unutar matrice rastojanja D tako što ćemo sve elemente matrice u redu j i koloni j, osim dijagonalnog, umanjiti za $LimbLength(j)$. Spoljne grane sa težinom nula zovemo *ogoljenim* spoljnim granama (eng. *bald*), a matricu rastojanja dobijenu opisanim oduzimanjem označićemo sa D^{bald}. Nadalje, pretpostavljamo da je ogoljena grana uklonjena iz stabla a da su iz matrice D^{bald} uklonjene j-ta vrsta i j-ta kolona čime je dobijena matrica $D^{trimmed}$ dimenzije $(n-1) \times (n-1)$. Na taj način, dimenzija polaznog problema je smanjena za jedan. Rekurzivno možemo rekonstruisati stablo $Tree(D)$ na sledeći način:

1. izaberemo proizvoljno list j, izračunamo $LimbLength(j)$ i konstruišemo matricu rastojanja $D^{trimmed}$
2. konstruišemo stablo za problem manje dimenzije, za matricu $D^{trimmed}$
3. identifikujemo mesto u stablu $Tree(D^{trimmed})$ gde treba dodati list j
4. na mesto identifikovano u prethodnom koraku dodati granu dužine $LimbLength(j)$ na čijem je kraju list j, čime je formirano $Tree(D)$

Primetimo da iz rekurzije kod opisanog postupka izlazimo kada matricu svedemo na dimenziju 2×2. Ovakva matrica rastojanja odgovara stablu koje ima dva lista i jednu granu (slika 1.14).

Dodavanje lista u potkresano stablo

Razmotrimo sada korak 3. koji podrazumeva određivanje mesta gde ćemo dodati list koji smo izbacili. List sa stablom možemo povezati samo preko njegovog roditeljskog čvora pa je stoga u ovom koraku zadatak odrediti roditeljski čvor izbačenog lista. Taj čvor se može već nalaziti u stablu, i tada je potrebno identifikovati da je to traženi čvor i dodati granu koja povezuje taj čvor i list (korak 4). Nekada se traženi čvor ne nalazi u stablu i tada ga je potrebno uvesti i pravilno povezati sa ostalim čvorovima stabla. Postupak će biti prikazan na primeru na slici 1.14 u kom izgradnju stabla počinjemo od grane koja spaja listove v_1 i v_2 a onda dodajemo listove v_3 i v_4.

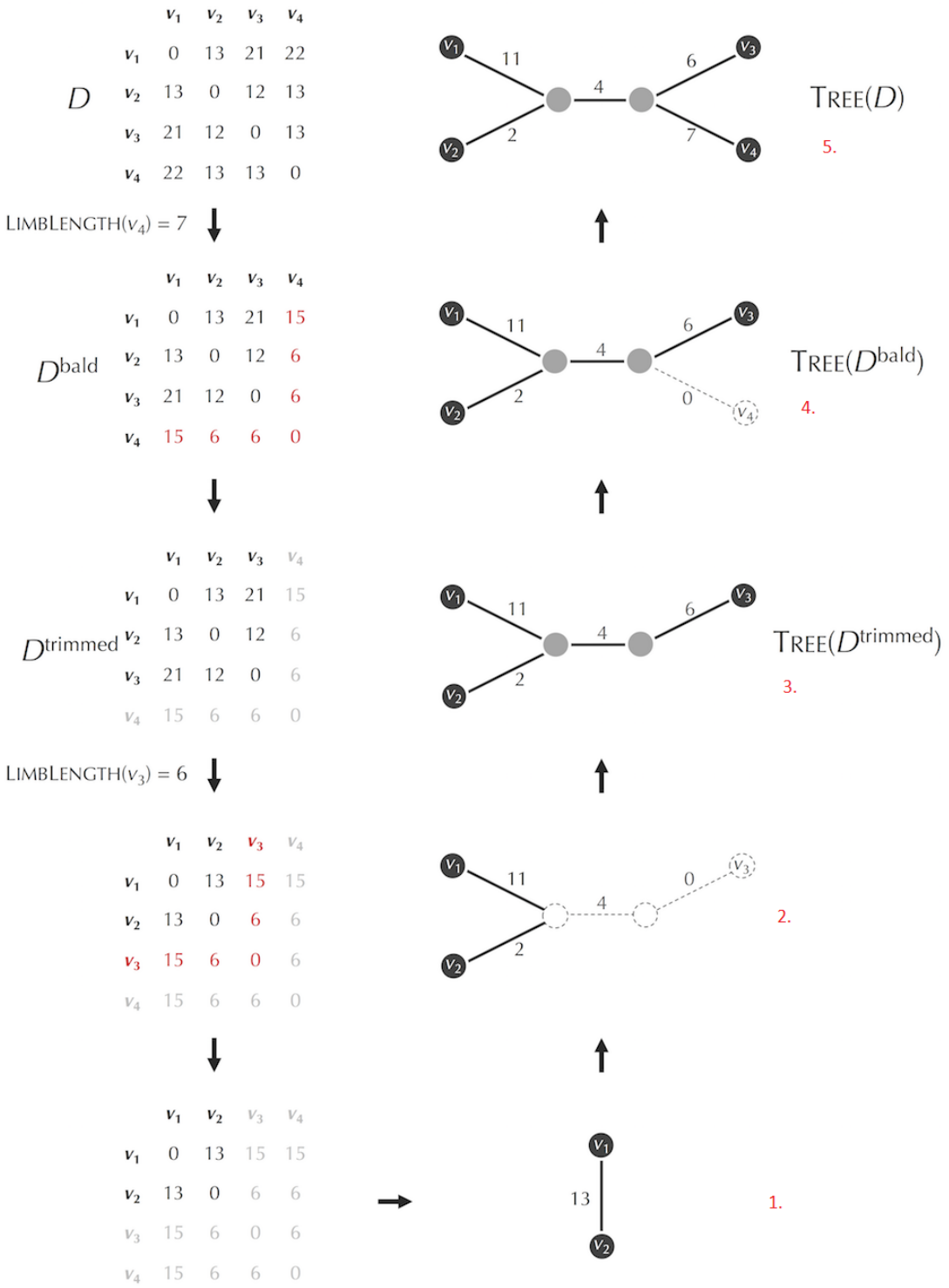
Posmatrajmo stablo $Tree(D^{bald})$ (označeno sa 4.). Ovo stablo je, kao i $Tree(D)$, filogenetsko stablo koje odgovara aditivnoj matrici rastojanja D^{bald} i na njega možemo primeniti teoremu o dužini spoljnih grana za spoljnu granu koja odgovara listu j:

$$LimbLength(j) = \min_{i,k} \frac{D^{bald}_{i,j} + D^{bald}_{j,k} - D^{bald}_{i,k}}{2}$$

S obzirom da je u $Tree(D^{bald})$ dužina spoljne grane koja odgovara listu j jednaka nuli, to znači da postoje listovi (i,k) takvi da važi:

$$\frac{D^{bald}_{i,j} + D^{bald}_{j,k} - D^{bald}_{i,k}}{2} = 0,$$

odnosno



Slika 1.14: Rekontrukcija filogenetskog stabla na osnovu date aditivne matrice rastojanja pristupom preko spoljnih grana

$$D^{bald}_{i,j} + D^{bald}_{j,k} = D^{bald}_{i,k}$$

Tako, mesto dodavanja lista j mora biti na rastojanju $D^{bald}_{i,j}$ od lista i na putanji koja povezuje i i k u potkresanom stablu $Tree(D^{trimmed})$ (označen sa 3.). Pogledajmo zašto je to baš prikazana pozicija (a nije, na primer, list dodat na putanji između v_1 i v_2). Videli smo da jednakost $D^{bald}_{i,j} + D^{bald}_{j,k} = D^{bald}_{i,k}$ važi za jedan par listova. Imamo sledeće opcije:

1. (v_1, v_2) Da bi $Parent(v_4)$ bio dodat na ovoj putanji, treba da važi $D^{bald}_{1,4} + D^{bald}_{4,2} = D^{bald}_{1,2}$, što nije ispunjeno $(15 + 6 \neq 13)$

2. (v_1, v_3) Da bi $Parent(v_4)$ bio dodat na ovoj putanji, treba da važi $D^{bald}_{1,4} + D^{bald}_{4,3} = D^{bald}_{1,3}$, što jeste ispunjeno $(15 + 6 = 21)$

3. (v_2, v_3) Da bi $Parent(v_4)$ bio dodat na ovoj putanji, treba da važi $D^{bald}_{2,4} + D^{bald}_{4,3} = D^{bald}_{2,3}$, što jeste ispunjeno $(6 + 6 = 12)$

U ovom slučaju, uslov je ispunjen za dva para čvorova i te putanje se preklapaju. Unutrašnji čvor $Parent(v_4)$ treba da bude na rastojanju $D^{bald}_{2,4}(= 6)$ od v_2 i $D^{bald}_{4,3}(= 6)$ od v_3. Na tom rastojanju u putanji već postoji unutrašnji čvor za koji su, takođe, odgovarajuća rastojanja i ka čvoru v_2. To znači da je ovaj čvor roditeljski čvor za list v_4 i da na njega treba dodati spoljnu granu težine nula. U narednoj iteraciji (stablo označeno sa 5), težina se povećava na $LimbLength(v_4)(= 7)$ i time se dobija stablo koje odgovara polaznoj matrici rastojanja D.

Razmotrimo i dodavanje lista v_3 (označeno sa 2). U ovom slučaju nema drugog izbora nego da $Parent(v_3)$ bude dodat na putanji između v_1 i v_2. Ta putanja se sastoji od samo jedne grane dužine 13. Prema matrici D^{bald} (korak 2), $D^{bald}(v_1, v_3) = 15$ i stoga se $Parent(v_3)$ ne može naći na sredini grane koja povezuje v_1 i v_2 već se ta grana mora negde prekinuti. Postavlja se pitanje gde tačno. Pošto grane imaju celobrojne težine, moguće pozicije prekida su takve da dodavanje $Parent(v_3)$ proizvede sledeća rastojanja $(v_1, Parent(v_3))$ i $(v_2, Parent(v_3))$: $(1, 12), (2, 11), (3, 10), (4, 9), (5, 8), \ldots, (11, 2), (12, 1)$. Ispitajmo redom počevši od poslednjeg. Ako uzmemo poslednje, $(12, 1)$, potrebno je dodatni granu dužine 3 na roditeljski čvor da bi rastojanje od v_1 do v_3 bilo 15 koliko je u D^{bald}. Istovremeno, ta bi grana morala da bude dužine 5 da bi rastojanje od v_2 do v_3 bilo 6 koliko je u D^{bald}. Jasno je da tako nešto nije moguće. Sledeća podela $(11, 2)$ će omogućiti da dodata grana bude dužine 4 i da se na taj način dobiju odgovarajuća rastojanja iz D^{bald}. Ovde, za razliku od prethodnog slučaja, moramo da dodajemo nove unutrašnje čvorove: jedan koji će podeliti granu (v_1, v_2) i drugi, $Parent(v_3)$, koji će sa jedne strane biti spojen sa tim čvorom a sa druge sa listom v_3 (korak 2). Nakon toga, spoljnoj grani za v_3 dodamo odgovarajuću težinu $LimbLength(v_3)$ izračunatu tokom rekurzivnog postupka.

Opisani algoritam za rekonstrukciju filogenetskog stabla na osnovu rastojanja se naziva algoritam aditivne filogenije (eng. *AdditivePhylogeny*).

1. Izaberemo proizvoljno list, npr. j.
2. Izračunamo dužinu njegove krajnje grane, LimbLength(j).
3.OduzmemoLimbLength(j) od svake grane i dobijemo matricu D^{bald} u kojoj do lista j vodi ogoljena (bold) grana (dužine 0).
4. Uklonimo j-ti red i kolonu iz matrice i dobijemo (n - 1) x (n - 1) matricu D^{trim}.
5. Konstruišemo Tree(D^{trim}).
6. Identifikujemo tačku u Tree(D^{trim}) gde list j treba da se nalazi.
7. Dodamo list j povezujuci ga granom dužine LimbLength(j) kako bismo formirali Tree(D).

1.2.3 Šta ako matrica nije aditivna?

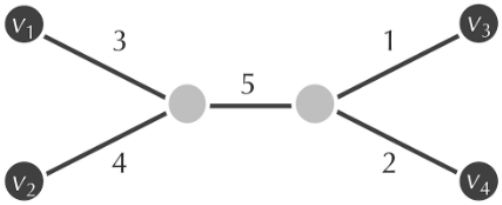
U realnim primenama, retko će se dešavati da je matrica rastojanja aditivna. U slučaju da ovaj uslov nije ispunjen, AdditivePhylogeny ne može rekonstruisati odgovarajuće filogenetsko stablo, tako da su rastojanja između listova u stablu jednaka odgovarajućim vrednostima u matrici rastojanja. Možemo postaviti sledeće pitanje: ako već nije moguće rekonstruisati odgovarajuće filogenetsko stablo, da li je moguće konstruisati filogenetsko stablo sa rastojanjima između listova koja se *minimalno* razlikuju od odgovarajućih vrednosti u matrici rastojanja. Ako bi to bilo moguće, dobijeno stablo ne bi odgovaralo datoj matrici rastojanja već bi predstavljalo *najbolju aproksimaciju* takvog stabla.

Jedan način da se izmeri kvalitet ovakve aproksimacije je preko sume kvadrata greške koju ćemo oznaciti sa *Discrepancy*:

$$Discrepancy(T, D) = \sum_{1 \le i < j \le n} (d_{i,j}(T) - D_{i,j})^2$$

Na slici 1.15 prikazana je neaditivna matrica rastojanja D i jedna aproksimacija njenog filogenetskog stabla sa sumom kvadrata greške:

$$Discrepancy(T, D) = (7-3)^2 + (9-4)^2 + (10-4)^2 + (10-4)^2 + (11-5)^2 + (3-2)^2 = 170$$



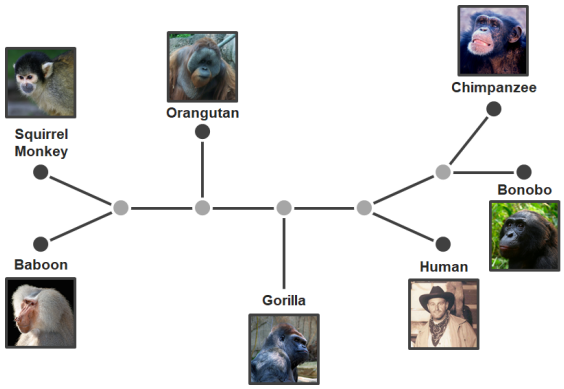
	v_1	v_2	v_3	v_4
v_1	0	3	4	3
v_2	3	0	4	5
v_3	4	4	0	2
v_4	3	5	2	0

Slika 1.15: Neaditivna matrica rastojanja i filogenetsko stablo

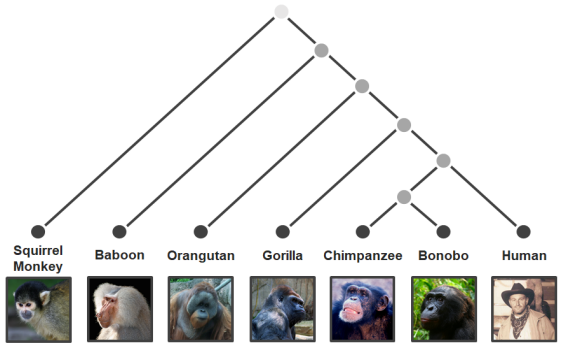
Razmotrimo pitanje kako dodeliti dužine granama u stablu T tako da suma $Discrepancy(T, D)$ bude minimalna. U opštem slučaju, za stablo date topologije postoji algoritam polinomijalne složenosti koji će dodeliti dužine granama stabla tako da diskrepanca bude minimalna. Međutim, u praktičnim primenama neće biti poznata topologija stabla pa stoga moramo računati minimum po svim mogućim stablima. Sa dodavanjem svakog lista u stablo, broj različitih topologija stabala raste eksponencijalno. Problem minimizacije diskrepance po svim mogućim stablima je NP kompletan. U nastavku, razmotrićemo dve heuristike za konstrukciju stabla na osnovu neaditivnih matrica.

1.2.4 UPGMA algoritam

Do sada smo razmatrali *nekorena* evolutivna stabla i to prosta, kod kojih nije bilo čvorova stepena 2 (slika 1.16). U filogenetskim istraživanjima se često koriste i *korena* stabla. Jedno nekoreno stablo je lako zameniti korenim tako što se izabere jedna grana i podeli na dve dodavanjem novog čvora. Taj čvor nazivamo *korenom* i njegov stepen bi bio 2 (slika 1.17).



Slika 1.16: Nekoreno filogenetsko stablo

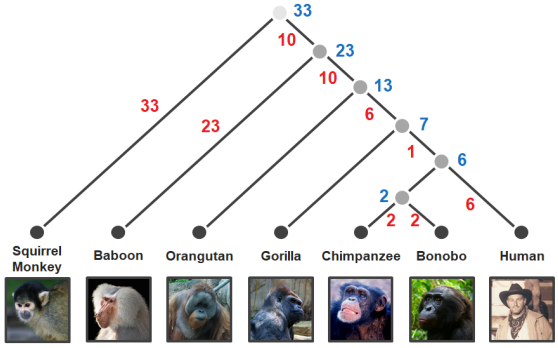


Slika 1.17: Koreno filogenetsko stablo dobijeno od nekorenog stabla sa prethodne slike tako što je spoljna grana lista *Squirrel Monkey* podeljena novim (korenim) čvorom na dve

Kod korenih filogenetskih stabala moguće je dodeliti težine granama na sledeći način:

- Svakom *čvoru* se dodeli ceo broj koji predstavlja njegovu *starost* u smislu kada se vrsta koju taj čvor predstavlja razdvojila na druge vrste
- Listovima je dodeljena nula
- Težina grana se određuje kao razlika starosti čvorova koje grana spaja

Koreno filogenetsko stablo konstruisano na opisani način nazivamo *ultrametričnim* stablima (slika 1.18). Ovakva stabla imaju osobinu da je svaki list jednako udaljen od korena.



Slika 1.18: Ultrametrično stablo

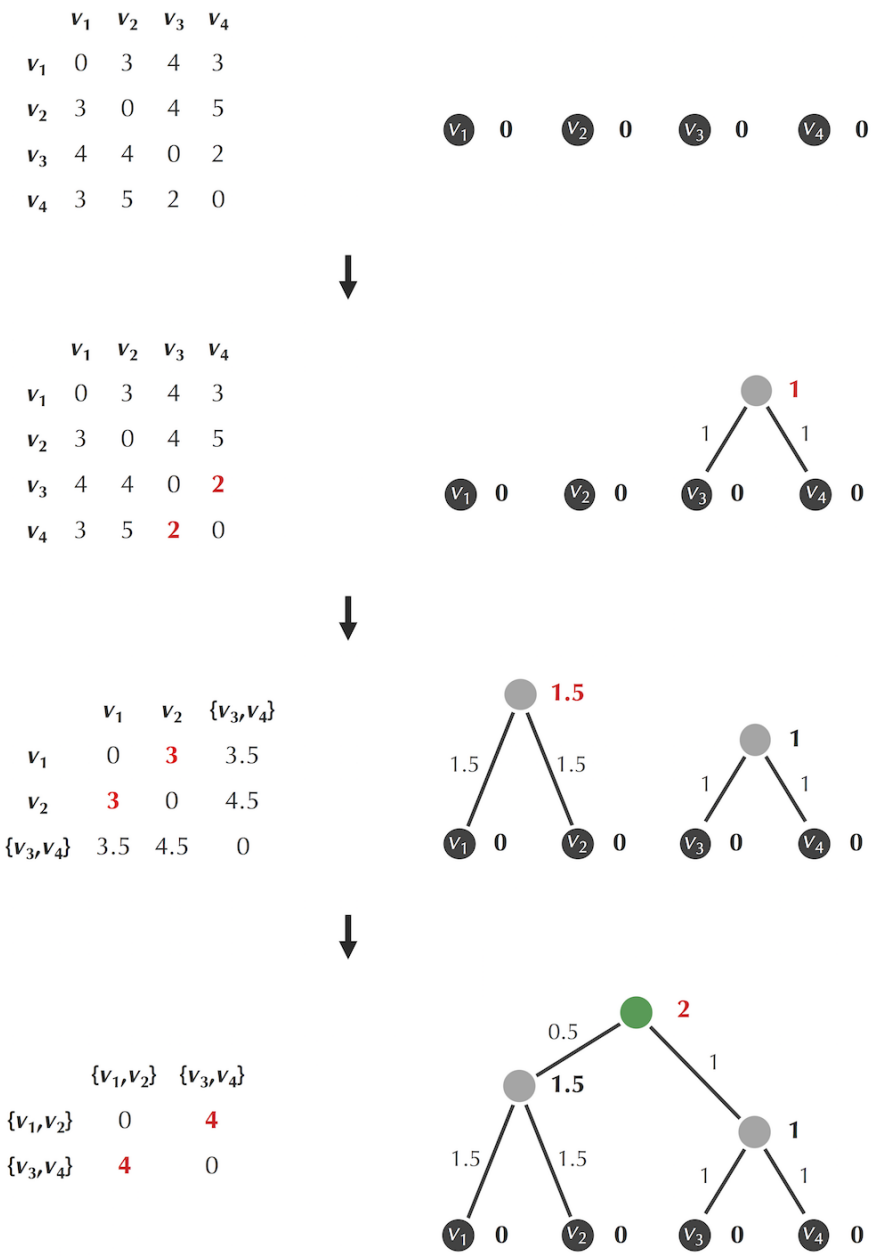
Jedan način za konstruisanje ultrametričnog stabla od matrice rastojanja je primenom UPGMA algoritma (eng. *Unweighted Pair Group Method with Arithmetic Mean*) koji primenjuje jednostavnu heuristiku klasterovanjem. Koraci algoritma su sledeći:

1. Formirati klaster za svaku današnju vrstu. Svaki klaster sadrži jedan list.
2. Naći dva *najbliža* klastera C_1 i C_2 pri čemu se rastojanje između klastera meri kao prosečno rastojanje između njihovih članova

$$D_{avg}(C_1, C_2) = \sum_{i \in C_1, j \in C_2} \frac{D_{i,j}}{\mid C_1 \mid \cdot \mid C_2 \mid}$$

gde $\mid C \mid$ označava broj elemenata u klasteru C.
3. Spojiti C_1 i C_2 u jedinstveni klaster C.
4. Uvesti novi čvor C kao roditeljski za C_1 i C_2 i odgovarajuće grane. Postaviti starost čvora C na $D_{avg}(C_1, C_2)/2$.
5. Ažurirati matricu rastojanja tako što izbacimo redove/kolone koji se odnose na klastere C_1 i C_2 i ubacimo red/kolonu koja se odnosi na novi klaster C
6. Iteriramo dok matricu rastojanja ne svedemo na dimenziju 2x2 i ne povežemo sve čvorove u jedan klaster

Primer izgradnje filogenetskog stabla UPGMA algoritmom je prikazan na slici 1.19. Obratimo pažnju da UPGMA algoritam radi i sa aditivnim i sa neaditivnim matricama rastojanja i to je njegova dobra strana. Međutim, uočimo da filogenetsko stablo na prikazanom primeru *ne odgovara* polaznoj matrici rastojanja. Naime, rastojanje između listova v_1 i v_4 u stablu je 4 a u matrici rastojanja je 3. Ova mana UPGMA algoritma koja potiče od činjenice da se spajaju najbliži klasteri što u početnom koraku podrazumeva da se listovi na najmanjem rastojanju proglašavaju za susedne, što ne mora biti slučaj kao što smo razmatrali u potpoglavlju 1.2.1.



Slika 1.19: Izgradnja ultrametričnog filogenetskog stabla pomoću UPGMA algoritma

Do sada smo se susreli sa algoritmom koji konstruiše odgovarajuća filogenetska stabla ali radi samo za aditivne matrice (AdditivePhylogeny) i sa algoritmom koji radi i za aditivne i za neaditivne matrice ali dobijeno filogenetsko stablo ne mora biti odgovarajuće (UPGMA). U nastavku ćemo prikazati algoritam koji ispunjava oba kriterijuma.

1.2.5 Neighbour-Joining algoritam

U prethodnim razmatranjima se ispostavilo da bi bilo korisno da na osnovu matrice rastojanja znamo koja dva lista će biti susedna. Ipak, kao što smo videli u više primera, minimalna vrednost

u matrici rastojanja ne znači da su listovi koji joj odgovaraju susedni. Ideja Neighbour-Joining (u nastavku NJ) algoritma je da, ako već ne možemo na osnovu minimuma matrice rastojanja da utvrdimo koji listovi će biti susedni, da odredimo matricu čiji minimum će ukazati na susedne listove. U nastavku navodimo bez dokaza NJ teoremu koja se bavi ovim problemom. Teorema nije intuitivna a njen dokaz nije trivijalan.

NJ teorema. Neka je data aditivna matrica rastojanja D dimenzija $n \times n$ i sledeće veličine:

- $TotalDistance_D(i)$ koja predstavlja sumu rastojanja lista i od ostalih listova

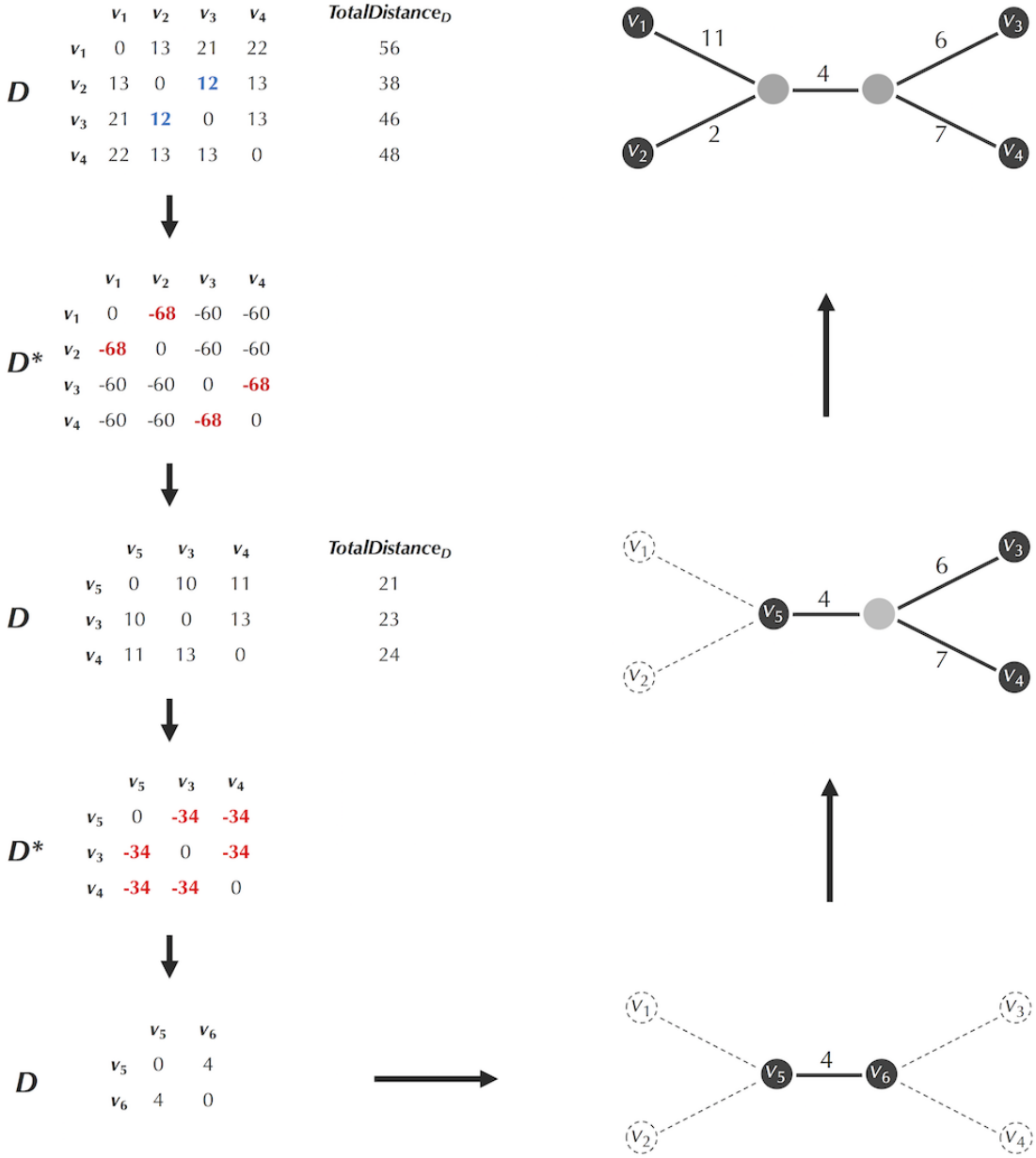
$$TotalDistance_D(i) = \sum_{1 \leq k \leq n} D_{i,k}$$

- NJ matrica $D^\star$ čiji su elementi na glavnoj dijagonali nule a na ostalim pozicijama

$$D^\star_{i,j} = (n-2) \cdot D_{i,j} - TotalDistance_D(i) - TotalDistance_D(j)$$

Tada, ako je $D^\star_{i,j}$ minimalni element u NJ matrici $D^\star$, listovi i i j će biti susedni u $Tree(D)$ ■.

Analizirajmo primer izgradnje filogenetskog stabla za datu aditivnu matricu rastojanja D na slici 1.20.



Slika 1.20: Izgradnja filogenetskog stabla pomoću NJ algoritma

Ako je $n = 2$, tada $NeighborJoining(D, n)$ vraća stablo sastavljeno od jedne grane dužine $D_{1,2}$, Ako je $n > 2$, tada bira najmanji element u NJ matrici (pretpostavimo da se nalazi na poziciji (i, j)), zamenjuje susedne listove i i j sa novim listom m i računa rastojanje od m do bilo kog lista k kao $D_{k,m} = (D_{k,i} + D_{k,j} - D_{i,j})/2$ (kao što je razmatrano na slici 1.11). Pošto znamo koji listovi su susedni i kako da izračunamo rastojanje od unutrašnjeg čvora do ostalih čvorova, možemo ukloniti vrste/kolone koje odgovaraju listovima i i j i uvesti novu vrstu/kolonu koja odgovara novom čvoru m. Time dobijamo matricu D' dimenzije $(n-1) \times (n-1)$. Rekurzivnom primenom NJ algoritma na matricu D' dobijamo stablo $Tree(D')$ koje se od stabla $Tree(D)$ razlikuje samo po tome što nema listove i i j povezane sa čvorom m. Dužine spoljnih grana

(i, m) i (j, m) određujemo na sledeći način:

$$\Delta_{i,j} = \frac{TotalDistance_D(i) - TotalDistance_D(j)}{n-2}$$

$$LimbLength(i) = \frac{D_{i,j} + \Delta_{i,j}}{2}, LimbLength(j) = \frac{D_{i,j} - \Delta_{i,j}}{2}$$

Pokažimo zašto se spoljne grane računaju ovako. Pretpostavimo najpre da je D aditivna matrica. Neka je k list različit od listova i i j a m njihov roditeljski čvor. Sa slike 1.11 možemo uočiti da za aditivne matrice važi $d_{i,m} = d_{i,j} - d_{j,m}$ i $d_{i,m} = d_{i,k} - d_{m,k}$. Kada saberemo ove dve jednačine, dobijamo:

$$2 \cdot d_{i,m} = d_{i,j} + d_{i,k} - d_{j,m} - d_{m,k} = d_{i,j} + d_{i,k} - d_{j,k}$$

Odatle možemo izračunati dužinu spoljne grane za list i, njemu susedni list j i od njih različit list k:

$$LimbLength(i) = \frac{D_{i,j} + D_{i,k} - D_{j,k}}{2}$$

Ova formula važi kada je D aditivna matrica i za proizvoljni list k (koji god list da odaberemo različit od i i j, vrednost $LimbLength(i)$ će biti ista). U slučaju da D nije aditivna matrica, vrednosti će biti različite za različit izbor k i tada uzimamo prosek po svim takvim vrednostima (ako matrica ima n elemenata, stablo ima n listova pa potencijalnih listova različitih od i i j ima $n-2$):

$$\begin{aligned}
LimbLength(i) &= \frac{D_{i,j}}{2} + \frac{1}{n-2} \cdot \sum_{\text{all leaves k differing from i and j}} \frac{D_{i,k} - D_{j,k}}{2} \\
&= \frac{D_{i,j}}{2} + \frac{1}{n-2} \cdot \Big(\sum_{\text{all leaves k differing from i and j}} \frac{D_{i,k}}{2} - \sum_{\text{all leaves k differing from i and j}} \frac{D_{j,k}}{2} \Big) \\
&= \frac{1}{2} \cdot (D_{i,j} + \frac{1}{n-2} \cdot \Big(\sum_{\text{all leaves k differing from i}} D_{i,k} - \sum_{\text{all leaves k differing from j}} D_{j,k} \Big)) \\
&= \frac{1}{2} \cdot (D_{i,j} + \frac{(TotalDistance_D(i) - TotalDistance_D(j))}{n-2}) \\
&= \frac{1}{2} \cdot (D_{i,j} + \Delta_{i,j})
\end{aligned}$$

Koraci algoritma su sumirani u narednoj listi.

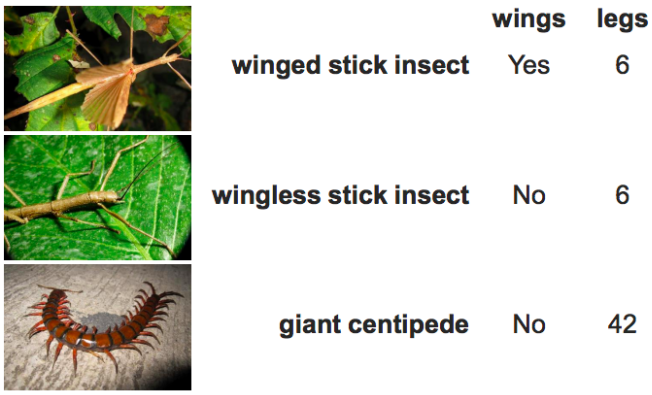
1. Konstruišemo neighbour-joining matricu D* na osnovu matrice D.
2. Nađemo minimalni element $D^*_{i,j}$ matrice D*.
3. Izračunamo $\triangle_{i,j} = (TotalDistance_d$(i) - $TotalDistance_d$(j)) / (n - 2).
4. Postavimo LimbLength(i) na $1/2(D_{i,j} + \triangle_{i,j})$ i LimbLength(j) na $1/2(D_{i,j} + \triangle_{i,j})$.
5. Formiramo matricu D' tako što uklonimo i-ti i j-ti red/kolonu iz D i dodamo m-ti red/kolonu tako da za svako k važi $D_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j})$ / 2..
6. Primenimo *NeighborJoining* rekurzivno na D' da dobijemo Tree(D').
7. Vratimo krajnje grane do čvorova i i j i dobijemo Tree(D).

1.2.6 Mane metoda zasnovanih na rastojanju

Kada višestruko poravnanje zamenimo matricom rastojanja, gubimo informacije o sekvencama iz poravnanja. Zbog toga ne možemo da odredimo kakva je sekvenca odgovarala vrstama iz unutrašnjih čvorova što je nekada u istraživanjima od značaja. Zbog toga u nastavku razmatramo rekonstrukciju evolutivnih stabala bez korišćenja matrice rastojanja.

1.3 Različiti pristupi rešavanju problema filogeneze na osnovu karakteristika

Do sada smo razmatrali rekonstrukciju evolutivnih stabala na osnovu rastojanja. Za formiranje matrica rastojanja bilo je neophodno poznavati sastav DNK sekvence. Znamo da je sekvenciranje genoma započelo krajem 70tih godina 20. veka, a masovno tek početkom 21. veka pa možemo postaviti pitanje da li je i pre toga bilo rezultata u rekonstrukciji evolutivnih stabala. Sredinom 20. veka istraživači su konstruisali filogenetska stabla na osnovu anatomsko-fizioloških osobina organizama koje ćemo nazivati *karakteristikama*. Na primer, za analizu evolucije beskičmenjaka korišćen je broj nogu kao i činjenica da li vrsta ima krila ili ne. Ove karakteristike određuju *matricu karakteristika* za tri vrste prikazanu na slici 1.21).



Slika 1.21: Primer matrice karakteristika za utvrđivanje filogeneze kod beskičmenjaka

Svaki red u matrici karakteristika dimenzije $n \times m$ predstavlja vektor karakteristika koji se sastoji od m karakteristika za jednu od n postojećih vrsta. Na osnovu takve matrice potrebno je konstruisati evolutivno stablo čiji listovi odgovaraju postojećim vrstama u kom su vrste sa sličnim vektorima karakteristika blizu. Takođe, potrebno je svakom unutrašnjem čvoru dodeliti odgovarajući vektor karakteristika i tako rekonstruisati kakve karakteristike su imale eventualne izumrle vrste.

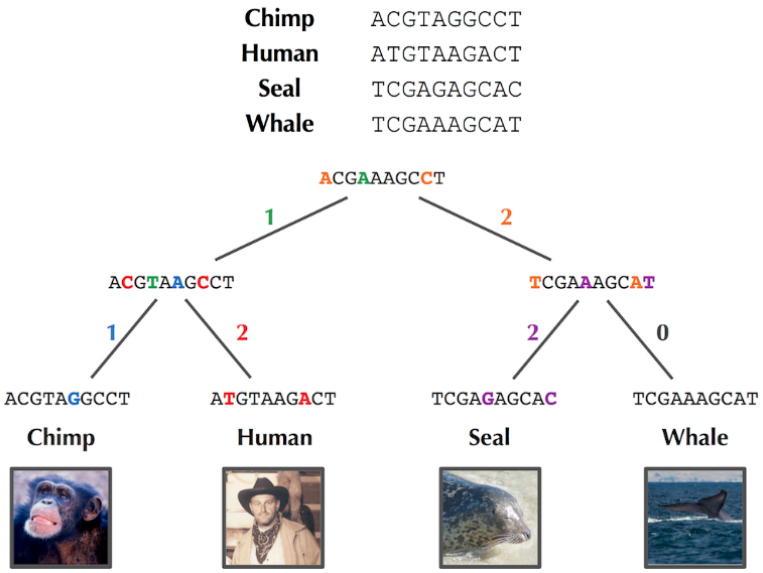
Korišćenje karakteristika za rekonstrukciju evolutivnih stabala nije pokazalo dobre rezultate u odnosu na korišćenje DNK sekvenci. Ipak, dodeljivanje vektora karakteristika (koje god one bile) unutrašnjim čvorovima je preostalo kao zahtev istraživača i uslovilo razvoj više algoritama.

1.3.1 Problem male parsimonije

Pretpostavimo da je poznata topologija evolutivnog stabla i da je svakom listu dodeljena jedna postojeća vrsta. Za svaku vrstu je poznat deo DNK sekvence (jedan gen) i za ove sekvence je

urađeno višestruko poravnanje. Razmotrimo kako možemo da dodelimo odgovarajuće niske svakom unutrašnjem čvoru koji odgovara potencijalnom izumrlom pretku. Da bismo odredili takve niske, potrebno je uvesti neku *funkciju skora* koja meri kako stablo obeleženo na takav način odgovara datom višestrukom poravnanju. U nastavku ćemo pretpostaviti da višestruko poravnanje ne sadrži indele već samo supstitucije. U praktičnim primerima se kolone sa indelima iz poravnanja mogu izostaviti.

Uvođenje funkcije skora analiziraćemo na primeru na slici 1.22. Dati su poravnati delovi DNK sekvenci iz četiri vrste i dato je filogenetsko stablo. Listovima su dodeljene sekvence postojećih vrsta a unutrašnjim čvorovima sekvence za koje pretpostavljamo da su pripadale izumrlim vrstama. S obzirom da su čvorovi označeni DNK sekvencama, ima smisla grane označiti Hamingovim rastojanjem između oznaka krajnjih čvorova. Jedan način za računanje funkcije skora filogenetskog stabla bi mogao da bude zbir težina svih grana stabla. Ovakvu funkciju skora nazivmo *skorom parsimonije*.



Slika 1.22: Primer filogenetskog stabla izgrađenog na osnovu matrice poravnanja

Definišimo problem male parsimonije.

Problem male parsimonije: Označiti unutrašnje čvorove datog filogenetskog stabla tako da njegov skor parsimonije za datu matricu poravnanja bude minimalan.

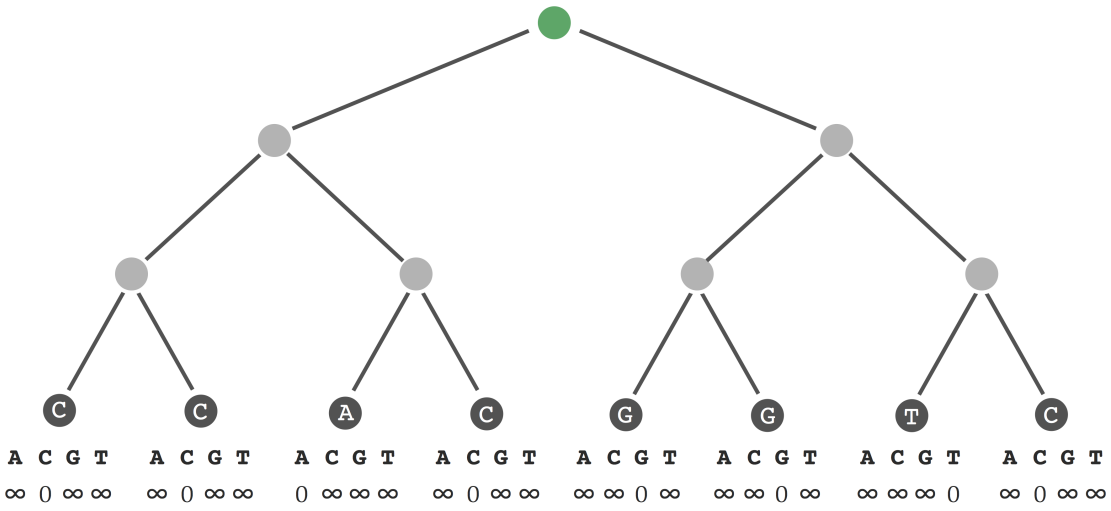
Ulaz: Koreno binarno stablo gde je svaki list obeležen niskom dužine m.

Izlaz: Označavanje svakog unutrašnjeg čvora niskama dužine m tako da skor parsimonije stabla bude minimalan.

S obzirom da je cilj dodeliti svakom čvoru stabla nisku dužine m, kao i da su karakteri niske međusobno nezavisni, problem male parsimonije možemo rešavati pojedinačno za svaku kolonu matrice poravnanja. Preciznije, možemo umesto jednog stabla i niske dužine m u svakom čvoru posmatrati m stabala $T_1, \ldots, T_m$ u kojima je svaki čvor označen karakterom i iz oznaka unutar stabla T. Tako, grane ovakvog stabla mogu imati težinu 1, ako spajaju čvorove različitih oznaka,

ili 0 u suprotnom. U nastavku ćemo opisati *SmallParsimony* algoritam za rešavanje jednoslovnog problema male parsimonije.

Neka je k simbol u datoj azbuci i v čvor u stablu T. Dalje, neka je $s_k(v)$ minimalni skor parsimonije podstabla T_v (podstablo stabla T sa korenom u v) takvo da je v označeno sa k. Listovi definišu podstablo koje se sastoji od jednog čvora (tog lista). Tako, ako je v list, $s_k(v)$ će biti 0 ako je v označeno sa k, a u suprotnom ∞ (slika 1.23).

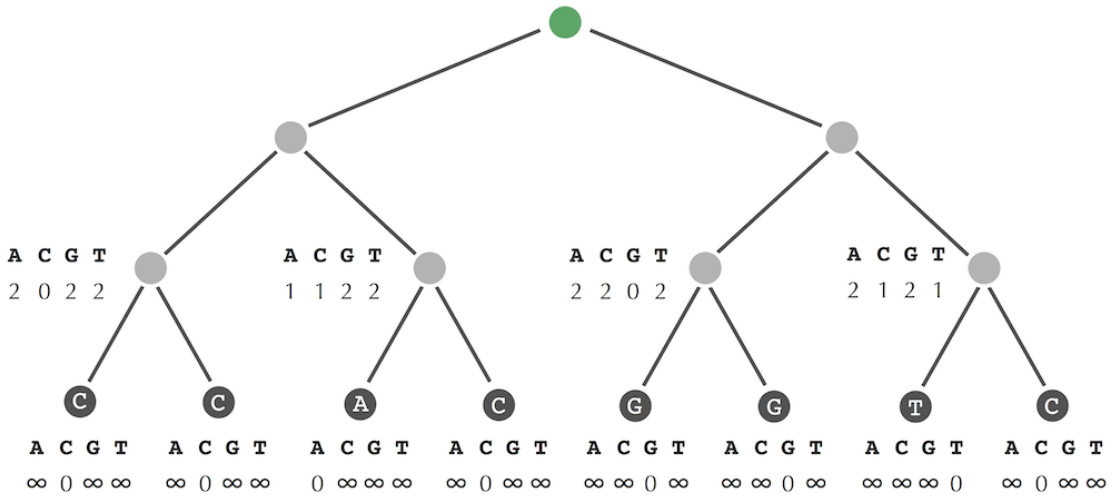


Slika 1.23: Inicijalizacija vrednosti $s_k(v)$ za listove

Razmotrimo sada $s_k(v)$ kada je v unutrašnji čvor. Pošto je stablo binarno, v ima dva potomka koja možemo označiti sa $Daughter(v)$ i $Son(v)$. Tada skor $s_k(v)$ možemo izračunati na sledeći način:

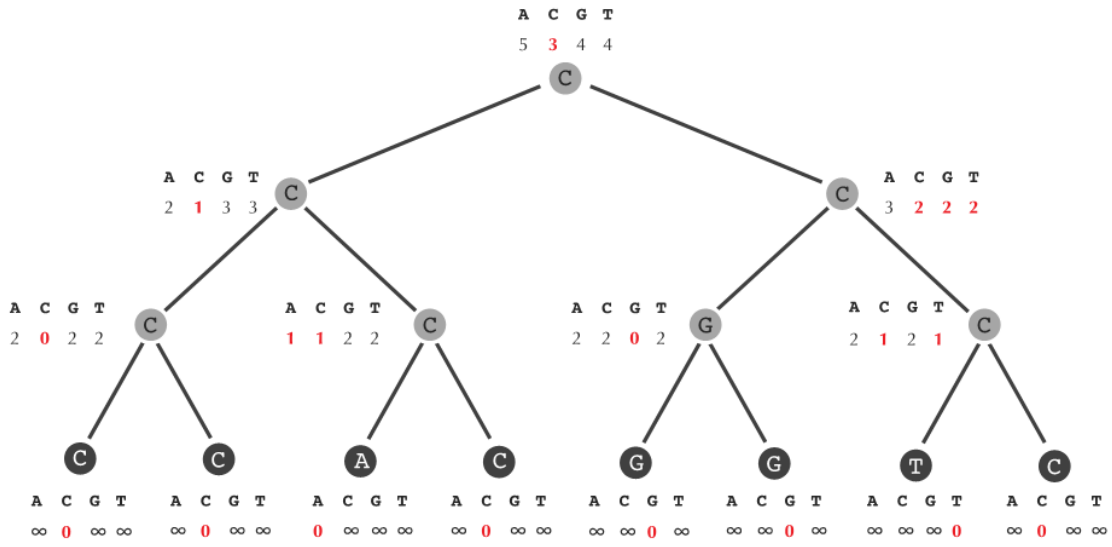
$$s_k(v) = \min_{all\ symbols\ i}\{s_i(Daughter(v)) + \delta_{i,k}\} + \min_{all\ symbols\ j}\{s_j(Son(v)) + \delta_{j,k}\},$$

gde δ predstavlja Kronekerov $\delta-$simbol. Na slici 1.24 su predstavljene su vrednosti $s_k(v)$ za unutrašnje čvorove stabla za svako k. Podstablo korena predstavlja celo stablo T pa je stoga najmanji skor parsimonije jednak minimalnoj vrednosti $s_k(root)$ po svim simbolima k.



Slika 1.24: Vrednosti $s_k(v)$ za unutrašnje čvorove

Nakon što smo izračunali skor parsimonije za stablo T, možemo dodeliti simbole unutrašnjim čvorovima. Minimum za $s_k(root)$ iznosi 3 i postignut je za k, pa je stoga u korenu C. Na sličan način označimo i ostale unutrašnje čvorove. Ukoliko se minimum postiže za više različitih vrednosti k, skor parsimonije stabla će biti isti bez obzira koje od njih izaberemo pa možemo izabrati proizvoljno (slika 1.25).



Slika 1.25: Oznake za unutrašnje čvorove

Pseudokod za algoritam *SmallParsimony* je prikazan u nastavku. Algoritam računa skor parsimonije za binarno koreno stablo T čiji su listovi označeni simbolima sačuvanim u nizu $Character$. U svakoj iteraciji, algoritam bira čvor v i računa $s_k(v)$ za svaki simbol k u azbuci. Za svaki čvor v algoritam u nizu Tag čuva informaciju o tome da li je obrađen ili nije. Kažemo da je unutrašnji čvor v *zreo* (eng. *ripe*) ako je $Tag(v) = 0$ a vrednosti niza Tag za oba njegova potomka su 1. *SmallParsimony* obrađuje čvorove od listova prema gore tako što pronalazi zreli čvor v za koji

je moguće izračunati $s_k(v)$ (slika 1.26).

```
SmallParsimony(T, Character)
    for each node v in tree T
        Tag(v) ← 0
        if v is a leaf
            Tag(v) ← 1
            for each symbol k in the alphabet
                if Character(v) = k
                    s_k(v) ← 0
                else
                    s_k(v) ← ∞
    while there exist ripe nodes in T
        v ← a ripe node in T
        Tag(v) ← 1
        for each symbol k in the alphabet
            s_k(v) ← minimum_{all symbols i} {s_i(Daughter(v))+α_{i,k}} + minimum_{all symbols j} {s_j(Son(v))+α_{j,k}}
    return minimum over all symbols k {s_k(v)}
```

Slika 1.26: Algoritam *SmallParsimony*

1.3.2 Problem velike parsimonije

Kod problema male parsimonije bila je poznata topologija stabla. Kada nije poznato kako stablo izgleda, potrebno je ispitati sva moguća stabla i odrediti ono sa najmanjim skorom parsimonije. Ovaj problem se naziva problemom velike parsimonije.

Problem velike parsimonije: Za dati skup niski, naći stablo čiji su listovi označeni ovim niskama koje ima najmanji skor parsimonije.

Ulaz: Kolekcija niski jednake dužine.

Izlaz: Koreno binarno stablo T koje minimizuje skor parsimonije po svim mogućim korenim binarnim stablima čiji su listovi označeni datim niskama.

S obzirom da ukupan broj različitih prostih filogenetskih stabala eksponencijalno raste sa povećanjem broja listova, ovaj problem je NP-kompletan. Heuristika pod nazivom *heuristika zamene najbližih suseda* koja ispituje neka od stabala od svih mogućih i daje dobre rezultate, izvan je okvira ovog kursa.

Literatura