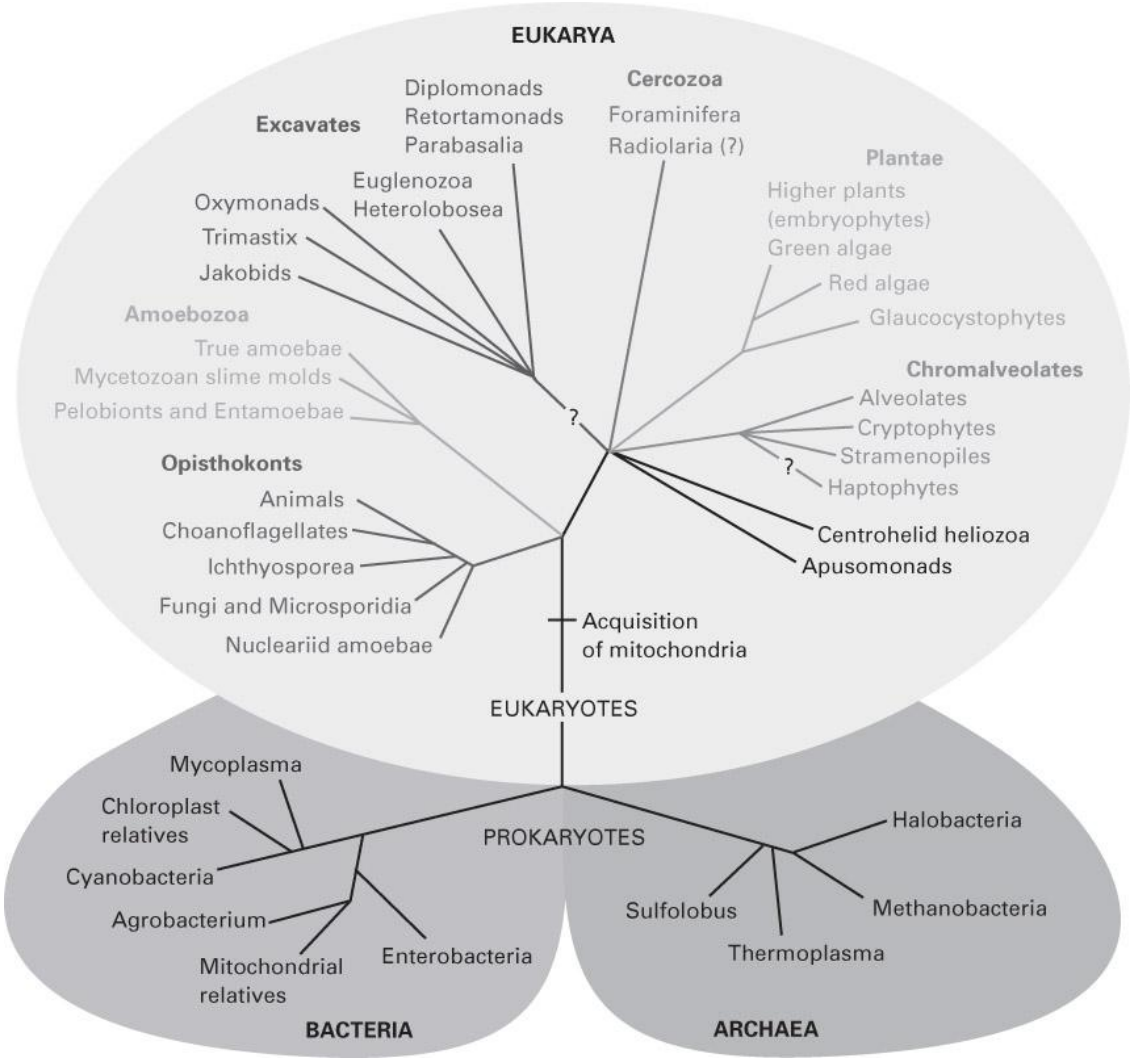
Filogenetska analiza

Molekulska evolucija



Molekulska evolucija

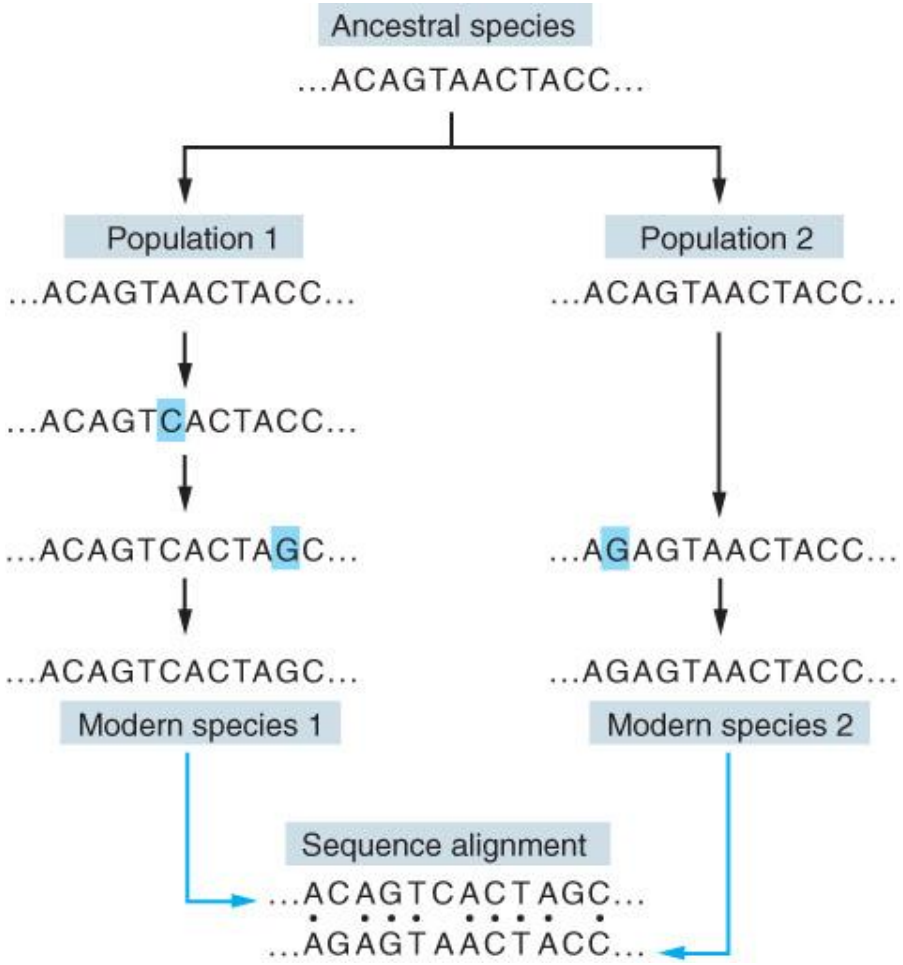
- Kako možemo utvrditi da li dve vrste potiču od istog pretka?
  - Starije metode: preko fosilnih ostataka i osobina organizama
  - Novije metode: na osnovu određenih delova DNK

Molekulska evolucija

- Uporedimo njihove DNK (najčešće jedan gen, pažljivo odabran), poravnamo ih i utvrdimo koliko su slične
- mere sličnosti mogu biti različite – najjednostavnija je brojanje pozicija na kojima imamo poklapanje
- *veća sličnost* – bliža evolutivna povezanost

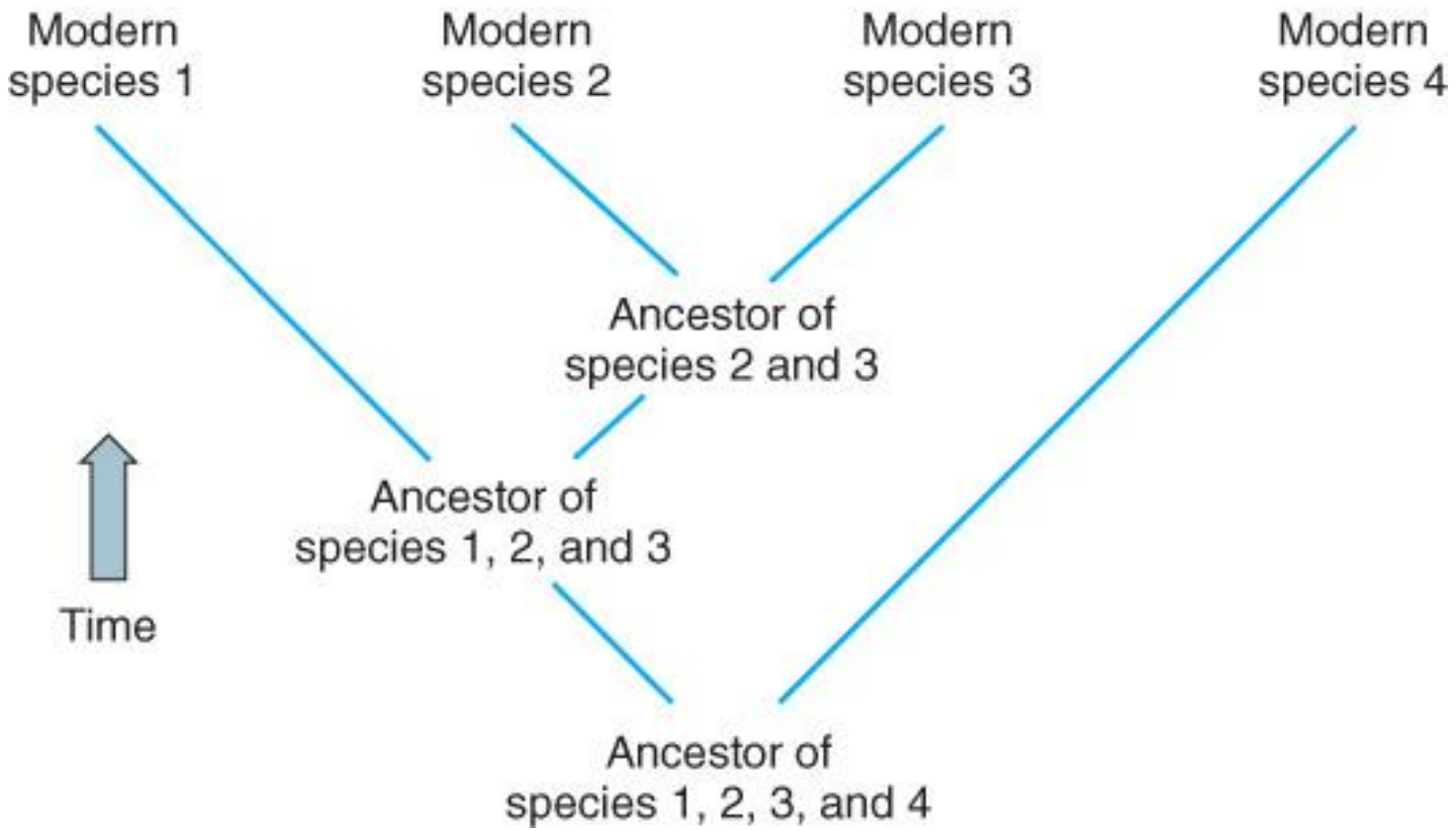
Molekulska evolucija

- Pozicije na kojima nemamo poklapanje – *mismatches* – mutacije nukleotida u DNK



Molekulska evolucija

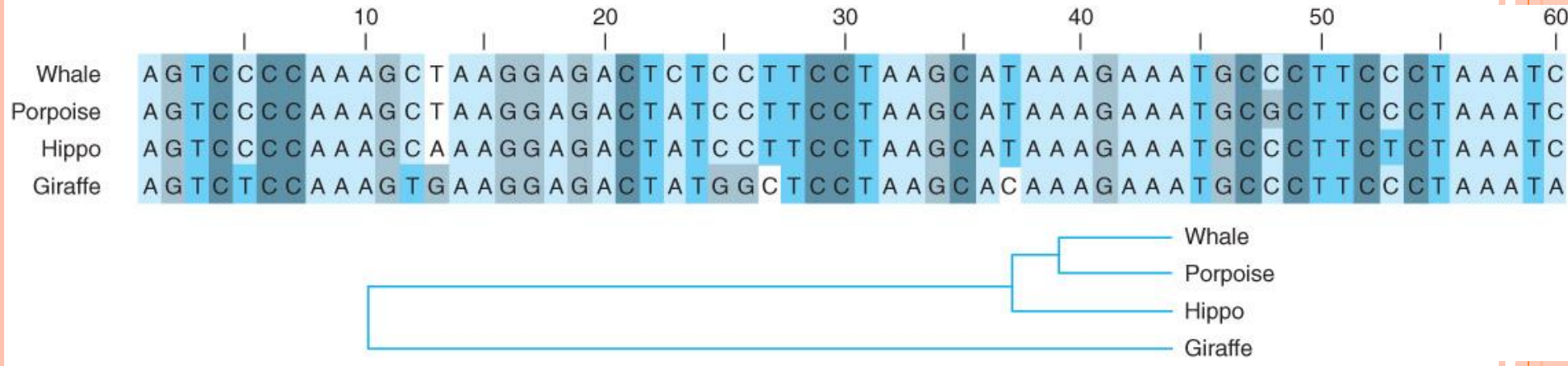
- Filogenetsko stablo pokazuje evolutivnu povezanost između dve vrste



Molekulski sat

- Gen koji koristimo da bismo rekonstruisali evolutivno stablo za grupu organizama
- Ideja: Ako dva organizma imaju isti gen, taj gen mora da je bio prisutan i kod njihovog zajedničkog pretka
- Primeri:
  - HBB kod kičmenjaka koji imaju hemoglobin
  - 16s rRNA kod svih živih bića
  - kazein – kod sisara

Molekulski sat



Molekulski sat

- Što je više *mismatches* u molekulskom satu, to je evolutivna veza između dve vrste dalja

Species	Substitutions in cytochrome c	Time (m.y.) since divergence
Human	–	–
Chimpanzee	1	5
Mouse	12	80
Yeast	62	800

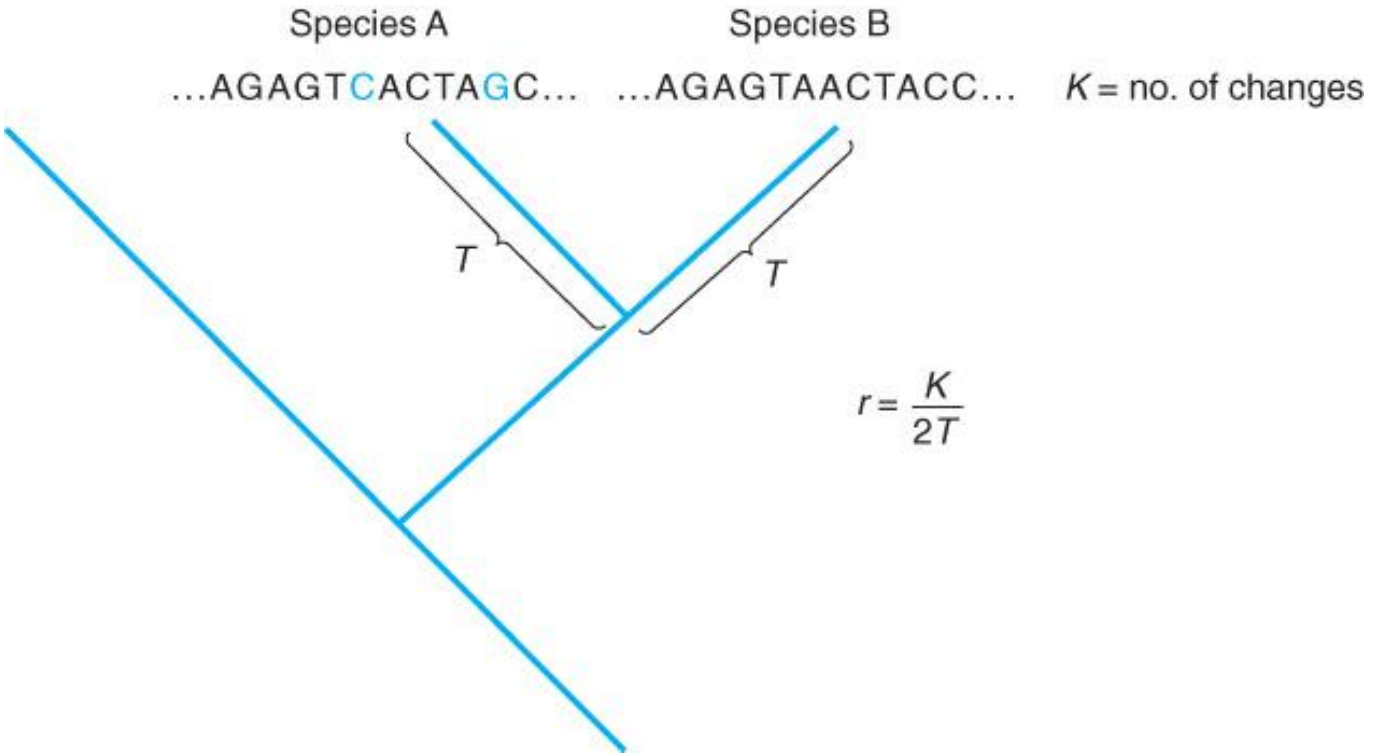
Molekulski sat

- Da li je broj *mismatches* dovoljan da odredi koliko su vrste evolutivno udaljene, tj koliko je miliona godina prošlo od njihovog razdvajanja?

I	II	III
ACCTG	**ACCTG**	**ACCTG**
ACTTG	TCCTG	ACTTG
ATTTG	ACCTG	**ACTTA**
ATTTA	ACTTG	
TTTTA	**ACTTA**	
TCTTA		
ACTTA		

Brzina supstitucije

K – broj mismatches
T – količina proteklog vremena

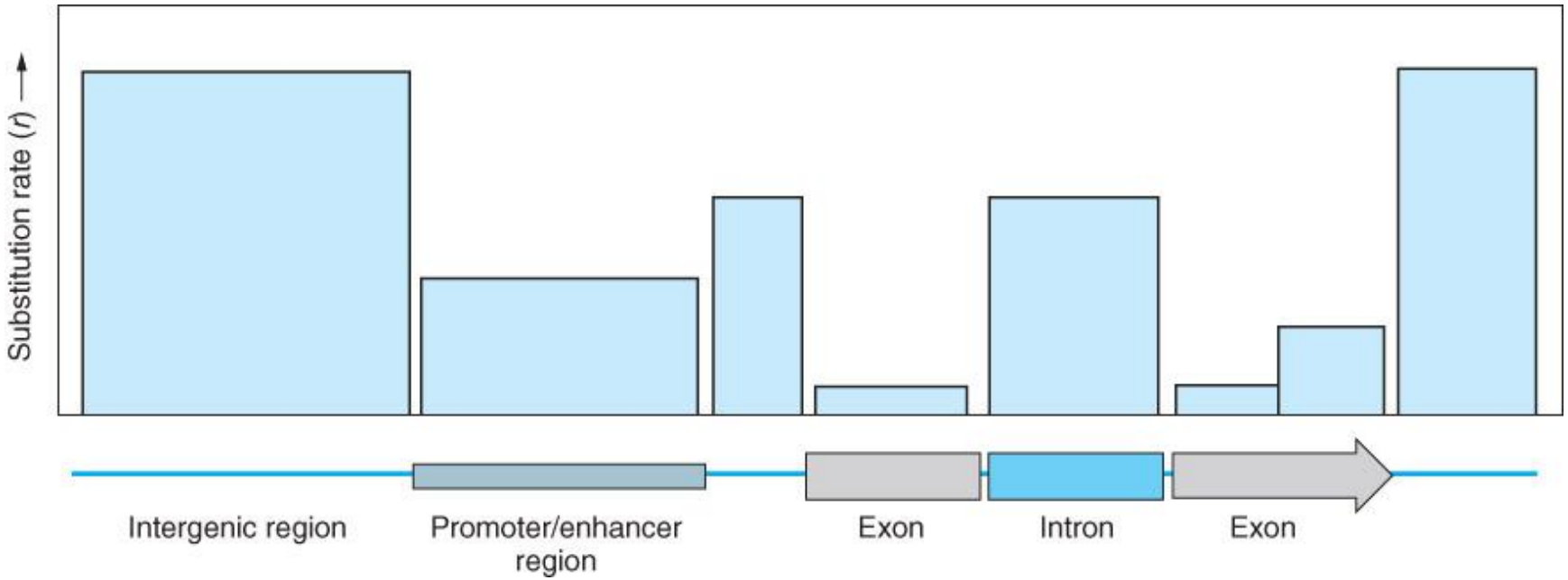


Brzina supstitucije

- Problem naći K i T
- T se može izračunati iz eksternih podataka (fosilni ostaci, radiometričko datiranje)
- K je teško odrediti – ne znamo na koji je način jedna sekvenca mutirala u drugu
- Ako pretpostavimo da je r konstantna, a znamo T, možemo izračunati K

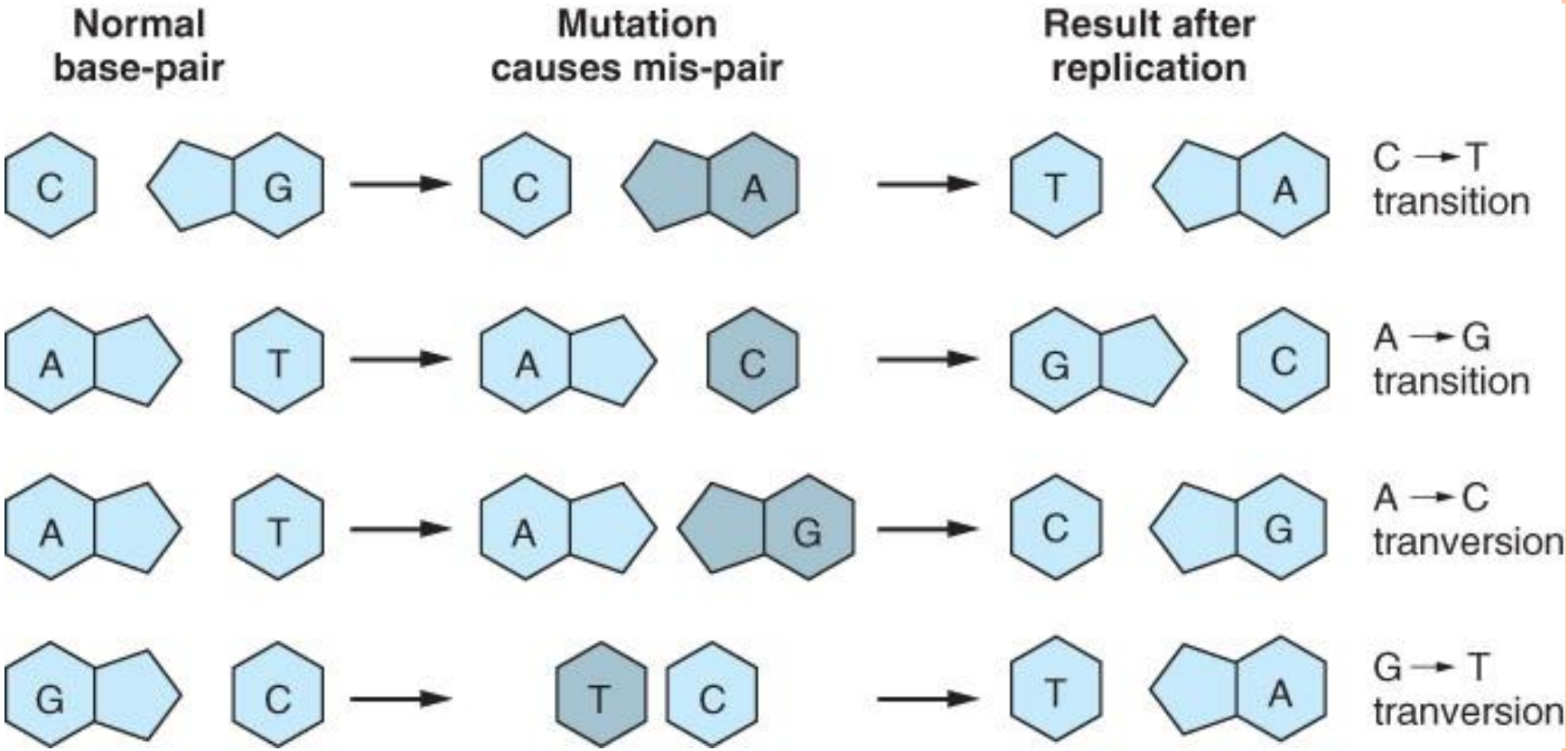
Da li su sve supstitucije zaista jednako verovatne?

- Zavisno od pozicije unutar DNK sekvence



Da li su sve supstitucije zaista jednako verovatne?

- Zavisno od tipa supstitucije



Evolutivni modeli

- K možemo proceniti na osnovu evolutivnih modela

- *Jukes-Cantor-ov model*:

$K_{AB} = -3/4\ln(1-4/3D_{AB})$, gde je D_{AB} procenat različitih nukleotida na istim pozicijama
  - podrazumeva da su sve supstitucije jednako verovatne

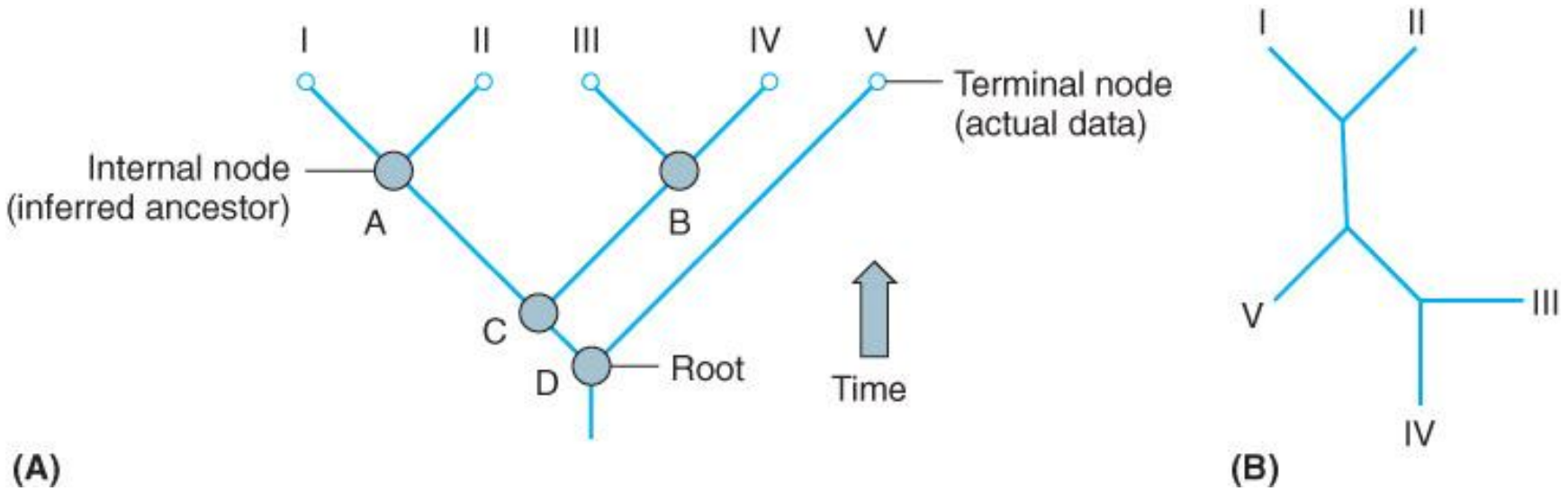
- *Kimurin dvoparametarski model*:

$K_{AB} = 1/2\ln(1/(1-2S-V))+3/4\ln(1/(1-2V))$,
gde je S procenat tranzicija (mutacija iz jedne purinske/pirimidinske baze u drugu) a V procenat transverzija (mutacija iz purinske baze u pirimidinsku ili obrnnuto)
  - podrazumeva da su tranzicije češće od transverzija

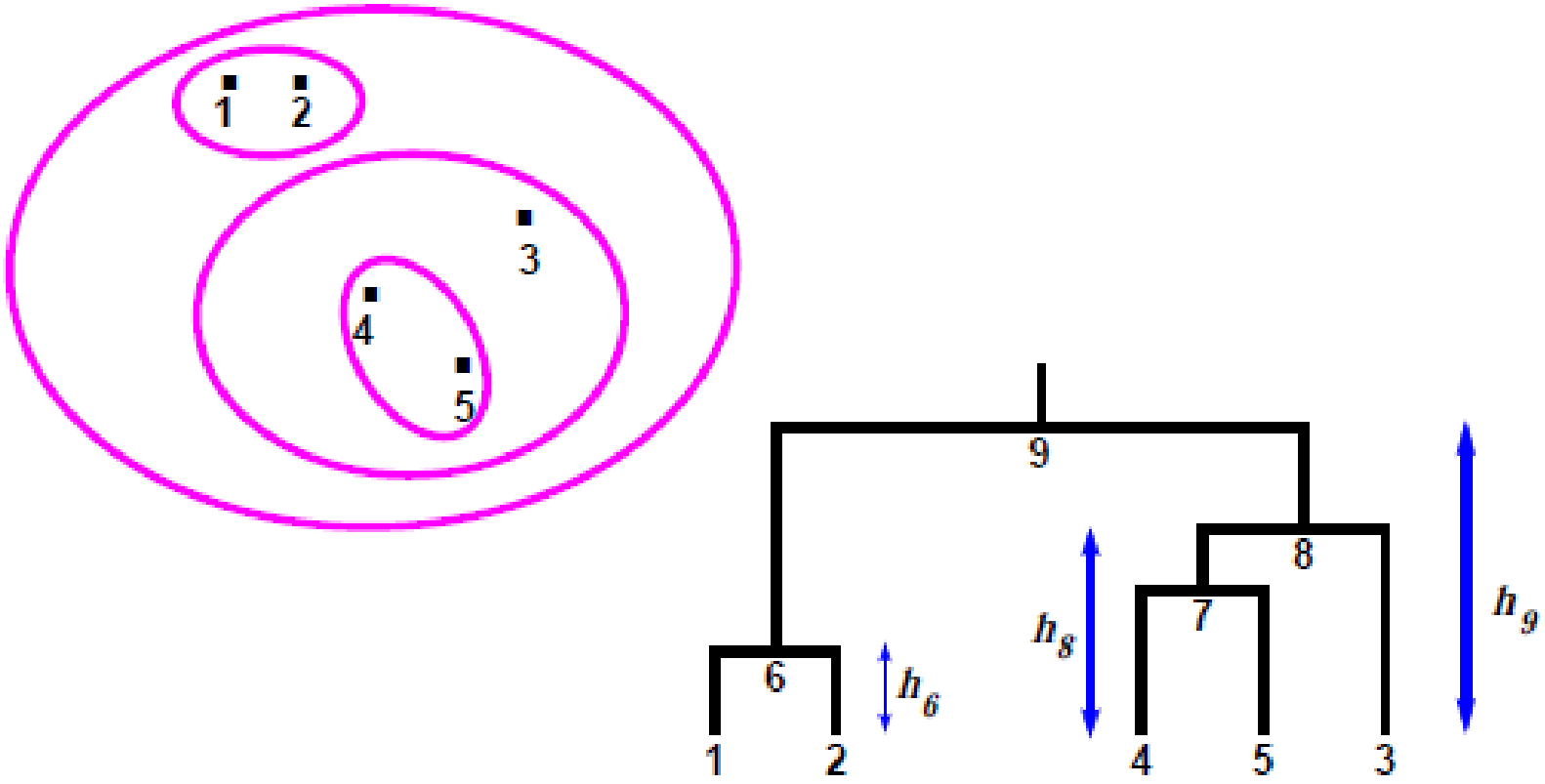
Konstrukcija filogenetskog stabla

1. Odabrati molekulski sat za datu grupu organizama
2. Izračunati sličnost između svaka dva organizma na osnovu odabrane mere
3. Primeniti odgovarajući algoritam
   - pristupi zasnovani na rastojanju
   - pristupi zasnovani na parsimoniji

Konstrukcija filogenetskog stabla



Algoritmi aglomerativnog hijerarhijskog klasterovanja



$$h_6 = \frac{1}{2}d_{12},\ h_7 = \frac{1}{2}d_{45},\ h_8 = \frac{1}{2}d_{37},\ h_9 = \frac{1}{2}d_{68}$$

Algoritmi aglomerativnog hijerarhijskog klasterovanja

- *Uopštena deja*: objekti na manjem rastojanju su bliskiji od objekata na većem rastojanju
- Inicijalno, svaki objekat je klaster za sebe
- Prvi nivo klastera se formiraju na osnovu rastojanja između objekata
- Naredni nivoi klasteri se podrazumevaju udruživanje postojećih klastera na osnovu rastojanja između njih (min, max, avg)

Algoritmi aglomerativnog hijerarhijskog klasterovanja

- **UPGMA** (Unweighted Pair Group Method using Arithmetic Averages) – podrazumeva konstantnu brzinu supstitucija – *ultrametričnost*

Za bilo koja tri klastera u tabeli rastojanja mora da važi

dist AC <= max {distAB, distBC}

- **NJ** (Neighbour joining) – NE podrazumeva konstantnu brzinu supstitucija – *aditivnost*

Za bilo koja četiri klastera u tabeli rastojanja mora da važi

dist AB + dist CD <= max {(distAC + distBD),(distAD + distBC)}

- Razlika: računanje rastojanja između klasera

UPGMA

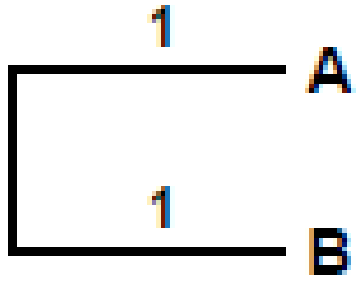
- Date su sekvence sa sledećom tabelom rastojanja

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

•Spajamo dva najbliža klastera: A i B

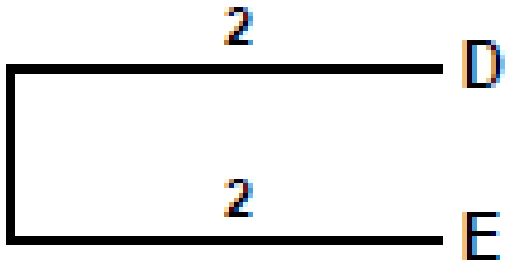
•Rastojanje između novog klastera AB i ostalih klastera Računamo po formuli:

$$d_{kl} = \frac{d_{il}\,|C_i| + d_{jl}\,|C_j|}{|C_i| + |C_j|}$$

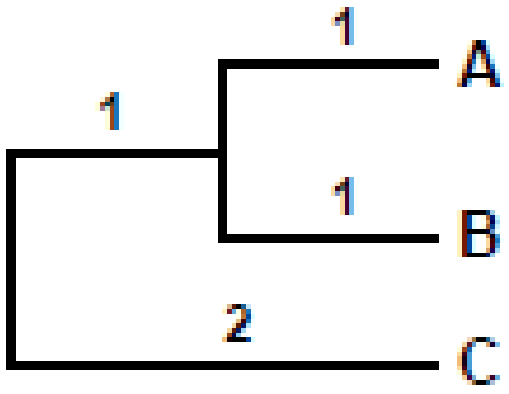


novi klaster AB se postavlja na visinu $d_{AB}/2$

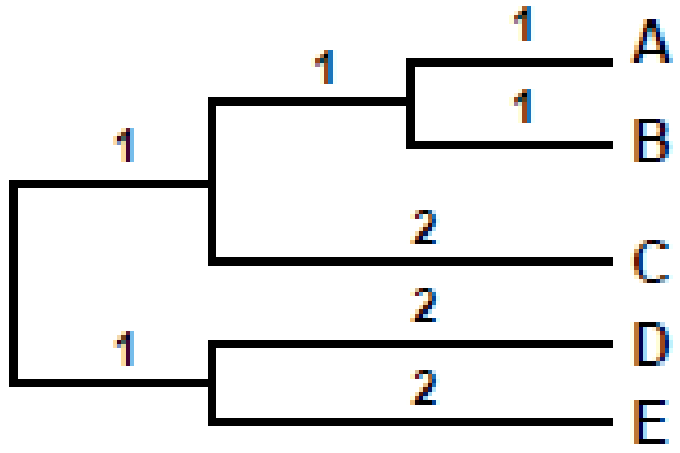
	AB	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



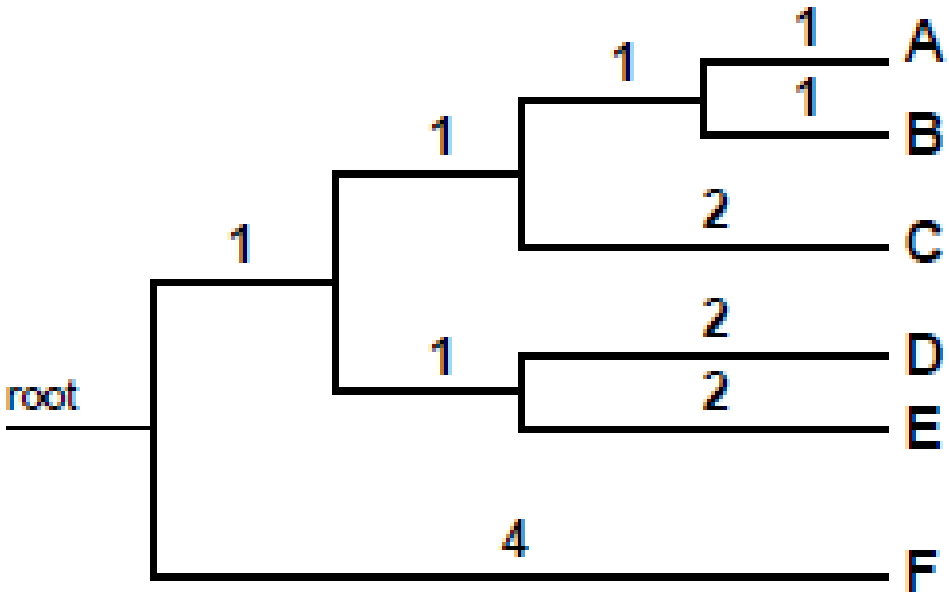
	AB	C	DE
C	4		
DE	6	6	
F	8	8	8



	ABC	DE
DE	6	
F	8	8

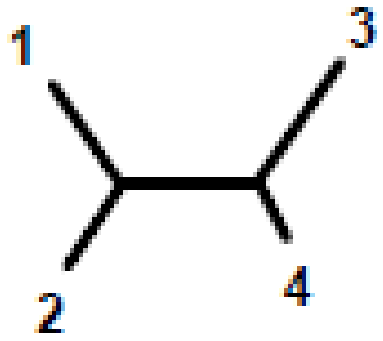


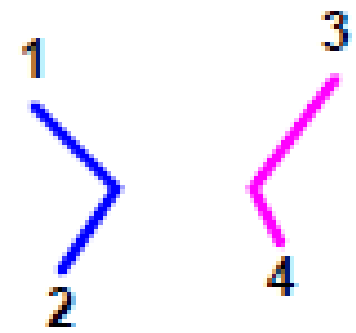
	ABCDE
F	8



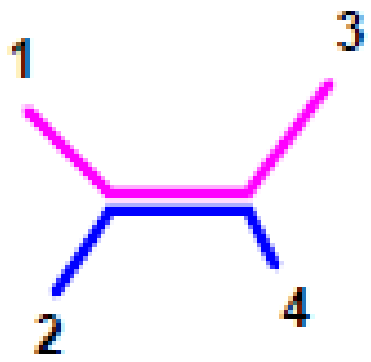
NJ – SPAJANJE SUSEDA

- Aditivnost rastojanja – uslov 4 tačke

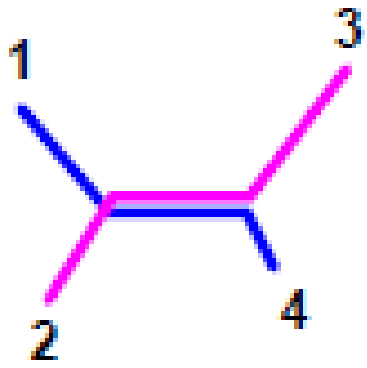




$d_{12} + d_{34}$



$d_{13} + d_{24}$



$d_{14} + d_{23}$

NJ

- Data je matrica rastojanja:

d	A	B	C	D	E	F
A		2	4	6	6	8
B	2		4	6	6	8
C	4	4		6	6	8
D	6	6	6		4	8
E	6	6	6	4		8
F	8	8	8	8	8	

Formiramo tabelu transformisanih rastojanja po formuli:

$$D_{ij} = d_{ij} - (r_i + r_j)$$

gde je: L – skup klastera

$$r_i = \frac{1}{|L|-2} \sum_{k \in L} d_{ik}$$

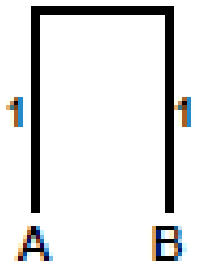
r	
A	6.5
B	6.5
C	7
D	7.5
E	7.5
F	10

NJ

- Rezultujuća matrica transformisanih rastojanja:

D	A	B	C	D	E	F
A						
B	−11					
C	−9.5	−9.5				
D	−8	−8	−8.5			
E	−8	−8	−8.5	−11		
F	−8.5	−8.5	−9	−9.5	−9.5	

Spajamo klastere sa najmanjim Dij



Izračunamo rastojanja novog klastera k (dobijenog spajanjem i i j) od ostalih klastera (m) po formuli:

$$d_{km} = \frac{1}{2}(d_{im} + d_{jm} - d_{ij})$$

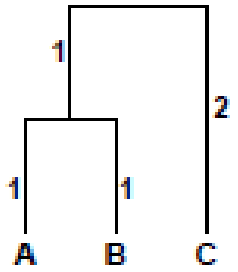
U stablo dodamo čvor k na visini d_{ik} od čvora i i d_{jk} od čvora j:

$$d_{ik} = \frac{1}{2}(d_{ij} + r_i - r_j) \qquad d_{jk} = \frac{1}{2}(d_{ij} - r_i + r_j)$$

d	C	D	E	F	AB
C		6	6	8	3
D	6		4	8	5
E	6	4		8	5
F	8	8	8		7
AB	3	5	5	7	

r	
C	7.67
D	7.67
E	7.67
F	10.33
AB	6.67

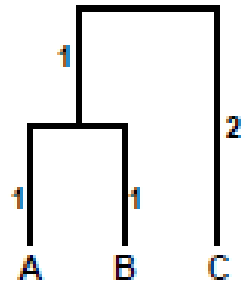
D	C	D	E	F	AB
C					
D	−9.33				
E	−9.33	−11.33			
F	−10	−10	−10		
AB	**−11.33**	−9.33	−9.33	−10	

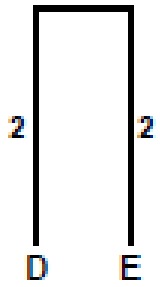


d	D	E	F	ABC
D		4	8	4
E	4		8	4
F	8	8		6
ABC	4	4	6	

r	
D	8
E	8
F	11
ABC	7

D	D	E	F	ABC
D				
E	**−12**			
F	−11	−11		
ABC	−11	−11	−12	

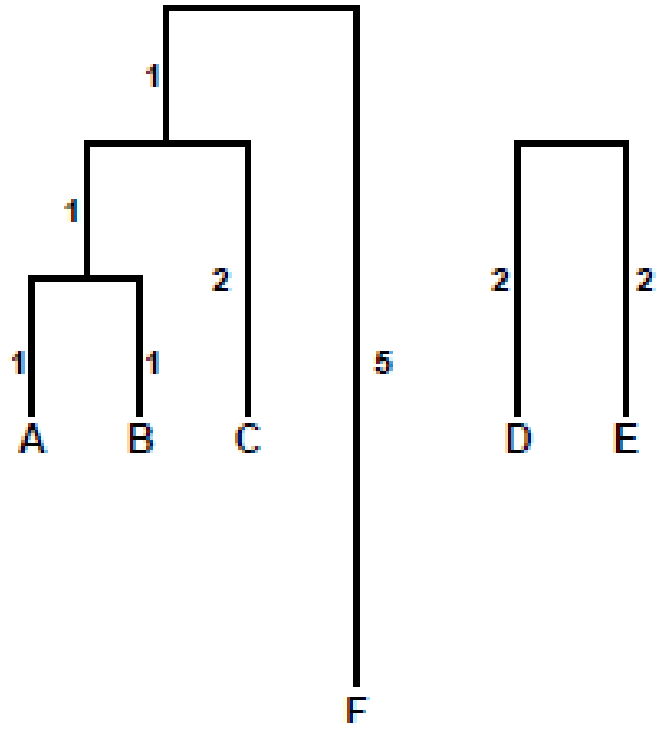




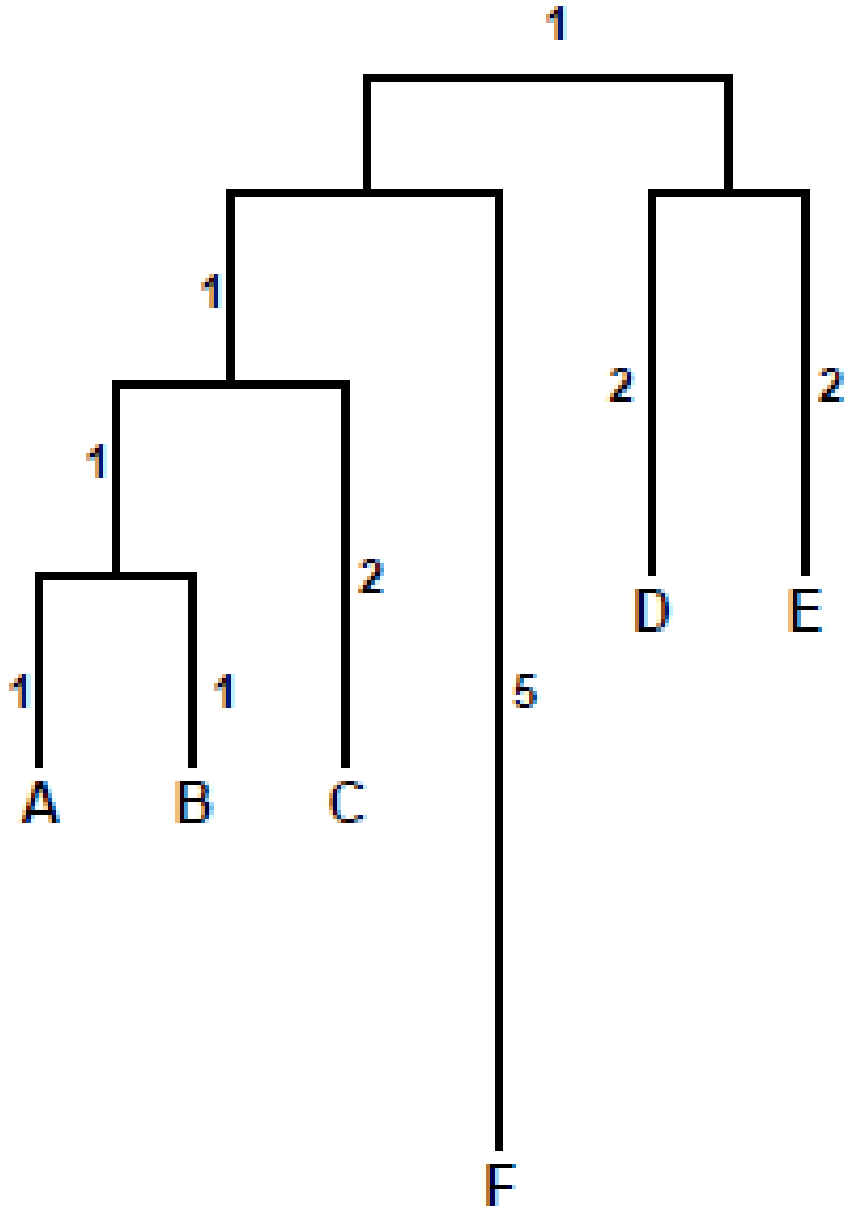
d	F	ABC	DE
F		6	6
ABC	6		2
DE	6	2	

r	
F	12
ABC	8
DE	8

D	F	ABC	DE
F			
ABC	-14		
DE	-14	-14	

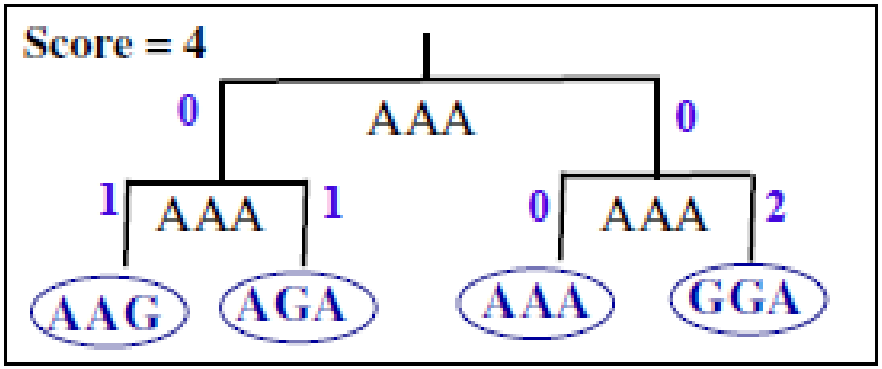


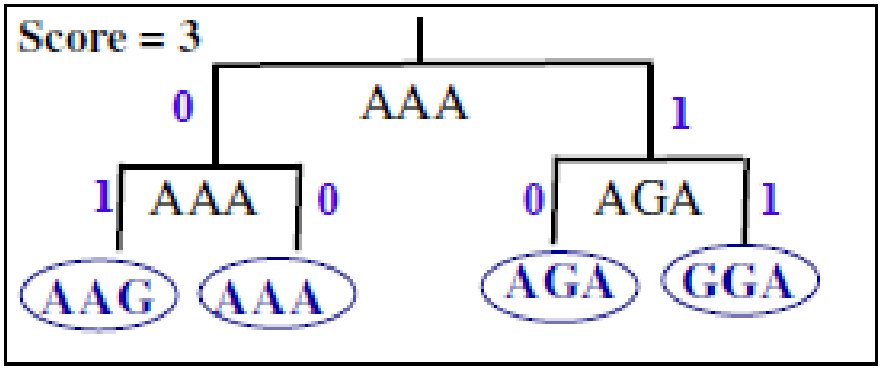
KONAČNO STABLO BEZ KORENA



Pristupi zasnovani na parsimoniji

- Skor parsimonije: broj mutacija u filogenetskom stablu
- Za isti niz sekvenci moguća su različita filogenetska stabla. Primer za sekvence: AAG, AGA, AAA i GGA.





- Princip minimalne evolucije: *najmanji broj mutacija*
- **Zadatak**: za dati niz sekvenci naći filogenetsko stablo sa najmanjim brojem mutacija

Fičov algoritam

- Date su sekvence:
  - **A**jkula: CAGGTA
  - **B**izon: CAGACA
  - **C**vrčak: CGGGTA
  - **D**abar: TGCACT
  - **E**mu: TGCGTA
- Izračunati skor parsimonije za stablo

  (((A,B),C),(D,E))
- Skor parsimonije za jednu poziciju: broj operacija unije
- Skor parsimonije za stablo: zbir skorova parsimonije po svim pozicijama

Sankofov algoritam

- Data je jedna pozicija u poravnanju sekvenci:

A: A

B: C

C: T

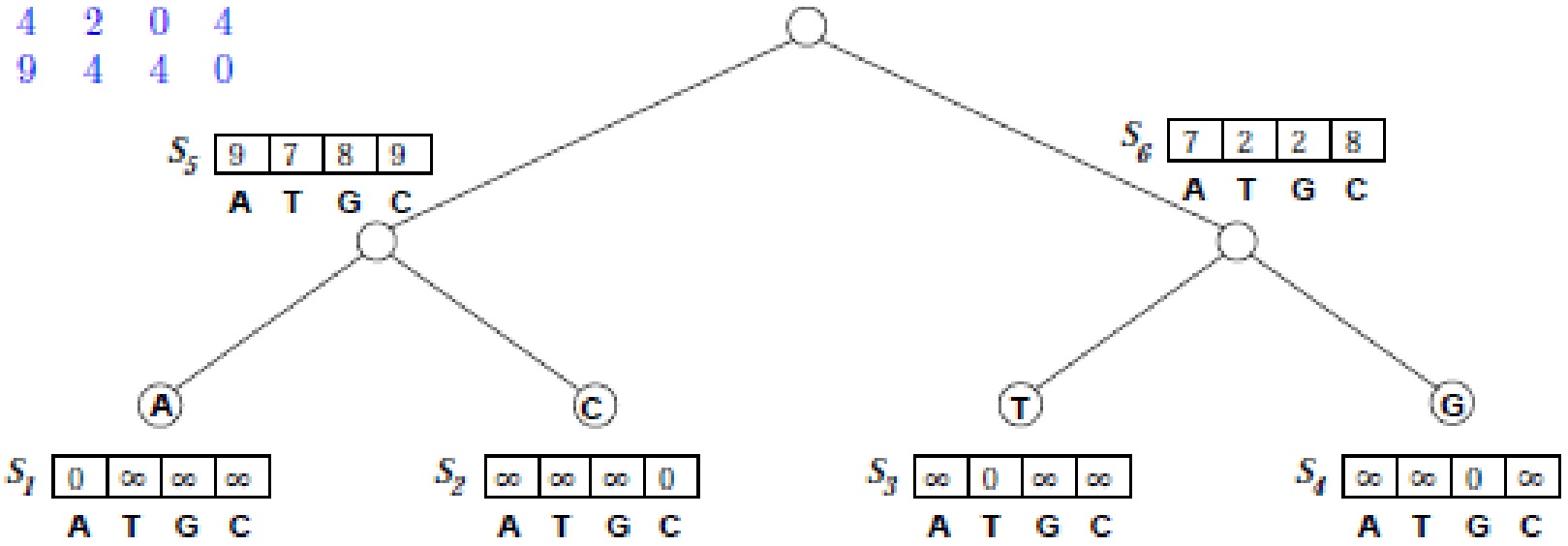
D: G

i filogenetsko stablo ((A,B),(C,D)).

Izračunati skor parsimonije za datu poziciju na osnovu sledeće tabele:

S	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

S	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0



$S_5(\mathsf{A}) = S(\mathsf{A}, \mathsf{A}) + S(\mathsf{A}, \mathsf{C}) = 0 + 9 = 9$
$S_5(\mathsf{T}) = S(\mathsf{T}, \mathsf{A}) + S(\mathsf{T}, \mathsf{C}) = 3 + 4 = 7$
...

LC: $S_{7,5} = S + [\ S_5^T\ S_5^T\ S_5^T\ S_5^T\ ]$
$S_{7,6} = S + [\ S_6^T\ S_6^T\ S_6^T\ S_6^T\ ]$

- Backtracking – na osnovu izračunatih matrica Sk, određujemo nukleotid u svakom unutrašnjem čvoru
- Koren: $\bar{a} = \ argmin_{a \in \{A,C,G,T\}} S_{root}(a)$
  - Ako ima više vrednosti za $argmin$, to znači da postoji više rešenja sa minimalnim skorom parsimonije
- Ako označimo potomke korena sa $\bar{b}$ i $\bar{c}$:

$$\bar{b} = \ argmin_{b \in \{A,C,G,T\}} (S(\bar{a}, b) + S_{left}(b)),$$

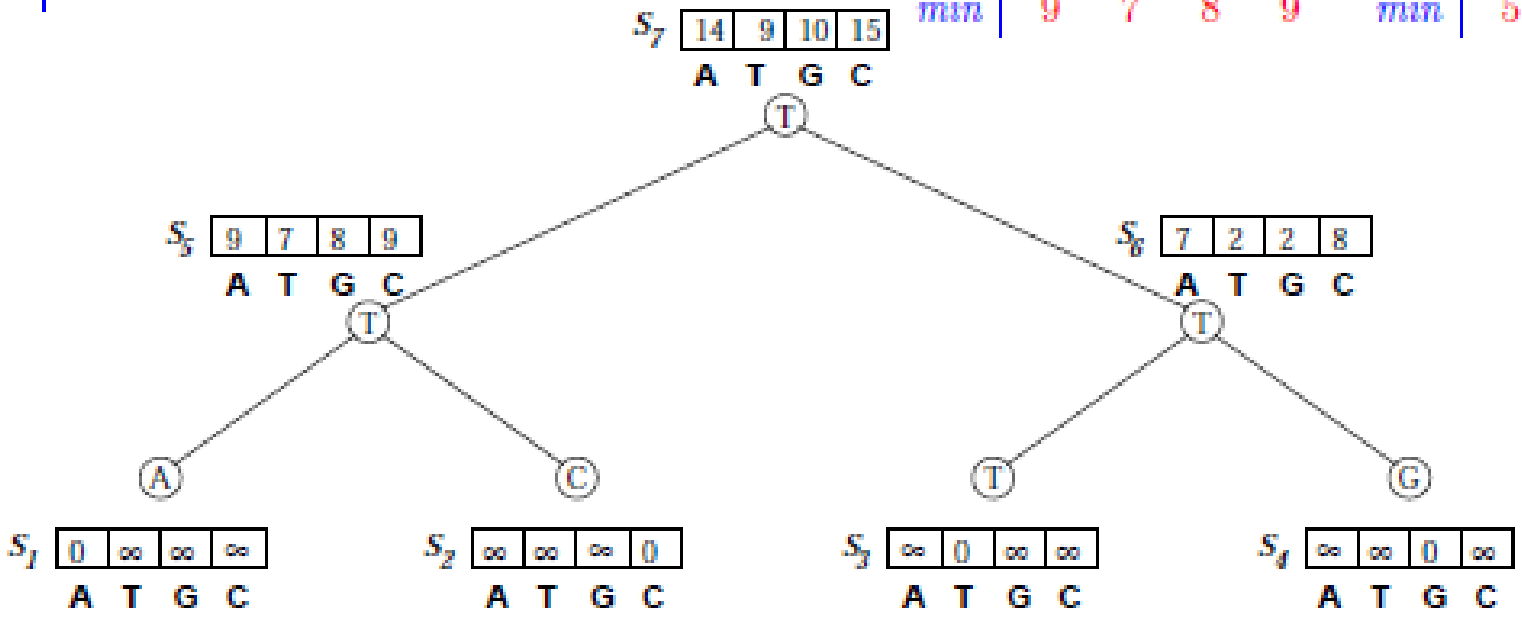
$$\bar{c} = \ argmin_{c \in \{A,C,G,T\}} (S(\bar{a}, c) + S_{right}(c))$$

I tako redom za sve unutrašnje čvorove

S	A	T	G	C
A	0	3	4	9
T	3	0	2	4
G	4	2	0	4
C	9	4	4	0

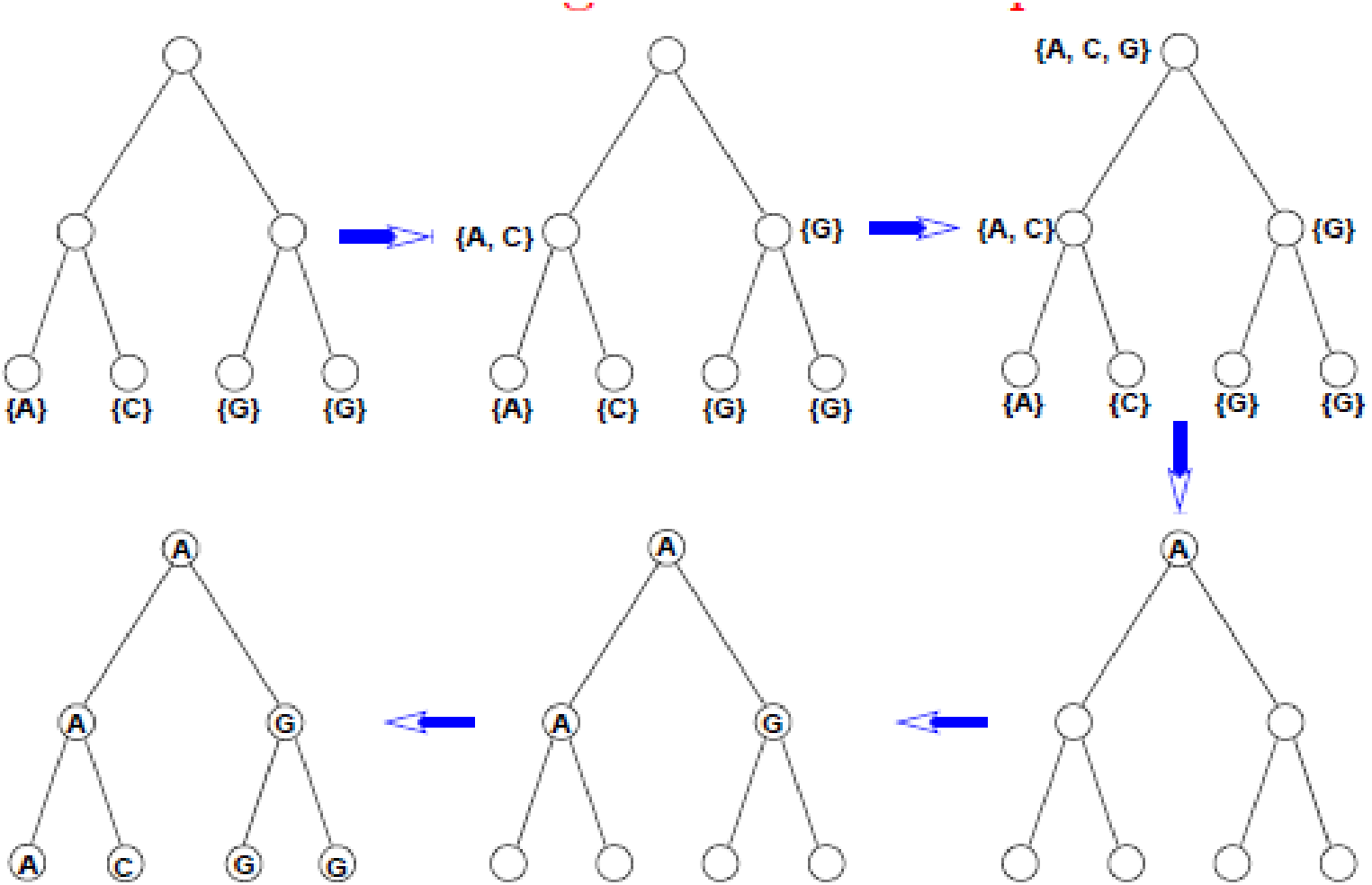
$S_{7,5}$	A	T	G	C
A	9	12	13	18
T	10	7	9	11
G	12	10	8	12
C	18	18	13	9
min	9	7	8	9

$S_{7,6}$	A	T	G	C
A	7	10	11	16
T	5	2	4	6
G	6	4	2	6
C	17	12	17	8
min	5	2	2	6



$$S_7(\mathrm{A}) = min_b(S_5(b) + S(\mathrm{A}, b)) + min_c(S_6(c) + S(\mathrm{A}, c)) = 9 + 5 = 14$$

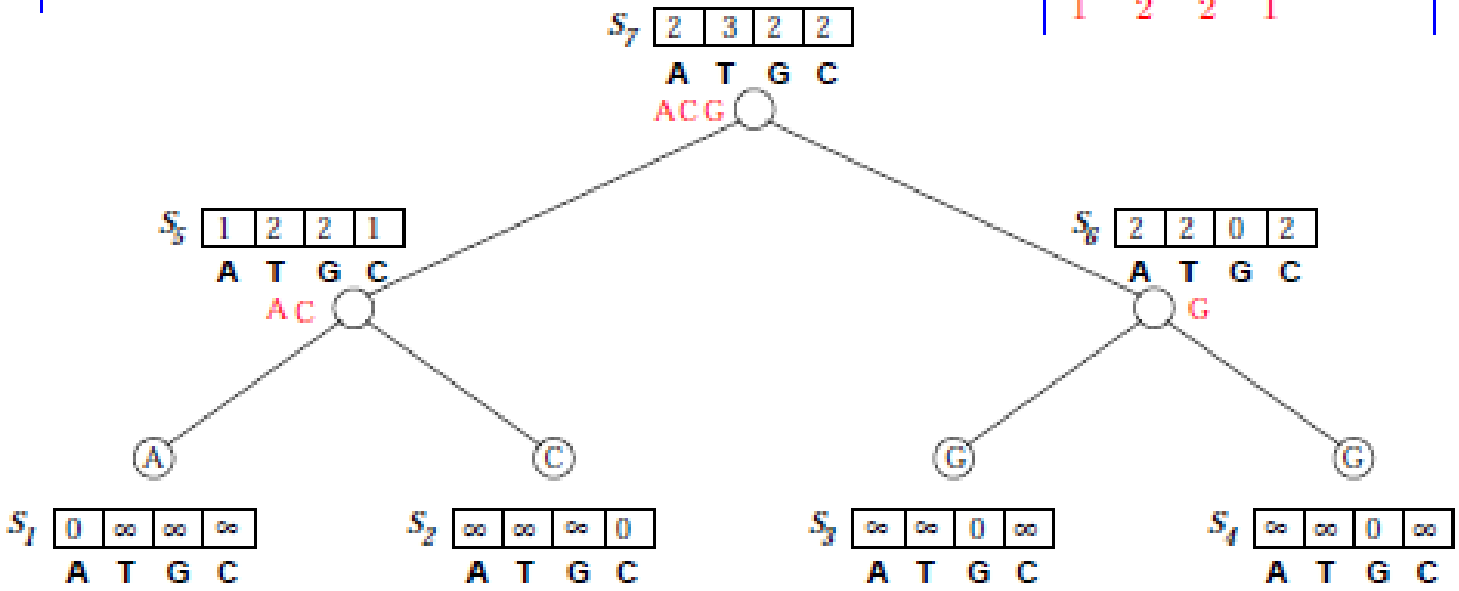
Fičov algoritam



S	A	T	G	C
A	0	1	1	1
T	1	0	1	1
G	1	1	0	1
C	1	1	1	0

$S_{7,5}$	A	T	G	C
A	1	2	2	2
T	3	2	3	3
G	3	3	2	3
C	2	2	2	1
	1	2	2	1

$S_{7,6}$	A	T	G	C
A	2	3	3	3
T	3	2	3	3
G	1	1	0	1
C	3	3	3	2
	1	1	0	1



- Kada koristimo 0-1 matricu rastojanja:
  - I Fičov i Sankofov algoritam računaju isti skor parsimonije
  - Fičov algoritam ne može backtracking-om da proizvede sva optimalna stabla; *na primer*, u prethodnom primeru skor dva bi imalo i stablo sa nukleotidom C u levom potomku korena, što je dobijeno backtracking-om u Sankofovom algoritmu, ali ne i u Fičovom