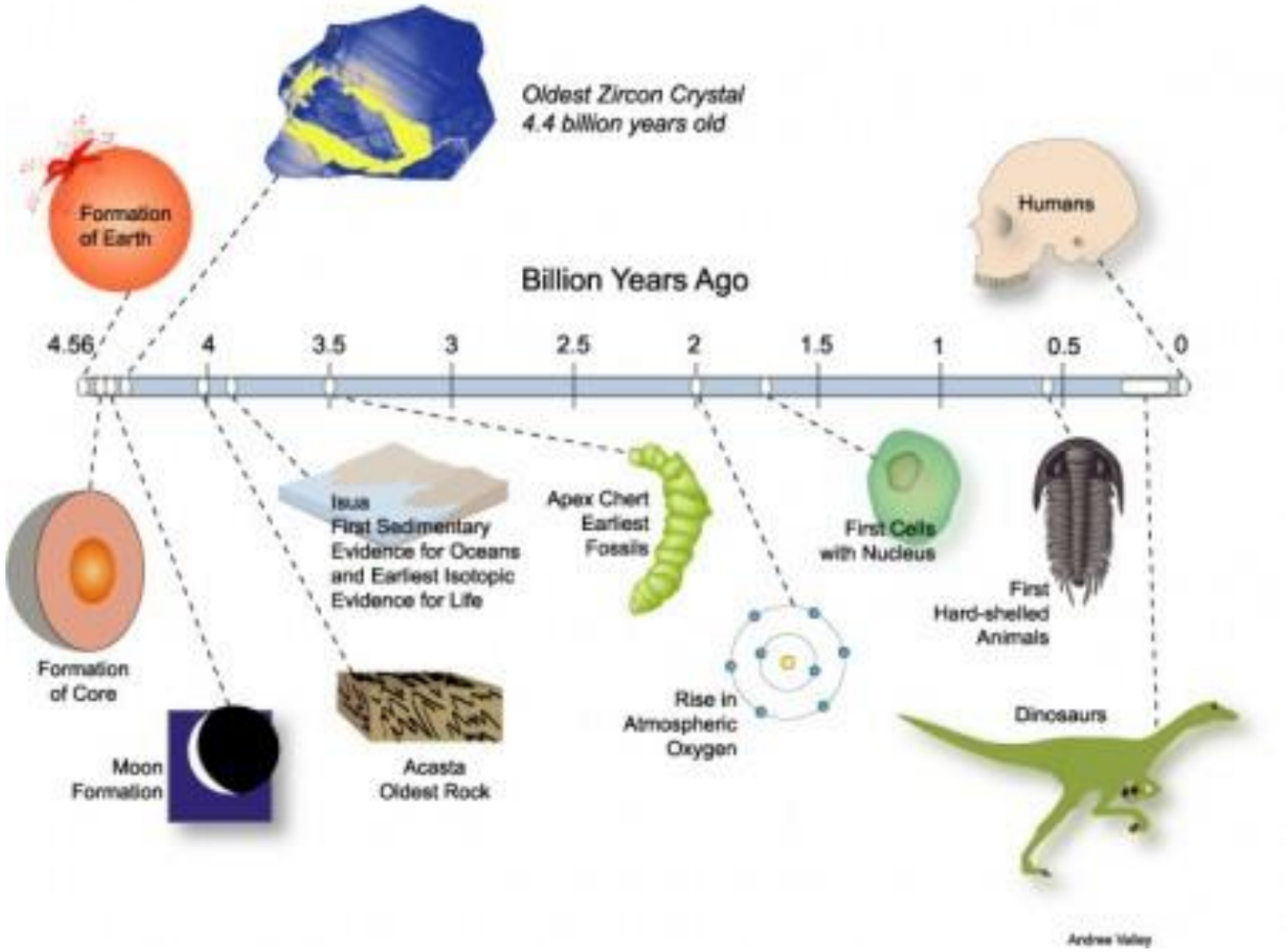
Bioinformatika 1

Filogenija

Mirjana Domazet-Lošo
FER, 2020./2021.

Sličnost među organizmima

- sličnosti među organizmima temeljene na morfološkim i molekularnim podatcima
- pretpostavka o zajedničkom pretku (engl. *common ancestor*) svih organizama na Zemlji

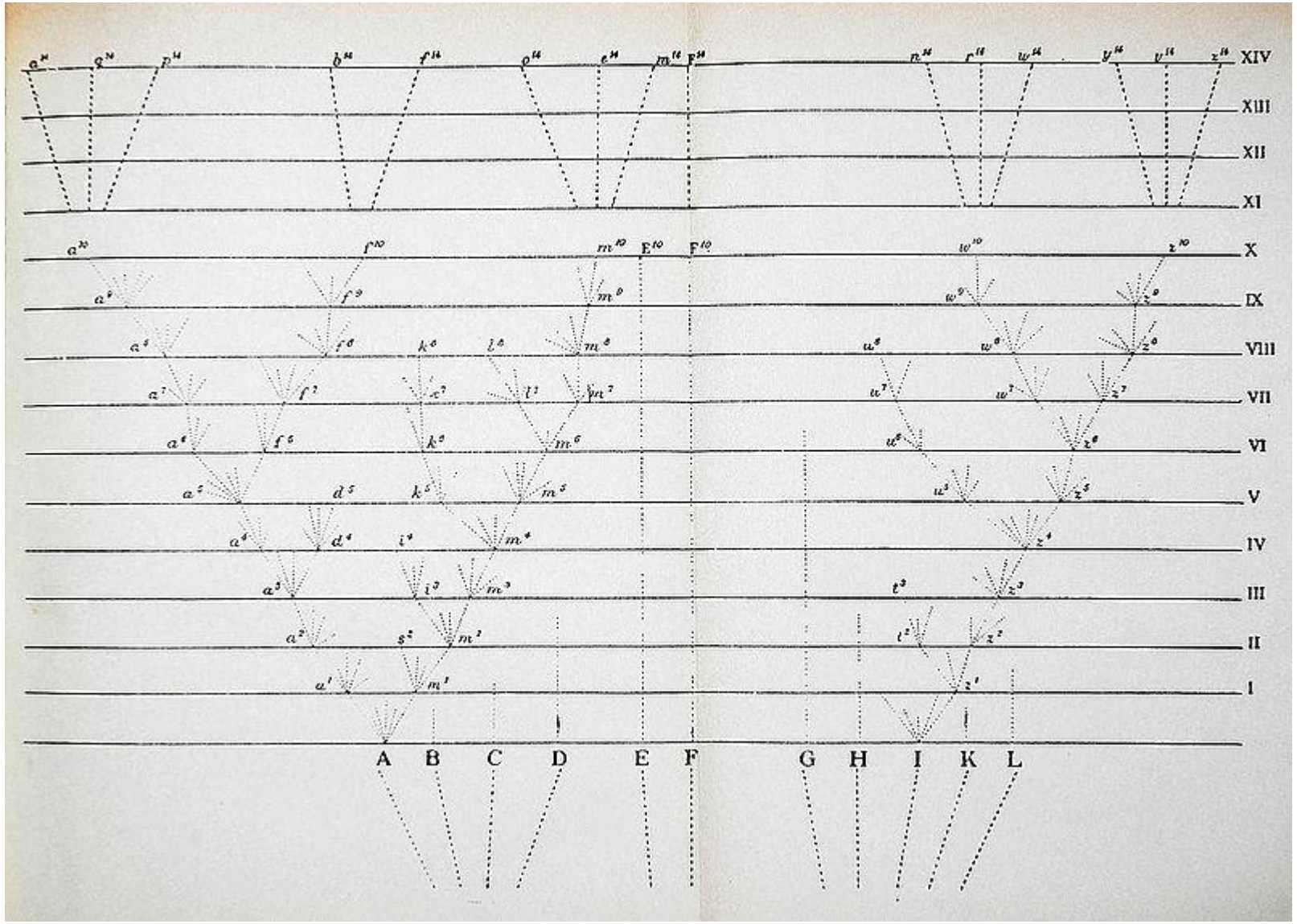


Urbano, L., 2010. Toilet Paper Timeline of Earth History

Povijesni razvoj ideja (1)

- C. Darwin (1809. –1882.) i A. R. Wallace (1823. – 1913.)
  - *On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection.* **1858.** Journal of the Proceedings of the Linnean Society of London. Zoology 3: 45-50.

- C. Darwin: *On the Origin of Species,* **1859.**
  - puni naziv 1. izdanja: *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*

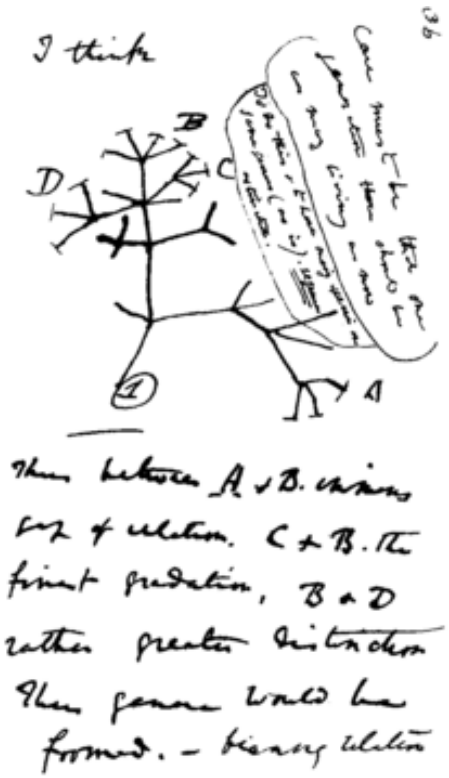


C. Darwin, *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*, 1859.

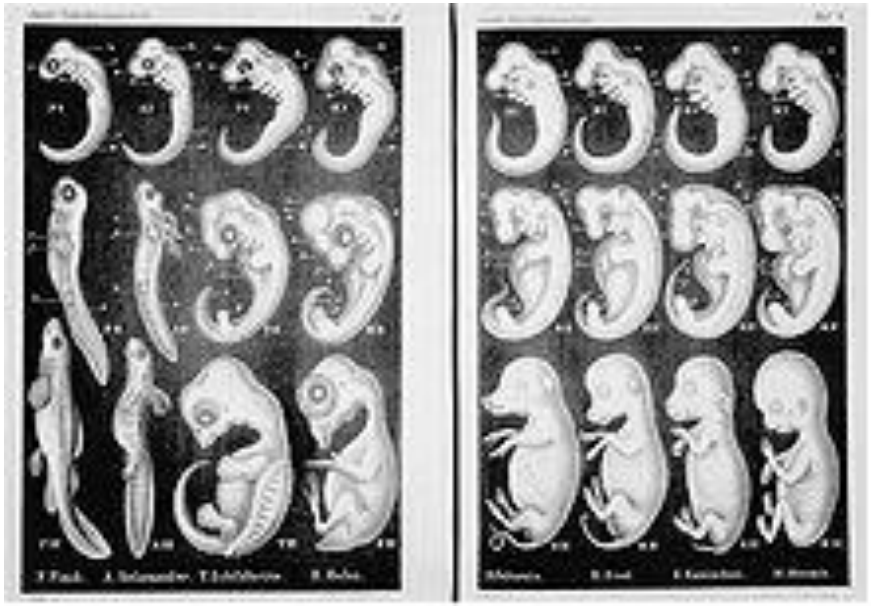
Povijesni razvoj ideja (2)

- G. Mendel (1822. – 1884.)
  - istraživanje nasljednih osobina
  - organizmi nasljeđuju osobine preko jedinica za nasljeđivanje koje danas nazivamo genima (dominantni i recesivni gen)
  - *Versuche über Pflanzenhybriden* (Eksperimenti u hibridizaciji biljaka), **1866.** (objavljeno u znanstvenom časopisu Prirodoslovnog društva iz Brna)

- E. Haeckel (1834. – 1919.)
  - **1866.** – uveo je pojam filogenija (i ekologija)
  - rekapitulacijska teorija (ontogenija je rekapitulacija filogenije)



Darwin, evolucijsko stablo (1837)



Haeckel, *Anthropogenie*,
prikaz embrionalnog razvoja (1874)

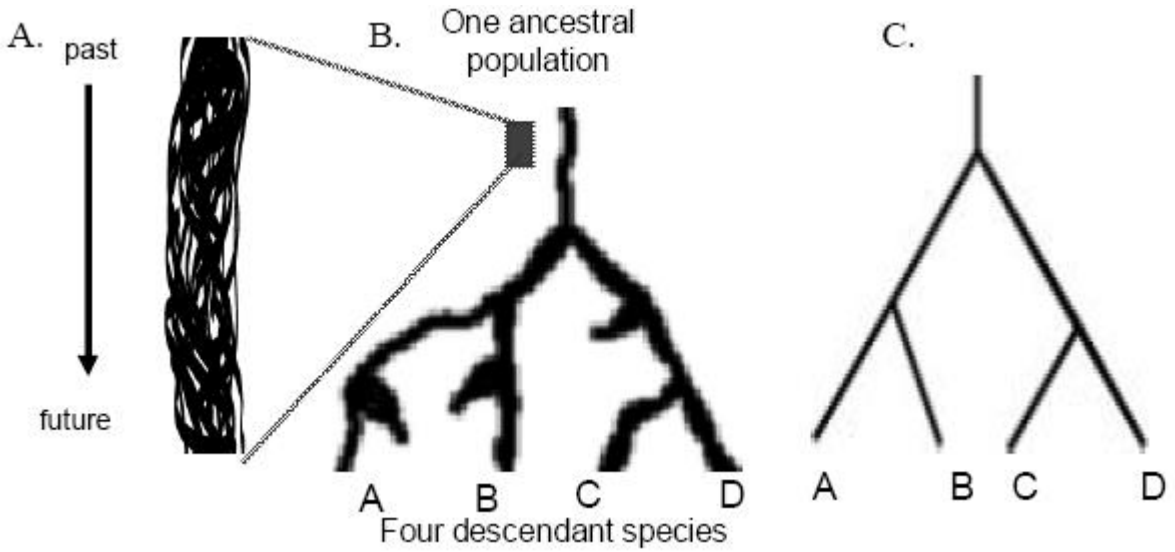
Filogenija i ontogenija (1)

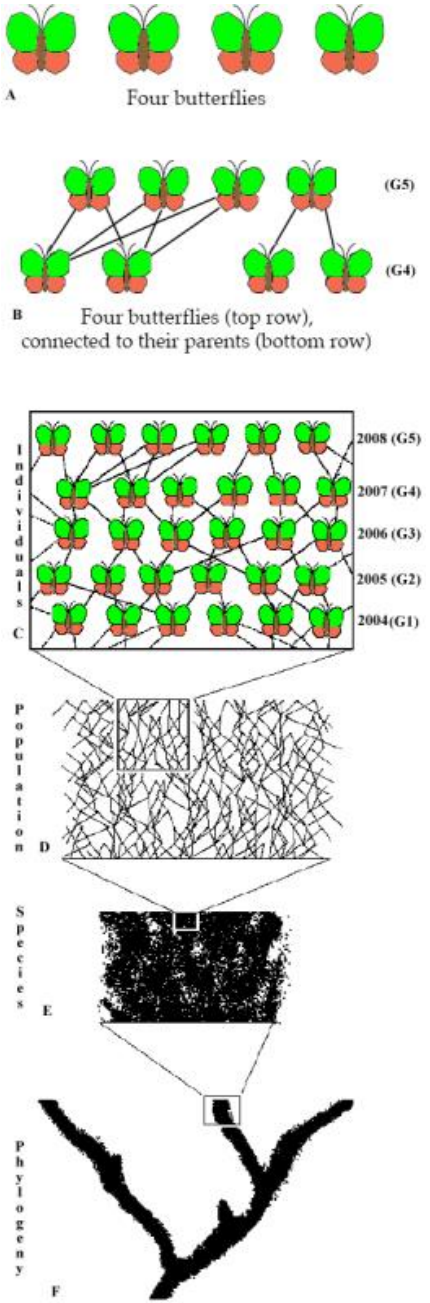
- Filogenija
  - prikazuje evolucijske odnose između vrsta (ili nekih drugih taksonomskih jedinica organizama)
  - slijed događaja uključenih u evoluciju neke vrste (ili druge taksonomske jedinice organizama)
  - pretpostavka o zajedničkom pretku

Filogenija i ontogenija (2)

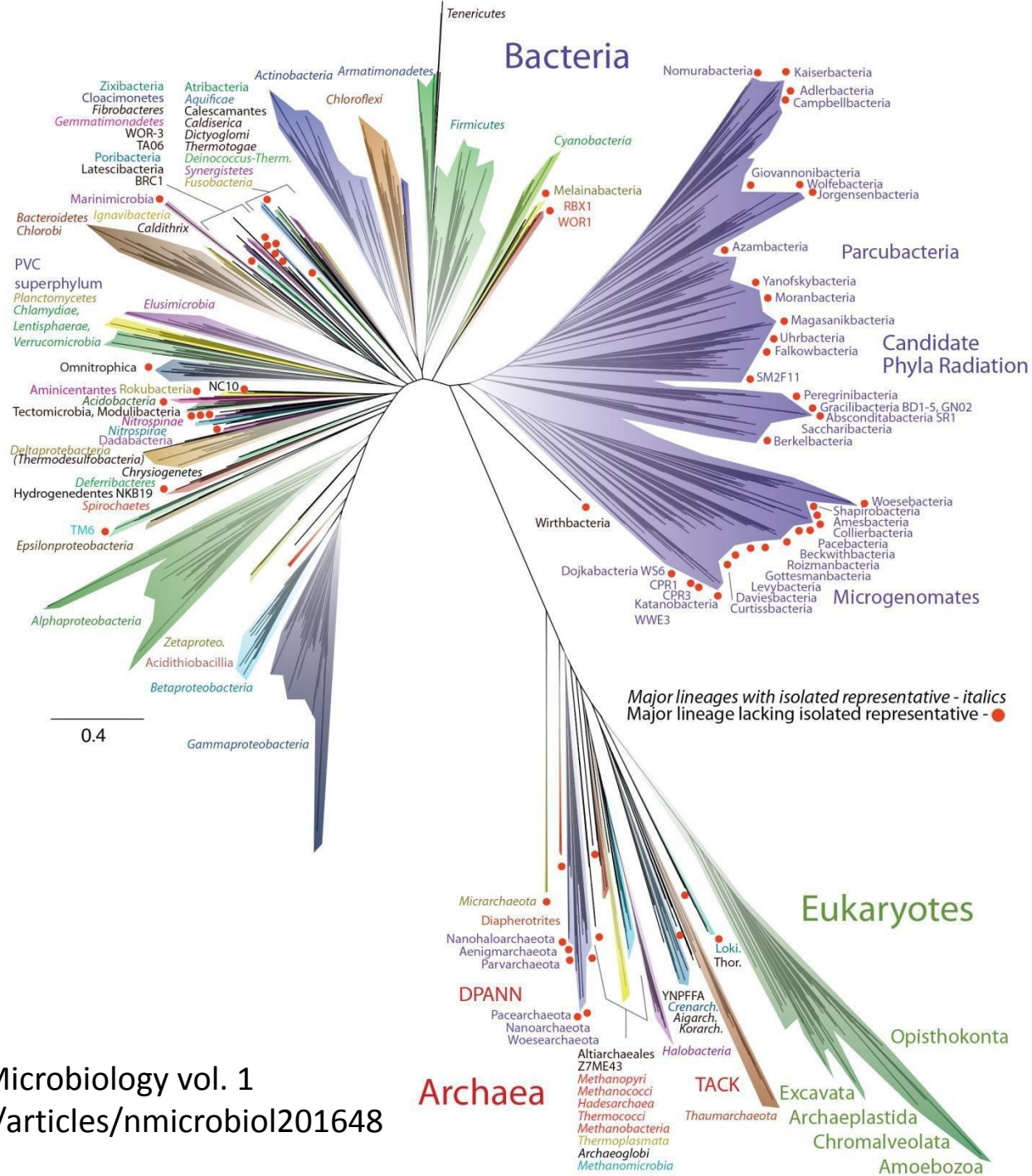
- Ontogenija
  - životni ciklus jedinke; evolucija jedinke
  - danas: smatra se da postoji složena veza između filogenije i ontogenije

- Taksonomija
  - izučava klasifikaciju organizama temeljenu na filogenetskim odnosima
  - takson: hijerarhijska jedinica





Baum, D. (2008) *Reading a phylogenetic tree: The meaning of monophyletic groups.* Nature Education 1(1):190



Hug *et al.* 2016 Nature Microbiology vol. 1
http://www.nature.com/articles/nmicrobiol201648

Stupanj	**Čovjek**	**Vinska mušica**	**Grašak**	***E. coli***
Domena	Eukarya	Eukarya	Eukarya	Bacteria
Carstvo	Animalia	Animalia	Plantae	Bacteria
Koljeno/ odjeljak	Chordata	Arthropoda	Magnoliophyta	Proteobacteria
Potkoljeno/ pododjeljak	Vertebrata	Hexapoda	Magnoliophytina	
Razred	Mammalia	Insecta	Magnoliopsida	γ-Proteobacteria
Podrazred	Placentalia	Pterygota	Magnoliidae	
Red	Primate	Diptera	Fabales	Enterobacteriales
Podred	Haplorrhini	Brachycera	Fabineae	
Porodica	Hominidae	Drosophilidae	Fabaceae	Enterobacteriaceae
Potporodica	Homininae	Drosophilinae	Faboideae	
Rod	*Homo*	*Drosophila*	*Pisum*	*Escherichia*
Vrsta	*H. sapiens*	*D. melanogaster*	*P. sativum*	*E. coli*

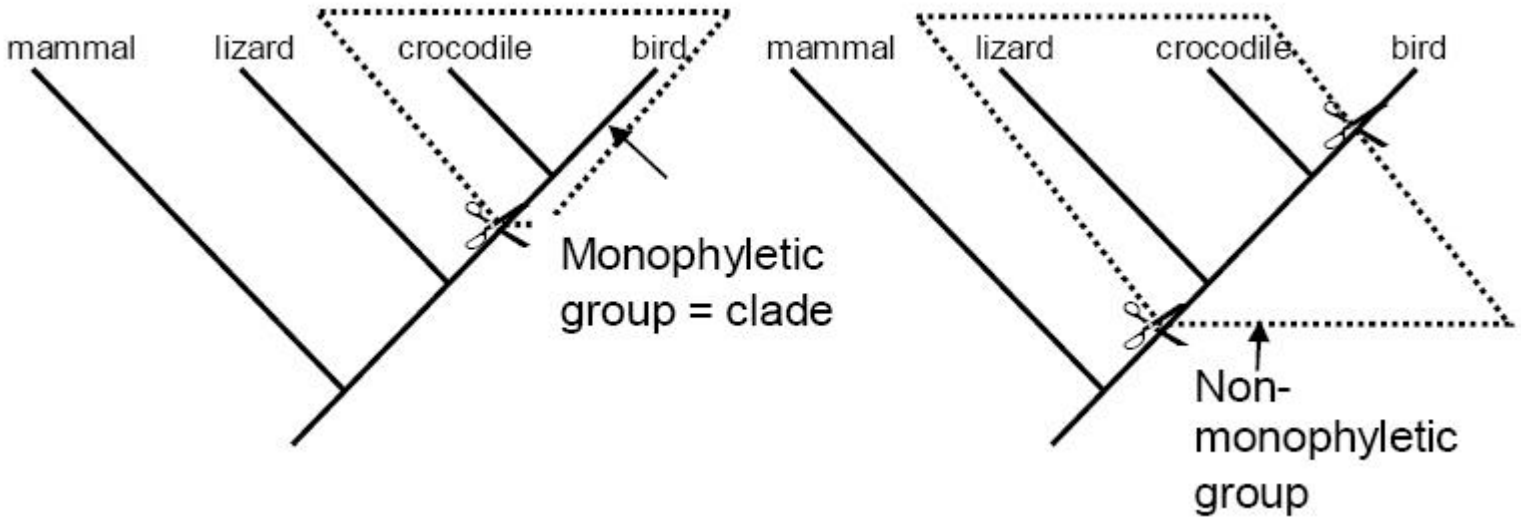
http://hr.wikipedia.org/wiki/Sistematika

Filogenetsko stablo i kladogram

- filogenetsko stablo - koristi se za prikaz srodstvenih odnosa među organizmima (topologija + vrijeme)
  - duljina grana je proporcionalna broju evolucijskih promjena
  - *Watch antibiotic resistance evolve* (Science News – YouTube) https://m.youtube.com/watch?v=yybsSqcB7mE

- kladistika – određivanje odnosa između organizama na temelju sličnosti
- kladogram – (*stablast*) dijagram koji prikazuje evolucijske odnose utvrđene kladističkom analizom (topologija)
  - organizmi ili njihovi biološki sljedovi predstavljeni su kao listovi stabla
  - iz svakog čvora izlaze dvije grane
  - klada – poddrvo

Monofiletska skupina



Baum, D. (2008) *Reading a phylogenetic tree:*
The meaning of monophyletic groups. Nature Education 1(1):190

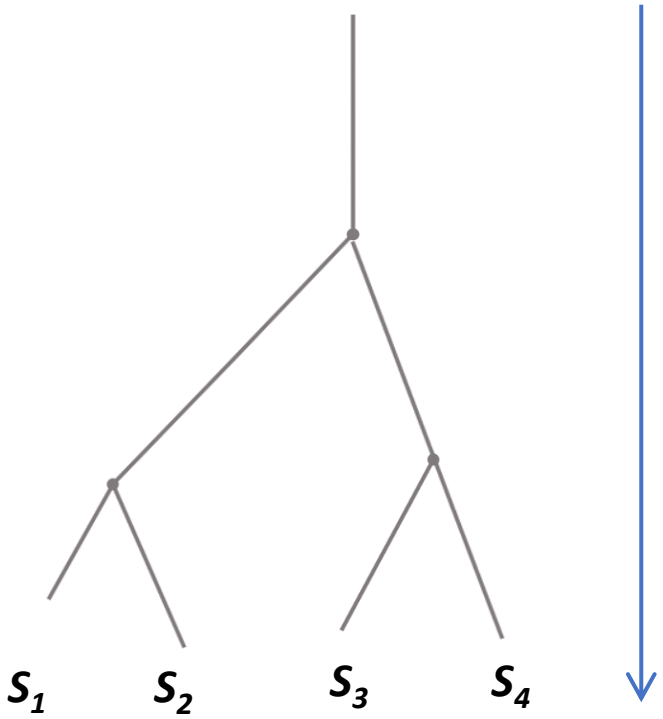
Tipovi filogenetskih stabala

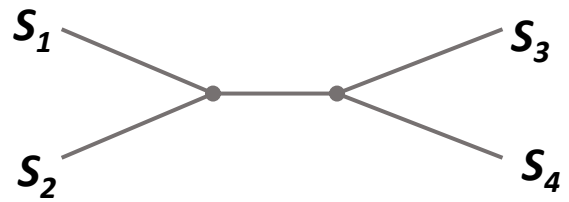
- ukorijenjena (engl. *rooted*)
  - pokazuje smjer evolucijskog procesa određujući odnos predak – potomak
  - ako se nukleotidni/aminokiselinski slijed nekog organizma ne može promatrati, onda pripadajući unutarnji čvor predstavlja hipotetski takson

- neukorijenjena (engl. *unrooted*)
  - nije određen niti korijen, niti smjer evolucije
  - prikazuje samo relativne odnose

ukorijenjeno stablo (engl. *rooted tree*)

neukorijenjeno stablo (engl. *unrooted tree*)



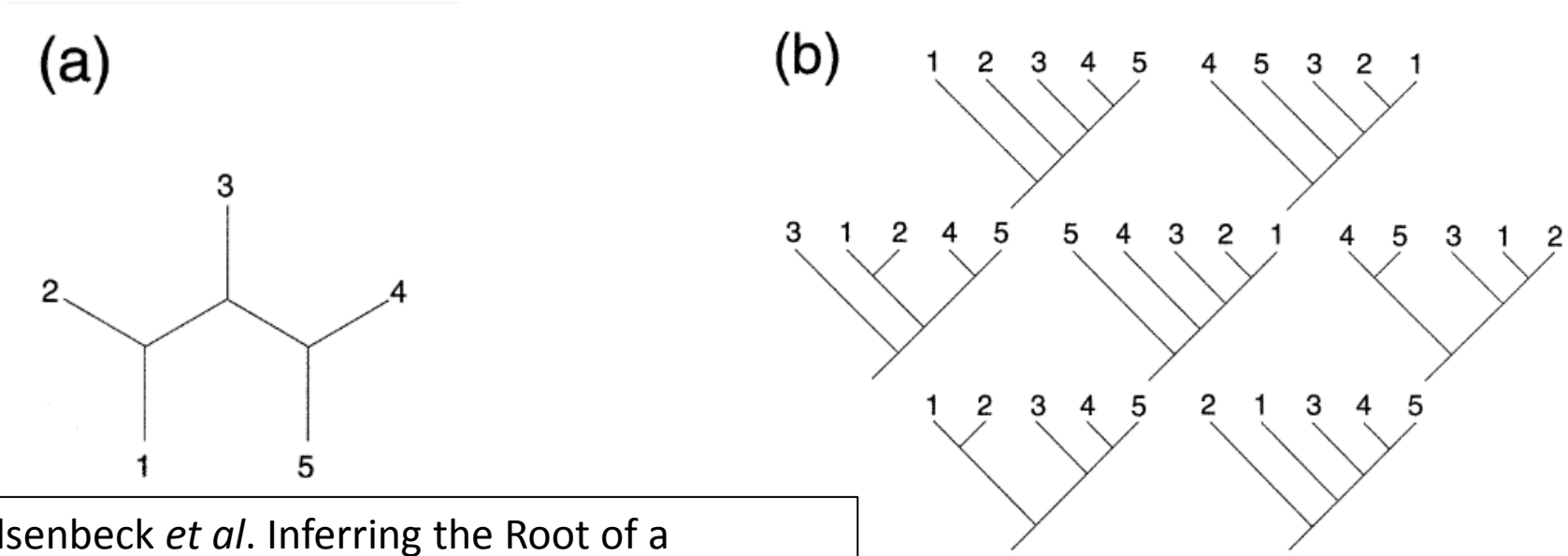


<u>Duljine grana</u> predstavljaju mjeru evolucije, a obično su izražene kao očekivani relativni broj supstitucija.

vrijeme diferencijacije (engl. *differentiation time*)

Dodavanje korijena u neukorijenjeno stablo (1)

1. dodaje se vanjska taksonomska jedinica (engl. *outgroup; outgroup rooting*)
   - taj takson treba biti najudaljeniji takson u odnosu na sve ostale taksone u početnom neukorijenjenom stablu



Huelsenbeck *et al*. Inferring the Root of a Phylogenetic Tree. Syst. Biol.51(1):32–43, 2002.

Dodavanje korijena u neukorijenjeno stablo (2)

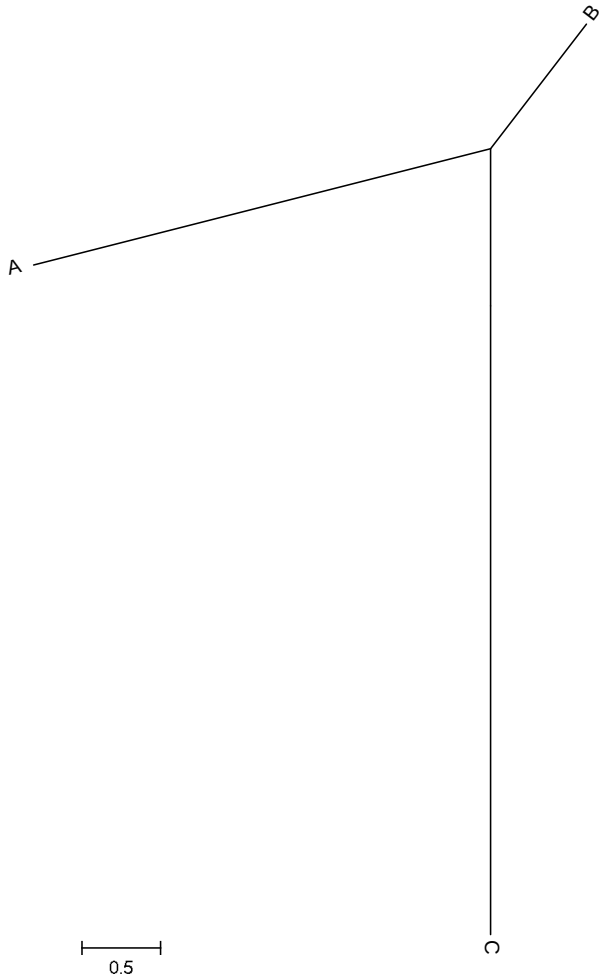
2. dodaje se korijen na sredinu između grana koje povezuju 2 najudaljenija lista postojećeg neukorijenjenog stabla (engl. *mid-point rooting*)
   - pretpostavka: vrijedi princip molekularnog sata, tj.
     u svim je granama ista konstantna evolucijska stopa
   - metoda je neprikladna, ako evolucijska stopa nije konstanta u svim granama stabla

A)

	A	B	C
A	0	4	8
B	4	0	6
C	8	6	0

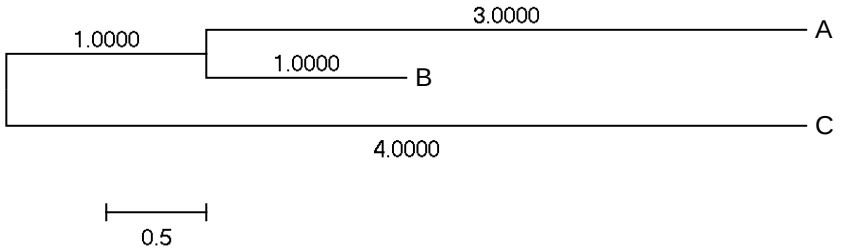
	A	B	C	D
A	0	4	8	12
B	4	0	6	10
C	8	6	0	12
D	12	10	12	0

B)



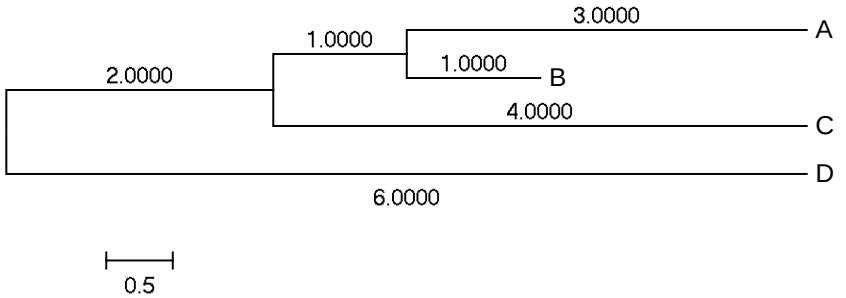
C)

mid-point rooting



D)

outgroup routing



Metode za izgradnju filogenetskih stabala

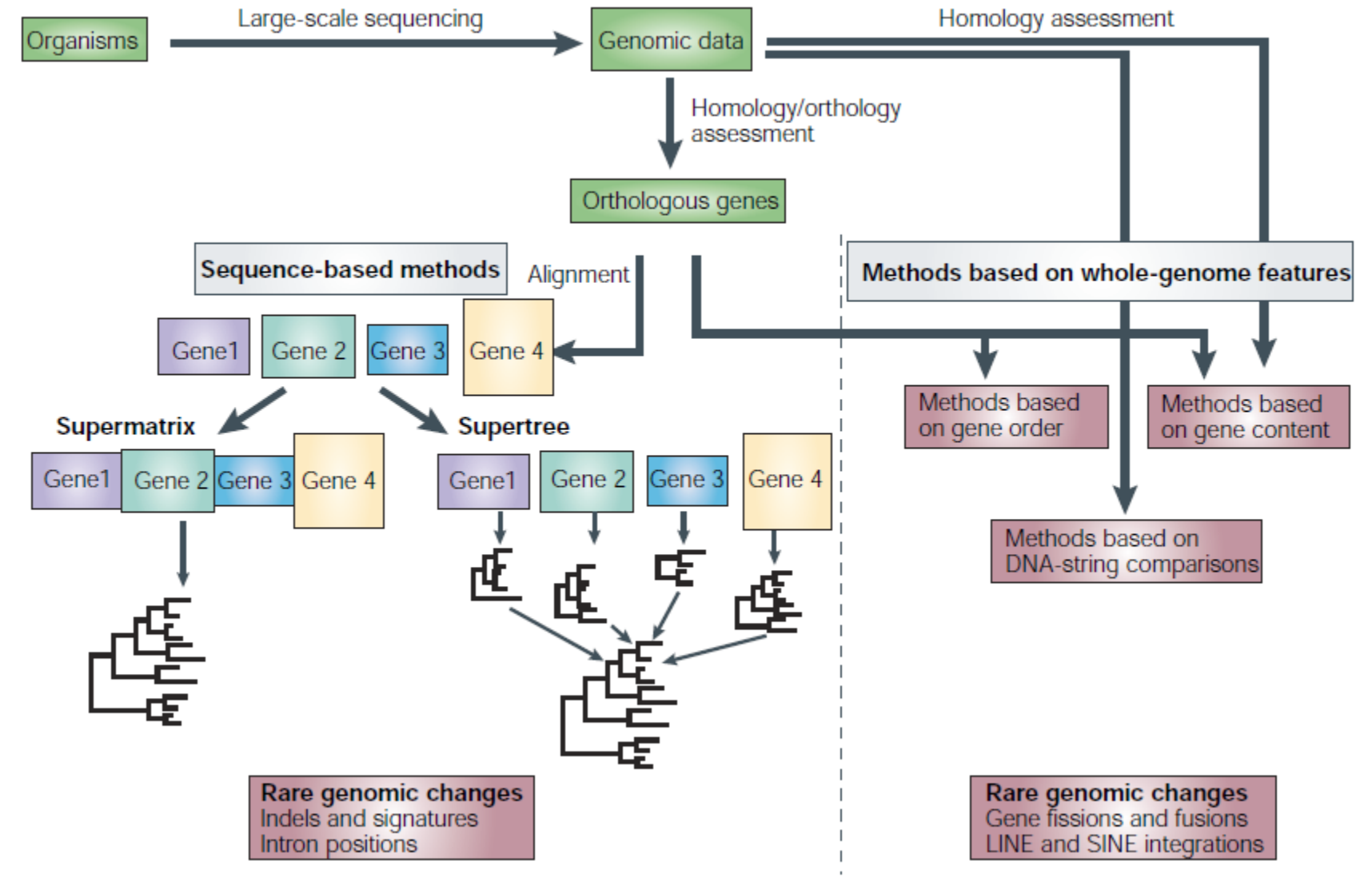
- izgradnja stabla temeljem homolognih gena

  → izgradnja genskog/proteinskog stabla (engl. *gene/protein tree*)

  - geni imaju zajedničkog pretka
  - ortologni geni (engl. *orthologous gene*) – nastali specijacijom
  - paralogni geni (engl. *paralogous gene*) – nastali duplikacijom

- stabla nastala na temelju cijelih genoma (engl. *species tree*)

  - mogu se razlikovati od stabala izgrađenih temeljem skupova gena ili proteina
  - glavni razlog: horizontalni prijenos gena (HGT; engl. *horizontal gene transfer*)



Delsuc *et al.* Phylogenomics and the reconstruction of the tree of life. *Nat Rev Genetics* 2005

Broj stabala u ovisnosti o broju taksona

- n – broj taksona
- N_U - broj mogućih neukorijenjenih stabla
- N_R - broj mogućih ukorijenjenih stabla
- N_U za $n \geq 3$
  (Cavalli-Sforza & Edwards, 1967):

$$\boldsymbol{N_U} = \frac{(2n-5)!}{2^{n-3}(n-3)!}$$

- N_R za $n \geq 2$:

$$\boldsymbol{N_R} = \frac{(2n-3)!}{2^{n-2}(n-2)!}$$

n	N_R	N_U
2	1	1
3	3	1
4	15	3
5	105	15
10	3.4×10^7	2×10^6
20	8×10^{21}	2×10^{20}
50*	2.8×10^{76}	3×10^{74}

***Broj protona u svemiru: ~ 10^{79} (Pevsner, 2009)

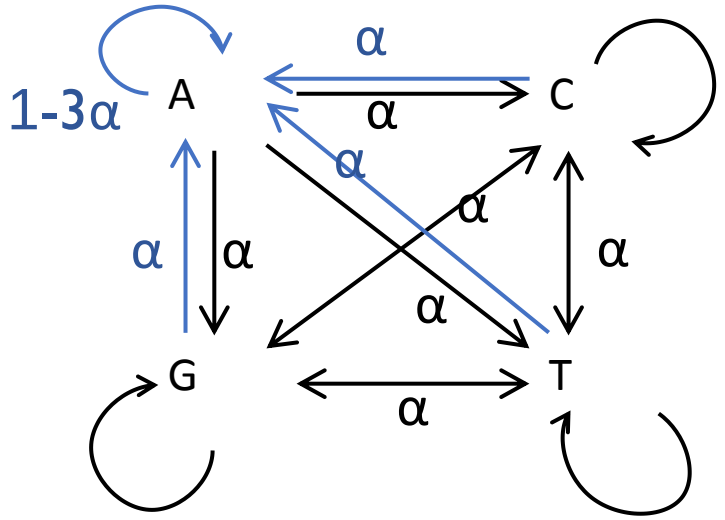
Evolucijski DNA modeli (1)

- broj promjena koji je vidljiv je u pravilu manji od stvarnog broja promjena/supstitucija (engl. *substitution rate*)
- supstitucijske modele razlikujemo prema parametrima
- **Jukes-Cantorov model** (Jukes i Cantor, 1969)
  - najjednostavniji model: pretpostavlja se jednaka frekvencija i jednaka mutacijska stopa svih nukleotida te neovisnost između nukleotida → *i.i.d*. (*independently identically distributed*)
  - evolucijska udaljenost ***d*** između 2 slijeda (*p* - relativni broj nukleotida po kojima se sljedovi razlikuju):

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

Jukes-Cantorov model (1)

- α – stopa mutacije (engl. *substitution rate*) u jedinici vremena
- pretpostavka: početno stanje je A, tj. za $t = 0$: $P_{A(0)} = 1$
- u trenutku $t = 1$, $P_{AA(1)} = P_{A(1)} = p_{AA}(t=1) = 1 - 3\alpha$
  - $P_{AA(1)}$ znači da je početno stanje A i da je trenutno (ovdje za t=1) isto A
  - ukupna vjerojatnost svih supstitucija X → Y (X ≠ Y): 3α
    ($p_{AC}(t=1) = p_{AG}(t=1) = p_{AT}(t=1) = \alpha$)



Jukes-Cantorov model (2)

- $p_{ij}(t)$ – vjerojatnost supstitucije $i \rightarrow j$ ovisno o vremenu t
- $p_{ij}(t)$ – vjerojatnost da će nukleotid biti j u trenutku t, uz pretpostavku da je bio i u $t = 0$

***P*(*t*)** – matrica vjerojatnosti prijelaza (engl. *transition probability matrix*)

$$p_{ij}(t) = \frac{1}{4} - \frac{1}{4}e^{-4\alpha t}$$
$$p_{ii}(t) = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

	A	**T**	**G**	**C**
A	$1 - 3p_{ij}(t)$	$p_{ij}(t)$	$p_{ij}(t)$	$p_{ij}(t)$
T	$p_{ij}(t)$	$1 - 3p_{ij}(t)$	$p_{ij}(t)$	$p_{ij}(t)$
G	$p_{ij}(t)$	$p_{ij}(t)$	$1 - 3p_{ij}(t)$	$p_{ij}(t)$
C	$p_{ij}(t)$	$p_{ij}(t)$	$p_{ij}(t)$	$1- 3p_{ij}(t)$

Jukes-Cantorov model (3)

- α – stopa supstitucije nukleotida X u Y, tj. X → Y
- $P_{A(t)}$ – vjerojatnost da je nukleotid u stanju A u trenutku *t*
- pretpostavka, tj. početni uvjet: $P_{A(0)} = 1$
- određujemo $P_{A(t+1)}$ (vjerojatnost da je nukleotid u stanju A u *t*+1):

$$P_{A(t+1)} = (1 - 3\alpha)\, P_{A(t)} + \alpha(1 - P_{A(t)})$$

$$\Delta P_{A(t)} = P_{A(t+1)} - P_{A(t)} = -3\alpha\, P_{A(t)} + \alpha(1 - P_{A(t)}) = -4\alpha\, \mathbf{P_{A(t)}} + \alpha$$

kraće:

$$dp / dt = -4\alpha p + \alpha$$

$$dt = dp / (-4\alpha p + \alpha)$$

$$\int dt = \int \frac{dp}{-4\alpha p + \alpha} \rightarrow t = \frac{-1}{4\alpha} \ln(-4\alpha p + \alpha) + C \quad (1)$$

Jukes-Cantorov model (4)

- određivanje konstante C:

$$t = \frac{-1}{4\alpha}\ln(-4\alpha p + \alpha) + C$$

- početni uvjet: $P_{A(0)} = 1$

$$C = t + \frac{1}{4\alpha}\ln(-4\alpha p + \alpha)$$

$$= |t = 0, p = 1| = \frac{1}{4\alpha}\ln(-4\alpha + \alpha) = \frac{1}{4\alpha}\ln(-3\alpha) \quad (2)$$

- uvrstimo (2) u (1):

$$\rightarrow t = \frac{-1}{4\alpha}\ln(-4\alpha p + \alpha) + \frac{1}{4\alpha}\ln(-3\alpha)$$

$$\rightarrow -4\,\alpha\, t = \ln(-4\alpha p + \alpha) - \ln(-3\alpha) = \ln\left(\frac{-4\alpha p + \alpha}{-3\alpha}\right) = \ln\left(\frac{-4p+1}{-3}\right)$$

$$\rightarrow p = P_{A(t)} = \frac{1}{4} + \frac{3}{4}e^{-4\alpha t}$$

Jukes-Cantorov model (5)

- $P_{A(t)} = P_{AA(t)} = \boldsymbol{p_{ii}(t)} = \frac{1}{4} + \frac{3}{4} e^{-4\alpha t}$ (3)

- $P_{AY(t)} = 1 - P_{AA(t)} = \frac{3}{4} - \frac{3}{4} e^{-4\alpha t}$ (za sve nukleotide Y ≠ A zajedno)
  ili
  $P_{AC(t)} = P_{AG(t)} = P_{AT(t)} = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t}$

Odnosno:

$$\boldsymbol{p_{ij}(t)} = \frac{1}{4} - \frac{1}{4} e^{-4\alpha t} \qquad (4)$$

Jukes-Cantorov model (6)

Određivanje evolucijske udaljenosti između 2 slijeda

- pretpostavimo da je zajednički predak 2 slijeda u $t = 0$ na nekom mjestu u slijedu bio u stanju A
- neka je $I_{(t)}$ vjerojatnost da su u oba slijeda-potomka isti nukleotidi na promatranom mjestu u trenutku t
- u trenutku t:
  - svaki od sljedova-potomaka bit će u stanju A s $P_{AA(t)}$, odnosno oba zajedno s $P_{AA(t)}^2$
  - analogno, oba će biti u stanju C s $P_{AC(t)}^2$, itd.
- koristimo (3) i (4):

$$I_{(t)} = P_{AA(t)}^2 + P_{AC(t)}^2 + P_{AG(t)}^2 + P_{AT(t)}^2 = \frac{1}{4} + \frac{3}{4} e^{-8\alpha t} \qquad (5)$$

Jukes-Cantorov model (7)

- vjerojatnost da su u sljedovima-potomcima na promatranom mjestu različiti nukleotidi: $p = 1 - I_{(t)} = \frac{3}{4}(1 - e^{-8\alpha t})$

$$8\alpha t = -\ln\left(1 - \frac{4}{3}p\right) \qquad (6)$$

- 3αt – broj supstitucija u vremenu t na jednoj poziciji u jednom slijedu
- *d* – broj supstitucija na jednoj poziciji od divergencije promatranoga para sljedova: $d = 2 \cdot 3\alpha t$

$$\boldsymbol{d = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)} \qquad (7)$$

Evolucijski DNA modeli (2)

- **Kimurin model** (Kimura, 1980)
  - model razlikuje
    tranziciju (purin ↔ purin: A ↔ G; pirimidin-pirimidin: C ↔ T)
    i transverziju (purin ↔ pirimidin)
  - evolucijska udaljenost *d* između 2 slijeda (*p* - relativan broj tranzicija; *q* – relativan broj transverzija):

$$d = -\frac{1}{2}\ln(1-2p-q) - \frac{1}{4}ln(1-2q)$$

- **GTR model** (engl. *General Time Reversible*) (Tavaré, 1986)
  - frekvencije nukleotida se međusobno razlikuju (4 parametra)
  - 6 parametara za različite supstitucije

Metode za izgradnju filogenetskog stabla

- <u>Metode temeljene na udaljenosti:</u>
  manja udaljenost → veća evolucijska povezanost
  - UPGMA (***U****nweighted* ***P****air* ***G****roup* ***M****ethod with* ***A****rithmetic Mean*)
    (Sokal i Michener, 1958)
  - metoda povezivanja susjeda (engl. *neighbour-joining*; NJ)

- <u>Metode temeljene na obilježjima</u>
  (npr. znakovima: nukleotidima/aminokiselinama, ali mogu biti i neka druga obilježja)
  - princip najmanjeg broja evolucijskih promjena (engl. *maximum parsimony*)
  - metoda najveće izglednosti (engl. *maximum likelihood*)

Mjera udaljenosti

- da bi *M* bila mjera udaljenosti mora ispunjavati sljedeće uvjete:
  - udaljenost od nekog taksona *i* do sebe samog je 0: $M_{ii} = 0$
  - simetričnost: $M_{ij} = M_{ji}$ za $i \neq j$
  - nejednakost trokuta: $M_{ij} + M_{jk} \geq M_{ik}$

- dodatni uvjet za aditivnost:
  - nejednakost četverokuta: $M_{ik} + M_{jl} = M_{il} + M_{jk} \geq M_{ij} + M_{kl}$

UPGMA (1)

- pretpostavka: vrijedi molekularni sat
  - konstantna evolucijska stopa u cijelom stablu
  - udaljenost od svakog lista do korijena je jednaka
- najjednostavija metoda
  - koristi se za izgradnju inicijalnog stabla (engl. *guide tree*) kao pomoć drugim metodama, npr. kod izgradnje MSA
- vremenska složenost: $O(n^2)$ za n taksona

UPGMA (2)

- Ideja: u svakom koraku se dvije najsličnije skupine (engl. *cluster*) spajaju u jednu (skupina može imati samo jedan takson)
  - 2 mogućnosti: ili se postojećoj skupini dodaje novi takson ili se dva taksona međusobno spajaju
  - udaljenost između skupina - prosjek svih međusobnih udaljenosti članova skupine A i članova skupine B:

$$\frac{1}{|A||B|}\sum_{i\in A}\sum_{j\in B} d_{i,j}$$

UPGMA (3)

za svaki takson i /* inicijalizacija */

C_i = {i} /* inicijalno svaka skupina C_i sadrži samo i-ti takson */

$d(C_i, C_j) = d_{i,j}$

$h(i) = 0$ /* visina i-tog taksona u stablu */

kraj

Ponavljaj sve dok ima taksona koji nisu dodani u stablo

Između svih skupina pronaći C_i i C_j tako da je $d(C_i, C_j)$ minimalna

Dodati novu skupinu C_k koja zamjenjuje C_i i C_j

Dodati u stablo novi čvor $N_{i,j}$ tako da je visina $h(N_{i,j}) = d(C_i, C_j) / 2$

$d(C_i, N_{i,j}) = h(N_{i,j}) - h(C_i)$ /* povezati C_i i $N_{i,j}$*/

$d(C_j, N_{i,j}) = h(N_{i,j}) - h(C_j)$ /* povezati C_j i $N_{i,j}$*/

za sve $C_l \neq C_k$

$d(C_k, C_l) = (|C_i| \cdot d(C_i, C_l) + |C_j| \cdot d(C_j, C_l)) / (|C_i| + |C_j|)$

kraj

kraj

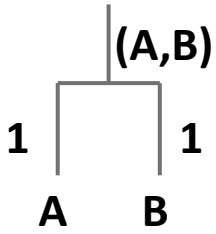
Primjer: UPGMA (1)

1. korak

	A	B	C	D
A	0	**2**	4	6
B	**2**	0	8	10
C	4	8	0	12
D	6	10	12	0

$d(C_k, C_l) = (|C_i| \cdot d(C_i, C_l) + |C_j| \cdot d(C_j, C_l)) / (|C_i| + |C_j|)$

$h(N_{A,B}) = d(A, B) / 2 = 2 / 2 = 1$

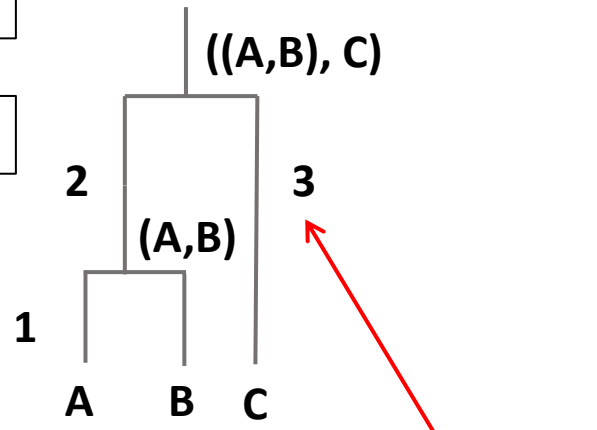


2. korak

	(A,B)	C	D
(A,B)	0	**6**	**8**
C	**6**	0	12
D	**8**	12	0

$(1 \cdot 4 + 1 \cdot 8) / 2 = 6$

$(1 \cdot 6 + 1 \cdot 10) / 2 = 8$



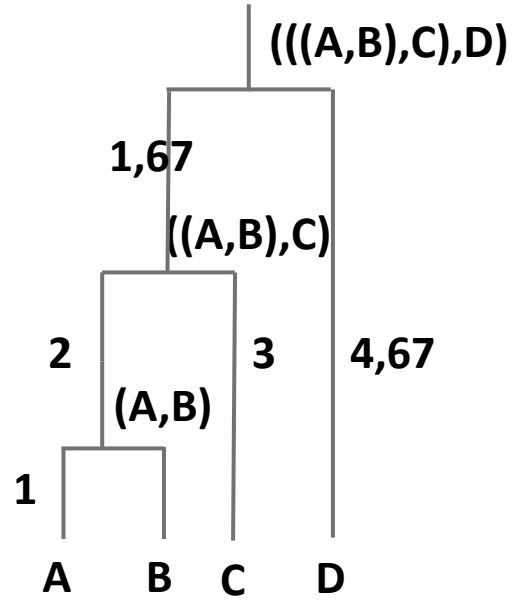
$h(N_{AB,C}) = d(AB, C) / 2 = 3$

Primjer: UPGMA (2)

3. korak

	((A,B),C)	**D**
((A,B),C)	0	**9,33**
D	**9,33**	0

$(2 \cdot 8 + 1 \cdot 12) / (2 + 1) = 9{,}33$



Povezivanje susjeda (1)

- *Neigbor-Joining* (Saitou i Nei, 1987) → neukorijenjeno stablo
  - određuje i topologiju i duljinu grana filogenetskog stabla
  - ova mjera minimizira zbroj duljina grana u svakom koraku, ali konačan rezultat ne mora biti minimalan zbroj duljina svih grana
  - vremenska složenost: $O(n^3)$ za n taksona

Povezivanje susjeda (2)

- Ideja:
  - prvo se generira inicijalno stablo koje uključuje svih *n* taksona
  - zatim se napravi $n(n - 1)/2$ usporedbi kako bi se pronašle dva najbliža taksona, npr. *A* i *B*, koji se zatim tretiraju kao jedan u daljnjim usporedbama
    → dva najbliža taksona određuju se odabirom najmanje vrijednosti M_{ij} iz matrice *M* (*objašnjeno na sljedećoj stranici*); npr. za par (A, B) to je vrijednost M_{AB}
  - → u stablo se dodaje novi čvor koji povezuje par (*A*, *B*)
    - određuju se udaljenost *A* i *B* od novog čvora
    - određuju se udaljenost ostalih čvorova (≠ *A*, *B*) od novog čvora
  - postupak se ponavlja dok se ne odrede sve grane, tj. dok *n* ne postane 2

Povezivanje susjeda (3)

- D_{ij} - originalne udaljenosti između taksona *i* i *j*
- S_i - zbroj udaljenosti čvora *i* do ostalih čvorova
  - kada se S_i podijeli s $n - 2$ dobijemo prosječnu "korigiranu" udaljenost
- M_{ij} - udaljenosti za izgradnju filogenetskog stabla → traži se minimum
- $S_i = \sum_{k=1}^{n} D_{ik}$
- $S_j = \sum_{k=1}^{n} D_{jk}$
- $M_{ij} = D_{ij} - \frac{1}{n-2}(S_i + S_j)$

Povezivanje susjeda (4)

- ako su, prema udaljenostima iz matrice *M,* najbliži taksoni *A* i *B*, onda treba spojiti *A* i *B* i zamijeniti ih s čvorom *X*
- udaljenost taksona *A* do novog čvora *X* (analogno i za *B*):

$$D_{AX} = \frac{1}{2} D_{AB} + \frac{1}{2(n-2)} (S_A - S_B)$$

- određujemo udaljenost ostalih taksona *Y*, $Y \neq A$ i $Y \neq B$, do novog čvora *X*:

$$D_{XY} = \frac{1}{2} (D_{AY} - D_{XA}) + \frac{1}{2} (D_{BY} - D_{XB}) = \frac{1}{2} (D_{AY} + D_{BY} - D_{AB})$$

- dalje računamo s čvorom *X*, a bez čvorova *A* i *B*

Primjer – metoda povezivanja susjeda (1)

1. korak

	A	***B***	***C***	***D***
A	0	6	10	6
B	6	0	8	10
C	10	8	0	12
D	6	10	12	0

$n = 4$, $S_A = 22$, $S_B = 24$, $S_C = 30$, $S_D = 28$

$M_{AB} = D_{AB} - (S_A + S_B) / 2 = 6 - 23 = -17$

$M_{AC} = -16$, $M_{AD} = \mathbf{-19}$ → odaberemo npr. $\boldsymbol{M_{AD}}$

$\boldsymbol{M_{BC}} = -19$, $M_{BD} = -16$

$M_{CD} = -17$

2. korak

	X	***B***	***C***
X	0	5	8
B	5	0	8
C	8	8	0

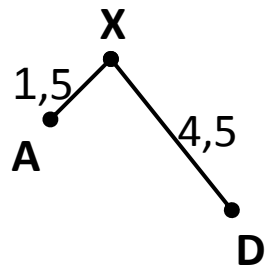
Uvodimo novi čvor **X:**

$D_{AX} = D_{AD} / 2 + (S_A - S_D) / 4 = 1,5$

$D_{DX} = D_{AD} / 2 + (S_D - S_A) / 4 = 4,5$

$D_{CX} = (D_{AC} + D_{DC} - D_{XA} - D_{XD}) / 2$
$= (10 + 12 - 1,5 - 4,5) / 2 = 8$

$D_{BX} = (D_{AB} + D_{DB} - D_{XA} - D_{XB}) / 2$
$= (6 + 10 - 1,5 - 4,5) / 2 = 5$



Primjer – metoda povezivanja susjeda (2)

3. korak

	X	***B***	***C***
X	0	5	8
B	5	0	8
C	8	8	0

$n' = 3$, $S_B = 13$, $S_C = 16$, $S_X = 13$

$M_{XB} = D_{XB} - (S_X + S_B) / 1 = 5 - (13 + 13) / 1 = \mathbf{-21}$

$M_{XC} = D_{XC} - (S_X + S_C) / 1 = 8 - (13 + 16) / 1 = -21$

$M_{BC} = D_{BC} - (S_B + S_C) / 1 = 8 - (13 + 16) / 1 = -21$

Uvodimo novi čvor **Y:**

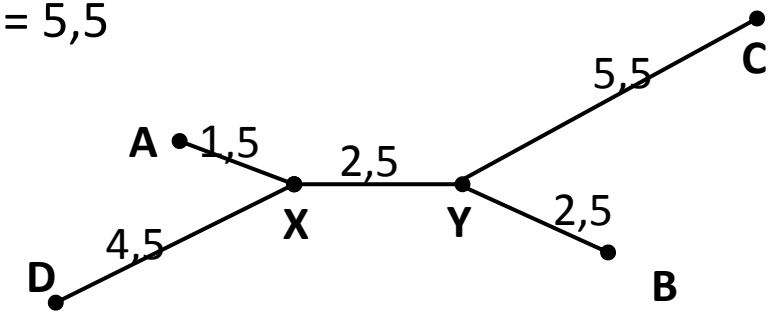
4. korak

	Y	***C***
Y	0	5,5
C	5,5	0

$D_{XY} = D_{BX} / 2 + (S_X - S_B) / 2 = 2,5$

$D_{BY} = D_{BX} / 2 + (S_B - S_X) / 2 = 2,5$

$D_{CY} = (D_{BC} + D_{XC} - D_{BX}) / 2 = 5,5$



Metoda najmanjeg broja evolucijskih promjena (1)

- metoda najmanjeg broja evolucijskih promjena/mutacija (engl. *maximum parsimony*)
  - princip Occamove oštrice (engl. *Occam razor's principle*)
    - najjednostavnije rješenje je vjerojatno najispravnije
  - pronaći filogenetsko stablo kojim se mogu opisati promatrani sljedovi tako da stablo uključuje minimalan broj evolucijskih promjena
  - NP-problem
  - obično se koristi za manje skupove sličnih sljedova
  - nema eksplicitne mjere udaljenosti

Metoda najmanjeg broja evolucijskih promjena (2)

- Ideja:
  - za zadani skup sljedova izgraditi sva moguća stabla te pronaći ono koje uključuje najmanji broj evolucijskih promjena
  - ulazni skup su poravnati sljedovi, tj. matrica *n* x *m*, gdje je *n* broj sljedova, a *m* duljina poravnatih sljedova
  - za svaki stupac u poravnanju odabrati ono stablo koje uključuje minimalan broj promjena
  - "pobjednik" je ono stablo za koje je ukupno (po svim stupcima) bilo najmanje promjena

Informativna mjesta

- promatramo stupce u poravnanju *n* sljedova
- informativni stupac u poravnanju (engl. *informative site*)
  i. kada su barem 2 vrste nukleotida u stupcu i
  ii. kada je svaka od tih vrsta nukleotida zastupljena u barem 2 slijeda u poravnanju
- <u>Primjer:</u>
  - stupci 1 i 3 su informativni
  - stupac 4 nije informativan, jer su svi nukleotidi u stupcu isti (tj. C)
  - stupac 2 nije informativan, jer pretpostavljamo

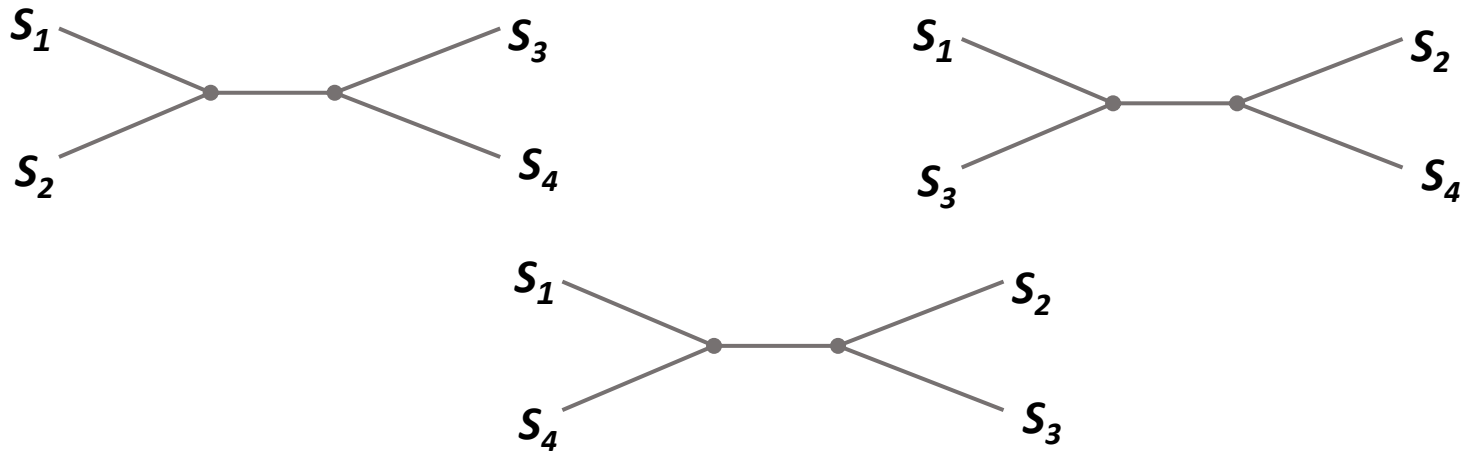
  da je u S_1 bila promjenu C → A
  (i svako od tri moguća stabla bi uključivalo jednu promjenu na tom mjestu)

	1	**2**	**3**	**4**
S_1	A	A	C	C
S_2	A	C	A	C
S_3	C	C	A	C
S_4	C	C	C	C

Primjer - metoda najmanjeg broja evolucijskih promjena (1)

	1	**2**	**3**
S_1	A	A	C
S_2	A	C	A
S_3	C	C	A
S_4	C	C	C

Tri moguća filogenetska stabla za zadani skup od 4 slijeda:



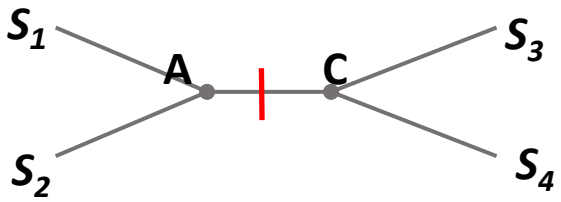
Primjer - metoda najmanjeg broja evolucijskih promjena (2)

	1	**2**	**3**
S_1	A	A	C
S_2	A	C	A
S_3	C	C	A
S_4	C	C	C

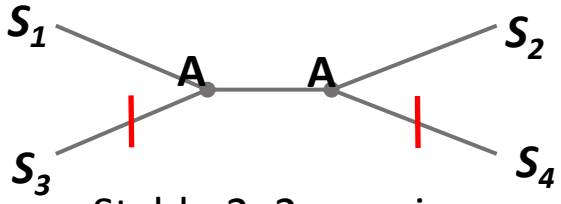
Težina stabla (engl. *parsimony score*)
→ 1 za supstituciju, a 0 inače

	1	**2**	**3**	**Ukupno**
Stablo 1	1			
Stablo 2	2			
Stablo 3	2			

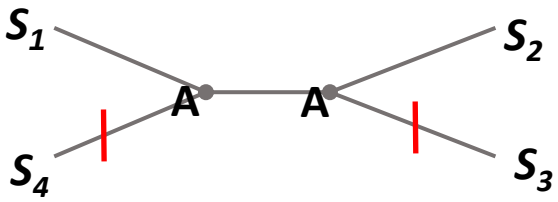
Promatramo tri moguća filogenetska stabla za **1. mjesto** u poravnanju:



Stablo 1: 1 promjena



Stablo 2: 2 promjene



Stablo 3: 2 promjene

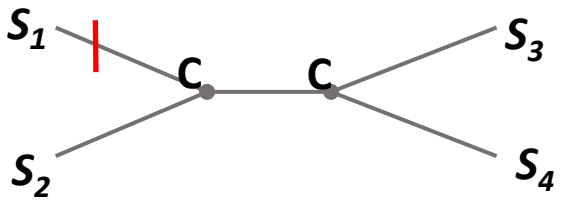
Primjer - metoda najmanjeg broja evolucijskih promjena (3)

	1	2	3
S_1	A	A	C
S_2	A	C	A
S_3	C	C	A
S_4	C	C	A

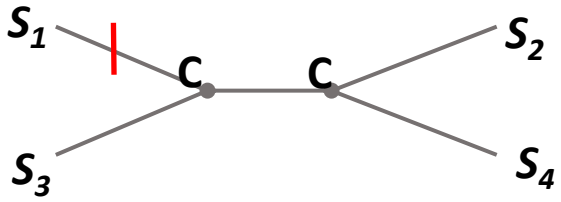
Stupac 2 nije informativan (tj. ista promjena u svim stablima), pa se može izostaviti iz analize.

	1	2	3	Ukupno
Stablo 1	1	1		
Stablo 2	2	1		
Stablo 3	2	1		

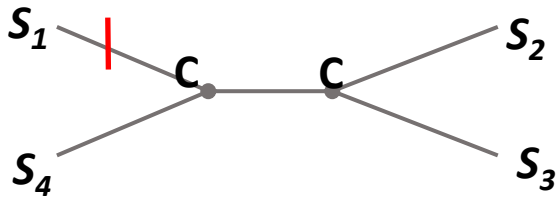
Promatramo tri moguća filogenetska stabla za **2. mjesto** u poravnanju:



Stablo 1: 1 promjena



Stablo 2: 1 promjena



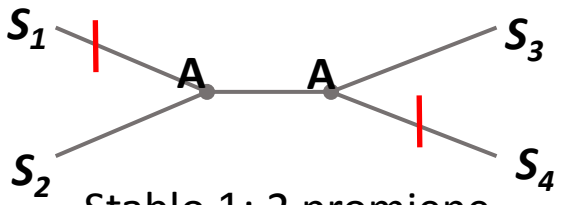
Stablo 3: 1 promjena

Primjer - metoda najmanjeg broja evolucijskih promjena (4)

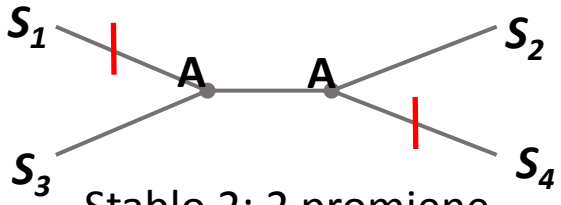
	1	**2**	**3**
S_1	A	A	C
S_2	A	C	A
S_3	C	C	A
S_4	C	C	C

	1	**3**	**Ukupno**
Stablo 1	1	2	**3**
Stablo 2	2	2	4
Stablo 3	2	1	**3**

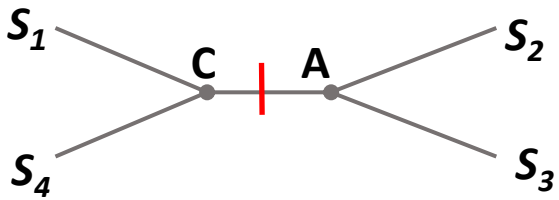
Promatramo tri moguća filogenetska stabla za **3. mjesto** u poravnanju:



Stablo 1: 2 promjene



Stablo 2: 2 promjene



Stablo 3: 1 promjena

Metoda najmanjeg broja evolucijskih promjena (3)

Problemi:

- mutacijska stopa različita po granama
- broj mogućih stabala raste eksponencijalno ovisno o broju sljedova
  → kako pronaći optimalno rješenje?
  - pretraživanje svih mogućih rješenja (engl. *exhaustive search*)
  - *granaj i ogradi* princip (engl. *branch and bound*)
  - heuristički pristup → pohlepni algoritam (engl. *greedy algorithm*)
    - proizvoljno se odabere početno stablo
    - na sljedeće stablo se prelazi samo ako to stablo ima manju težinu (engl. *parsimony score*) od trenutnoga

Popis literature

- J. Pevsner, 2009. Bioinformatics and Functional Genomics, 2nd edition, Ch. 7 (Molecular Phylogeny and Evolution)
- J. Xiong, 2006. Essential Bioinformatics, Ch. 10 & 11
- Baum, D. (2008) *Reading a phylogenetic tree: The meaning of monophyletic groups.* Nature Education 1(1):190
- http://www.southampton.ac.uk/~re1u06/teaching/upgma/
- http://en.wikipedia.org/wiki/Models_of_DNA_evolution
- http://www.megasoftware.net/
- Hug *et al. (*2016) Nature Microbiology vol. 1