Metódy tvorby evolučných stromov

metóda tvorby stromov	typ dát: vzdialenosti	typ dát: DNA sekvencie (alebo iné znaky)
zhlukovací algoritmus	→ UPGMA → neighbor-joining tree	
optimalizačné kritérium	→ minimum evolution tree	parsimónia maximum likelihood Bayesova analýza

Shluková analýza

Shluk (klastr, *cluster*) je skupina objektů, které uvnitř nějaké větší skupiny nemají ani nahodilý ani rovnoměrný výskyt a jejich vzájemná vzdálenost resp. nepodobnost je menší než vzdálenost resp. nepodobnost s objekty, které patří do jiných shluků.

Těžiště (*centroid*) shluku je hypotetický (nikoliv nutně existující) prvek, jehož souřadnice ve znakovém prostoru jsou dány průměrnými hodnotami souřadnic jednotlivých objektů.

Shluková analýza

způsob tvorby shluků: aglomerativní metody – divizivní metody

uspořádání shluků: hierarchické metody – nehierarchické metody

překryv shluků: nepřekrývající nebo překrývající se shluky (*fuzzy clustering*)

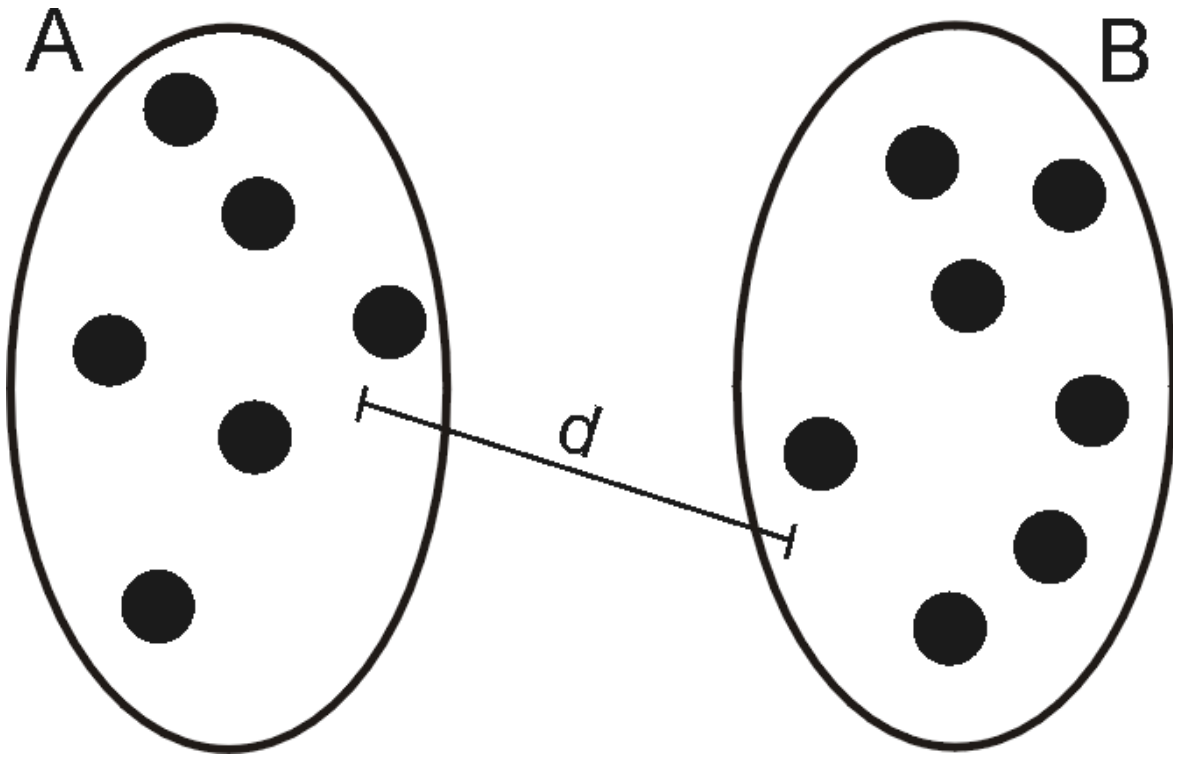
postup shlukování: sekvenční metody – simultánní metody

Shlukovací metody kategorie SAHN:

(a) metody založené na minimalizaci vzdálenosti mezi shluky

(b) metody založené na optimalizaci homogenity shluků podle určitého kritéria

Metoda nejbližšího souseda (jednospojná metoda, metoda jediné vazby, *single linkage*, *the nearest neighbor method*)



$D_1 =$

	1	2	3	4	5
1	0,0	1,0	7,0	4,0	12,0
2	1,0	0,0	2,0	5,0	9,0
3	7,0	2,0	0,0	8,0	3,0
4	4,0	5,0	8,0	0,0	6,0
5	12,0	9,0	3,0	6,0	0,0

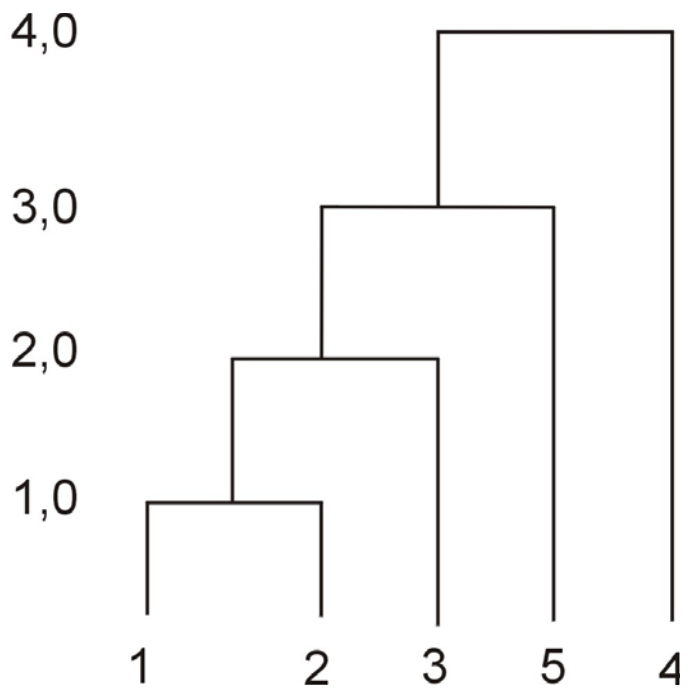
$d_{(1,2)3} = \min \{d_{1,3}, d_{2,3}\} = d_{2,3} = 2,0$

$d_{(1,2)4} = \min \{d_{1,4}, d_{2,4}\} = d_{1,4} = 4,0$

$d_{(1,2)5} = \min \{d_{1,5}, d_{2,5}\} = d_{2,5} = 9,0$

$D_2 =$

	(1, 2)	3	4	5
(1, 2)	0,0	2,0	4,0	9,0
3	2,0	0,0	8,0	3,0
4	4,0	8,0	0,0	6,0
5	9,0	3,0	6,0	0,0

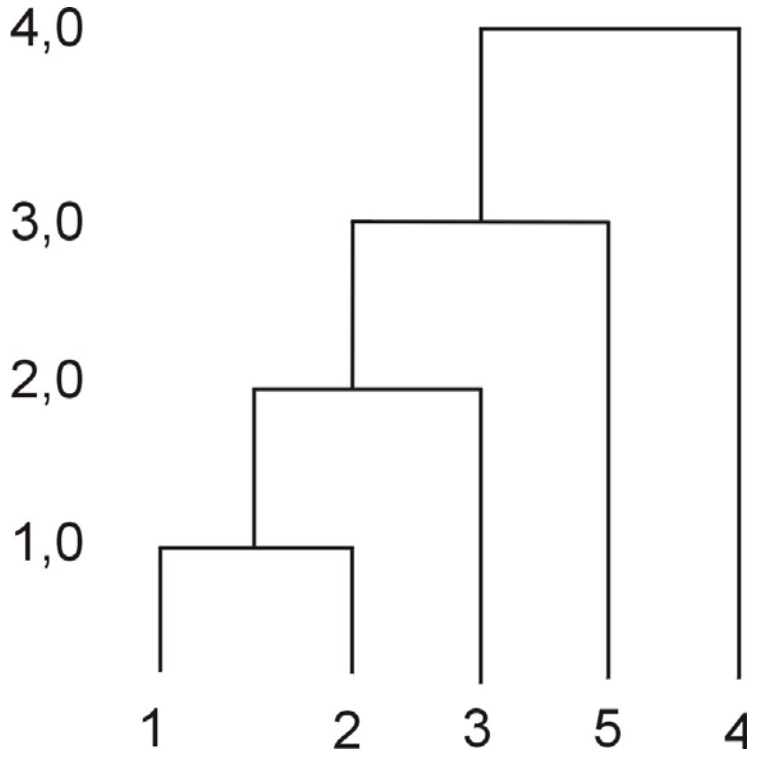


$d_{(1, 2, 3)4} = \min \{d_{(1, 2)\,4}, d_{3, 4}\} = d_{(1, 2)\,4} = 4{,}0$

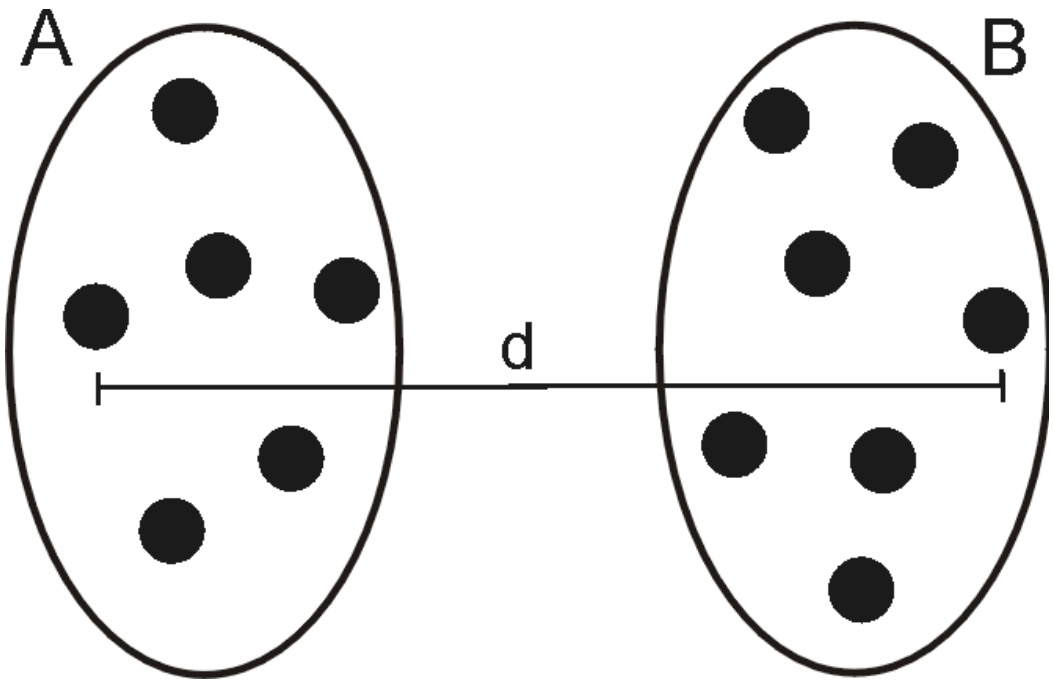
$d_{(1, 2, 3)5} = \min \{d_{(1, 2)\,5}, d_{3, 5}\} = d_{3, 5} = 3{,}0$

$D_3 =$

	(1, 2, 3)	4	5
(1, 2, 3)	0,0	4,0	3,0
4	4,0	0,0	6,0
5	3,0	6,0	0,0



Metoda nejvzdálenějšího souseda (všespojná metoda, metoda úplné vazby, *complete linkage*, *the furthest neighbor method*)



$D_1 =$

	1	2	3	4	5
1	0,0	1,0	7,0	4,0	12,0
2	1,0	0,0	2,0	5,0	9,0
3	7,0	2,0	0,0	8,0	3,0
4	4,0	5,0	8,0	0,0	6,0
5	12,0	9,0	3,0	6,0	0,0

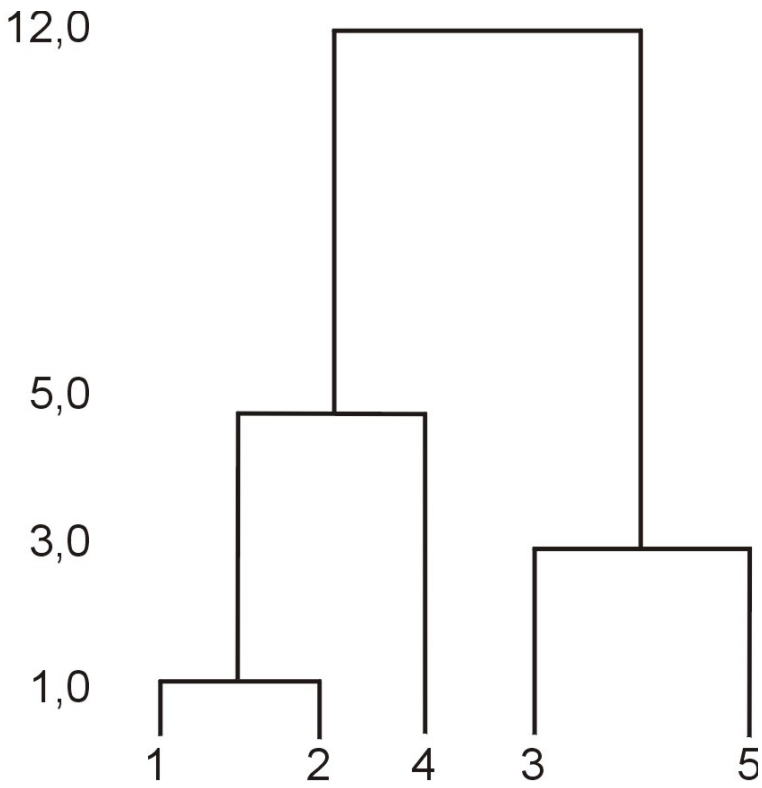
$d_{(1,2)3} = \max \{d_{1,3}, d_{2,3}\} = d_{1,3} = 7,0$

$d_{(1,2)4} = \max \{d_{1,4}, d_{2,4}\} = d_{2,4} = 5,0$

$d_{(1,2)5} = \max \{d_{1,5}, d_{2,5}\} = d_{1,5} = 12,0$

$D_2 =$

	(1, 2)	3	4	5
(1, 2)	0,0	7,0	5,0	12,0
3	7,0	0,0	8,0	3,0
4	5,0	8,0	0,0	6,0
5	12,0	3,0	6,0	0,0

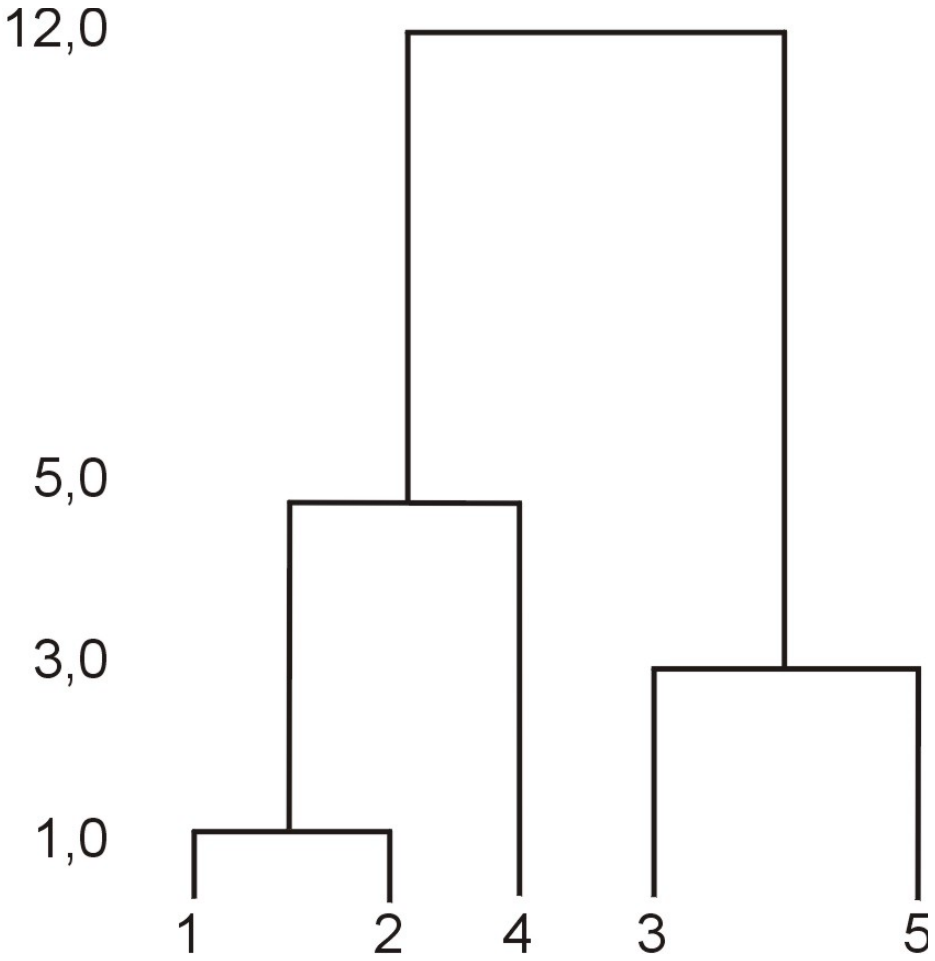


$d_{(1,\ 2)(3,\ 5)} = \max\{d_{(1,\ 2)\ 3}, d_{(1,\ 2)\ 5}\} = d_{(1,2),\ 5} = 12{,}0$

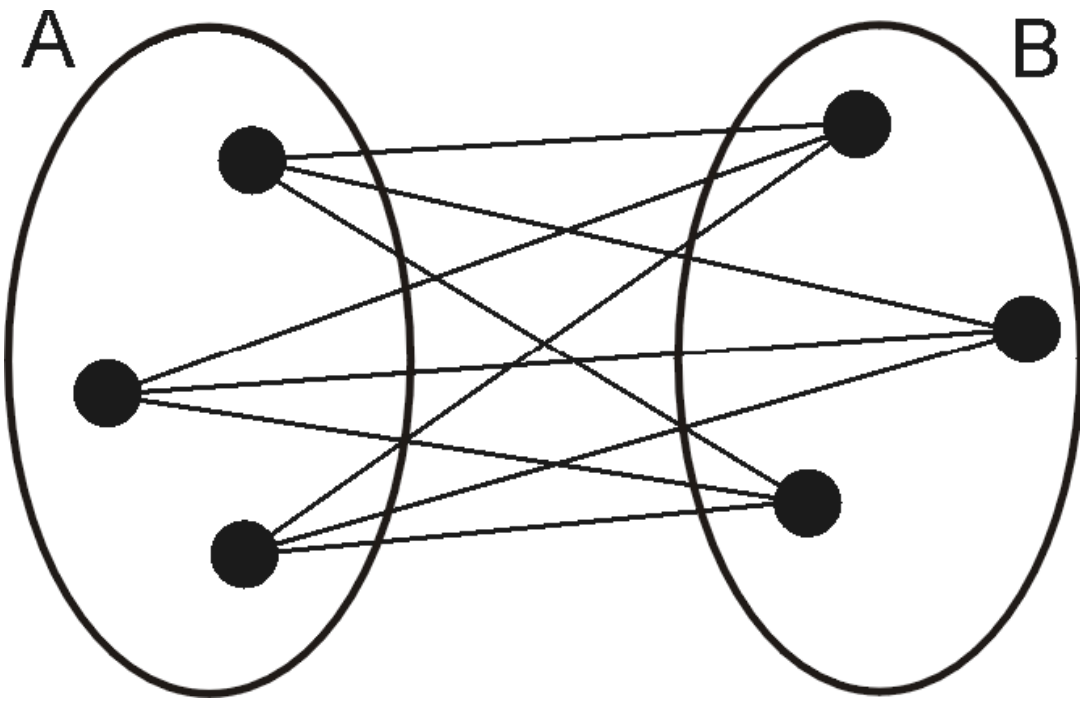
$d_{(3,\ 5)4} = \max\{d_{3,\ 4}, d_{3,\ 5}\} = d_{3,\ 4} = 8{,}0$

$D_3 =$

	(1, 2)	(3, 5)	4
(1, 2)	0,0	12,0	5,0
(3, 5)	12,0	0,0	8,0
4	5,0	8,0	0,0



Metoda průměrné vzdálenosti (středospojná metoda, metoda průměrné vazby, *average linkage*, *UPGMA – unweighted pair-group method using arithmetic averages*)



		1	2	3	4	5
	1	0,0	1,0	7,0	4,0	12,0
	2	1,0	0,0	2,0	5,0	9,0
$D_1 =$	3	7,0	2,0	0,0	8,0	3,0
	4	4,0	5,0	8,0	0,0	6,0
	5	12,0	9,0	3,0	6,0	0,0

$d_{(1, 2)3} = 1/2\ (d_{1, 3} + d_{2, 3}) = 4,5$

$d_{(1, 2)4} = 1/2\ (d_{1, 4} + d_{2, 4}) = 4,5$

$d_{(1, 2)5} = 1/2\ (d_{1, 5} + d_{2, 5}) = 10,5$

		(1, 2)	3	4	5
	(1, 2)	0,0	4,5	4,5	10,5
$D_2 =$	3	4,5	0,0	8,0	3,0
	4	4,5	8,0	0,0	6,0
	5	10,5	3,0	6,0	0,0

Metóda spájania susedných objektov (neighbor-joining method)

Metóda je založená na genetickej vzdialenosti, ktorá napr. pri hodnotení AFLP dát závisí na počte zhodujúcich sa prúžkov v príslušných porovnávaných vzorkách. Pri použití dát zo sekvencií DNA sa genetická vzdialenosť počíta iným spôsobom.

Je do určitej miery príbuzná zhlukovacím metódam. Pri výpočte vzdialenosti vytvorených zhlukov od zostávajúcich objektov sa postupuje podobne ako pri metóde priemernej vzdialenosti.

Analógia však nie je úplná, pretože ako „susedné objekty“ sa nespájajú tie, které ležia najbližšie, ale tie, výsledkom spojenia ktorých (resp. výberu ktorých) je čo najkratší dendrogram (strom). Tieto dendrogramy se skladajú z uzlov (*node*) spojených medziuzlami (*internode*) a z vetiev (*branch*).

Genetická vzdialenosť pre AFLP dáta

Koeficient podľa Nei & Li (1979): $NL_{xy} = 1 - (2\,N_{xy} / N_x + N_y)$

kde

N_{xy} = počet prúžkov (fragmentov) spoločných vzorkám *x* a *y*

N_x = celkový počet prúžkov (fragmentov) prítomných vo vzorke *x*

N_y = celkový počet prúžkov (fragmentov) prítomných vo vzorke *y*

Príklad:

vzorka *x*: 1010100011

vzorka *y*: 1010111101

$N_x = 5$; $N_y = 7$; $N_{xy} = 4$

$NL_{xy} = 1 - (2 * 4 / 5 + 7) = 0{,}333$

Koeficient podľa Link et al. (1995):

$L_{xy} = (N_x' + N_y') / (N_x' + N_y' + N_{xy})$

kde

N_{xy} = počet prúžkov (fragmentov) spoločných vzorkám *x* a *y*

N_x' = počet prúžkov (fragmentov) prítomných vo vzorke *x*, ale neprítomných ve vzorku *y*

N_y' = počet prúžkov (fragmentov) prítomných vo vzorke *y*, ale neprítomných vo vzorke *x*

Príklad:

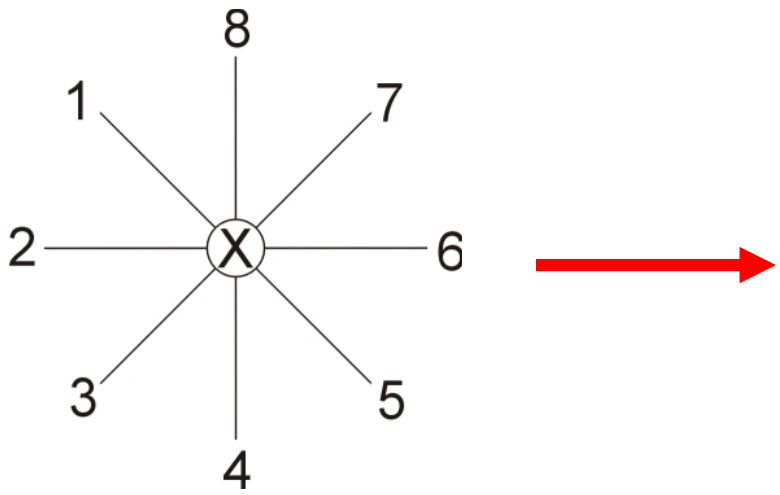
vzorka *x*: 1010100011

vzorka *y*: 1010111101

$N_x' = 1$; $N_y' = 3$; $N_{xy} = 4$

$L_{xy} = 1+3 / 1+3+4 = 0,5$

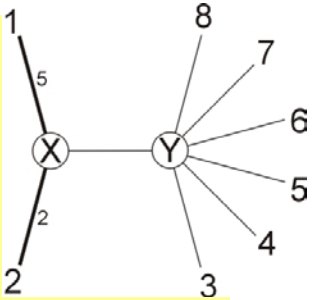
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8



Celková dĺžka dendrogramu S sa počíta podľa nasledovného vzorca (vzorec je uvedený pre dvojicu objektov 1 a 2, v ostatných prípadoch sa postupuje analogicky, pričom sa menia hodnoty „$i = 3$“, „$k = 3$“ a „$3 \le i < j$“, ktoré sú špecificky stanovené tak, aby sa vylúčili objekty 1 a 2):

$$S_{12} = L_{XY} + (L_{1X} + L_{2X}) + \sum_{i=3}^{N} L_{iY} = \frac{1}{2(N-2)} \sum_{k=3}^{N} (D_{1k} + D_{2k}) + \frac{1}{2} D_{12} + \frac{1}{N-2} \sum_{3 \le i < j} D_{ij}$$

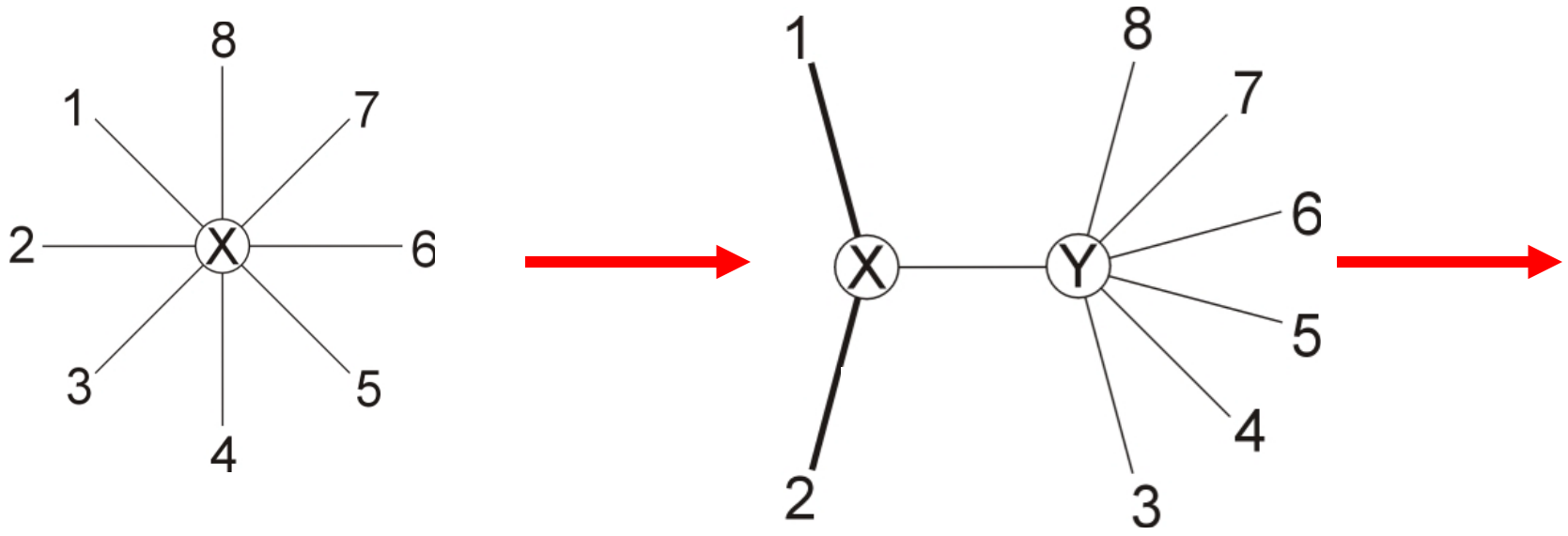
OTU	**1**	**2**	**3**	**4**	**5**	**6**
2	**7**					
3	**8**	**5**				
4	**11**	**8**	**5**			
5	**13**	**10**	**7**	**8**		
6	**16**	**13**	**10**	**11**	**5**	
7	**13**	**10**	**7**	**8**	**6**	**9**
8	**17**	**14**	**11**	**12**	**10**	**13**	**8**



$$S_{12} = \frac{1}{2(8-2)}(8+5+11+8+13+10+16+13+13+10+17+14) + \frac{7}{2} +$$

$$\frac{1}{8-2}(5+7+10+7+11+8+11+8+12+5+6+10+9+13+8) = 36{,}67$$

OTU	1	2	3	4	5	6	7
2	**36,67**						
3	38,33	38,33					
4	39,00	39,00	38,67				
5	40,33	40,33	40,00	39,67			
6	40,33	40,33	40,00	39,67	37,00		
7	40,17	40,17	39,83	39,50	38,83	38,83	
8	40,17	40,17	39,83	39,50	38,83	38,83	37,67



Vzdialenosti L_{1X} a L_{2X} sa počítajú podľa vzorcov:

$$L_{1X} = \frac{D_{12} + D_{1Z} - D_{2Z}}{2} \quad L_{2X} = \frac{D_{12} + D_{2Z} - D_{1Z}}{2} \text{ kde } D_{1Z} = \frac{\sum_{i=3}^{N} D_{1i}}{N-2} \text{ a } D_{2Z} = \frac{\sum_{i=3}^{N} D_{2i}}{N-2}$$

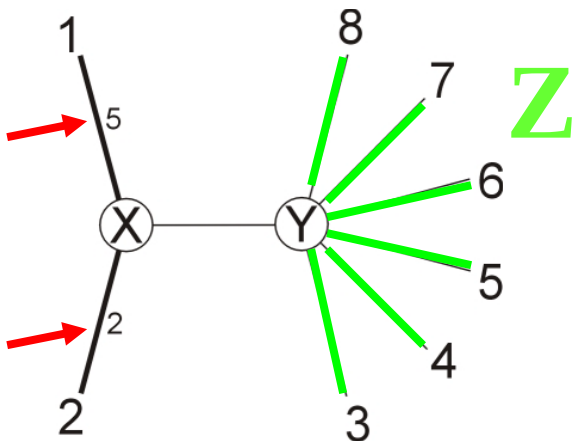
OTU	1	2	3	4	5	6	7
2	7						
3	8	5					
4	11	8	5				
5	13	10	7	8			
6	16	13	10	11	5		
7	13	10	7	8	6	9	
8	17	14	11	12	10	13	8

Pre prípad objektov 1 a 2:

$$D_{1Z} = \frac{8+11+13+16+13+17}{8-2} = 13$$

$$D_{2Z} = \frac{5+8+10+13+10+14}{8-2} = 10$$

$$L_{1X} = \frac{7+13-10}{2} = 5 \quad L_{2X} = \frac{7+10-13}{2} = 2$$

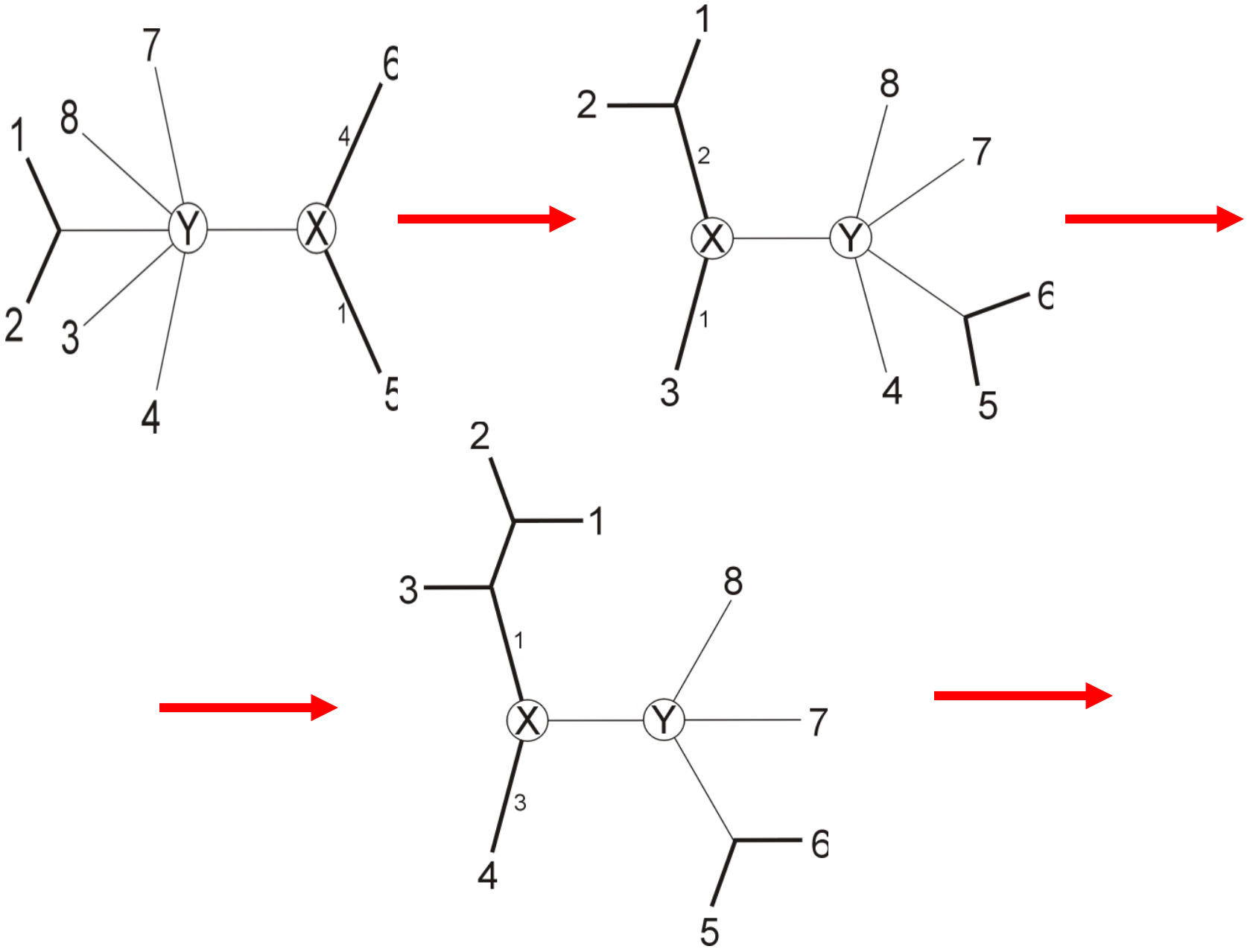


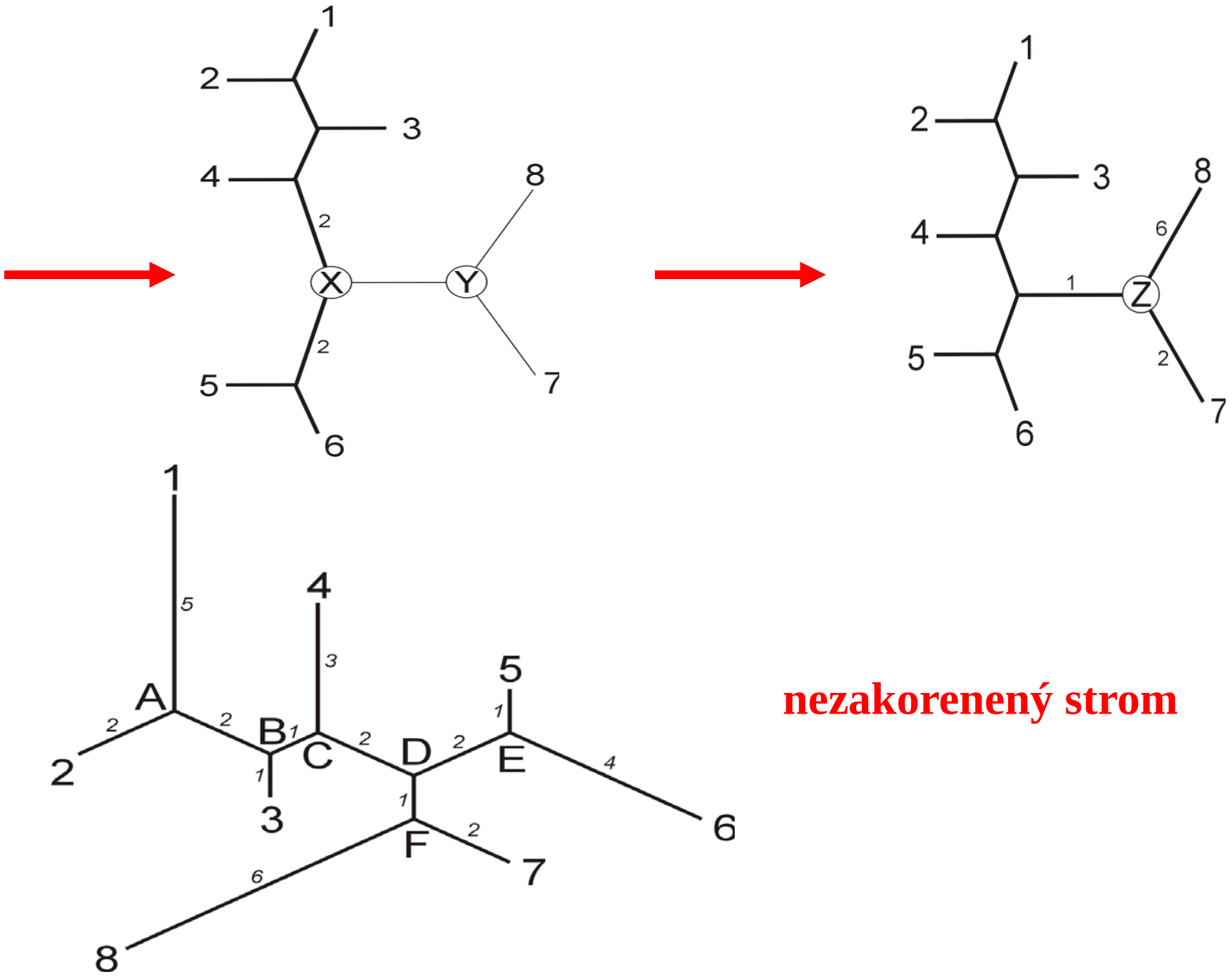
V ďalšom cykle sa postupuje podobným spôsobom s tým, že objekty 1 a 2 sa posudzujú ako jeden objekt (resp. zhluk). Vzdialenosť zhluku 1 a 2 od ostatných objektov sa počíta podobne ako pri metóde priemernej vzdialenosti zhlukovej analýzy:

$$D_{(1-2)\,j} = \frac{D_{1\,j} + D_{2\,j}}{2}$$, kde $3 \leq j \leq \mathrm{N}$.

Matica hodnôt S_{ij} v druhom cykle (ako susedné objekty boli vybrané objekty 5 a 6):

OTU	**1-2**	**3**	**4**	**5**	**6**	**7**
3	31,50					
4	32,30	32,30				
5	33,90	33,90	33,70			
6	33,90	33,90	33,70	**31,30**		
7	33,70	33,70	33,50	33,10	33,10	
8	33,70	33,70	33,50	33,10	33,10	31,90


7
6
8
4
1
Y
X
2
3
1
4
5
1
2
8
2
7
X
Y
1
6
3
4
5
2
1
3
8
1
X
Y
7
3
6
4
5



nezakorenený strom

Bootstrap

- Zo znakov vytvoríme variácie n-tej triedy z n prvkov s opakovaním (kde n je počet znakov) – „resampling with replacement“ (t.j. vytvoríme nové replikované dátové súbory)
- Analyzuje sa každý replikovaný dátový súbor
- Hľadáme, ktoré zhluky sa v jednotlivých analýzach opakujú

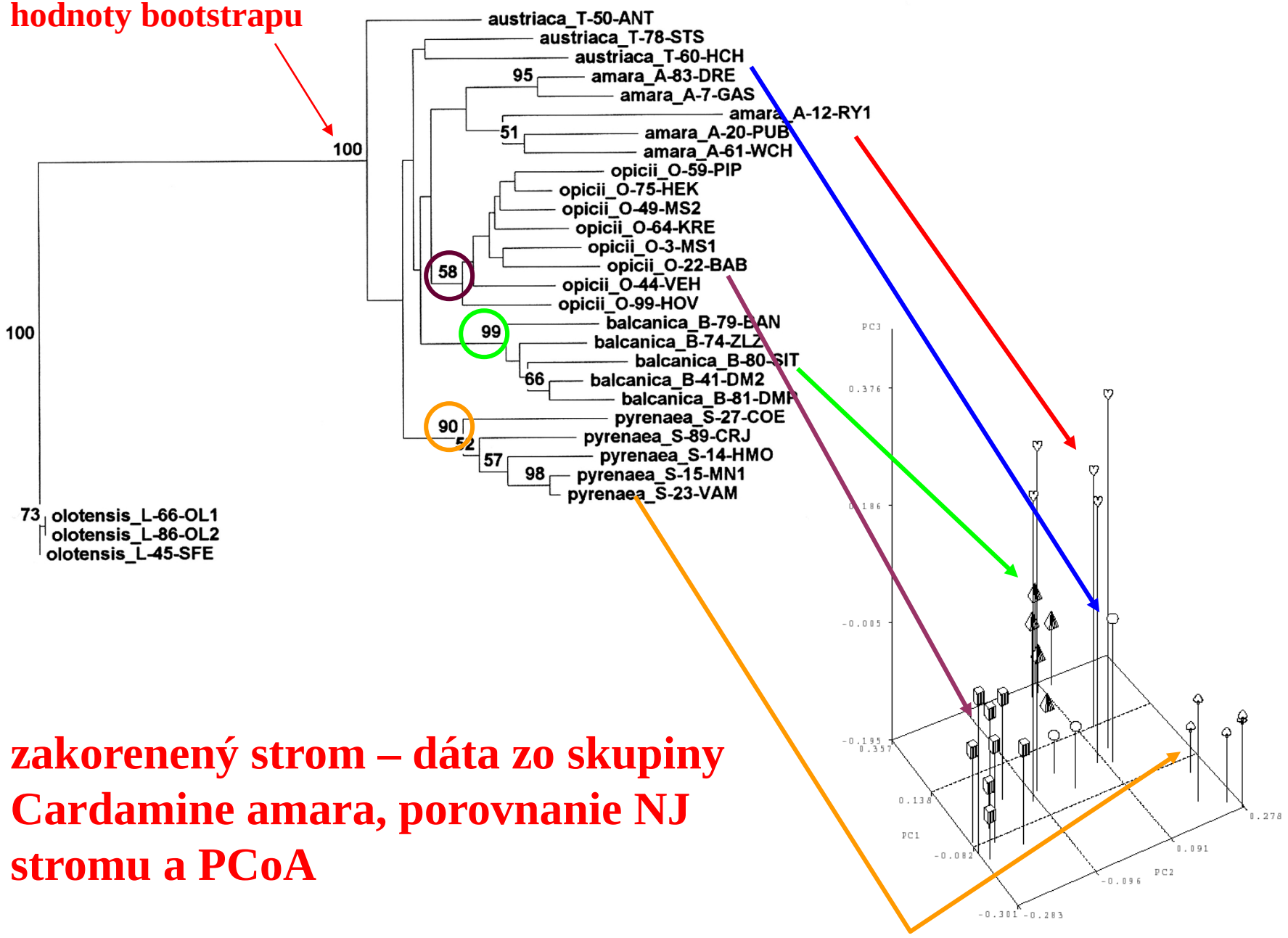
pôvodná matica dát

	znaky							
taxóny	1	2	3	4	5	6	7	8
A	R	R	Y	Y	Y	Y	Y	Y
B	R	R	Y	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	R	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

replikovaný dátový súbor

	znaky							
taxóny	1	2	2	5	5	6	6	8
A	R	R	R	Y	Y	Y	Y	Y
B	R	R	R	Y	Y	Y	Y	Y
C	Y	Y	Y	Y	Y	R	R	R
D	Y	Y	Y	R	R	R	R	R
Outgp	R	R	R	R	R	R	R	R

hodnoty bootstrapu



zakorenený strom – dáta zo skupiny Cardamine amara, porovnanie NJ stromu a PCoA

Metódou **spájania susedných objektov** **(neighbor-joining method) je možné vyhodnocovať aj DNA sekvencie**

Genetickú vzdialenosť v takomto prípade počítame dvomi spôsobmi:

(1) jednoduchá vzdialenosť
= počet rozdielnych pozícií / celkový počet pozícií nukleotidov (najmä v prípade vysokej substitučnej rýchlosti môže značne podhodnocovať počet substitúcií)

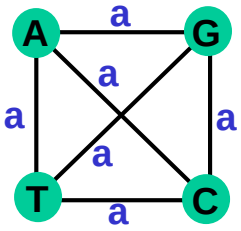
(2) vzdialenosť vypočítaná na základe substitučných modelov

Substitučné modely

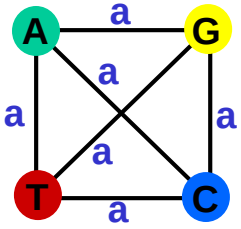
- substitučné modely
  - Jukes Cantor, Kimura, Tajima & Nei, Transversion analyses
  - General time reversable model (GTR): zahŕňa odlišnú pravdepodobnosť pre každý typ zmeny
  - LogDet / Paralinear distance model: umožňuje zohľadniť odlišné frekvencie báz v rôznych sekvenciách
- všetky tieto modely zahŕňajú korekciu pre opakované substitúcie na tej istej pozícii
- všetky tieto modely (okrem Logdet/paralinear distances) sa môžu modifikovať tak, aby zahŕňali gama korekciu pre heterogenitu rýchlostí zmien na rôznych pozíciách

Substitučné modely

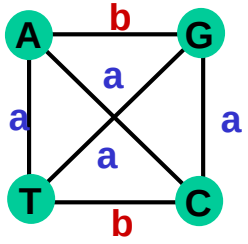
Zvyšujúci sa počet parametrov modelu



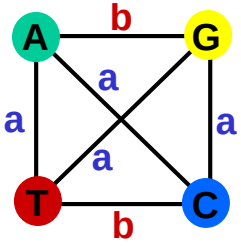
JC – rovnaké rýchlosti substitúcie; rovnaké frekvencie báz



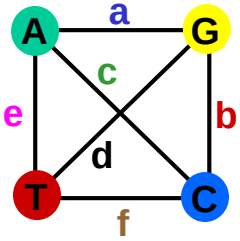
F81 – rovnaké rýchlosti substitúcie; nerovnaké frekvencie báz



K2P – dve rôzne rýchlosti substitúcie; rovnaké frekvencie báz



HKY – dve rôzne rýchlosti substitúcie; nerovnaké frekvencie báz



GTR – 6 rôznych rýchlostí substitúcie; nerovnaké frekvencie báz

Jukes & Cantor model: $d_{xy} = - (3/4) \mathit{ln} (1- 4/3 D)$

- d_{xy} = vzdialenosť medzi sekvenciou x a sekvenciou y vyjadrená ako počet zmien pripadajúcich na jednu pozíciu
- (ak by sme d_{xy} počítali podľa vzorca $d_{xy} = r/n$ kde r je počet zmien a n je celkový počet pozícií v sekvencii, potom sa predpokladá, že sa môžu meniť nukleotidy na všetkých pozíciách; ak však existujú v dvoch sekvenciách invariantné pozície, potom tento model podhodnotí množstvo zmien na variabilných pozíciách)
- D = je pozorovaná proporcia nukleotidov, ktoré sa líšia medzi dvomi sekvenciami (fractional dissimilarity)
- *ln* = prirodzený logaritmus korigujúci (zohľadňujúci) opakované substitúcie na tej istej pozícii
- 3/4 a 4/3 vyjadrujú skutočnosť, že sú celkove štyri nukleotidy a tri možnosti, ktorými sa jeden nukleotidu môže líšiť od druhého – všetky typy zmien sú rovnako pravepodobné (t.j. nepríbuzné sekvencie sú z 25% identické čistou náhodou)

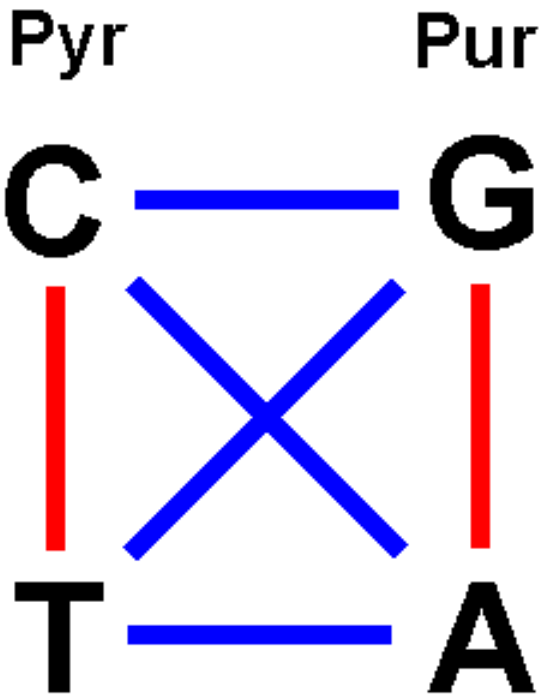
A. Wilmotte, Uni Liège

Prirodzený logaritmus *ln* sa používa na korekciu opakovaných zmien nukleotidov na tej istej pozícii

- Ak sú dve sekvencie na 95% identické, potom sú rozdielne v 5% pozícií, potom D = 0,05 a $d_{xy} = -3/4 \ln [1 – (4/3 \times 0,05)] = 0,0517$
- Pozorovaná proporcia líšiacich sa nukleotidov 0.05 sa zvýšila len málo - na 0,0517 – toto je logické, lebo v prípade dvoch podobných sekvencií predpokladáme len malé množstvo opakovaných substitúcií na tej istej pozícii, pretože uplynula len krátka doba, od momentu keď sa tieto sekvencie vývojovo oddelili
- Ale, ak sú dve sekvencie podobné len na 50% a teda odlišné 50% nukleotidových pozícií, potom D = 0,5 a potom

  $d_{xy} = -3/4 \ln [1 – (4/3 \times 0,5)] = 0,824$
- Pri sekvenciách, ktoré sa vo vývoji oddelili už dávno predpokladáme väčší počet zmien na tej istej pozícii

A. Wilmotte, Uni Liège



Tranzície (**Ti**) sú zmeny medzi **pyrimidínmi** (**C T**), alebo medzi **purínmi** (**A G**)

Tranzverzie (**Tv**) sú zmeny medzi **pyrimidínmi** a **purínmi**

A. Wilmotte, Uni Liège

Kimurov dvojparametrový model

V Kimurovom dvojparametrovom modeli (Kimura 1980) sa pri výpočte genetickej vzdialenosti tranzície a tranzverzie hodnotia oddelene:

$$d_{AB} = -\frac{1}{2}\ln\left[(1-2P-Q)\sqrt{1-2Q}\right]$$

kde P je časť nukleotidových pozícií, ktorá sa líši tranzíciou a Q je časť nukleotidových pozícií, ktorá sa líši tranzverziou.

Keďže je dvakrát viac možností pre transverzie ako pre tranzície hodnota **K transition bias = [Ti] / [Tv]**

by sa mala blížiť **0.5**

Tv sú ale pri blízko príbuzných sekvenciách zriedkavé, častejšie sa objavujú pri porovnaní nepríbuzných sekvencií.

Pre blízko príbuzné sekvencie sa uvádza **K > 6**

Tajima & Nei

Pri general correction podľa Tajima a Nei (1984), sa evolučná/genetická vzdialenosť odhaduje podľa vzorca:

$$d_{AB} = -b \ln\left(1 - \frac{1}{b} f_{AB}\right) \qquad \text{kde} \qquad b = 1 - \sum_{i \in N} f_i^2$$

fi je frekvencia i-teho typu nukleotidu, ktorý patrí do súboru možných typov nukleotidov N (= A, G, C, U alebo T) v porovnávaných sekvenciách. Tento vzorec platí pre model nukleotidových substitúcií s rovnakou rýchlosťou zmien medzi nukleotidmi a neberie do úvahy možné rôzne rýchlosti substitúcií medzi rôznymi pármi nukleotidov. V programe TREECON, vypočítané zloženie báz je priemerom zo všetkých analyzovaných sekvencií (ako navrhli Swofford et al. 1996). Ak sú frekvencie všetkých štyroch nukleotidov 0.25 potom tento vzorec zodpovedá Jukes-Cantorovmu modelu.

A. Wilmotte, Uni Liège

Transversion analysis

Niekedy môže byť užitočné odhadovať genetickú (evolučnú) vzdialenosť iba na základe transverzií (Woese et al. 1991, Van de Peer et al. 1996). Táto vzdialenosť sa potom počíta podľa nasledovného vzorca (Tajima & Nei 1984, Swofford et al. 1996):

$$d_{AB} = -b \ln\left(1 - \frac{1}{b} Q\right)$$

kde Q je podiel transverzií

$$b = 1 - \left[\left(f_A + f_G\right)^2 + \left(f_C + f_{(T,U)}\right)^2 \right]$$

a fA + fG sú podiely purínov a fC + fU alebo fT sú podiely pyrimidínov počítané v celom alignmente

A. Wilmotte, Uni Liège

logDet/paralinear distance

LogDet/paralinear distance bola navrhnutá pre situácie, kde sú rôzne frekvencie báz pri každom páre porovnávaných sekvencií – dovoľuje teda aby sa zloženie báz menilo v rôznych častiach stromu

Tým sa táto vzdialenosť líši od GTR distance model, ktorý berie do úvahy len priemerné zloženie báz a aplikuje ho na všetky porovnávané sekvencie

LogDet/paralinear distance predpokladá, že sa môžu meniť nukleotidy na všetkých pozíciách – preto je dobré vypustiť z analýzy pozície, na ktorých sa nukleotidy nemenia (označované invariable v programe PAUP)

A. Wilmotte, Uni Liège

logDet/Paralinear Distances

$$d_{xy} = -\ln (\det F_{xy})$$

- d_{xy} = odhadovaná vzdialenosť medzi sekvenciou x a sekvenciou y
- *l*n = prirodzený logaritmus korigujúci (zohľadňujúci) opakované substitúcie na tej istej pozícii
- F_{xy} = 4 x 4 (štyri bázy v DNA) matica rozdielov sekvencií X a Y – matica sumarizuje relatívne frekvencie báz pri párovom porovnaní
- det = je determinant (matematická hodnota) matice

A. Wilmotte, Uni Liège

LogDet – príklad pre dve sekvencie A a B

Matica sumarizuje párové porovnania zodpovedajúcich pozícií sekvencií A a B s 900 pozíciami

		Sekvencia B			
		a	c	g	t
Sekvencia A	a	224	5	24	8
	c	3	149	1	16
	g	24	5	230	4
	t	5	19	8	175

Matica $\mathbf{F_{xy}}$ vyjadruje tieto dáta ako proporciu (napr. 224/900 = 0.249) pozícií:

$\mathbf{F_{xy}}$ =

	a	c	g	t
a	.249	.006	.027	.009
c	.003	.166	.001	.018
g	.027	.006	.256	.004
t	.006	.021	.009	.194

- d_{xy} = -*l*n [det Fxy] = -*l*n [.002] = 6.216 (logDet vzdialenosť medzi sekvenciami A a B)

A. Wilmotte, Uni Liège

logDet/paralinear distance: výhody

Veľmi užitočné ak pomer báz je medzi sekvenciami veľmi výrazne odlišný

Aj keď zloženie báz nie je výrazne odlišné LogDet / paralinear distances model dáva výsledky aspoň tak isto dobré ako ostatné metódy výpočtu vzdialeností

Nedostatok je v tom, že model predpokladá že nukleotidy na rôznych pozíciách sa menia rovnakým spôsobom a že rýchlosti sú rovnaké pre všetky pozície

Vylúčenie invariabilných miest však v simuláciách dávalo dobré výsledky

A. Wilmotte, Uni Liège

Vzdialenosti - výhody

Rýchla metóda – vhodná pre analýzu dátových súborov, ktoré sú príliš veľké na to, aby sa analyzovali metódou ML

K dispozícii je veľký počet modelov s mnohými parametrami - toto zlepšuje odhad vzdialeností

A. Wilmotte, Uni Liège

Vzdialenosti - nevýhody

Stráca sa istá časť informácie – keďže pracujeme so vzdialenosťami výsledok nie je možné dávať do vzťahu k pôvodným sekvenciám

Evolučne najvýznamnejšie pozície nukleotidov sa dajú zistiť iba analýzou, ktorá je založená na znakoch (maximum likelihood, parsimónia)

Metóda maximum likelihood dáva vo všeobecnosti lepšie výsledky – je úspešnejšia pri hľadaní správneho stromu v počítačových simuláciách (ale logDet vzdialenosť dáva niekedy lepšie výsledky)

A. Wilmotte, Uni Liège

Problémom môže byť heterogenita rýchlostí zmien nukleotidov na rôznych pozíciách

Poblém nastáva keď sa nukleotidy na rôznych pozíciách molekuly menia rôznou rýchlosťou v dôsledku funkčných obmedzení

Viaceré modely (Jukes Cantor, Parsimony, LogDet, some ML models) predpokladajú, že nukleotidy na všetkých pozíciách sa možu meniť a že k zmenám dochádza rovnakou rýchlosťou

Toto podhodnocuje množstvo zmien, ku ktorým došlo a tým aj vzdialenosti medzi sekvenciami, čo má za následok nesprávne stromy

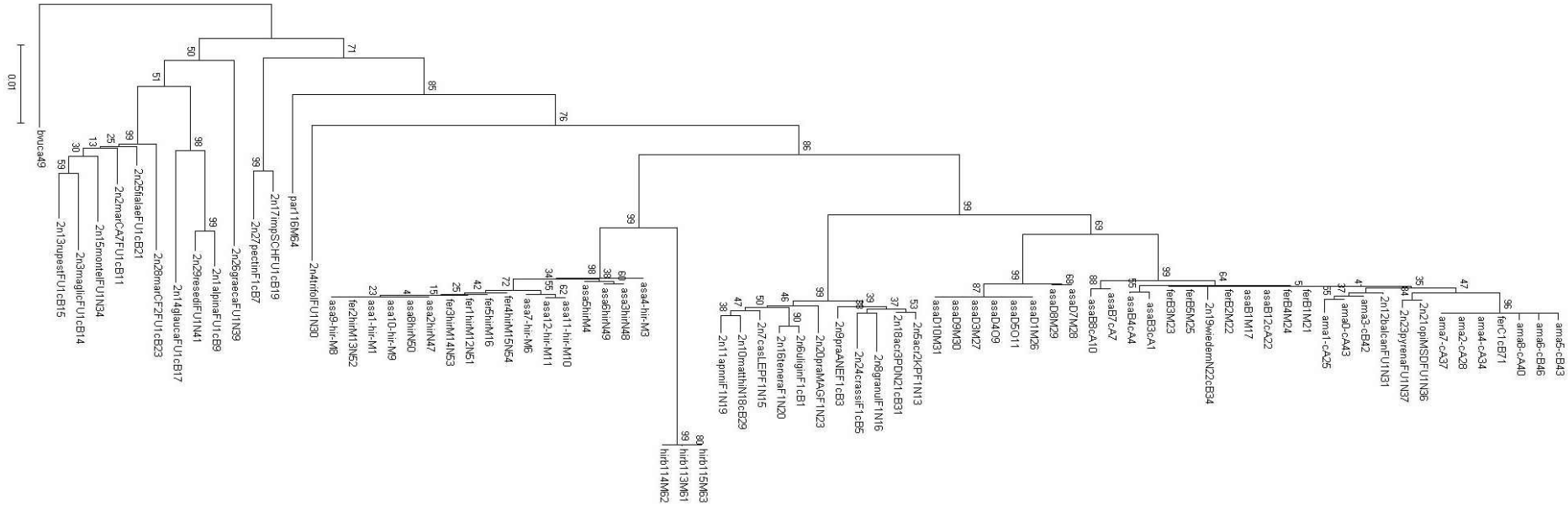
A. Wilmotte, Uni Liège

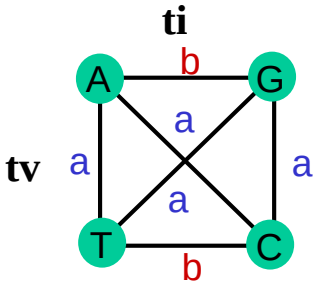
Problémom môže byť heterogenita rýchlostí zmien nukleotidov na rôznych pozíciách

Uvedené problémy možno odstrániť gama korekciou na heterogenitu rýchlostí zmien na jednotlivých pozíciách – za predpokladu, že to príslušný model dovoľuje

Druhá možnosť je editovať dáta - vylúčiť zo sekvencie tie pozície, ktoré sú v celom alignmente konštantné (t.j. tie, ktoré sa najpomalšie menia) alebo tie pozície, na ktorých sa nukleotidy menia rýchlejšie ako u ostatných.

A. Wilmotte, Uni Liège





K2P – dve rôzne rýchlosti substitúcie; rovnaké frekvencie báz

Cardamine asarifolia a príbuzné druhy
Neighbor joining, Kimura 2 parameter, 1000 bootstraps

Minimum Evolution

Pre každý možný alternatívny strom odhadneme dĺžku každého konára z odhadovanej vzdialeností medzi taxónmi a potom počítame sumu (S) všetkých konárov na strome. Kritérium minimum evolution má za cieľ nájsť strom s najnižšou hodnotou S.