MOLEKULÁRNÍ TAXONOMIE - 6 (2019)

Uvolňování parametrů v substitučních modelech (opakování z minula, trochu jinak)

Nyní si ukážeme obecný princip, jakým se obohacují substituční modely o parametry tak, aby se lépe podobaly evoluci sekvencí. Vyjdeme z jednoduchého DNA modelu Jukes-Cantora. Ten předpokládá, že všechny typy substitucí probíhají se stejnou rychlostí. Můžeme to vyjádřit maticí v jejíž sloupce a řádky představují nukleotidy a členy matice rychlost substituce pro danou dvojici nukleotidů.

$$Q = \begin{array}{c} \begin{matrix} A & C & T & G \end{matrix} \\ \begin{pmatrix} * & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & * & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & * & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & * \end{pmatrix} \begin{matrix} A \\ C \\ T \\ G \end{matrix} \end{array}$$

Když tuto matici vynásobíme časem t získáme očekávaný počet substitucí v intervalu t. Násobení matice jedním číslem je ještě snadná věc, stačí vynásobit tímto číslem každý člen. Matici pravděpodobností, s jakými v časovém intervalu t dojde k různým typům záměn, získáme, když Eulerovo číslo umocníme maticí očekávaného počtu různých typů substitucí.

$P(t)=e^{Qt}$

Umocňování maticí již není tak snadné jako násobení a neznamená to, že každým členem mocníme Eulerovo číslo. I v případě takto jednoduchého Jukes-Cantorova modelu vypadá výsledná matice poměrně složitě

$$P = \begin{pmatrix} \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} \\ \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} - \frac{1}{4}e^{-t\mu} & \frac{1}{4} + \frac{3}{4}e^{-t\mu} \end{pmatrix}$$

Z tvaru členů této matice cítíme, že jsme blízko rovnicím, se kterými jsem se setkali při odvozování výpočtu distance u Jukes-Cantorova modelu. Všimněte si také, že sloupce a řádky této matice dávají součet 1, což je správně, protože pravděpodobnost, že se nukleotid buď změní nebo zůstane sám sebou je 1 (jiná možnost neexistuje). U general time reversible modelu jsem dovolili, aby různé typy záměn měly různé relativní rychlosti (α,β,γ,δ,ε,ζ). Pokud

chceme ještě navíc vzít v potaz frekvence s jakými sekvence používají jednotlivé typy bazí, vynásobíme výrazy v matici frekvencí báze v řádku (π_i). Matice rychlostí bude vypadat takto.

$$Q = \begin{bmatrix} -(\alpha\pi_G + \beta\pi_C + \gamma\pi_T) & \alpha\pi_A & \beta\pi_A & \gamma\pi_A \\ \alpha\pi_G & -(\alpha\pi_A + \delta\pi_C + \varepsilon\pi_T) & \delta\pi_G & \varepsilon\pi_G \\ \beta\pi_C & \delta\pi_C & -(\beta\pi_A + \delta\pi_G + \eta\pi_T) & \eta\pi_C \\ \gamma\pi_T & \varepsilon\pi_T & \eta\pi_T & -(\gamma\pi_A + \varepsilon\pi_G + \eta\pi_C) \end{bmatrix}$$

Stále musíme dbát na to, aby součet řádku byl 0 a průměr členů řádku mimo diagonálu 1. Matici pravděpodobností, že přes větev dojde ke změně odvodíme opět jako $P(t)=e^{Qt}$ a tvar jejích členů bude dost komplikovaný.

Proteinové substituční modely

Obdobou Jukes-Cantorova substitučního modelu pro proteinové sekvence je Poissonův model

$$D = -19/20 \ln(1- 20/19p)$$

kde p je počet rozdílných nukleotidů mezi dvěma sekvencemi proteinů. Vzorec je stejný jako u Jukes-Cantorova modelu až na to, že hodnoty 4 a 3 označující počet různých nukleotidů a počet typů změn na jiný nukleotid, byly nahrazeny čísly 20 a 19, které odpovídají situaci u aminokyselin. Stejně jako Jukes-Cantorův model ani Poissonův model nepředpokládá různou substituční rychlost různých aminokyselinových substitucí ani vliv frekvence aminokyselin v proteinu, tedy ochotu používat určitou aminokyselinu. Proteinové modely, které toto v potaz berou, vychází z pozorovaných počtů záměn v alignmentech velkého množství konzervovaných proteinů a jedná se o podobné substituční tabulky, jaké se používají pro skórování alignmentů proteinových sekvencí (PAM, Blosum). Modelům, které používají tyto tabulky odvozené z reálných dat nazýváme empirické.

Níže uvádím tabulku PAM001, která odpovídá počtu pozorovaných rozdílů mezi sekvencí 1 (řádek) a sekvencí 2 (levý sloupec). Tyto sekvence se v průměru lišily v 1% aminokyselin. Hodnoty jsou vztaženy na 10 000 aminokyselin.

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
		ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
A	ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	his	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

Protože se sekvence lišily velmi málo, substituční saturace je zanedbatelná a počet pozorovaných rozdílů lze tedy považovat za počet substitučních událostí. 1% rozdílů odovídá délce větve (genetické vzdálenosti) D=0,01. Máme tedy před sebou matici pozorovaného množství veškerých typů záměn, ke kterým došlo na větvi dlouhé 0,01 mezi 10 000 aminokyselin dlouhými sekvencemi. Z této matice budeme vycházet. Všimněte si, že matice není symetrická, což znamená, že pravděpodobnost záměn určité dvojice aminokyselin jedním směrem není stejná jako druhým směrem. Sekvence tedy zřejmě substituují podle nereverzibilního modelu, jenže s tím si zatím nedokážeme poradit a chceme pracovat s reverzibilními modely. Proto si hodnoty nejprve zesymetrizujeme přes diagonálu (vypočteme průměry odpovídajích hodnot), pak je vydělíme 10 000 takže získáme čísla menší než 1 (procenta změn). Součty ve sloupcích by se měly rovnat 1,00 a hodnoty představují empirický odhad pravděpodobnosti záměny aminokyseliny v řádku na aminokyselinu ve sloupci. Takto upravená matice (platná pro D=0,01) se dá převést na matici odpovídající vyššímu D umocněním. Z Poissonova rozdělení přece víme, že pravděpodobnost, že k události dojde souvisí s očekávaným počtem událostí, tj. s délkou větve (D) skrze mocninu Eulerova čísla.

$P=e^D$

Pokud se 10x prodlouží větev, umocní se pravděpodobnost na 10. Matici pravděpodobností pro 10x vyšší D=0,1 tedy získáme, když námi známou matici pro D=0,01 umocníme na 10. Pozor umocnění matice neznamená umocnit hodnoty matic (viz. násobení matic http://cs.wikipedia.org/wiki/N%C3%A1soben%C3%AD_matic). V našem případě mocnění číslem >1 způsobí to, že se hodnoty na diagonále (pravděpodobnost, že aminokyselina zůstane nezměněna) budou snižovat a hodnoty mimo diagonálu (pravděpodobnost změn) se budou zvyšovat, ale součet řádků i sloupců zůstane 1. U matice, kde by hodnoty mimo diagonálu byly stejné (Poissonův model) by se po umocnění na ∞ (dvě nepříbuzné sekvence) všechny hodnoty

v matici (na diagonále i mimo) změnily na 1/20, což odpovídá předpokládané maximální saturaci u tohoto jednoduchého modelu.

Vraťme se však k reálnější PAM matici. Pokud tedy chceme zjistit genetickou vzdálenost D_{xy}, která dělí dvojici analyzovaných sekvencí X z Y, stačí si sestavit matici počtu různých typů záměn, které pozorujeme mezi sekvencemi X a Y, tu symetrizovat a normalizovat, jak bylo uvedeno výše. Pak budeme hledat hodnotu mocnitele, který převede empiricky odvozenou PAM matici pro D=0,01 na matici, která se bude nejvíce blížit matici získané z našich dat. Tímto mocnitelem vynásobíme 0,01 a získáme D_{xy} pro naše sekvence. V praxi se k tomuto používá metoda maximum likelihood, kterou si představíme o dvě přednášky později.

Je třeba mít na paměti, že hodnoty v empiricky odvozené matici vychází hodnoty pravděpodobností z pozorovaných rozdílů a ovlivňuje je tedy frekvence jednotlivých typů aminokyselin (tedy ochotu je používat) v sekvencích, ze kterých byly odvozeny. Námi analyzované sekvence mohou být v tomto ohledu jiné. Pokud bychom si chtěli empirickou matici lépe přizpůsobit našim sekvencím, je potřeba nejprve odfiltrovat frekvence aminokyselin sekvencí, ze kterých byla empirická matice vytvořena – vydělit hodnoty v řádcích, frekvencí příslušné aminokyseliny, s jakou se vyskytovala v datech. Takto oproštěnou matici můžeme uzpůsobit našim sekvencím tak, že její hodnoty naopak vynásobíme frekvencemi aminokyselin pozorovanými v našich sekvencích. Pokaždé je matici třeba normalizovat a symetrizovat, aby suma sloupců a řádků byla rovna jedné.

Odfiltrování vlivu frekvence aminokyselin a další operace již provedli bioinformatici za nás a připravili nám nejrůznější typy matic, které mohou mít různá použití, třeba v závislosti na tom, z jaké sady proteinů byly odvozeny. Analyzujeme-li kupříkladu proteiny kódovány mitochondriálním genomem, bude vhodnější použít matice odvozené ze sady mitochondriálních proteinů (mtREV). Z obecných matic je dnes široce používána a zdá se, že nejlépe funguje, matice LG, která má také několik variant. Dříve to byly matice WAG, JTT a další. Jen pro úplnost třeba dodat, že tyto matice vznikly logaritmováním pravděpodobnostních matic a jedná tedy o matice relativních substitučních rychlostí (Q).

Q=lnP

Při výpočtu pravděpodobnostních matic tedy musíme tyto připravené matice nejen vynásobit distancí, ale upět jimi umocnit eulerovo číslo.

Jak najít nejlepší strom

Obecné pravidlo říká, že nejlepší strom je takový, který nejlépe vysvětlí data, která pozorujeme, ať je to alignment sekvencí, fingerprintingový vzor nebo matice distancí. Co v tomto konkrétním případě znamená formulka “nejlépe vysvětlí” ponechám zatím stranou, protože to vyplyne z následujícího textu a “měřením”, jak dobře stromy vysvětlují data, se budeme zabývat v i následujících přednáškách.

Existují v zásadě dva přístupy, jak nalézt takový “nejlepší strom”.

1. **Algoritmus** – najde jen jeden strom postupným “skládáním” OTU - klastrovací analýza UPGMA, neighbour-joining.
2. **Prohledávání stromového prostoru** – heuristické hledání, Marcov chain Monte Carlo – a skórování stromů podle různých kritérií, která jsou měřítkem toho, jak dobře strom “vysvětluje” pozorovaná data.

V následujícím textu si představíme dvě metody algoritmické (UPGMA a neighbor-joining) a dvě metody, které skórují stromy jim předložené podle svých kritérií (nejmenší čtverce, minimální evoluce). Všechny metody, které si nyní ukážeme vychází z matice genetických distancí, říkáme jim distanční metody a jejich úspěch je založen do značné míry na tom, jak přesně jsou distance spočítány.

Jedna taková matice genetických vzdáleností je uvedena níže. Pod A, B, C a D se skrývají taxony neboli obecně řečeno operačně taxonomické jednotky (OTU). Hodnoty v buňkách představují genetické vzdálenosti.

	A	B	C	D
A	-			
B	0.5	-		
C	0.45	0.15	-	
D	0.55	0.4	0.35	-

UPGMA (Unweighted Pair Group Method with Arithmetic mean)

Nazývá se také shlukovací analýza. Tato metoda postupuje následovně.

1. V matici distancí najde dvojici s nejmenší distancí a shlukne ji dohromady. V našem případě je to BC.

2. Vypočítá vzdálenost této společné OTU od ostatních:

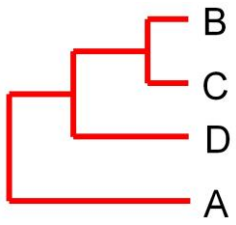
$$D_{(BC)A} = (D_{AB} + D_{AC})/2 = (0,5 + 0,45)/2 = 0,475$$
$$D_{(BC)D} = (D_{BD} + D_{CD})/2 = (0,4 + 0,35)/2 = 0,375$$

Počítá to jako aritmetický průměr ze vzdáleností všech dvojic jednoduchých OTU (druhů), kde jeden člen dvojice pochází z jedné porovnávané OTU (v našem případě BC) a druhý z druhé OTU (v našem případě pouze A)

3. Z vypočtených genetických vzdáleností vytvoří novou (menší) matici. D_{AD} se přepíše z původní matice beze změny.

	A	BC	D
A	-		
BC	0.475	-	
D	0.55	0.375	-

4. Postup se opakuje. Tentokrát je nemenší distance mezi D a BC, takže D připojíme k BC. Protože nám zbývá už jediné OTU - A - připojíme A jako poslední.

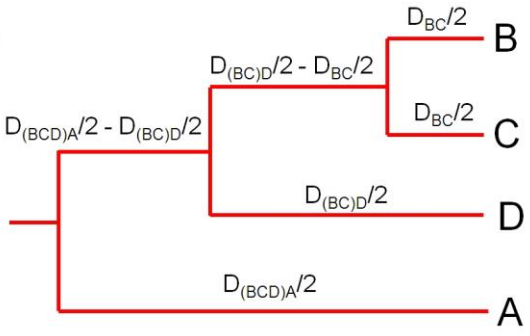


Vypočítáme vzdálenost BCD od A.

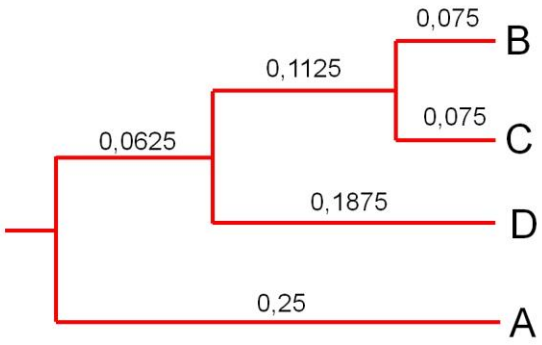
$D_{(BCD)A} = (D_{AB} + D_{AC} + D_{AD})/3 = (0,5 + 0,45 + 0,55)/3 = 0,5$

5. Ze známých distancí vypočteme délky větví následovně

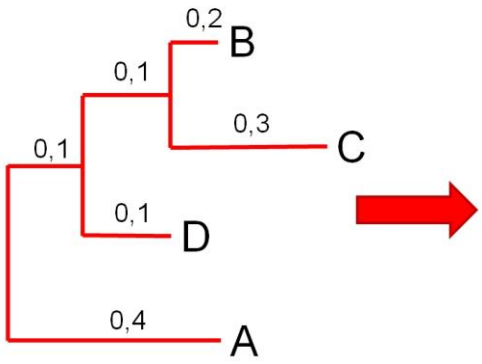
$D_{BC} = 0,15$
$D_{(BC)D} = 0,375$
$D_{(BCD)A} = 0,5$



a dospějeme k tomuto výsledku.



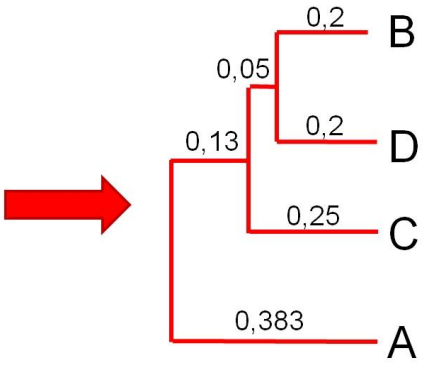
UPGMA je to nejjednodužší metoda konstrukce fylogenetických stromů. Předpokládá, že substituční rychlost je konstantní v celém stromu, takže distance (D) je přímo úměrná času (T) - naprosto přesně platí molekulární hodiny. Proto UPGMA předpokládá, že strom je ultramerický, všechny dnešní taxony „dosubstituovaly“ stejně daleko a na opačné straně leží kořen stromu. Tyto předpoklady jsou však téměř vždy porušeny. Pokud jsou předpoklady porušeny výrazně metoda vytvoří strom s nesprávnou topologií. Má tendenci posouvat divergentnější (rychle substituující) sekvence blíže ke kořeni stromu – artefakt **přitahování dlouhých větví (LBA).** LBA je jedno z největších úskalí molekulární fylogenetiky. Působení LBA si můžeme ukázat na následujícím příkladu. Představte si, že evoluce proběhla podle níže uvedeného stromu a délek větví. Všimněte si, že na větvi k taxonu C se zvýšila substituční rychlost, což se projevilo délkou větve. Změřením větví oddělujících OTU dospějeme k matici genetických distancí, která přesně odpovídá skutečnosti. Schválně, jestli UPGMA dokáže z matice zpětně zrekonstruovat strom, ze kterého byla matice odvozena.



	A	B	C	D
A	-			
B	**0,8**	-		
C	**0,9**	**0,5**	-	
D	**0,6**	**0,4**	**0,5**	-

Nedokáže. Hned v prvním kroku označí za nejbližší dvojici taxonů BD a shlukne je dohromady. Celý výsledný strom bude pak vypadat následovně.

	A	B	C	D
A	-			
B	0,8	-		
C	0,9	0,5	-	
D	0,6	0,4	0,5	-



Taxon C považovala UPGMA vinou jeho zrychlené substituční rychlosti za taxon, který se odvětvil dříve, než tomu bylo ve skutečnosti.

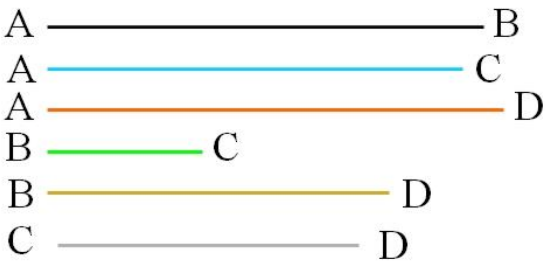
Metoda nejmenších čtverců (Least squares)

Tato metoda používá druhou strategii pro hledání "nejlepšího stromu". Nepostupuje pomocí algoritmu, ale tak, že jednotlivé topologie, které ji předložíme, ohodnotí podle kritéria, které si představíme níže, a pak z topologií vybere tu, která v tomto kritériu dopadne nejlépe. Ideální by bylo samozřejmě ohodnotit všechny možné topologie, ale to není vždy možné v reálném čase udělat. Proto byly vyvinuty algoritmy, které chytře vybírají vzorek topologií ze všech možných tak, aby pokud možno neminuly ten nejlepší strom. Tyto algoritmy si představíme v příští přednášce.

Nyní k používanému kritériu kvality. Pro danou topologii zkouší měnit délky větví (vnitřních i terminálních) dokud neminimalizuje parametr Q, který se počítá následovně

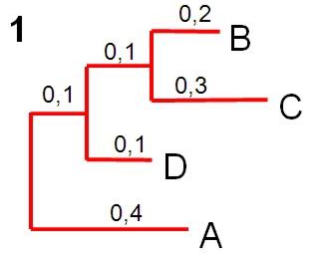
$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \, (D_{ij} - d_{ij})^2$$

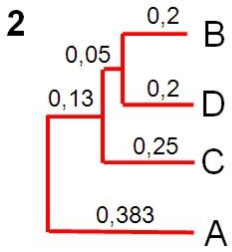
kde D_{ij} je pozorovaná distance mezi dvojicí taxonů i a j (distance z matice) a d_{ij} je vzdálenost mezi taxony i a j naměřená na hodnocené topologii s momentální délkou větví. Můžeme si to představit graficky tak, že se snažíme napasovat úsečky odpovídající pozorovaným genetickým distancím z matice (Dji) do topologie stromu a měříme, jak moc přesahují či nedosahují. Čím přesněji do topologie padnou, tím menší je Q a tím, kvalitnější je strom. Při tomto napasovávání („fitování") měníme podle libosti délky větví stromu, ale nesmíme změnit topologii (pořadí odvětvování).



Minimální Q pro topologii se stává jejím skóre a délky větví, které toto minimální Q poskytly, se stávají délkami jejích větví. Topologie, která dosáhla celkově nejmenšího skóre, je zvolena jako nejlepší.

Níže uvádím výpočet skóre podle metody nejmenších čtverců pro správnou topologii č.1, podle které proběhla evoluce a ze které byly odvozena tabulka distancí, a pro topologii č. 2, se kterou přišla metoda UPGMA. Asi nikoho nepřekvapí, že topologii 1 má skóre Q=0, topologie 2 má skóre vyšší, a proto je podle metody nejmenších čtverců horší.





	A	B	C	D
A	-			
B	0,8	-		
C	0,9	0,5	-	
D	0,6	0,4	0,5	-

$$\mathbf{Q_1} = (0,8-0,8)^2+(0,9-0,9)^2+(0,6-0,6)^2+(0,5-0,5)^2 +(0,4-0,4)^2+(0,5-0,5)^2 = \mathbf{0,0}$$

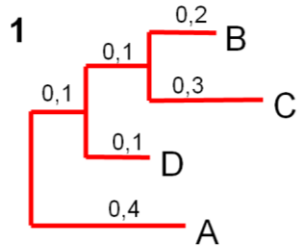
$$\mathbf{Q_2} = (0,8-0,763)^2+(0,9-0,763)^2+(0,6-0,763)^2 +(0,5-0,5)^2+(0,4-0,4)^2+(0,5-0,5)^2 = \mathbf{0,046707}$$

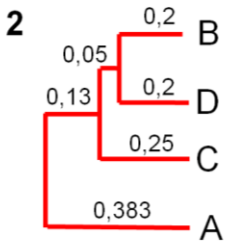
Vidíte, že metodě nejmenších čtverců nevadí, že substituční rychlost je v různých větvích různě veliká. Metoda nejmenších čtverců dokonce garantuje nalezení správné topologie, pokud jí poskytneme správně spočítané genetické vzdálenosti.
Parametr w_{ij} ve vzorci nejmenších čtverců umožňuje vážit příspěvek distancí vzhledem k jejich délce. Je smysluplné úměrně snížit příspěvek ke Q pocházejících od dlouhých větví, kde lze

očekávat větší odchylky, tak, aby nepřevážil nad vlivem odchylek u krátkých větví. Jako w se nejčastěji používá $1/D_{ij}^2$, pak se této metodě říká Fitch-Margoliash.

Minimální evoluce (Minimum evolution)

Metoda minimální evoluce funguje na podobném principu jako metoda nejmenších čtverců. U každé topologie nejprve optimalizuje délky větví pomocí nejmenších čtverců. Poté, na rozdíl od metody nejmenších čtverců, spočítá skóre topologie jako celkovou délku stromu (součet délek všech větví). Také metoda mimimální evoluce preferuje strom č. 1.





$$Q = \sum_{i=1}^{n} \sum_{j=1}^{n} D_{ij}$$

$\mathbf{Q_1}$= 0,2+0,3+0,1+0,1+0,1+0,4= **1,2**

$\mathbf{Q_2}$= 0,2+0,2+0,05+0,25+0,13+0,383=**1,213**

Neighbor-joining

Jedná se o algoritmizovanou verzi minimální evoluce. Princip spočívá v tom, že se postupně rozkládá hvězdicový strom tak, aby se v každém kroku maximálně snížila celková délka stromu. Stejně jako v případě UPGMA se vychází z matice genetických vzdáleností.

	A	B	C	D
A	-			
B	0,8	-		
C	0,9	0,5	-	
D	0,6	0,4	0,5	-

1. Pro každý vrcholový uzel (OTU) spočítáme koeficient u podle vzorce

$$u_i = \sum_{j:\, j \neq 1}^{n} D_{ij} / (n-2)$$

kde i a j označují vrcholové uzly. Pro uzly A, B, C a D vypadá v našem případě výpočet následovně

u_a = 0,8/2+0,9/2+0,6/2=1,15
u_b = 0,8/2+0,5/2+0,4/2=0,85
u_C = (0,9+0,5+0,5)/2=0,95
u_D = (0,6+0,4+0,5)/2=0,75

2. Matici genetických vzdáleností transformujeme následovně

$nD_{AB} = D_{AB} - u_A - u_B$ = 0,8-1,15-0,85=-1,2

a transformovaná matice vypadá takhle.

	A	B	C	D
A	-			
B	-1,2	-		
C	-1,2	-1,3	-	
D	-1,3	-1,2	-1,2	-

3. V transformované matici vybereme nejmenší distanci. V našem případě je to BC i AD. Je to způsobeno tím, že star decomposition stromu o čtyřech taxonech proběhne jen v jednom kroku, kdy vznikne plně rozlišený nezakořeněný strom – v našem případě bude mít tvar ((BC)(AD)). Je jedno, kterou dvojici si nyní zvolíme jako tu nejbližší. My si v našem výpočtu zvolíme dvojici BC.

4. Spojíme BC do jednoho uzlu a spočítáme délky větví od společného uzlu BC k vrcholu B (v_B) a vrcholu C (v_C).

$v_B = ½ D_{BC}+1/2(u_B - u_C) = ½$ 0,5+1/2(*0,85 – 0,95*) = 0,2
$v_c = ½ D_{BC}+1/2(u_C - u_B) = ½$ 0,5+1/2(*0,95 – 0,85*) = 0,3

Všimněte si, že na rozdíl od metody UPGMA nemusí tyto větve mít stejnou délku, čím metoda neighbor-joining umožňuje různou substituční rychlost v různých větvích.

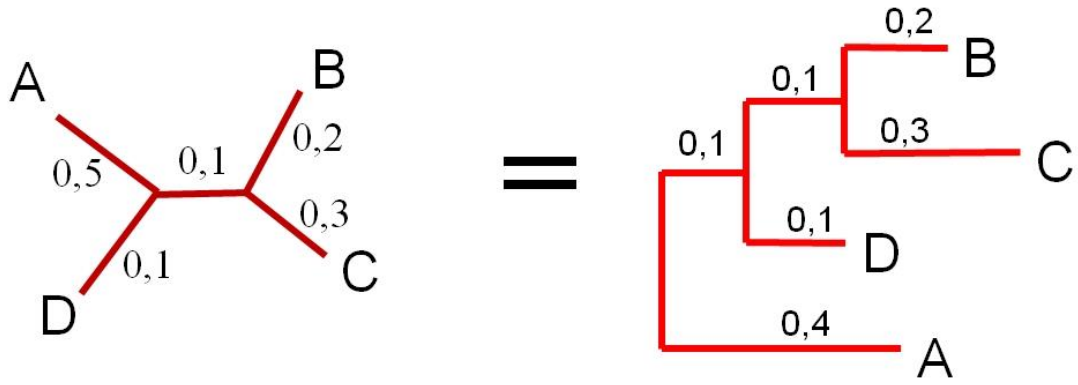
5. Nyní již víme, že budeme vytvářet společný uzel pro A a D. Délky větví od nového nodu AD k vrcholovým uzlům A a D určíme podobně jako předchozím bodě.

$v_A = ½ D_{AD}+1/2(u_A - u_D) = ½$ 0,6+1/2(1,15 – 0,75) = 0,5
$v_D = ½ D_{AD}+1/2(u_D - u_A) = ½$ 0,6+1/2(0,75 – 1,15) = 0,1

7. Prostřední větev $D_{(AD)(BC)}$ vypočítáme čistě geometricky

$$\mathbf{D_{(AD)(BC)} = (D_{AB} - v_A - v_B) = 0,1}$$

Stejně jako nejmenší čtverce garantuje neighbor-joining nalezení správného stromu pokud jsou přesně vypočítány genetické distance. Metoda neighbor-joining produkuje nezakořeněný neultramerický strom (vlevo)



.