# MOLEKULÁRNÍ TAXONOMIE – 10

## Molekulární hodiny

Skutečnost, že počet substitučních událostí vzrůstá s časem, vedla velmi brzy ke snahám využít sekvencí k datování stáří uzlů na fylogenetických stromech. Jako první s touto myšlenkou přišli pánové Emile Zuckerkandl a Linus Pauling v šedesátých letech. Populační genetika přinesla dokonce argument, proč by to mohlo fungovat. Lze totiž ukázat, že tikání molekulárních hodin (rychlost fixace mutací) je nezávislá na velikosti populace, a tak by neměla být ovlivňována různou populační velikostí u různých taxonů ani fluktuacemi ve velikostech populací, kterými druhy během evoluce procházejí.

**μ** si označíme mutační rychlost (počet nově vzniklých mutací za jednotku času u jednoho jedince). Počet nově vzniklých mutací za jednotku času v populaci je poté $\mu * N_e$ ($N_e$ = efektivní velikost populace). Pravděpodobnost, že tato vzniklá mutace bude v populaci fixována a stane se z ní substituce je $1/N_e$. Rychlost vzniku substitucí (vzniku těch mutací, které se fixují) je součinem

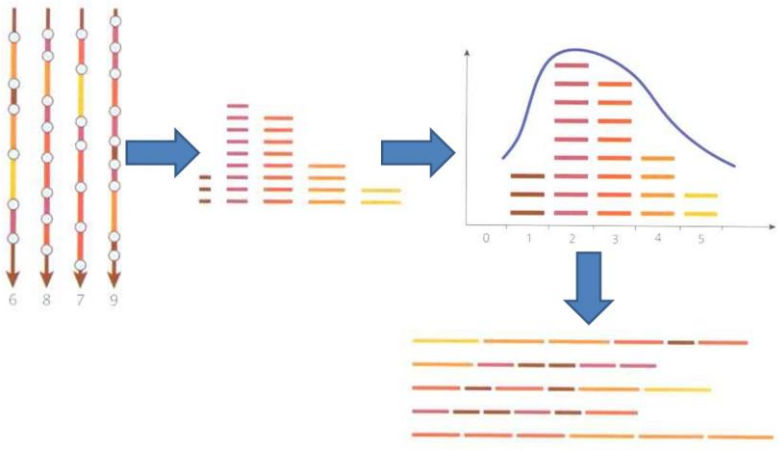
$$\mu * N_e * 1/N_e = \mu$$

a je tedy shodná s mutační rychlostí. Jinými slovy ve větší populaci sice vznikne více mutací, ale menší procento z nich dosáhne fixace, aby se z nich staly substituce, kterých si všímáme my v našich analýzách.

Substituční rychlost tedy nezávisí na efektivní velikosti populace. Přesto se velmi brzy ukázalo, že s aplikací molekulárních hodin souvisí mnoho problémů. Níže si představíme zdroje chyb, se kterými musíme při používání molekulárních hodin počítat.

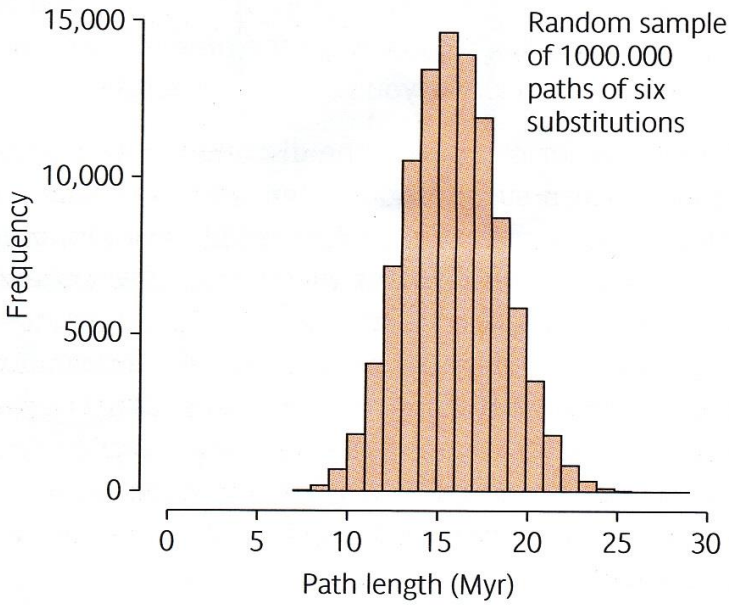
<u>Molekulární hodiny tikají nepravidelně</u>

Tyto hodiny jsou nepravidelně tikajícím chronometrem. Stejně jako u rozpadu nestabilních izotopů atomů je jejich rychlost uváděna jen jako odhad získaný z dlouhodobého průměru. Na obrázku níže je uveden příklad čtyř linií, jejich molekulární hodiny běží stejně rychle. Přesto v časovém úseku, který je tam znázorněn, došlo v jednotlivých liniích k různému počtu substitucí (kolečka). Kdybychom si nalámali časové úseky mezi substitucemi, zjistili bychom, že jejich frekvence (počty výskytu) jejich délek má určité rozdělení. V našem případě se tam vyskytovalo




nejvíce tmavě červených úseček, které budou zřejmě nejblíže odpovídat dlouhodobě průměrné rychlosti, nicméně vyskytují se tam i úsečky jiných délek. Co to pro nás znamená? I když víme, že mezi dvěma organismy došlo k 6 substitucím a známe průměrnou rychlost, nemůžeme přesně zjistit, jaký evoluční čas je odděluje. Když náhodně sáhneme pro 6 úseček do našeho rozdělení a poskládáme je na sebe, můžeme totiž získat velmi

rozdílné časové úseky. Kdybychom to udělali mnohokrát, zjistili bychom, že délky našich úseček budou opět mít rozložení jako na obrázku na následující straně.




Opět platí, že nejčastěji se budou vyskytovat délky odpovídající průměrné rychlosti, ale s nemalou frekvencí narazíme i na výrazně jinou délku. Z nepravidelného chodu molekulárních hodin vyplývá, že i kdybychom znali přesně průměrnou substituční rychlost, odhad uplynulého času bude vždy mít jistý rozptyl.

## Molekulární hodiny tikají různě rychle v různých genech

Tabulka níže ukazuje rychlost chodu molekulárních hodin pro různé geny (počet substitucí za rok). Je patrné, že genetické vzdálenosti naměřené pro různé geny nelze srovnávat, pokud nepřihlédneme k tomu, že u různých genů tikají hodiny jinak. Příklad je uveden v tabulce vpravo. To přináší naštěstí také výhodu v tom, že máme možnost zvolit si pro analýzu určité časové hloubky vhodně gen s vhodnou rychlostí chodu hodin.

TABLE 2. Characteristics affecting the phylogenetic performance of mitochondrial genes.

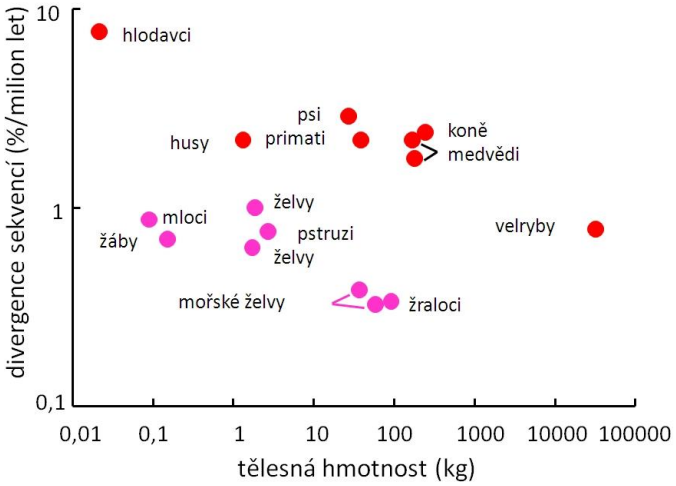
| Gene | Length[a] | Rate of evolution[b] | $\alpha$[c] | Amino acid ML distance[d] |
|---|---|---|---|---|
| *rrnL* | 1054 bp/1630 bp | .07 (.03) | 0.32 | NA |
| *nad4* | 1281 bp/1378 bp | .44 (.08) | 0.24 | 0.4 (0.14) |
| *nad2* | 1000 bp/1044 bp | .37 (.09) | 0.34 | 0.6 (0.2) |
| *nad5* | 1698 bp/1833 bp | .42 (.09) | 0.25 | 0.42 (0.12) |
| *tRNAs* | 1282 bp/1656 bp | .12 (.03) | 0.51 | NA |
| *cob* | 1141 bp/1141 bp | .62 (.16) | 0.16 | 0.26 (0.06) |
| *rrnS* | 726 bp/958 bp | .07 (.02) | 0.32 | NA |
| *cox1* | 1530 bp/1530 bp | 1.04 (.27) | 0.11 | 0.06 (0.03) |
| *atp6* | 683 bp/683 bp | .45 (.11) | 0.27 | 0.45 (0.13) |
| *cox3* | 784 bp/784 bp | .64 (.12) | 0.14 | 0.19 (0.06) |
| *nad1* | 939 bp/970 bp | .43 (.11) | 0.22 | 0.28 (0.09) |
| *cox2* | 687 bp/687 bp | .47 (.22) | 0.22 | 0.18 (0.1) |
| *nad6* | 396 bp/522 bp | .24 (.05) | 0.38 | 0.99 (0.44) |
| *nad3* | 330 bp/354 bp | .42 (.09) | 0.22 | 0.5 (0.13) |
| *nad4L* | 290 bp/290 bp | .31 (.06) | 0.31 | 0.41 (0.17) |
| *atp8* | 78 bp/158 bp | .16 (.02) | 0.29 | (too short) |

[a] Length shows the number of unambiguously alignable base pairs/total length of the alignment.
[b] Mean rates of evolution are in number of substitutions per nucleotide site per 100 million years. Standard deviations are given in parentheses.
[c] $\alpha$ Describes the among-site rate heterogeneity of the nucleotide sequence, estimated with no invariant sites. Length, rate of evolution, and $\alpha$ were included in the logistic regression analysis to determine their impact on phylogenetic performance.
[d] See Discussion for the relevance of amino acid ML distances.

Molekulární hodiny tikají různě rychle u různých organizmů

Asi největším problémem je, že molekulární hodiny pro tentýž gen tikají různě rychle u různých skupin, jak ilustruje graf.

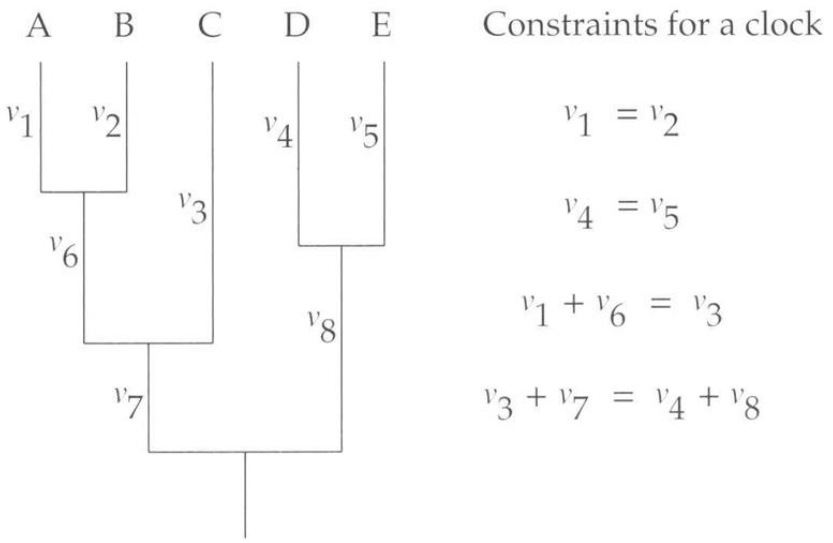
Martin a Palumbi, PNAS USA 90: 4087-4091, 1993




Je tedy zřejmé, že rychlost chodu molekulárních hodin se mění v evolučních liniích. Fylogenetické metody, které jsme si představovali dosud, se s tímto problémem vypořádaly tak, že každé větvi na stromu přisoudili jinou délku $\mathbf{t}_i$. Délky větví se staly parametry určujícími pravděpodobnosti záměn a podstatnou součástí analýzy maximum likelihood je právě hledání takové kombinace délek větví, která maximalizuje pravděpodobnost dat. Samotná délka větve $\mathbf{t}_i$ ovšem neodráží čas, ale je součinem $\mathbf{t}_i=\mathbf{u}_i*\mathbf{t}$ času a substituční rychlosti pro příslušnou větev. Aby se nám to nepletlo, budu odteď délku větve označovat raději **v**, takže $\mathbf{v}_i=\mathbf{u}_i*\mathbf{t}_i$. Metody, kterými jsme se zabývali na předchozích přednáškách, umožnily substituční rychlosti měnit se na různých větvích, ale musely rezignovat na to, abychom se dozvěděli, jaký podíl na délce větve tvoří čas **t** a jaký substituční rychlost **u**. Délky větví těchto stromů obvykle nekončí stejně daleko, což je důkazem, že neodráží čas a, že se substituční rychlosti skutečně mění. Další nevýhodou je, že tyto předešlé metody produkují nezakořeněné stromy. Metody, které se představíme níže, se pokouší čas a substituční rychlost osamostatnit. Přitom se musí jednak vypořádat s nestálostí substituční rychlosti napříč stromem a jednak jim musíme poskytnout informaci o stáří minimálně jednoho uzlu na stromu, strom kalibrovat.

Testování rovnoměrnosti chodu molekulárních hodin

Zda je substituční rychlost napříč stromem proměnlivá lze testovat v likelihoodovském rámci pomocí likelihood ratio testu (viz. předchozí přednáška) Při takovém testování porovnáváme dvě hypotézy. První složitější hypotéza $H_1$ předpokládá, že délky větví v1-v8 na stromu níže jsou nezávislé parametry. Jednodušší hypotéza $H_0$ předpokládá existenci homogenní substituční rychlosti (globálních molekulárních hodin) v naší sadě dat, z čehož plyne, že délky větví nejsou nezávislé, ale platí mezi nimi vztahy uvedené vpravo.

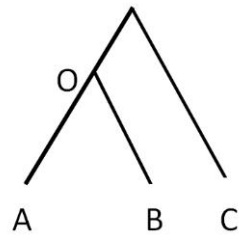



Hypotéza $H_0$ má méně parametrů, protože nám stačí znát délky větví v1, v4, v6 a v7. Délky větví v2, v3, v5 a v8 jsou pak spočitatelné. Nulová hypotéza je tedy jednodušším případem hypotézy H1, a má o 4 parametry méně. Pro jejich porovnání můžeme použít **likelihood ratio test** s tím, že statistika tohoto testu

## $\delta=2(\ln L_1-\ln L_0)$

bude mít rozložení $X^2$ se čtyřmi stupni volnosti. Pokud rozdíl v likelihoodech nebude signifikantní, můžeme předpokládat, že se substituční rychlost mezi liniemi výrazně nemění a že na naši fylogenezi lze aplikovat globální molekulární hodiny.

Dalším testem, který porovnává substituční rychlost dvojice taxonů je **relative rate test**. Tento test se provádí vždy pro dvojici taxonů. Nejprve se spočítá rozdíl délky větve (genetické distance) každého z nich A a B od společného předka (O). Sekvenci společného předka neznáme, ale může nám posloužit sekvence outgroupu (C), tedy organizmu ležícího mimo porovnávanou dvojici. Rozdíl genetických distancí od každého srovnávaného taxonu a společného outgroupu je totiž totéž, protože část stromu O-C je shodná pro oba taxony a rozdíl tedy vzniká jen na větvích O-A a O-B.




Spočítáme tedy statistiku **d**

**$d = D_{AC} - D_{BC}$**

tato statistika bude mít opět nějaké rozložení charakterizované rozptylem V a směrodatnou odchylkou SE. Rozptyl této statistiky V(d) si můžeme spočítat z rozptylu genetických distancí (vzpomeňte si, že každá genetická distance má rozptyl, přednáška 5) a směrodatná odchylka (SE) je odmocninou rozptylu.

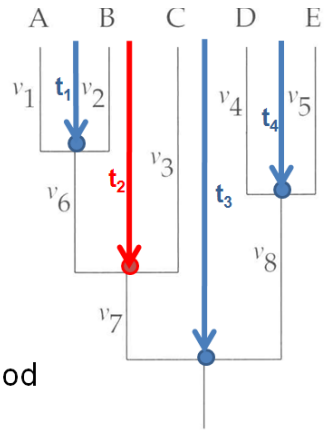
$$\mathbf{V(d) = V(D_{AC}) + V(D_{BC}) + 2V(D_{OC})}$$

Platí, že pokud je d >= 2*SE je tento rozdíl signifikantní na hladině pravděpodobnosti 5%, pokud je d >= 2,7*SE je tento rozdíl signifikantní na hladině pravděpodobnosti 1%. Tento test je jednoduchý, ale má poměrně malou sílu. To znamená, že i když v něm vyjde rozdíl nesignifikantní, může být ve skutečnosti významný a negativně ovlivňovat výsledky analýz.

Globální molekulární hodiny

Pokud tedy vyházíme taxony, které porušují jednotně tikající hodiny, a budeme předpokládat, že ve zbytku našich dat tikají molekulární hodiny stejně nacházíme se v poměrně jednoduché situaci. Máme-li topologii stromu a známe-li stáří alespoň jednoho uzlu (kalibrační bod) můžeme v metodou maximum likelihood snadno odhadnout stáří ostatních uzlů. V takovém případě totiž platí na stromu vztahy uvedené níže.

$$v_1=v_2=t_1\mu$$
$$v_4=v_5=t_4\mu$$
$$v_3=v_6+v_1=t_2\mu$$
$$v_8+v_4=v_7+v_6+v_1=t_3\mu$$

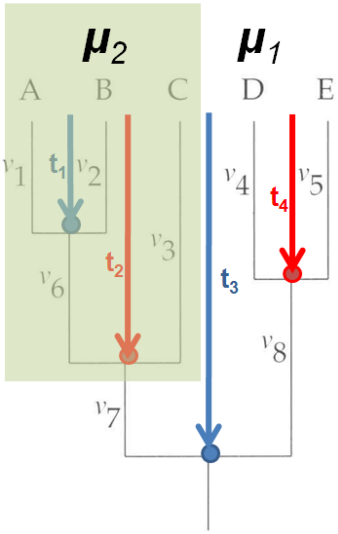



$t_2$ je v tomto příkladu kalibrační bod

Při výpočtu likelihoodu budeme optimalizovat parametry substitučního modelu, ale místo délek větví $\mathbf{v_1}$-$\mathbf{v_8}$ budeme optimalizovat stáří uzlů $\mathbf{t_1}$, $\mathbf{t_3}$ a $\mathbf{t_5}$ a jednotnou celkovou substituční rychlost **μ.** Tato celková rychlost však stále může být, v závislosti na použitém substitučním modelu, rozložena do rychlostní matice Q, specifické rychlosti pozic alignmentu atd. Platí však, že průměrná rychlost substituce je zmíněné **μ**, které platí globálně pro celý strom. $\mathbf{t_1}$, $\mathbf{t_3}$ a $\mathbf{t_5}$, **μ** a parametry substitučního modelu postačí k výpočtu pravděpodobnosti alignmentu (likelihoodu). Hodnoty, které poskytnou nejvyšší likelihood, budou nejlepšími odhady stáří uzlů. Všimněte si, že tato metoda skutečně rozpřáhla substituční rychlost a čas. Výhody globálních hodin spočívají v tom, že odhady na nich založené mají užší intervaly spolehlivosti (model má méně parametrů). Dále nám stačí znát méně kalibrační bodů. Dokonce stačí jen jeden, ale čím více tím lépe. Pokud však globální hodiny neplatí (a my s nimi počítáme) výsledky budou zcela špatně.

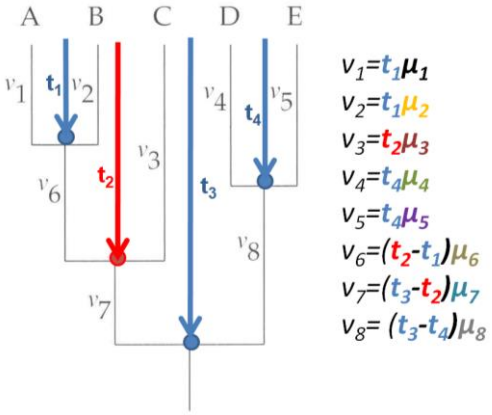
Lokální molekulární hodiny

Alternativou globálních hodin jsou **lokální molekulární hodiny**. V takovém případě strom rozdělíme na několik oblastí a v každé předpokládáme platnost lokálních hodin. Pro každou oblast stromu ovšem potřebujeme kalibrační bod. Navíc musíme vědět, na kolik a na jaké části strom rozdělit. Lokální molekulární hodiny navíc obsahují nereálný předpoklad, že se substituční rychlost mění skokově z jedné části stromu na druhou. Přitom je zřejmé, že substituční rychlost se mění plynule.

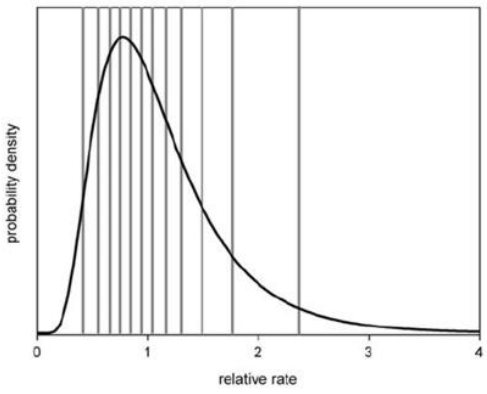



Relaxované molekulární hodiny

Nejrealističtějším modelem molekulárních hodin jsou **relaxované molekulární hodiny**. Ty předpokládají, že každá větev na stromu má svoji vlastní substituční rychlost. Dělá se to tím způsobem, že se substituční rychlosti větví tahají náhodně z rozložení jejich frekvence. Oblíbeným rozložením v tomto případě, je třeba lognormální rozložení na obrázku níže. (Je to obdobné, jako použití Γ rozložení pro tahání relativních rychlostí jednotlivých pozic alignmentu)




$v_1=t_1\mu_1$
$v_2=t_1\mu_2$
$v_3=t_2\mu_3$
$v_4=t_4\mu_4$
$v_5=t_4\mu_5$
$v_6=(t_2-t_1)\mu_6$
$v_7=(t_3-t_2)\mu_7$
$v_8=(t_3-t_4)\mu_8$



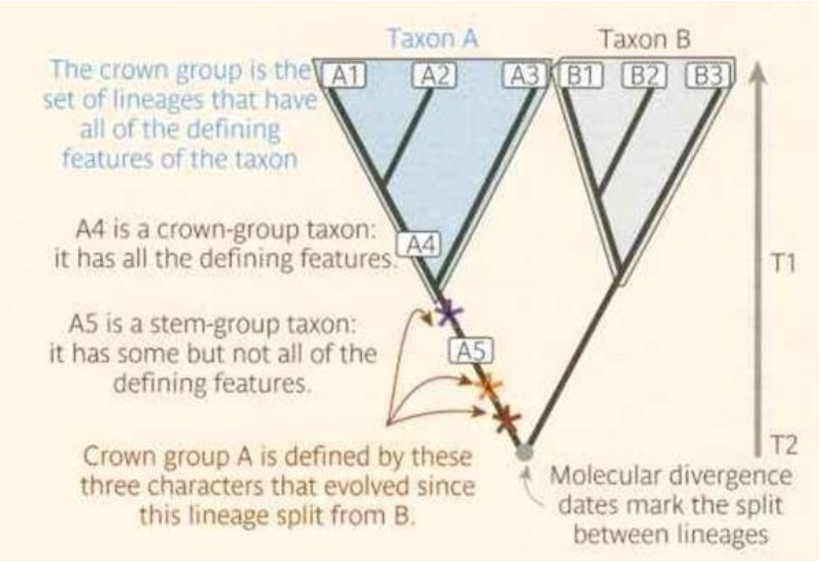

Lognormální rozložení si rozdělíme na 12 diskrétních kategorií podobně, jako jsme to dělali s rozložením funkce gamma při modelování různé substituční rychlosti pozic. Plochu pod křivkou rozdělíme na 12 stejně velkých ploch (rychlostních kategorií) a každou kategorii bude zastupovat průměrná hodnota. Kromě těchto nezávislých relaxovaných hodin se používají také autokorelované relaxované hodiny. V takovém případě je rozložení rychlostí dceřiné větve závislé na substituční rychlosti mateřské větve podle nějakého vztahu.

Relaxované hodiny se podařilo úspěšně implementovat do Bayesiánského rámce a používají je programy BEAST a Phylobayes. Tyto programy nechávají běžet Marcov Chain Monte Carlo prostorem hypotéz. Hypotézy sestávají, kromě topologie, také z hloubek (stáří) nodů a různých kombinací substitučních rychlostí větví. Všechny tyto parametry se mění a podle nám již známého scénáře dospěje MCMC dříve nebo později do rovnovážného stavu, který určuje

posteriorní pravděpodobnosti hypotéz. Tato metoda tedy umožňuje nejen různé substituční rychlosti v různých částech stromu, ale také umožňuje optimalizovat hodnoty stáří nodů společně s topologií. Rekonstruuje tedy fylogenezi, dokonce zakořeněnou a zároveň poskytne stáří uzlů.

Kalibrace

Rovněž kalibrační body je třeba interpretovat opatrně. Jako kalibrační bod může sloužit nejčastěji fosílie, fosilní DNA nebo chemická látka, kterou produkuje výhradně určitá skupina organizmů. Stáří takového kalibračního bodu není přesné číslo, ale hodnota, která má jistou odchylku a z ní vyplývající konfidenční interval. Tyto intervaly nám poskytnou metody, kterými bylo stáří vzorků určeno a je s nimi možné a dobré počítat při analýze pomocí molekulárních hodin.
Pokud je kalibračním bodem fosílie, pak vstupuje do hry ještě nejistota ohledně její pozice na stromu. Představte si, že taxon A je definován přítomností tří znaků (barevné hvězdičky na obrázku níže). Nejstarší fosílie, která má všechny tři znaky a patří tedy do taxonu A, na našem obrázku je to A4, pravděpodobně neleží přímo na uzlu společného předka taxonu A. Její stáří nám, říká, že taxon A již v tuto dobu existoval, ale jak dlouho předtím vznikl, nevíme, protože nemáme k dispozici všechny fosílie. Stáří takové fosílie nám tedy udává minimální stáří společného předka taxonu A. Naopak fosílie A5, která má některé ze znaků taxonu A, ale ne všechny, leží snad někde na stonkové větvi taxonu A a v takovém případě udává maximální stáří jeho společného předka. Může ovšem také ležet na vyhynulé větvi paralelní se stonkem taxonu A a být tak dokonce mladší než společný předek A. Nejistota ohledně stáří fosílie a jejím postavení na stromu přispívá k šíři konfidenčního intervalu.

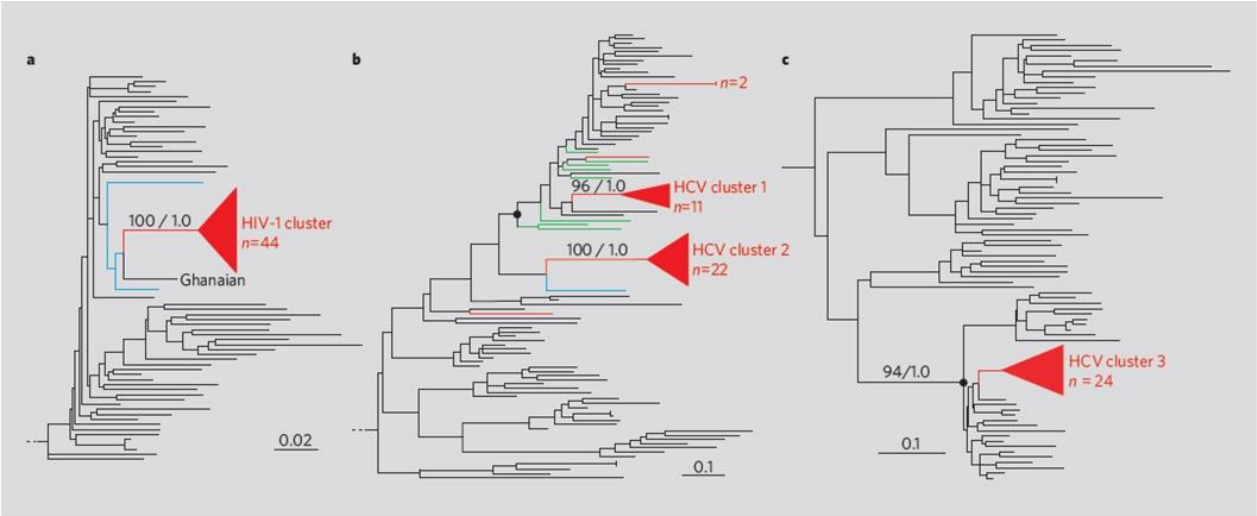



Konfidenční intervaly

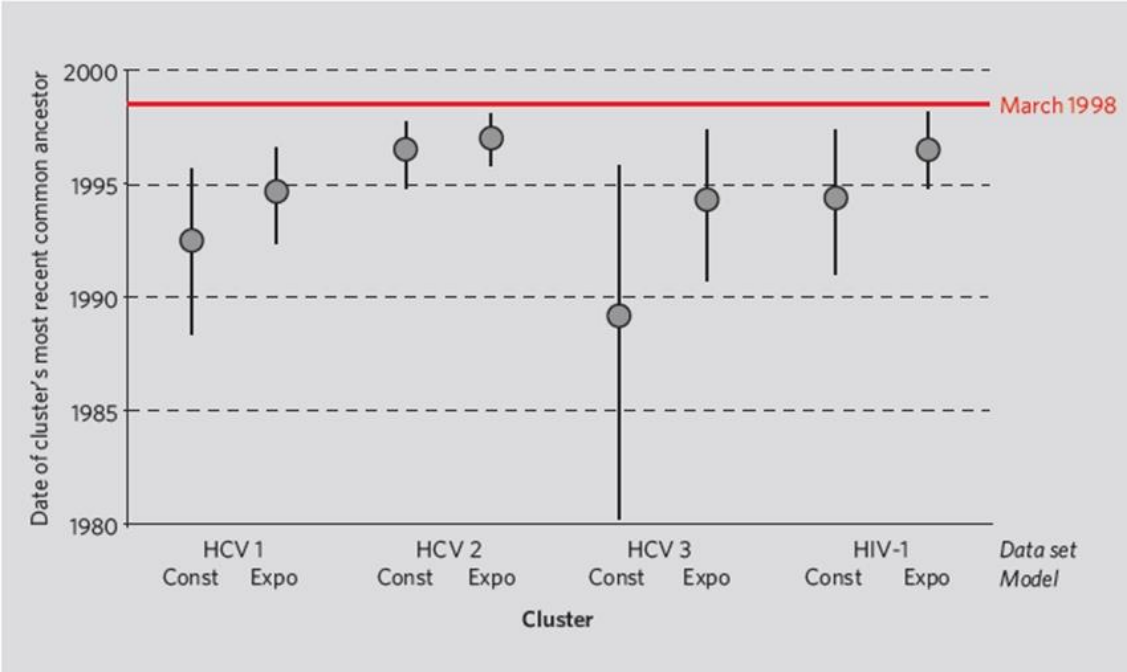
Je třeba zdůraznit, že hodnoty hloubek uzlů u všech zmíněných metod nejsou přesná čísla, ale odhady, které mají jistou chybu. Ta se vyjadřuje nejčastěji formou konfidenčních intervalů (intervalů spolehlivosti) - tj. intervalů hodnot, mezi kterými se nachází skutečná hodnota se zvolenou pravděpodobností (95%, 99%). Obecně platí, že konfidenční intervaly odhadů vytvořených složitějšími metodami s více parametry (relaxované a lokální hodiny) jsou širší než u jednoduchých metod typu globální hodiny.

Epidemie HIV v nemocnici Al-Fateh v Benhazi

Široké konfidenční intervaly jsou nepříjemné, ale pokud jsou pravdivé, tj. můžeme se spolehnout, že skutečná hodnota leží uvnitř, pak mohou někdy přinést odpověď na otázku, kterou si klademe. Příkladem může být studie, která reagovala na případ epidemie HIV a hepatitidy C k libyjské nemocnici Al-Fateh v Benhazi. Tam se poté, co v březnu 1998 přišel zahraniční personál (palestinský doktor a bulharské sestry), začaly vyskytovat případy těchto onemocnění u dětí. Na základě toho byli zahraniční pracovníci obviněni, vězněni a poté odsouzeni k trestu smrti, který byl naštěstí odložen, a v roce 2007 byli všichni po diplomatickém nátlaku propuštěni. Přes 200 nakažených dětí bylo hospitalizováno v Evropě, a proto měl vědecký tým přístup k jejich vzorkům. Ze 44 těchto vzorků získali sekvence jednoho protein kódujícího genu pro viry HIV a hepatitidy C. Fylogenetickou analýzou ukázali, že viry HIV pocházejí ze společného předka (obr. dole a), kdežto viry hepatitidy C se seskupily do tří větších skupin a několika osamělých větví (obr. dole b a c).




Protože sekvence viru se vyvíjí velmi rychle a mění se rok od roku, bylo možné sekvence referenčních vzorků odebraných v různých letech použít jako kalibrační body. S použitím různých metod datování podle molekulárních hodin (globálních a relaxovaných) dospěli k odhadu doby, kdy žil společný předek těchto virů. Tyto odhady měly velké intervaly spolehlivosti, ale březen 1998 ležel vždy nad nimi, což vylučuje zavinění zahraničního personálu. Viz obrázek níže.


Date of cluster's most recent common ancestor
2000
1995
1990
1985
1980
March 1998
HCV 1
Const
Expo
HCV 2
Const
Expo
HCV 3
Const
Expo
HIV-1
Const
Expo
Data set
Model
Cluster