MOLEKULÁRNÍ TAXONOMIE - 5 (2018)

Single nucleotide polymophism - SNP

Polymorfimus DNA, kdy se jedinci nebo druhy liší v jedné nukleotidové záměně

AAGC**C**TA
AAGC**T**TA

V tomto případě mluvíme o alelách C a T. Téměř všechny SNPy mají jen 2 alely, protože je málo pravděpodobné, že by v populaci konkrétní nukleotid zmutoval hned dvakrát. Genom dvou lidí se liší zhruba ve 3 mil. bází, ale ne všechno jsou SNP. Databáze SNP v rámci NCBI (http://www.ncbi.nlm.nih.gov/projects/SNP/) k březnu 2012 eviduje přes 52 miliónů různých SNP u člověka, přes 15 miliónů SNP u myši a tak dále.

SNP se využívají pro mapování genomu a hledání genů souvisejících s fenotypy, které nás zajímají (například choroby). Pokud přítomnost SNP v populaci přesně odpovídá výskytu choroby nebo s ní signifikantně koreluje, je zřejmé, že tento SNP je nějak svázán s genem, jehož porucha tuto chorobu vyvolává. Buď je tento SNP uvnitř tohoto genu a jedna jeho alela přímo způsobuje tuto chorobu nebo, a to je častější případ, SNP leží poblíž genu a je s ním v genetické vazbě. Pátrání po genech v okolí takového SNP může vést k odhalení problematického genu. V molekulární taxonomii mají SNP podobné využití jako mikrosatelity. Stejně jako mikrosatelity se Mendelovsky dědí, kombinace SNP je specifická pro jedince a ukazuje na jeho příbuznost s jinými jedinci v populaci. Proto můžeme SNP využít v identifikaci jedinců, určování rodičovství nebo v populačních studiích.

Pokud chceme nalézt SNP v genomech organizmů, pro které zatím SNP známy nejsou, je nejjednodušší způsob osekvenovat genomy od více jedinců a porovnat je vzájemně. Problém je, že velké množství zdánlivých SNP budou ve skutečnosti chyby. Z tohoto důvodu je potřebovat sekvenovat s vyšším pokrytím. Pokud nemůžeme investovat tolik prostředků do celogenomového sekvenování, můžeme postupovat tak, že si připravíme směsný vzorek DNA dvou jedinců a budeme sekvenovat náhodné části smíchaných genomů. V některých místech uvidíme dvojitý signál. Některé dvojité signály budou opět chyby nebo nepřesnosti sekvenace. Některé mohou představovat SNP - místa, ve kterých mají dva použité vzorky DNA jiný nukleotid.

Pokud však hodláme genotypizovat na SNP jedince modelového organizmu, pak máme k dispozici nepřeberné množství metod SNP genotypizace, které se neustále rozvíjejí a mění. Některé z nich si představíme. Výhodou mnoha z nich je, že nám v jenom kroku otypují obrovské množství SNP. Tyto metody jsou založeny na hybridizačních technikách, enzymatických reakcích nebo na jiných principech.

Hybridizační technikou je například Molecular beacon. U této metody se používají fluorescenčně značené próby, které mají uprostřed své molekuly místo přesně komplementární

s lokusem obsahujícím konkrétní alelu SNP a na krajích mají sekvence komplementární navzájem. Na jednom konci próby je navázána fluorescenční barva na druhém konci tzv. quencher, který inhibuje fluorescenci. V “klidovém” stavu próba vytváří formu vlásenky s kličkou. U této formy jsou fluorochrom a quencher blízko sebe a sonda nesvítí. Pokud se dostane do kontaktu s komplementární DNA, tedy se “svou” alelou, její struktura se otevře a my můžeme detekovat fluorescenci. Pokud necháme vzorek DNA inkubovat s próbami na všechny známé alely SNP obarvené různými barvičkami, můžeme podle barevného signálu odečíst, která alela/alely se u jedince vyskytuje.

Hybridizační metody založené na mikroarray čipech dovolují skrínovat velké množství SNP najednou. Oligonukleotidy komplementární s lokusem, kde je známý SNP jsou imobilizovány na sklíčku těsně vedle a známe jejich polohu. Oligonukleotidy představující čtyři možné formy daného SNP, tj. lišící se v jednom nukleotidu představujícím SNP, jsou obvykle vedle sebe. Vzorek DNA je naštípán na krátké fragmenty, fluorescenčně označen, hybridizován na čip a pak je odečten světelný signál. Čipy firmy Affimetrix umožňují naráz oskrínovat 906 tis. známých lidských SNP. Problém hybridizačních metod spočívá především v jemném vyladění podmínek hybridizace tak, aby próby na alely SNP vzájemně nekrosreagovaly. U hromadného skrínování je tento problém ještě větší, protože různé oligonukleotidy vyžadují různé optimální podmínky hybridizace. Proto je na mikroarray čipech každý SNP analyzován víckrát na různých místech čipu v rámci různých oligonukleotidů z jeho lokusu.

Z enzymatických metod si představíme metodu Infinium od firmy (Illumina). Na sklíčku jsou těsně vedle sebe do známých míst připevněny kuličky, na kterých jsou navázány oligonukleotidy. Každá kulička nese oligonukleotidy jednoho typu, které jsou komplementární se sekvencí známého SNP lokusu a jejich volný 3’ konec končí o jeden nukleotid před SNP. Vzorek DNA se naštípe na náhodné fragmenty určitých délek, které se denaturují a hybridizují na sklíčko, takže fragment se SNP lokusem se zachytí na oligonukleotidu čouhajícím z kuličky. Následně dojde k polymeraci a DNA polymeráza prodlouží oligonukleotid navázaný na kuličce o jeden nukleotid podle templátu z DNA vzorku - jedná se právě o polymorfní SNP nukleotid. Pomocí fluorescenčně značených protilátek proti čtyřem možným nukleotidům se detekuje, jaký nukleotid, v případě heterozygota jaké dva nukleotidy, byly připolimerovány. Ten představuje SNP genotyp.
Starou a jednoduchou, avšak stále používanou metodou analýzy SNP polymorfismu v jednom lokusu je SSCP (Single Strand Conformation Polymorphism). SNP lokus amplifikujeme pomocí PCR. Produkty pak denaturujeme teplotou, aby se rozdělily na jednotlivé řetězce, a necháme je renaturovat. Odstraníme dsDNA a zbudou nám ssDNA řetězce, které renaturovaly samy se sebou, přičemž vytvořily komplexní 3D struktury. Ty rozdělíme na elektroforéze. Rychlost jejich migrace je dána ani ne tak délkou, jako tvarem, který zaujmou. Na tvaru těchto 3D strukturách, a tedy i na jejich elektromobilitě, se projeví i substituce v jednom nukleotidu, která by elektromobilitu dsDNA neovlivnila. U homozygotů nalezneme dva pruhy, každý odpovídá jednomu vláknu DNA. U heterozygotů 4 pruhy.

Na posledních dvou snímcích najdete srovnání všech probíraných “nesekvenačních” metod získávání molekulárních dat. Metodu “microcomplement fixation” jsem v přednášce vynechal.

VÝPOČET GENETICKÝCH DISTANCÍ

Alignované sekvence dvojice taxonů stejně jako výstupy nesekvenačních metod lze převést na genetickou distanci. Genetická distance je mírou odlišnosti dvou organismů a podle toho, o jaký typ distance se jedná, vyjadřuje procento rozdílných nukleotidů, počet substitucí na jeden nukleotid, podíl odlišných pruhů ve fingerprintovém vzoru nebo rozdíly ve frekvencích alel mezi populacemi. Z genetických distancí lze konstruovat fylogenetické stromy, což bude tématem další přednášky.

Distance z podobnosti vzorů

Bylo navrženo několik koeficientů, které převádí podobnost fingerprintového vzoru na genetickou distanci. Na snímku uvádím jednoduchý a poměrně intuitivní koeficient podle Nei a Li (1979).

Pro každou dvojici (na snímku dráhy X a Y) spočteme počet všech fragmentů v dráze (**Mx, My**) a dále počet fragmentů, které se vyskytují v obou drahách (**Mxy**). Vypočteme podíl shodných fragmentů

I = 2Mxy/(Mx + My)

distance je doplňkem rozdílu.

D= 1- I

V našem příkladě Mx=8, My=7, Mxy=7 a D=0,06666. Při porovnávání vzorů nás zajímá pouze délka pruhu, nevšímáme si jeho tloušťky. Pochopitelně, že odhad genetické distance založený na 8 pruzích bude velmi nepřesný. Pro vzorky X a Y je potřeba vytvořit větší počet fingerprintů a hodnoty Mx, My a Mxy sečíst pro všechny fingerprinty.

Koeficient Nei-Li patří do kategorie geometrických koeficientů, které neberou v potaz mechanizmy, které stojí na pozadí vzniku či zániku fragmentů ve vzoru - pravděpodobnosti zániku nebo naopak vzniku restrikčního místa nebo místa nasedání primeru. Byly vyvinuty také koeficienty, které toto dokáží. Ty pracují s délkou restrikčního místa, případně délky fragmentů, a jejich odvození je složitější a přesahuje náplň této přednášky.

Distance z frekvence alel

Genetickou vzdálenost mezi populacemi můžeme spočítat z frekvence alel (mikrosatelitů, SNP, alozymů) v těchto populacích například pomocí Rogersovy vzdálenosti. Pro každý lokus spočítáme distanci D následovně

$$D= (0{,}5\ \Sigma(x_{Ai} - x_{Bi})^2)^{0,5}$$

kde $\mathbf{x_{Ai}}$ a $\mathbf{x_{Bi}}$ jsou frekvence alely i v populacích A a B.
Příklad:

Frekvence alel v jednom lokusu

Alela	Populace A	Populace B
1	0,12	0,20
2	0,48	0,30
3	0,40	0,50

$D= (0,5((0,12-0,20)^2+(0,48-0,30)^2+(0,40-0,50)^2))^{0,5}= (0,5(0,0064+0,0324+0,01))^{0,5}= 0,156$

Dalším oblíbeným koeficientem je distance Cavali-Svorza a Edwardse (1967)

$$D_{CH} = \frac{2}{\pi}\sqrt{2(1-\sum_{u}\sqrt{X_u.Y_u})}$$

kde X_u a Y_u jsou frekvence alely u v populacích X a Y.

Podobně jako v případě distancí z fingerprintových vzorů. Rogersova i Cavali-Svorza a Edwardsova distance neberou v potaz biologické pozadí stojící za změnami ve frekvencích alel v populaci případně za mutacemi alel (např. prodlužování a zkracování mikrosatelitů). Distance, které toto umějí, byly také vyvinuty, např. Reynoldsova distance (1983) nebo Neiova distance (1972, 1978), avšak na tomto místě se jimi zabývat nebudeme.

Pokud porovnáváme mezi populacemi více lokusů, spočítáme celkovou vzdálenost jako aritmetický průměr vzdáleností pro jednotlivé lokusy.

Frekvence rozdílných nukleotidů

Označuje se p. Známe-li tyto dvě sekvence, můžeme jej vypočítat jednoduše jako

$$p=n_d/n$$

kde n_d je počet rozdílných nukleotidů a n je počet všech nukleotidů.

Podíl rozdílných nukleotidů můžeme také odhadnout pomocí DNA - DNA hybridizace (viz předchozí přednáška). V tomto případě

$$p = \Delta Tm \,.\, 0,01\ (0,015)$$

Podíl rozdílných nukleotidů můžeme odhadnout také z počtu shodných restrikčních míst (v přednášce jsem to neuváděl).

1. Sestavíme restrikční mapy pro každou OTU
2. Pro každou dvojici sekvencí (x, y) spočteme všechna restrikční místa (Mx, My) a dále místa vyskytující se v obou sekvencích (Mxy)
3. Vypočteme podíl shodných restrikčních míst

$$S = 2Mxy/(Mx + My)$$

4. Vypočteme odhad podílu nukleotidů, ve kterých se sekvence neshodují

$p = 1 - S^{1/r}$ r-délka restrikčního místa

Vzdálenosti mezi genomy

Genetické distance mezi celými genomy lze spočítat také na základě srovnání třeba formou celogenomových alignmentů pomocí algoritmu BLAST. Byla navržena celá řada koeficientů, ze který ty používané u indexu zvaného GBDP (Genome Blast Distance Phylogeny) pro bakteriální genomy, viz. 4. přednáška. Postupuje se tak, že se provede BLAST(XY), tj. BLAST proti genomu X genomem Y jako query, a BLAST(YX), totéž v opačné gardu, tj. proti genomu Y s X jako query. BLASTy nám poskytnou sady lokálních alignmentů s nadprahovou shodou, tzv. HSPs (high scoring sequence pairs viz. přednáška 3). Nechť H_{XY} a H_{YX} jsou celkové délky HSPs pro BLAST(XY) a BLAST(YX), I_{XY} a I_{YX} jsou celkové počty identických nukleotidů v těchto HSPs, a λ (X, Y) se suma délek obou genomů. Tři navržené koeficienty pro výpočet genetických vzdáleností mezi genomy vycházející z těchto veličin jsou uvedeny níže.

$$d_0(X,Y) = 1 - \frac{H_{XY} + H_{YX}}{\lambda(X,Y)}$$

$$d_4(X,Y) = 1 - \frac{2 \cdot I_{XY}}{H_{XY} + H_{YX}}$$

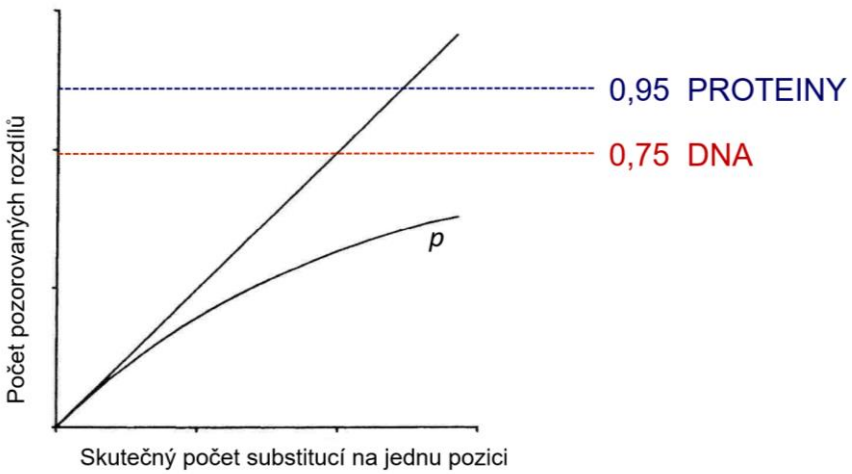
$$d_6(X,Y) = 1 - \frac{2 \cdot I_{XY}}{\lambda(X,Y)}$$

Je patrné, že některé z nich neberou v potaz množství identických párů (d_0), jiné zase vůbec neuvažují velikost genomů (d_4), d_6 zachováná nejvíce informací a je ve fylogenetickém kontextu nejvhodnější[1].

[1] https://www.ncbi.nlm.nih.gov/pubmed/23432962

Odhad frekvence substitucí

Počet substitucí, ke kterým došlo v sekvenci dvou organizmů během doby, která uplynula od jejich společného předka, je ovšem obvykle vyšší než počet rozdílů, který pozorujeme. Důvody jsou shrnuty na obrázku níže. Protože znaky v DNA mohou nabývat jen 4 různých stavů (4 nukleotidy) je nezanedbatelná pravděpodobnost, že jedna pozice v sekvenci projde více substitucemi (**vícenásobná substituce**), přitom my pozorujeme jen tu poslední nebo v případě **zpětné substituce** nepozorujeme dokonce žádnou. Stejně tak je možné, že stejná pozice prošla substitucí v obou sekvencích (**koincidence**), ale my vidíme opět jen jeden rozdíl nebo dokonce žádný pokus sekvence nakonec **konvergovaly** ke stejnému nukleotidu nebo u nich došlo **paralelně** k substituci na stejný nukleotid**.** Čím delší doba uplynula od společného předka a čím větší počet substitucí se odehrál, tím častěji (vztaženo k počtu substitucí) docházelo k podobným jevům a tím větší je rozdíl mezi počtem pozorovatelných rozdílů a počtem substitučních událostí. Tomuto jevu se říká **substituční saturace**.

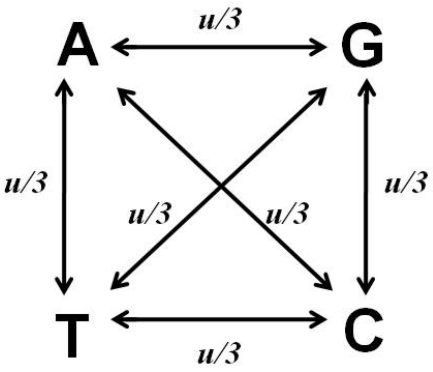


Vinou substituční saturace není samotný podíl rozdílných nukleotidů (p) vhodnou mírou genetické vzdálenosti použitelnou pro rekonstrukci fylogeneze. Problémem p je skutečnost, že podléhá saturaci, a proto tato míra není aditivní. To znamená, že vznikne-li za čas t mezi sekvencemi A a B rozdílů v nukleotidech x, za čas 2t vznikne méně než 2x rozdílů, protože p neroste s časem lineárně. Blíží-li se čas, po který sekvence divergují nekonečnu, procento rozdílných nukleotidůse bude blížit k hodnotě 0,75. Pokud bychom srovnali velké množství náhodně vytvořených dvojic zcela náhodných sekvencí, zjistili bychom, že se budou v průměru shodovat v ¼ nukleotidů. Je-li v pozici 1 sekvence A nukleotid T, je 25% pravděpodobnost, že ve zcela nepříbuzné sekvenci B je v pozici 1 také nukleotid T. Mnohem lepší míra genetické vzdálenosti než p je počet substitučních událostí, respektive počet substitučních událostí vztažený na jednu pozici alignmentu. Tato míra je aditivní a můžeme ji získat, pokud p zkorigujeme na “neviditelné” substituce. K tomu budeme potřebovat pravděpodobnostní substitučního modely, které dokáží vyjádřit vztah mezi počtem substitucí a p pomocí funkce, která vychází z reálného základu. Tj. vystihuje podstatu substitučního procesu a její parametry

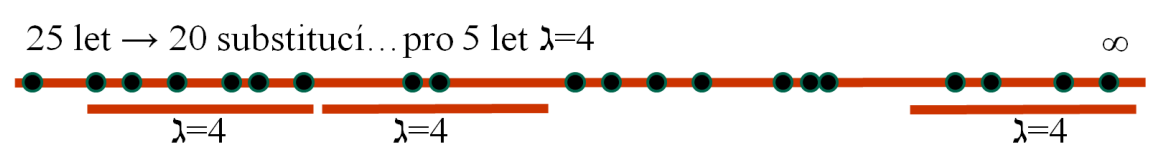
mají biologický základ. Nejjednodušším substitučním modelem, který si představíme je model **Jukes-Cantor (1961)**. Pravděpodobnostní substituční modely mají velký význam nejen pro korekci p na počet substitucí, ale také pro konstrukci stromů metodami maximum likelihood a Bayéskou metodou. Proto budeme jejich principu věnovat zvýšenou pozornost.

Substituční model **Jukes-Cantor**

Množství substitucí, ke kterým došlo v evoluci od sekvence A k sekvenci B si můžeme představit jako úsečku (větev) oddělující obě sekvence. Její délka je určována dvěma parametry u (substituční rychlost) a t (čas). **Jukes-Cantor** předpokládá, že subtituční rychlosti jsou stejné pro všechny typy záměn. Je tedy celková rychlost substituce za jiný nukleotid u, pak rychlost změny za konkrétní jeden ze tří odlišných nukleotidů je u/3.

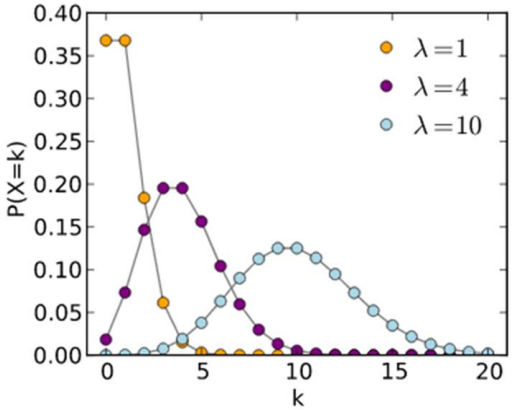


Nyní si představíme si, že kromě změn na tři jiné nukleotidy, dochází rychlostí u/3 také k substitucím na identický nukleotid. Můžeme si představit, že s frekvencí polymeráza zapomene „zkontrolovat“ nukleotid na protějším vláknu a sáhne po náhodném. Ve třech případech ze čtyř se dopustí chyby (s celkovou frekvencí u) v jednom případě sáhne náhodou po tom vhodném. Celková rychlost, s jakou dochází ke všem zaváháním polymerázy, je tedy 4/3u. My ale potřebujeme znát nikoli rychlost (neboli frekvenci) událostí, ale pravděpodobnost s jakou k nim dochází, protože pravděpodobnosti změn se odráží v pozorovaných datech. Pokud by k substitucím docházelo pravidelně, věděli bychom, že ta čas t dojde s pravděpodobností 100% ke 4/3ut substitucím. Ve skutečnosti však dochází k událostem nepravidelně. Jako příklad uvažujme situaci, kdy z dlouhodobého pozorování víme, že za 25 let došlo k 20 substitucím. Rychlost je tedy 4/5 substituce za rok. V pětiletém intervalu tedy očekáváme vznik 4 substitucí. Naše očekávání označíme λ=4. Ve skutečnosti, zdaleka ne ve všech pětiletých intervalech dojde právě ke 4 substitucím. Nalezneme intervaly s větším i menším počtem substitucí. Pravděpodobnost, s jakou dojte určitému počtu (k) nám pomůže určit Poisonovo rozdělení.



Poissonovo rozdělení bývá označováno jako *rozdělení řídkých jevů*, neboť se podle něj řídí četnosti jevů, které mají velmi malou pravděpodobnost výskytu (substituce v sekvencích, rozpady radioizotopů). Pravděpodobnost, že dojde právě ke k událostem závisí jen na očekávaném počtu λ

$$f(k,\lambda) = (\lambda^k e^{-\lambda})/k!$$



Jak je patrné z křivky pro λ=4 (fialová) je pravděpodobnost výskytu 4 substitucí v pětiletém úseku 0,2 (k=4, λ=4). Pravděpodobnost výskytu 3 substitucí je například úplně stejná (0,2). Nenulovou pravděpodobnost má i výskyt 0 substitucí nebo 10 substitucí.

Vraťme se k našemu substitučnímu procesu. Očekávaný počet substitucí za čas t, oddělující sekvenci A od sekvence B, je λ=4/3ut. Dosadíme-li to do funkce Poissonova rozdělení a spočítáme pravděpodobnost, že k žádné substituci nedojde (k=0; 0!=1) a dospějeme k výrazu

$$e^{-4/3ut}$$

pravděpodobnost, že dojde k jedné nebo více událostem je potom doplněk

$$1- e^{-4/3ut}$$

Pravděpodobnost, že dojde k události (nebo více událostem), které skončí jedním konkrétním nukleotidem ze čtyř možných, např. C, je čtvrtina tohoto výrazu

$$P(C|A) = 1/4\ (1- e^{-4/3\ ut})$$

Protože jsou 3 možnosti, jak může dojít ke změně (tři jiné nukleotidy), je pravděpodobnost, že dojde ke změně je trojnásobkem

$$Ds = 3/4\ (1- e^{-4/3\ ut})$$

Ds (v přednášce jsem to označoval p, Ds je podle mě lepší) velmi těsně souvisí s procentem rozdílných nukleotidů, které jsme označovali p. U nekonečně dlouhých sekvencí se pravděpodobnost, se kterou dojde ke změně na větvi oddělující sekvence, přesně rovná procentu nukleotidů, ve kterých se sekvence liší. Pravděpodobnost ⅓ přece znamená, že ke změně dojde u jednoho nukleotidu ze tří. V případě sekvencí s konečnou délkou je procento rozdílných nukleotidů aproximací pravděpodobnosti Ds, která se vinou omezeného souboru pozic v sekvencích může od pravděpodobnosti lišit. Můžeme ji však použít a dosadit do vzorce.

Člen exponentu ut odpovídá délce větve oddělující obě sekvence (neboli počtu substitucí, ke kterým došlo) a je tedy kýženou genetickou distancí korigovanou na "neviditelné substituce", kterou jsme chtěli získat. Jednoduchou úpravou rovnice získáme

$$D = ut = -3/4 \ln(1- 4/3\ p)$$

Protože procento rozdílných nukleotidů, které jsem dosadili za p je u konečných sekvencí pouhým odhadem pravděpodobnosti Ds, nezískali jsme přesnou hodnotu D, ale její odhad, který má rozptyl

$$V(D) = (p(1 -p))/(L(1 - 4/3\ p)2)$$

kde L je délka sekvencí.

Všimněte si, že jsme v naší úvaze rezignovali na to určit, jak k D (délce větve) přispívá délka časového intervalu a jak substituční rychlost. Pár sekvencí substituující rychlostí 2 po dobu 0,5 bude oddělený stejně dlouhou větví (D), jako sekvence substituující rychlostí 1 po dobu 1, a my z délky větve samotné nemůžeme poznat o jaký případ jde. Jak uvidíme v následujících přednáškách, přináší to sebou několik nevýhod pro rekonstrukci fylogeneze na základě této a většiny ostatních genetických distancí, které stejně rezignovaly - čas v různých místech stromu neběží stejně rychle, konce větví na stromu označující současné taxony nedosahují stejně daleko, nevíme kde je kořen stromu. Pokud ovšem neznáme substituční rychlost u nebo čas t, musíme se s tímto spokojit.

Příklad z prezentace:

Sekvence A a B se liší ve 3 nukleotidech ze 14.

p=3/14=0,2148

D= -3/4 ln(1- 4/3 * 0,2148)
D= 0,246

D je vyšší než *p* o množství očekávaných “neviditelných” substitucí.

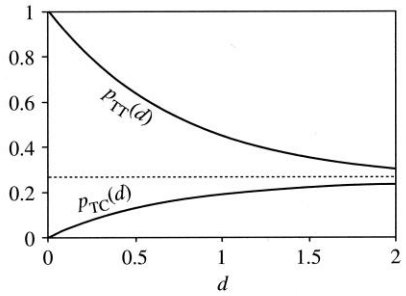
Nyní si ukážeme obecnou metodu, jak odvodit Jukes Cantorův pravděpodobnostní model. Při tomto postupu nejprve vytvoříme matici (Q) substitučních rychlostí pro všechny typy záměn. Substituční rychlosti budou opět u/3 pro všechny typy záměn. Na diagonálu doplníme členy *-u* proto, aby součet řádků matice byl 0[2]. Z rychlostní matice Q můžeme získat pravděpodobnostní matici P(t) umocněním (opět vycházíme z Poissonova rozdělení). Mocnění matic je složitější matematická operace a tvar členů matice *p0(t)* a *p1(t)* je uveden vpravo.

$$Q = \{q_{ij}\} = \begin{array}{c} \begin{matrix} \text{A} & \text{G} & \text{C} & \text{T} \end{matrix} \\ \begin{bmatrix} -u & u/3 & u/3 & u/3 \\ u/3 & -u & u/3 & u/3 \\ u/3 & u/3 & -u & u/3 \\ u/3 & u/3 & u/3 & -u \end{bmatrix} \end{array} \begin{matrix} \text{A} \\ \text{G} \\ \text{C} \\ \text{T} \end{matrix}$$

$$P(t) = \mathrm{e}^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_1(t) & p_1(t) \\ p_1(t) & p_0(t) & p_1(t) & p_1(t) \\ p_1(t) & p_1(t) & p_0(t) & p_1(t) \\ p_1(t) & p_1(t) & p_1(t) & p_0(t) \end{bmatrix} \quad \text{with} \begin{cases} p_0(t) = \frac{1}{4} + \frac{3}{4}\mathrm{e}^{-4/3\,ut} \\ p_1(t) = \frac{1}{4} - \frac{1}{4}\mathrm{e}^{-4/3\,ut} \end{cases}$$

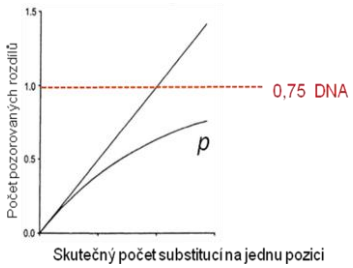
Můžete si všimnout, že výraz mají stejný tvar k jakému jsme dospěli při prvním odvozování pravděpodobnostního modelu Jukes Cantora.

Model Jukes Cantora má několik pozoruhodných vlastností, které nás ubezpečují že jdeme správnou cestou. Součet členů matice P(t) v každém řádku a sloupci je vždy roven 1. To odpovídá skutečnosti, že v pozici DNA sekvence je s pravděpodobní 1 přítomen jeden ze čtyř nukleotidů a ten se při substituční události s pravděpodobností 1 buď změní nebo ne. Dále si všimněme, že pokud necháme sekvenci mutovat nekonečně dlouho (ut=∞), pak pravděpodobnost že nukleotid zůstane sám sebou P_{TT} bude 1/4 a bude stejná jako pravděpodobnost, že se změní na jiný konkrétní nukleotid (P_{TC}, P_{TA} i P_{TG} budou 1/4). Jinými slovy, pokud necháme sekvenci mutovat nekonečně dlouhou, vznikne nám náhodná sekvence složená ze 4 nukleotidů o frekvencích 1/4. Hovoříme o stacionárním rozložení.

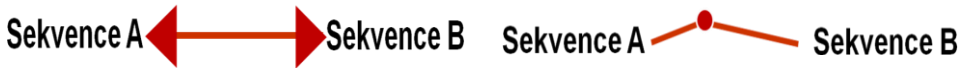


[2] Proč by měl být součet členů v řádcích a sloupcích a vlastně i v celé matici roven nule. Hodnoty mimo diagonálu udávají rychlost, s jakou daný nukleotid vzniká z ostatních nukleotidů (první sloupec udává rychlost vzniku A). Diagonální členy udávají rychlost, s jakou daný nukleotid zaniká, tj. mění se na jiný. Pokud je suma sloupce nula znamená to, že nukleotidy stejně rychle vznikají i zanikají a tedy, že jich v sekvenci zůstává pořád stejný počet. Pokud by Suma všech členů matice byla vyšší než jedna, model by předpokládal, že se nám sekvence bude prodlužovat.

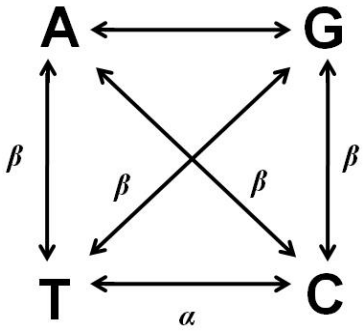
To odpovídá naší představě, že díky substituční saturaci se dvě náhodné DNA sekvence se v průměru podobají v 1/4 nukleotidů a nikoli v 0 nukleotidech.



Všimněme si také, že model je symetrický přes diagonálu. Pravděpodobnost záměny tam a zpět je stejná. Symetrické modely mají tu vlastnost, že poskytují stejné výsledky nezávisle na tom kterým směrem evoluce ve skutečnosti šla. Nezáleží na tom, jestli sekvence A byla předkem sekvence B nebo naopak, či zda jsou obě potomkem společného předka ležícího kdekoli na jejich spojnici. Výsledek výpočtu to neovlivní. Nevýhodou symetrických modelů je, že poskytují nezakořeněné dendrogramy.



Pro svoji jednoduchost Jukes-Cantorův model není příliš realistický, opomíjí například skutečnost, že rychlosti různých typů záměn jsou různé a že v některých sekvencích se vyskytují některé báze častěji než jiné (jsou třeba GC bohatší), a proto se v takových sekvencích zvyšuje pravděpodobnost změny na báze, které jsou tam frekventované. První problém se snaží řešit o něco komplikovanější **Kimurův 2-parametrový model.**



Tento model předpokládá, že rychlost transverzí (substitucí purin za purin nebo pyrimidin za pyrimidin) je jiná než rychlost tranzic (substitucí purin za pyrimidin), označuje je alfa a beta. Rychlostní matice Q vypadá následovně

$$Q = \begin{bmatrix} -(\alpha+2\beta) & \alpha & \beta & \beta \\ \alpha & -(\alpha+2\beta) & \beta & \beta \\ \beta & \beta & -(\alpha+2\beta) & \alpha \\ \beta & \beta & \alpha & -(\alpha+2\beta) \end{bmatrix}$$

matice pravděpodobností záměn pak

$$P(t) = \mathrm{e}^{Qt} = \begin{bmatrix} p_0(t) & p_1(t) & p_2(t) & p_2(t) \\ p_1(t) & p_0(t) & p_2(t) & p_2(t) \\ p_2(t) & p_2(t) & p_0(t) & p_1(t) \\ p_2(t) & p_2(t) & p_1(t) & p_0(t) \end{bmatrix}$$

a členy této matice mají tvary

$$p_0(t) = \tfrac{1}{4} + \tfrac{1}{4}\mathrm{e}^{-4\beta t} + \tfrac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t}$$

$$p_1(t) = \tfrac{1}{4} + \tfrac{1}{4}\mathrm{e}^{-4\beta t} - \tfrac{1}{2}\mathrm{e}^{-2(\alpha+\beta)t}$$

$$p_2(t) = \tfrac{1}{4} - \tfrac{1}{4}\mathrm{e}^{-4\beta t}$$

Z toho lze o Distance se vypočítá následovně

D = 0,5 ln(a) + 1/4 ln(b)
a = 1/(1 - 2P - Q) b = 1/(1 -2Q)

Pro výpočet distance tímto modelem potřebujeme znát podíl tranzic (P) a transverzí (Q). I tato distance (jako každá jiná) má rozptyl, který se v tomto případě rovná

V(D) = [a2P + c2Q -(aP +cQ)2]/L
c = (a + b)/2, L=délka sekvence

Příklad z prezentace:

Sekvence A a B se liší ve 2 tranzicích a jedné transverzi.

P=2/14
Q=1/14

a = 1/(1 – 2*2/14 – 1/14) = 1,54
b = 1/(1 -2*1/14) = 1,16
D = 0,5 ln(1,54) + 1/4ln(1,16)=0,254

D je vyšší než p i D podle Jukes-Cantora, protože K2P model je realističtější a umožňuje odhalit více “neviditelných” substitucí.

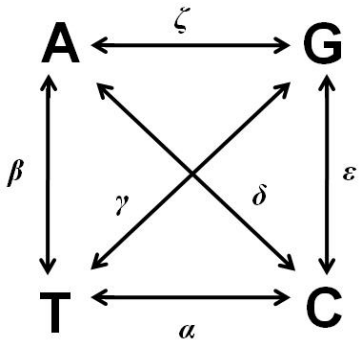
I Kimurův 2-parametrový model značně zjednodušuje, a proto bylo vyvinuto několik dalších modelů, které se snaží více přiblížit průběhu substitučních procesů v sekvencích DNA.

Dalším krokem je zohlednění skutečnosti, že frekvence nukleotidů v reálných sekvencích není 1/4. Genomy se přece liší obsahem GC. Je tedy nerealistické, aby stacionární rozložení nukleotidů bylo 1/4, 1/4, 1/4 a 1/4 jak předpokládá Jukes Cantorův a Kimura 2P model. Zohledněné této skutečnosti se provádí tak, že se členy v rychlostní matici násobí parametry π_A, π_C, π_T, π_G, které představují očekávané stacionární rozložení nukleotidových frekvencí a jejich odhad získáme tak, že si spočítáme frekvenci jednotlivých nukleotidů v našem alignmentu.

$$Q = \begin{bmatrix} -(\alpha_1\pi_C + \beta\pi_R) & \alpha_1\pi_C & \beta\pi_A & \beta\pi_G \\ \alpha_1\pi_T & -(\alpha_1\pi_T + \beta\pi_R) & \beta\pi_A & \beta\pi_G \\ \beta\pi_T & \beta\pi_C & -(\alpha_2\pi_G + \beta\pi_Y) & \alpha_2\pi_G \\ \beta\pi_T & \beta\pi_C & \alpha_2\pi_A & -(\alpha_2\pi_A + \beta\pi_Y) \end{bmatrix}$$

Tato konkrétní matice náleží modelu F84. Koeficinty π způsobí, že pravděpodobnosti změn na vzácnější nukleotidy budou nižší a nekonečně dlouhou substituující sekvence bude mít frekvence nukleotidů π_A, π_C, π_T, π_G.

Pomyslným vrcholem je **General time reversible model (GTR)**



Tento model umožňuje přidělit všem typům záměn jinou substituční rychlost. Zároveň umožňuje také zohlednit to, jak často se ve zkoumaných sekvencích jednotlivé nukleotidy vyskytují a tedy jak ochotně v daných sekvencích dochází k substituci na jednotlivé konkrétní nukleotidy. Substituční rychlost z A na C se u tohoto modelu skládá jednak z rychlosti této záměny (δ) a také zase z “ochoty” použít jako nový nukleotid právě C (π_c). Odhadem této ochoty je frekvence C v sekvenci. Konkrétní hodnoty veškerých parametrů (rychlosti α,β,γ,δ,ε,ζ a frekvence π_A, π_C, π_G, π_T) předem neznáme a odhadujeme je z analyzovaných sekvencí. Za odhady π_A, π_C, π_G, π_T

považujeme frekvence nukleotidů v sekvencích. Rychlost substitucí odvozujeme z pozorovaných záměn. Pro dvojici sekvencí si zapíšeme pro všechny možné kombinace nukleotidů kolikrát se v sekvenci A vyskytoval nukleotid X a v sekvenci B nukleotid Y.

Sekvence A

Sekvence B	A	C	G	T
A	224	5	24	8
C	3	149	1	16
G	24	5	230	4
T	5	19	8	175

Protože GTR je “time reversible” tedy předpokládá stejné rychlosti substitucí a zpětných substitucí (G na A jako A na G), musíme tabulku nejprve zesymetrizovat (počet G na A musí být stejný jako A na G) průměrováním hodnot. Potom tabulku znormalizujeme, aby hodnoty v sloupcích dávaly součet 1 (hodnoty v buňkách vydělíme sumou sloupce). Dalšími několika úpravami, například logaritmováním, této tabulky můžeme dospět k odhadu rychlostí α,β,γ,δ,ε,ζ.

Odhady hodnot parametrů vnášejí do odhadu chyby, takže rozptyl hodnot GTR je vyšší než u K2P nebo Jukes-Cantorovy distance a jako vždy vzrůstá s klesající délkou sekvence.

Z grafu na snímku 25 vyplývá, že čím komplikovanější model použijeme, tj. čím více parametrů mu uvolníme, tím více se přiblížíme k pravděpodobnostem skutečných substitučních dějů a tím přesněji, za předpokladu že dobře odhadneme hodnoty našich parametrů, jsme schopni odhadnout počet “neviditelných” substitucí, tedy potlačit substituční saturaci. Křivky se napřimují, protože odhad počtu substitucí se u komplikovanějších modelů přibližuje skutečnému počtu substitucí. Model GTR v grafu zahrnut není.

Poslední genetická vzdálenost, která není založena na substitučním modelu, ale přesto se jí daří poměrně dobře bojovat jak se substituční saturací, tak s vlivem nerovnoměrného zastoupení nukleotidů v jednotlivých sekvencím, který může analýzy ovlivnit, je **LogDet distance**. Stejně jako odhadu rychlostí u GTR si nejprve vyplníme tabulku

Sekvence A

Sekvence B	A	C	G	T
A	224	5	24	8
C	3	149	1	16
G	24	5	230	4
T	5	19	8	175

Tuto tabulku normalizujeme tak, aby součet všech buněk činil 1. Hodnoty buněk vydělíme sumou všech hodnot. Tabulku potom budeme považovat za matici čísel a určíme její determinant. Záporný logaritmus tohoto determinantu je LogDet distance.