

Domáca úloha č. 2 pre študentov prírodovedných zameraní
2-AIN-501 Metódy v bioinformatike
Termín: streda 3.12.2024 22:00
Odrovzdajte pdf cez Moodle

1. Fylogenetické stromy. V tejto úlohe zostrojíme fylogenetický strom niekoľkých druhov mušiek *Drosophila* (*D. melanogaster*, *D. simulans*, *D. erecta*, *D. ananassae*, *D. persimilis*, *D. willistoni*, *D. virilis*, *D. mojavensis*, *D. grimshawi*), pričom ako outgroup (vonkajšiu skupinu) použijeme komára *A. gambiae*.

- a) V dvoch publikáciách uvedených nižšie (obidve sú voľne dostupné) sa nachádza strom týchto druhov, ale tieto stromy sa mierne líšia. Vyberte z publikovaných stromov len vyššie uvedené druhy a nakreslite ich do vašej úlohy (netreba dodržať dĺžky vetiev). Do obidvoch stromov tiež doplňte aj komára.

Obr.1 v článku Stark et al. 2007, Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures, Nature

Obr.4 v článku Obbard et al. 2012, Estimating Divergence Dates and Substitution Rates in the Drosophila Phylogeny, Mol Biol Evol

- b) Odobratím jednej vetvy v strome sa nám strom vždy rozpadne na dve skupiny, ktoré nie sú spojené. Napríklad v oboch stromoch sa nachádza vetva, ktorá oddelí *D.mel.*, *D.sim.* a *D.ere.* od ostatných druhov. Tieto dve vetvy teda môžeme považovať za zhodné medzi oboma stromami. Na cvičeniach sme podobne porovnávali vetvy stromov pri výpočte bootstrapu.

Ktoré vetvy sa medzi týmto dvomi stromami nezhodujú? Vyznačte ich graficky (napr. inou farbou) v obrázku v časti a).

- c) V UCSC browseri nájdite genóm pre *D. melanogaster*, verzia genómu dm6 a v nej nájdite gén hd (humpty dumpty) (zadajte toto meno ako search term). Na ktorom chromozóme sa nachádza?
- d) K tomuto génu získajte zarovnanie zodpovedajúcich kódujúcich sekvencií a proteínov nasledujúcim spôsobom:

Na hornej modrej lište browsera zvolíme Tools->Table browser. Na ďalšej obrazovke nastavíme track: NCBI RefSeq, output format: CDS FASTA alignment a ako output file: vami zvolené meno súboru, kam sa vám uloží výsledok. Navyše nechceme vypísať všetky gény, ale len ten nás a aj preň iba jednu formu zostrihu, takže stlačíme ešte filter: create a na novej obrazovke v riadku name does match zadáme identifikátor génu NM_141274.3. Potom už zostáva len stlačiť tlačidlá submit a get output. Na ďalšej obrazovke si zvolíme, ktoré organizmy chceme v zarovnaní (ako uvedené vyššie) a či chceme nukleotidové alebo proteínové sekvencie (vyrobte si obe verzie zarovanania).

Výsledné súbory si otvorte v editore a zmenťte mená sekvencií tak, aby začínali skratkou mena príslušného organizmu (zmažte text NM_141274.3_ z každého mena). V opačnom prípade by vám fylogenetické programy pokrátli mená tak, že nebudeť vedieť, ktorý list je ktorý.

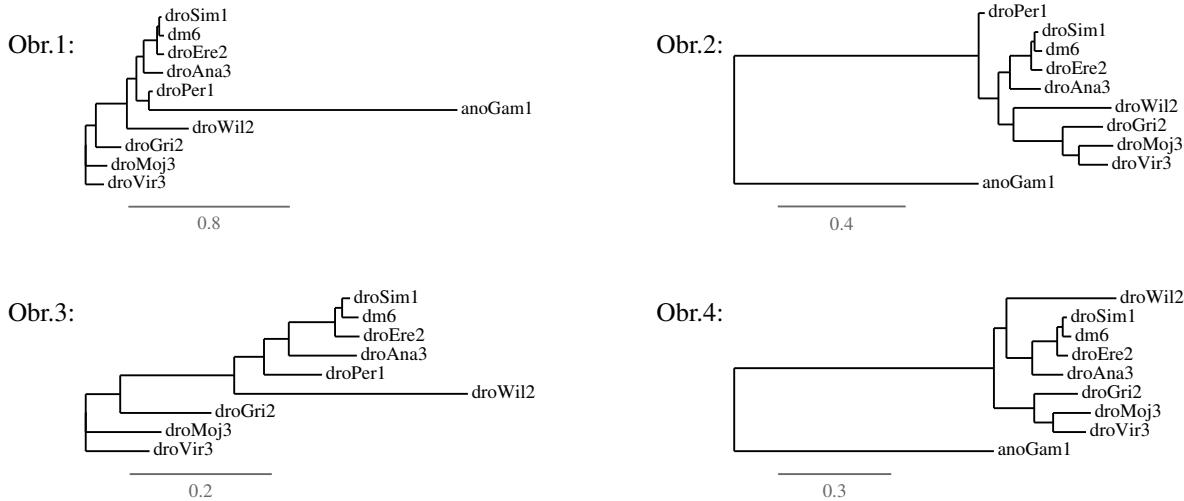
Pre obidve zarovnania zistite, koľko percent zachovaných báz resp. aminokyselín majú jednotlivé druhy s *D. melanogaster*. Môžete to zistiť napr. programom mvview, stĺpec pid. Alternatíva je program Dist.seqs na Galaxy, ktorý spočíta celú maticu vzdialenosťí, z ktorej môžete zobrať príslušné položky.

- e) Pre obidve zarovnania zostavte strom nejakým dostupným nástrojom. Pokiaľ možno použite metódu spájania susedov (neighbour joining), ktorá sa zvykne označovať aj skratkou NJ, prípadne menom programu BioNJ/RapidNJ. Vo vašej úlohe uved'te, aký program ste použili, aké ste zvolili nastavenia alebo či už boli nejaké predvolené. Ďalej v úlohe uved'te stromy. V prípade, že vám program vykreslí strom nesprávne zakorenéný (komár nebude outgroup), uved'te v úlohe pôvodné zakorenenie z programu, ale aj správne zakorenenie (podobne ako na obr.1 a 2 v časti f nižšie). Správne zakorenenie môžete bud' nakresliť ručne, ale programy na kreslenie stromov často umožňujú používateľovi zvoliť outgroup a prekreslia strom.

Príklady dostupných programov:

https://www.ebi.ac.uk/jdispatcher/phylogeny/simple_phylogeny (webserver na spájanie susedov)

https://usegalaxy.eu/?tool_id=toolshed.g2.bx.psu.edu%2Frepos%2Fiuc%2Frapidnj%2Frapidnj%2F2.3.2&version=latest (nástroj RapidNJ v Galaxy)



<https://phylotree.hyphy.org/> (zobrazovač stromov, umožňuje zmeniť zakorenenie)

<https://www.megasoftware.net/> (sofver nainštalovateľný na váš počítač)

- f) Proteínové zarovnanie z časti d) sme použili v programe PhyML, ktorý zostaví strom metódou maximálnej vieročnosti. Dostali sme strom na obr.1, ktorý sme potom správne zakorenili na obr.2.

Potom sme ten istý program spustili na zmenených verziach zarovnania, z ktorého sme vynechali komára resp. *D.per*. Výsledné stromy sú uvedené na obrázkoch 3 a 4.

Strom na obr.3 nevieme správne zakoreníť, lebo neobsahuje vonkajšiu skupinu. Ak by sme zo stromu na obr.2 vyhodili komára a porovnávali vetvy so stromom na obr.3 (podobne ako ako v časti b), budú sa niektoré vetvy lísiť? (Pri tomto porovnaní obidva stromy považujeme za nezakorenene).

- g) Po vynechaní *D.per* sme na obr.4 dostali inú polohu komára. Porovnajte tento strom s publikovanými stromami z časti a). Na ktorý sa viac podobá? Líšia sa niektoré vetvy?
- h) Za max. 2 bonusové body môžete skúsiť zostaviť strom aj iným spôsobom (napr. iné gény, iný program a pod.). Popíšte váš postup, uveďte výsledky a okomentujte ich.

2. Evolučné modely. V tomto príklade budeme uvažovať model substitúcií Kimura 1980, v ktorom sa tranzície dejú rýchlosťou α a transverzie rýchlosťou β , ale všetky bázy sa vyskytujú s rovnakou frekvenciou.

V tomto modeli sa pravdepodobnosť, že báza X sa zmení na bázu Y za čas t , liší podľa toho, či ide o tranzíciu (pravdepodobnosť označíme ako u_t), transverziu (označíme ako s_t), alebo nezmenenú bázu (označíme ako r_t):

$$\begin{aligned} u_t &= \frac{1}{4}(1 + e^{-4\beta t} - 2e^{-2(\alpha+\beta)t}) && \text{(tranzícia, napr. } \Pr(A|G,t)\text{)} \\ s_t &= \frac{1}{4}(1 - e^{-4\beta t}) && \text{(transverzia, napr. } \Pr(C|G,t)\text{)} \\ r_t &= 1 - 2s_t - u_t && \text{(žiadna zmena, napr. } \Pr(G|G,t)\text{)} \end{aligned}$$

Ak máme zarovnanie s n_u tranzíciami, n_s transverziami a n_r nezmenenými bázami, jeho pravdepodobnosť v tomto modeli bude

$$0.25^{n_u+n_s+n_r} \cdot u_t^{n_u} \cdot s_t^{n_s} \cdot r_t^{n_r}.$$

Nakoľko takáto pravdepodobnosť je väčšinou číslo veľmi blízke 0, používame namiesto toho jej logaritmus vypočítaný pomocou vzorca

$$(n_u + n_s + n_r) \cdot \log(0.25) + n_u \cdot \log(u_t) + n_s \cdot \log(s_t) + n_r \cdot \log(r_t).$$

Ak na nukleotidové zarovnanie z úlohy 1 spustíme program PhyML s modelom Kimura 1980, spočíta, že najvieročnejšia hodnota pomeru α/β je 3.458. My teda nastavíme $\alpha = 3.458$ a $\beta = 1$. Ak v tomto zarovnaní

porovnáme *D. melanogaster* a *D. ananassae*, vidíme $n_u = 218$ tranzícíí, $n_s = 171$ transverzií a $n_r = 1315$ báz bez zmeny.

- a) V Exceli alebo inom tabuľkovom procesore (LibreOffice a pod.) vytvorte tabuľku, v ktorej využijete uvedené vzorce na výpočet rôznych hodnôt Kimura 80 modelu pre vyššie uvedené α , β a pre hodnoty t od 0.01 po 0.3 s krokom 0.01.

V stĺpcu A si vytvorte hodnoty t a v ďalších troch stĺpcach hodnoty u_t , s_t a r_t vzorcami uvedenými vyššie. Exponenciálna funkcia e^x sa v Exceli napíše ako =exp(x) a desiatkový logaritmus ako =log(x).

V stĺpci E spočítajte podiel očakávaného počtu tranzícíí spomedzi všetkých zmenených báz po čase t , ktorý dostaneme ako $u_t/(u_t + 2s_t)$, lebo daná zmenená báza sa bud' zmenila tranzíciovou pravdepodobnosťou u_t alebo jednou z dvoch možných tranverzií, pričom každá má pravdepodobnosť s_t , a teda celková pravdepodobnosť zmeny je $(u_t + 2s_t)$.

Napokon v stĺpci F spočítajte logaritmus pravdepodobnosti nášho zarovnania s vyššie uvedenými hodnotami n_u , n_s a n_r , pričom použijete už známe hodnoty u_t , s_t a r_t zo stĺpcov B-D.

Prvé riadky vašej tabuľky by mali vyzerat' nejako takto:

	A	B	C	D	E	F
1	t	u_t	s_t	r_t	podiel tranzícíí	log pravdepodobnosť
2	0.01	0.0328	0.0098	0.95	0.63	-1723.56
3	0.02	0.0624	0.0192	0.90	0.62	-1642.70

Do domácej úlohy vypíšte hodnoty z tabuľky pre $t = 0.1$ a $t = 0.3$.

- b) Pre ktorú hodnotu t má naše zarovnanie najväčšiu pravdepodobnosť? (Toto t by mohlo byť odhadom vzdielenosti týchto dvoch organizmov) Vykreslite priebeh tejto pravdepodobnosti ako graf (os x je t, os y logaritmus pravdepodobnosti).
- c) V našom zarovnaní máme podiel tranzícíí medzi všetkými zmenami $218/(218 + 171) = 0.56$. Aký podiel tranzícíí predpovedá model pre hodnotu t zistenú v časti b)? Sú tieto hodnoty podobné?
- Pridajte do tabuľky riadky pre $t = 10$ a $t = 100$ a uvedťte, aké hodnoty ste dostali pre u_t , s_t , r_t a podiel tranzícíí. Prečo intuitívne dostávame pre veľké t takéto hodnoty?