III

MOLEKULAARSE EVOLUTSIOONI VALITUD PROBLEEME

3.1 Fisheri loodusliku valiku teoreem

3.1.1 Kohastumus (kohasus, kohanemus) evolutsiooni tegurina

Charles Darwin („The origin of species", 1872) näitas, et looduse mitmekesisust ja otstarbekust saab seletada nn loodusliku valikuga, mis soodustab konkreetse keskkonna suhtes paremini kohastunud indiviidide paljunemist. Ronald Fisher jt konstrueerisid sellele nähtusele („teooriale") matemaatilise esituse (R. Fisher, „The Genetical Theory of Natural Selection", 2nd ed., Oxford, 1958).

Järgnevalt vaatame lähemalt üht Fisheri loomingu näidet. Tähistame alleelide a_1, a_2, …, a_k jaotuse $\mathbf{p} = \mathbf{p}(t)$ populatsioonis ajahetkel t vektoriga

$$(p_1, p_2, \ldots, p_k)^{\mathrm{T}} = \mathbf{p} .$$

Olgu nende alleelide oodatav arv, mis genotüübiga $a_i a_j$ indiviidi järglastes elama jäävad, tähistatud $2w_{ij}$. Suurust w_{ij} nimetatakse genotüübi $a_i a_j$ **kohastumuseks** (*fitness*)[1]. Kõigi genotüüpide komplekti tarvis kirjeldab kohastumusi nn kohastumusmaatriks **W**:

$$\mathbf{W} = \begin{pmatrix} w_{11} & \cdots & w_{1k} \\ \vdots & \ddots & \vdots \\ w_{k1} & \cdots & w_{kk} \end{pmatrix},$$

kus $\mathbf{W} = \mathbf{W}^{\mathrm{T}}$.

Keskmine kohastumus populatsioonis on Hardy-Weinbergi tasakaaluseaduse kehtides

$$\bar{w} = \sum_{i,\,j=1}^{k} (w_{ij} \times p_i p_j) = \sum_{i=1}^{k} (w_{ii} \times p_i^2) + \sum_{i<j} (w_{ij} \times 2 p_i p_j) .$$

Edasi teeme loomuliku eelduse, et iga üksiku genotüübi kohastumus ajas ei muutu, st $w_{ij}(t) = w_{ij}(t+1) = w_{ij}$. Küll aga muutuvad valiku toimel genotüübisagedused ja seeläbi ka populatsiooni keskmine kohastumus.

Teoreem (looduslikust valikust). Hardy-Weinbergi tasakaalus populatsiooni keskmine kohastumus $\bar{w}$ ei ole järglaspõlvkonnas väiksem kui eellastel.

Tõestus. Olgu $N(t)$ indiviidide arv generatsioonis t. Näitame esmalt, et

$$\bar{w} = N(t+1)/N(t) . \tag{3.1}$$

Olgu n_i alleeli a_i kordsus (arv) populatsioonis ja p_i, nagu eelnevalt tähistatud, alleeli a_i tõenäosus populatsioonis hetkel t. Siis on Hardy-Weinbergi tasakaaluseaduse kehtides populatsioonis hetkel t $N(t)p_1^2$ indiviidi genotüübiga $a_1 a_1$, ja igaüks neist jätab järglasele keskmiselt $2w_{11}$ alleeli. Kõik need

[1] Tänapäevases eestikeelses evolutsiooniteooriaalases kirjanduses on ingliskeelne mõiste *fitness* tõlgitud kui **kohasus** (või **kohanemus** – Richard Villems, http://www.ebc.ee/tymri00/loengud/evol/) ja tähendab see pigem suhtelist kohastumust. Näiteks:

Mart Viikmaa. Klassikalise geneetika leksikon. http://biomedicum.ut.ee/~martv/genolex.html
Kohasus (*fitness*) – e valikuväärtus (*selective value*) e adaptiivväärtus (*adaptive value*), mingi genotüübiga isendite suhteline sigimisedukus võrreldes antud populatsiooni teiste genotüüpidega. Kohasust tähistatakse populatsioonigeneetikas sümboliga *w*. Iga genotüübi kohasuse määramisel hinnatakse selle sigimisedukust kõige võimekama suhtes, kusjuures kriteeriumiks on sigimisvõimeliste järglaste arv. Maksimaalse sigimisedukuse korral on kohasuse väärtus 1 ($w = 1$), geneetilise letaalsuse korral (puuduvad sigivõimelised järglased) on kohasus null ($w = 0$). Kohasus sõltub eluvõimest ja viljakusest. Kohasuse määranguid kasutatakse loodusliku valiku intensiivsuse hindamisel eri genotüüpide suhtes (tavaliselt üksikute lookuste kaupa) populatsiooni geneetilise dünaamika e mikroevolutsiooni uurimisel. Kohasuse vastandväärtuseks on selektsioonikoefitsient (*s*), millega hinnatakse mingi genotüübi suhtelise eliminatsiooni intensiivsust. Kohasus ja selektsioonikoefitsient on täiendsuurused ($w + s = 1$; $w = 1 - s$; $s = 1 - w$), nii et populatsioonigeneetilistes võrrandites võib kasutada üht teise asemel.

Toomas Tammaru. Evolutsiooniline ökoloogia. BGZH 03.006. http://uuslepo.it.da.ut.ee/~tammarut/evol.htm
Kohasuse mõistet võib vaadelda loodusliku valiku "pöördmõistena", st väljendeid „looduslik valik soosib alleeli *A*" ja „alleelil *A* on kõrgem kohasus" võib vaadelda sünonüümsetena. Kohasust võib seega defineerida kvalitatiivsel tasemel kui (suhtelist) paljunemisedukust ehk paljuneja panust järglaspõlvkondadesse. Kohasuse mõiste on põhimõtteliselt rakendatav alleelile, genotüübile, fenotüübile ja isendile. Kõrgema kohasusega („kohasem") alleel on see, millel on rohkem koopiaid populatsiooni järglaspõlvkondade genofondis.

alleelid on muidugi a_1, sest teisi alleele ei ole vastavatel indiviididel anda. Seega jätavad a_1a_1-indiviidid järglaspõlvkonda $N(t)p_1^2 2w_{11}$ a_1 alleeli. Analoogselt on lähtepopulatsioonis oodatavalt $N(t)\times 2p_1p_2$ indiviidi genotüübiga a_1a_2 ja need jätavad järglaspõlvkonda $2w_{12}$ alleeli, millest pooled on a_1. Sedasi edasi arutledes saame võrdused alleelide arvude tarvis generatsioonis $t+1$:

$$\begin{cases} n_1(t+1) = N(t)\left[p_1^2 2w_{11} + 2p_1p_2\frac{2w_{12}}{2} + \ldots + 2p_1p_k\frac{2w_{1k}}{2} \right], \\ \quad \cdots, \\ n_k(t+1) = N(t)\left[2p_kp_1\frac{2w_{k1}}{2} + 2p_kp_2\frac{2w_{k2}}{2} + \ldots + p_k^2 2w_{kk} \right]. \end{cases}$$

Liidame kirjutatud avaldised ja arvestame, et iga indiviid kannab kahte alleeli, saame

$$\sum_{i=1}^k n_i(t+1) = 2N(t+1) = 2N(t)\sum_{i,j=1}^k p_ip_jw_{ij} = 2N(t)\bar{w}.$$

Mis tõestabki võrduse (3.1).

Saadud võrdusi kasutades leiame $p_i(t+1)$:

$$p_i(t+1) = \frac{n_i(t+1)}{2N(t+1)} = \frac{n_i(t+1)}{2N(t)\bar{w}} = \frac{2N(t)\sum_{j=1}^k p_ip_jw_{ij}}{2N(t)\bar{w}} = \frac{p_i(t)}{\bar{w}}\sum_{j=1}^k p_jw_{ij}. \tag{3.2}$$

Tähistades

$$w_{i\cdot} = \sum_j p_jw_{ij}$$

($w_{i\cdot}$ on alleeli a_i keskmine kohastumus), võime kirjutada valemi (3.2) kujul

$$p_i(t+1) = \frac{1}{\bar{w}}p_i(t)w_{i\cdot}.$$

Kuna ilmselt

$$p_i(t) = \frac{1}{\bar{w}}p_i(t)\bar{w},$$

saame

$$\Delta p_i = p_i(t+1) - p_i(t) = \frac{1}{\bar{w}}p_i(t)[w_{i\cdot} - \bar{w}] \tag{3.3}$$

(alleeli tõenäosuse muutus on võrdeline alleeli keskmise kohastumuse ja populatsiooni keskmise kohastumuse erinevusega).

Teoreemi tõestuse lõpetamiseks näitame, et $\bar{w}$ võib ainult suureneda:

$$\Delta\bar{w} = \bar{w}(t+1) - \bar{w}(t) \geq 0.$$

Kasutades diferentseerimise[2] reegleid, saame

$$\begin{aligned} \Delta\bar{w} &= \Delta\sum_{i,j} w_{ij}p_ip_j = \sum_{i,j} w_{ij}[p_i\Delta p_j + p_j\Delta p_i] \\ &= 2\sum_{i,j} w_{ij}p_i\Delta p_j = 2\sum_j \Delta p_j\sum_i w_{ij}p_i = 2\sum_j \Delta p_jw_{j\cdot}. \end{aligned}$$

Asendades siia Δp_i valemist (3.3) saame

$$\Delta\bar{w} = 2\sum_j w_{j\cdot}\frac{1}{\bar{w}}p_j(w_{j\cdot} - \bar{w}) = \frac{2}{\bar{w}}\sum_j w_{j\cdot}(w_{j\cdot} - \bar{w})p_j = \frac{2}{\bar{w}}\sum_j (w_{j\cdot} - \bar{w})(w_{j\cdot} - \bar{w})p_j,$$

sest

$$\sum_j \bar{w}(w_{j\cdot} - \bar{w})p_j = \bar{w}\underbrace{\sum_j w_{j\cdot}p_j}_{\bar{w}} - \bar{w}\underbrace{\sum_j p_j}_{1} = 0.$$

Oleme saanud võrduse

$$\Delta\bar{w} = \frac{2}{\bar{w}}\sum_j (w_{j\cdot} - \bar{w})^2 p_j.$$

Kuna $\bar{w}$ koosneb mittenegatiivsetest liidetavatest, järeldub siit teoreemi väide:

$$\Delta\bar{w} \geq 0.$$

■

On loomulik defineerida alleelide kohastumuse dispersioon σ_w^2:

$$\sigma_w^2 = \sum_j (w_{j\cdot} - \bar{w})^2 p_j.$$

[2] Diferentseerimisoperaator Δ seab funktsioonile f vastavusse tema diferentsi $\Delta f(x) = f(x+1) - f(x)$.

Seda kasutades võib Fisheri teoreemi väljendada ka võrdusega

$$\Delta\bar{w} = \frac{2}{\bar{w}}\sigma_w^2 .$$

3.1.2 Fisheri loodusliku valiku teoreem valiku teooria mudelina

Võtame vaatluse alla dialleelse lookuse alleelidega *A* ja *a* esinemistõenäostega *i*-ndas põlvkonnas vastavalt p_i ja q_i ($p_i + q_i = 1$). Populatsioonis toimib valik, mida iseloomustavad selektsiooniindeksid s_1, s_2 ja s_3, st et genotüüpe *AA*, *Aa* ja *aa* kõrvaldatakse populatsioonist vastavalt proportsioonides s_1, s_2 ja s_3. Seega on loomulik eeldada, et alleelide *A* ja *a* kohastumusmaatriks

$$\mathbf{W} = \begin{pmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \end{pmatrix},$$

kus w_{11} on genotüübi *AA* kohastumus (genotüübiga *AA* indiviidi oodatav järglaste arv on $2w_{11}$), w_{12} on genotüübi *Aa* kohastumus jne, on selektsiooniindeksite kaudu esitatav maatriksiga

$$\mathbf{W} = c\begin{pmatrix} 1-s_1 & 1-s_2 \\ 1-s_2 & 1-s_3 \end{pmatrix},$$

st et mingi genotüübiga indiviidide järglaste arv on proportsionaalne selle genotüübiga indiviidide allesjäämise tõenäosusega.

Vastavalt teoreemile populatsiooni statsionaarsetest seisunditest (pt 1.5.2, valem 1.7) avaldub alleeli *A* sageduse muutus $\Delta p = p_i - p_{i-1}$ kujul

$$\Delta p = \frac{p_{i-1}^3(s_1 + s_3 - 2s_2) + p_{i-1}^2(3s_2 - s_1 - 2s_3) + p_{i-1}(s_3 - s_2)}{p_{i-1}^2(2s_2 - s_1 - s_3) + 2p_{i-1}(s_3 - s_2) + 1 - s_3}.$$

Näitame, et analoogse valemi, modelleerimaks alleeli A sageduse muutust, saame ka Fisheri teoreemist lähtudes. Modifitseerides pisut tähistusi Fisheri loodusliku valiku teoreemis (saavutamaks paremat kooskõla valiku mõju uurimisel kasutatud tähistustega, eeldame nüüd, et erinevalt Fisheri loodusliku valiku teoreemist ei tähista p_i mitte *i*. alleeli tõenäosust vaid hoopis alleeli *A* tõenäosust põlvkonnas *i*, $1 - p_i = q_i$ on seega alleeli *a* tõenäosus põlvkonnas *i*), saame, et

$$\Delta p = \frac{1}{\bar{w}} p_{i-1}(\bar{w}_{1.} - \bar{w}),$$

kus $\bar{w}_{1.} = p_{i-1}w_{11} + q_{i-1}w_{12} = cp_{i-1}(1 - s_1) + c(1 - p_{i-1})(1 - s_2)$ on alleeli *A* keskmine kohastumus ja

$$\begin{aligned}
\bar{w} &= p_{i-1}^2 w_{11} + 2p_{i-1}(1 - p_{i-1})w_{12} + (1 - p_{i-1})^2 w_{22} \\
&= cp_{i-1}^2(1 - s_1) + c2p_{i-1}(1 - p_{i-1})(1 - s_2) + c(1 - p_{i-1})^2(1 - s_3) \\
&= c[\underbrace{p_{i-1}^2 + 2p_{i-1}(1 - p_{i-1}) + (1 - p_{i-1})^2}_{[p_{i-1} + (1 - p_{i-1})]^2 = 1} - p_{i-1}^2 s_1 - 2p_{i-1}(1 - p_{i-1})s_2 - (1 - p_{i-1})^2 s_3] \\
&= c[p_{i-1}^2(2s_2 - s_1 - s_3) + 2p_{i-1}(s_3 - s_2) + 1 - s_3].
\end{aligned}$$

Seega

$$\begin{aligned}
\bar{w}_{1.} - \bar{w} &= c[p_{i-1}(1 - s_1) + (1 - p_{i-1})(1 - s_2) - p_{i-1}^2(2s_2 - s_1 - s_3) - 2p_{i-1}(s_3 - s_2) - 1 + s_3] \\
&= c[p_{i-1}^2(s_1 + s_3 - 2s_2) + p_{i-1}(3s_2 - s_1 - 2s_3) + s_3 - s_2].
\end{aligned}$$

Pannes saadud keskmiste kohastumuste avaldised alleelisageduse muudu valemisse, saame

$$\begin{aligned}
\Delta p &= \frac{1}{\bar{w}} p_{i-1}(\bar{w}_{1.} - \bar{w}) \\
&= \frac{cp_{i-1}[p_{i-1}^2(s_1 + s_3 - 2s_2) + p_{i-1}(3s_2 - s_1 - 2s_3) + s_3 - s_2]}{c[p_{i-1}^2(2s_2 - s_1 - s_3) + 2p_{i-1}(s_3 - s_2) + 1 - s_3]} \\
&= \frac{p_{i-1}^3(s_1 + s_3 - 2s_2) + p_{i-1}^2(3s_2 - s_1 - 2s_3) + p_{i-1}(s_3 - s_2)}{p_{i-1}^2(2s_2 - s_1 - s_3) + 2p_{i-1}(s_3 - s_2) + 1 - s_3}.
\end{aligned}$$

Viimane avaldis on identne valiku mõju uurimisel saadud valemiga. Seega viivad Robert Fisheri poolt evolutsiooniteooria tarvis genotüüpide kohastumuse kaudu defineeritud alleelisageduste muutust generatsioonist generatsiooni kajastavad valemid sama tulemuseni, kui genotüüpide mittevõrdse valiku alusel tuletatud valemid (eeldada on vaja üksnes seda, et mistahes genotüübiga indiviidide järglaste arv on samas proportsioonis vastava genotüübiga indiviidide allesjäämise tõenäosusega, kusjuures alleelisageduste muutus ei sõltu selle proportsiooni suurusest).

3.2 Molekulaarne evolutsioon ja fülogenees

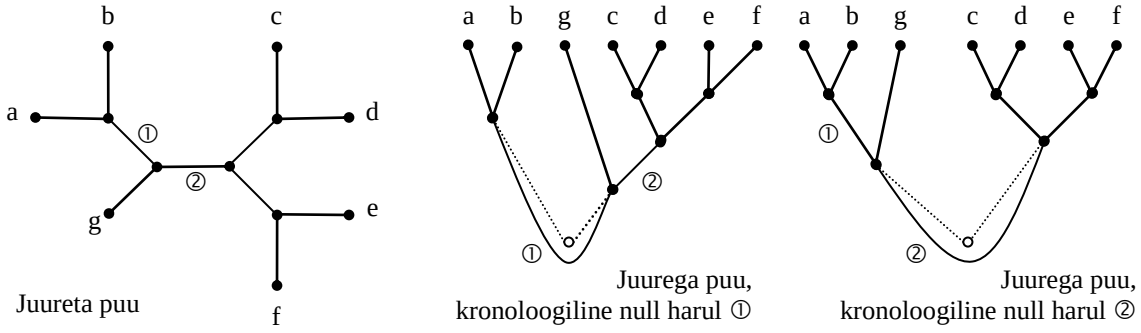
3.2.1 Sissejuhatus

Molekulaarsel tasemel realiseerub evolutsioon päritavat informatsiooni kandvate DNA molekulide või nende mingite piirkondade pikaajalise muutumisena. Kuna kõik indiviidid populatsioonis ei ole geneetiliselt identsed, on õigem rääkida üksnes DNA molekulide jaotuse muutumisest. Lihtsuse huvides piirdume aga edaspidi idealisatsiooniga, kus liigi kõikidel isenditel on ühesugune DNA molekulide komplekt (pank), mis evolutsiooni käigus muutub. Seda suhteliselt stabiilset pikaajalist muutumist moonutavad üksnes **bifurkatsioonid**, kus liigid (või geenid) lõhenevad/lahknevad uuteks liikideks (või geenideks). Et liigi või geeni sisu evolutsioonilises bifurkatsioonis muutub (ühest saab kaks uut, millest kumbki ei ole eellas-DNA „õigusjärglane"), nimetame liiki või geeni identifitseerimisperioodil **operatiivseks taksonoomiliseks**[3] **ühikuks** (OTÜ, *operational taxonomy unit*)[4]. Liigist (geenist) erinevad OTÜ-d selle poolest, et peale bifurkatsiooni nad lakkavad olemast (kaotavad identiteedi), andes ruumi järgmistele, nendest tekkinud OTÜ-dele. Liigi võib aga mõnikord samastada üksteisele ajaliselt järgnevate OTÜ-de blokiga[5].

OTÜ-de ajalist (evolutsioonilist) järgnevust e **fülogeneesi** saab väljendada **fülogeneesipuuga** (põlvnemis- e arengu- e evolutsioonipuuga), kus põlvnemist näitavad puu harud, mis ühendavad OTÜ-sid. Selline puu võib olla nii orienteeritud (suunaga, **juurega** [*rooted*], kronoloogilise nulliga) kui ka orienteerimata (suunata, **juureta** [*unrooted*]).

Joonisel 3.1 on näitena kujutatud juureta fülogeneesipuu ja kaks selle orienteeritud varianti.

Tänapäevased OTÜ-d (liigid või geenid) on puu „lehtedeks" (**tippudeks**) – joonistel märgitud tähtedega a, …, g. Puu tippe koos bifurkatsioonipunktidega nimetatakse ka puu **sõlmedeks** ja neid ühendavaid teid **harudeks**. OTÜ-de vahelise kauguse hindamisel ei oma puu juur (nn kronoloogiline nullpunkt) tähtsust, mistõttu jäetakse see enamasti eraldi välja toomata ja räägitakse nn nullharust.



Joonis 3.1. Juureta 7-tipuline fülogeneesipuu ja kaks selle juurega varianti.

Fülogeneesipuude konstrueerimine on oluline taksonoomias (süstemaatikas), sest võimaldab seletada liikide omavahelist lähedust evolutsioonilise põlvnemisega. Bioloogi huvitabki enamasti liikide põlvnemisel (fülogeneesil) rajanev taksonoomia – **kladistika**, kus kronoloogilise nulli fikseerimine on oluline; vastavaid (orienteeritud) fülogeneesipuid nimetatakse ka **klodogrammideks**[6]. Kladistika vastandiks on **feneetika**, mis piirdub liikide sarnasusel põhineva läheduspuu – fenogrammi – konstrueerimisega ja ei tähtsusta evolutsioonilist ajafaktorit. Fenogramm on sisuliselt juureta (orienteerimata) fülogeneesipuu.

Molekulaarse evolutsiooni uurimine seab kaks üldist eesmärki:

(a) teha kindlaks fülogeneesipuu topoloogia ja

(b) hinnata puu harude ajalised pikkused.

[3] **Taksonoomia** = süstemaatika.

[4] Kirjanduses nimetatakse OTÜ-deks fülogeneesipuu „lehti", mis tavaliselt kujutavad tänapäevaseid liike, bifurkatsioonipunktidele vastavaid OTÜ-sid nimetatakse siis **hüpoteetilisteks taksonoomilisteks ühikuteks** (HTÜ).

[5] Ülevaate saamiseks biosüstemaatika (ja seeläbi näiteks ka teatud isendigruppide liikideks kuulutamise) olemusest vt prof Erast Parmasto loengut:
http://www.botany.ut.ee/lectures/mukoloogia/biosustemaatika.teooria.ja.meetodid.pdf

[6] Vt näiteks http://lepo.it.da.ut.ee/~mprous/Zoa/

Esimese ülesande lahendamiseks on tänapäeval kasutusel kaks peamist printsiipi – maksimaalse parsimoonia printsiip[7] ja maksimaalse sarnasuse printsiip[8], mille matemaatiline taust väljub käesoleva kursuse raamidest.

Teise ülesande lahendamiseks eeldame, et

1. kõigil vaatluse alla võetavatel kaasaegsetel liikidel või geenidel (OTÜ-del) on teada nukleotiidide järjestus mingites kindlates DNA lõikudes (seeläbi on teada ka nukleotiidide järjestuste alusel genereeritavate valke moodustavate aminohapete järjestus);
2. järjekordne evolutsiooniline nukleotiidi asendumine jadas saab toimuda nn **üritusel**, mis seisneb järgmises: kõigi lookuste hulgast valitakse huupi (ühtlase diskreetse jaotusega) mingi lookus ja selles lookuses olev nukleotiid „kirjutatakse üle" kõigi võimalike nukleotiidide hulgast valitud nukleotiidiga;
3. ürituste toimumise intensiivsus kogu vaadeldaval evolutsiooniperioodil on ühesugune (**molekulaarse kella printsiip**[9]), st et oodatav ürituste arv evolutsioonilise aja ühikus on konstantne.

3.2.2 Liikide lahknemisaja hindamine suurima tõepära meetodil

Fülogeneesipuu harude ajalise pikkuse abil saab hinnata tänapäevaste liikide lahknemise aega. Oletame, et huvipakkuvail (fülogeneesipuu tippudele vastavail) liikidel on määratus s valgu aminohappeline järjestus. Iga valgu kohta on teada seda moodustavate aminohapete arv v_i, erinevate aminohapete arv võrreldavail liikidel k_i (nn muutuste arv) ja valgu muutumise kiirus w_i (näitab, kui mitu aminohapete muutust on valgus toimunud ühe aasta jooksul; hinnanguliselt on muutumise kiirus umbes 1 muutus miljoni aasta kohta).

Eeldame, et muutused toimuvad ajas kogu aeg ühe kiirusega, sellisel juhul võib muutuste arvu i.-ndas valgus (kui väga harva aset leidvate sündmuste arvu) jaotuseks lugeda Poissoni jaotust, $k_i \sim P(\lambda_i)$, tõenäosusfunktsiooniga $\mathrm{P}(k_i \text{ muutust}) = e^{-\lambda_i}\lambda_i^{k_i}/k_i!$.

Samas on kahe võrreldava liigi i.-nda valguga aset leidnud keskmine muutuste arv esitatav korrutisena

$$\mathrm{E}(k_i) = 2t \times w_i \times v_i$$

(t on ajavahemik liikide lahknemist märkivast bifurkatsioonipunktist tänapäevani; et selle aja jooksul on eeldatavalt muutused toimunud mõlema liigiga, tuleb kõigi kaht liiki eristavate muutuste arvu leidmiseks see ajavahemik korrutada kahega), mistõttu $k_i \sim P(2tw_iv_i)$.

Kõigi s valguga sõltumatult aset leidnud muutuste $k_1, \ldots, k_s$ ühisjaotusele vastav tõepärafunktsioon avaldub kujul

$$L(t; k_1,\ldots,k_s, v_1,\ldots,v_s, w_1,\ldots,w_s) = \prod_{i=1}^{s} e^{-2tw_iv_i}\frac{(2tw_iv_i)^{k_i}}{k_i!}.$$

Lahknemise aja t hindamiseks tuleb see tõepärafunktsioon maksimeerida t suhtes. Vastav logaritmiline tõepärafunktsioon on

$$\ln L(t; k_1,\ldots,k_s, v_1,\ldots,v_s, w_1,\ldots,w_s) = -\sum_{i=1}^{s} 2tw_iv_i + \sum_{i=1}^{s} k_i \ln(t) + const.$$

ja selle tuletis t järgi

$$\frac{\partial \ln L}{\partial t} = -2\sum_{i=1}^{s} w_iv_i + \frac{1}{t}\sum_{i=1}^{s} k_i.$$

Võrdsustades viimase avaldise 0-ga, saame suurima tõepära hinnangu liikide lahknemisajale kujul

[7] Maksimaalse parsimoonia printsiip – püütakse valitakse niisugune puu, mis eeldab minimaalse arvu evolutsioonilisi sündmusi (säästuprintsiip).

[8] Maksimaalse sarnasuse printsiip – OTÜ (liik) paigutatakse kokku (grupeeritakse) selle OTÜ-ga (liigiga), kellega ta jagab suurima arvu ühiseid tunnuseid.

[9] R. Villems. Loengukursus evolutsioonilisest bioloogiast. http://www.ebc.ee/tymri00/loengud/evol/
Molekulaarsete andmete kasutamine fülogeneetilisteks rekonstruktsioonideks baseerub ja on võimalik vaid molekulaarse kella olemasolu kontekstis. Molekulaarse kella kontseptsioon oma tavatähenduses väidab, et makromolekulide järjestus muutub evolutsioonilises ajaskaalas lineaarselt. Molekulaarse kella kontseptsioon ei väida, et kõik makromolekulid muutuvad ühesuguse kiirusega. Tuleb aga veel lisada, et lineaarsus ei ole iseenesest mingi absoluutne nõue – oluline on hoopis selle kiiruse teadmine. Et aga ebalineaarsusi, eriti veel suisa edasi-tagasi kõikumisi on üsna raske näha, siis on lineaarsus tavaliselt piisavaks lähenduseks seni, kuni pole näidatud ebalineaarsus.

$$\hat{t} = \frac{\sum_{i=1}^{s} k_i}{2\sum_{i=1}^{s} w_i v_i}. \tag{3.4}$$

Näide. Püüame hinnata inimese ja šimpansi lahknemise aega, võttes vaatluse alla 6 valku. Järgmises tabelis on kirjas nende valkude hinnangulised muutumise kiirused w_i, aminohapete arvud v_i ning inimese ja šimpansi vahelised erinevuste (erinevate aminohapete) arvud k_i.

Valk	w_i	v_i	k_i
Fibrünopeptiid	45×10^{-10}	30	0
Tsütokroom C	$2{,}5\times10^{-10}$	104	0
Hemoglobiin α	$9{,}9\times10^{-10}$	141	0
Hemoglobiin β	13×10^{-10}	146	0
Hemoglobiin δ	10×10^{-10}	146	1
Müoglobiin	10×10^{-10}	153	1

Vastavalt valemile (3.4) saame lahknemisaja hinnanguks $\hat{t} \approx 1{,}3\times10^{6}$ aastat, mis on vägagi ebatäpne.

Võttes vaatluse alla enam valke, on inimese ja šimpansi lahknemisaja (ehk siis vastava fülogeneesipuu haru pikkuse) hinnanguks saadud $\hat{t} \approx 4{,}6\times10^{6}$ aastat – mida rohkem valke ja mida pikemad ajavahemikud, kus muutused said toimuda, seda täpsem hinnang.

Kirjeldatud suurima tõepära hinnang on küll suhteliselt lihtsalt leitav, aga nõuab väga palju lihtsustavaid eeldusi, mistõttu saadud hinnangud ei pruugi reaalsusega kokku sobida. Järgnevalt vaatame pisut teistel alustel välja töötatud metoodikat fülogeneesipuu haru ajalise pikkuse hindamiseks.

3.2.3 Fülogeneesipuu haru ajalise pikkuse hindamine teadaoleva muutuste arvu järgi

Kui DNA panka (nukleotiidide või aminohapete jada) tabab mingil evolutsioonilisel ajavahemikul x **üritust** (atakki), mille all võib mõista nii ühe nukleotiidi asendumist teisega kui ka ühe aminohappe asendumist teisega, siis toovad need üritused kaasa y muutust. Ilmselt $0 \le y \le x$, sest esiteks ei tarvitse iga üritus põhjustada jada muutust (nukleotiid A asendub nukleotiidiga A, näiteks) ja teiseks võib üht ja sama positsiooni tabada mitu üritust nii, et viimane neist taastab algseisu viies muutuste arvu 0-ks. Muutuse tõenäosus jada mingis positsioonis on ühesugune sõltumata sellest, kas positsiooni tabab üks või mitu üritust. Seega sõltub y-i jaotus ainult ürituste poolt tabatud positsioonide arvust.

Oletades, et toimunud muutuste arv y on teada, hindame toimunud ürituste arvu x. Olgu jada pikkus V (= geeni moodustavate nukleotiidide arv või valgu moodustavate aminohapete arv või …) ja võimalike elementide arv, millega mingis positsioonis paiknev nukleotiid või aminohape asenduda võib, N (kui tegu on nukleotiididega, siis $N = 4$, kui aga aminohapetega, siis $N = 20$).

Tõenäosus, et üks juhuslikult valitud positsioon saab ühe üritusega pihta, on $1/V$, tõenäosus, et ei saa pihta, on $1-1/V$. Kui toimub x üritust, siis tõenäosus, et juhuslikult valitud positsioon ei saa pihta, on $(1-1/V)^x$ ja tõenäosus, et saab pihta, on $1-(1-1/V)^x$. Keskmine tabatud positsioonide arv on sellisel juhul V korda suurem.

Kuna iga ürituse korral on tõenäosus, et üritusega kaasnes ka muutus $(N-1)/N = 1-1/N$ (so tõenäosus, et nukleotiid või aminohape asendus teisega), siis on x ürituse korral oodatav muutuste arv jadas pikkusega V positsiooni

$$E(y) = V\left(1-\frac{1}{N}\right)\left[1-\left(1-\frac{1}{V}\right)^{x}\right]. \tag{3.5}$$

Paneme tähele, et ürituste arvu x suurenemisel suureneb ka oodatav muutuste arv E(y), kuid kehtib võrratus $E(y) \le V\ 1-1/N$.

Reaalsuses on meil teada muutuste arv y ja molekulaarse kella printsiibist lähtuvalt leitud hinnang ürituste arvu intensiivsusele. Leidmaks neile andmetele tuginedes fülogeneesipuu harude pikkusi ja seeläbi tänapäevaste liikide lahknemisaegu, peame leidma funktsiooni (3.5) pöördfunktsiooni.

Kasutades momentide meetodi ideoloogiat, võrdsustame $\mathrm{E}(y) = y$ ja avaldame valemist (3.5) ürituste arvu x. Saame

$$\hat{x} = \frac{\ln\left(1-\dfrac{y}{V(1-1/N)}\right)}{\ln(1-1/V)}. \tag{3.6}$$

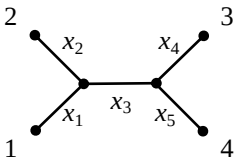
Olles leidnud fülogeneesipuu harudest moodustuvale teele vastava ürituste arvu hinnangu $\hat{x}$, saame hinnata sellele teele vastava evolutsioonilise ajavahemiku pikkust $\Delta t = \kappa\hat{x}$, kus κ on ürituste toimumise intensiivsus (näiteks κ üritust 10^6 aasta kohta).

3.2.4 Terve fülogeneesipuu hindamine

Rakendamaks kirjeldatud metoodikat L tänapäevase liigi (OTÜ) vahelise kauguse hindamiseks ürituste arvude skaalas, tuleb esmalt leida kõigi liikide paaride vahelised muutuste (molekulaarsete erinevuste) arvud – koondame need $L(L-1)/2$-komponendilisse vektorisse $\mathbf{y} = (y_{12}\ y_{13}\ \ldots\ y_{L-1,L})^{\mathrm{T}}$. Tähistades tänapäeva liikide molekulaarsete erinevuste tekitamiseks kulunud ürituste arvude hinnangute vektori $\mathbf{x} = (x_{12}\ x_{13}\ \ldots\ x_{L-1,L})^{\mathrm{T}}$, saame iga elemendi x_{ij} hindamiseks kasutada valemit (3.6) argumendiga y_{ij}.

Osutub, et liikide vaheliste ürituste arvude hinnangute vektorit **x** saab kasutada, hindamaks ürituste arvu ka nendel fülogeneesipuu harudel, mille kohta muutuste (ja seeläbi ka ürituste) arvud ei ole vahetult teada. Põhjuseks on asjaolu, et ürituste arvud puu harudel on erinevalt muutuste arvudest aditiivsed.

Hindamise olemuse selgitamiseks võtame vaatluse alla juuresoleval skeemil kujutatud juureta puu, kus x_1, …, x_5 on ürituste arvud puu harudel. Tänapäevaste liikide molekulaarsete erinevuste põhjal hinnatavad liikidevahelised ürituste arvud (ürituste arvud fülogeneesipuu tippude vahel) avalduvad siis võrranditena



$$\begin{aligned} x_{12} &= x_1 + x_2 \\ x_{13} &= x_1 + x_3 + x_4 \\ x_{14} &= x_1 + x_3 + x_5 \\ x_{23} &= x_2 + x_3 + x_4 \\ x_{24} &= x_2 + x_3 + x_5 \\ x_{34} &= x_4 + x_5 \end{aligned},$$

ehk maatrikskujul

$$\mathbf{x} = \mathbf{K}\boldsymbol{\beta}\,, \tag{3.7}$$

kus **K** on $L(L-1)/2$-realine (uuritavate liikide paaride arv) ja $(2L-3)$-veeruline (kõigi fülogeneesipuu harude arv) plaanimaatriks ja **β** on vektor puu harudel mõjunud ürituste arvudest,

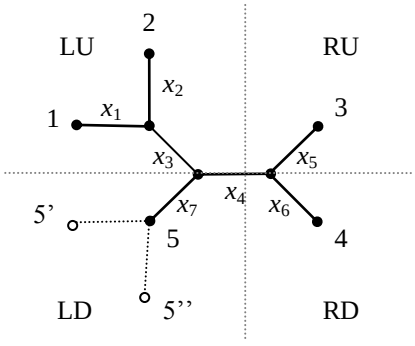
$$\mathbf{K} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix},\ \boldsymbol{\beta} = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{pmatrix}.$$

Vähimruutude hinnangud ürituste arvudele üksikharudel on leitavad maatriksvõrdusest

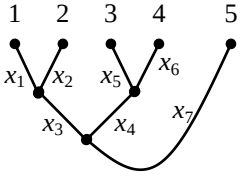
$$\hat{\boldsymbol{\beta}} = (\mathbf{K}^{\mathrm{T}}\mathbf{K})^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{x}\,. \tag{3.8}$$

Pöördmaatiks valemis (3.8) eksisteerib alati, kuid selle arvutamine suurte puude korral võib olla tehniliselt raske – näiteks 200 tänapäevase OTÜ korral on $\mathbf{K}^{\mathrm{T}}\mathbf{K}$ dimensioon 397×397, hoopis võimatu on aga piirprotsesside uurimine, kus OTÜ-de arv piiramatult suureneb.

Siiski on puu harude hindamiseks ka alternatiivne tee. Oletame näiteks, et vaja on hinnata ürituste arv x_4 kõrval kujutatud juurega fülogeneesipuul. Esimese sammuna joonistatakse puu üles juureta puuna, kusjuures hinnatav haru paigutatakse keskele, jagades selleks puu neljaks kvadrandiks (LD, LU, RU ja RD).



Juhul, kui puu vasakpoolsed või parempoolsed nurgad ei ole omavahel sümmeetrilised, laiendatakse puud (antud näite puhul on tipule 5 lisatud kaks fiktiivset haru tippudeni 5' ja 5'').



Järgnevalt võetakse vaatluse alla kõik võimalikud teed puu vasakpoolsetest tippudest parempoolsetesse tippudesse – kokku on selliseid teid $N_L \times N_R$, kus N_L on puu vasaku poole tippude arv ja N_R on puu parempoolsete tippude arv (koos lisatud fiktiivsete tippudega). Liites kokku kõigile teedele vastavad ürituste arvud, saame summa

$$\sum_{i\in L,\, j\in R} x_{ij}\,, \tag{3.9}$$

kusjuures laiendatud **x**-i komponendid võetakse võrdseks reaalse **x**-i vastavate komponentidega: $x_{i5'} = x_{i5''} = x_{i5}$. Meie näite puhul on summa (3.9) järgmine:

$$x_{13} + x_{14} + x_{23} + x_{24} + 2x_{53} + 2x_{54}\,.$$

Otsitav ürituste arv sisaldub summas (3.9) täpselt $N_L \times N_R$ korda, kõik puu vasakusse poolde kuuluvad harud sisalduvad kordusega $2^{h_i} N_R$ ja kõik paremasse poolde kuuluvad harud kordusega $2^{h_i} N_L$, h_i märgib harude arvu i.-st harust lähima tipuni (näiteks haru x_2 puhul $h_2 = 0$, haru x_3 puhul $h_3 = 1$ jne).

Analoogselt saab leida uurimise alla võetud harust vasakul pool paiknevate tippude kõikvõimalikele omavahelistele teedele vastavad ürituste arvud

$$\sum_{i \in LU,\, j \in LD} x_{ij}$$

(iga vasakpoolne haru sisaldub summas kordsusega $2^{h_i} N_L / 2$) ja parempoolsete tippude vahelistele teedele vastavad ürituste arvud

$$\sum_{i \in RU,\, j \in RD} x_{ij}$$

(iga parempoolne haru sisaldub summas kordsusega $2^{h_i} N_R / 2$).

Jagades summa (3.9) läbi liidetavate arvuga $N_L \times N_R$, saame avaldise, kus otsitav tsentraalharu pikkus sisaldub täpselt üks kord, kõik tsentrist vasakule jäävad harud kordajaga $2^h / N_L$ ja kõik tsentrist paremale jäävad harud kordajaga $2^h / N_R$. Edasi on ilmne, et otsitav tsentraalharul toimunud ürituste arv on leitav, lahutades puu kõigi vasak- ja parempoolsete tippude vaheliste teede sobivalt kaalutud summast kõigi vasakpoolsete kvadraatide ja parempoolsete kvadraatide tippude vaheliste teede sobivalt kaalutud summad:

$$x_\alpha = \frac{1}{N_L N_R} \sum_{\substack{i \in L \\ j \in R}} x_{ij} - \frac{2}{N_L^2} \sum_{\substack{i \in LU \\ j \in LD}} x_{ij} - \frac{2}{N_R^2} \sum_{\substack{i \in RU \\ j \in RD}} x_{ij}\,, \tag{3.10}$$

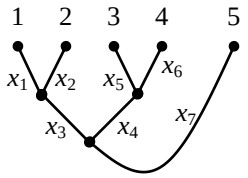
x_α tähistab suvalist juureta puu tsentraalharul toimunud ürituste arvu.

Meie näite puhul avaldub ürituste arv x_4 kujul

$$x_4 = \tfrac{1}{4\times 2}(x_{13} + x_{14} + x_{23} + x_{24} + 2x_{53} + 2x_{54}) - \tfrac{2}{4^2}(2x_{15} + 2x_{25}) - \tfrac{2}{2^2}(x_{34})\,.$$

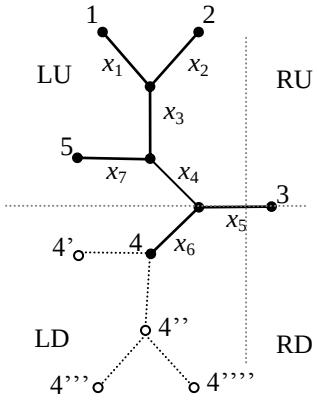
Näide. Võtame vaatluse alla kõrvaloleva fülogeneesipuu ja hindame ürituste arvu x_5, eeldades, et liikide 1 kuni 5 vahelised muutuste arvud y_{ij}, $i,j = 1, \ldots, 5$, on teada.

y_{12}	y_{13}	y_{14}	y_{15}	y_{23}	y_{24}	y_{25}	y_{34}	y_{35}	y_{45}
2	3	1	5	4	4	6	1	3	3



Puu tippude vahelised ürituste arvud x_{ij} hindame valemist (3.6).

$\hat{x}_{12}$	$\hat{x}_{13}$	$\hat{x}_{14}$	$\hat{x}_{15}$	$\hat{x}_{23}$	$\hat{x}_{24}$	$\hat{x}_{25}$	$\hat{x}_{34}$	$\hat{x}_{35}$	$\hat{x}_{45}$
2,12	3,19	1,05	5,38	4,28	4,28	6,49	1,05	3,19	3,19

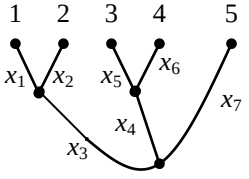


Järgnevalt joonistame välja vastava juureta puu, paigutades huvipakkuva haru keskele, täiendame seda fiktiivsete harude ja tippudega ning jagame neljaks kvadraadiks.

Vastavalt valemile (3.10) avaldub ürituste arv x_5 kujul

$$\begin{aligned}\hat{x}_5 &= \tfrac{1}{6\times 1}(x_{13} + x_{23} + x_{53} + 3x_{43}) - \tfrac{2}{6^2}(3x_{14} + 3x_{24} + 3x_{54}) - \tfrac{2}{1^2}\times 0 \\ &= \tfrac{1}{6}(3{,}19 + 4{,}28 + 3{,}19 + 3\times 1{,}05) \\ &\quad - \tfrac{1}{18}(3\times 1{,}05 + 3\times 4{,}28 + 3\times 3{,}19) \\ &= 0{,}883\,.\end{aligned}$$

Ülesanne 15. Hinnake kõrvaloleva fülogeneesipuu baasil ürituste arvu x_3, eeldades, et liikide 1 kuni 5 vahelised ürituste arvud x_{ij}, $i,j = 1, \ldots, 5$, on teada:



x_{12}	x_{13}	x_{14}	x_{15}	x_{23}	x_{24}	x_{25}	x_{34}	x_{35}	x_{45}
2	3	1	5	4	4	6	1	3	3