3.2 Klassikaline aheldusanalüüs

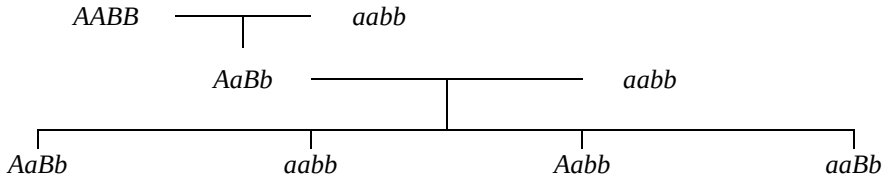
3.2.1 Aheldusanalüüsi olemus, markerite informatiivsus

Oma klassikaliselt definitsioonilt on aheldusanalüüs meetod, mis püüab teha järeldusi uuritud lookuste suhtelise paiknemise kohta genoomis baseeruvana Morgani seadustel. Lihtsaimal, vaid kahe lookuse analüüsil, taandub probleem nende lookuste aheldumise testimisele ja nendevahelise rekombineerumise tõenäosuse hindamisele. Juhul, kui juhuslikult valitud indiviidide genoomi moodustanud sugurakkude kohta on täpselt kokku loetavad vaadeldavate lookuste suhtes rekombinantsed gameedid, seisneb aheldumise test kontrollimises, kas rekombinantsete gameetide proportsioon võrdub ½-ga (ahelduse puudumisele vastav nullhüpotees) või on sellest väiksem (lookuste aheldust kinnitav sisukas hüpotees).

Et igal indiviidil on võimalik määrata vaid genotüüp uuritavate markerlookuste osas, on genotüübi kahe haplotüübina esitamiseks vaja teada ka vanemate genotüüpe – et tuvastada, kummalt vanemalt konkreetne alleel pärandus, ehk millised alleelid moodustavad ühe haplotüübi. Selleks, et osata öelda, kas uuritava indiviidi ühe haplotüübina identifitseeritud alleelid paiknevad samas haplotüübis ka selle indiviidi vanemal, ehk kas sugurakkude moodustumisel vanema organismis on nende alleelide vahel aset leidnud rekombinatsioon või mitte, on vaja teada ka vanavanemate genotüüpe.

Täpsustamaks, kas genotüübi $A_1A_2B_1B_2$ moodustavad haplotüübid A_1B_1 ja A_2B_2 või A_1B_2 ja A_2B_1, kasutatakse tähistusi $A_1B_1|A_2B_2$ või $A_1B_2|A_2B_1$, mis aheldusanalüüsi seisukohast märgivad erinevaid genotüüpe. Juhul, kui võtta arvesse ka see, kummalt vanemalt mingi haplotüüp pärandus, võib esialgne genotüübi kirjeldus vastata neljale erinevale genotüübile: $A_1B_1|A_2B_2$, $A_2B_2|A_1B_1$, $A_1B_2|A_2B_1$ või $A_2B_1|A_1B_2$, kus traditsiooniliselt on esimesena märgitud haplotüüp pärit isalt. Viimati defineeritud genotüüpe nimetatakse ka **järjestatud genotüüpideks**.

Näide. Vaatleme kolme generatsiooni ühes perekonnas, kus genotüüp on määratud 2 dialleelse markeri osas.



Toodud põlvnemisskeem hõlmab 8 indiviidi ja seega potentsiaalselt 16 erinevat haplotüüpi. Esimeses nn vanavanemate generatsioonis on indiviidide haplotüübid tänu homosügootsusele tuvastatavad (*AB*|*AB* vanaisal ja *ab*|*ab* vanaemal), kuid võimalikke rekombinatsioone, mille tulemusel sellised haplotüübid võisid tekkida, pole eelneva generatsiooni mitteteadmise tõttu võimalik leida. Teises nn vanemate generatsioonis on haplotüübid samuti leitavad (*AB*|*ab* isal ja *ab*|*ab* emal), kuid jällegi pole võimalik tuvastada rekombinatsioone. Isa puhul on see tingitud tema vanemate homosügootsusest mõlema lookuse suhtes – ükskõik, kas rekombinatsioon toimus või mitte, meioosi lõpuks moodustunud sugurakud on uuritud lookuste suhtes ikka ühesugused; ema kohta pole aga jällegi teada tema vanemate genotüüpi.

Kolmanda generatsiooni indiviidid pärivad kõik emalt haplotüübi *ab* ega anna seega mingit informatsiooni võimaliku rekombinatsiooni kohta selle haplotüübi moodustumisel. Seega peavad nelja järglase teised haplotüübid pärandununa isalt sisaldama alleele *AB*, *ab*, *Ab* ja *aB*. Et isa genotüüp on juba jagatud haplotüüpideks kujul *AB*|*ab*, siis on selge, et kahe esimese järglase isalt päritud haplotüübid on mitterekombinantsed, aga kahe ülejäänud järglase haplotüüpe moodustavate sugurakkude kujunemisel on uuritavate lookuste vahel toimunud rekombinatsioon.

Kokkuvõttes ei olnud vaatluse all olnud 8 indiviidi 16 haplotüübist tervelt 12 korral mingit võimalust tuvastada kas rekombinatsioon toimus või mitte ning seega on need 12 haplotüüpi aheldusanalüüsi seisukohast ebainformatiivsed. Ülejäänud 4 haplotüüpi öeldakse olevat **täielikult informatiivsed** (*fully informative*), sest nende rekombineeritus (*recombination status*) on täpselt tuvastatav (analoogse põhimõtte alusel defineeritakse mõnes kirjutises ka

informatiivsed meioosid ja/või **gameedid**).

Kuigi sarnaselt kompleksele segregatsioonanalüüsile on ka aheldusanalüüsi puhul välja kirjutatavad üldised tõepärafunktsioonid arvestamaks mitteteadaolevate haplotüüpide puhul kõigi võimalike alleelikombinatsioonidega, püütakse enamasti uuritavaiks valida vaid informatiivseid indiviide (so siis neid, kelle puhul on täpselt teada, millise alleeli/haplotüübi ta milliselt vanemalt päris). Informatsiooni sellest, millistest haplotüüpidest vanemate sugurakud (ehk siis uuritava järglase haplotüübid) meioosi protsessil kokku pannakse, nimetatakse **meioosi faasiks** (ka **ahelduse faasiks** – *phase of meioses*, *linkage phase*).

Lisaks uuritavate indiviidide valikule on oluline ka kasutatavate markerite polümorfsus, sest pole ju erilist tolku geneetilisest markerist, mille suhtes on enamus indiviide homosügootsed – siis ei ole mitte kuidagi võimalik tuvastada, kumb kahest identsest DNA-järjestusest järglasele edasi kandus. Markerite kasumlikkust aheldusanalüüsi tarvis püütakse sageli kirjeldada mõne kordajaga.

- Üks neist, nn **heterosügootsuse kordaja** (*heterozygosity*), väljendab tõenäosust, et panmiktilisest populatsioonist juhuslikult valitud indiviid on heterosügootne (so heterosügootsete indiviidide osakaal taolises populatsioonis):

$$H = 1 - \sum_{i=1}^{n} p_i^2 ,$$

siin p_i on vaatlusaluses markerlookuses paikneva i.-nda alleeli sagedus ja n on alleelide koguarv.

- Teine, nn **polümorfismi informatsiooni mahu kordaja** (*polymorphism information content*, *PIC*), mis defineeritakse seosega

$$PIC = 1 - \sum_{i=1}^{n} p_i^2 - \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} 2 p_i^2 p_j^2 ,$$

püüab lisaks vaatlusaluses lookuses heterosügootsete indiviidide osakaalule arvesse võtta ka seda, et sugugi mitte kõik heterosügootse vanempaari järglased pole aheldusanalüüsi seisukohalt informatiivsed. Nimelt on vanempaari $A_iA_j \times A_iA_j$ järglastest keskmiselt vaid pooltel üheselt tuvastatav, kumb isa ja ema genotüüpi kuulunud alleelidest järglasele pärandus (so juhul, kui järglane on homosügootne: A_iA_i või A_jA_j), pooltel juhtudel aga ei ole võimalik öelda, kumba isa ja ema alleelidest järglane kannab (so juhul, kui järglane on heterosügootne: A_iA_j). Vanempaari $A_iA_j \times A_iA_j$ esinemise tõenäosus panmiktilises populatsioonis on $2p_i p_j \times 2p_i p_j = 4p_i^2 p_j^2$, nende järglastest omakorda ½ ei ole aheldusanalüüsi seisukohast informatiivsed, mistap nende esinemise sagedus lahutatakse kõigi heterosügootide esinemise sagedusest.

3.2.2 Klassikalise aheldusanalüüsi ülesande püstitus – tõepärasuhe ja LOD-skoor

Klassikalise aheldusanalüüsi korral võetakse vaatluse alla kaks lookust ning uuritakse, kas mingid alleelide kombinatsioonid (haplotüübid) päranduvad järglastele sagedamini, kui võiks eeldada lookuste sõltumatuse (ahelduse puudumise) korral. Viimasel juhul on nende kahe lookuse/geeni vaheline rekombinatsioonimäär $r = 0{,}5$ (taoline on olukord alati, kui kaks geeni paiknevad erinevates kromosoomides). Kui aga kaks lookust on nö täielikult aheldunud, päranduvad neis paiknevad alleelid alati koos ja nende lookuste vaheline rekombinatsioonimäär $r = 0$.

Kuna aheldusanalüüsi eesmärk on leida vaid potentsiaalse haigusgeeni (või mõnda muud tunnust märkimisväärselt mõjutava geeni) ligikaudne asukoht kromosoomi geneetilisel kaardil, siis on otsimiskriteeriumiks 0,5-st väiksem rekombinatsioonimäär (ehk markerlookus, mis on potentsiaalse haigusgeeni lookusega aheldunud). Hüpoteeside kontrollimise kontekstis soovitakse testida ühepoolset hüpoteeside paari kujul

$$H_0: r = ½,$$
$$H_1: r < ½.$$

Analüüsi teostamise eelduseks on mingi hulga informatiivsete meiooside olemasolu, st et genotüpiseeritud peab olema mitu põlvkonda indiviide, kelle alusel saab tuvastada vanemailt järglastele pärandunud haplotüübid ning lugeda kokku, kui paljudel juhtudel leidis aset rekombinatsioon ja kui paljudel juhtudel mitte.

Edasi rakendatakse enamasti suurima tõepära meetodit hindamaks andmetest rekombinatsioonimäära väärtust ning testitakse tõepärasuhte testi abil rekombinatsioonimäära erinevust ½-st (st, testitakse ahelduse statistilist olulisust).

Seejuures defineeritakse geneetilises statistikas tõepärasuhe enamasti kulul

$$T = \mathrm{LR} = \frac{\max\limits_{0 \le r \le 0{,}5} L(r)}{L(r = \tfrac{1}{2})},$$

millest tõepärasuhte statistik avaldub seosena

$$\Lambda(r) = 2\ln(T) = 2\ln\left(\frac{\max\limits_{0 \le r \le 0{,}5} L(r)}{L(r = \tfrac{1}{2})}\right) = 2\left[\ln L(\hat{r}) - \ln L(\tfrac{1}{2})\right].^{1}$$

Kõrvuti tõepärasuhte statistikuga ja sageli ka selle asemel kasutatakse aheldusanalüüsil nn **LOD-skoori** (ingl. *LOD-score*, *LOD = logarithm of odds*), mis kujutab endast kümnendlogaritmi tõepärasuhtest:

$$\mathrm{LOD}(r) = \log_{10}(T) = \log_{10}\left(\frac{\max\limits_{0 \le r \le 0{,}5} L(r)}{L(r = \tfrac{1}{2})}\right) = \left[\log_{10} L(\hat{r}) - \log_{10} L(\tfrac{1}{2})\right].$$

Seega on LOD-skoor statistik, mille täisarvulised väärtused 1, 2, 3 jne vastavad "ümmargustele" tõepärasuhetele 10, 100, 1000 jne.

Tähistades $\mathrm{LOD}(r) = \log_{10}(T) = Z$ ja avaldades siit tõepärasuhte $T = 10^Z$, tuleneb tõepärasuhte statistiku definitsioonist seos

$$\Lambda(r) = 2\ln(T) = 2\ln(10) \times \mathrm{LOD}(r),$$

ehk

$$\mathrm{LOD}(r) = \Lambda(r)/2\ln(10) = \Lambda(r)/4{,}6.$$

Traditsiooniliselt loetakse geneetikas kaks lookust/geeni aheldunuks, kui $\mathrm{LOD}(r) > 3$. Tõepärasuhte testi kontekstis on see kriteerium samaväärne olukorraga $\Lambda(\mathrm{r}) > 4{,}6 \times 3 = 13{,}8$, millele vastab tõepärasuhe $T = \mathrm{LR} > 1000$ ja olulisuse tõenäosus $p < 0{,}0002$.

Et rekombinantsete gameetide arv k kõigi informatiivsete meiooside hulgas n on modelleeritav binoomjaotusega, avaldub rekombinatsioonimäära r tõepärafunktsioon kujul

$$L(r \mid n,k) = \mathrm{C}_n^k r^k (1-r)^{n-k}.$$

Suurima tõepära hinnang rekombinatsioonimäärale r on rekombinantsete gameetide suhteline sagedus:

$$\hat{r} = k/n.$$

Et aga aheldusanalüüsil loetakse rekombinatsioonimäära 0,5-st suuremad väärtused bioloogiliselt sobimatuteks, defineeritakse rekombinatsioonimäära hinnang enamasti kujul

$$\hat{r} = \begin{cases} k/n, & \text{kui } k/n \le 1/2; \\ 0{,}5, & \text{kui } k/n > 1/2. \end{cases}$$

Võtame vaatluse alla joonisel 16 kujutatud sugupuu, kus mõlema lookuse osas heterosügootset isast indiviidi ristati 13 mittegenotüpiseeritud emase indiviidiga, igast ristamisest sündis üks järglane, kes on genotüpiseeritud. Kas lookused *A* ja *B* võivad olla aheldunud?

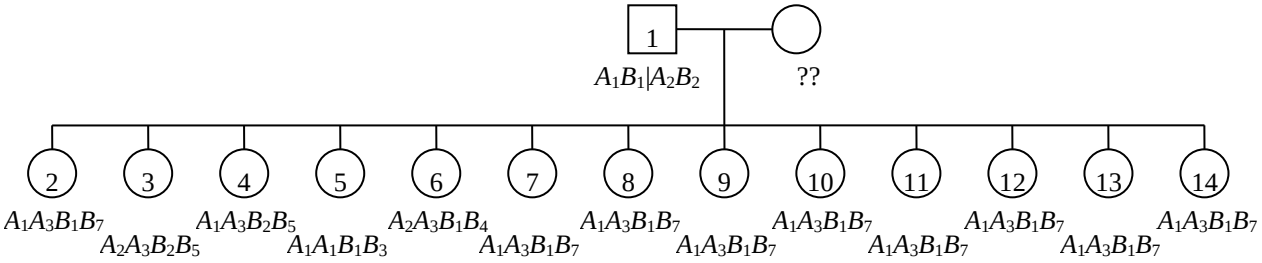
Lookuste ahelduvuse testimiseks tuleb esmalt hinnata andmetest rekombinatsioonimäär ja seejärel testida selle erinevust ½-st.

---

[1] Statistikaraamatutes on tõepärasuhte tähistuseks sageli D ja suhe ise on pöördväärtus geneetikas kasutavast suhtest T: $D = L(r = \tfrac{1}{2}) \Big/ \max\limits_{0 \le r \le 0{,}5} L(r)$.

Sestap on ka tõepärasuhte statistiku valem statistikas enamasti pisut teine:

$$\Lambda(r) = -2\ln(D) = -2\ln\left(\frac{L(r = \tfrac{1}{2})}{\max\limits_{0 \le r \le 0{,}5} L(r)}\right) = -2\left[\ln L(\tfrac{1}{2}) - \ln L(\hat{r})\right].$$



Joonis 16. Sugupuu ühe isa ja 13 järglasega.

Et enamasti pole sugugi kõik indiviidid aheldusanalüüsi seisukohalt informatiivsed, paneme joonisel 16 esitatud sugupuu selguse mõttes kirja ka tabelina (tabel 2). Tabelist on näha, et kokku 13-st meioosist (kuna emad ei ole genotüpiseeritud, ei saa nende sugurakkude moodustumise protsessidega rekombinantsete gameetide tuvastamisel arvestada) on informatiivsed vaid 10.

Tabel 2. Joonisel 16 esitatud sugupuu koos sellelt välja loetava informatsiooniga rekombinatsioonide toimumise kohta.

Isa genotüüp	Järglase nr	Järglase genotüüp	Isapoolne haplotüüp	Rekombinatsioon
$A_1B_1\|A_2B_2$	2	$A_1A_3B_1B_7$	A_1B_1	Ei
	3	$A_2A_3B_2B_5$	A_2B_2	Ei
	4	$A_1A_3B_2B_5$	A_1B_2	Jah
	5	$A_1A_1B_1B_3$	A_1B_1	Ei
	6	$A_2A_3B_1B_4$	A_2B_1	Jah
	7	$A_2A_3B_2B_2$	A_2B_2	Ei
	8	$A_1A_2B_2B_5$	?	-
	9	$A_1A_1B_2B_2$	A_1B_2	Jah
	10	$A_2A_3B_1B_2$	?	-
	11	$A_1A_2B_1B_2$	?	-
	12	$A_2A_3B_2B_7$	A_2B_2	Ei
	13	$A_2A_5B_2B_2$	A_2B_2	Ei
	14	$A_2A_3B_2B_3$	A_2B_2	Ei

Rekombinantsete gameetide arv k kõigi informatiivsete meiooside hulgas n on modelleeritav binoomjaotusega ja rekombinatsioonimäära r tõepärafunktsioon esitub kujul

$$L(r \mid n,k) = \mathrm{C}_n^k r^k (1-r)^{n-k} .$$

Suurima tõepära hinnang rekombinatsioonimäärale r avaldub siit suhtena

$$\hat{r} = k/n .$$

Joonisel 16 ja tabelis 2 toodud näites tuleb rekombinatsioonimäära hinnanguks 3/10 = 0,3.

Tõepärafunktsiooni väärtus antud andmete korral on

$$L(\hat{r} = 0,3 \mid n = 10, k = 3) = \mathrm{C}_{10}^3 \times (0,3)^3 \times (1-0,3)^{10-3} \approx 0,2668$$

ning nullhüpoteesi (r = ½) korral

$$L(r = 0,5 \mid n = 10, k = 3) = \mathrm{C}_{10}^3 \times (0,5)^3 \times (1-0,5)^{10-3} \approx 0,1172 .$$

Tõepärasuhe

$$\mathrm{LR} = \frac{\max\limits_{0 \le r \le 0,5} L(r)}{L(½)} = \frac{0,2668}{0,1172} \approx 2,2769$$

ning tõepärasuhte statistik

$$\Lambda(r) = 2\left[\ln L(\hat{r}) - \ln L(½)\right] \approx 1,646 .$$

Et tõepärasuhte statistiku väärtused peaksid nullhüpoteesi kehtides jaotuma ligikaudu ühe vabadusastmega χ^2-jaotuse järgi, saame siit leida ka lookuste A ja B ahelduse statistilist olulisust näitava p-väärtuse: $\mathrm{P}(\chi^2_1 > 1,646) = 0,200$. Muidugi ei kehti nii väikese andmestiku korral teststatistiku asümptootilised omadused, mistõttu ei pruugi ka χ^2-jaotuse baasil leitud p-väärtus olla korrektne.

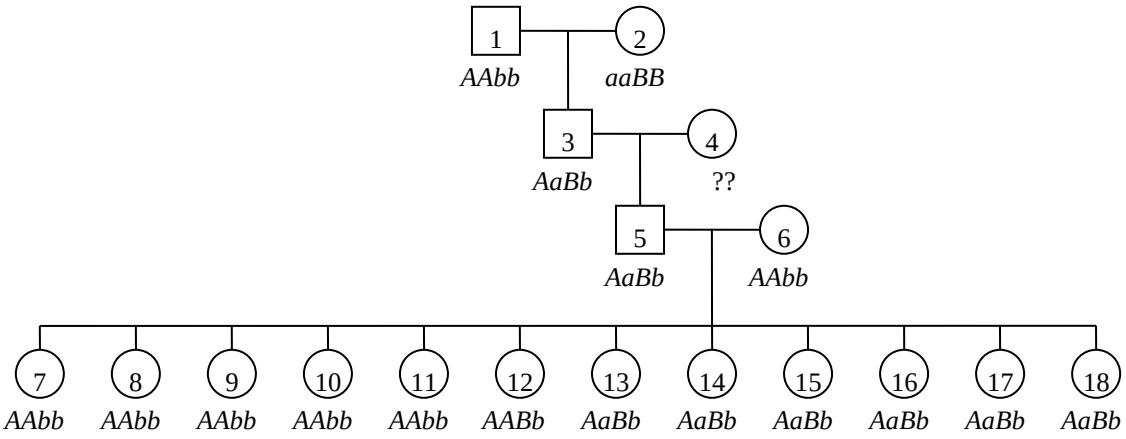
Alternatiivina kasutatakse tänapäeval ka tõepärasuhte statistiku nullhüpoteesile vastava jaotusega võrdlemiseks permutatsioonitesti või selle ligikaudseid lähendeid, nn Monte Carlo meetodeid.

Antud andmetele vastav *LOD*-skoori väärtus tuleb

$$\mathrm{LOD}(r) = \left[\log_{10} L(\hat{r}) - \log_{10} L(\tfrac{1}{2})\right] = 0{,}357 .$$

Ülesanne 5.

Võtame vaatluse alla järgneva sugupuu (sigade oma). Sugupuu alumises osas on kahekordselt heterosügootne kult (5) paaritatud kahekordselt homosügootse emisega (6), tulemusena on sündinud 12 põrsast, kes on kõik ka genotüpiseeritud.



Arvutage tõepärafunktsiooni $L(r; n, k)$ väärtus põrsaste genotüüpide tarvis, võttes rekombinatsioonimäära r väärtusteks 0; 0,01; 0,1; 0,2 ja 0,5. Leidke ka vastavad tõepärasuhte statistiku $\Lambda(r)$ ja LOD-skoori $\mathrm{LOD}(r)$ väärtused.

Arvutused teostage järgimistel tingimustel.

(a) Eeldage, et kuldi 5 järjestatud genotüüp on kujul *Ab|aB*.
Arvutage selle juhu jaoks ka rekombinatsioonimäära suurima tõepära hinnang.

(b) Eeldage, et kuldi 5 genotüübi haplotüüpideks jagunemine (ahelduse faas) ei ole teada (eirake kuldi 5 vanemaid).
Arvutage täiendavalt (järglaste genotüüpide alusel) kuldi 5 järjestatud genotüüpide *Ab|aB* ja *AB|ab* tõenäosused, võttes rekombinatsioonimääraks $r = 0{,}4$.

(c) Kasutage kogu sugupuu informatsiooni, eeldades täiendavalt, et kuldi 5 mitteteadaolev ema 4 on valitud juhuslikult populatsioonist, kus $\mathrm{P}(A) = 0{,}8$ ja $\mathrm{P}(B) = 0{,}2$.

3.2.3 Tõepärafunktsioon sugupuu-andmete korral

Tõepärafunktsioon n indiviidist koosneva sugupuu tarvis väljendab tõenäosust saada vaadeldud fenotüübiväärtused $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ andmete kirjeldamiseks valitud mudeli korral. Mudel eeldatakse sõltuvat parameetrite vektorist $\boldsymbol{\theta}$ (mis lihtsamal juhul sisaldab vaid rekombinatsioonimäära r ning alleelide sagedusi marker- ja haiguslookuses) ning penetrantsuse parameetrite vektorist $\mathbf{f}$ (viimase elemendid näitavad, kuidas mingi genotüüp fenotüübiväärtusena avaldub).

Eeldame, et sugupuusse kuuluvate indiviidide fenotüübiväärtused on täielikult määratud nende genotüüpidega, st et

$$\mathrm{P}(\mathbf{y}\,|\,\mathbf{g}) = \prod_{i=1}^{n} \mathrm{P}(y_i\,|\,g_i), \tag{3.1}$$

kus $g_i = (m_i, d_i)$ on i.-nda indiviidi markerlookuse ja hüpoteetilise haiguslookuse ühisgenotüüp, milles omakorda $m_i = (m_{i_1}, m_{i_2})$ on markergenotüüp ja $d_i = (d_{i_1}, d_{i_2})$ on genotüüp haiguslookuses.

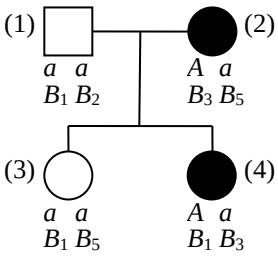
Parameetervektori $\boldsymbol{\theta}$ tõepärafunktsioon on siis esitatav kujul

$$L(\boldsymbol{\theta}) = \mathrm{P}(\mathbf{y}\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{g}} \mathrm{P}(\mathbf{y},\mathbf{g}\,|\,\boldsymbol{\theta}) = \sum_{\mathbf{g}} \mathrm{P}(\mathbf{y}\,|\,\mathbf{g})\,\mathrm{P}(\mathbf{g}\,|\,\boldsymbol{\theta}), \tag{3.2}$$

kus summeerimine toimub üle kõigi andmetega sobivate ning ahelduse faasi arvestavate marker- ja haiguslookuse ühisgenotüüpide vektorite **g**.

Näiteks joonisel 17 kujutatud sugupuu korral on marker- ja haiguslookuse (ahelduse faasi arvestavad) järjestatud ühisgenotüübid nelja indiviidi tarvis järgmised:

isa (1)	$g_1 = (m_1, d_1)$:	$aB_1\|aB_2$,
ema (2)	$g_2 = (m_2, d_2)$:	$aB_5\|AB_3$ või $aB_3\|AB_5$,
esimene tütar (3)	$g_3 = (m_3, d_3)$:	$aB_1\|AB_5$,
teine tütar (4)	$g_4 = (m_4, d_4)$:	$aB_1\|AB_3$.



Joonis 17. Sugupuu nr 1 – perekond kahe haige indiviidiga (täidetud sümbolid). Hüpoteetiline dominantne haigusalleel on tähistatud tähega *A*.

Isa ja esimene tütar on mõlemad haiguslookuse osas homosügootsed, mistap on nende marker- ja haiguslookuse ühisgenotüübid üheselt tuvastatavad. Seevastu ema korral, kes on heterosügootne nii haigus- kui ka markerlookuses, on võimalikud kaks erinevat marker- ja haiguslookuse ühisgenotüüpi. Teine tütar on samuti mõlemas lookuses heterosügootne, aga tänu vanematele on tema ahelduse faas üheselt tuvastatav, mistõttu vastab talle vaid üks andmetega sobiv ahelduse faasi arvestav marker- ja haiguslookuse ühisgenotüüp.

Seega vastab antud andmetele kaks võimalikku marker- ja haiguslookuse ühisgenotüüpide vektorit **g** (üle mille toimub summeerimine avaldises (3.2)):

$\boldsymbol{g}_1$: ($aB_1|aB_2$, $aB_5|AB_3$, $aB_1|AB_5$, $aB_1|AB_3$),

$\boldsymbol{g}_2$: ($aB_1|aB_2$, $aB_3|AB_5$, $aB_1|AB_5$, $aB_1|AB_3$).

- Võrduse (3.2) paremal poolel oleva summa esimene liidetav, $\mathrm{P}(\mathbf{y}|\mathbf{g})$, sõltub vaid penetrantsuse parameetritest, mida dialleelse haiguslookuse korral on kolm: $\mathbf{f} = (f_1, f_2, f_3)$, ning kus sagedused f_j näitavad, kui tõenäoliselt on *j*.-nda haigusgenotüübiga indiviid haige.

Näiteks eeldades joonisel 17 kujutatud sugupuu korral haigusalleeli täielikku dominantsust – indiviidid genotüüpidega *AA* ja *Aa* on haiged ja indiviidid genotüübiga *aa* on terved –, on penetrantsuse parameetreid sisaldav vektor kujul $\mathbf{f} = (1, 1, 0)$. Kuna kõigi sugupuusse kuuluvate indiviidide korral on nende fenotüüp genotüübi alusel üheselt määratud, siis võrduvad korrutise (3.1) kõik liikmed ühega: $\mathrm{P}(y_i|g_i)=1$ iga *i* korral, ning tõepärafunktsioon (3.2) sõltub vaid suurustest $\mathrm{P}(\mathbf{g}|\boldsymbol{\theta})$.

- Summa (3.2) teine liidetav, $\mathrm{P}(\mathbf{g}|\boldsymbol{\theta})$, sõltub rekombinatsioonimäärast *r*, haigust põhjustava alleeli (näiteks *A*) sagedusest $p = \mathrm{P}(A)$ ja markeralleelide sagedustest $\mathbf{p}_M$.

Tõenäosuse $\mathrm{P}(\mathbf{g}|\boldsymbol{\theta})$ avaldamiseks märgime esmalt, et mingi andmetega sobiva ja ahelduse faasi arvestava marker- ja haiguslookuse ühisgenotüüpide vektori **g** tõenäosus avaldub sugupuusse kuuluvate indiviidide ühisgenotüüpide kaudu kujul

$$\mathrm{P}(\mathbf{g}) = \mathrm{P}(g_1 \cap g_2 \cap \ldots \cap g_n),$$

mis vastavalt tingliku tõenäosuse definitsioonile[2] on omakorda avaldatav korrutisena

$$\mathrm{P}(\mathbf{g}) = \mathrm{P}(g_1)\mathrm{P}(g_2|g_1)\mathrm{P}(g_3|g_1,g_2)\ldots\mathrm{P}(g_n|g_1,\ldots,g_{n-1}). \qquad (3.3)$$

Viimane avaldis on edasi lihtsustatav arvestades, et iga indiviidi *i* genotüüp g_i sõltub vaid tema isa ja ema genotüüpidest g_{Fi} ja g_{Mi}.

Baas- ehk **asutajaindiviidid** (indiviidid, kelle vanemad on teadmata; ingl. *founders*, *base individuals*) eeldatakse suguluses mitteolevaiks, mistap loetakse nende genotüübid üksteisest sõltumatuks ja määratuiks üksnes alleelisageduste läbi populatsioonis.

Jagades kõik sugupuusse kuuluvad *n* indiviidi baasindiviididdeks (*F*) ja mitte-baasindiviididdeks (ingl. *non-founders*, *NF*), on tõenäosus (3.3) esitatav kahe korrutisena:

$$\mathrm{P}(\mathbf{g}) = \prod_{i\in F}\mathrm{P}(g_i)\prod_{j\in NF}\mathrm{P}(g_j|g_{Fj},g_{Mj}).$$

[2] $\mathrm{P}(A\cap B) = \mathrm{P}(A)\mathrm{P}(B|A)$

Baasindiviidide kohta eeldatakse täiendavalt, et nende mistahes eri lookustes (seega ka marker-ja haiguslookuses) paiknevad alleelid on tasakaalulises ahelduses[3]. Sellisel juhul on marker- ja haiguslookuse ühistõenäosus kirjutatav kujul

$$\mathrm{P}(g_i)=\mathrm{P}((m_i,d_i)|\mathbf{p}_M,p)=\mathrm{P}(m_i|\mathbf{p}_M)\mathrm{P}(d_i|p),$$

siin m_i ja d_i märgivad i.-nda indiviidi markergenotüüpi ja genotüüpi haiguslookuses, st $g_i=(m_i,d_i)=(m_{i1},m_{i2},d_{i1},d_{i2})$, $\mathbf{p}_M$ on markeralleelide sageduste vektor ja $p = \mathrm{P}(A)$ on haigusalleeli sagedus. Lisaks eeldatakse baasindiviidide alleeli- ja genotüübisageduste vastamist Hardy-Weinbergi tasakaalule.

Indiviidide, kelle vanemad on teada, genotüübisagedus sõltub nende vanemate genotüüpidest ning marker- ja haiguslookuse vahelisest rekombinatsioonimäärast:

$$\mathrm{P}(g_j|g_{Fj},g_{Mj})=\mathrm{P}(m_i,d_i|m_{Fi},m_{Mi},d_{Fi},d_{Mi},r).$$

Kokkuvõttes avaldub tõepärafunktsioon (3.2) kujul

$$L(\boldsymbol{\theta})=\sum_{\mathbf{g}}\left[\prod_{i=1}^{n}\mathrm{P}(y_i|g_i,\mathbf{f})\prod_{i\in F}\mathrm{P}(g_i)\prod_{j\in NF}\mathrm{P}(g_j|g_{Fj},g_{Mj})\right]. \qquad (3.4)$$

Pöördume tagasi joonisel 17 kujutatud sugupuu juurde. Tähistame markeralleelide sagedused $p_j=\mathrm{P}(m=B_j)$ (üleskirjutuste lihtsuse huvides jätame üksikute markeralleelide sageduste tähistest ära alaindeksi M, st $p_j=p_{Mj}$) ning haigusalleeli sageduse $\mathrm{P}(A) = p$ ja normaalse alleeli sageduse $\mathrm{P}(a) = q = 1 – p$.

Isa ja ema kui baasindiviidide marker- ja haiguslookuse ühisgenotüüpide tõenäosused avalduvad tänu marker- ja haiguslookuse eeldatavalt tasakaalulisele aheldusele ning Hardy-Weinbergi seadusele järgnevalt:

isa (1): $\mathrm{P}(g_1) = \mathrm{P}(aaB_1B_2) = \mathrm{P}(aa)\times\mathrm{P}(B_1B_2) = q^2\times 2p_1p_2$,

ema (2): $\mathrm{P}(g_2) = \mathrm{P}(AaB_3B_5) = \mathrm{P}(Aa)\times\mathrm{P}(B_3B_5) = 2pq\times 2p_3p_5$.

Tütarde marker- ja haiguslookuse ühisgenotüüpide tõenäosused sõltuvad nende vanemate järjestatud genotüüpidest ja rekombinatsioonimäärast ning avalduvad järgnevalt:

$\mathbf{g}_1$: $\mathrm{P}(g_3|g_1,g_2)=\frac{1}{2}\times\frac{1-r}{2}$, $\mathrm{P}(g_4|g_1,g_2)=\frac{1}{2}\times\frac{1-r}{2}$;

$\mathbf{g}_2$: $\mathrm{P}(g_3|g_1,g_2)=\frac{1}{2}\times\frac{r}{2}$, $\mathrm{P}(g_4|g_1,g_2)=\frac{1}{2}\times\frac{r}{2}$.

Kõigis tõenäosustes näitab esimene tegur ½ tõenäosust pärida isalt haplotüüp aB_1 (isalt saavad tütardele päranduda haplotüübid aB_1 ja aB_2, mõlemad tõenäosusega ½). Teine tegur vastab pärandumisele emalt tütrele. Näiteks marker- ja haiguslookuse ühisgenotüüpide vektori $\mathbf{g}_1$ korral on ema järjestatud genotüüp kujul $aB_5|AB_3$, esimene tütar (3) sai emalt haplotüübi aB_5 ja vastava sündmuse tõenäosus avaldub kujul:

$$\mathrm{P}(aB_5|\mathbf{g}_1) = \mathrm{P}(a|\mathbf{g}_1)\,\mathrm{P}(B_5|a,\mathbf{g}_1) = \frac{1}{2}\times(1-r).$$

Tõepärafunktsioon (3.4) avaldub joonisel 17 esitatud sugupuu korral kujul

$$\begin{aligned}L(\boldsymbol{\theta}) &= \sum_{\mathbf{g}\in\{\mathbf{g}_1,\mathbf{g}_2\}}\left[\prod_{i=1}^{2}\mathrm{P}(g_i)\prod_{j=3}^{4}\mathrm{P}(g_j|g_1,g_2)\right]\\ &=(q^2\times 2p_1p_2)\times(2pq\times 2p_3p_5)\times(\tfrac{1}{2}\times\tfrac{1-r}{2})\times(\tfrac{1}{2}\times\tfrac{1-r}{2})\\ &\quad+(q^2\times 2p_1p_2)\times(2pq\times 2p_3p_5)\times(\tfrac{1}{2}\times\tfrac{r}{2})\times(\tfrac{1}{2}\times\tfrac{r}{2})\\ &=C\times[(1-r)^2+r^2],\end{aligned}$$

kus $C=\frac{1}{2}q^3p\times p_1p_2p_3p_5$.

Kuna $L(r=0)=C$ ja $L(r=½)=C/2$, on *LOD*-skoor kohal r = 0:

[3] Markeralleel M_i ja haigust põhjustav alleel D_j on tasakaalulises ahelduses, kui haplotüübi M_iD_j sagedus võrdub alleelide M_i ja D_j sageduste korrutisega: $\mathrm{P}(M_iD_j) = \mathrm{P}(M_i)\mathrm{P}(D_j)$.

$$LOD(r=0) = \log_{10}\left(\frac{L(0)}{L(\tfrac{1}{2})}\right) = \log_{10}(2) = 0{,}3\,.$$

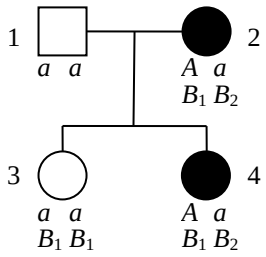
3.2.4. Tõepärafunktsioon puuduvate marker-andmete korral

Võtame vaatluse alla joonisel 18 kujutatud sugupuu ning eeldame nagu eelnevaski näites, et uuritav tunnus (haigus) on täielikult määratud ühe dominantse alleeli *A* poolt. Ema ja kaks tütart on genotüpiseeritud markerlookuse *B* osas, aga isa (indiviid 1) on mingil põhjusel jäänud genotüpiseerimata. Et isa on terve, peab ta eelduste kohaselt olema haiguslookuse osas homosügootne genotüübiga *aa*. Lisaks järeldub esimese tütre (indiviidi 3) homosügootsusest genotüübiga $aB_1|aB_1$, et isal peab olema vähemalt üks alleel B_1. Teist isa genotüüpi kuuluvat markeralleeli ei ole antud andmete korral võimalik üheselt tuvastada, see võib olla nii alleel B_1, B_2 või hoopis mingi mistahes kolmas alleel B_w. Taolisel puuduvate marker-andmete juhul tuleb lisaks tinglikustamisele ema (2) genotüübi haplotüüpideks jagunemise suhtes arvestada ka kõigi võimalike isa genotüüpidega.

Andmetega ja püstitatud geneetilise mudeliga sobivaid ahelduse faasi arvestavaid marker- ja haiguslookuse järjestatud ühisgenotüüpe on antud juhul kaheksa. Isa puhul tuleb arvestada kolme erineva genotüübiga: $aB_1|aB_1$, $aB_1|aB_2$ või $aB_1|aB_w$, ema korral on võimalikud kaks varianti: $AB_1|aB_2$ või $aB_1|AB_2$. Juhul, kui isal ei ole alleeli B_2, on tütarde genotüübid $aB_1|aB_1$ ja $aB_1|AB_2$, aga kui isal on markeralleel B_2, võib haige tütre (4) genotüüp haplotüüpideks jaguneda kahel viisil: $aB_1|AB_2$ või $AB_1|aB_2$.

Kokkuvõttes on kaheksa võimalikku marker- ja haiguslookuse ühisgenotüüpide vektorit **g** (üle mille toimub summeerimine avaldises (3.2)) järgmised:

$\mathbf{g}_1$: ($aB_1|aB_1$, $AB_1|aB_2$, $aB_1|aB_1$, $aB_1|AB_2$)
$\mathbf{g}_2$: ($aB_1|aB_2$, $AB_1|aB_2$, $aB_1|aB_1$, $aB_1|AB_2$)
$\mathbf{g}_3$: ($aB_1|aB_2$, $AB_1|aB_2$, $aB_1|aB_1$, $AB_1|aB_2$)
$\mathbf{g}_4$: ($aB_1|aB_w$, $AB_1|aB_2$, $aB_1|aB_1$, $aB_1|AB_2$)
$\mathbf{g}_5$: ($aB_1|aB_1$, $aB_1|AB_2$, $aB_1|aB_1$, $aB_1|AB_2$)
$\mathbf{g}_6$: ($aB_1|aB_2$, $aB_1|AB_2$, $aB_1|aB_1$, $aB_1|AB_2$)
$\mathbf{g}_7$: ($aB_1|aB_2$, $aB_1|AB_2$, $aB_1|aB_1$, $AB_1|aB_2$)
$\mathbf{g}_8$: ($aB_1|aB_w$, $aB_1|AB_2$, $aB_1|aB_1$, $aB_1|AB_2$)



Joonis 18. Sugupuu nr 2 – perekond kahe haige indiviidiga, isa genotüpiseerimata. Hüpoteetiline dominantne haigusalleel on tähistatud tähega *A*.

Tõepärafunktsioon (3.2) on summa kaheksast korrutisest, kus iga korrutis sisaldab teguritena kahte baasindiviididele (isa ja ema) ja kahte mitte-baasindiviididele vastavat tõenäosust.

Isa ja ema kui baasindiviidide marker- ja haiguslookuse ühisgenotüüpide sagedused avalduvad sarnaselt eelmisele punktile, vahe on vaid võimalike variantide arvus:

isa (1): $$\mathrm{P}(g_1) = \begin{cases} \mathrm{P}(aaB_1B_1) = \mathrm{P}(aa)\times\mathrm{P}(B_1B_1) = q^2\times p_1^2, & \mathbf{g}_1 \text{ ja } \mathbf{g}_5 \text{ korral,} \\ \mathrm{P}(aaB_1B_2) = \mathrm{P}(aa)\times\mathrm{P}(B_1B_2) = q^2\times 2p_1p_2, & \mathbf{g}_2, \mathbf{g}_3, \mathbf{g}_6 \text{ ja } \mathbf{g}_7 \text{ korral,} \\ \mathrm{P}(aaB_1B_w) = \mathrm{P}(aa)\times\mathrm{P}(B_1B_w) = q^2\times 2p_1p_w, & \mathbf{g}_4 \text{ ja } \mathbf{g}_8 \text{ korral;} \end{cases}$$

ema (2): $$\mathrm{P}(g_2) = \mathrm{P}(AaB_1B_2) = \mathrm{P}(Aa)\times\mathrm{P}(B_1B_2) = 2pq\times 2p_1p_2,\ \mathbf{g}_1, \mathbf{g}_2, \ldots, \mathbf{g}_8 \text{ korral}\,.$$

Esimene tütar (3) pärib $\mathbf{g}_1$ ja $\mathbf{g}_5$ korral tõenäosusega 1 isalt haplotüübi aB_1 (isa on homosügootne), ülejäänud juhtudel (isa on siis heterosügootne) on sama haplotüübi esimesele tütrele pärandumise tõenäosus ½.

Tõenäosus, et esimene tütar pärib emalt haplotüübi aB_1, avaldub ema genotüübi $AB_1|aB_2$ korral ($\mathbf{g}_1$, $\mathbf{g}_2$, $\mathbf{g}_3$ ja $\mathbf{g}_4$ korral) kujul

$$\mathrm{P}(aB_1) = \mathrm{P}(a)\,\mathrm{P}(B_1 \mid a) = \tfrac{1}{2}\times r\,,$$

ning ülejäänud juhtudel (ema genotüüp on $aB_1|AB_2$) kujul

$$\mathrm{P}(aB_1) = \mathrm{P}(a)\,\mathrm{P}(B_1 \mid a) = \tfrac{1}{2}\times(1-r)\,.$$

Analoogse aruteluga jätkates on leitavad kõik mitte baasindiviididele (kahele tütrele) vastavad tõenäosused:

esimene tütar (3): $\mathrm{P}(g_3 \mid g_1, g_2) = \begin{cases} 1\times\frac{r}{2}, & \mathbf{g}_1 \text{ korral,} \\ \frac{1}{2}\times\frac{r}{2}, & \mathbf{g}_2, \mathbf{g}_3, \text{ ja } \mathbf{g}_4 \text{ korral,} \\ 1\times\frac{(1-r)}{2}, & \mathbf{g}_5 \text{ korral,} \\ \frac{1}{2}\times\frac{(1-r)}{2}, & \mathbf{g}_6, \mathbf{g}_7, \text{ ja } \mathbf{g}_8 \text{ korral;} \end{cases}$

teine tütar (4): $\mathrm{P}(g_4 \mid g_1, g_2) = \begin{cases} 1\times\frac{r}{2}, & \mathbf{g}_1 \text{ korral,} \\ \frac{1}{2}\times\frac{r}{2}, & \mathbf{g}_2, \mathbf{g}_4, \text{ ja } \mathbf{g}_7 \text{ korral,} \\ \frac{1}{2}\times\frac{(1-r)}{2}, & \mathbf{g}_3, \mathbf{g}_6, \text{ ja } \mathbf{g}_8 \text{ korral,} \\ 1\times\frac{(1-r)}{2}, & \mathbf{g}_5 \text{ korral.} \end{cases}$

Leitud tõenäosuste kaudu on nüüd viimaks kirja pandav ka tõepärafunktsioon (3.2):

$$\begin{aligned} L(\boldsymbol{\theta}) &= \sum_{\mathbf{g}\in\{\mathbf{g}_1,\ldots,\mathbf{g}_8\}} [\mathrm{P}(g_1)\times\mathrm{P}(g_2)\times\mathrm{P}(g_3 \mid g_1, g_2)\times\mathrm{P}(g_4 \mid g_1, g_2)] \\ &= p_1q^2\times 4pqp_1p_2\times\left[p_1\tfrac{r^2}{4} + 2p_2\left(\tfrac{r^2}{16} + \tfrac{r(1-r)}{16}\right) + 2p_w\tfrac{r^2}{16} + p_1\tfrac{(1-r)^2}{4} + 2p_2\left(\tfrac{(1-r)^2}{16} + \tfrac{r(1-r)}{16}\right) + 2p_w\tfrac{(1-r)^2}{16}\right] \\ &= p_1q^2\times 4pqp_1p_2\times\left[\left(p_1 + \tfrac{p_w}{2}\right)\left(r^2 + (1-r)^2\right) + \tfrac{p_2}{2}\right]. \end{aligned}$$

Et kõigi võimalike marker- ja haiguslookuse ühisgenotüüpide vektorite $\mathbf{g}_i$ korral sisaldavad summeeritavad suurused sama emale (2) vastavat komponenti $\mathrm{P}(g_2) = 4pqp_1p_2$ ning kõik isale (1) vastavad komponendid ühist tegurit p_1q^2, jäetakse geneetika raamatutes need tegurid enamasti kirjutamata, sest *LOD*-skoori arvutamisel taanduvad nad nagunii ära. Seega võib tõepärafunktsiooni esitada ka kujul:

$$L(\boldsymbol{\theta}) \propto \left(p_1 + \tfrac{p_w}{2}\right)\left(r^2 + (1-r)^2\right) + \tfrac{p_2}{2}\,.$$

LOD-skoor kohal $r = 0$ on

$$LOD(r=0) = \log_{10}\left(\frac{L(0)}{L(\tfrac{1}{2})}\right) = \log_{10}\left(\frac{p_1 + \frac{(p_2+p_w)}{2}}{\frac{p_1+p_2}{2} + \frac{p_w}{4}}\right). \tag{3.5}$$

Tabelis 3 on toodud valemiga (3.5) esitatud *LOD*-skoori väärtused mõningate alleelisageduste korral. *LOD*-skoori väärtus on maksimaalne, kui markeralleeli B_2 sagedus on null – see on üsna loomulik tulemus, sest siis on teise tütre (4) ahelduse faas üheselt teada ($aB_1|AB_2$, emalt on pärandunud haplotüüp AB_2) ning situatsioon on analoogne eelmises näites vaadeldule. Teine ekstreemne situatsioon on markeralleeli B_2 sageduse lähenemine ühele – sellisel ei ole vaadeldav perekond aheldusanalüüsi seisukohast informatiivne (võimatu on öelda, kas haige tütar päris koos haigusalleeliga *A* markeralleeli B_1 või markeralleeli B_2), millele vihjab ka nullilähedane *LOD*-skoori väärtus.

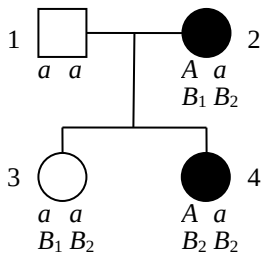
Tabel 3. Valemiga (3.5) esitatud *LOD*-skoori väärtused mõningate alleelisageduste korral.

p_1	p_2	p_w	$LOD(0)$
0,50	0,00	0,50	0,301
0,45	0,10	0,45	0,272
0,30	0,40	0,30	0,185
0,10	0,80	0,10	0,064
0,01	0,99	0,00	0,004
0,50	0,50	0,00	0,176
0,25	0,50	0,25	0,155
0,10	0,50	0,40	0,138

Ülesanne 6.

Võtame vaatluse alla kahe haige indiviidiga perekonna, kus isa ei ole genotüpiseeritud. Eeldame, et haigus on põhjustatud täielikult dominantse alleeli *A* poolt (kõik *Aa* ja *AA* indiviidid on haiged) ning haiged indiviidid on haiguslookuse osas heterosügootsed.

Tuletage tõepärafunktsiooni

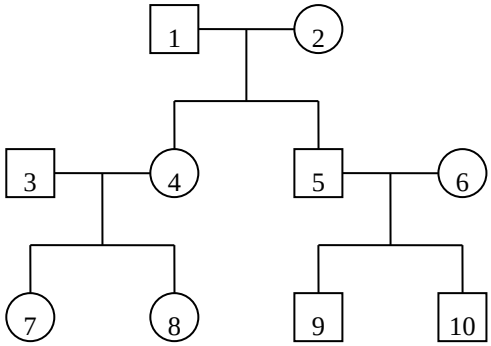


$$L(\boldsymbol{\theta}) = \sum_{\mathbf{g}} \prod_{i=1}^{n} \mathrm{P}(y_i \mid g_i, \mathbf{f}) \prod_{i \in F} \mathrm{P}(g_i) \prod_{j \in NF} \mathrm{P}(g_j \mid g_{F_j}, g_{M_j})$$

avaldis toodud sugupuu tarvis ning leidke *LOD*-skoori väärtus kohal $r = 0$.

3.2.5 Klassikaline aheldusanalüüs kompleksse sugupuu korral – Elston-Stewart'i algoritm

Võtame vaatluse alla järgmise sugupuu:



$$L(\boldsymbol{\theta}) = \sum_{g_1} \sum_{g_2} \cdots \sum_{g_n} \mathrm{P}(\mathbf{y} \mid \mathbf{g}) \mathrm{P}(\mathbf{g} \mid \boldsymbol{\theta})$$

$$\begin{aligned} \mathrm{P}(y_3, y_7, y_8 \mid g_4) &= \sum_{g_3} \sum_{g_7} \sum_{g_8} \mathrm{P}(y_3, y_7, y_8, g_3, g_7, g_8 \mid g_4) \\ &= \sum_{g_3} \sum_{g_7} \sum_{g_8} [\mathrm{P}(y_3 \mid g_3) \mathrm{P}(y_7 \mid g_7) \mathrm{P}(y_8 \mid g_8) \mathrm{P}(g_7 \mid g_3, g_4) \mathrm{P}(g_8 \mid g_3, g_4) \mathrm{P}(g_3)]. \end{aligned}$$

$$\begin{aligned} \mathrm{P}(y_6, y_9, y_{10} \mid g_5) &= \sum_{g_6} \sum_{g_9} \sum_{g_{10}} \mathrm{P}(y_6, y_9, y_{10}, g_6, g_9, g_{10} \mid g_5) \\ &= \sum_{g_6} \sum_{g_9} \sum_{g_{10}} [\mathrm{P}(y_6 \mid g_6) \mathrm{P}(y_9 \mid g_9) \mathrm{P}(y_{10} \mid g_{10}) \mathrm{P}(g_9 \mid g_5, g_6) \mathrm{P}(g_{10} \mid g_5, g_6) \mathrm{P}(g_6)]. \end{aligned}$$

$$\begin{aligned} \mathrm{P}(y_3, \ldots, y_{10} \mid g_1, g_2) &= \left[\sum_{g_4} \mathrm{P}(y_3, y_7, y_8 \mid g_4) \mathrm{P}(y_4 \mid g_4) \mathrm{P}(g_4 \mid g_1, g_2)\right] \\ &\quad \times \left[\sum_{g_5} \mathrm{P}(y_6, y_9, y_{10} \mid g_5) \mathrm{P}(y_5 \mid g_5) \mathrm{P}(g_5 \mid g_1, g_2)\right] \end{aligned}$$

$$\mathrm{P}(y_1, \ldots, y_{10}) = \sum_{g_1} \sum_{g_2} \mathrm{P}(y_1 \mid g_1) \mathrm{P}(y_2 \mid g_2) \mathrm{P}(y_3, \ldots, y_{10} \mid g_1, g_2) \mathrm{P}(g_1) \mathrm{P}(g_2)$$