TARTU ÜLIKOOL
LOODUS- JA TÄPPISTEADUSTE VALDKOND
Matemaatika ja statistika instituut

Anna Laaneväli

Statistilised mudelid bakterite segude määramiseks

Matemaatika ja statistika õppekava
Matemaatilise statistika eriala
Magistritöö (30 EAP)

Juhendaja: Märt Möls, PhD

Tartu 2020

Statistilised mudelid bakterite segude määramiseks

Magistritöö

Anna Laaneväli

Lühikokkuvõte. Proovis sisalduva bakterite DNA sekveneerimisel on võimalik öelda, milliste bakterite DNAd proovis nähti ning mis koguses.
Sekveneerimisel saadud andmeid on võimalik analüüsida vähimruutude, mittenegatiivsete vähimruutude või suurima tõepära meetodit kasutades, tuvastamaks bakterite tüvesid ning hindamaks nende k-meeride katvust. Magistritöö eesmärgiks on võrrelda kolme nimetatud meetodi käitumist sekveneerimisvigade olemasolu korral, vastavalt vigadega arvestades kui arvestamata. Vaatlusaluseid meetodeid rakendatakse ka ühe reaalse proovi sekveneerimisel kogutud andmete (saadud Tartu Ülikooli molekulaar- ja rakubioloogia instituudist) analüüsil.
CERCS teaduseriala:
B110 Bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika;
P160 Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.
B220 Geneetika, tsütogeneetika;
Märksõnad: hindamine, vähimruutude meetod, mittenegatiivne vähimruutude meetod, suurima tõepära meetod, DNA, k-meerid, sekveneerimine, R (programmeerimiskeel)

Statistical models for determining bacterial mixture

Master's thesis

Anna Laaneväli

Abstract. Bacterial DNA sequencing enables us to detect bacterial isolates and to estimate its abundance in the sample. The least squares method, non-negative least squares and maximum likelihood estimation can be used to analyze sequencing data, to identify bacterial isolates and to estimate their k-mer abundance. The aim of this thesis is to compare these methods by observing them on data set with sequencing errors which accordingly have been taken into consideration or not. Afterwards, the methods are applied to an actual DNA sequencing data set from the Institute of Molecular and Cell Biology, University of Tartu.

CERCS research specialisation:
B110 Bioinformatics, medical informatics, biomathematics, biometrics;
P160 Statistics, operation research, programming, actuarial mathematics.
B220 Genetics, cytogenetics; **Keywords**: estimation, least squares method, non-negative least squares method, maximum likelihood method, DNA, k-mers, sequencing, R (programming language)

Sisukord

Sissejuhatus

Patogeenseteks nimetatakse baktereid, mis patsiendi organismi tungides põhjustavad haigusi. Enamik neist bakteritest moodustavad toksiine, mis kutsuvad esile koekahjustusi. Määramaks patsiendile sobivat ravi, tuleb enne kindlaks teha, millised bakterid organismi vaevavad.

Bakterite olemasolu hindamiseks ning nende liikide määramiseks kasutatakse ühe variandina bakterite kasvatamist selektiivsetel söötmetel. Teine levinud variant on kasutada ELISA või PCR testi, mis võimaldavad testida ühe või paari kahtlusaluse bakteritüve olemasolu proovis. Kui meil puudub eelinformatsioon proovis esineda võivate bakterite kohta, on nimetatud meetodid kahjuks üsna kasutud.

Tänapäeval saab bakterite määramiseks kasutada ka DNA sekveneerimisele tuginevaid meetodeid - lähenemine, mida proovitakse rakendada ka käesolevas magistritöös. Sekveneerimissageduste analüüsimisel on enamasti kasutatud vähimruutude meetodit. Bakterite osakaalude määramiseks sekveneeritavas proovis on kasutatud ka mittenegatiivset vähimruutude meetodit [10]. Lisaks proovib autor kasutada ka suurima tõepära meetodit. Hindamaks bakterite olemasolu ja osakaalu on vaadeldud antud magistritöös nimetatud kolme meetodit ning püütud leida, millisel meetodil on võimalik saada kõige täpsemad tulemused.

Töö esimeses peatükis kirjeldatakse andmete tausta, simulatsioonides kasutatud andmete genereerimist ning tuuakse vajalikke bioloogiamõisteid koos näidetega. Teises peatükis selgitatakse lähemalt probleemi olemust. Kolmandas peatükis on esitatud ülevaade antud töös kasutatavatest statistilistest meetoditest. Neljandas ja viiendas peatükis uuritakse sekveneerimisvigu sisaldavaid ja sekveneerimisvigadeta genereeritud andmeid kasutades selleks kolme eri meetodit. Kuuendas peatükis analüüsitakse reaalseid DNA sekveneerimisel saadud andmeid.

Magistritöö on vormistatud tekstitöötlusprogrammi LaTeX veebiversioonis Overleaf. Andmete analüüsimiseks, mudelite sobitamiseks ja jooniste koostamiseks on kasutatud statistikatarkvara R versiooni 3.6.1.

Autor tänab juhendajat Märt Mölsi toetuse, asjakohaste nõuannete, paranduste ning kasulike ideede eest. Ühtlasi tänab autor bioinformaatika nooremteadurit Mihkel Vaherit bakterite sekveneeritud DNA andmestiku jagamise, eeltöötluse ning asjakohase ülevaate andmise eest.

1 Töös kasutatavad andmed

Käesolev peatükk on kirjutatud kasutades [4], [8] ja [9] ning lähtudes Mihkel Vaheri ja Märt Mölsi ülevaadetest.

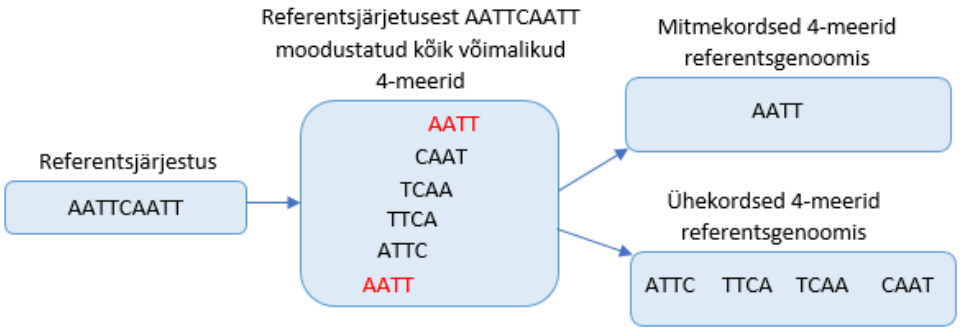
1.1 Andmete taustast

Bakterid on lihtsa ehitusega üherakulised mikroorganismid, kelle pärilikkusaineks on enamasti üks rõngaskromosoom. Mikroorganismid, sealhulgas ka bakterid, on kõige vanemad eluvormid Maal, kelle ilmumisaeg ulatub hinnanguliselt nelja miljardi aasta tagusesse aega. Bakterite laialdane levik on olnud võimalik nende kiire paljunemisvõime ja muutlikkuse tulemusena. Mikroobid paljunevad pooldudes, mis on mittesuguline paljunemine, mille käigus tekivad kaks uut ekvivalentset tütarrakku. Pooldumisprotsessi käigus, mis võtab soodsatel tingimustel aega 20–40 minutit [4], toimub rõngaskromosoomi replikatsioon nii, et tütarrakus on esialgse raku genoomi duplikaat.

Paljunemise käigus toimunud mutatsioonid mikroorganismide geenides võivad olla indutseeritud või tekkida spontaanselt, tekkesagedusega umbes 10^{-6}–10^{-9} nukleotiidi kohta [4]. Neid põhjustavad muutused DNA-d või RNA-d moodustavates nukleotiidides, kusjuures enamus mutatsioone on kahjulikud, mõningad neutraalsed ja väga väike osa kasulikud. Mutatsioonide ulatus võib olla küllaltki erinev, hõlmates üksikuid nukleotiidide paare, kuid ka suuri DNA lõike. Geenides toimuvad mutatsioonid loovad eelduse bakterite evolutsioneerumiseks. Kuna bakterid on üherakulised, siis avalduvad muutused neis kiiresti ning võivad tekkida uued (alam)liigid. Põhjalikult on uuritud, kui lähedalt on erinevad bakteriliigid omavahel suguluses ning kui sarnased on sama liigi muteerunud tüved.

Uurimaks bakterite olemasolu ja esinemissagedust võetakse valitud keskkonnast proov ning eraldatakse sellest bakterid. Seejärel eraldatakse bakteritest kogu genoomne DNA, mis sekveneeritakse kasutades Illumina sekveneerimisplatvormi (MiSeq, NextSeq500 või HiSeq2500). Tulemusena saadakse lugemid, mis on sekveneeritud piirkonnad genoomist pikkusega 100–150 aluspaari antud sekveneerimismeetodi puhul.

Bakteriliigi referentsjärjestus (kokkuleppeline antud liigi DNA kirjeldus) või sekveneerimisel saadud lugemid võib lõigata veelgi lühemateks k-tähe pikkusteks juppideks ehk k-meerideks. Käesolevas töös on vaadeldud *32*-meere ehk genoomi lõike pikkusega 32 aluspaari. Järgnev näide on saadud kasutades [8] tulemusi, kus vaadeldud on lihtsuse mõttes *4*-meere. Joonisel 1 on DNA referentsjärjestusest AATTCAATT moodustatud kõik võimalikud *4*-meerid.

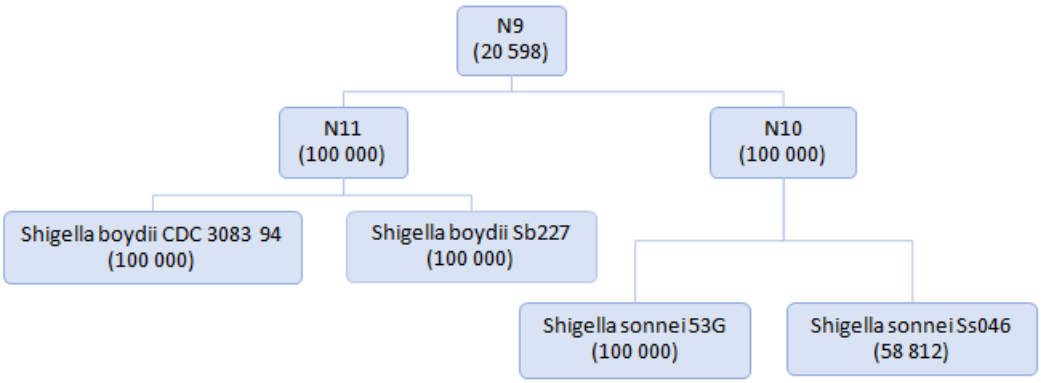


Joonis 1: DNA referentsjärjestusest AATTCAATT saadud 4-meerid, mis on jagatud kordsuse järgi ühe- ning mitmekordseteks 4-meerideks.

Toodud joonise 1 vasakul poolel on DNA järjestus jagatud kuueks 4-meeri kombinatsiooniks, kasutades toodud näites 4-nukleotiidist „libisevat akent". Iga 4-meer kattub eelneva ja järgneva 4-meeriga $4-1$ ehk 3 nukleotiidi ulatuses. Antud referentsjärjestuses esineb 4-meer AATT kahekordselt ning ülejäänud 4-meerid ühekordselt.

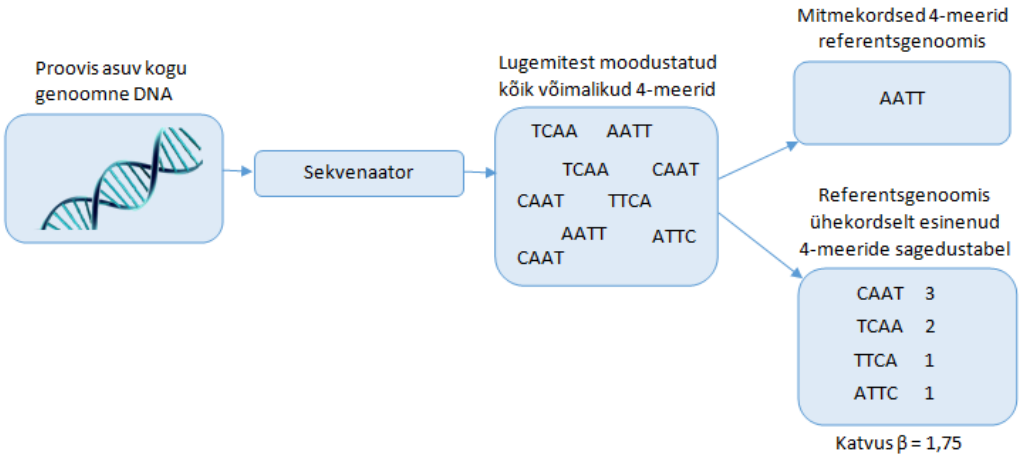
Referentsgenoomides nähtud k-meerid paigutatakse bakterite fülogeneesipuule. Fülogeneesipuu lehtedes on bakterile vastavate unikaalsete k-meeride arv ning igas sõlmes on sellele alluvate sõlmede või bakterite ühiste k-meeride arv. Unikaalseteks peame k-meere, mida teadaolevalt ei esine ühelgi teisel bakteriliigil või tüvel. Juhul, kui ühekordses korduses esinenud unikaalseid k-meere on rohkem kui 100 000, eemaldatakse üleliigsed juhuslikult eri programmide arvutuskiiruse huvides. Ühe bakteriliigi genoomis esinevad k-meerid on seega kirjas kas fülogeneesipuu lehes (unikaalsed k-meerid), sellele lehele eelnevates sõlmedes (ühekordsed k-meerid, mida võib kohata ka teistes sugulasliikides) või on üldse tabelist välja jäänud (pole ühekordsed, esinevad ka mittesugulas-liikidel või on välja jäänud 100 000 k-meeri ülempiiri tõttu).

Järgneval joonisel 2 on näitena toodud fülogeneesipuu, mille lehtedes on bakterite *Shigella boydii* ja *Shigella sonnei* kahe alamliigi unikaalsete 32-meeride arv. Mõlemad bakterid tekitavad düsenteeriat, mida teatakse ka shigelloosina. Sõlmes *N10* on 100 000 32-meeri, mis on ühised bakteri *Shigella sonnei* kahele alamliigile ning sõlmes *N11* on 100 000 32-meeri, mis on ühised bakteri *Shigella boydii* kahele alamliigile. Juures *N9* on 20 598 32-meeri, mida esineb nii sõlme *N10* kui sõlme *N11* kuuluvates bakterites.



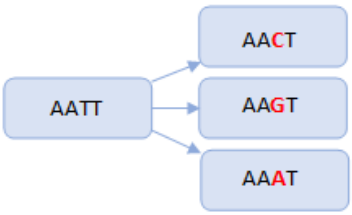
Joonis 2: Bakterite *S. Sonnei* ning *S. Boydii 32*-meeride fülogeneesipuu. Mõlemad bakterid tekitavad düsenteeriat, mida teatakse ka shigelloosina.

Sekveneerimisel loetakse bakteriproovis esinevat genoomi juhuslikest kohtadest alates. Ka sekveneerimisel saadud lugemid võime teisendada k-tähe pikkusteks k-meerideks. Bakteri genoomis võib vaadeldav k-meer esineda ühe- või mitmekordselt. Kuna sama kohta genoomis võime lugeda korduvalt (sama liigi baktereid võib olla proovis palju), siis võib üks esialgselt unikaalne k-meer proovis esineda mitmeid kordi. Bakteril ühekordselt esinevate k-meeride sageduste keskmist nimetatakse k-meeride katvuseks. Järgnev joonis 3 illustreerib olukorda, kus 4-meeride keskmine katvus $\beta = \frac{7}{4} = 1{,}75$.



Joonis 3: Proovis asuva genoomse DNA sekveneerimisel ühekordselt esinenud 4-meeride katvus $\beta = 1{,}75$.

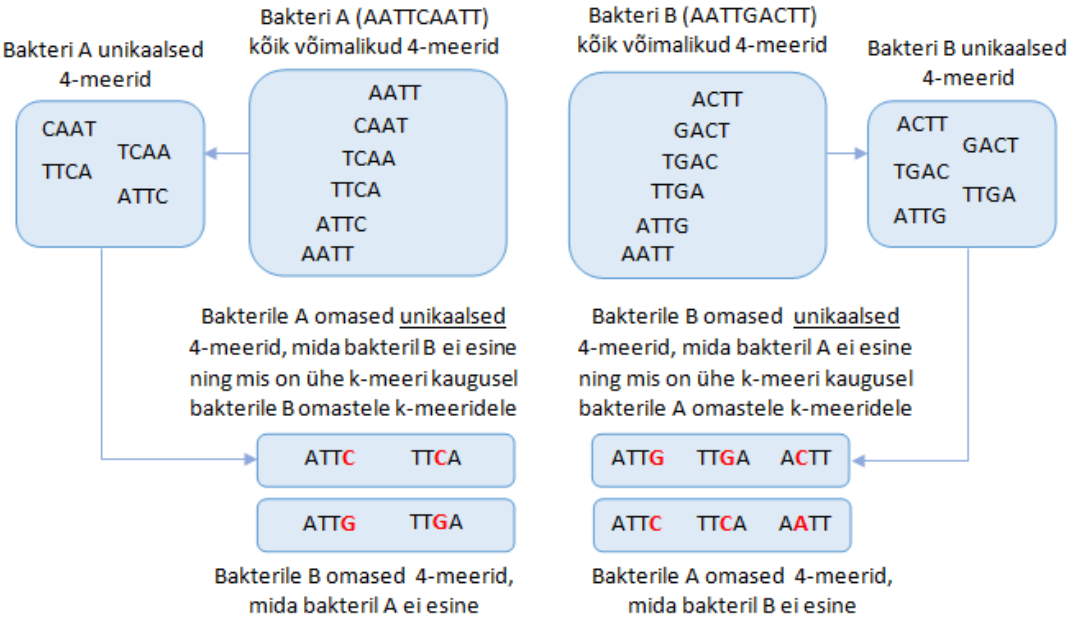
Tuleb arvestada ka võimalusega, et sekveneerimisel võib esineda vigu. Järgnevalt toome näite 4-meeri mutatsioonist ühe nukleotiidi ulatuses.



Joonis 4: K-meeri AATT kolmanda positsiooni lugemisel tehtava vea tõttu võime näha kolme erinevat k-meeri: AACT, AAGT ning AAAT. Toodud 4-meeri nukleotiidi mutatsioon on tähistatud punasega.

Joonisel 4 toodud k-meeri AATT eelviimase nukleotiidi T lugemisel tehtud vea tõttu saame uue k-meeri AACT. Vastaval fülogeneesipuul satub muteerunud k-meer AACT teise lehte (on ühe teise bakteriliigi unikaalne k-meer), tõstes lehes olevate bakterile omaste unikaalsete k-meeride arvu ühe võrra.

Sekveneerimisel tehtud võimalike vigade arv on vahemikus 0,5%-1% ühe nukleotiidi kohta [11]. Valesti loetud k-meeri tõttu võime näha ka selliste bakteriliikide unikaalseid k-meere, keda proovis tegelikult ei esine. Kasutades puu struktuuri on mudeleid kasutades võimalik hinnata lehtedes ekslikult paiknevate k-meeride osakaalu. Järgnev joonis 5 näitlikustab, kuidas leitakse ühe nukleotiidi kaugusel olevad k-meerid.



Joonis 5: Bakteril A (AATTCAATT) leidub ühe nukleotiidi kaugusel kaks 4-meeri, mis ei ole sellele bakterile omane ning bakteril B (AATTGACTT) leidub ühe nukleotiidi kaugusel kolm 4-meeri, mis ei ole sellele bakterile omane.

1.2 Vaatlusandmed

Uuritavad kolm vaatlusandmestikku on koostanud bioinformaatika nooremteadur Mihkel Vaher Tartu Ülikooli molekulaar- ja rakubioloogia instituudist, bioinformaatika õppetoolist.

Esimeses, referentsgenoomi põhjal koostatud andmestikus kirjeldatakse, kui mitu uuritava bakteri või sõlme k-meeri sisaldub teises bakteris ning ühel toimetamiskaugusel olevate k-meeride arv. Vaadeldavas andmestikus on 6 tunnust ja 10 878 vaatlust. Tunnusteks on:

- **V1** - bakteri nimetus;
- **V2** - huvipakkuva bakteri või sõlme nimetus;
- **V3** - huvipakkuva bakteri või sõlme k-meeride arv, mis sisalduvad tunnuse V1 bakteri k-meeride hulgas (loendatud ilma kordusteta);
- **V4** - huvipakkuva bakteri või sõlme k-meeride arv, mis sisalduvad tunnuse V1 bakteri k-meeride hulgas (loendatud koos kordustega);

- **V5** - k-meeride arv, mis on huvipakkuva bakteri või sõlme k-meerist ühel toimetamiskaugusel (loendatud ilma kordusteta);
- **V6** - k-meeride arv, mis on huvipakkuva bakteri või sõlme k-meerist ühel toimetamiskaugusel (loendatud koos kordustega).

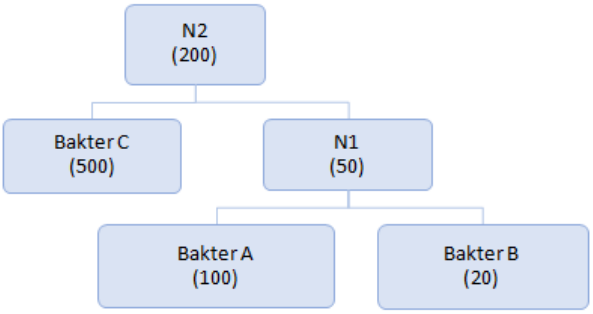
Tunnustes V1 ja V2 on vastavalt 74 ja 147 unikaalset bakterit või sõlme.

Teises ja kolmandas, sekveneerimisandmete põhjal koostatud andmestikes on kirjeldatud bakteri või sõlme k-meere ning nende jaotust. Mõlemas andmestikus on 504 kattuvat uuritavat tunnust ning andmestikes on vastavalt 10 ja 100 vaatlust. Tunnusteks on:

- **V1** - bakteri või sõlme nimetus;
- **V2** - bakterile või sõlmele omaste k-meeride arv;
- **V3** - valimis nähtud bakteri või sõlme k-meeride arv;
- **V4** - valimis mitte esinenud k-meeride arv;
- **V5** - valimis ühe korra esinenud k-meeride arv;
- **V6** - valimis kaks korda esinenud k-meeride arv;
- ...
- **V504** - valimis viissada korda esinenud k-meeride arv.

1.3 Simuleeritud andmed

Simuleeritud andmete puhul genereerisime andmestiku, mis on uuritava vaatlusandmestikuga võimalikult sarnase ülesehitusega. Antud juhul teame keskmise k-meeride katvuse β tegelikke väärtusi ning me saame neid võrrelda mudeli poolt hinnatud parameetrite $\hat{\beta}$ vastu. Parameetrite tegelikeks väärtusteks on võetud $\beta = (\beta_A, \beta_B, \beta_C) = (0{,}1;\ 0;\ 0{,}02)$, kui pole väidetud teisiti. Kolme bakteri A, B ja C unikaalsete k-meeride arvuks on vastavalt 100, 20 ning 100 k-meeri. Bakteritel A ja B on ühiseid k-meere 50 ning bakteritel A, B ja C on ühiseid k-meere 200. Kirjeldatud seosed on kujutatud joonisel 6.



Joonis 6: Simuleeritud andmetele vastav k-meeride fülogeneesipuu.

Tähistame sümboliga n_A proovi sekveneerimisel nähtud bakterile A omaste unikaalsete k-meeride arvu. Sümbolitega n_B ja n_C tähistame vastavalt, kui mitut bakteritele B ja C vastavat unikaalset k-meeri nägime ning sümbolitega n_{N1} ja n_{N2} vastavalt, kui mitut sõlmedele N_1 ja N_2 vastavat unikaalset k-meeri nägime.

Kõigil positsioonidel on lugemi alguseks saamisel sama tõenäosus. Kui lugemid jaotuvad üle genoomi juhuslikult, tuleks [2] põhjal sageduste vektori $\boldsymbol{y} = (n_A, n_B, n_C, n_{N1}, n_{N2})'$ elemendid genereerida kasutades Poissoni jaotust. Vastavalt fülogeneesipuu seostele

- $n_A \sim P(100 \cdot \beta_A)$;
- $n_B \sim P(20 \cdot \beta_B)$;
- $n_C \sim P(500 \cdot \beta_C)$;
- $n_{N1} \sim P(50 \cdot (\beta_A + \beta_B))$;
- $n_{N2} \sim P(200 \cdot (\beta_A + \beta_B + \beta_C))$.

Simuleeritud andmeid saab kirjeldada lineaarse mudeli $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ abil

$$\begin{bmatrix} n_A \\ n_B \\ n_C \\ n_{N1} \\ n_{N2} \end{bmatrix} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 500 \\ 50 & 50 & 0 \\ 200 & 200 & 200 \end{bmatrix} \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \end{bmatrix} + \begin{bmatrix} \varepsilon_A \\ \varepsilon_B \\ \varepsilon_C \\ \varepsilon_{N1} \\ \varepsilon_{N2} \end{bmatrix} \tag{1}$$

$$\boldsymbol{y} \quad = \quad \boldsymbol{X} \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon},$$

kus $\boldsymbol{y}$ on 5-mõõtmeline sageduste vektor, $\boldsymbol{X}$ on 5×3 mudelimaatriks, $\boldsymbol{\beta}$ on 3-mõõtmeline hindamist vajav k-meeride keskmiste katvuste vektor ja $\boldsymbol{\varepsilon}$ on 5-mõõtmeline juhuslike vigade vektor. Vektori $\boldsymbol{\beta}$ elemendid β_A, β_B ja β_C kirjeldavad bakterite A, B ja C genoomide keskmist sekveneerimiskatvust. Siis bakterile A vastavate unikaalsete k-meeride arv sekveneerimisandmetes avaldub kujul

$$n_A = 100 \cdot \beta_A + \varepsilon_A. \tag{2}$$

Siin vektorid $\boldsymbol{y}$ ja $\boldsymbol{\varepsilon}$ on juhuslikud vektorid. Eeldades, et tehtud mõõtmistulemused pole süstemaatiliselt valed ehk nihkega ($E\boldsymbol{\varepsilon} = 0$) saame, et

$$\boldsymbol{Ey} = \boldsymbol{X\beta}(:= \boldsymbol{\lambda}). \tag{3}$$

2 Probleemi kirjeldus

Meenutame, et baktereid, mis inimese organismi tungides põhjustavad haigusi, nimetati patogeenseteks. Enamik neist moodustavad orgaanilisi mürkaineid ehk toksiine, mis kutsuvad esile koekahjustusi. Toksiinidel on väga spetsiifiline toime. Näiteks mõjutab teetanuse ehk kangestuskramptõve tekitaja poolt sünteesitav teetanusetoksiin närvisüsteemi. Tegemaks kindlaks, milline bakter organismi vaevab, tuleb teha vastav test patsiendi sümptomeid arvestades. Seni on kasutatud metoodikat, kus patsiendi põletikukeskkonnast võetakse proov ehk külv, mis saadetakse edasi laborisse. Kindel kogus külvi kantakse söötmega Petri tassidele ja seda inkubeeritakse teatava ajaühiku (näiteks 72 tundi). Seejärel kolooniad loendatakse ja tassil kasvanud kolooniate arvu järgi arvutatakse mikroorganismide arv põletikukolde ühes ühikus. Antud meetod on väga ajamahukas ning annab ainult osalise ülevaate võimalikest bakteritest, mis patsiendile vaevusi tekitavad. Iga järgnev test on nii aja- kui rahakulukas, mis osutub probleemiks. Üldiselt kasutatakse bakterite kindlakstegemiseks ka PCR või ELISA testi, kuid ka antud meetodid kontrollivad ühe bakteri olemasolu, seega antud meetodid ei lahenda nimetatud probleeme.

Alternatiivne viis on võtta külv ning rakkude suuruste järgi filtreerides eraldatakse külvist kõik bakterid ning neist omakorda eraldatakse genoomne DNA. Saadud DNA järjestused sekveneeritakse ning võetakse sellest valim. Saadud valmist loendatakse bakteritele omaste k-meeride arv ning nende sagedused. Ühtlasi on võimalik kindlaks teha, kui palju on k-meere, mis on antud bakterile unikaalsed ning k-meeride arvu, mis on teise bakteri k-meerile väga sarnased ehk erinevus on ainult ühes nukleotiidis. Eesmärgiks on saadud k-meeride valimi põhjal hinnata, milliseid baktereid ning kui palju neid võrreldes teiste bakteritega esineb proovis. Antud meetodi puhul saab ühe testi tulemusena teada kõik põletikukoldes esinevad bakterid ning nende hinnangulised sagedused.

Bakterite suhtelise arvukuse (mida saab leida sekveneerimiskatvuste võrdlemise teel) leidmiseks võime kasutada kolme erinevat meetodit. Järgnevalt uuritakse, milline neist meetoditest võiks olla kõige täpsem ja usaldusväärsem.

3 Kasutatud metoodika

Käesolevas peatükis on alampeatükid 3.1 ja 3.2 on kirjutatud kasutades vastavalt allikate [7] ja [1] materjale ning alampeatükid 3.3 ja 3.4 on kirjutatud kasutades allika [12] materjale.

3.1 Vähimruutude hinnang

Lineaarne mudel on mudeli parameetrite lineaarne funktsioon ning see esitub kujul

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{4}$$

kus $\boldsymbol{y}$ on $n \times 1$ uuritava tunnuse vektor, $\boldsymbol{X}$ on $n \times p$ mudeli- või disainmaatriks, mille elementideks olevad konstandid on meile teada, $\boldsymbol{\beta}$ on $p \times 1$ hindamist vajavate (tundmatute) parameetrite vektor ning $\boldsymbol{\varepsilon}$ on $n \times 1$ juhuslike vigade vektor. Nii $\boldsymbol{y}$ kui $\boldsymbol{\varepsilon}$ on juhuslikud vektorid. Maatrikskujul kirjapandult näeks lineaarne mudel välja järgmine

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{12} & \dots & x_{1p} \\ 1 & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}, \tag{5}$$

$$\boldsymbol{y} \quad = \quad \boldsymbol{X} \quad \boldsymbol{\beta} \quad + \quad \boldsymbol{\varepsilon}.$$

Fikseeritud i-nda vaatluse korral y_i avaldub kujul

$$y_i = \beta_1 + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i.$$

Antud rakenduses on vaatlusvektori $\boldsymbol{y}$ väärtusteks sõlmede või lehtedele omaste k-meeride sagedused (mitu korda me antud sõlme või lehte kuuluvaid k-meere proovis nägime) ja hinnatavaks parameetervektoriks $\boldsymbol{\beta}$ eri bakteriliikide sekveneerimiskatvused. Mudelimaatriksi $\boldsymbol{X}$ kuju sõltub fülogeneesipuu kujust ning sõlmedesse sattunud k-meeride arvust. Paremaks arusaamiseks vaata joonist 6 ja vastavat lineaarset mudelit (1).

Lineaarse mudeli parameetrid $\boldsymbol{\beta}$ on võimalik hinnata vähimruutude meetodil. Antud meetodi korral pole vaja teha eeldusi mudelis olevate tunnuste jaotuste kohta, meetod töötab alati ja on teatud mõttes parim. Antud meetodi puhul

leitakse mudeli parameetrite $\boldsymbol{\beta}$ hinnang $\hat{\boldsymbol{\beta}}$ selliselt, et tekkivate prognoosivigade $\boldsymbol{\varepsilon} = \boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}$ ruutude summa $SSE = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^T\boldsymbol{\varepsilon}$ oleks minimaalne. Teisisõnu minimiseeritakse avaldise

$$\boldsymbol{S}(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \tag{6}$$

väärtus parameetrite $\boldsymbol{\beta}$ järgi. Funktsiooni $\boldsymbol{S}(\boldsymbol{\beta})$ miinimumi leidmiseks leitakse esmalt tuletis vektori $\boldsymbol{\beta}$ järgi

$$\begin{aligned}
\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^{\boldsymbol{T}}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial (\boldsymbol{y}^{\boldsymbol{T}} - \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}})(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial (\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{y} - \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= \frac{\partial (\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{y} - 2\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \\
&= 2\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - 2\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}.
\end{aligned}$$

Leitud tuletis võrdsustatakse nulliga ja teisendatakse saadud võrrandisüsteem

$$\begin{aligned}
\left.\frac{\partial \boldsymbol{S}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} &= 0 \\
2\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} - 2\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X} &= 0 \\
\hat{\boldsymbol{\beta}}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} &= \boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X} \\
\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} &= \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}.
\end{aligned} \tag{7}$$

Kui maatriks $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ on pööratav, siis saadakse (7) lahendiks

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}. \tag{8}$$

Tihtipeale pole maatriks $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ pööratav. Kui otsitakse lahendeid võrrandisüsteemile $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, siis juhul, kui lahend leidub, avaldub see lahend kujul $\boldsymbol{x} = \boldsymbol{A}^{-}\boldsymbol{b}$, kus $\boldsymbol{A}^{-}$ tähistab maatriksi $\boldsymbol{A}$ üldistatud pöördmaatriksit. Kuna üldistatud pöördmaatriks pole üheselt määratud, siis pole ka antud lahend üldjuhul üheselt määratud. Seega kui võrrandisüsteem (7) on lahenduv, siis avaldub lahend kujul

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})^{-}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}. \tag{9}$$

Saadud vähimruutude hinnang $\hat{\boldsymbol{\beta}}$ on üheselt määratud vaid juhul, kui maatriks $(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X})$ on pööratav. Kui maatriks $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ on kõdunud, siis lahendiks sobib iga

$\hat{\boldsymbol{\beta}}$, mis avaldub kujul

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X^T X})^- \boldsymbol{X^T y} + (\boldsymbol{I} - (\boldsymbol{X^T X})^- \boldsymbol{X^T X})\boldsymbol{u}, \tag{10}$$

kus $\boldsymbol{u}$ on suvaline n-mõõtmeline vektor, mis osutub samuti võrrandisüsteemi (7) lahendiks. Vähimruutude meetodil saadud hinnangud $\boldsymbol{y}$-le on

$$\begin{aligned} \hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X^T X})^- \boldsymbol{X^T y}, \end{aligned} \tag{11}$$

ning prognoosivead on

$$\begin{aligned} \boldsymbol{\varepsilon} &= \boldsymbol{y} - \hat{\boldsymbol{y}} \\ &= (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X^T X})^- \boldsymbol{X^T})\boldsymbol{y}. \end{aligned} \tag{12}$$

Kui soovitakse testida, kas i-ndat bakteriliiki proovis esines või ei esinenud ($H_0 : \beta_i = 0$), siis eeldatakse enamasti, et uuritava tunnuse tinglikuks jaotuseks on normaaljaotus, $\boldsymbol{y}|\boldsymbol{X} \sim N(\boldsymbol{X\beta; I\sigma^2})$. Kuigi antud eeldus on sekveneerimisandmete puhul rikutud, uurime siiski, kas vastav test võiks praktikas rakendamiseks piisavalt usaldusväärne olla.

Praktikas leiti parameetrite hinnangud kasutades R-is autori poolt kirjutatud funktsiooni (toodud lisas B) ning parameetrite olulisuse testimiseks paketi **stats** käsku **lm()**.

3.2 Mittenegatiivne vähimruutude meetod

Keerulisemaks osutub ülesanne, kui nõutakse, et kõikide parameetrite väärtused peavad olema mittenegatiivsed $\boldsymbol{\beta_i} \geq 0$,

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}, \quad \text{kus } \boldsymbol{\beta} \geq 0. \tag{13}$$

Meeldetuletuseks, $\boldsymbol{y}$ oli $n \times 1$ uuritava tunnuse vektor, $\boldsymbol{X}$ on $n \times p$ mudelimaatriks, mille elementideks olevad konstandid on teada, $\boldsymbol{\beta}$ oli $p \times 1$ hindamist vajavate (tundmatute) parameetrite vektor ning $\boldsymbol{\varepsilon}$ on $n \times 1$ juhuslike vigade vektor.

Hinnangute leidmiseks lahendatakse järgmine minimeerimise ülesanne

$$\min_{\beta,\ \beta \geq 0} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2. \tag{14}$$

Siinkohal $\boldsymbol{\beta}$ hinnangute leidmiseks ilmutatud kujul lahendit pole, tuleb kasutada numbrilisi meetodeid. Praktikas leiti parameetrite hinnangud kasutades R-is paketi **nnls** käsku **nnls()** ning paketi **penalized** käsku **penalized()**. Parameetrite statistilise olulisuse testimisel on eeldatud, et vaatlused on normaaljaotusega. Käsus **nnls()** on hinnangute leidmiseks kasutatud Lawson-Hanson algoritmi implementatsiooni [6].

Lawson-Hanson'i algoritm

Olgu antud maatriks $\boldsymbol{X}$ mõõtmetega $n \times p$ ning vektor $\boldsymbol{y}$ pikkusega n, kus $n, p \in \mathbb{N}$. Vektoreid w ja z pikkusega p ($p \in \mathbb{N}$) kasutame tulemuste salvestamiseks. Indeksite hulgad P ning z defineeritakse algoritmis töö käigus. Muutujad indeksite hulgas P võivad võtta kõiki täisarvulisi väärtusi, välja arvatud 0. Juhul kui muutuja võtab mittepositiivse väärtuse, liigub algoritm selle positiivse väärtuse juurde või väärtustab muutuja nulliga ning liigutab selle indeksi hulgast P hulka Z. Algoritmi töö lõppedes on $\boldsymbol{\beta}$ lahendivektor ning w on duaalne vektor (*dual vector*). Antud alampeatükis mittenegatiivne vähimruutude meetod esitub kujul $\min\limits_{\beta,\ \beta \geq 0} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2$. Järgnev algoritm on võetud allikast [3].

Algoritm 3.2 NNLS($\boldsymbol{X}, n, p, \boldsymbol{y}, \boldsymbol{\beta}, w, z, P, Z$)

1: Olgu $P \coloneqq NULL$, $Z \coloneqq 1, 2, \ldots, p$ ja $\boldsymbol{\beta} \coloneqq 0$.
2: Leia pikkusega p vektor $w \coloneqq \boldsymbol{X}^{\intercal}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$.
3: Kui hulk Z on tühi või kui $w_j \leq 0$ iga $j \in Z$ korral, siis algoritm jätkub punktis 12.
4: Leia indeks $t \in Z$ nii, et $w_t = \max\{w_j : j \in Z\}$.
5: Liiguta indeks t hulgast Z hulka P.
6: Tähistagu $\boldsymbol{X}_P$ $n \times p$ maatriksit, mis on defineeritud kui

$$\text{Maatriksi } \boldsymbol{X}_P \text{ veerg } j \coloneqq \begin{cases} \text{maatriksi } \boldsymbol{X} \text{ veerg } j, \text{ kui } j \in P \\ 0, \text{ kui } j \in Z \end{cases}$$

Leia vektor z pikkusega p vähimruutude meetodil võrrandist $\min\|\boldsymbol{y} - \boldsymbol{X_p z}\|_2$. Paneme tähele, et ainult z_j, $j \in P$ komponendid on määratud antud võrrandi poolt. Kui $j \in Z$, siis $z_j \coloneqq 0$.
7: Kui $z_j > 0$ iga $j \in P$ korral, siis $\boldsymbol{\beta} \coloneqq z$ ning algoritm jätkub punktis 2.
8: Leia indeks $q \in P$ nii, et $\boldsymbol{\beta}_q/(\boldsymbol{\beta}_q - z_q) = \min\{\boldsymbol{\beta}_j/(\boldsymbol{\beta}_j - z_j) : z_j \leq 0, j \in P\}$.
9: Olgu $\alpha \coloneqq \boldsymbol{\beta}_q/(\boldsymbol{\beta}_q - z_q)$.
10: Olgu $\boldsymbol{\beta} \coloneqq \boldsymbol{\beta} + \alpha(z - \boldsymbol{\beta})$.
11: Liiguta hulgast P hulka Z kõik indeksid $j \in P$, mille korral $\boldsymbol{\beta}_j = 0$. Algoritm jätkub punktis 6.
12: Protsess on lõppenud.

Algoritmi NNLS töö lõppedes vektor $\boldsymbol{\beta}$ rahuldab tingimusi

$$\boldsymbol{\beta}_j > 0, \text{ kui } j \in P, \tag{15}$$

$$\boldsymbol{\beta}_j = 0, \text{ kui } j \in Z \tag{16}$$

ning on vähimruutude meetodi lahendivektor võrrandile (14). Algoritmi koonduvus on tõestatud allikas [3] alates lk 163.

3.3 Suurima tõepära meetod

Tihtipeale on võimalik teha eelduseid uuritava tunnuse jaotuse kohta. Otsitavaid parameetreid $\boldsymbol{\beta}$ võime leida suurima tõepära meetodil, kui uuritava tunnuse jaotus on teada. Käesolevas töös uuritakse, millise hinnangu parameetritele saame,

kui uuritava n-elemendilise valimi $\boldsymbol{y}$ elemendi y_i, kus $i \in \{1, 2, 3 \ldots, n\}$ jaotuseks on Poissoni jaotus. Tuletame meelde eeldust, et tehtud mõõtmistulemused ei olnud nihkega ning tähistasime, et $\boldsymbol{Ey} = \boldsymbol{X\beta} = \boldsymbol{\lambda}$.

Elemendi y_i, kus $i \in \{1, 2, \ldots, n\}$ tõepärafunktsiooniks on

$$\begin{aligned} L(\lambda_i, y_i) &= f_y(\lambda_i, y_i) \\ &= \frac{e^{-\lambda_i}}{y_i!} \cdot (\lambda_i)^{y_i}. \end{aligned} \tag{17}$$

Olgu x_i, kus $i \in \{1, 2, 3 \ldots, n\}$ mudelimaatriksi $\boldsymbol{X}$ i-s rida. Uuritava n-elemendilise valimi $\boldsymbol{y}$ tõepärafunktsiooniks on

$$\begin{aligned} L(\boldsymbol{\lambda}, \boldsymbol{y}) &= L(\boldsymbol{X\beta}, \boldsymbol{y}) \\ &= \prod_{i=1}^{n} f_Y(x_i \cdot \boldsymbol{\beta}, y_i) \\ &= \prod_{i=1}^{n} \frac{e^{-x_i \cdot \boldsymbol{\beta}}}{y_i!} \cdot (x_i \cdot \boldsymbol{\beta})^{y_i} \end{aligned} \tag{18}$$

ning log-tõepära funktsiooniks saame

$$\begin{aligned} l(\boldsymbol{\lambda}, \boldsymbol{y}) &= l(\boldsymbol{X\beta}, \boldsymbol{y}) \\ &= \ln L(\boldsymbol{X\beta}, \boldsymbol{y}) \\ &= \ln(\prod_{i=1}^{n} \frac{e^{-x_i \cdot \boldsymbol{\beta}}}{y_i!} \cdot (x_i \cdot \boldsymbol{\beta})^{y_i}) \\ &= \sum_{i=1}^{n} \ln(\frac{e^{-x_i \cdot \boldsymbol{\beta}}}{y_i!} \cdot (x_i \cdot \boldsymbol{\beta})^{y_i}) \\ &= \sum_{i=1}^{n} y_i \cdot \ln(x_i \cdot \boldsymbol{\beta}) - x_i \cdot \boldsymbol{\beta} - \ln(y_i!). \end{aligned} \tag{19}$$

Simuleeritud andmete puhul

$$\boldsymbol{y} = \begin{bmatrix} n_A \\ n_B \\ n_C \\ n_{N1} \\ n_{N2} \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 500 \\ 50 & 50 & 0 \\ 200 & 200 & 200 \end{bmatrix}$$

$$\text{ning} \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_A \\ \beta_B \\ \beta_C \end{bmatrix}.$$

Saame, et

$$\boldsymbol{\lambda} = \boldsymbol{X}\boldsymbol{\beta} = \begin{bmatrix} 100 \cdot \beta_A \\ 20 \cdot \beta_B \\ 500 \cdot \beta_C \\ 50 \cdot (\beta_A + \beta_B) \\ 200 \cdot (\beta_A + \beta_B + \beta_C) \end{bmatrix}.$$

Siis simuleeritud andmete log-tõepära funktsiooniks saame

$$\begin{aligned} l(\boldsymbol{\lambda}, \boldsymbol{y}) = &\ln L(\boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{y}) = \ln(\prod_{i=1}^{n} \frac{e^{-x_i \cdot \boldsymbol{\beta}}}{y_i!}(x_i \cdot \boldsymbol{\beta})^{y_i}) \\ = &\sum_{i=1}^{n} \ln(\frac{e^{-x_i \cdot \boldsymbol{\beta}}}{y_i!}(x_i \cdot \boldsymbol{\beta})^{y_i}) \\ = &y_1 \cdot \ln(x_1 \cdot \boldsymbol{\beta}) - x_1 \cdot \boldsymbol{\beta} - \ln(y_1!) + \\ &y_2 \cdot \ln(x_2 \cdot \boldsymbol{\beta}) - x_2 \cdot \boldsymbol{\beta} - \ln(y_2!) + \\ &y_3 \cdot \ln(x_3 \cdot \boldsymbol{\beta}) - x_3 \cdot \boldsymbol{\beta} - \ln(y_3!) + \\ &y_4 \cdot \ln(x_4 \cdot \boldsymbol{\beta}) - x_4 \cdot \boldsymbol{\beta} - \ln(y_4!) + \\ &y_5 \cdot \ln(x_5 \cdot \boldsymbol{\beta}) - x_5 \cdot \boldsymbol{\beta} - \ln(y_5!) \\ = &n_A \cdot \ln(100 \cdot \beta_A) - 100 \cdot \beta_A - \ln(n_A!) + \\ &n_B \cdot \ln(20 \cdot \beta_B) - 20 \cdot \beta_B - \ln(n_B!) + \\ &n_C \cdot \ln(500 \cdot \beta_C) - 500 \cdot \beta_C - \ln(n_C!) + \\ &n_{N1} \cdot \ln(50 \cdot (\beta_A + \beta_B)) - 50 \cdot (\beta_A + \beta_B) - \ln(n_{N1}!) + \\ &n_{N2} \cdot \ln(200 \cdot (\beta_A + \beta_B + \beta_C)) - 200 \cdot (\beta_A + \beta_B + \beta_C) - \ln(n_{N2}!). \end{aligned} \tag{20}$$

Suurima tõepära hinnangu saamiseks tuleb log-tõepära maksimiseerida. Selleks tuleb esmalt leida osatuletised parameetrite β_A, β_B ja β_C järgi

$$\begin{aligned}\frac{\partial l(\boldsymbol{\lambda},\boldsymbol{y})}{\partial \beta_A} &= \frac{\partial \ln L(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y})}{\partial \beta_A} \\ &= \frac{100\cdot n_A}{100\cdot \beta_A} - 100 + \frac{50\cdot n_{N1}}{50\cdot(\beta_A+\beta_B)} - 50 + \frac{200\cdot n_{N2}}{200\cdot(\beta_A+\beta_B+\beta_C)} - 200 \\ &= \frac{n_A}{\beta_A} + \frac{n_{N1}}{\beta_A+\beta_B} + \frac{n_{N2}}{\beta_A+\beta_B+\beta_C} - 350,\end{aligned} \tag{21}$$

$$\begin{aligned}\frac{\partial l(\boldsymbol{\lambda},\boldsymbol{y})}{\partial \beta_B} &= \frac{\partial \ln L(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y})}{\partial \beta_B} \\ &= \frac{20\cdot n_B}{20\cdot \beta_B} - 20 + \frac{50\cdot n_{N1}}{50\cdot(\beta_A+\beta_B)} - 50 + \frac{200\cdot n_{N2}}{200\cdot(\beta_A+\beta_B+\beta_C)} - 200 \\ &= \frac{n_B}{\beta_B} + \frac{n_{N1}}{\beta_A+\beta_B} + \frac{n_{N2}}{\beta_A+\beta_B+\beta_C} - 270\end{aligned} \tag{22}$$

ja

$$\begin{aligned}\frac{\partial l(\boldsymbol{\lambda},\boldsymbol{y})}{\partial \beta_C} &= \frac{\partial \ln L(\boldsymbol{X}\boldsymbol{\beta},\boldsymbol{y})}{\partial \beta_C} \\ &= \frac{500\cdot n_C}{500\cdot \beta_C} - 500 + \frac{50\cdot n_{N1}}{50\cdot(\beta_A+\beta_B)} - 50 + \frac{200\cdot n_{N2}}{200\cdot(\beta_A+\beta_B+\beta_C)} - 200 \\ &= \frac{n_C}{\beta_C} + \frac{n_{N1}}{\beta_A+\beta_B} + \frac{n_{N2}}{\beta_A+\beta_B+\beta_C} - 750\end{aligned} \tag{23}$$

ning seejärel saadud tuletised võrdsustada nulliga ning avaldada vastavalt parameetrid β_A, β_B ja β_C, mis olekski antud parameetri suurima tõepära hinnang. Praktikas leiti parameetrite suurima tõepära hinnangud kasutades R-is paketi **stats** käsku **optim()** tagamaks, et sama programmi saaks kasutada ka rohkemate parameetritega näiteandmestiku analüüsimiseks.

Leiame teist järku osatuletised parameetrite β_A, β_B ja β_C järgi, mis osutuvad hiljem vajalikuks suurima tõepära meetodi usaldusintervallide leidmisel.

$$\frac{\partial l^2(\boldsymbol{\lambda},\boldsymbol{y})}{\partial \beta_A^2} = -\frac{n_A}{\beta_A^2} - \frac{n_{N1}}{(\beta_A+\beta_B)^2} - \frac{n_{N2}}{(\beta_A+\beta_B+\beta_C)^2}, \tag{24}$$

$$\frac{\partial l^2(\boldsymbol{\lambda}, \boldsymbol{y})}{\partial \beta_B^2} = -\frac{n_B}{\beta_B^2} - \frac{n_{N1}}{(\beta_A + \beta_B)^2} - \frac{n_{N2}}{(\beta_A + \beta_B + \beta_C)^2} \tag{25}$$

ja

$$\frac{\partial l^2(\boldsymbol{\lambda}, \boldsymbol{y})}{\partial \beta_C^2} = -\frac{n_C}{\beta_C^2} - \frac{n_{N1}}{(\beta_A + \beta_B)^2} - \frac{n_{N2}}{(\beta_A + \beta_B + \beta_C)^2}. \tag{26}$$

3.4 Tõepärasuhte test

Jagame valimiruumi S kaheks osaks S_1 ja S_2, kus jagamise teostab sobivalt valitud funktsioon valimipunktidest $\boldsymbol{Y}$. Olgu meil kahepoolne hüpoteesipaar

$$\begin{aligned} H_0 &: \theta \in \Omega_0 \\ H_1 &: \theta \in \Omega_1, \end{aligned} \tag{27}$$

kus θ on tundmatu parameeter. Tõepärasuhte teststatistik antud hüpoteeside paari testimiseks on

$$\Lambda(\boldsymbol{Y}) = \frac{\sup\limits_{\theta \in \Omega_0} L(\boldsymbol{\theta}, \boldsymbol{Y})}{\sup\limits_{\theta \in \Omega} L(\boldsymbol{\theta}, \boldsymbol{Y})}, \tag{28}$$

kus $L(\boldsymbol{\theta}, \boldsymbol{Y})$ on valimi tõepärafunktsioon. Kui $\hat{\theta}$ on parameetri θ suurima tõepära hinnang, siis $\sup\limits_{\theta \in \Omega} L(\boldsymbol{\theta}, \boldsymbol{Y}) = L(\hat{\boldsymbol{\theta}}, \boldsymbol{Y})$. Olgu $\sup\limits_{\theta \in \Omega_0} L(\boldsymbol{\theta}, \boldsymbol{Y}) = L(\boldsymbol{\theta_0}, \boldsymbol{Y})$. Siis

$$\Lambda(\boldsymbol{Y}) = \frac{L(\boldsymbol{\theta_0}, \boldsymbol{Y})}{L(\hat{\boldsymbol{\theta}}, \boldsymbol{Y})}. \tag{29}$$

Osutub, et suurusel $-2 \ln \Lambda(\boldsymbol{Y})$ on lihtne tuntud piirjaotus, mida saab testimiseks kasutada suurte valimimahtude korral. Moodustame tõepärasuhte teststatistikust (29) teisendatud suuruse

$$T_n = -2 \ln \Lambda(\boldsymbol{Y}) = -2 \ln \frac{L(\boldsymbol{\theta_0}, \boldsymbol{Y})}{L(\hat{\boldsymbol{\theta}}, \boldsymbol{Y})}. \tag{30}$$

Antud juhul eeldatakse, et $f(\theta, y)$ omab teist järku pidevaid osatuletisi θ järgi, kui $\theta \in \Omega_0$ ning Fisheri informatsioonimaatriks on pööratav. Olgu nullhüpotees H_0 õige. Siis Wilks'i teoreemi kohaselt protsessis $n \to \infty$

$$T_n \xrightarrow{\mathrm{D}} \chi^2_{df}. \tag{31}$$

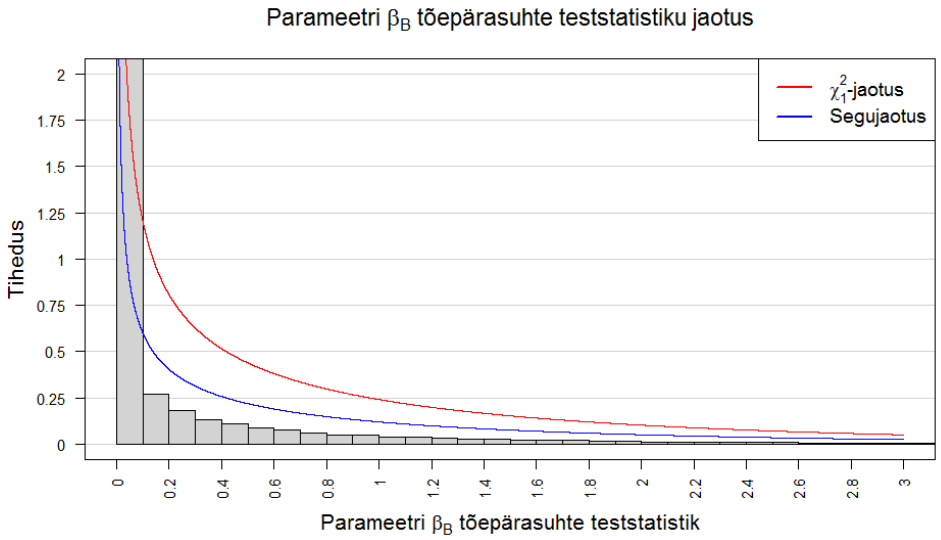
Vabadusastmete arv *df* on vabade parameetrite arv $\theta \in \Omega$, millest on lahutatud vabade parameetrite arv $\theta \in \Omega_0$ korral.

Olgu meil kaks mudelit - lihtsam ning keerulisem. Antud juhul on lihtsama mudeli korral hinnatud $p_0 = 2$ parameetrit ja keerulisema mudeli korral $p_1 = 3$ parameetrit. Siinkohal on lihtsama mudeli tõepärafunktsiooniks L_1 ja keerulisema mudeli tõepära funktsiooniks L_2. Meid huvitab, kas keerulisem mudel sobitub oluliselt paremini kui lihtsam mudel. Mudelite tõepärasuhte teststatistik on defineeritud kujul

$$T_n = -2 \cdot \ln(\frac{L_0}{L_1}) = -2 \cdot (\ln L_0 - \ln L_1) = 2 \cdot (\ln L_1 - \ln L_2). \tag{32}$$

Antud juhul ei saa Wilks'i teoreemi rakendada, sest parameetrite vektoril θ on seatud alumine kitsendus ($\theta \geq 0$) ning samuti on vaatluste arv $n = 5$ märkimisväärselt väike. Siiski on antud olukorras tõepärasuhte testi võimalik kasutada ning see on optimaalne Neymann-Personi lemma kohaselt, kuid olulisuse tõenäosust ei saa leida kasutades χ^2- jaotust vabadusastmetega *df*, mis on määratud Wilks'i teoreemi poolt.

Artiklis [5] on näidatud, et tõepärasuhte testistatistiku jaotus nullhüpoteesi kehtides vastab segujaotusele, kus 50% jaotusest on χ^2_1-jaotusega ning ülejäänud 50% jaotusest on võrdne nulliga, teatud situatsioonis. Kuigi käesolev olukord ei vasta täpselt artiklis [5] uuritud olukorrale, võime siiski katsetada, kas väljapakutud segujaotus võiks kirjeldada teststatistiku jaotust ka antud juhul. Seega pakub huvi, kas antud juhul tõepära suhte testi jaotused vastavad segujaotusele, mille puhul on pool jaotusest χ^2-jaotus punktmassiga punktis 0 ja pool χ^2-jaotus punktmassiga punktis 1. Järgmisel joonisel 7 on toodud parameetri β_B tõepärasuhte teststatistiku jaotus. Hüpoteesi $H_0 : \beta_B = 0$ testimiseks on kasutatud tõepärasuhte testi teststatistiku tegelikku jaotust nullhüpoteesi kehtides (simulatsioonide põhjal).



Joonis 7: Parameetri β_B tõepärasuhte teststatistiku jaotus nullhüpoteesi kehtides. Punase joonega on lisatud χ^2_1-jaotuse tihedus ning sinise joonega segujaotuse tihedus. Antud segujaotuse puhul on pool jaotusest punktmassiga punktis 0 ja pool χ^2-jaotus vabadusastmete arvuga 1.

Antud jooniselt 7 on näha, et tõepärasuhte teststatistiku jaotus ei vasta χ^2_1-jaotusele. Võrreldes χ^2_1-jaotusega on segujaotus lähedasem tõepärasuhte teststatistiku jaotusele, kuid ei vasta sellele. Seega tõepärasuhte testi p-väärtuse arvutamisel χ^2_1-jaotust kasutades saame valed tulemused, kuid saadud p-väärtused võiksid olla pigem liiga konservatiivsed (leitud p-väärtused on suuremad, kui nad olema peaksid).

4 Meetodite võrdlus sekveneerimisvigadeta andmete korral

4.1 Parameetrite hinnangud

Kordasime alampeatükis (1.3) kirjeldatud andmete genereerimise protsessi 1000 korda. Nende simulatsioonide pealt hindasime parameetervektori $\boldsymbol{\beta}$ väärtusi kolmel meetodil - vähimruutude meetodil (lühend *VR*), mittenegatiivsete vähimruutude meetodil (*Non-negative least squares*, lühend *NNLS*) ning suurima tõepära meetodil (lühend *STH*). Siin parameetervektori $\boldsymbol{\beta}$ tegelikeks väärtuseks on võetud $\boldsymbol{\beta} = (\beta_A, \beta_B, \beta_C)' = (0{,}1;\ 0;\ 0{,}02)'$. Võrdleme kolme mudelit hinnangu keskmise ruutvea (*Mean squared error*, lühend *MSE*) ja keskmise absoluutvea (*Mean absolute error*, lühend *MAE*) põhjal. Neist esimene on täpsuse mõõdik, mis mõõdab hinnangu $\hat{\beta}_B$ täpsust. Antud statistik avaldub kujul

$$MSE(\beta_B) = \frac{1}{n}\sum_{i=1}^{n}(\beta_B - \hat{\beta}_{Bi})^2, \tag{33}$$

kus n on simulatsioonide arv, antud juhul $n = 1000$. Teine statistik, keskmine absoluutviga näitab tegeliku parameetri β_B ja hinnatud parameetri $\hat{\beta}_B$ erinevust. Antud statistik avaldub kujul

$$MAE(\beta_B) = \frac{1}{n}\sum_{i=1}^{n}|\beta_B - \hat{\beta}_{Bi}|. \tag{34}$$

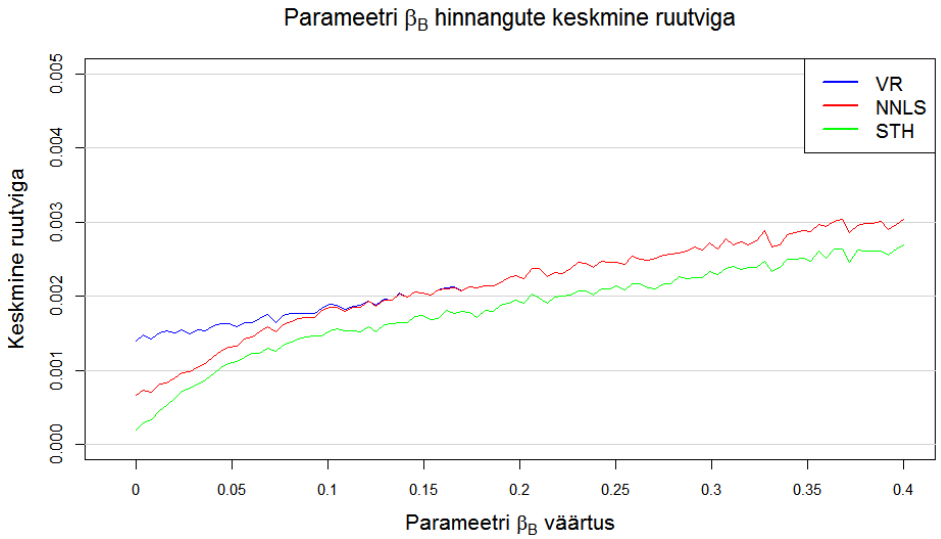
Kui andmete simuleerimise protsessi on korratud 1000 korda, siis saime järgmised tulemused

Meetod	Keskmine ruutviga	Keskmine absoluutviga
Vähimruutude meetod	$1{,}444 \cdot 10^{-3}$	$3{,}006 \cdot 10^{-2}$
Mittenegatiivne vähimruutude meetod	$6{,}892 \cdot 10^{-4}$	$1{,}517 \cdot 10^{-2}$
Suurima tõepära meetod	$1{,}928 \cdot 10^{-4}$	$5{,}048 \cdot 10^{-3}$

Tabel 1: Parameetri β_B hinnangute keskmine ruut- ja absoluutviga.

Keskmise ruutvea ja keskmise absoluutvea puhul vaatame, et mõlemad statistikud oleks võimalikult lähedal nullile. Saadud keskmise ruutvea $MSE(\beta_B)$ ja keskmise absoluutvea $MAE(\beta_B)$ põhjal on antud kolmest meetodist täpseim suurima tõepära meetod, millele järgnevad vastavalt mittenegatiivne vähimruutude

meetod ja vähimruutude meetod. Järgmisel joonisel on kujutatud parameetri β_B keskmist ruutviga kõigil kolmel meetodil, kui β_B varieerub lõigus $[0;\ 0{,}4]$.

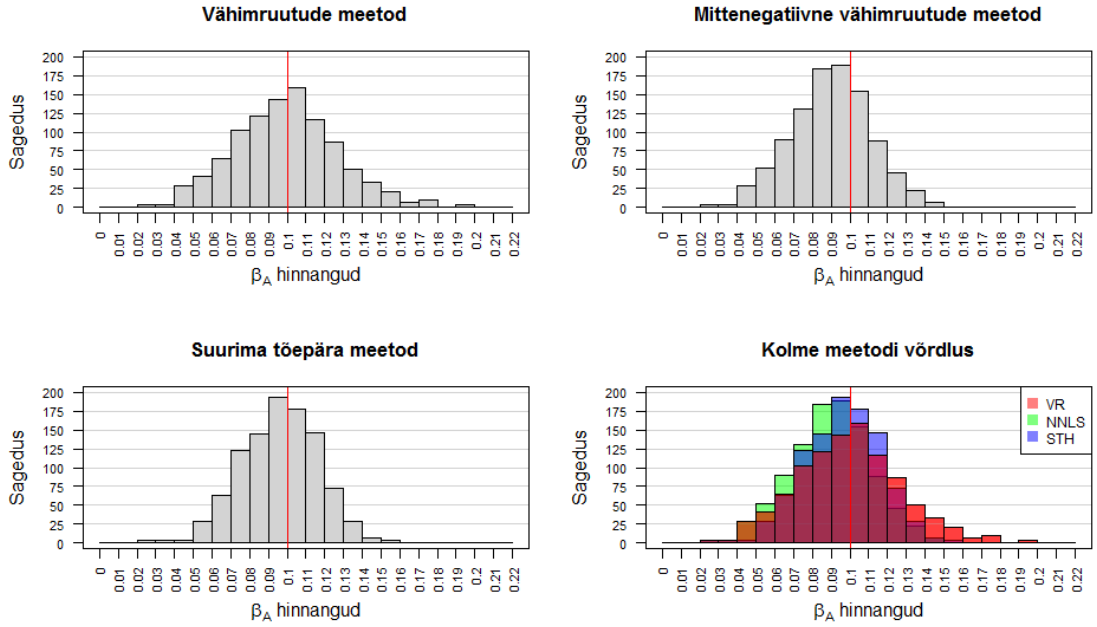


Joonis 8: Parameetri β_B hinnangute vähimruutude meetodi, mittenegatiivsete vähimruutude meetodi ning suurima tõepära meetodi keskmine ruutviga, kui β_B varieerub lõigus $[0;\ 0{,}4]$.

Antud joonisel 8 on lõigust $[0;\ 0{,}4]$ võetud ühtlaste vahedega 100 punkti ning iga punkti kohta on arvutatud 5000 hinnangut kõigil kolmel meetodil. Kui β_B varieerub lõigus $[0;\ 0{,}4]$, siis parima tulemuse kolmest meetodist annab suurima tõepära hinnang, mille keskmine ruutviga on väikseim kogu lõigu ulatuses. Mittenegatiivsete vähimruutude meetodi keskmine ruutviga on pisut parem vähimruutude meetodi keskmisest ruutveast lõigus $[0;\ 0{,}1]$, kuid lõigus $[0{,}1;\ 0{,}4]$ langevad meetodite keskmised ruutvead kokku ning mittenegatiivne vähimruutude meetod kaotab enda eelise vähimruutude meetodi ees. Parameetervektori $\boldsymbol{\beta}$ hinnangute keskmise absoluutvea joonis 21 on toodud lisas A.

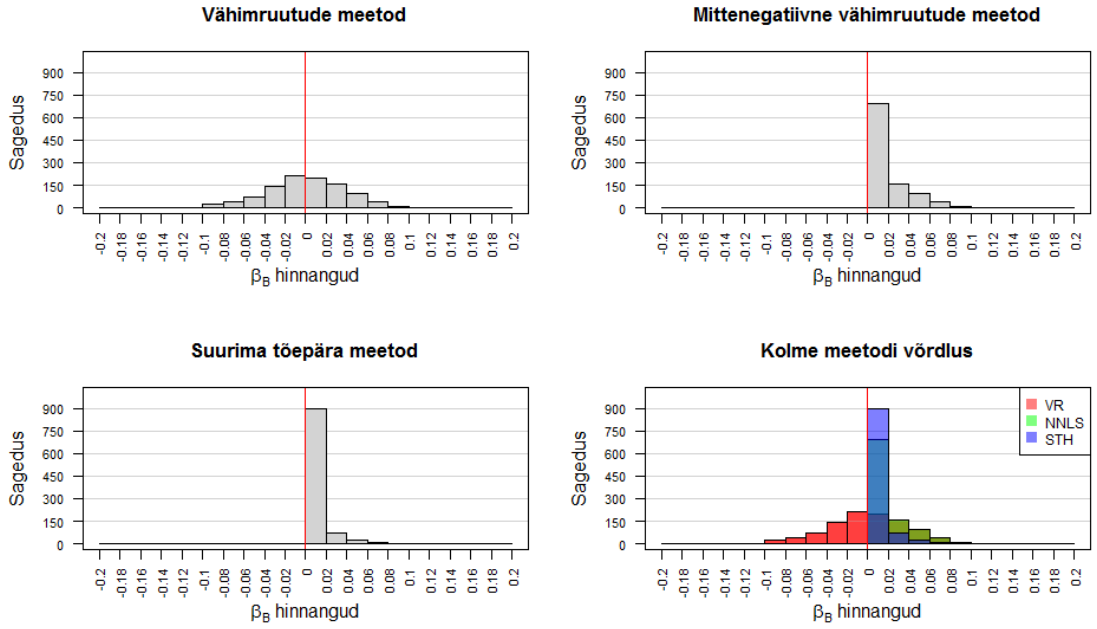
Esitame vähimruutude, mittenegatiivse vähimruutude ja suurima tõepära meetodil saadud parameetervektori $\boldsymbol{\beta}$ hinnangud joonistel 9, 10 ja 11. Parameetervektori $\boldsymbol{\beta}$ hinnangute hajuvusdiagrammid on toodud lisas A (joonisel 22, 23 ja 24).

Järgmiselt jooniselt 9 märkame, et parameetri $\beta_A = 0{,}1$ tegeliku väärtuse katab kõige rohkem suurima tõepära meetodil saadud hinnangud. Vähimruutude meetodil saadud hinnangute jaotus on ebasümmeetriline - rohkem esineb tegelikust väärtusest väiksemaid hinnanguid.

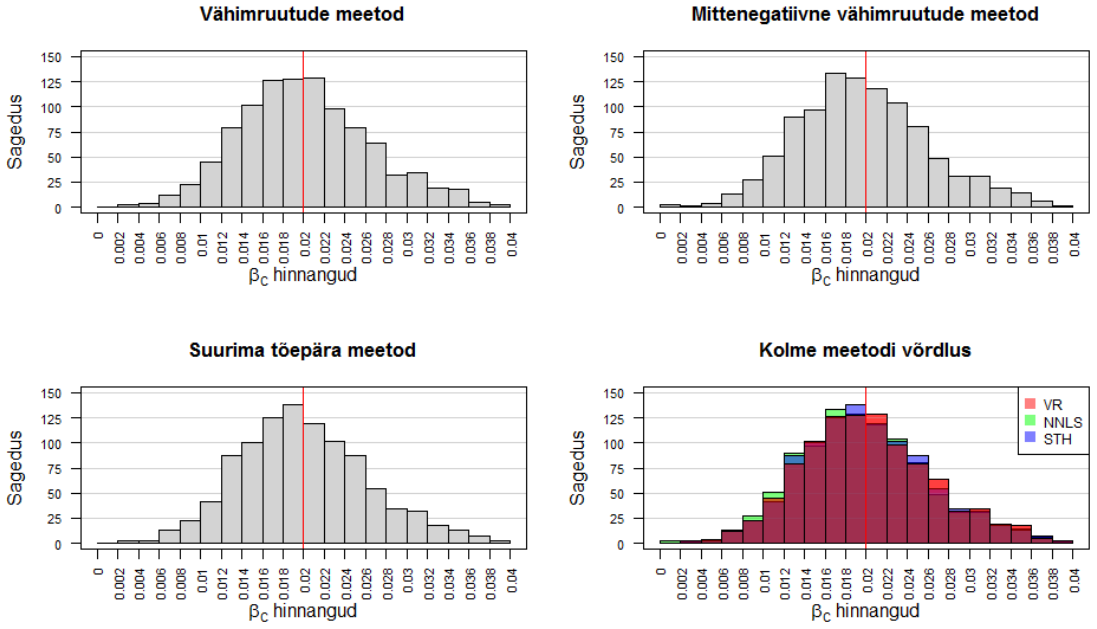


Joonis 9: Parameetri β_A hinnangud kolmel meetodil. Parameetri $\beta_A = 0{,}1$ tegelik väärtus on märgitud punase joonega.

Järgmiselt jooniselt 10 märkame, et parameetri β_B tegeliku väärtuse 0 hindamisel saame negatiivseid hinnanguid vaid vähimruutude meetodil. Suurima tõepära meetod ning mittenegatiivne vähimruutude meetod, nagu nimigi ütleb, annavad rangelt mittenegatiivseid hinnanguid. Parameetri β_B tegeliku väärtuse lähedale on kõige sagedamini sattunud suurima tõepära meetodil leitud hinnangud.

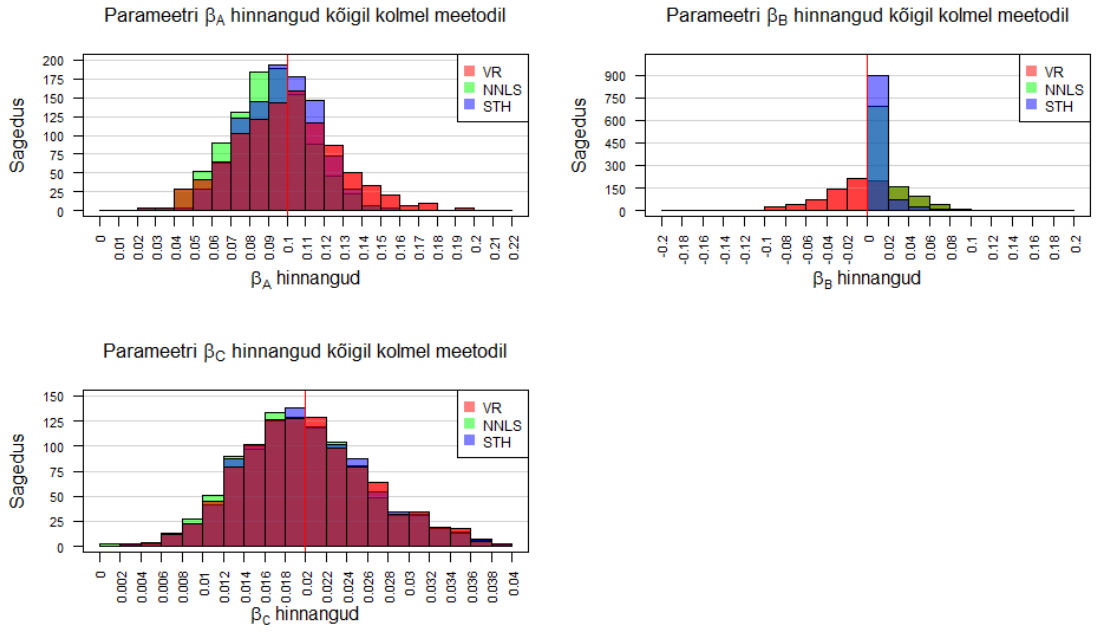


Joonis 10: Parameetri β_B hinnangud kolmel meetodil. Parameetri $\beta_B = 0$ tegelik väärtus on märgitud punase joonega.



Joonis 11: Parameetri β_C hinnangud kolmel meetodil. Parameetri $\beta_C = 0{,}02$ tegelik väärtus on märgitud punase joonega.

Jooniselt 11 näeme, et parameetri β_C hindamisel, mille tegelik väärtus on 0,02 ei ole antud meetoditel saadud hinnangutes niivõrd suuri erinevusi, kui parameetrite $\beta_A = 0{,}1$ ja $\beta_B = 0$ hindamisel. Antud meetoditel saadud hinnangujaotused on võrdlemisi sarnased. Mittenegatiivsete vähimruutude meetodil saadud hinnangud on pisut raskemate vasakpoolsete sabadega ning suurima tõepära meetodil saadud hinnangud pisut raskemate parempoolsete sabadega, kuid üleüldiselt on erisused väikesed.

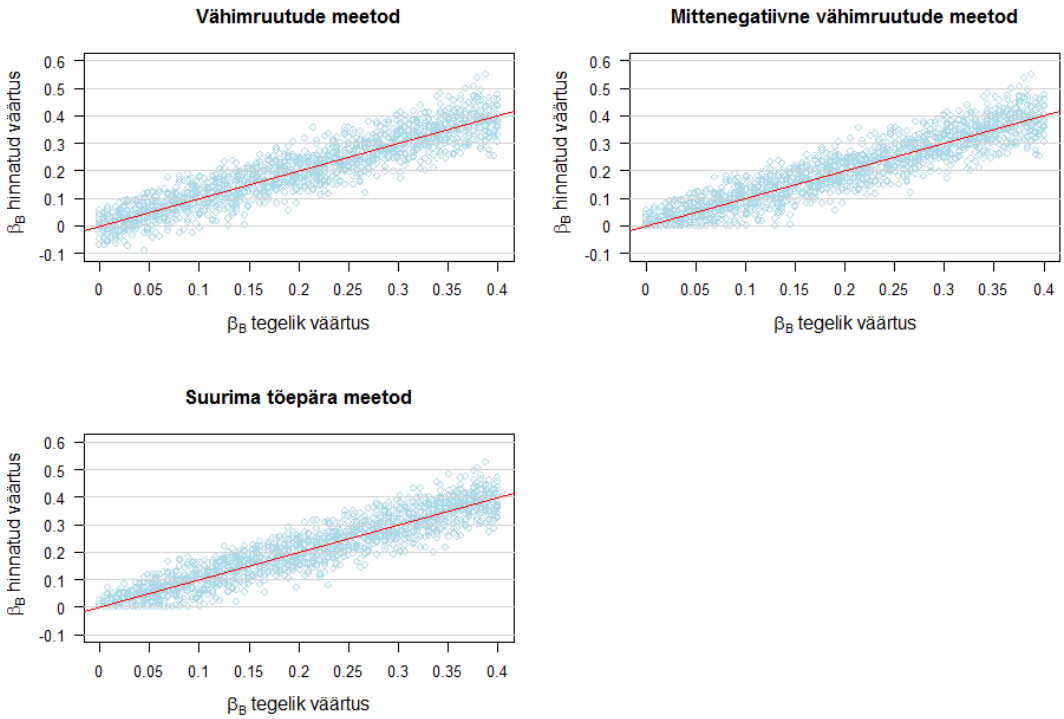


Joonis 12: Parameetrite β_A, β_B ja β_C hinnangud kõigil kolmel meetodil. Parameetri tegelik väärtus on märgitud punase joonega.

Jooniselt 12 märkame, et suurim erinevus on parameetri $\beta_B = 0$ hinnangute jaotuse puhul. Kõige täpsemini hindab parameetrit β_B suurima tõepära meetod, millele järgneb vahetult mittenegatiivne vähimruutude meetod. Mõlema meetodi puhul saadud hinnangud on rangelt positiivsed. Seevastu vähimruutude meetod hindab parameetri β_B väärtusteks nii positiivseid kui negatiivseid väärtusi ning jääb täpsusega märgatavalt alla kahele eelnevalt mainitud meetodile. Jooniselt 12 näeme parameetri $\beta_A = 0{,}1$ hinnangutes suuremaid erinevusi, kui parameetri $\beta_C = 0{,}02$ hinnangutes, mis langevad peaaegu kokku kõigil kolmel meetodil.

Järgmisel joonisel 13 on kujutatud parameetri β_B hinnanguid kõigil kolmel meetodil, kui β_B tegelik väärtus varieerub lõigus $[0;\ 0{,}4]$. Lõigust $[0;\ 0{,}4]$ on võetud ühtlaste vahedega 100 punkti ning iga punkti kohta on arvutatud 15 hinnangut kõigil kolmel meetodil. Joonistele on punase joonega märgitud olukord, kus $\beta_B = \hat{\beta}_B$.

Märkame, et kui β_B tegelik väärtus kuulub lõiku $[0;\ 0{,}1]$, siis vähimruutude meetod hindab vahel parameetri β_B väärtuse negatiivseks, kaotades sellega täpsust. Mittenegatiivne vähimruutude meetod ja suurima tõepära meetod annavad parameetri hinnanguteks alati ka parameetri nullilähedaste tegelike väärtuste korral mittenegatiivsed hinnangud, mis osutub nende meetodite eeliseks.



Joonis 13: Parameetri β_B hinnangud vähimruutude meetodil, mittenegatiivsete vähimruutude meetodil ning suurima tõepära meetodil, kui β_B varieerub lõigus $[0;\ 0{,}4]$. Joonistel punane joon tähistab võrdust $\beta_B = \hat{\beta}_B$.

4.2 Testi korrektsus ja võimsuse analüüs

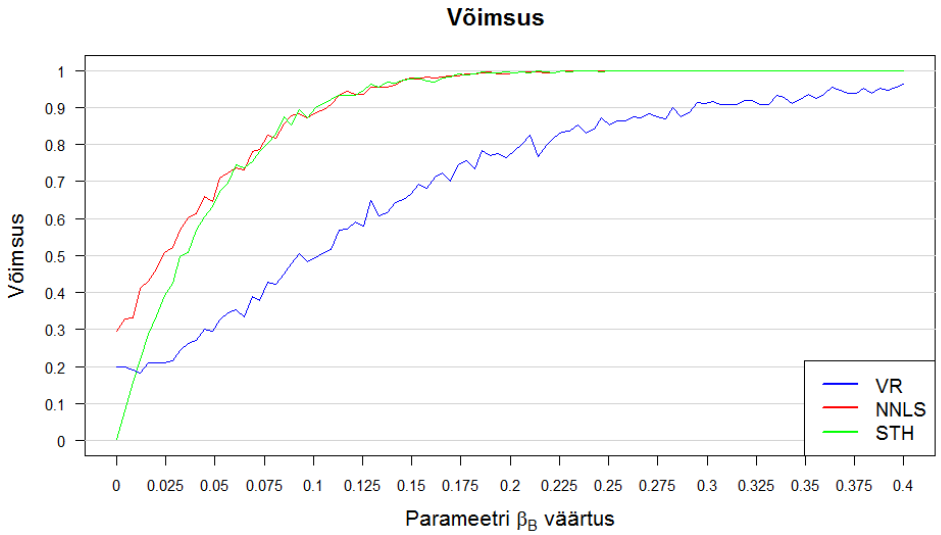
Statistiliste hüpoteeside juures räägitakse kaht liiki vigadest ja nende tegemise tõenäosusest

- I liiki vea tegemise tõenäosus $\alpha = \mathbf{P}(\text{võetakse vastu } H_1 \mid H_0 \text{ on õige})$
- II liiki vea tegemise tõenäosus $\beta = \mathbf{P}(\text{jäädakse } H_0 \text{ juurde} \mid H_1 \text{ on õige})$

Kui hüpoteesi testimisel kasutatakse olulisuse nivood 0,05, siis ei tohiks I liiki vea tegemise tõenäosus ületada 0,05. Kui kehtib alternatiivne hüpotees, siis soovitakse,

et testi võimsus $1-\beta$ oleks võimalikult suur. Võimsuse kasvades väheneb tõenäosus teha II liiki viga. Kui võimsus on väike, võib tulemuste osas jõuda valede järeldusteni, mistõttu sageli nõutakse, et realistlike alternatiivide korral oleks testi võimsus 0,8 või suurem. See tähendab, et on 20% või väiksem tõenäosus teha II liiki viga.

Järgmisel joonisel 14 on toodud vähimruutude meetodi, mittenegatiivsete vähimruutude meetodi ja suurima tõepära meetodi võimsused, kui testime hüpoteesi $H_0 : \beta_B = 0$ ja kui β_B tegelik väärtus varieerub lõigus $[0;\ 0{,}4]$. Joonisel on lõigust $[0;\ 0{,}4]$ võetud ühtlaste vahedega 100 punkti ning iga punkti kohta on korratud teste 1000 korda kõigil kolmel meetodil.



Joonis 14: Vähimruutude meetodi (t-test), mittenegatiivsete vähimruutude meetodi (tõepärasuhte test) ja suurima tõepära meetodi (tõepärasuhte test) võimsused, kui β_B varieerub lõigus $[0;\ 0{,}4]$.

Kui $\beta_B = 0$ on suurima tõepära meetodi I liiki vea tegemise tõenäosus 0, vähimruutude meetodil 0,2 ning mittenegatiivsete vähimruutude meetodil 0,3. Seega statistikatarkvarasse sisse kirjutatud testid (vähimruutude meetod, mis kasutab t-testi ning mittenegatiivne vähimruutude meetod, mis kasutab tõepärasuhte testi) ei tööta korrektselt. Kui $\beta_B = 0{,}1$ on suurima tõepära meetodi ja vähimruutude meetodi võimsus 0,9 ning mittenegatiivsete vähimruutude meetodi võimsus ligikaudu 0,5. Seega vaadeldud testidest töötab korrektselt (ei tee I liiki viga liiga sageli) vaid suurima tõepära meetod (kuigi on ka liiga konservatiivne). Samas on selle võimsus alternatiivse hüpoteesi kehtides võrreldav mittenegatiivsete vähimruutude meetodiga (ehk parima alternatiivse meetodiga).

4.3 Hinnangute nihketus

Võrdlemaks meetodite täpsust, tahetakse teada, kas $\hat{\boldsymbol{\beta}} = (\hat{\beta}_A, \hat{\beta}_B, \hat{\beta}_C)$ on nihketa. Et $\hat{\boldsymbol{\beta}}$ oleks nihketa, peab kehtib võrdus $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$. Leiame parameetervektori $\boldsymbol{\beta} = (\beta_A, \beta_B, \beta_C)$ igale parameetrile hinnangute keskmisele $\bar{\hat{\beta}}_i$, kus $i \in (A, B, C)$, vastava $(1-\alpha)$ usaldusvahemiku, mis tõenäosusega $(1-\alpha)$ sisaldab parameetri tegelikku väärtust. Teaduslikus kirjanduses võetakse tihti $\alpha = 0{,}05$. Siis leitav usaldusintervall katab parameetri tegelikku väärtust 95% tõenäosusega. Kuna antud juhul on simulatsioonide arv ($n = 1000$) piisavalt suur, võime usaldusintervalli leida järgnevalt

$$\bar{\hat{\beta}}_i \pm t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}, \tag{35}$$

kus $i \in (1, 2, 3)$ ja $t_{1-\frac{\alpha}{2}} = 1{,}96$ on t-jaotuse $(1 - \frac{\alpha}{2})$-kvantiil.

Juhul kui $\beta_i \in [\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}};\ \bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}]$, kus $i \in (A, B, C)$, siis võib vaadeldav hinnang olla nihketa. Leidsime vaadeldud kolmel meetodil hinnatud parameetri $\boldsymbol{\beta} = (\beta_A, \beta_B, \beta_C)$ hinnangute keskmisele usaldusintervallid, mille abil hindame, kas parameetrite hinnangud on nihkega või nihketa.

Parameeter	Tegelik väärtus β_i	Hinnangute keskmine $\bar{\hat{\beta}}_i$	$\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	Nihketa
β_A	0,1	0,1	$0,098$	$0,101$	+
β_B	0	0	-0,002	0,003	+
β_C	0,02	0,0201	0,0197	0,0205	+

Tabel 2: Vähimruutude meetodil saadud hinnangute nihketuse kontroll.

Parameeter	Tegelik väärtus β_i	Hinnangute keskmine $\bar{\hat{\beta}}_i$	$\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	Nihketa
β_A	0,1	0,088	0,087	0,089	−
β_B	0	0,015	0,014	0,016	−
β_C	0,02	0,0197	0,0193	0,0201	+

Tabel 3: Mittenegatiivsete vähimruutude meetodil saadud hinnangute nihketuse kontroll.

Parameeter	Tegelik väärtus β_i	Hinnangute keskmine $\bar{\hat{\beta}}_i$	$\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	Nihketa
β_A	0,1	0,096	0,095	0,097	$-$
β_B	0	0,005	0,004	0,006	$-$
β_C	0,02	0,02	0,0196	0,0204	$+$

Tabel 4: Suurima tõepära meetodil saadud hinnangute nihketuse kontroll.

Vaadeldud meetodite puhul on vähimruutude meetodil leitud hinnangud ainsana nihketa, mis on ootuspärane. Mittenegatiivne vähimruutude ja suurima tõepära meetodil leitud hinnangud on väikse nihkega.

5 Meetodite võrdlus sekveneerimisvigadega andmete korral

5.1 Parameetrite hinnangud

Kasutasime varem simulatsioonideks mudelimaatriksit

$$\boldsymbol{X} = \begin{bmatrix} 100 & 0 & 0 \\ 0 & 20 & 0 \\ 0 & 0 & 500 \\ 50 & 50 & 0 \\ 200 & 200 & 200 \end{bmatrix}.$$

Sekveneerimisvigade olemasolu korral aga mudel

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{36}$$

ei sobi. Näiteks, kui $\beta_C \geq 0$ ning $\beta_A = 0$ ja $\beta_B = 0$ on $E(y_A) > 0$ ja $(Ey_B) > 0$, sest bakteri C DNA sekveneerimisel tekkivate vigade tõttu näeme ka bakteritele A ja B ainuomaseid k-meere. Nende hulk sõltub sellest, kui lähedased on proovis olevate bakterite DNA kontrollitava bakteri DNA-st.

Oletame, et loeme kokku iga bakteri referentsgenoomis olevad k-meerid, mis on ühe nukleotiidi kaugusel antud lehe või sõlme k-meeridest. Antud olukorda aitab näitlikustada joonis 5. Saame järgmise maatriksi

$$\boldsymbol{X_{1mm}} = \begin{bmatrix} 105 & 12 & 89 \\ 8 & 22 & 45 \\ 0 & 4 & 557 \\ 55 & 67 & 78 \\ 200 & 250 & 234 \end{bmatrix}.$$

Maatriksi $\boldsymbol{X_{1mm}}$ reas 1 ja veerus 2 olev element $\boldsymbol{X_{1mm}}[1,2] = 12$ näitab, et bakteri B referentsgenoomis on 12 k-meeri kõigest ühe tähemärgi kaugusel liigi A unikaalsetest k-meeridest. Kui liigi B k-meeridest 12 on ühe tähemägi kaugusel lehte A kuuluvatest k-meeridest, siis ootuspäraselt tehakse nende 12 k-meeri lugemisel

$$12 \cdot \beta_B \cdot \frac{viga\%}{100}$$

sekveneermisviga ja neist sekveneerimisvigadest keskmiselt $\frac{1}{3}$ ehk

$$12 \cdot \beta_B \cdot \frac{1}{3} \cdot \frac{viga\%}{100}$$

satuvad olema liigi A jaoks unikaalsed k-meerid. Käesolevas töös *viga* on 4%. Seega sekveneerimisviga arvestava mudelimaatriksi saaksime leida järgmisel viisil

$$\boldsymbol{X}^* = \boldsymbol{X} + \boldsymbol{X_{1mm}} \cdot \frac{1}{3} \cdot \frac{4}{100}. \tag{37}$$

Simulatsioonide abil uurime, mis muutub, kui kasutada mudelimaatriksit $\boldsymbol{X}^*$ või asendades mudelimaatriks $\boldsymbol{X}^*$ mudelimaatriksiga $\boldsymbol{X}$. Kui andmete simuleerimise protsessi on korratud 1000 korda, siis saime järgmised tulemused

Meetod	Keskmine ruutviga	Usaldusintervall ruutveale
Vähimruutude meetod	$1{,}568 \cdot 10^{-3}$	$(1{,}429 \cdot 10^{-3};\ 1{,}706 \cdot 10^{-3})$
Mittenegatiivne vähimruutude meetod	$7{,}143 \cdot 10^{-4}$	$(6{,}183 \cdot 10^{-4};\ 8{,}102 \cdot 10^{-4})$
Suurima tõepära meetod	$2{,}149 \cdot 10^{-4}$	$(1{,}694 \cdot 10^{-4};\ 2{,}604 \cdot 10^{-4})$

Tabel 5: Parameetri β_B hinnangute täpsus kasutades mudelimaatriksit $\boldsymbol{X}$.

Meetod	Keskmine absoluutviga	Usaldusintervall absoluutveale
Vähimruutude meetod	$3{,}133 \cdot 10^{-2}$	$(2{,}983 \cdot 10^{-2};\ 3{,}283 \cdot 10^{-2})$
Mittenegatiivne vähimruutude meetod	$1{,}493 \cdot 10^{-2}$	$(1{,}356 \cdot 10^{-2};\ 1{,}631 \cdot 10^{-2})$
Suurima tõepära meetod	$5{,}686 \cdot 10^{-3}$	$(4,848 \cdot 10^{-3};\ 6,525 \cdot 10^{-3})$

Tabel 6: Parameetri β_B hinnangute erinevus kasutades mudelimaatriksit $\boldsymbol{X}$.

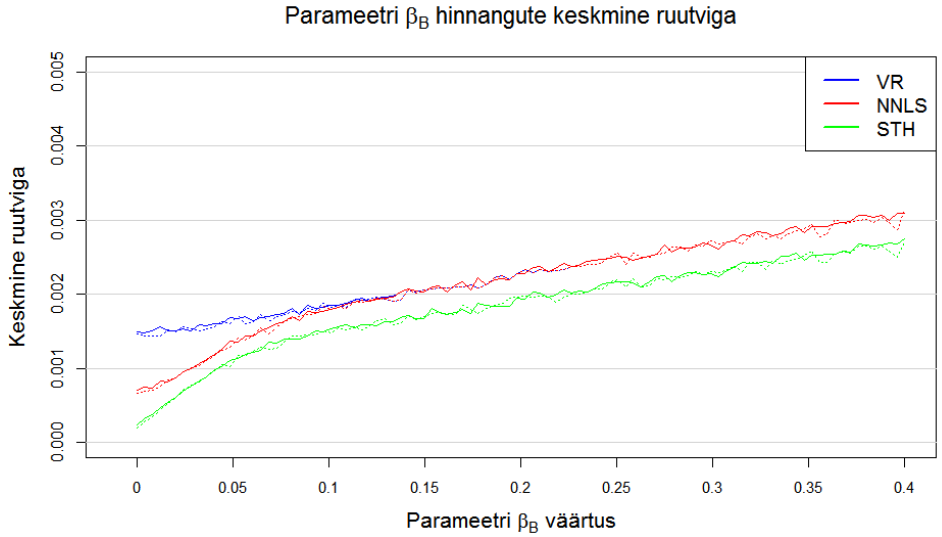
Meetod	Keskmine ruutviga	Usaldusintervall ruutveale
Vähimruutude meetod	$1{,}521 \cdot 10^{-3}$	$(1{,}387 \cdot 10^{-3};\ 1{,}655 \cdot 10^{-3})$
Mittenegatiivne vähimruutude meetod	$6{,}987 \cdot 10^{-4}$	$(6{,}05 \cdot 10^{-4};\ 7{,}923 \cdot 10^{-4})$
Suurima tõepära meetod	$2{,}149 \cdot 10^{-4}$	$(1{,}694 \cdot 10^{-4};\ 2{,}604 \cdot 10^{-4})$

Tabel 7: Parameetri β_B hinnangute täpsus kasutades mudelimaatriksit $\boldsymbol{X}^*$.

Meetod	Keskmine absoluutviga	Usaldusintervall absoluutveale
Vähimruutude meetod	$3{,}087 \cdot 10^{-2}$	$(2{,}939 \cdot 10^{-2};\ 3{,}235 \cdot 10^{-2})$
Mittenegatiivne vähimruutude meetod	$1{,}48 \cdot 10^{-2}$	$(1{,}344 \cdot 10^{-2};\ 1{,}616 \cdot 10^{-2})$
Suurima tõepära meetod	$5{,}686 \cdot 10^{-3}$	$(4{,}848 \cdot 10^{-3};\ 6{,}525 \cdot 10^{-3})$

Tabel 8: Parameetri β_B hinnangute erinevus kasutades mudelimaatriksit $\boldsymbol{X}^*$.

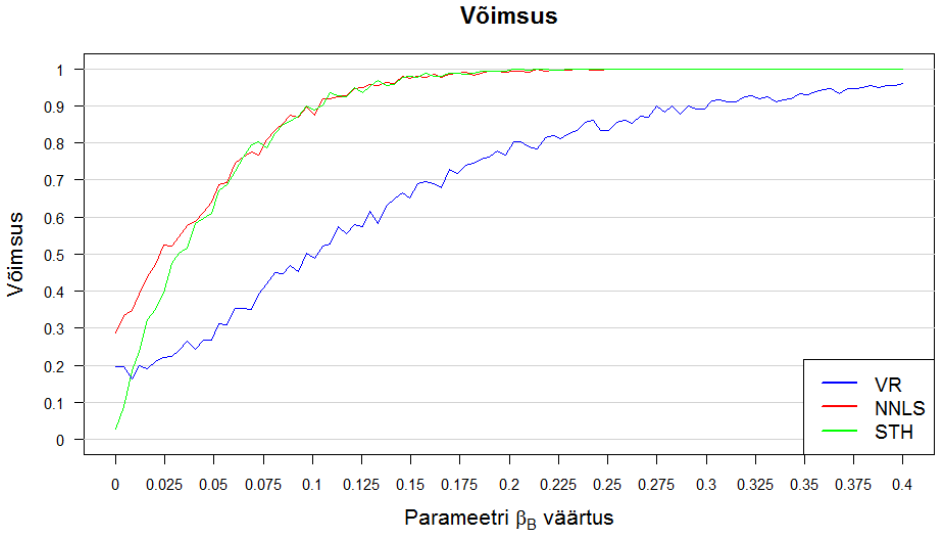
Sekveneerimisveaga andmetel kasutades mudelimaatriksit $\boldsymbol{X}^*$ on parameetri β_B hinnangu täpsus vähimruutude ning mittenegatiivsete vähimruutude meetodil pisut parem, kui kasutades mudelimaatriksit $\boldsymbol{X}$. Suurima tõepära meetodil jääb hinnangu täpsus samaks. Üldiselt on kolmest meetodist täpseim suurima tõepära meetod, millele järgnevad vastavalt mittenegatiivne vähimruutude meetod ja vähimruutude meetod. Samadele tulemustele jõudsime ka sekveneerimisvigadeta andmetel.



Joonis 15: Parameetri β_B hinnangute vähimruutude meetodi, mittenegatiivsete vähimruutude meetodi ning suurima tõepära meetodi keskmine ruutviga sekveneerimisveaga andmetel, kui β_B varieerub lõigus $[0;\ 0{,}4]$. Punktiirjoonega on tähistatud meetodite keskmised ruutvead sekveneermisveata andmetel.

Joonisel 15 on kujutatud parameetri β_B keskmist ruutviga kõigil kolmel meetodil, kui β_B varieerub lõigus $[0;\ 0{,}4]$, kuhu on lisatud vastavat värvi punktiirjoonega keskmised ruutvead sekveneerimisvigadeta andmetel. Antud jooniselt on hästi näha, et sekveneerimisviga ei mõjuta mudeli hinnangute täpsust oluliselt.

Meenutame, et sekveneerimisel tehtud võimalike vigade arv jääb vahemikku 0,5–1 ühe nukleotiidi kohta [11]. Järgmisel joonisel 16 on toodud vähimruutude meetodi, mittenegatiivsete vähimruutude meetodi ja suurima tõepära meetodi võimsused. Joonisel on lõigust $[0;\ 0{,}4]$ võetud ühtlaste vahedega 100 punkti ning iga punkti kohta on korratud teste 1000 korda kõigil kolmel meetodil. Saadud tulemused ei erine oluliselt joonisel 14 esitatud tulemustest ning võib järeldada, et sekveneerimisviga ei avalda suurt mõju meetodite võimsusele.



Joonis 16: Vähimruutude meetodi (t-test), mittenegatiivsete vähimruutude meetodi ja suurima tõepära meetodi (tõepärasuhte test) võimsused sekveneerimisveaga andmetel, kui β_B varieerub lõigus $[0;\ 0{,}4]$.

Leidsime sekveneerimisveaga andmetel kolmel meetodil hinnatud parameetri $\boldsymbol{\beta} = (\beta_A, \beta_B, \beta_C)$ hinnangute keskmisele usaldusintervallid, mille abil hindame, kas parameetrite hinnangud on nihkega või nihketa.

Parameeter	Tegelik väärtus β_i	Hinnangute keskmine $\bar{\hat{\beta}}_i$	$\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	Nihketa
β_A	0,1	0,103	0,102	0,105	−
β_B	0	-0,001	-0,004	0,001	+
β_C	0,02	0,02	0,0199	0,0207	+

Tabel 9: Vähimruutude meetodil saadud hinnangute nihketuse kontroll sekveneerimisveaga andmete korral.

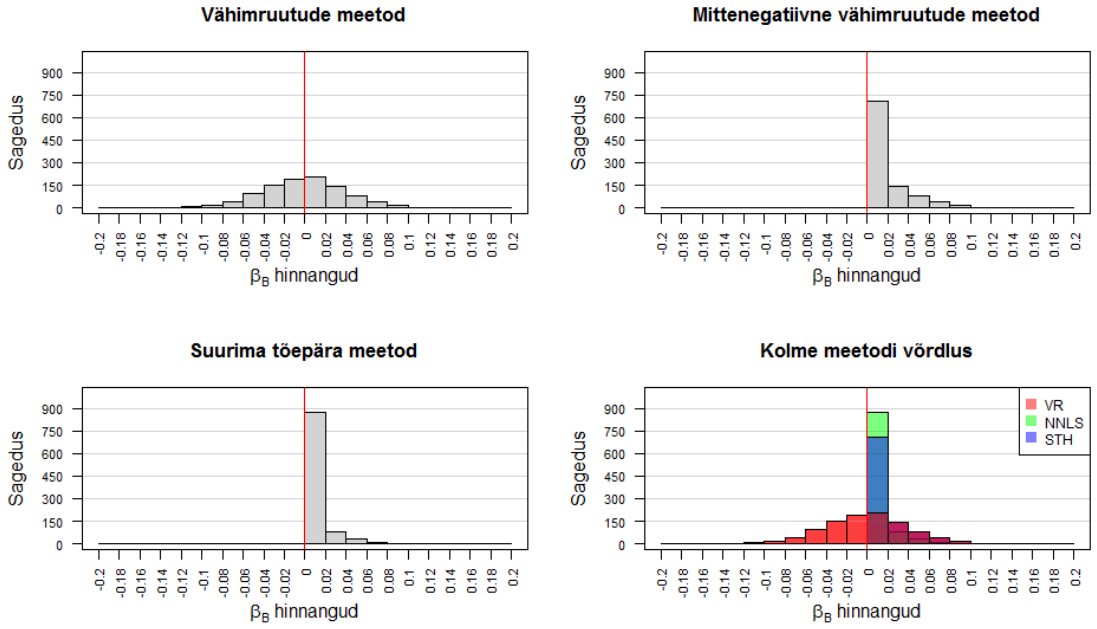
Parameeter	Tegelik väärtus β_i	Hinnangute keskmine $\bar{\hat{\beta}}_i$	$\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	Nihketa
β_A	0,1	0,091	0,089	0,092	−
β_B	0	0,015	0,014	0,016	−
β_C	0,02	0,0198	0,019	0,02	+

Tabel 10: Mittenegatiivne vähimruutude meetodil saadud hinnangute nihketuse kontroll sekveneerimisveaga andmete korral.

Parameeter	Tegelik väärtus β_i	Hinnangute keskmine $\bar{\hat{\beta}}_i$	$\bar{\hat{\beta}}_i - t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	$\bar{\hat{\beta}}_i + t_{1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$	Nihketa
β_A	0,1	0,098	0,097	0,099	−
β_B	0	0,006	0,005	0,007	−
β_C	0,02	0,020	0,020	0,021	+

Tabel 11: Suurima tõepära meetodil saadud hinnangute nihketuse kontroll sekveneerimisveaga andmete korral.

Kui sekveneerimisvigadeta andmete puhul oli vähimruutude meetodi hinnangud nihketa ning mittenegatiivne vähimruutude ja suurima tõepära meetodi hinnangud väikse nihkega, siis sekveneerimisvigadega andmetel on kõik kolm meetodit väikse nihkega.



Joonis 17: Parameetri β_B hinnangud sekveneerimisveaga andmetel kolmel meetodil. Parameetri $\beta_B = 0$ tegelik väärtus on märgitud punase joonega.

Jooniselt 17 näeme, et parameetri β_B õige väärtuse (0) lähedale on kõige sagedamini sattunud jällegi suurima tõepära meetodil leitud hinnangud. Ka β_A ja β_C hinnangud on sarnased peatükis 4.1 toodud hinnangutele ning nendel hetkel pikemalt ei peatu. Vastavad hinnangute joonised on toodud lisas A joonisel 25, 26 ja 27. Ühtlasi on lisas A toodud parameetervektori $\boldsymbol{\beta}$ hinnangute hajuvusdiagrammid (joonisel 29, 30 ja 31).

6 Tegelike sekveneerimisandmete analüüs

Uuritavad vaatlusandmestikud on koostanud bioinformaatika nooremteadur Mihkel Vaher ning ülevaade andmetest on toodud peatükis 1.2. Uuritavad andmed on 74 bakterit ning 73 sõlmekohta. Esmalt koostame maatriksi $\boldsymbol{X}$, mille reanimedeks määrame bakterite ja sõlmede nimetused ning veerunimedeks määrame bakterite nimetused. Reas i ning veerus j asub element $\boldsymbol{X}[i,j]$, mis näitab mitu i-nda rea bakteri või sõlme k-meeri sisaldub j-veeru bakteris. Maatriks $\boldsymbol{X}_{1mm}$ on mõõtmetega 147×74, mille rea- ning veerunimed kattuvad maatriksi $\mathbf{X}$ omadega. Antud maatriksi i-ndas reas ja j-ndas veerus $\boldsymbol{X}_{1mm}[i,j]$ asub element, kui mitu i-nda rea bakteri või sõlme omast k-meeri on ühe nukleotiidi kaugusel veerus j asuvale bakterile omastest k-meeridest (olukorda näitlikustab joonis 5). Vektori $\mathbf{Y}$ pikkusega 147 reanimedeks määrame samad reanimed, mis maatriksi $\mathbf{X}$ puhul. Vektori $\mathbf{Y}$ i-nda rea $\mathbf{Y}[i]$ elemendiks on bakteri või sõlme nähtud k-meeride arv. Siinkohal

$$\boldsymbol{X^*} = (\frac{viga}{3})\boldsymbol{X_{1mm}} + \boldsymbol{X} = (\frac{0{,}04}{3})\boldsymbol{X_{1mm}} + \boldsymbol{X}. \tag{38}$$

Varasemate tulemuste põhjal kõige ebatäpsemalt hindab bakterite sekveneerimiskatvust vähimruutude meetod. Nimetatud meetodi puhul oli kolme bakteri katvuse hinnang statistiliselt nullist oluliselt erinev. Nendeks oli

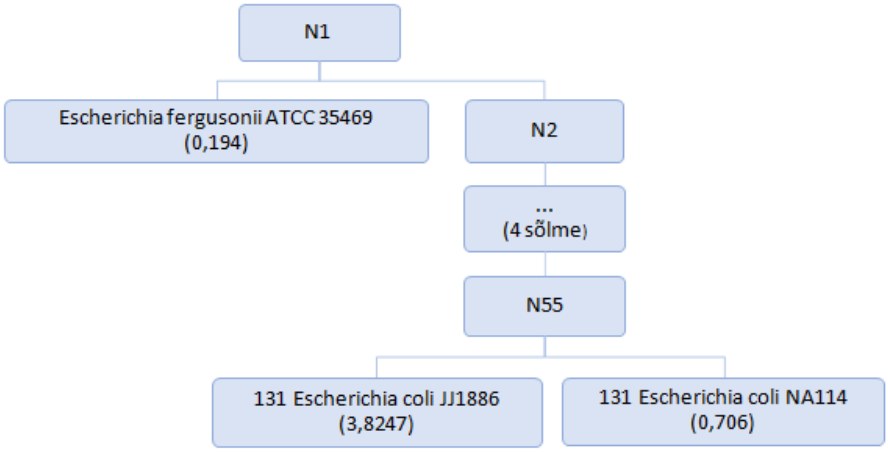
Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,864
131 Escherichia coli NA114	0,714
Escherichia fergusonii ATCC 35469	0,195

Tabel 12: Vähimruutude meetodi statistiliselt oluliste sekveneerimiskatvuse hinnangud, kasutades mudelimaatriksit $\boldsymbol{X}$.

Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,825
131 Escherichia coli NA114	0,706
Escherichia fergusonii ATCC 35469	0,194

Tabel 13: Vähimruutude meetodi statistiliselt oluliste sekveneerimiskatvuse hinnangud, kasutades mudelimaatriksit $\boldsymbol{X^*}$.

Kasutades mudelimaatriksit $\boldsymbol{X}^*$ on bakteritele vastav fülogeneesipuu järgmine



Joonis 18: Vähimruutude meetodi statistiliselt oluliste sekveneeritud katvuste hinnangute fülogeneesipuu.

Kuna teostatud teste oli palju, siis uurime milliste bakterite sekveneerimiskatvused on nullist suuremad, kui rakendame Bonferroni korrektsiooni (74 parameetrit). Vähimruutude meetodil leitud statistiliselt oluliste sekveneeritud katvuste hinnangud ei pruugi olla õiged. Seetõttu uurime, millised sekveneeritud katvuse hinnangud saame, kui kasutame Bonferroni korrektsiooni (74 parameetrit). Nendeks oli

Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,864
131 Escherichia coli NA114	0,714

Tabel 14: Vähimruutude meetodi sekveneeritud katvuste hinnangud vaatlusandmetel kasutades Bonferroni korrektsiooni (mudelimaatriksit $\boldsymbol{X}$ puhul).

Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,825
131 Escherichia coli NA114	0,706

Tabel 15: Vähimruutude meetodi sekveneeritud katvuste hinnangud vaatlusandmetel kasutades Bonferroni korrektsiooni (mudelimaatriksit $\boldsymbol{X}^*$ puhul).

Vähimruutude meetodi puhul kasutades Bonferroni korrektsiooni saame, et leiduvad bakterid *131 Escherichia coli JJ1886* ja *131 Escherichia coli NA114*. Antud juhul jäi kõrvale bakter *Escherichia fergusonii ATCC 35469*, mille sekveneeritud katvuse hinnang oli statistiliselt mitteoluline.

Simulatsioonide põhjal otsustades hindab sekveneerimiskatvuseid täpsemalt mittenegatiivne vähimruutude meetod. Antud meetod jääb täpsuselt alla küll suurima tõepära meetodile, kuid on täpsem vähimruutude meetodist. Nimetatud meetodi puhul kasutades mudelimaatriksit $\boldsymbol{X}$ on kaheksa bakteri katvuse hinnang nullist erinev ning kasutades mudelimaatriksit $\boldsymbol{X}^*$ on kuue bakteri katvuse hinnang nullist erinev. Nendeks oli

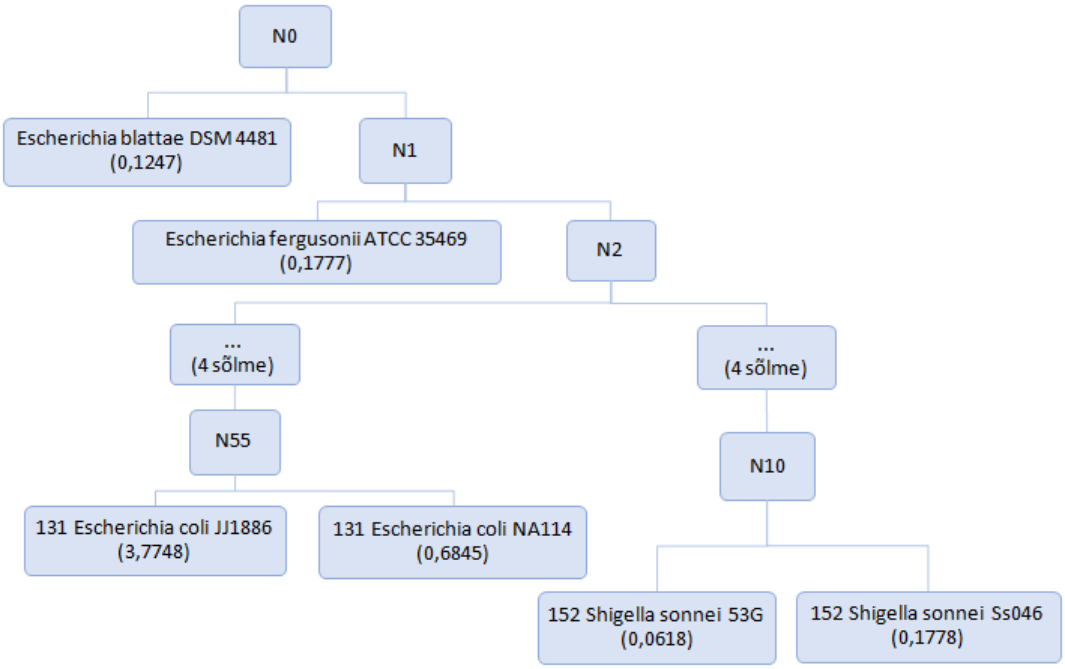
Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,821
131 Escherichia coli NA114	0,695
152 Shigella sonnei 53G	0,065
152 Shigella sonnei Ss046	0,183
62 Escherichia coli IAI39	0,007
62 Escherichia coli O7 K1 CE10	0,002
Escherichia blattae DSM 4481	0,127
Escherichia fergusonii ATCC 35469	0,179

Tabel 16: Mittenegatiivsete vähimruutude meetodi nullist erinevate sekveneeriskatvuste hinnangud, kasutades mudelimaatriksit $\boldsymbol{X}$.

Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,775
131 Escherichia coli NA114	0,685
152 Shigella sonnei 53G	0,062
152 Shigella sonnei Ss046	0,178
Escherichia blattae DSM 4481	0,125
Escherichia fergusonii ATCC 35469	0,178

Tabel 17: Mittenegatiivsete vähimruutude meetodi nullist erinevate sekveneeriskatvuste hinnangud, kasutades mudelimaatriksit $\boldsymbol{X}^*$.

Kasutades mudelimaatriksit $\boldsymbol{X}^*$ on bakteritele vastav fülogeneesipuu järgmine



Joonis 19: Mittenegatiivsete vähimruutude meetodi nullist erinevate sekveneeritud katvuste hinnangute fülogeneesipuu.

Kõige täpsemini kolmest meetodist hindab bakterite sekveneeritud katvust suurima tõepära meetod. Antud juhul kasutades mudelimaatriksit $\boldsymbol{X}$ neljakümne üheksa bakteri katvuse hinnang nullist erinev ning kasutades mudelimaatriksit $\boldsymbol{X}^*$ kuue bakteri katvuse hinnang nullist erinev. Nendeks oli

Bakter	Katvuse hinnang
100 Escherichia coli UMNK88	0,001
1079 Escherichia coli W	0,001
10 Escherichia coli P12b	0,002
1120 Escherichia coli ATCC 8739	0,001
1128 Escherichia coli IAI1	0,001
1129 Shigella boydii CDC 3083 94	0,001
1130 Shigella boydii Sb227	0,001
1132 Escherichia coli E24377A	0,001
11 Escherichia coli O157 H7 Sakai	0,001
127 Escherichia coli 536	0,001
131 Escherichia coli JJ1886	2,013
131 Escherichia coli NA114	1,503
131 Escherichia coli SE15	0,609
135 Escherichia coli LF82	0,001
146 Shigella dysenteriae 1617	0,001
146 Shigella dysenteriae Sd197	0,001
152 Shigella sonnei 53G	0,001
152 Shigella sonnei Ss046	2,013
156 Escherichia coli SE11	1,503
15 Escherichia coli O127 H6 E2348 69	0,609
16 Escherichia coli O111 H 11128	0,0004
17 Escherichia coli O103 H2 12009	0,001
21 Escherichia coli O26 H11 11368	0,001
23 Escherichia coli APEC O78	0,002
245 Shigella flexneri 2002017	0,001
245 Shigella flexneri 2a 301	0,001
335 Escherichia coli O55 H7 CB9615	0,001
354 Escherichia coli SMS 3 5	0,005
414 Escherichia coli 042	0,068
452 Escherichia coli ED1a	0,001
46 Escherichia coli HS	0,002
48 Escherichia coli ETEC H10407	0,009
597 Escherichia coli UMN026	0,016
62 Escherichia coli IAI39	0,002
62 Escherichia coli O7 K1 CE10	0,001
634 Shigella flexneri 5 8401	0,001
678 Escherichia coli 55989	0,001
678 Escherichia coli O104 H4 2009EL 2050	0,001

(Tabel jätkub järgmisel lehel)

(Tabel jätkub)

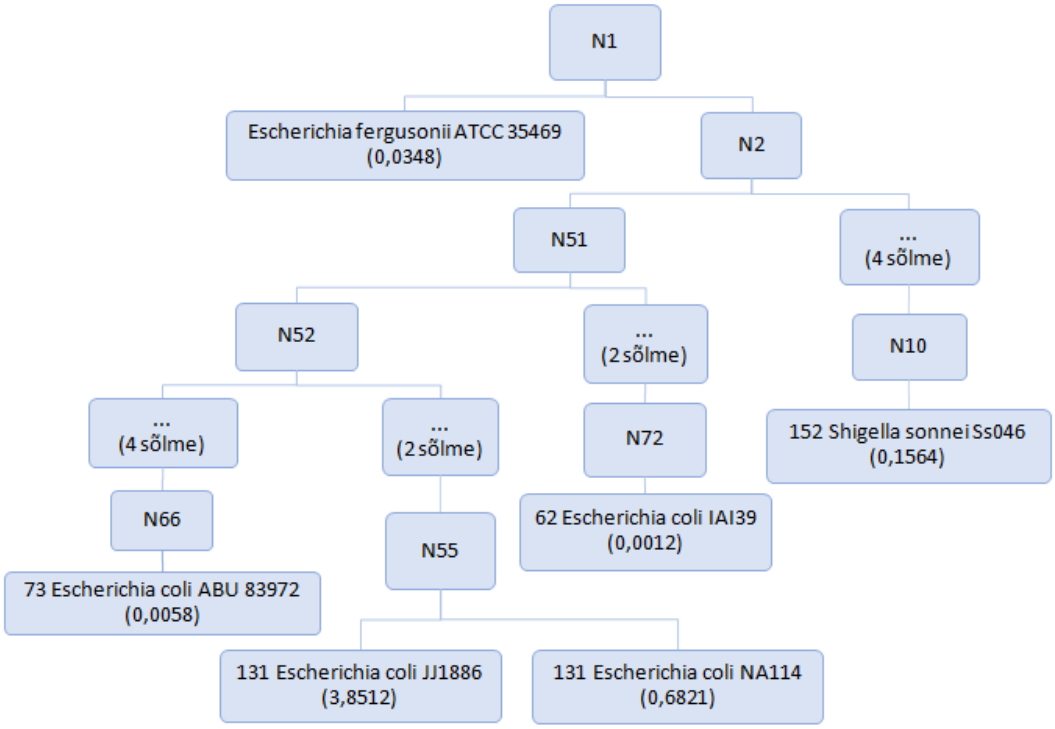
Bakter	Katvuse hinnang
678 Escherichia coli O104 H4 2009EL 2071	0,001
678 Escherichia coli O104 H4 2011C 3493	0,018
73 Escherichia coli ABU 83972	4,394
73 Escherichia coli CFT073	0,001
73 Escherichia coli clone D i14	0,001
93 Escherichia coli BL21 DE3	0,001
95 Escherichia coli PMV	0,0001
95 Escherichia coli S88	0,002
95 Escherichia coli UTI89	0,001
Escherichia blattae DSM 4481	0,001
Escherichia fergusonii ATCC35469	0,073

Tabel 18: Suurima tõepära meetodi nullist erinevate sekveneerimiskatvuste hinnangud, kasutades mudelimaatriksit $\boldsymbol{X}$.

Bakter	Katvuse hinnang
131 Escherichia coli JJ1886	3,851
131 Escherichia coli NA114	0,682
152 Shigella sonnei Ss046	0,156
62 Escherichia coli IAI39	0,001
73 Escherichia coli ABU 83972	0,006
Escherichia fergusonii ATCC 35469	0,035

Tabel 19: Suurima tõepära meetodi nullist erinevate sekveneerimiskatvuste hinnangud, kasutades mudelimaatriksit $\boldsymbol{X}^*$.

Kasutades mudelimaatriksit $\boldsymbol{X}^*$ on bakteritele vastav fülogeneesipuu järgmine



Joonis 20: Suurima tõepära meetodi nullist erinevate sekveneeritud katvuste hinnangute fülogeneesipuu.

Tähele tasub panna, et kaks täpsemat meetodit - mittenegatiivne vähimruutude meetod ja suurima tõepära meetod, hindasid kuus bakterit sekveneerimiskatvusega erinevaks nullist. Vähimruutude meetodil oli kõigi bakterite hinnangulised katvused nullist erinevad, 43 bakteri sekveneerimiskatvus leiti olevat negatiivne. Samas hinnatud bakteritest luges t-test nullist statistiliselt oluliseks vaid 3 bakteri katvused (*131 Escherichia coli JJ1886*, *131 Escherichia coli NA114* ja *Escherichia fergusonii ATCC 35469*). Ühtlasi nimetatud bakterite sekveneermiskatvust hindasid nullist erinevaks kõik kolm meetodit. Kõige kõrgemaks hindas bakteri *131 Escherichia coli JJ1886* katvuse suurima tõepära meetod 3,85. Bakteri *131 Escherichia coli NA114* katvuse hindas kõrgemaiks vähimruutude meetod sekveneeritud katvusega 0,71. Bakteri *Escherichia fergusonii ATCC 35469* katvuse hindas kõrgemaiks vähimruutude meetod sekveneeritud katvusega 0,18.

7 Kokkuvõte

Käesolevas töös vaadeldi kolme meetodit bakterite sekveneerimiskatvuste hindamiseks. Võrreldi nii hinnangute täpsust kui ka erinevatele hindamismeetoditele tuginevate statistiliste testide töökindlust ja headust. Simulatsioonides osutusid täpseimaks suurima tõepära meetodi abil saadud hinnangud, kuid saadud hinnangud olid väikese nihkega.

Testides juhtu, kus bakteriliigi sekveneerimiskatvus oli suurem kui 0 (antud bakteri DNA-d esines sekveneeritud proovis) osutus ainsana usaldusväärseks tõepärasuhte test, mis oli konservatiivne, kuid ta ei teinud lubatust sagedamini I-liiki viga. Teised vaadeldud testid - t-test (vähimruutude hinnang) ja normaaljaotuse eeldust kasutav tõepärasuhte test (lisamoodul **penalized**, mittenegatiivsete vähimruutude meetodi implementatsioon), arvutasid olulisustõenäosuse selgelt valesti tehes I liiki viga lubatust märksa sagedamini.

Antud meetodeid rakendati ka ühe reaalse proovi sekveneerimisel saadud andmete analüüsimisel.

Kaks täpsemat meetodit- mittenegatiivne vähimruutude meetod ja suurima tõepära meetod, hindasid vaatlusandmetel kuus bakterit sekveneerimiskatvusega erinevaks nullist. Vähimruutude meetodil oli kõigi bakterite hinnangulised katvused nullist erinevad, 43 bakteri sekveneerimiskatvus leiti olevat negatiivne. Samas hinnatud bakteritest luges t-test nullist statistiliselt oluliseks vaid 3 bakteri katvused.

Käesoleva töö autoril õnnestus ka näidata, et antud suurusega ülesande puhul (kontrolliti 74 bakteri olemasolu proovis ehk tegemist oli 74 hinnatava parameetriga) on täiesti võimalik ka pärisandmete korral kasutada suurima tõepära meetodit - kuigi parameetrite hinnangute leidmiseks tuli kasutada numbrilisi meetodeid.

Viited

[1] **A. Björck**, *„Numerical Methods for Least Squares Problems"*, Society for Industrial and Applied Mathematics (1996).

[2] **E. S. Lande, M. S. Waterman**, *"Genomic Mapping by Fingerprinting Random Clones: A Mathematical Analysis"*, Genomics (1988).

[3] **C. L Lawson, R. J. Hanson**, *"Solving Least Squares Problems"*, Society for Industrial and Applied Mathematics (1995).

[4] **M. Mihkelsaar, T. Karki, I. Lutsar, R. Mändar**, *„Meditsiiniline mikrobiloogia, I osa"*, Tartu Ülikooli kirjastus (2006).

[5] **G. Molenberghs, G. Verbeke**, *"Likelihood Ratio, Score, and Wald Tests in a Constrained Parameter Space"*, The American Statistician (2012).

[6] **K. M. Mullen, I. H. M. van Stokkum**. *Package 'nnls'*, [www]`https://cran.r-project.org/web/packages/nnls/nnls.pdf` (avaldatud 19.03.2012, vaadatud 01.05.2019).

[7] **M. Möls**, *Kursuse „Lineaarsed mudelid" loengukonspekt*, Tartu Ülikooli matemaatika ja statistika instituut (2018).

[8] **M. Roosaare**, *„K-mer based methods for the identification of bacteria and plasmids"*, University of Tartu (2018).

[9] **M. Roosaare, M. Vaher, L. Kaplinski, M. Mols, R. Andreson, M. Lepamets, T. Kõressaar, P. Naaber, S. Kõljalg, M. Remm** , *„StrainSeeker: fast identification of bacterial strains from raw sequencing reads using user-provided guide trees"*, PeerJ (2019).

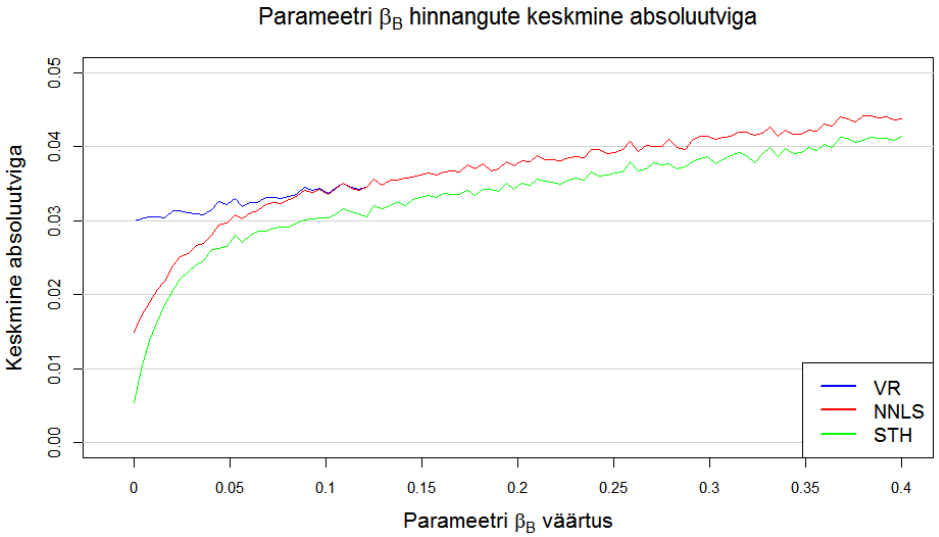
[10] **G. G. Z. Silva, D. A. Cuevas, B. E. Dutilh, R. A. Edwards**, *"FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares"*, PeerJ (2014).

[11] **G. Tan, L. Opitz, R. Schlapbach, H. Rehrauer**, *"Long fragments achieve lower base quality in Illumina paired-end sequencing"*, Scientific Reports (2019).

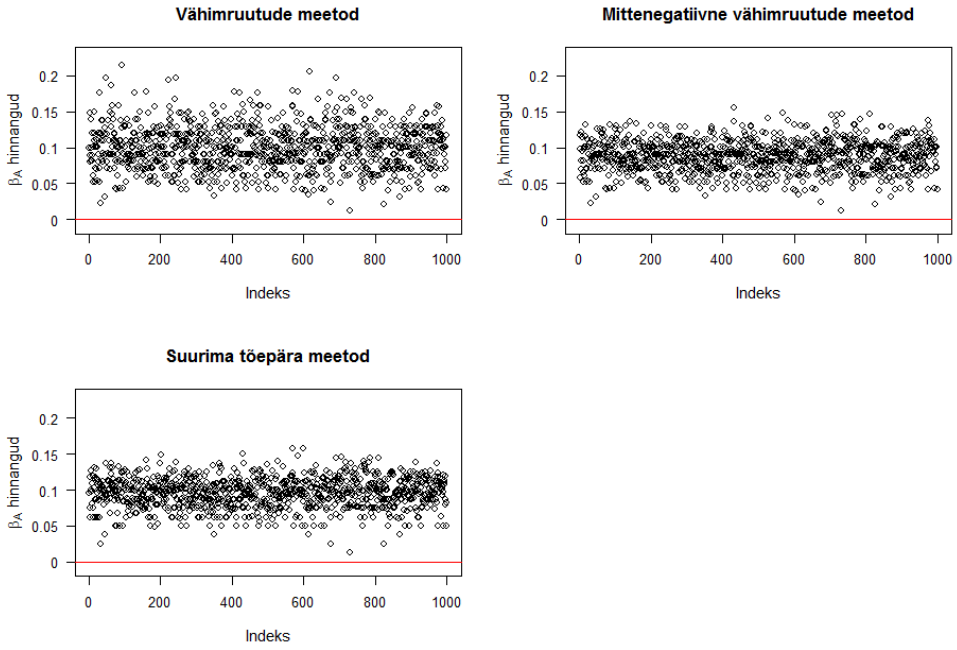
[12] **I. Traat**, *Kursuse „Matemaatiline statistika II" loengukonspekt*, Tartu Ülikooli matemaatika ja statistika instituut (2018).

Lisad

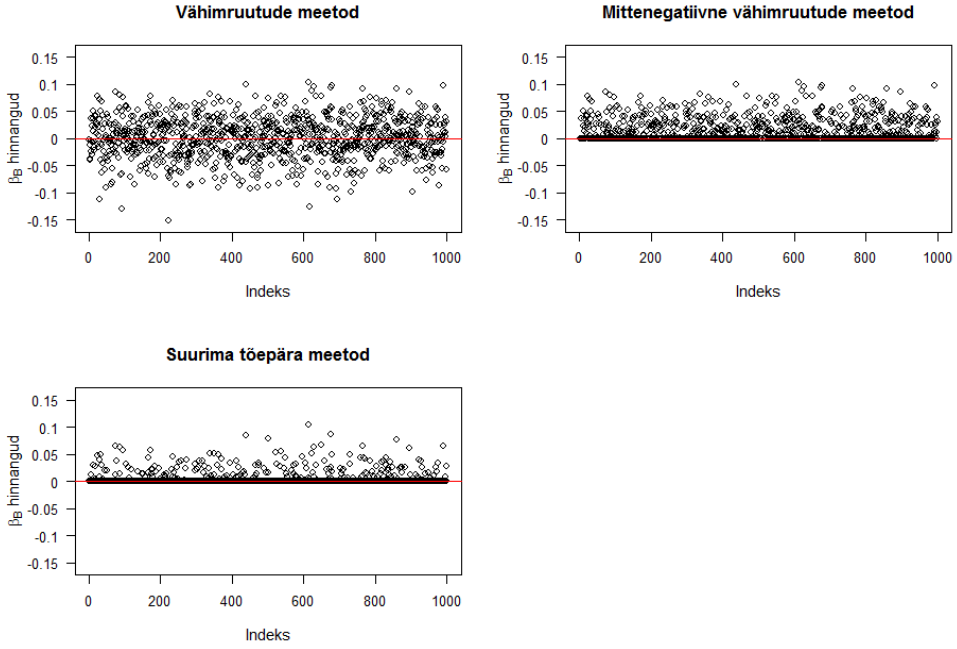
A Joonised



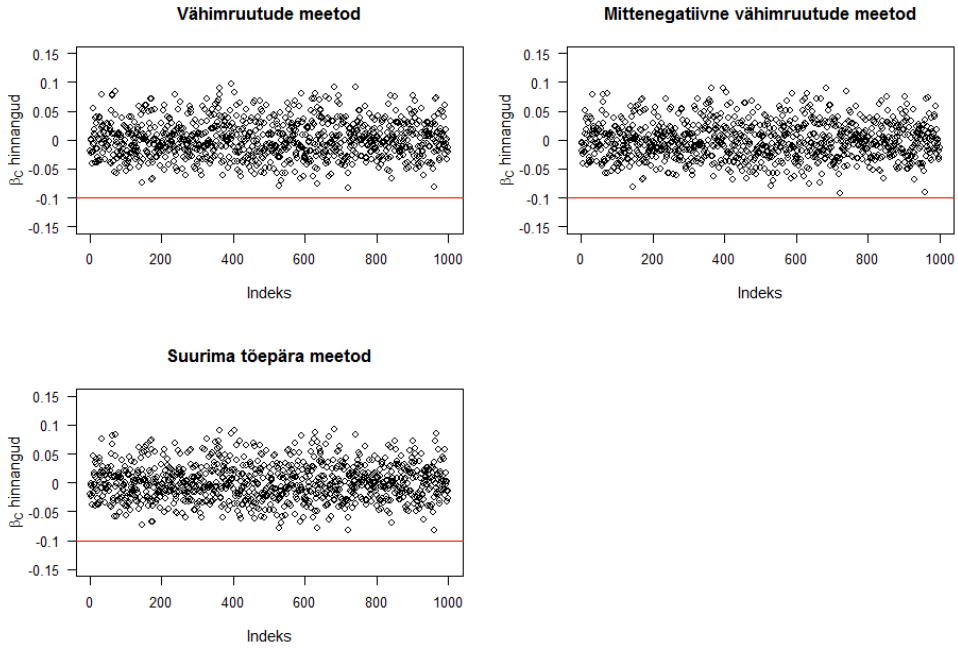
Joonis 21: Parameetri β_B hinnangute vähimruutude meetodi, mittenegatiivsete vähimruutude meetodi ning suurima tõepära meetodi keskmine absoluutviga, kui β_B varieerub lõigus $[0;\ 0{,}4]$.



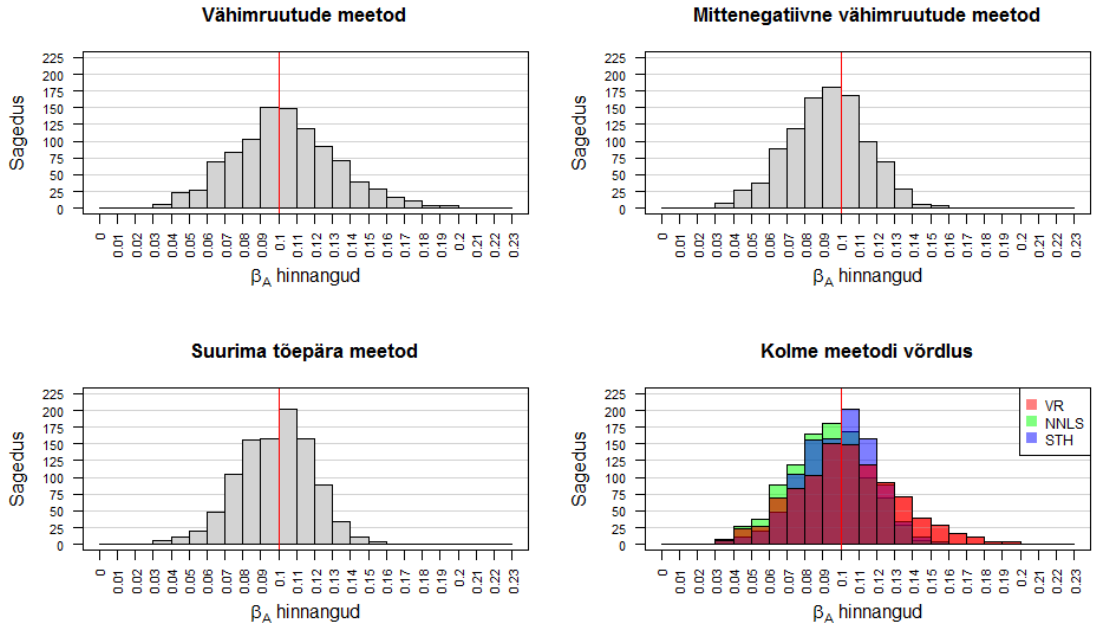
Joonis 22: Parameetri β_A hinnangud kolmel meetodil, kus $\beta_A = 0{,}1$ tegelik väärtus on märgitud punase joonega.



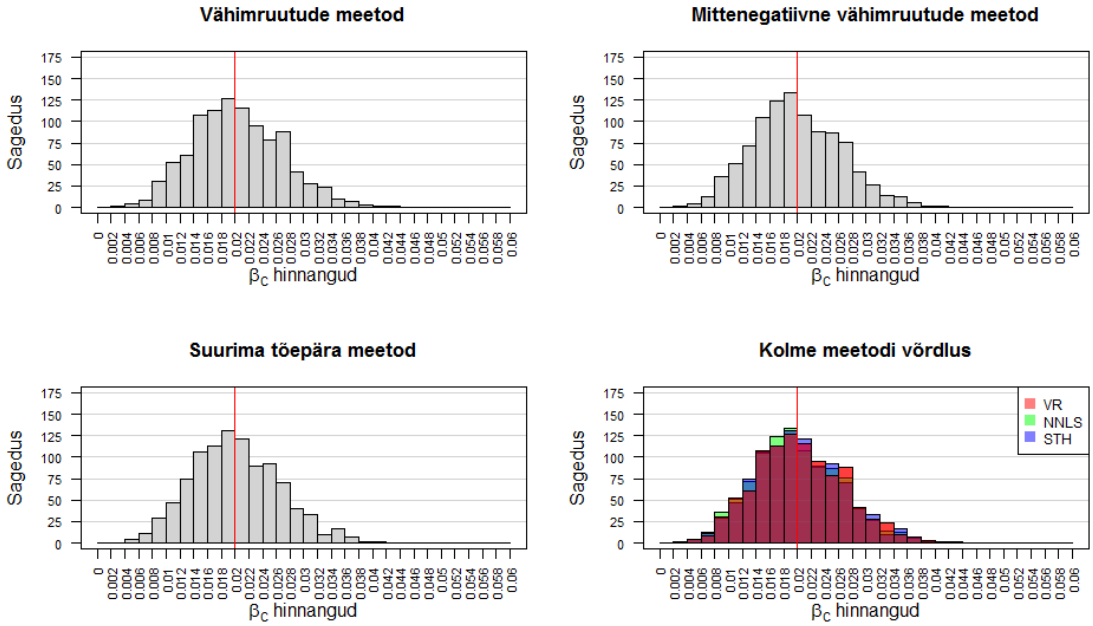
Joonis 23: Parameetri β_B hinnangud kolmel meetodil, kus $\beta_B = 0$ tegelik väärtus on märgitud punase joonega.



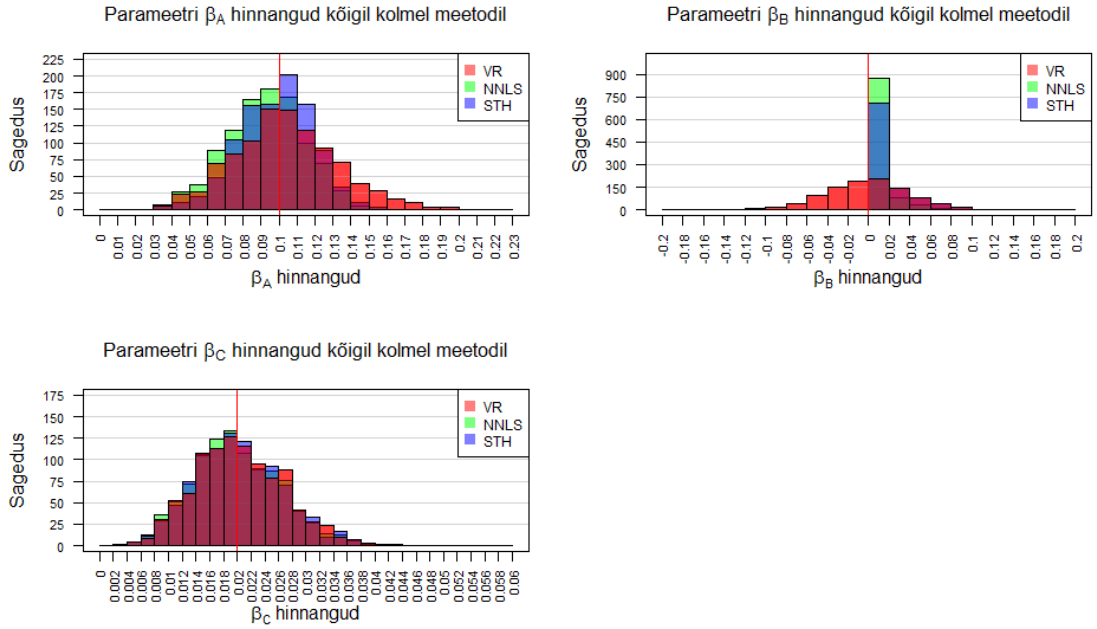
Joonis 24: Parameetri β_C hinnangud kolmel meetodil, kus $\beta_C = 0{,}02$ tegelik väärtus on märgitud punase joonega.



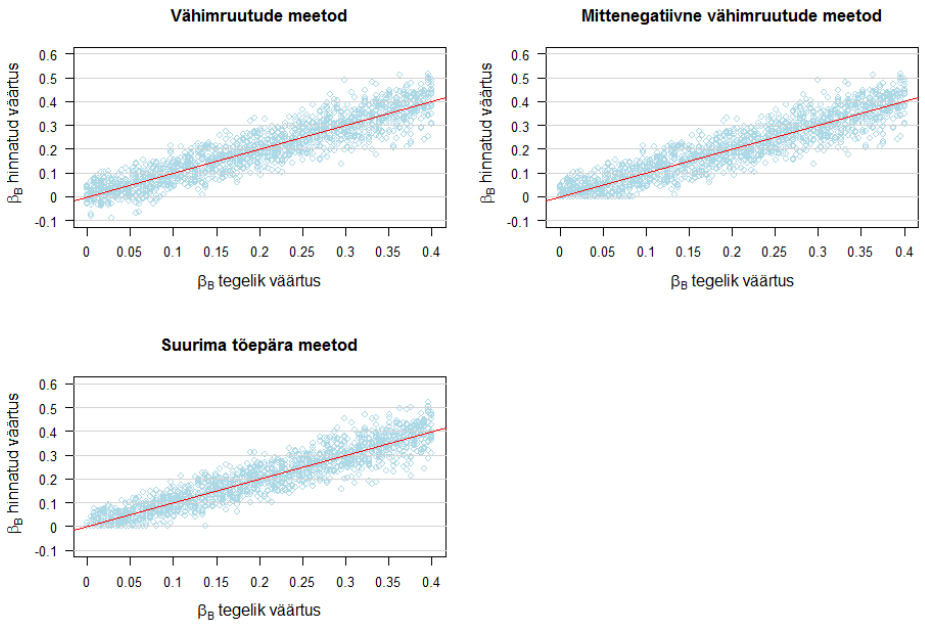
Joonis 25: Parameetri β_A hinnangud sekveneerimisveaga andmetel kolmel meetodil. Parameetri $\beta_A = 0{,}1$ tegelik väärtus on märgitud punase joonega.



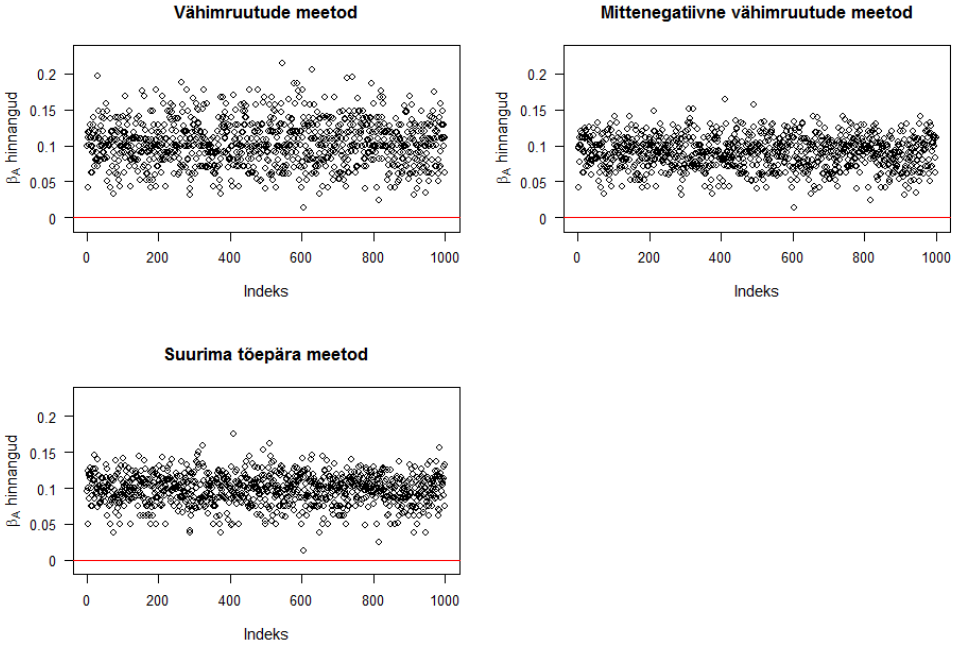
Joonis 26: Parameetri β_C hinnangud sekveneerimisveaga andmetel kolmel meetodil. Parameetri $\beta_C = 0{,}02$ tegelik väärtus on märgitud punase joonega.



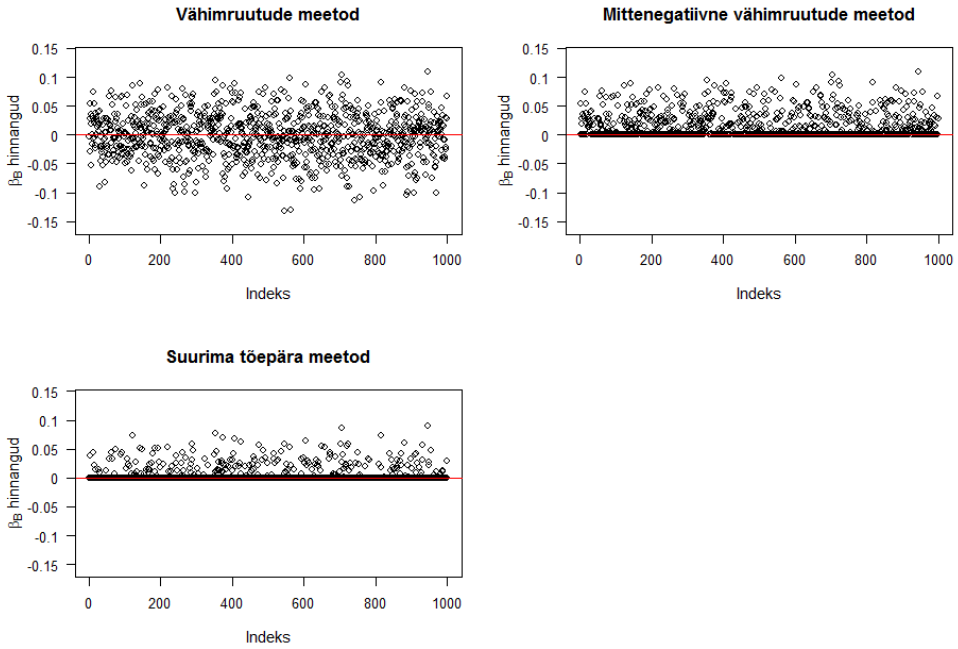
Joonis 27: Parameetrite β_A, β_B ja β_C hinnangud sekveneerimisveaga andmetel kõigil kolmel meetodil. Parameetri tegelik väärtus on märgitud punase joonega.



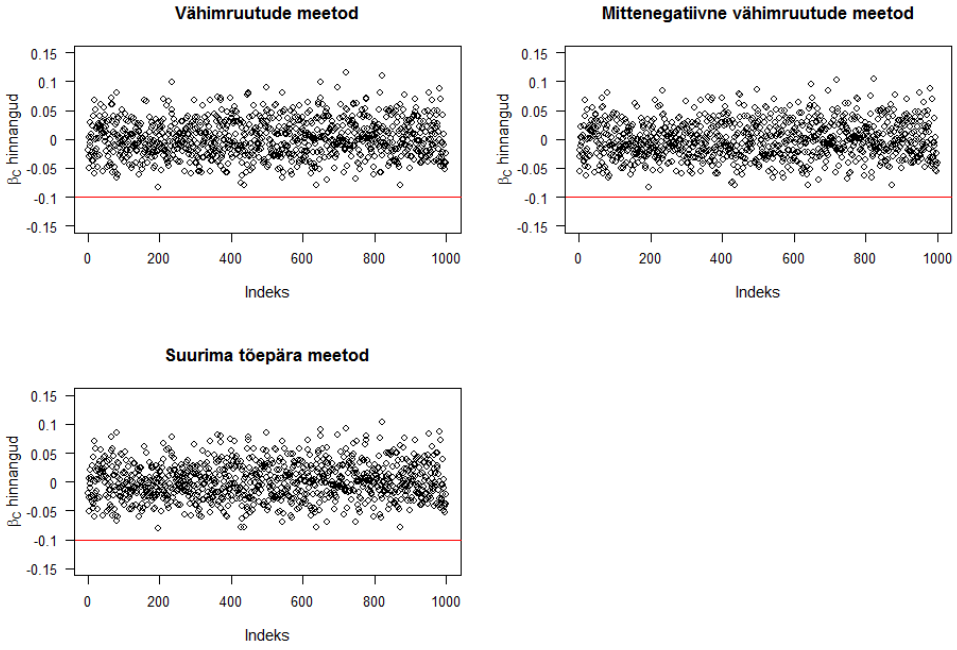
Joonis 28: Parameetri β_B hinnangud kolmel meetodil sekveneerimisveaga andmetel, kui β_B varieerub lõigus $[0;\ 0{,}4]$. Joonistel punane joon tähistab võrdust $\beta_B = \hat{\beta}_B$.



Joonis 29: Parameetri β_A hinnangud sekveneerimisveaga andmetel kolmel meetodil, kus $\beta_A = 0{,}1$ tegelik väärtus on märgitud punase joonega.



Joonis 30: Parameetri β_B hinnangud sekveneerimisveaga andmetel kolmel meetodil, kus $\beta_B = 0$ tegelik väärtus on märgitud punase joonega.



Joonis 31: Parameetri β_C hinnangud sekveneerimisveaga andmetel kolmel meetodil. Parameetri $\beta_C = 0{,}02$ tegelik väärtus on märgitud punase joonega.

B Kood

Programm on kirjutatud programmeerimiskeeles R (versioon 3.6.1).

Autori poolt defineeritud funktsioonid

```
keskmine_ruutviga<- function(tegelik, hinnatud){
  n<- length(tegelik)
  mse<- 1/n*sum((tegelik-hinnatud)**2)
  return(mse)}

keskmine_absoluutviga<- function(tegelik, hinnatud){
  n<- length(tegelik)
  mae<- 1/n*sum(abs(tegelik-hinnatud))
  return(mae)}

vahimruutude_meetod<- function(matX,Y){
  return(ginv(t(matX)%*%matX)%*%t(maX)%*%Y)}

logtoepara_pois<-function(beeta, y){
  konstant<- c(beeta[1]*350+270*beeta[2]+700*beeta[3])
  n<- length(y)
  logl<- log(beeta[1]*100)*y[1]+log(beeta[2]*20)*y[2]+
    log(beeta[3]*500)*y[3]+log(50*(beeta[1]+beeta[2]))*y[4]+
    log(200*(beeta[1]+beeta[2]+beeta[3]))*y[5]-konstant-
    lgamma(y[1]+1)-lgamma(y[2]+1)-lgamma(y[3]+1)-
    lgamma(y[4]+1)-lgamma(y[5]+1)
  return(logl)}

#Vektoris beeta on algselt kaks elementi. Funktsiooni sees
#lisatakse juurde beeta1=1e-8.
logtoepara_poisbeeta1<- function(beeta, y){
  beeta<- c(1e-8, beeta)
  konstant<- c(beeta[1]*350+270*beeta[2]+700*beeta[3])
  n<- length(y)
  logl<- log(beeta[1]*100)*y[1]+log(beeta[2]*20)*y[2]+
    log(beeta[3]*500*y[3]+log(50*(beeta[1]+beeta[2]))*y[4]+
    log(200*(beeta[1]+beeta[2]+beeta[3]))*y[5]-konstant-
    lgamma(y[1]+1)-lgamma(y[2]+1)-lgamma(y[3]+1)-
    lgamma(y[4]+1)-lgamma(y[5]+1)
  return(logl)}
```

```
#Vektoris beeta on algselt kaks elementi. Funktsiooni sees
#lisatakse juurde beeta2=1e-8.
logtoepara_poisbeeta2<- function(beeta, y){
  beeta=c(beeta[1], 1e-8, beeta[2])
  konstant<- c(beeta[1]*350+270*beeta[2]+700*beeta[3])
  n<- length(y)
  logl<- log(beeta[1]*100)*y[1]+log(beeta[2]*20)*y[2]+
    log(beeta[3]*500)*y[3]+log(50*(beeta[1]+beeta[2]))*y[4]+
    log(200*(beeta[1]+beeta[2]+beeta[3]))*y[5]-konstant-
    lgamma(y[1]+1)-lgamma(y[2]+1)-lgamma(y[3]+1)-
    lgamma(y[4]+1)-lgamma(y[5]+1)
  return(logl)}

#Vektoris beeta on algselt kaks elementi. Funktsiooni sees
#lisatakse juurde beeta3=1e-8.
logtoepara_poisbeeta3<-function(beeta, y){
  beeta<- c(beeta, 1e-8)
  konstant<- c(beeta[1]*350+270*beeta[2]+700*beeta[3])
  n<- length(y)
  logl<- log(beeta[1]*100)*y[1]+log(beeta[2]*20)*y[2]+
    log(beeta[3]*500)*y[3]+log(50*(beeta[1]+beeta[2]))*y[4]+
    log(200*(beeta[1]+beeta[2]+beeta[3]))*y[5]-konstant-
    lgamma(y[1]+1)-lgamma(y[2]+1)-lgamma(y[3]+1)-
    lgamma(y[4]+1)-lgamma(y[5]+1)
  return(logl)}
```

Sekveneerimisveata andmete genereerimine ja hinnangute leidmine

```
set.seed(111)

hinnangud_lm<- list()
lm_mse<- c()
lm_mae<- c()
hinnangud_nnls<- list()
nnls_mse<- c()
nnls_mae<- c()
nnls_beeta1<- c()
nnls_beeta2<- c()
nnls_beeta3<- c()
tulemused_penalized<-matrix(rep( 0, len=3*1000), nrow=1000,
   ncol=3)
hinnangud_sth<- list()
```

```
sth_mse<- c()
sth_mae<- c()
loglik1_vordlus<- c()
loglik2_vordlus<- c()
loglik3_vordlus<- c()
hinnangud_vr<- list()
vr_mse<- c()
vr_mae<- c()
loglik_beeta1_count<- c()
loglik_beeta2_count<- c()
loglik_beeta3_count<- c()
beeta<- c(0.1, 0, 0.02)
matX<- rbind(c(100,0,0), c(0,20,0), c(0,0,500), c(50,50,0),
    c(200,200,200))
colnames(matX)<- c("x1", "x2", "x3")

for (i in 1:1000){
  A<- rpois(1, beeta[1]*100)
  B<- rpois(1, beeta[2]*20)
  C<- rpois(1, beeta[3]*500)
  N1<- rpois(1, beeta[1]*50)+rpois(1, beeta[2]*50)
  N2<- rpois(1, beeta[1]*200)+rpois(1, beeta[2]*200)+
    rpois(1, beeta[3]*200)
  Y=c(A,B,C, N1,N2)

  lineaarne_mudel<- lm(Y ~ matX-1)
  hinnangud_lm[[i]]<- lineaarne_mudel$coefficients
  lm_mse[i]<- keskmine_ruutviga(
    lineaarne_mudel$coefficients[2], beeta[2])
  lm_mae[i]<- keskmine_absoluutviga(
    lineaarne_mudel$coefficients[2], beeta[2])
  nnls_mudel<- nnls(matX, Y)
  hinnangud_nnls[[i]]<- nnls_mudel$x
  nnls_mse[i]<- keskmine_ruutviga(nnls_mudel$x[2], beeta[2])
  nnls_mae[i]<- keskmine_absoluutviga(nnls_mudel$x[2],
    beeta[2])
  nnls_mudel_pen<- penalized(Y, penalized=matX,
    unpenalized=~0, model="linear", positive=TRUE, beeta1=0,
    beeta2=0, trace=FALSE)
  vek<- rep(0,3)
  vek[as.numeric(substr(names(coef(nnls_mudel_pen)),2,2))]=
    coef(nnls_mudel_pen)
  tulemused_penalized[i, ]=vek
```

```
nnls_mudel_pen1<- penalized(Y, penalized=matX[,-1],
  unpenalized=~0, model="linear", positive=TRUE, beeta1=0,
  beeta2=0, trace=FALSE)
nnls_mudel_pen2<- penalized(Y, penalized=matX[,-2],
  unpenalized=~0, model="linear", positive=TRUE, beeta1=0,
  beeta2=0, trace=FALSE)
nnls_mudel_pen3<- penalized(Y, penalized=matX[,-3],
  unpenalized=~0, model="linear", positive=TRUE, beeta1=0,
  beeta2=0, trace=FALSE)

nnls_beeta1[i]<- 2*(loglik(nnls_mudel_pen)-
  loglik(nnls_mudel_pen1))
nnls_beeta2[i]<- 2*(loglik(nnls_mudel_pen)-
  loglik(nnls_mudel_pen2))
nnls_beeta3[i]<- 2*(loglik(nnls_mudel_pen)-
  loglik(nnls_mudel_pen3))

hinnangud_sth[[i]]<- optim(c(1,1,1), fn=logtoepara_pois,
  y=Y, control=list(fnscale=-1), method="L-BFGS-B",
  lower=10^{-8})$par
sth_mse[i]<- keskmine_ruutviga(hinnangud_sth[[i]][2],
  beeta[2])
sth_mae[i]<- keskmine_absoluutviga(hinnangud_sth[[i]][2],
  beeta[2])

a1_beeta1<- optim(c(1,1), fn=logtoepara_poisbeeta1, y=Y,
  control=list(fnscale=-1), method="L-BFGS-B",
  lower=10^{-8})
a2_beeta2<- optim(c(1,1), fn=logtoepara_poisbeeta2, y=Y,
  control=list(fnscale=-1), method="L-BFGS-B",
  lower=10^{-8})
a3_beeta3<- optim(c(1,1), fn=logtoepara_poisbeeta3, y=Y,
  control=list(fnscale=-1), method="L-BFGS-B",
  lower=10^{-8})
a4<- optim(c(1e-8,a1_beeta1$par), fn=logtoepara_pois, y=Y,
  control=list(fnscale=-1), method="L-BFGS-B",
  lower=10^{-8})
a5<- optim(c(1,a1_beeta1$par), fn=logtoepara_pois, y=Y,
  control=list(fnscale=-1), method="L-BFGS-B",
  lower=10^{-8})
a6<- optim(c(a2_beeta2$par[1], 1e-8, a2_beeta2$par[2]),
  fn=logtoepara_pois, y=Y, control=list(fnscale=-1),
  method= "L-BFGS-B", lower=10^{-8})
```

```
  a7<- optim(c(a2_beeta2$par[1],1,a2_beeta2$par[2]),
    fn=logtoepara_pois, y=Y, control=list(fnscale=-1),
    method="L-BFGS-B", lower=10^{-8})
  a8<- optim(c(a3_beeta3$par,1e-8), fn=logtoepara_pois, y=Y,
    control=list(fnscale=-1), method="L-BFGS-B",
    lower=10^{-8})
  a9<- optim(c(a3_beeta3$par,1), fn=logtoepara_pois, y=Y,
    control=list(fnscale=-1), method="L-BFGS-B",
    lower=10^{-8})

  if (a5$value>a4$value){a4=a5}
  loglik1_vordlus[[i]]<- 2*(a4$value-a1_beeta1$value)
  if (loglik1_vordlus[[i]]>(1.96)**2){
    loglik_beeta1_count[i]<- loglik1_vordlus[[i]]}

  if (a7$value>a6$value){a6=a7}
  loglik2_vordlus[[i]]<-2*(a6$value-a2_beeta2$value)
  if (loglik2_vordlus[[i]]>(1.96)**2){
    loglik_beeta2_count[i]<- loglik2_vordlus[[i]]}

  if (a9$value>a8$value){a8=a9}
  loglik3_vordlus[[i]]<-2*(a8$value-a3_beeta3$value)
  if (loglik1_vordlus[[i]]>(1.96)**2){
    loglik_beeta3_count[i]<- loglik2_vordlus[[i]]}

  hinnangud_vr[[i]]<- vahimruutude_meetod(matX, Y)
  vr_mse[i]<- keskmine_ruutviga(hinnangud_vr[[i]][2],
    beeta[2])
  vr_mae[i]<- keskmine_absoluutviga(hinnangud_vr[[i]][2],
    beeta[2])
}

#Programm jätkub:
#1)tõepära teststatistiku jaotuse leidmisega (lõigus [0;3]),
#2)keskmise ruut- ja absoluutvea leidmisega (lõigus [0;0.4]),
#3)võimsuste leidmisega (lõigus [0;0.4]),
#4)nihketuse kontrolliga,
#5)keskmisele ruut- ja absoluutveale usaldusintervallide
   leidmisega,
#6)jooniste koostamisega.

#Jälgitavuse huvides jäetakse antud koodiread välja.
```

Sekveneerimisveaga andmete genereerimine ja hinnangute leidmine (veaga mitte arvestades)

```
#Kasutab sama koodi, mis sekveneerimisveata andmete
#genereerimise ja hinnangute leidmise programm, kuid väikeste
#erisustega. Oluliseim on enne for tsükli algust
#defineerida järgmised muutujad.

viga<- 0.04
beeta <- c(0.1, 0, 0.02)
matX<- rbind(c(100,0,0), c(0,20,0), c(0,0,500), c(50,50,0),
    c(200,200,200))
colnames(matX)<- c("x1", "x2", "x3")
uhe_kmeeri_kaugusel<- matrix(c(105, 12, 89, 8, 22, 45, 0, 4,
   557, 55, 67, 78, 200, 250, 234), ncol=3, byrow=TRUE)
korrutis<- uhe_kmeeri_kaugusel%*%(lambda*viga*1/3)

#Tsükli sees andmete genereerimine toimub järgnevalt
A<- rpois(1, lambda[1]*100)
B<- rpois(1, lambda[2]*20)
C<- rpois(1, lambda[3]*500)
N1<- rpois(1, lambda[1]*50)+rpois(1, lambda[2]*50)
N2<- rpois(1, lambda[1]*200)+rpois(1, lambda[2]*200)+
    rpois(1, lambda[3]*200)
Y_A<- A+rpois(1,korrutis[1])
Y_B<- B+rpois(1,korrutis[2])
Y_C<- C+rpois(1,korrutis[3])
Y_N1<- N1+rpois(1,korrutis[4])
Y_N2<- N2+rpois(1,korrutis[5])
Y=c(Y_A,Y_B, Y_C, Y_N1, Y_N2)
```

Sekveneerimisveaga andmete genereerimine ja hinnangute leidmine (veaga arvestades)

```
#Kasutab sama koodi, mis sekveneerimisveata andmete
#genereerimise ja hinnangute leidmise programm, kuid väikeste
#erisustega. Oluliseim on enne for tsükli algust defineerida
#järgmised muutujad (siinkohal eriti oluline matX
```

```
#väärtustamine).

viga<- 0.04
beeta <- c(0.1, 0, 0.02)
matX<- rbind(c(100,0,0), c(0,20,0), c(0,0,500), c(50,50,0),
    c(200,200,200))
matX<- matX+uhe_kmeeri_kaugusel*(viga*1/3)
colnames(matX)<- c("x1", "x2", "x3")
uhe_kmeeri_kaugusel<- matrix(c(105, 12, 89, 8, 22, 45, 0, 4,
   557, 55, 67, 78, 200, 250, 234), ncol=3, byrow=TRUE)
korrutis<- uhe_kmeeri_kaugusel%*%(lambda*viga*1/3)

#Tsükli sees andmete genereerimine toimub järgnevalt
A<- rpois(1, lambda[1]*100)
B<- rpois(1, lambda[2]*20)
C<- rpois(1, lambda[3]*500)
N1<- rpois(1, lambda[1]*50)+rpois(1, lambda[2]*50)
N2<- rpois(1, lambda[1]*200)+rpois(1, lambda[2]*200)+
    rpois(1, lambda[3]*200)
Y_A<- A+rpois(1,korrutis[1])
Y_B<- B+rpois(1,korrutis[2])
Y_C<- C+rpois(1,korrutis[3])
Y_N1<- N1+rpois(1,korrutis[4])
Y_N2<- N2+rpois(1,korrutis[5])
Y=c(Y_A,Y_B, Y_C, Y_N1, Y_N2)
```

Tegelikud sekveneerimisandmed

```
#Lugeda sisse andmestikud 1-3
#Eemaldame andmed1 tunnuste V1 ja V2 lõpust "_32.list"
andmed1$V1<- str_sub(andmed1$V1, 1, str_length(andmed1$V1)-8)
andmed1$V2<- str_sub(andmed1$V2, 1, str_length(andmed1$V2)-8)

unik_V2<- length(unique(andmed1$V2))
unik_V1<- length(unique(andmed1$V1))

#Koostame mudelimaatriksi X
X<- matrix(rep( 0, len=unik_V2*unik_V1), nrow=unik_V2,
    ncol=unik_V1)
rownames(X)<- unique(andmed1$V2)
colnames(X)<- unique(andmed1$V1)
a<- 1
```

```
for(i in rownames(X)){
  b<- 1
  for(j in colnames(X)){
    if (andmed1[(andmed1[,2]==i)&(andmed1[,1]==j),][1,3]>0){
        X[a,b]<-andmed1[(andmed1[,2]==i)&
            (andmed1[,1]==j),][1,3]}
    b<- b+1}
  a<- a+1}

#Koostame sageduste vektori Y
Y<- matrix(rep(0, 147), nrow=147, ncol=1)
rownames(Y)<- rownames(X)
jada <- seq(1, 500, 1)
a<- 1
for(i in rownames(Y)){
    if (is.na(andmed3[(andmed3[,1]==i),][1,3])!= TRUE){
      Y[a,1]<-sum(andmed3[andmed3[,1]==i,][5:504]*jada)}
  a<- a+1}

#Koostame maatriksi U[i,j], mis näitab, kui mitu i-nda rea
#bakteri või sõlme omast k-meeri on ühe nukleotiidi
#kaugusel veerus j asuvale bakterile omastest k-meeridest.
U<- matrix(rep(0, 147*74), nrow=147, ncol=74)
rownames(U)<- rownames(X)
colnames(U)<- colnames(X)
a<- 1
for(i in rownames(U)){
  b<- 1
  for(j in colnames(U)){
    if (andmed1[(andmed1[,2]==i)&(andmed1[,1]==j),][1,5]>0){
        U[a,b]<-
            andmed1[(andmed1[,2]==i)&(andmed1[,1]==j),][1,5]}
    b<- b+1}
  a<- a+1}

#Maatriksi X, programmiaknas 1
mudelimat<- X

#Maatriks X*, programmiaknas 2
X_tarn<- 0.04/3*U+X
mudelimat<- X_tarn

#Vähimruutude meetodi nullist erinevad hinnangud
```

```
mudel_lm <- lm(Y~mudelimat-1)
hinnangud_lm<-
    mudel_lm$coefficients[summary(mudel_lm)$coef[,4]<=0.05]

#Vähimruutude meetodi statistiliselt olulised hinnangud
hinnangud_lm_statolulised<-
    mudel_lm$coefficients[summary(mudel_lm)$coef[,4]<=
    0.05/74]

#Mittenegatiivsete vähimruutude meetodi nullist erinevad
#hinnangud
nimed_nnls<- colnames(mudelimat)[(nnls(mudelimat,Y)$x)>0]
hinnangud_nnls<- nnls(mudelimat,Y)$x[(nnls(mudelimat,Y)$x)>0]
cbind(nimed_nnls, hinnangud_nnls)

#Mittenegatiivsete vähimruutude meetodi statistiliselt
#olulised hinnangud
stat_olulised_nnls<- rep(FALSE, 74)
loglik_vordlus_nnls<- list()
for (b in 1:74){
  nnls_mudel_pen<- penalized(Y, penalized=mudelimat,
    unpenalized=~0, model="linear", positive=TRUE,
    lambda1=0, lambda2=0, trace = FALSE)
  nnls_mudel_pen_lambda<- penalized(Y,
    penalized=mudelimat[,-b], unpenalized=~0,
    model="linear", positive=TRUE,lambda1=0,
    lambda2=0, trace= FALSE)
  loglik_vordlus_nnls[[i]]<- 2*(loglik(nnls_mudel_pen)-
    loglik(nnls_mudel_pen_lambda))
  if (loglik_vordlus_nnls[[i]]>(1.96)**2){
        stat_olulised_nnls[b]<- TRUE}
}
nimed_nnls_statolulised<-
    colnames(mudelimat)[stat_olulised_nnls]
hinnangud_nnls_statolulised<-
    nnls(mudelimat,Y)$x[stat_olulised_nnls]
cbind(nimed_nnls_statolulised, hinnangud_nnls_statolulised)

#Suurima tõepära meetodi nullist erinevad hinnangud
logtoepara_pois_parisandmed<- function(par, y){
  tulemus<- (y*log(mudelimat%*%par)-mudelimat%*%par-
    lgamma(y+1))
  tulemus1<- tulemus[-64]
```

```
  tagasta<- sum(tulemus1)
  return(tagasta)}
alg<- rep(0.1, 74)
STH<- optim(alg, fn=logtoepara_pois_parisandmed, y=Y,
    control=list(fnscale=-1), method="L-BFGS-B",
    lower=10^{-13})$par)
nimed_sth<- colnames(mudelimat)[STH>0.0000001]
hinnangud_sth<- STH[STH>0.0000001]
cbind(nimed_sth, hinnangud_sth)

#Suurima tõepära meetodi statistiliselt olulised hinnangud
stat_oluline<- rep(FALSE, 74)
loglik_vordlus<- rep(NA,74)
logtoepara_pois_parisandmed_lambda<-function(alg, y, indeks){
  par<- insert(alg, indeks, 1e-8)
  tulemus<- (y*log(mudelimat%*%par)-mudelimat%*%par-
    lgamma(y+1))
  tulemus1<- d[-64]
  tagasta<- sum(tulemus1)
  return(tagasta)}
for (a in 1:74){
  alg<- rep(0.1, 73)
  a1=optim(alg, fn=logtoepara_pois_parisandmed_lambda, y=Y,
    indeks=a, control=list(fnscale=-1), method="L-BFGS-B",
    lower=10^{-13})
  alg=insert(a1$par, a, 1e-8)
  a2=optim(alg, fn=logtoepara_pois_parisandmed, y=Y,
    control=list(fnscale=-1), method="L-BFGS-B",
    lower=10^{-13})
  loglik_vordlus[a] <-2*(a2$value-a1$value)
  loglik_vordlus[a]<- loglik_vordlus[a]
  if (loglik_vordlus[a]>(1.96)**2){
    stat_oluline[a]<- TRUE}
}
nimed_sth_statoluline<- colnames(mudelimat)[stat_oluline]
hinnangud_sth_statoluline<- STH[stat_oluline]
cbind(nimed_sth_statoluline, hinnangud_sth_statoluline)

#Programm jätkub hinnangute fülogeneesipuu struktuuri
#leidmisega. Jälgitavuse huvides jäetakse antud
#koodiread välja.
```

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Anna Laaneväli,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Statistilised mudelid bakterite segude määramiseks", mille juhendaja on Märt Möls,

   reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Anna Laaneväli

30.01.2020