Mikrobiomiaineistojen analysointi CoDaCoRe-menetelmän avulla

Laura Perasto

Pro gradu -tutkielma
Toukokuu 2023

Ohjaajat:
Prof. Kari Auranen
FM Juho Pelto

MATEMATIIKAN JA TILASTOTIETEEN LAITOS

TURUN YLIOPISTO
Matematiikan ja tilastotieteen laitos

Laura Perasto: Mikrobiomiaineistojen analysointi CoDaCoRe-menetelmän avulla
Pro gradu -tutkielma, 54 s., 14 liites.
Tilastotiede
Toukokuu 2023

Mikrobiomiaineistot muodostuvat mikrobiominäytteissä havaittujen eri mikrobien lukumääristä. Aineistojen tyyppillisiä piirteitä ovat suuri dimensio, nollasolut sekä mikrobien lukumäärien jakaumien vinous. Lisäksi aineistojen voidaan ajatella olevan kompositionaalisia. Tämän tutkielman tavoitteena on esitellä vastikään julkaistu uusi menetelmä mikrobiomiaineistojen analysointiin, CoDaCoRe. CoDaCoRe-menetelmän tarkoituksena on tunnistaa mahdollisimman vähälukuiset (eli "harvat") kaksi mikrobien osajoukkoa, jotka ennustavat valittua vastemuuttujaa parhaiten. Optimaaliset osajoukot määritetään aineiston perusteella joko ns. balanssien tai yhdelmien avulla. Balanssit ja yhdelmät määritellään kahteen osajoukkoon kuuluvien mikrobien havaittujen lukumäärien suhteina.

Tutkielman alussa kuvataan, kuinka mikrobiomiaineistot muodostuvat sekä niiden erityispiirteitä ja haasteita. Lisäksi esitellään mikrobikoostumuksen monimuotoisuutta kuvaavia indeksejä sekä sitä, miten biomarkkereita eli biologista ilmiötä ennustavia mikrobeja voidaan tunnistaa. CoDaCoRe-menetelmän vaiheet esitellään ensin teoreettisesti. Tämän jälkeen menetelmää sovelletaan mikrobiomiaineistoon, joka on kerätty FinnBrain-syntymäkohorttitutkimuksessa 2.5 kuukauden ikäisiltä vauvoilta. CoDaCoRe-menetelmän avulla voidaan tutkia sekä binääristä että jatkuvaa vastemuuttujaa. Binäärisenä eli kaksiarvoisena vasteena käytetään synnytystapaa (alatiesynnytys vs. sektio). Jatkuvana vastemuuttujana käytetään 30 kuukautisten lasten silmänliikkeemittauksista johdettua muuttujaa, joka mittaa, kuinka usein lapsi käänsi katseensa pois pelokkaista kasvokuvista.

Tutkielman tulosten perusteella alatiesynnyttäneet ja sektiolla synnyttäneet erosivat toisistaan selkeästi löydettyjen osajoukkojen suhteen. Myös silmänliikemmuuttujan tapauksessa tunnistettiin osajoukkoja, jotka pystyivät ennustamaan vastemuuttujaa. Sekä binäärisen että jatkuvan vastemuuttujan kohdalla tutkittiin, miten hyvin tunnistetut osajoukot ennustivat uusia havaintoja. Yhdelmien ennustuskyky oli vain hieman parempi verrattuna balansseihin. Sensitiivisyysanalyysina tarkasteltiin optimointialgoritmin tarvitseman siemenluvun vaikutusta löydettyihin osajoukkoihin. Valitulla siemenluvulla oli vaikusta löydettyjen osajoukkojen sisältämiin mikrobeihin.

Asiasanat: mikrobiomiaineisto, CoDaCoRe-menetelmä, balanssit, monimuotoisuus, DA-analyysi, DR-analyysi, FinnBrain-kohortti.

Sisällys

1 Johdanto

Mikrobeja (*microorganisms*) eli pieneliöitä ovat muun muassa bakteerit ja arkit. Mikrobeja on joka puolella maailmaa erilaisissa elinympäristöissä. Osa selviää erittäin kuumissa lämpötiloissa, osa erittäin suolaisissa ympäristöissä ja osa pH-arvon ääripäissä. Mikrobeita löytyy runsaasti myös ihmisistä esimerkiksi iholta, genitaalialueilta sekä ruuansulatuskanavasta. Ihmisessä elävät mikrobit muodostavat yhdessä ihmisen mikrobiston (*microbiota*). Ihmisen suurin mikrobisto sijaitsee suolistossa. Mikrobien analysointi on erittäin tärkeää, koska niiden tiedetään olevan yhteydessä erilaisiin sairauksiin, esimerkiksi suolistosairauksiin. Myös mikrobien yhteyttä allergioihin, syöpiin ja neurologisiin sairauksiin on tutkittu. [1, 2]

Käsitteitä mikrobiomi (*microbiome*) ja mikrobisto käytetään usein sekaisin, vaikka molemmilla on oma merkityksensä. Mikrobisto viittaa varsinaisiin mikrobeihin. Mikrobiomi taas tarkoittaa mikrobeita ja niiden geneettistä materiaalia. Pelkästä geneettisestä materiaalista käytetään sanaa metagenomi (*metagenome*). [2]

Mikrobiomiaineistot muodostuvat mikrobiominäytteissä havaittujen eri mikrobien lukumääristä, eli mikrobiomiaineistot ovat lähtökohtaisesti lukumääräaineistoja. Aineistojen tyyppillisiä piirteitä ovat suuri dimensio, nollasolut, havaittujen lajirunsauksien jakaumien vinous ja aineistojen kompositionaalisuus. Mikrobiomien analysointiin tarkoitetut työkalut kehittyvät nopeasti. Tämän takia parhaan menetelmän valitseminen tilastollisiin analyyseihin on haastavaa.

Tämän tutkielman tavoitteena on esitellä vastikään julkaistu uusi menetelmä mikrobiomiaineistojen analysointiin, CoDaCoRe [3]. Menetelmä on kehitetty vastaamaan tehokkaasti mikrobiomiaineistojen haasteisiin. CoDaCoRe-menetelmä viittaa CoDa-menetelmiin eli menetelmiin, jotka ottavat huomioon aineiston kompositionaalisen luonteen (*Compositional Data Analysis*). CoDaCoRe-menetelmä tarjoaa helposti tulkittavia biomarkkereita, koska sen tavoitteena on löytää suuresta määrästä lajeja sellaiset kaksi harvaa (*sparse*) osajoukkoa, jotka ennustavat valittua vastemuuttujaa parhaiten. Harvat osajoukot tarkoittavat, että vasteen ennustamiseen ei käytetä aineiston kaikkia dimensioita eli lajeja. Menetelmä siis pienentää mikrobiomiaineistojen dimensioita ja on kuitenkin laskennallisesti erittäin tehokas. Menetelmässä käytetty vastemuuttuja voi olla binäärinen tai jatkuva. Binäärisenä eli kaksiarvoisena vasteena voi olla esimerkiksi sairaus vs. ei sairautta, tai, kuten tässä tutkimuksessa, synnytystapa (alatiesynnytys vs. sektio). Mielenkiinnon kohteena voivat olla myös jatkuvana käsiteltävät muuttujat, esimerkiksi erilaiset tulehduksista kertovat merkkiaineet.

Tutkielman luvussa 2 esitellään mikrobiomiaineistojen tyyypillinen rakenne ja aineistojen erityispiirteitä. Luvussa 2 käsitellään myös mikrobiomiaineistojen monimuotoisuuden tutkimista sekä biomarkkereiden etsimistä. CoDaCoRe-menetelmän eri vaiheita selvitetään kolmannessa luvussa. Luvussa 4 sovelletaan CoDaCoRe-menetelmää oikeaan aineistoon ja tutkimuskysymykseen. Tutkielman luvussa 5 on yhteenveto sekä pohdintaa CoDaCoRe-menetelmän hyvistä ja huonoista puolista. Koodit esitetään tutkielman liitteessä C.

2 Mikrobiomiaineistojen erityispiirteitä

Mikrobiomeita voidaan tutkia erilaisista biologisista näytteistä, esimerkiksi ulosteesta tai kudosnäytteestä. Kun näyte on kerätty, siitä erotetaan geneettinen materiaali ja suoritetaan sekvensointi. Näytteen sekvensoinnin tarkoituksena on selvittää geneettisen materiaalin molekyylirakenne ja digitoida se. Sekvensointitapoja on erilaisia, mutta pääsääntöisesti käytetään joko amplikonisekvensointia (*amplicon sequencing*) tai shotgun-sekvensointia. Amplikonisekvensoinnissa tarkastellaan vain yhtä tiettyä geeniä. Usein geeni, jota sekvensoidaan, on 16S rRNA-geeni. Shotgun-sekvensoinnin idea on hajottaa geenimateriaali satunnaisesti DNA-fragmenteiksi. Fragmentit sekvensoidaan yksitellen. Tämän jälkeen tietokoneohjelma etsii päällekkäisyyksiä DNA-sekvensseistä ja kokoaa niiden avulla fragmentit oikeaan järjestykseen. [2]

Hyvin yleinen sekvensointitapa on 16S rRNA-sekvensointi. Menetelmä on suosittu, koska kyseinen geeni löytyy kaikista bakteereista. Menetelmä on myös halvempi ja yksinkertaisempi kuin shotgun-sekvensointi. Lisäksi 16S rRNA-sekvensointiin on tarjolla laajoja tietokantoja. Esimerkiksi SILVA-tietokannassa on laatukontrolloituja geenisekvenssitietoja, joiden avulla havaittu sekvenssi voidaan yhdistää oikeaan bakteeriin [4]. Suurimpana erona kahden sekvensointimenetelmän välillä on niiden mahdollistamat tulkinnat. Shotgun-sekvensointi tarjoaa laajemman käsityksen mikrobiston koostumuksesta sekä mikrobiomien toiminnasta. Shotgun-sekvensointi voi siis antaa vastauksen siihen, mitä lajeja näytteessä on, sekä siihen, mikä on lajien toiminta eli funktionaalisuus. Sen sijaan 16S rRNA-sekvensointia voidaan käyttää vain näytteen taksonomiseen profilointiin. Liitteen A taulukkoon A1 on koottu sekä 16S rRNA- että shotgun-sekvensoinnin keskeisimmät hyvät ja huonot puolet. [2]

Mikrobiomianalyyseissa tulee valita, millä taksonomisen asteikon tasolla eli taksonilla mikrobeja tutkitaan. Taksonominen asteikko on biologinen järjestelmä, jonka perusteella mikrobeja voidaan luokitella hierarkkisesti. Taksonomian ylin taso on domeeni ja alin laji (kuva 1). Esimerkiksi nykyihmisen domeeni on aitotumaiset, suku Homo ja laji sapiens. [2]

Mikrobiomianeistot ovat lukumääräaineistoja, joiden rakenne on niin sanottu näyte-ominaisuus-ristiintaulukko (*Sample-by-Feature Contingency Table*). Ominaisuudella tarkoitetaan taksonomisen asteikon valitun tason mukaista luokitusta (kuva 1). Tämän tutkielman teoriaosuudessa käytetään selvyyden vuoksi sanaa laji, vaikka ei tarkoitettaisikaan taksonomisen asteikon tasoa laji. Lukumääräaineisto kertoo, kuinka monta kertaa jotain lajia on havaittu näytteessä eli mikä sen havaittu lajirunsaus (*abundance*) on. Tavallisesti lajit ovat sarakkeilla ja näytteet riveillä. Aineisto on siis $n \times p$ -matriisi, jossa n on näytteiden lukumäärä ja p on lajien lukumäärä. Datamatriisin rakenne voi myös olla toisin päin riippuen siitä, millä tilastollisella menetelmällä aineistoa analysoidaan. [2]

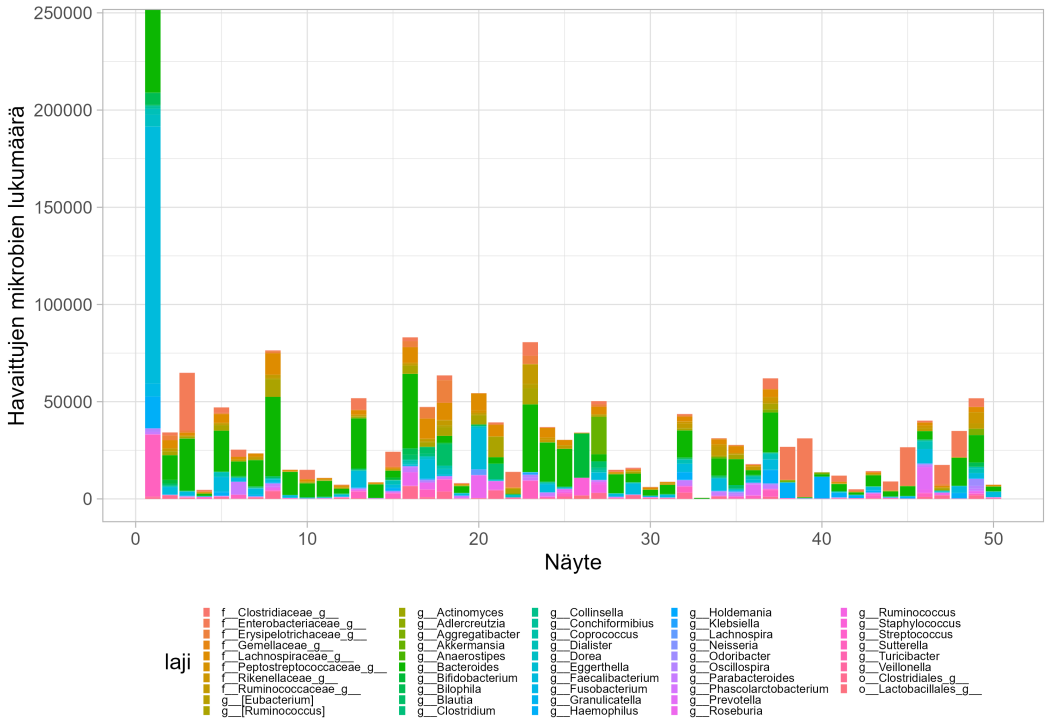
Kuva 1: *Mikrobiomianalyyseissa tulee valita taksonomisen asteikon taso, jolla mikrobeja tutkitaan.*

Tyypillisiä mikrobiomiaineiston ominaisuuksia ja haasteita ovat suuri dimensio, nollasolut sekä havaittujen lajirunsauksien vino jakauma. Lajien lukumäärä on yleensä suurempi kuin näytteiden lukumäärä eli $p > n$. Mikrobiomiaineistoissa jopa suurin osa soluista voi olla nollia. Nollasolut voivat aiheuttaa hankaluuksia esimerkiksi silloin, jos valittu tilastollinen menetelmä käyttää hyväkseen havaittujen lajirunsauksien log-muunnoksia. Nollasoluja voidaan poistaa lisäämällä jokaiseen soluun jokin pieni luku, esimerkiksi 0.5 tai 1. Lisättyä lukua kutsutaan pseudolukumääräksi (*pseudo count*). Lisätyillä luvuilla voi olla suuri painoarvo tilastollisissa analyyseissa. [2]

Mikrobiomiaineistojen rakennetta on havainnollistettu taulukossa 1, jossa on esitetty pieni osa Crohnin sairautta kuvaasta aineistosta [3]. Aineistossa on 975 näytettä ja 48 eri lajia. Lisäksi siinä on kategorinen muuttuja, joka kertoo, onko näytteen yksilöllä Crohnin sairaus (tapaus) vai ei (verrokki). Tapauksia on 662 ja verrokkeja on 313. Eri näytteissä havaitut lajien kokonaislukumäärät vaihtelevat aineistossa suuresti (kuva 2). Liitteen kuvassa A1 on koko aineistoa kuvaava pylväsdiagrammi.

Taulukko 1: *Esimerkki mikrobiomiaineiston rakenteesta.* Riveillä on esitetty kuusi ensimmäistä näytettä ja sarakkeilla neljä ensimmäistä lajia Crohnin sairautta kuvaavasta aineistosta. Yksinkertaistamisen vuoksi aineiston oikeat näytekoodit on korvattu numeroilla yhdestä eteenpäin ja lajien niemet on korvattu sanalla laji. Lajien lukumäärät kuvaavat biologisen näytteen havaittuja lajirunsauksia. Lisäksi rivisummasarakkeelle on laskettu jokaiselle näytteelle lajien kokonaislukumäärä eli näytteen kirjastokoko. [3]

	laji 1	laji 2	laji 3	laji 4	...	rivisumma
näyte 1	0	1190	2257	31711	...	380827
näyte 2	0	16	1912	22	...	34186
näyte 3	2	151	867	141	...	64875
näyte 4	0	47	369	845	...	4669
näyte 5	0	17	2829	37	...	47059
näyte 6	0	6607	278	1919	...	25336



Kuva 2: *Pylväsdiagrammi Crohnin sairautta kuvaavasta aineistosta.* Viisikymmentä ensimmäistä näytettä on x-akselilla ja y-akselilla on lajien havaitut lukumäärät. Ensimmäisessä näytteessä havaittujen lajien yhteenlaskettu lajirunsaus on suurempi kuin 250 000. Eri lajit ovat eri värein pylväissä. Tämän esimerkin yhteydessä lajit ovat todellisuudessa sukuja. Esityksen yhtenäisyyden vuoksi tässä käytetään kuitenkin termiä laji. [3]

Luvuissa 2.1 ja 2.2 esitellään mikrobiomiaineistojen vääristymiä ja tapoja, joilla niitä voidaan ottaa huomioon tilastollisia analyysejä tehdessä. Luvussa 2.3 tutustutaan hyvin yleisiin tapoihin, joilla mikrobiomiaineistoja voidaan kuvailla. Lopuksi luvussa 2.4 vertaillaan, miten biomarkkereita voidaan tunnistaa aineistoista ns. DA- ja DR-analyysien avulla.

2.1 Aineistojen harhaisuus ja kompositionaalisuus

Mikrobiomiaineistoihin voi kehittyä erilaisia vääristymiä eli harhoja. Harhoja voi syntyä esimerkiksi näytteiden keräämisen, säilyttämisen ja sekvensoinnin aikana. Kerätyn biologisen näytteen tulisi vastata mikrobien elinympäristöä eli ekosysteemiä. Biologinen näyte voidaan joissain tapauksissa kerätä useilla tavoilla mikrobien elinympäristöstä, esimerkiksi suunäytteitä voidaan kerätä vaikka syljen tai suuveden avulla. Näytteiden säilyttäminen vaikuttaa siihen, miten hyvin näytteiden mikrobikoostumus säilyy sekvensointiin saakka. Näytteet tulisi jäädyttää heti näytteen ottamisen jälkeen -80°C:seen [5]. Tämä on käytännössä hyvin haastavaa, joten tavallisesti näytteisiin lisätään säilöntäaine. Säilöntäaine estää mikrobien lisääntymisen huoneenlämmössä ja säilyttää mikrobien geneettisen materiaalin. Säilöntäaineita on erilaisia ja ne saattavat vaikuttaa eri tavoin eri mikrobeihin. Niiden vaikutusta havaittuihin lajirunsauksiin on tutkittu esimerkiksi vuonna 2015 julkaistussa artikkelissa [6].

Myös sekvensoinnin takia aineistoihin syntyy vääristymiä. Sekvensointi löytää eri lajeja (eli mikrobeja) eri tarkkuuksilla, esimerkiksi jokin laji voidaan havaita helpommin verrattuna johonkin toiseen. Tätä kutsutaan taksonimiseksi harhaksi (*taxonomic bias*). Seuraavaksi tarkastellaan taksonomista harhaa McLarenin ym. julkaiseman artikkelin mukaisesti [7]. Tehdään muutamia oletuksia, joiden ansiosta laskukaavat yksinkertaistuvat. Oletetaan, että tarkastellaan lajitasoa. Jos havaittua lajirunsautta ei voida yksiselitteisesti määrätä lajille j, se hylätään. Oletetaan, että havaitut lajirunsaudet yhdistetään oikeaan lajiin sekvensoinnissa ja taksonominen harha vaikuttaa johdonmukaisesti näytteisiin lajitasolla. Lisäksi oletetaan, että sekvensointimittaukset ovat deterministisiä eli sivuutetaan satunnainen harha.

Määritellään vielä muutamia merkintöjä, joita tarvitaan taksonomisen harhan tarkastelussa. Absoluuttinen eli todellinen lajirunsaus tarkoittaa tietyn lajin solujen lukumäärää yksikkötilavuutta kohti biologisessä näytteessä. Merkitään lajin j absoluuttista lajirunsautta näytteessä i notaatiolla a_{ij}. Olkoon T_j lajikohtainen mittaustehokkuus (*measurement efficiency*). Mittaustehokkuudella tarkoitetaan sitä osuutta, joka lajin todellisesta solujen lukumäärästä havaitaan näytteessä. Taksonominen harha seuraa siitä, että mittaustehokkuus vaihtelee eri lajien välillä. Kullekin lajille harha kuitenkin oletetaan yhtä suureksi kaikissa näytteissä. Olkoon F_i näytekohtainen tekijä (*sequencing effort*). Näytekohtainen tekijä on havaittujen runsauksien määrä todellista runsautta kohti, joka saataisiin lajille, jonka mittaustehokkuus on yksi. Havaittu lajirunsaus (*read count*) x_{ij} lajille j näytteessä i voidaan kirjoittaa todellisen lajirunsauden, mittaustehokkuuden sekä näytekohtaisen tekijän avulla seuraavasti:

$$x_{ij} = a_{ij} \cdot T_j \cdot F_i. \tag{1}$$

Näytteen i keskimääräinen mittaustehokkuus $\tilde{T}_i$ on

$$\tilde{T}_i = \frac{\sum_{j=1}^{p} (a_{ij} \cdot T_j)}{\sum_{j=1}^{p} a_{ij}}, \tag{2}$$

jossa $\sum_{j=1}^{p} a_{ij}$ on näytteen i kaikkien lajien j todellinen yhteenlaskettu lajirunsaus.

Lajin j todellinen osuus (*proportion*) näytteessä i saadaan jakamalla sen todellinen lajirunsaus kaikkien lajien yhteenlasketuilla runsauksilla:

$$P_{ij} = \frac{a_{ij}}{\sum_{j=1}^{p} a_{ij}}. \tag{3}$$

Vastaavasti lajin j havaittu osuus näytteessä i saadaan jakamalla sen havaittu lajirunsaus kaikkien lajien yhteenlasketulla havaituilla runsauksilla:

$$\tilde{P}_{ij} = \frac{x_{ij}}{\sum_{j=1}^{p} x_{ij}}, \tag{4}$$

jossa $\sum_{j=1}^{p} x_{ij}$ on lajien yhteenlaskettu havaittu lajirunsaus näytteessä i. Kaavoista (1)–(4) seuraa havaittujen ja todellisten osuuksien välinen yhteys:

$$\tilde{P}_{ij} = P_{ij} \cdot \frac{T_j}{\tilde{T}_i}. \tag{5}$$

Taksonominen harha luo siis taittovirheen FE (*a fold-error*) lajin j havaittuun osuuteen. Koska keskimääräinen mittaustehokkuus vaihtelee näytteiden välillä, myös lajin j osuuteen vaikuttava taittovirhe vaihtelee näytteiden välillä. Mikäli kaavan (5) kerroin $T_j/\tilde{T}_i$ on pienempi kuin 1, taittovirhe on negatiivinen eli lajin j runsaus on aliarvioitu. Vastaavasti jos kerroin on suurempi kuin 1, taittovirhe on positiivinen eli lajin j runsaus on yliarvioitu.

Merkitään kahden lajin l ja k välistä suhdetta (*ratio*) näytteessä i niiden todellisten lajirunsauksien suhteena:

$$R_{il/ik} = \frac{a_{il}}{a_{ik}}.$$

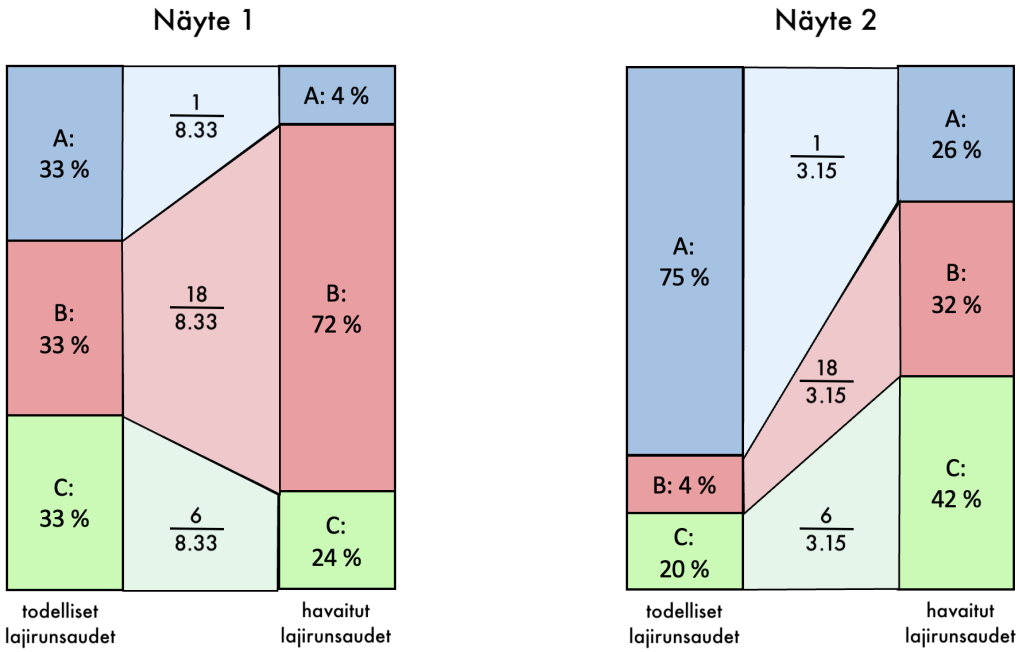
Vastaavasti kahden lajin l ja k havaittu suhde näytteessä i on

$$\tilde{R}_{il/ik} = \frac{x_{il}}{x_{ik}}. \tag{6}$$

Kaavoista (1) ja (6) seuraa:

$$\tilde{R}_{il/ik} = R_{il/ik} \cdot \frac{T_l}{T_k}, \tag{7}$$

jossa T_l on lajin l mittaustehokkuus ja T_k on lajin k mittaustehokkuus. Kaavan (7) mukaan siis havaittujen runsauksien suhde on oikeiden runsauksien suhde kerrottuna näytteestä i riippumattomalla kertoimella T_l/T_k. Kerroin kuitenkin riippuu lajeista l ja k. Lajien välinen suhteellinen runsaus on siis kaikissa näytteissä saman verran yli- tai aliarvioitu oikeasta runsaudesta. Taittovirhettä on havainnollistettu kuvassa 3.



Kuva 3: *Esimerkki taksonomisesta harhasta mikrobiomiaineistoissa kaavan (4) perusteella.* Näytteen 1 keskimääräinen mittaustehokkuus $\tilde{T}_1$ on 8.33 ja näytteessä 2 mittaustehokkuus $\tilde{T}_2$ on 3.15. Yksittäisen lajin mittaustehokkuus on sama kummassakin näytteessä, esimerkiksi lajin C mittaustehokkuus on 6. Lajin C lajirunsaus on aliarvioitu näytteessä 1 ($FE < 1$). Näytteessä 2 lajin C havaittu lajirunsaus taas on yliarvioitu ($FE > 1$). Tämä johtuu siitä, että kun näytteen 1 todelliset lajirunsaudet ovat tasapainossa, eli kaikkia on 33 prosenttia, keskimääräinen mittaustehokkuus on korkeampi verrattuna näytteeseen 2, joka on epätasapainoinen. Lajien välinen suhteellinen runsaus on kaikissa näytteissä saman verran yli- tai aliarvioitu oikeasta runsaudesta. Esimerkiksi lajin C suhde lajiin A on yliarvoitu kuusinkertaisesti molemmissa näytteissä, vaikka näytteiden koostumukset ovat erilaiset. [7]

Mikrobiomiaineistoissa kunkin näytteen kirjastokoko (*library size*) ei vastaa suoranaisesti näytteessä olevien lajien runsauksia. Lisäksi kirjastokoko vaihtelee paljon eri näytteiden välillä. Näiden syiden vuoksi mikrobiomiaineistojen voidaan ajatella olevan kompositionaalisia. Määritellään kompositionaalisuus ensin yleisesti. Kompositionaalinen aineisto koostuu vektoreista $\mathbf{x}_i$, joiden komponentit ovat positiivisia lukuja, $i = 1, ..., n$. Lisäksi komponenttien summa vektorissa $\mathbf{x}_i$ on jokin vakio o_i jokaisella i eli $\sum_{j=1}^{p} x_{ij} = o_i$. Kompositiossa yksi komponentti ei ole merkityksellinen yksinään vaan kaikkien komponenttien suhteet toisiinsa. Kompositio kuvaa siis kvantitatiivisesti jonkin kokonaisuuden osia. [8]

Taulukossa 2 on esimerkki kompositionaalisesta aineistosta, jossa on kuvattuna kolmen yksilön vuorokausirytmi. Jokaisen yksilön vuorokaudessa on yhteensä 24 tuntia, mutta heidän vuorokautensa koostuvat eri osuuksista työtä, harrastuksia, vapaa-aikaa ja nukkumista. Vuorokausiesimerkissä jokaisen vektorin $\mathbf{x}_i$ komponenttien summa on sama (24h).

Taulukko 2: *Esimerkki kompositionaalisesta aineistosta, jossa kuvataan kolmen henkilön vuorokausirytmiä.*

henkilö	työt	harrastukset	vapaa-aika	nukkuminen	rivisumma
1	6h	3h	9h	6h	24h
2	10h	1h	8h	5h	24h
3	0h	10h	7h	7h	24h

Mikrobomiaineistot voidaan ajatella kompositionaalisina sekvensoinnin tuottaman mielivaltaisen kirjastokoon vuoksi. Mikrobiomiaineistojen kompositionaalinen tarkastelu on perusteltua, koska kaavan (7) mukaan lajien välinen suhteellinen runsaus on kaikissa näytteissä saman verran yli- tai aliarvioitu oikeasta runsaudesta (kuva 3). Taulukossa 3 on esimerkki tilanteesta, jossa kerroin $T_l/T_k = 1$ eli lajien suhteelliset osuudet lukumääräaineistossa vastaavat täydellisesti kerätyn näytteen suhteellisia osuuksia. Mikrobiomien koostumus ei siis ole ilmiönä kompositionaalinen, kuten taulukon 2 esimerkissä, vaan kompositionaalisuus johtuu sekvensoinnin aiheuttamasta kirjastokoon vaihtelusta. [9]

Taulukko 3: *Mikrobomiaineistot voidaan ajatella kompositionaalisina sekvensoinnin tuottaman mielivaltaisen kirjastokoon vuoksi.* Havaitut lajirunsaudet eivät vastaa todellisia lajirunsauksia. Näytteessä 1 todellisten lajirunsauksien yhteenlaskettu lukumäärä on 900, mutta kirjastokoko on 56. Tämän taulukon esimerkissä kaavan (7) kerroin $T_l/T_k = 1$, jolloin lukumääräaineiston lajien suhteelliset osuudet vastaavat täydellisesti kerätyn näytteen suhteellisia osuuksia. Kerätyssä näytteessä 1 lajin 1 todellisen lajirunsauden osuus on $500/900 \approx 0.56$. Vastaavasti havaitun lajirunsauden mukaaan näytteessä 1 lajin 1 osuus on $56/100 = 0.56$.

Lajien *todelliset* lajirunsaudet kerätyissä näytteissä

	laji 1	laji 2	yhteensä
näyte 1	500	400	900
näyte 2	450	650	1100

Lajien *havaitut* lajirunsaudet kerätyissä näytteissä

	laji 1	laji 2	kirjastokoko
näyte 1	56	44	100
näyte 2	82	118	200

Sanotaan, että kaksi vektoria ovat kompositionaalisesti ekvivalentit, jos ne ovat suhteellisia toisiinsa nähden. Sekä lajirunsaudet että niiden muunnokset suhteellisiksi lajirunsauksiksi kuuluvat samaan ekvivalenssiluokkaan ja sisältävät saman

suhteellisen tiedon. Kompositiota voidaan kertoa positiivisella vakiolla ja lukujen merkityksen tulee pysyä samana. [8, 10]

2.2 Normalisointi ja log-muunnokset

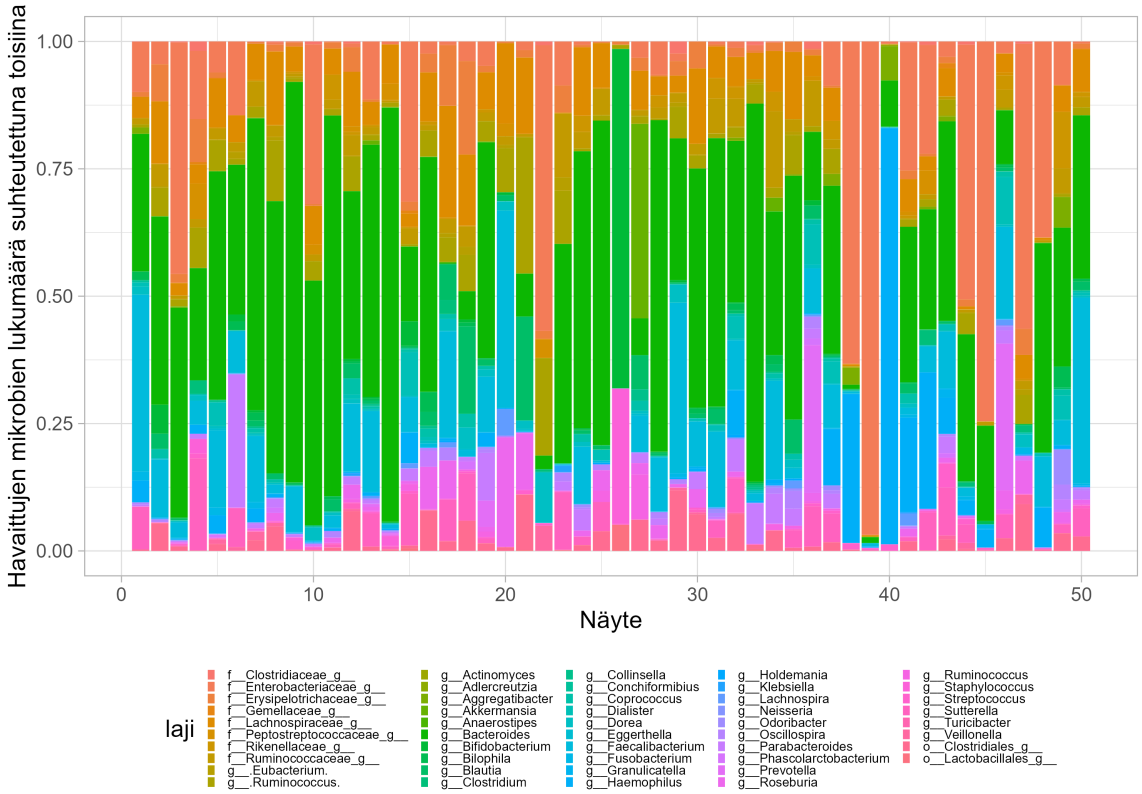
Normalisoinnin tarkoituksena on mahdollistaa mielekäs vertailu eri aineistojen ja näytteiden välillä, vaikka näytteiden kirjastokoot eroavat. Normalisointitapoja on monia, esimerkiksi skaalaaminen ja harvinaistaminen. [11, 12, 13]

Skaalaamisessa (*scaling*) lajirunsaus jaetaan skaalaus- eli normalisointikertoimella. Yksinkertaisin skaalaustapa on muuntaa lukumääräaineisto suhteellista lajirunsautta kuvaavaan muotoon kaavan (4) mukaisesti. Muunnoksen jälkeen jokaisen näytteen rivisumma on yksi. Taulukossa 4 on esitetty sama Crohnin sairautta kuvaavan aineiston alkuosa kuin taulukossa 1. Lajirunsaudet on nyt kuitenkin muutettu suhteellisiksi lajirunsauksiksi kaavan (4) mukaisesti. [12]

Taulukko 4: *Esimerkki Crohnin sairautta kuvaavan mikrobiomiaineiston rakenteesta, jossa havaitut lajirunsaudet on muunnettu suhteellisiksi lajirunsauksiksi niin, että kunkin näytteen rivisumma on yksi. [3]*

	laji 1	laji 2	laji 3	laji 4	...	rivisumma
näyte 1	0.000	0.003	0.006	0.083	...	1
näyte 2	0.000	0.000	0.056	0.001	...	1
näyte 3	0.000	0.002	0.013	0.002	...	1
näyte 4	0.000	0.010	0.079	0.181	...	1
näyte 5	0.000	0.000	0.060	0.001	...	1
näyte 6	0.000	0.261	0.011	0.076	...	1

Kuva 4 esittää Crohnin aineiston viisikymmentä ensimmäistä näytettä ja niiden lajien suhteelliset osuudet pylväsdiagrammina. Kuten kuvassa 2 myös kuvassa 4 vaaka-akselilla on viisikymmentä ensimmäistä näytettä. Pystyakselilla esitetään nyt suhteelliset lajirunsaudet eli havaittujen lajien lukumäärät suhteutettuna toisiinsa. Pylväsdiagrammien avulla on helppoa esittää kuvallisesti eri lajien osuuksia näytteissä ja näin saada yleiskuva tutkittavasta mikrobiomiaineistosta.



Kuva 4: *Pylväsdiagrammi Crohnin sairautta kuvaavasta aineistosta.* Viisikymmentä ensimmäistä näytettä on x-akselilla ja y-akselilla on lajien suhteelliset osuudet (vrt. kuva 2). Tämän esimerkin yhteydessä lajit ovat todellisuudessa sukuja. Esityksen yhtenäisyyden vuoksi tässä käytetään kuitenkin termiä laji. [3]

Normalisointi voidaan tehdä myös näytteiden harvinaistamisella (*rarefying*) yhteiseen kirjastokokoon L^*. Harvinaistamista käytetään tavallisesti silloin, kun aineiston näytteiden kirjastokoot ovat hyvin erilaisia. Harvinaistamisen tarkoituksena on taata, että harvinaisimmat lajit eivät esiinny havainnoissa todennäköisemmin ainoastaan ison kirjastokoon takia. [12].

Ensin valitaan jokin kirjastokoko L^* niin, että $L^* \leq \max_i(L_i)$, jossa L_i on näytteen i kirjastokoko. Seuraavaksi poistetaan kaikki näytteet, joille pätee $L_i < L^*$. Jäljelle jääneiden näytteiden kirjastokoot asetetaan saman kokoisiksi osa-otosten avulla. Yhteiseksi kirjastokooksi voidaan asettaa esimerkiksi pienin kirjastokoko L^*. Kirjastokoon L^* päättämisessä voidaan käyttää apuna harvinaistamiskäyrää (*rarefaction curves*). Siinä esitetään monimuotoisuus kirjastokoon funktiona. Harvinaistamista on kritisoitu muun muassa siksi, että siinä poistetaan dataa, kirjastokoko L^* valitaan mielivaltaisesti ja varianssin stabilisointi vaikeutuu. [14, 12, 11]

Edellä mainittujen normalisointitapojen sijaan voidaan käyttää log-muunnoksia. Ne ottavat paremmin huomioon aineistojen kompositionaaliseen rakenteen, koska ne riippuvat vain näytteessä olevien havaittujen runsauksien suhteista. Kuten luvussa 2.1 todettiin, lajien välinen suhteellinen runsaus on kaikissa näytteissä saman verran yli- ja aliarvioitu oikeasta runsaudesta. Log-muunnokset muuntavat havaitut lajirunsaudet näytekohtaisiksi log-suhteiksi. Log-suhteet voidaan määrittää muun muassa

additiivisella logaritmimuunnoksella (*additive log transfomation, alr*) tai keskitetyllä logaritmimuunnoksella (*centered log-ratio transformation, clr*). [12]

Additiivinen log-muunnos määritellään näytteelle i seuraavasti:

$$\text{alr}(\mathbf{x}_i) = \left[\log\left(\frac{x_{i1}}{x_{iR}}\right), ..., \log\left(\frac{x_{ip}}{x_{iR}}\right)\right],$$

jossa näytteen i jokainen lajirunsaus jaetaan jollain referenssilajirunsaudella x_{iR}. Keskitetty log-muunnos näytteelle i on taas:

$$\text{clr}(\mathbf{x}_i) = \left[\log\left(\frac{x_{i1}}{g(\mathbf{x}_i)}\right), ..., \log\left(\frac{x_{ip}}{g(\mathbf{x}_i)}\right)\right],$$

jossa $g(\mathbf{x}_i) = \left(\prod_{j=1}^{p} x_{ij}\right)^{\frac{1}{p}}$ on vektorin $\mathbf{x}_i$ alkioiden geometrinen keskiarvo. Additiivisen log-muunnoksen haasteena on määrittää lajirunsauksien referenssitaso x_{iR}. Keskitetty log-muunnos kiertää tämän haasteen jakamalla jokaisen lajirunsauden kaikkien lajien geometrisella keskiarvolla. [12]

2.3 Monimuotoisuuden tutkiminen

Näytteiden monimuotoisuutta voidaan tutkia alfa- ja beeta-monimuotoisuuden indekseillä. Alfa-monimuotoisuuden avulla tutkitaan yhden näytteen monimuotoisuutta. Kahden näytteen monimuotoisuuden eroja kuvataan beeta-monimuotoisuudella. Monimuotoisuuksien kuvaaminen on hyvin tyypillistä, kun tutkitaan mikrobiomiaineistoja. [8, 15]

2.3.1 Alfa-monimuotoisuus

Alfa-monimuotoisuus (*alpha diversity*) kuvaa, kuinka monipuolinen ja lajirikas yksi näyte on. Se tiivistää näytteen koko mikrobiomikoostumuksen yhdeksi arvoksi. Alfa-monimuotoisuutta voidaan kuvata erilaisten indeksien avulla, esimerkiksi Shannonin tai Simpsonin indeksillä, sekä lajirikkaudella ja lajien tasaisuudelle.

Lajirikkaus voidaan määritellä useilla indekseillä, esimerkiksi havaitulla lajirikkaudella. Havaittu lajirikkaus, R_{obs}, määritellään yksinkertaisesti niiden näytteessä havaittujen lajien lukumääränä, joiden havaittu lajirunsaus on suurempi kuin nolla. Havaittu lajirikkaus yleensä aliarvioi ympäristön todellista lajirikkautta, koska harvinaisimmat lajit jäävät todennäköisesti havaitsematta. [8, 12]

Shanonnin indeksi ottaa huomioon erot lajien suhteissa yhden näytteen sisällä. Shannonin indeksin määritelmä on

$$R_{Shannon} = -\sum_{j=1}^{p} \tilde{P}_{ij} \ln(\tilde{P}_{ij}),$$

jossa $\tilde{P}_{ij}$ on suhteellinen lajirunsaus kaavan (4) mukaisesti. Shannonin indeksi kasvaa, kun lajien määrä näytteessä kasvaa. Indeksillä ei ole ylärajaa, mutta tavallisesti indeksin arvot ovat alle viiden. Lähellä nollaa olevat arvot viittaavat alhaiseen monimuotoisuuteen. [2]

Simposonin indeksi määritellään seuraavasti:

$$R_{Simpson} = 1 - \sum_{j=1}^{p} \tilde{P}_{ij}^2,$$

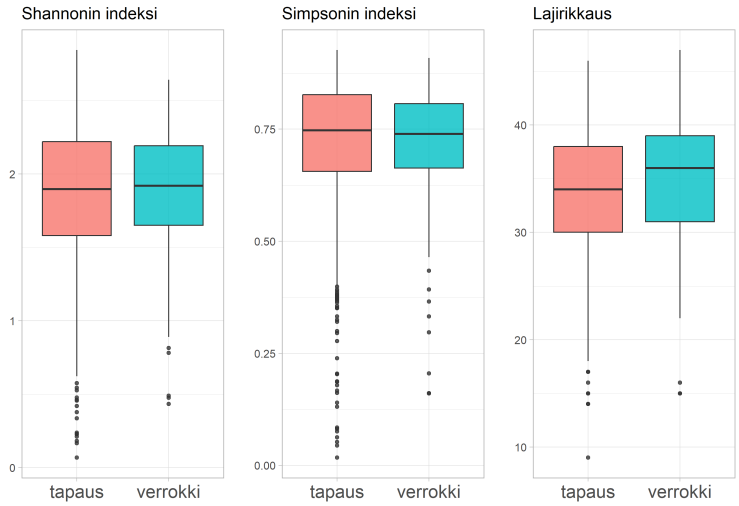
jossa $\tilde{P}_{ij}$ on määritelty kaavan (4) mukaisesti. Simpsonin indeksi voi saada arvoja nollasta lähes yhteen. Lähellä nollaa olevat arvot tarkoittavat sitä, että näytteen monimuotoisuus on alhainen. Simposonin indeksi antaa enemmän painoarvoa niille lajeille, joita esiintyy näytteessä enemmän. Shannonin indeksi taas antaa enemmän painoarvoa näytteen harvinaisimmille lajeille. [2]

Näytteen tasaisuudella tarkoitetaan sitä, kuinka lukumääräaineistossa eri näytteiden eli rivien havaitut lajirunsaudet jakautuvat. Näyte, jossa yksi tai muutama laji hallitsee lukumäärältään, ei ole tasaisesti jakautunut. Myös näytteen tasaisuutta voidaan tutkia useilla eri indekseillä, esimerkiksi käänteisellä Simpsonin indeksillä. Käänteinen Simpsonin indeksi määritellään seuraavasti:

$$R_{tasaisuus} = \frac{1}{p \sum_{j=1}^{p} \tilde{P}_{ij}^2}.$$

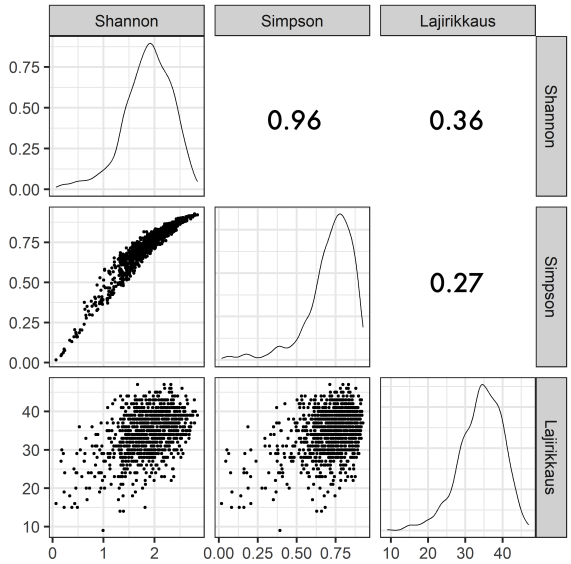
Käänteinen Simpsonin indeksi saa arvoja nollasta yhteen. Näytteen harvinaiset lajit eivät vaikuta suuresti indeksin arvoon. [2]

Kuvassa 5 on piirretty Crohnin aineiston perusteella kolme alfa-monimuotoisuuden indeksiä eli Shannonin indeksi, Simpsonin indeksi ja lajirikkaus. Shannonin ja Simpsonin indeksit ovat laskettu *vegan*-kirjaston avulla [16]. Shannonin ja Simpsonin indeksin mukaan alfa-monimuotoisuudessa ei ole suuria eroja tapauksien ja verrokkien välillä. Verrokeilla on vain hieman korkeampi lajirikkaus verrattuna tapauksiin (kuva 5).



Kuva 5: *Crohnin sairautta kuvaavan aineiston perusteella on laskettu kolme alfa-monimuoitoisuuden indeksiä; Shannonin indeksi, Simpsonin indeksi ja lajirikkaus.* Indeksien jakaumat on piirretty laatikko-janakuvioina erikseen aineiston tapauksille ja verrokeille. Sekä tapauksilla että verrokeilla on poikkeavan pieniä arvoja. Mediaanit ovat lähes samat kaikilla kolmella indeksillä, kun verrataan tapauksia ja verrokkeja. Lajirikkauden mediaani on kuitenkin verrokeilla hieman korkeammalla kuin tapauksilla eli verrokeilla on enemmän havaittuja lajeja aineistossa verrattuna tapauksiin. [3]

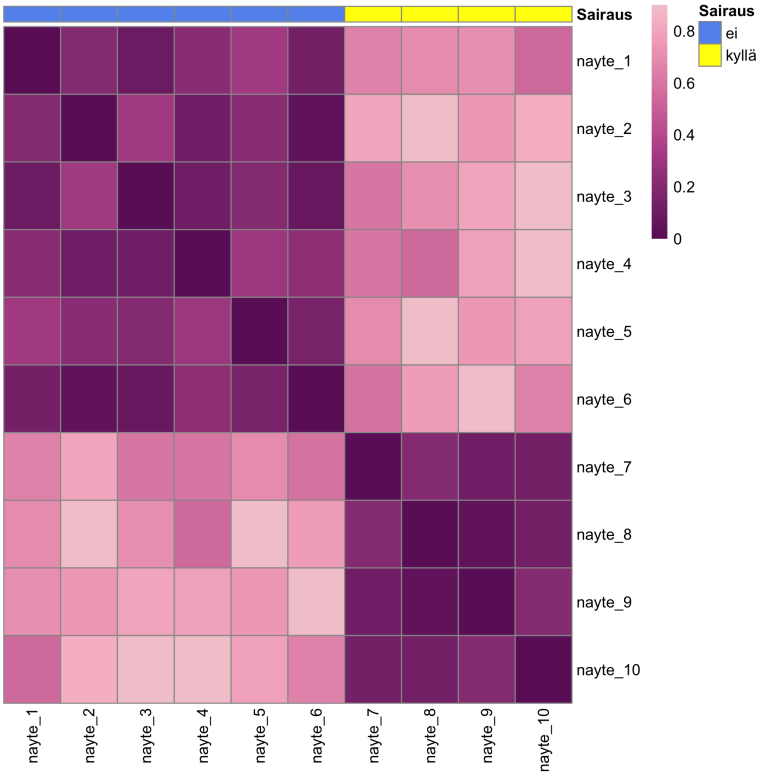
Kuvassa 6 on havainnollistettu kuvan 5 indeksien suhdetta toisiinsa. Jokaisen indeksin jakauma on hieman vino. Erityisesti Shannonin ja Simposonin indeksit ovat vahvasti korreloituneet. Niiden korrelaatiokerroin on 0.96 Spearmanin korrelaatiolla laskettuna.



Kuva 6: *Alfa-monimuotoisuuden indeksien välinen yhteys.* Crohnin sairautta kuvaavan aineiston perusteella laskettujen indeksien välinen yhteys on esitetty hajontakuvioiden avulla alakolmiolla. Yläkolmiolla on indeksien väliset korrelaatiokertoimet Spearmanin korrelaatiolla laskettuna. Indeksien jakaumat on esitetty kuvan diagonaalilla. [3]

2.3.2 Beeta-monimuotoisuus

Näytteiden monimuotoisuutta voidaan tutkia myös beeta-monimuotoisuudella. Sen avulla kuvataan, miten näytteiden välinen etäisyys eroaa ryhmien välillä (esim. sairaat vs. terveet). Beeta-monimuotoisuus ei tiivistä mikrobiomiaineistoja yhteen lukuarvoon kuten alfa-monimuotoisuus, vaan tarkoituksena on kuvata ryhmien välisiä eroja koko mikrobiston suhteen etäisyyksien avulla. Beeta-monimuotoisuutta on havainnollistettu lämpökartalla (*heatmap*) kuvassa 7.



Kuva 7: *Beeta-monimuotoisuuden havainnollistaminen lämpökartan ja kymmenen kuvitteellisen esimerkkinäytteen avulla.* Kuvaajassa on kymmenen keksittyä näytettä, joista ensimmäiset 6 on kerätty terveiltä yksilöiltä (näytteet 1–6, merkitty sinisellä) ja loput sairailta (näytteet 7–10, merkitty keltaisella). Lämpökartassa on esitetty kahden näytteen välinen etäisyys värin avulla. Väri on tumman liila, jos etäisyys on lähellä arvoa nolla. Väri vaalenee, kun etäisyys kasvaa. Kuvaajassa on nähtävissä selkeät alueet, joissa väri on aina tumman tai vaalean liila. Yhden ryhmän sisällä näytteiden väliset etäisyydet ovat siis samanlaisia (tumman liila). Ryhmien väliset etäisyydet taas eroavat toisistaan (vaalean liila). Tämä viittaa siihen, että mikrobistolla on yhteys sairauteen. Tämän esimerkin näytteiden etäisyydet keksittiin niin, että ero terveiden ja sairaiden välille saatiin selkeästi näkyviin beeta-monimuotoisuuden havainnolistamiseksi. Lämpökartta on tehty *pheatmap*-kirjaston avulla R-ohjelmistossa [17].

Etäisyys voidaan määrittää useiden indeksien avulla, esimerkiksi Bray-Curtisin etäisyydellä tai Jaccardin erilaisuuskertoimella. Jaccardin erilaisuuskerroin määritellään näytteiden A ja B välillä seuraavasti:

$$D_{Jaccard} = 1 - \frac{c}{c + b + a},$$

jossa c on niiden lajien yhteenlaskettu lukumäärä, jotka löytyvät sekä näytteestä A ja B, b on niiden lajien yhteenlaskettu lukumäärä, jotka löytyvät vain näytteestä B, ja a on niiden lajien yhteenlaskettu lukumäärä, jotka löytyvät vain näytteestä A. Jaccardin erilaisuuskerroin voi saada arvoja väliltä nollasta yhteen. Mitä pienempi kerroin on, sen samanlaisemmat kaksi näytettä ovat. [12]

Bray-Curtisin etäisyyden avulla voidaan laskea kahden näytteen kompositionaalinen ero. Määritellään Bray-Curtisin etäisyys kahden näytteen, A ja B, välillä seuraavasti:

$$D_{BC}(A, B) = \frac{\sum_{j=1}^{p} |x_{Aj} - x_{Bj}|}{\sum_{j=1}^{p} (x_{Aj} + x_{Bj})},$$

jossa x_{Aj} ja x_{Bj} ovat lajien j lukumäärät näytteissä A ja B. Lukumäärät x_{Aj} ja x_{Bj} tulee normalisoida (luku 2.1), jotta kirjastoko ei vaikuta laskettuun etäisyyteen. Bray-Curtisin etäisyys lasketaan siis summaamalla näytteiden väliset lukumäärien erot ja jakamalla se kahden näytteen lajien kokonaismäärällä. Nolla tarkoittaa sitä, että näytteet ovat täysin samanlaiset. [12, 15].

Mikäli näytteitä on enemmän kuin kaksi, Bray-Curtisin etäisyys voidaan esittää matriisina, jonka yhdessä solussa on ilmoitettu aina kahden näytteen välinen ero. Taulukossa 5 on havainnollistettu tätä Crohnin sairautta kuvaavassa aineistossa neljän ensimmäisen näytteen avulla.

Taulukko 5: *Esimerkki Bray-Curtisin etäisyysmatriisista.* Crohnin sairautta kuvaavasta aineistosta on laskettu Bray-Curtisin etäisyys neljän ensimmäisen näytteen avulla [3]. Etäisyysmatriisi laskettiin *vegan*-kirjaston avulla R-ohjelmistossa [16].

	näyte 1	näyte 2	näyte 3	näyte 4
näyte 1		0.49	0.54	0.45
näyte 2	0.49		0.47	0.49
näyte 3	0.54	0.47		0.54
näyte 4	0.45	0.49	0.54	

Alfa- ja beeta-monimuotoisuuden avulla on tutkittu erilaisia mikrobiomien ja terveysvasteiden välisiä yhteyksiä. Esimerkiksi vuonna 2019 julkaistussa artikkelissa tutkittiin alfa-monimuotoisuuden indeksien yhteyttä imettämiseen ja sylilapsen tempperamenttipiirteisiin [18]. Eräässä toisessa artikkelissa puolestaan tutkittiin suolistomikrobien ja persoonallisuuspiirteiden yhteyttä [19].

2.4 DA- ja DR-analyysit

Mikrobien tiedetään olevan yhteydessä erilaisiin sairauksiin, esimerkiksi suolistosairauksiin [1]. Usein tilastollisten analyysien tavoitteena onkin tunnistaa lajeja, jotka toimivat biomarkkereina. Biomarkkereita käytetään indikaattorina jonkin tietyn biologisen ilmiön läsnäolosta. Niitä voidaan käyttää esimerkiksi sairauksien diagnosointiin, seurantaan ja ennustamiseen. Muutokset mikrobistossa voivat olla sairauden syynä tai seurauksena.

Hyvin usein biomarkkereita etsitään ns. DA-analyysien (*differential abundance analysis*) avulla. Niiden tarkoituksena on tutkia, eroaako jonkin yksittäisen lajin runsaus eri ryhmien välillä. Esimerkiksi jos mikrobiominäytteitä on kerätty sekä sairailta että terveiltä kontrolleilta, voidaan olla kiinnostuneita, onko lajia j havaittu enemmän sairailla verrattuna kontrolleihin. Osa DA-menetelmistä pyrkii estimoimaan todellisten eli absoluuttisten runsauksien eroja. Osa taas estimoi suhteellisten lajirunsauksien eroja (kuva 8). DA-menetelmiä ovat esimerkiksi ALDEx [20], DESeq2 [13] ja ANCOM-BC [21].

DA-menetelmien vertailu on haastavaa ja parhaimman menetelmän nimeäminen on mahdotonta. Suositeltavaa onkin käyttää useampaa DA-menetelmää samaan tutkimuskysymykseen ja katsoa, miten hyvin eri menetelmien tulokset vastaavat toisiaan. Vuonna 2022 julkaistussa artikkelissa sovellettiin eri DA-menetelmiä useisiin eri mikrobiomiaineistoihin ja pyrittiin vertailemaan menetelmien tuloksia [22]. Artikkelin tulosten perusteella menetelmien löytämät biomarkkerit erosivat keskenään ja lisäksi aineistojen esikäsittelyllä (esim. normalisoinnilla) oli vaikutusta tuloksiin [22].

aineiston rakenne	*sairaat*			*kontrollit*			tulkinta
absoluuttinen lajirunsaus	*a*	*b*	*c*	*a*	*b*	*c*	kontrolleilla on enemmän lajia *a* verrattuna sairaisiin absoluuttisten lajirunsauksien mukaan
	0	0	53	15	0	93	
	5	2	28	27	0	107	
	14	0	47	14	7	65	
	13	0	33	0	5	88	
suhteellinen lajirunsaus	*a*	*b*	*c*	*a*	*b*	*c*	sairailla on enemmän lajia *a* verrattuna kontrolleihin suhteellisten lajirunsauksien mukaan
	0	0	1	0.14	0	0.86	
	0.14	0.06	0.80	0.20	0	0.80	
	0.23	0	0.77	0.16	0.08	0.76	
	0.28	0	0.72	0	0.05	0.95	

Kuva 8: *Mikrobiomiaineistoja voidaan tutkia absoluuttisina tai suhteellisina lajirunsauksina.* Riippuen valitusta analysointitavasta tulkinnat saattavat erota. Absoluuttinen lajirunsaus tarkoittaa näytteissä olevien lajien todellisia lukumääriä. Lukumääräaineiston perusteella havaitut lajirunsauden voidaan muuttaa suhteellisiksi osuuksiksi (kaava (4)). Absoluuttisen lajirunsauden mukaan kontrolleilla on enemmän lajia a verrattuna sairaisiin. Kun tutkitaan suhteellisia lajirunsauksia, tilanne onkin päinvastainen. [23]

DA-menetelmiä on kritisoitu erilaisten teoreettisten syiden takia. Jos tutkitaan absoluuttisten lajirunsauksien eroja ryhmien välillä, havaitut lajirunsaudet tulee normalisoida, jotta eri näytteitä on mielekästä vertailla (luku 2.2). Esimerkiksi DESeq2-menetelmässä käytettävä normalisointitapa olettaa, että suurin osa lajien lukumääristä on suurin piirtein samoja molemmissa ryhmissä. Oletuksen käyttäminen vaatii, että ongelman biologinen tausta ymmärretään hyvin. [23]

Myös suhteellisten lajirunsauksien tutkimisessa on ongelmia. Olkoot kolme lajia $[a, b, c]$. Lajin a suhteellinen osuus riippuu lajeista b ja c. Lajin a suhteellinen ero ryhmien välillä voi siis olla vääristynyt, koska näytteen kirjastoko vaikuttaa lajien suhteellisiin osuuksiin. Suhteellisten lajirunsauksien mukaan voi vaikuttaa siltä, että ryhmien välillä on eroja lajin a suhteen, vaikka absoluuttisten eli näytteessä olevien todellisten lukumäärien mukaan eroa ei ole (kuva 8).

Lisäksi yksittäisen lajin tutkiminen ei ota huomioon lajien biologisia yhteyksiä. Jos lajeja tutkitaan yksitellen, ei välttämättä huomata, miten useat eri lajit vaikuttavat yhdessä. Näiden syiden takia on ehdotettu vaihtoehtoisia analysointitapoja, joissa käytetään hyväksi lajien runsauksien (logaritmisia) suhteita sekä tutkitaan useita lajeja samanaikaisesti. Näitä tapoja kutsutaan ns. DR-analyyseiksi (*differential ratio analysis*). [23]

DR-analyyseissa ei tutkita yhtä lajia kerrallaan vaan tarkoituksena on löytää useiden lajien suhde, joka voidaan tulkita biomarkkerina. Koska tutkitaan lajien suhdetta, aineistojen kompositionaalinen luonne ei ole ongelma (luku 2.1). Kompositiossa $[a, b, c]$, lajin c lajirikkaus ei vaikuta lajien a ja b suhteeseen a/b (taulukko 3). Suhteita käytettäessä ei myöskään tarvitse tehdä oletusta siitä, miten lajien runsaudet ovat jakautuneet eri ryhmissä. Yksinkertaisin suhde muodostuu kahdesta lajista eli se on ns. parittainen suhde. Myös monimutkaisemmat suhteet ovat mahdollisia. [23]

DR-analyyseissä suhde voidaan muodostaa balanssien (*balances*) tai yhdelmien (*amalgamations*) avulla. Balanssi määritellään geometrisen keskiarvon avulla seuraavasti:

$$\begin{aligned} B(\mathbf{x}_i; J^+, J^-) &= \log \left(\frac{(\prod_{j \in J^+} x_{ij})^{\frac{1}{p^+}}}{(\prod_{j \in J^-} x_{ij})^{\frac{1}{p^-}}} \right) \\ &= \frac{1}{p^+} \sum_{j \in J^+} \log(x_{ij}) - \frac{1}{p^-} \sum_{j \in J^-} \log(x_{ij}), \end{aligned} \tag{8}$$

jossa J^+ ja J^- tarkoittavat toisensa poissulkevia osajoukkoja lajeista j ja p^+ ja p^- ovat osajoukkojen koot. Osajoukot eivät sisällä aineiston kaikkia lajeja, joten puhutaan harvoista osajoukoista. Balanssi on siis kahden geometrisen keskiarvon log-suhde. Jos $p^+ = p^- = 1$, kummassakin osajoukossa J^+ ja J^- on vain yksi laji. Tällöin balanssi sieventyy parittaiseksi log-suhteeksi. [23, 3]

Yhdelmä määritellään lajirunsauksien summien avulla seuraavasti:

$$A\left(\mathbf{x}_i; J^+, J^-\right) = \log \left(\frac{\sum_{j \in J^+} x_{ij}}{\sum_{j \in J^-} x_{ij}} \right). \tag{9}$$

Sekä balanssit että yhdelmät ovat joustavia siinä mielesssä, että niissä voi olla useampi laji sekä osoittajassa että nimittäjässä. Koska osajoukot muodostuvat vain pienestä osasta lajeja, balanssit ja yhdelmät tarjoavat helposti tulkittavia biomarkkereita. Lisäksi harvojen osajoukkojen ansiosta ylisovittamisen riski pienentyy, koska ei käytetä aineiston kaikkia lajeja eli dimensioita. Sekä balanssit että yhdelmät säilyttävät lajien väliset suhteet oikeina. Luvussa 2.1 esitelty taksonominen harha ei siis ole ongelma, koska tutkitaan lajien välisiä suhteita. [23, 3]

Balansseilla ja yhdelmillä on myös huonoja puolia. Greenacre ym. kuvaavat artikkelissaan geometrisen keskiarvon heikkoutta [24]. Heidän mukaansa geometriseen keskiarvoon saattaisivat vaikuttaa liian paljon aineiston harvinaisemmat lajit. Balanssien osajoukot voivat olla siis isoja, jolloin biologisten tulkintojen tekeminen vaikeutuu. Näin ollen summaus saattaisi tuottaa helpommin tulkittavia tuloksia, koska harvinaisemmat lajit eivät välttämättä esiinny yhdelmissä. Toisaalta harvinaisemmilla lajeilla voi olla tärkeä rooli tutkittavan biologisen ilmiön kannalta. Tämä tieto voidaan menettää, jos käytetään yhdelmiä. [25, 26, 3]

DR-analyysien haaste on määrittää, mistä lajeista osajoukot muodostuvat. Balanssit ja yhdelmät voidaan muodostaa joko aineiston avulla optimaalisiksi tai kirjallisuuden perusteella. CoDaCoRe-menetelmässä log-suhde määritellään vastemuuttujan kannalta parhaimmaksi aineiston avulla. Myös Rivera Pinto J. ym. kehittämä Selbal-menetelmä etsii optimaalisen balanssin aineiston avulla [27]. Koska mikrobiomiaineistojen dimensiot ovat suuria, optimaalisten osajoukkojen etsiminen aineistojen perusteella on usein laskennallisesti hidasta. CoDaCoRe-menetelmässä hyödynnetään ns. nopeaimman laskeutumisen menetelmää (luku 3.1), jonka ansiosta se on laskennallisesti erittäin tehokas. Jos log-suhde määritellään kirjallisuuden perusteella, tutkittava ilmiö tulee tuntea hyvin. [3, 23]

3 CoDaCoRe-menetelmä

CoDaCoRe-menetelmä kuuluu luvussa 2.4 esitellyihin DR-menetelmiin. Se ratkaisee hyvin mikrobiomiaineistojen tyypillisiä haasteita. CoDaCoRe-menetelmä ottaa huomioon aineistojen kompositionaalisen muodon (luku 2.1). Sekvensoinnissa syntyvä satunnainen kirjastokoko ei siis ole ongelma, koska menetelmässä tutkitaan lajien havaittujen lukumäärien log-suhteita balanssien tai yhdelmien avulla. Sekä balanssit että yhdelmät säilyttävät lajien väliset log-suhteet oikeina. CoDaCoRe-menetelmä pyrkii tehokkaasti pienentämään tutkittavien aineistojen dimensioita, koska tavoitteena on etsiä kaksi harvaa osajoukkoa, jotka ennustavat tutkittavaa vastemuuttujaa (esim. sairaat vs. terveet kontrollit) parhaiten. Harvojen osajoukkojen ansiosta CoDaCoRe-menetelmä tarjoaa helposti tulkittavia biomarkkereita. Menetelmän tärkeimpiä ominaisuuksia ovatkin sen laskennallinen tehokkuus, hyvä ennustustarkkuus ja helposti tulkittavat biomarkkerit. [3]

Luvussa 3.1 esitetään ensin muutamia määritelmiä ja aputuloksia, joita käytetään CoDaCoRe-menetelmän määrittelyssä. Tämän jälkeen selvitetään menetelmän eri vaiheita, kun vastemuuttuja on binäärinen tai jatkuva. Lukujen 3.2 ja 3.2.1 lähteinä on käytetty CoDaCoRe-menetelmän kehittäjien, Gordon-Rodriguez ym. kirjoittamaa artikkelia [3] sekä sen lisämateriaalia [28].

3.1 Määritelmiä

Määritellään sigmoid- ja ReLU-funktiot, nopeimman laskeutumisen menetelmä sekä ristiinvalidointi, joita hyödynnetään CoDaCoRe-menetelmässä.

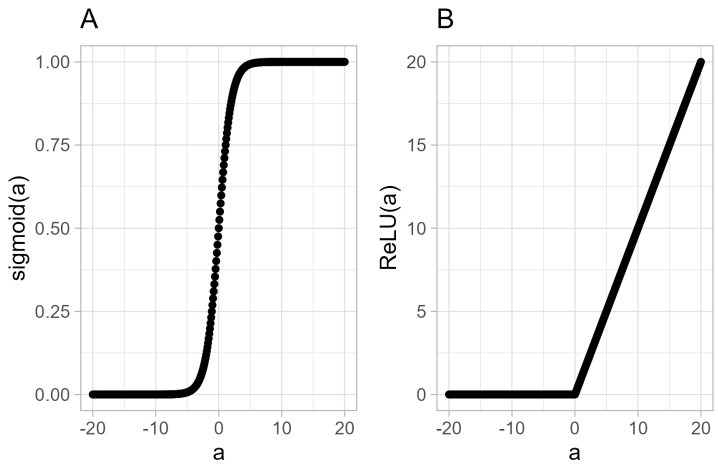
Määritelmä 3.1. Sigmoidifunktio määritellään seuraavasti [29]:

$$\text{sigmoid}(a) = \frac{1}{1+e^{-a}} = \frac{e^a}{e^a+1} = 1 - \text{sigmoid}(-a).$$

Määritelmä 3.2. ReLU-funktio (*Rectified Linear Unit*) määritellään seuraavasti [29]:

$$\text{ReLU}(a) = \max(a, 0).$$

Monissa laskennallisissa menetelmissä käytetään aktivointifunktioina sigmoidi- ja ReLU-funktioita (kuva 9). Sigmoidifunktiolla on monia hyviä puolia. Se saa arvoja nollan ja yhden välillä, jolloin se sopii hyvin todennäköisyyksien kuvaamiseen. Funktio on lisäksi differentioituva, mikä on laskennallisesti hyvä ominaisuus. ReLU-funktion avulla voidaan erotella jonkin joukon positiiviset ja negatiiviset alkiot omiksi joukoikseen. [29]



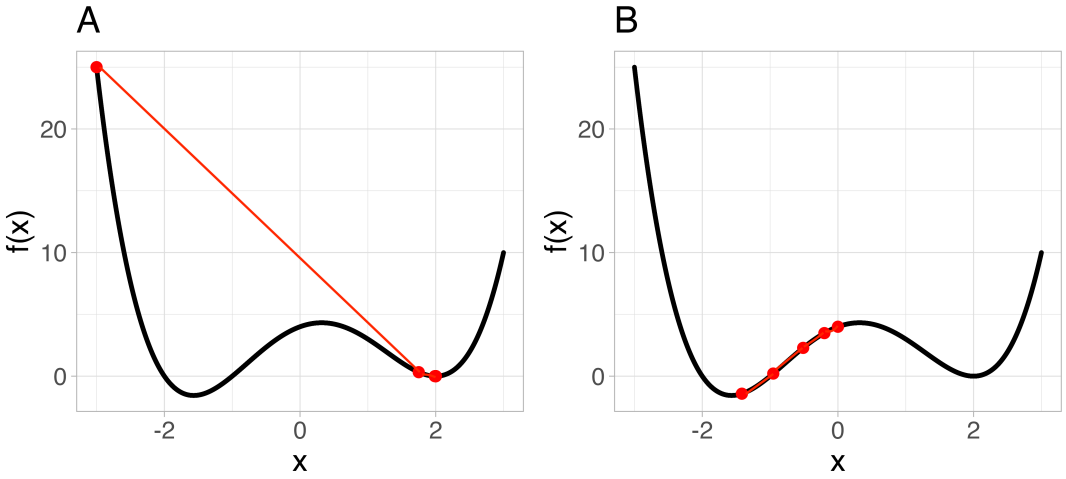
Kuva 9: *Kuvassa A on esitetty sigmoidifunktio ja kuvassa B on ReLU-funktio.* Molempia funktioita käytetään aktivointifunktioina erilaisissa menetelmissä, esimerkiksi CoDaCoRe-menetelmässä.

Nopeimman laskeutumisen menetelmä *(gradient descent)* käyttää hyväkseen derivaattaa ja on luonteeltaan ns. ahne algoritmi. Ahne algoritmi tarkoittaa nopeimman laskeutumisen menetelmän tapauksessa sitä, että jokaisella iteraatiokierroksella edetään mahdollisimman paljon siihen suuntaan, jossa funktion arvo pienenee eniten [30]. Menetelmän kyky löytää optimiarvo on sidonnainen lähtöpisteeseen, mitä on havainnollistettu esimerkissä 3.3. Lisäksi menetelmä ei takaa, että löydetään funktion globaalia optimipistettä vaan voidaan päätyä esimerkiksi lokaaliin optimiin. Nopeimman laskeutumisen menetelmä voidaan kirjoittaa seuraavasti:

$$x_{i+1} = x_i - \nu \cdot \nabla f(x_i),$$

jossa uusi piste x_{i+1} määritellään edellisen iteraatiokierroksen i pisteen x_i sekä derivaatan $\nabla f(x_i)$ arvon avulla. Parametria ν sanotaan askelpituudeksi (*learning rate*). Askelpituuden avulla hallitaan menetelmän etenemistä. Menetelmässä määritetään lopetuskriteeri $||\nabla f(x_i)|| < \epsilon$, jollain $\epsilon > 0$. [29, 30]

Esimerkki 3.3. Olkoon optimoitava funktio $f(x) = 0.5x^4 - 0.5x^3 - 3x^2 + 2x + 4$. Kuvan 10 perusteella funktion globaali minimi on kohdassa $x \approx -1.57$. Etsitään minimiarvo nopeimman laskeutumisen menetelmällä. Tutkitaan kahta eri alkuarvoa: $x_{01} = -3$ ja $x_{02} = 0$. Olkoon askelpituus $\nu = 0.1$. Lähdettäessä alkuarvosta $x_{01} = -3$ löydetään lokaali optimiarvo, mutta ei globaalia. Lähdettäessä alkuarvosta $x_{02} = 0$ algoritmi löytää globaalin optimin (kuva 10).



Kuva 10: *Funktion* $f(x) = 0.5x^4 - 0.5x^3 - 3x^2 + 2x + 4$ *minimiarvon etsiminen nopeimman laskeutumisen menetelmän avulla.* Kuvassa A lähtöarvona on $x_{01} = -3$ ja kuvassa B lähtöarvona on $x_{02} = 0$. Molemmissa tapauksissa askelpituus on $\nu = 0.1$. Lähdettäessä alkuarvosta x_{01} algortimi löytää lokaalin optimiarvon. Kun taas lähdetään alkuarvosta x_{02}, algoritmi löytää globaalin optimin.

Määritelmä 3.4. Ristiinvalidoinnin (*cross validation*) avulla voidaan tutkia muodostettujen mallien suorituskykyä. Menetelmässä aineisto jaetaan opetus- ja testiaineistoon. Opetusaineisolla muodostetaan mallit ja testiaineistolla arvoidaan, kuinka hyvin mallit ennustavat uusia havaintoja. Ristiinvalidoinnissa mallit koulutetaan useita kertoja käyttäen eri osia aineistosta opetus- ja testiaineistona. Kun kaikki mahdolliset opetus- ja testiaineiston jaot on käyty läpi, malleille lasketaan keskiarvo ennustevirheistä. Näin voidaan valita paras malli eli malli, jonka ennustevirhe on pienin tarkasteltujen mallien virheistä. [29]

3.2 Binäärinen vastemuuttuja

DR-menetelmien haasteena on löytää kaksi lajien vähälukuista eli harvaa osajoukkoa. CoDaCoRe-menetelmä ratkaisee tämän ongelman optimoimalla jokaiselle lajille oman painonsa. Painojen avulla lajit voidaan jakaa negatiiviseen ja positiiviseen joukkoon. Loppujen lopuksi balanssiin tai yhdelmään valitaan vain ne lajit, joiden painot ovat riittävän suuria. CoDaCoRe-menetelmän algoritmissa on kaksi vaihetta osajoukkojen määrittämiseen, jatkuva relaksointi ja diskretointi.

Olkoon lukumääräaineisto seuraavanlainen:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}, \tag{10}$$

jossa ei saa esiintyä nollasoluja, koska balanssit ja yhdelmät muodostetaan logaritmin avulla. Merkitään datamatriisin $\mathbf{X}$ riviä i notaatiolla $\mathbf{x}_i$ ja vastemuuttujaa

$$y = \begin{pmatrix} y_1 & y_2 & \dots & y_n \end{pmatrix}',$$

joka voi olla binäärinen tai jatkuva. Oletetaan ensin, että havainnot y_i, $i = 1, ..., n$, ovat binäärisiä eli voivat saada arvoja 0 tai 1. Luvussa 3.2.1 yleistetään tulokset myös jatkuvalle vastemuuttujalle.

CoDaCoRe-menetelmässä tapahtuman $y_i = 1$ todennäköisyys p_i kirjoitetaan seuraavasti:

$$\begin{aligned} \log\left(\frac{p_i}{1-p_i}\right) &= \alpha + \beta \cdot B(\mathbf{x}_i; J^+, J^-) \\ \Leftrightarrow \qquad p_i &= \frac{1}{1 + e^{-(\alpha+\beta \cdot B(\mathbf{x}_i; J^+, J^-))}}, \end{aligned} \tag{11}$$

jossa $B(\mathbf{x}_i; J^+, J^-))$ tarkoittaa lausekkeen (8) mukaisesti määritettyä balanssia sekä parametrit α ja β ovat skalaareja. Balanssin tilalla voidaan käyttää myös kaavan (9) mukaista yhdelmää.

Uskottavuusfunktion avulla voidaan mitata ennustettujen todennäköisyyksien ja todellisten havaintojen välistä eroa. Uskottavuusfunktio L (*binary cross-entropy*) on binäärisen vastemuuttujan tapauksessa muotoa:

$$\begin{aligned} L((\alpha, \beta, J^+, J^-); y) &= \sum_{i=1}^{n} \left[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i) \right] \\ &= \sum_{i=1}^{n} \left[y_i \cdot \log\left(\frac{1}{1 + e^{-(\alpha+\beta \cdot B_i)}}\right) \right. \\ &\quad \left. + (1 - y_i) \cdot \log\left(\frac{e^{-(\alpha+\beta \cdot B_i)}}{1 + e^{-(\alpha+\beta \cdot B_i)}}\right) \right], \end{aligned} \tag{12}$$

jossa y_i on havaittu arvo 0 tai 1, p on tapahtuman $y_i = 1$ todennäköisyys ja merkintä B_i tarkoittaa kaavan (8) mukaisesti määriteltyä balanssia, $B_i = B(\mathbf{x}_i; J^+, J^-)$.

CoDaCoRe-menetelmän esittelevässä artikkelissa on käytetty koneoppimisen termejä. Tyypillisesti koneoppimisen alalla minimoidaan virhefunktiota, joten CoDaCoRe-menetelmän tavoitteena on minimoida kaavan (12) mukaista negatiivista uskottavuusfunktiota:

$$\min_{(J^+, J^-, \alpha, \beta)} -L((\alpha, \beta, J^+, J^-); y). \tag{13}$$

Yhtälön (13) ratkaisua vaikeuttaa osajoukkojen J^+ ja J^- yhteinen optimointi. Osajoukot J^+ ja J^- määrittelevät ne lajit, jotka muodostavat balanssin. Lajien lukumäärä on p, joten mahdollisten osajoukkojen lukumäärä on 2^p. Mahdollisten balans-

sien määrä on siis vielä paljon suurempi kuin 2^p. Eksaktien optimiarvojen löytäminen on laskennallisesti raskasta. Kaavan (13) mukaista optimointiongelmaa helpotetaan jatkuvan relaksoinnin (*continuous relaxation*) avulla. Jatkuvan relaksoinnin tavoitteena on saattaa kaavan (12) mukainen uskottavuusfunktio sellaiseen muotoon, että se on derivoituva ja voidaan optimoida nopeimman laskeutumisen menetelmän avulla.

Jatkuva relaksointi. Olkoon $\mathbf{w} \in R^p$ painovektori. Vektorissa $\mathbf{w}$ jokaisella lajilla on oma reaaliarvoinen painonsa eli $-\infty < w_j < \infty$, $j = 1, ..., p$. Painojen avulla määritetään pehmeiden rajoitteiden (*soft assignment*) vektori $\tilde{\mathbf{w}}$. Määritellään vektorin $\tilde{\mathbf{w}}$ alkiot sigmoidifunktion (3.1) avulla:

$$\tilde{w}_j = 2 \cdot \text{sigmoid}(w_j) - 1 = \frac{2}{1 + \exp(-w_j)} - 1. \tag{14}$$

Kun painolla w_j on suuri positiivinen arvo, alkion $\tilde{w}_j$ arvo on lähellä arvoa 1. Kun taas painolla w_j on pieni negatiivinen arvo, alkion $\tilde{w}_j$ arvo on lähellä arvoa -1. Vektorin $\tilde{\mathbf{w}}$ perusteella lajit voidaan jakaa positiiviseen ja negatiiviseen osajoukkoon ReLU-funktion (3.2) avulla:

$$\begin{aligned} \tilde{w}_j^+ &= \text{ReLU}(\tilde{w}_j) \quad \text{ja} \\ \tilde{w}_j^- &= \text{ReLU}(-\tilde{w}_j). \end{aligned}$$

Kaavan (8) mukaisesti määriteltyä balanssia voidaan siis approksimoida seuraavasti:

$$\begin{aligned} \tilde{B}(\mathbf{x}_i; \tilde{\mathbf{w}}) &= \frac{\sum_j \tilde{w}_j^+ \log(x_{ij})}{\sum_j \tilde{w}_j^+} - \frac{\sum_j \tilde{w}_j^- \log(x_{ij})}{\sum_j \tilde{w}_j^-} \\ &= \frac{\tilde{\mathbf{w}}^+ \cdot \log(\mathbf{x}_i)}{\|\tilde{\mathbf{w}}^+\|_1} - \frac{\tilde{\mathbf{w}}^- \cdot \log(\mathbf{x}_i)}{\|\tilde{\mathbf{w}}^-\|_1}, \end{aligned} \tag{15}$$

jossa merkintä $\|\cdot\|_1$ tarkoittaa L_1-normia. Lausekkeessa (15) lasketaan osajoukkojen havaittujen lajirunsauksien geometrinen keskiarvo painottaen sitä vektoreilla $\tilde{\mathbf{w}}^+$ ja $\tilde{\mathbf{w}}^-$. Yhtälön (15) mukaista approksimointia kutsutaan *jatkuvaksi relaksoinniksi*. Vastaavanlainen relaksointi voidaan muodostaa myös yhdelmille:

$$\tilde{A}(\mathbf{x}_i; \tilde{\mathbf{w}}) = \log\left(\frac{\sum_j \tilde{w}_j^+ x_{ij}}{\sum_j \tilde{w}_j^- x_{ij}}\right).$$

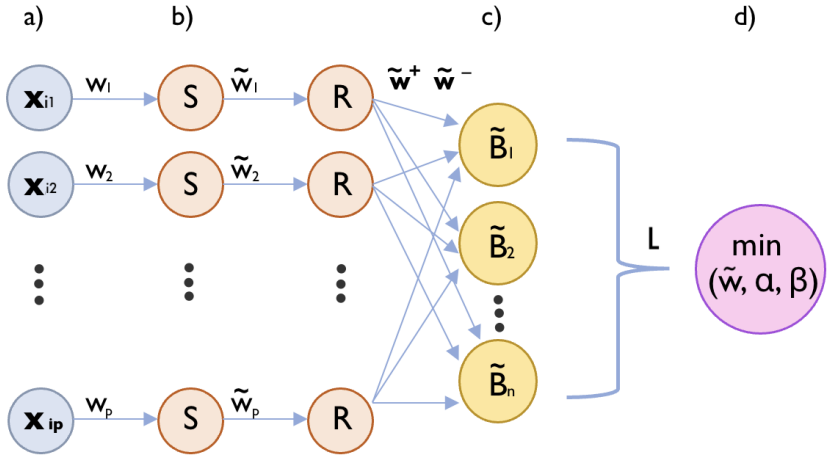
Jatkuvan relaksoinnin ansiosta kaavan (13) optimointiongelmassa esiintyvä uskottavuusfunktio saadaan muutettua helpommin optimoitavaan muotoon:

$$L((\alpha, \beta, \tilde{\mathbf{w}}); y) = \sum_{i=1}^{n} \left[y_i \cdot \log\left(\frac{1}{1 + e^{-(\alpha + \beta \cdot \tilde{B}_i)}}\right) + (1 - y_i) \cdot \log\left(\frac{e^{-(\alpha + \beta \cdot \tilde{B}_i)}}{1 + e^{-(\alpha + \beta \cdot \tilde{B}_i)}}\right) \right],$$

jossa balanssi $\tilde{B}_i$ on kaavan (15) mukainen. Uudelleen kirjoitetussa optimointiongelmassa riittää siis optimoida painovektori $\tilde{\mathbf{w}}$ ja skalaarit α ja β:

$$\min_{(\tilde{\mathbf{w}},\alpha,\beta)} -L((\alpha,\beta,\tilde{\mathbf{w}});y). \tag{16}$$

Optimiarvot etsitään nopeimman laskeutumisen menetelmällä, joka hyödyntää derivaattaa ja on siten laskennallisesti tehokas. Tämän ansiosta CoDaCoRe-menetelmän laskennallinen tehokkuus skaalautuu lineaarisesti eli menetelmän suoritusaika kasvaa lähes lineaarisesti, kun lajien lukumäärä kasvaa aineistossa $\mathbf{X}$. Jatkuvan relaksoinnin vaiheet on tiivistetty kuvassa 11.



Kuva 11: *Jatkuvan relaksoinnin vaiheet tiivistettynä.* Kohta a: jokaiselle lajille j asetetaan paino w_j. Kohta b: sigmoidi- ja ReLU-funktioiden avulla painot saadaan jaoteltua negatiivisiin ($\tilde{w}_j^-$) ja positiivisiin ($\tilde{w}_j^+$) osiin. Kohta c: approksimoidaan balanssit. Kohta d: uskottavuusfunktio L optimoidaan nopeimman laskeutumisen menetelmän avulla.

Diskretointi. Lausekkeen 15 mukaiset ratkaisut ovat haastavia tulkita, koska suurin osa vektorin $\tilde{\mathbf{w}}$ alkioista on lähellä nollaa. Pieni osa taas konvergoi kohti arvoja $+1$ tai -1. Tämän takia CoDaCoRe-menetelmässä suoritetaan diskretointi. Diskretoinnin tarkoitus on karsia pois ne lajit, joiden paino $\tilde{w}_j$ on lähellä nollaa. Näin voidaan tunnistaa harvoja osajoukkoja ja saavuttaa yksinkertaisempia biomarkkereita.

Diskretointi suoritetaan kynnysarvon $t \in (0,1)$ avulla. Merkitään

$$\begin{aligned} \tilde{J}^+ &= \{j : \tilde{w}_j > t\}, \\ \tilde{J}^- &= \{j : \tilde{w}_j < -t\}. \end{aligned} \tag{17}$$

Korkealla kynnysarvon t arvolla, osajoukkoihin $\tilde{J}^+$ ja $\tilde{J}^-$ määräytyy vähemmän lajeja eli tuloksena on harvempi malli. CoDaCoRe-menetelmässä käydään läpi kaksi-

kymmentä ehdokasarvoa optimaaliselle kynnysarvolle $\hat{t}$. Lopulta kynnysarvo $\hat{t}$ määritetään ristiinvalidoinnin avulla (määritelmä 3.4). Tarkastellaan ensin kuitenkin esimerkin 3.5 avulla, miten ehdokaskynnysarvot määritetään ja tämän jälkeen tarkennetaan ristiinvalidoinnin vaiheita.

Esimerkki 3.5. Tarkastellaan ehdokaskynnyarvojen muodostumista esimerkin avulla. Olkoot painot

$$\tilde{w}_j^+ = \{0.2, 0.4, 0.6, 0.8\} \quad \text{ja}$$
$$\tilde{w}_j^- = \{0.3, 0.5, 0.75\}.$$

Skaalataan arvot niin, että $\max \tilde{\mathbf{w}}^+ = \max \tilde{\mathbf{w}}^-$. Painoiksi saadaan tällöin

$$0.2,\ 0.4,\ 0.6,\ 0.8 \text{ ja } 0.32,\ 0.5333,\ 0.8.$$

Järjestetään arvot suuruusjärjestykseen:

$$\{t_{(1)}, t_{(2)}, ..., t_{(6)}\} = \{0.8, 0.6, 0.533, 0.4, 0.32, 0.2\},$$

jossa $t_{(1)}$ on ehdokaskynnysarvojen isoin arvo. Tässä esimerkissä ehdokaskynnysarvoja on siis vain 7 kappaletta. Skaalauksen ansiosta joukkoihin J^+ ja J^- tulee yhteensä kaksi lajia, kun kynnysarvona on $t_{(1)}$ eli suurin ehdokaskynnysarvo. Kun kynnysarvona on $t_{(2)}$, joukot muodostuvat kolmesta lajista ja niin edelleen.

Kun ehdokaskynnysarvot $\{t_{(1)}, t_{(2)}, ..., t_{(20)}\}$ on määritetty, aineisto jaetaan viiteen osaan ristiinvalidointia varten. Ristiinvalidoinnissa jokaista osaa käytetään vuorotellen testiaineistona ja jäljelle jääviä neljää osa-aineistoa käytetään opetusaineistona. Ristiinvalidoinnin aikana jokaiselle ehdokaskynnysarvolle $\{t_{(1)}, t_{(2)}, ..., t_{(20)}\}$ muodostetaan seuraava optimointiongelma:

$$\min_{(\alpha,\beta)} -L((\alpha, \beta, \tilde{\mathbf{w}}); y), \tag{18}$$

jossa uskottavuusfunktiossa esiintyvän approksimoidun balanssin (kaava (15)) painot $\tilde{\mathbf{w}}$ on jo optimoitu jatkuvan relaksoinnin vaiheessa (kuva 11:kohta d) ja osajoukot $\tilde{J}^+$ ja $\tilde{J}^-$ on muodostettu tutkittavan kynnysarvon $\{t_{(1)}, t_{(2)}, ..., t_{(20)}\}$ mukaisesti (kaava (17)). Käytännössä tämä tarkoittaa logistisen regression sovittamista. Logistisen regression sovittaminen on laskennallisesti tehokkaampaa kuin optimoida kaavan (16) mukaista mallia.

Kynnysarvoille $\{t_{(1)}, t_{(2)}, ..., t_{(20)}\}$ voidaan laskea ristiinvalidointiin perustuva keskimääräinen ennustevirhe, joka lasketaan viidestä ennustevirheestä. Optimaaliseksi kynnysarvoksi $\hat{t}$ valitaan suurin kynnysarvo siten, että sen mukaisesti muodostettu ennustevirhe on yhden standardipoikkeaman sisällä parhaimmasta ennustevirheestä (*1-standard-error rule*). Standardipoikkeama määritellään tarkasteltavalle kynnysarvolle jakamalla sen mukaiset ennustevirheiden keskihajonnat neliöjuuri viidellä. Yhden standardipoikkeaman ehtoa voidaan säädellä parametrin λ avulla, $\lambda \in [0, 1]$. Esimerkiksi jos asetetaan parametrin λ arvoksi 0.8, ennustevirheen tulee olla arvon 0.8 standardipoikkeaman sisällä parhaimmasta ennustevirheestä. Suurin kynnysarvo

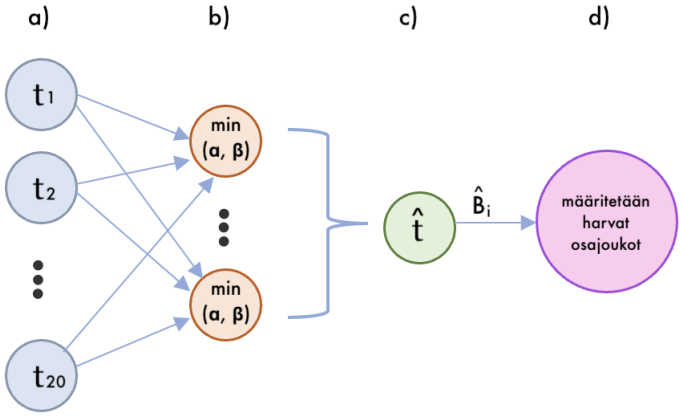
valitaan sen takia, että balanssi muodostuisi mahdollisimman harvoista osajoukoista. Parametrin λ valintaa voidaan pitää ristiinvalidoinnin regularisoimisena, koska sen avulla otetaan huomioon ristiinvalidointiin liittyvä epävarmuus ja valitaan se vaihtoehto, jossa keskimääräinen ennustevirhe voisi olla pienin. Mikäli valitaan alhainen parametrin λ arvo eli lähellä nollaa oleva arvo, tuloksena ei ole niin harvat osajoukot kuin parametrin λ korkeilla arvoilla (lähellä arvoa yksi). Toisaalta jos parametrin arvo on alhainen, malli sopii paremmin tutkittavaan mikrobiomiaineistoon, koska osajoukkojen tulisi muodostua suuremmasta määrästä lajeja.

Kun optimaalinen kynnysarvo $\hat{t}$ on määritetty, koko aineistoon sovitetaan kaavan 11 mukainen malli:

$$\log\left(\frac{p_i}{1-p_i}\right) = \hat{\alpha} + \hat{\beta} \cdot B(\mathbf{x}_i; \hat{J^+}, \hat{J^-})$$
$$\Leftrightarrow \qquad p_i = \frac{1}{1 + e^{-\hat{\alpha}+\hat{\beta}\cdot B(\mathbf{x}_i;\hat{J^+},\hat{J^-}))}}$$

jossa kaavan (8) balanssin osajoukot $\hat{J}^+$ ja $\hat{J}^-$ on muodostettu optimaalisen kynnysarvon $\hat{t}$ avulla.

CoDaCoRe-menetelmä tunnistaa sarjan balansseja tärkeysjärjestykssä. Menetelmän tuloksena voi siis olla esimerkiksi kolme balanssia, joista ensimmäinen on tärkein ja kolmas vähiten tärkeä. Kun ensimmäinen balanssi on määritetty, uusi ehdokasbalanssi sovitetaan ensimmäisen balanssin jäännöksiin. Ehdokasbalansseja sovitetaan peräkkäin, kunnes uusi balanssi ei paranna ristiinvalidoinnin tulosta. Mikäli menetelmä tunnistaa esimerkiksi kolme balanssia, kaikki balanssit yhdessä parantavat vastemuuttujan ennustamista. Ensimmäistä tunnistettua balanssia voidaan käyttää yksinään vasteen ennustamiseen, mutta muita ei. Tämä johtuu siitä, että jälkimmäiset balanssit on muodostettu edellisten jäännöksistä. CoDaCoRe-malli palauttaa estimaatit parametreille α ja β sekä lajit, joista löydetyt osajoukot muodostuvat. Diskretoinnin vaiheet on esitetty kuvassa 12.



Kuva 12: *Diskretoinnin vaiheet tiivistettynä.* Kohta a: valitaan ehdokaskynnysarvot $\{t_i, t_2, ..., t_{20}\}$. Kohta b: Optimaalinen kynnysarvo $\hat{t}$ etsitään ristiinvalidoinnin avulla, jossa aineisto jaetaan viiteen osaan. Jokaiselle kynnysarvolle muodostetaan siis kaavaa (18) vastaava optimointiongelma viisi kertaa. Kohta c: Ristiinvalidoinnin tuloksena valitaan kynnysarvo $\hat{t}$ niin, että se tuottaa mahdollisimman harvat osajoukot. Kohta d: muodostetaan optimaalisen kynnysarvon $\hat{t}$ avulla kaavan (3.2) mukainen regressiomalli koko aineistolle.

3.2.1 Jatkuva vastemuuttuja

CoDaCoRe-menetelmässä voidaan mallintaa myös jatkuvaa vastetta. Menetelmän vaiheet pysyvät samana, mutta optimoitava uskottavuusfunktio muuttuu keskineliövirheen (*mean-squared-error*) mukaiseksi.

Olkoon y jatkuva vastemuuttuja. Lineaarinen regressiofunktio on muotoa:

$$y_i = \alpha + \beta \cdot B(\mathbf{x}_i, J^+, J^-)$$

Log-uskottavuusfunktioksi saadaan:

$$L = \frac{1}{n} \sum_{i=1}^{n} (y_i - (\alpha + \beta \cdot B_i))^2, \tag{19}$$

jossa balanssi B_i on kaavan (8) mukainen. Balanssien jatkuva relaksointi ja osajoukkojen diskretointi tapahtuu luvun 3.2 mukaisesti.

4 Aineisto ja tutkimusmenetelmät

Tämän tutkielman soveltavassa osiossa käytetty mikrobiomiaineisto on kerätty FinnBrain-syntymäkohorttitutkimuksesta. FinnBrain on Turun yliopistossa vuonna 2010 aloitettu tutkimus, jonka tarkoituksena on selvittää ympäristön ja perimän vaikutusta lapsen kehitykseen. Tutkimuksessa ollaan erityisen kiinnostuneita lasten aivojen ja stressinsäätelyjärjestelmän kehittymisestä. FinnBrainiin on lähtenyt mukaan yli 4 000 perhettä Turusta sekä ympäristökunnista ja Ahvenanmaalta. Perheet on kutsuttu tutkimukseen ensimmäisen ultraäänitutkimuksen yhteydessä. Tarkoituksena on seurata tutkimuksessa mukana olevia lapsia aina aikuisikään saakka. FinnBrain-tutkimus tuottaa täysin uutta tietoa lapsen aivojen kehityksestä ja kehitykseen vaikuttavista tekijöistä. Tietoa sovelletaan mm. lasten ja nuorten terveyden hyväksi ja tehokkaiden tukitoimien kehittämiseen. [31]

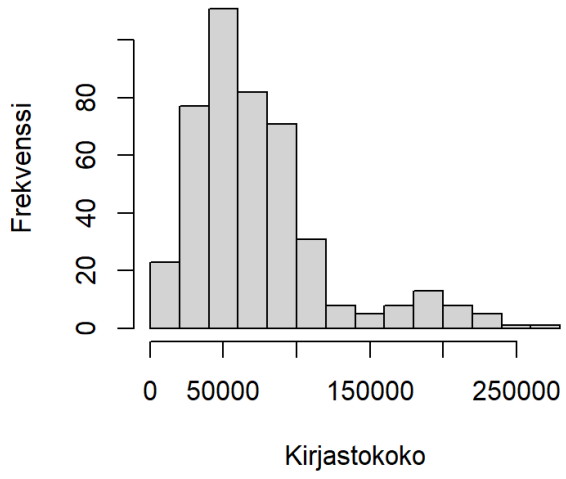
Tutkittavilta perheiltä on kerätty erilaisia tietoja mm. kyselylomakkeiden, psykologisten tutkimusten ja biologisten näytteiden avulla. Ulostenäytteitä on kerätty lapsilta eri ikäpisteissä. Tässä tutkielmassa käytetään 2.5 kuukauden ikäisten vauvojen näytteitä ($n = 444$). Näytteet on sekvensoitu 16S rRNA-sekvensoinnilla. Tarkempi kuvaus näytteiden keräämisestä, säilyttämisestä ja analysoimisesta on esitetty Aatsinki ym. artikkelissa [18].

Tässä tutkielmassa mikrobiomiaineistoa tutkittiin sukutasolla (kuva 1). Aineistossa oli yhteensä 97 sukua, joiden kaikkien kunta oli bakteeri. Kaikkia sukuja ei oltu tunnistettu sekvensoinnissa, joten lukumääräaineistoon niiden nimeksi oli asetettu Unidentified Genus. Analyyseissä pidettiin myös nämä 11 tunnistamatonta sukua mukana (taulukko 6). Lisäksi suvut Eubacterium, Clostridium ja Ruminococcus esiintyivät aineistossa useammin kuin kerrran, sillä niillä oli eri lahko- ja heimoluokittelu (taulukko 6). Näiden sukujen muuttujien nimet määriteltiin heimotason avulla. Esimerkiksi taulukon 6 ensimmäisen rivin tunnistamattoman suvun nimeksi asetettiin Enterobacteriaceae:Unidentified Genus. Rivin 12 suvun muuttujan nimeksi asetettiin Erysipelotrichaceae:Eubacterium. Jos heimoa ei oltu tunnistettu, käytettiin lahkoa nimeämisen apuna ja niin edelleen.

Taulukko 6: *Taulukossa on esitetty sekä tunnistamattomien sukujen että Eubacterium-, Clostridium- ja Ruminococcus-sukujen taksonominen asteikko.* Tunnistamaton taksoni on merkitty viivalla (-).

	luokka	**lahko**	**heimo**	**suku**
1	Gammaproteobacteria	Enterobacterales	Enterobacteriaceae	-
2	Bacilli	Lactobacillales	Enterococcaceae	-
3	Actinobacteria	Actinomycetales	Actinomycetaceae	-
4	Desulfovibrionia	Desulfovibrionales	Desulfovibrionaceae	-
5	Negativicutes	Veillonellales	-	-
6	-	-	-	-
7	Clostridia	Lachnospirales	Lachnospiraceae	-
8	Clostridia	Oscillospirales	Ruminococcaceae	-
9	Bacteroidia	Bacteroidales	-	-
10	Clostridia	Oscillospirales	-	-
11	Clostridia	-	-	-
12	Bacilli	Erysipelotrichales	Erysipelotrichaceae	Eubacterium
13	Clostridia	Lachnospirales	Lachnospiraceae	Eubacterium
14	Clostridia	Clostridiales	Clostridiaceae	Clostridium
15	Clostridia	Lachnospirales	Lachnospiraceae	Clostridium
16	Clostridia	Oscillospirales	DTU089	Clostridium
17	Clostridia	Lachnospirales	Lachnospiraceae	Ruminococcus
18	Clostridia	Oscillospirales	DTU089	Ruminococcus

Liitteessä A2 on esitetty aineiston kaikki lajit pylväsdiagrammin avulla. Aineiston viisi yleisintä sukua (*top taxa*) olivat Bifidobacterium, Escherichia, Veillonella, Bacteroides ja Clostridiaceae:Clostridium. Analyyseissa pidettiin kaikki näytteet mukana eli ei tehty harventamista kirjastokoon perusteella (kuva 13).

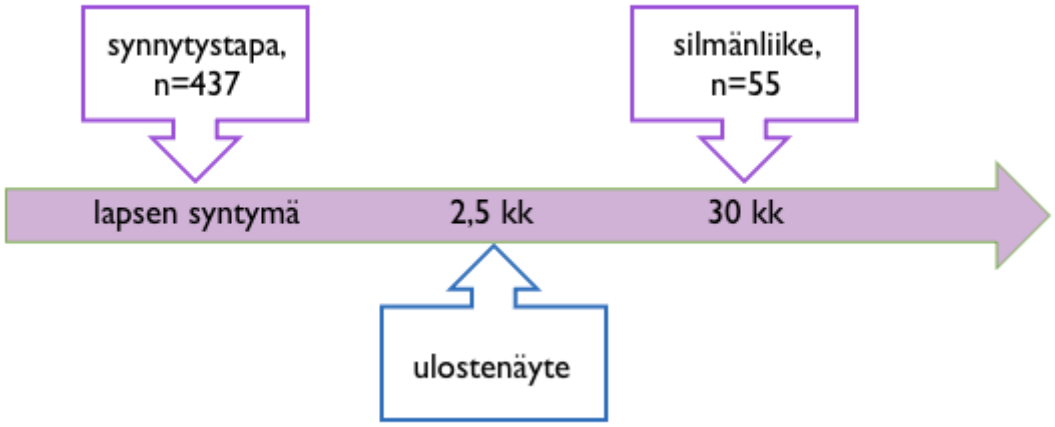


Kuva 13: *Aineiston eri näytteiden kirjastokoot on esitetty histogrammin avulla.* Kirjastokoot ovat x-akselilla ja frekvenssit y-akselilla. Aineiston pienin kirjastokoko oli 12 252 ja suurin 260 227. Kirjastokokojen mediaani oli 63672. Valtaosa näytteiden kirjastokoista oli välillä 40 000 – 90 000.

4.1 Soveltavan osion vastemuuttujat

Soveltavan osion vastemuuttujiksi valittiin vaihtoehtoisesti joko binäärinen muuttuja (synnytystapa) tai jatkuva muuttuja (katseen irtautumisen todennäköisyys). Aineistossa valtaosa oli synnyttänyt alateitse, 363 (83 %), ja sektion avulla synnyttäneitä oli 74 (17 %). Puuttuvia havaintoja oli seitsemän. Jatkuvana vastemuuttujana käytetty todennäköisyysmuuttuja on mitattu lapsen silmänliikkeen perusteella. Silmänliikemittauksissa lapsi katsoo tietokoneen ruutua, jossa on kuvia ihmisen kasvoista. Kasvoilla on erilaisia ilmeitä, esimerkiksi pelokkuus ja iloisuus. Kun lapsi katsoo ilmeikästä kasvokuvaa, ruudulle ilmestyy häiriönä jokin geometrinen muoto. Lapsen silmänliikkeestä lasketaan todennäköisyys sen perusteella, siirsikö lapsi katseensa pois kasvoista yhden sekunnin kuluessa. Maksimissaan lapselle näytetään yhtä ilmettä kuusi kertaa. Osa kerroista saattoi olla epäkelpoisia, jolloin niitä ei otettu huomioon todennäköisyysmuuttujassa. Todennäköisyys yksi tarkoittaa sitä, että lapsen katse siirtyi aina pois kasvoista. [32]

Silmänliike on mitattu, kun lapsi on ollut 30 kuukauden ikäinen. Tässä tutkielmassa tarkasteltiin irtautumisen todennäköisyyttä pelokkaasta ilmeestä. Niitä, joilta oli 30 kuukauden iässä mitattu silmänliike ja 2.5 kuukauden ikäisenä annettu ulostenäyte, oli 55. Todennäköisyysmuuttujan mediaani oli 0.67 sekä 1. kvartaali oli 0.5 ja 3. kvartaali oli 0.83. Aikajana eri muuttujien mittausajoista on esitetty kuvassa 14.



Kuva 14: *Aikana, jossa on kuvattu tämän tutkielman kannalta tärkeimmät muuttujat FinnBrain-syntymäkohortista.* Lapsia, joilta on tiedossa synnytystapa ja ulostenäyte on kerätty 2.5 kuukauden ikäisenä, oli 437. Lapsia, joilta on mitattu silmänliike 30 kuukauden iässä ja ulostenäyte on kerätty 2.5 kuukauden ikäisenä, oli 55. Silmänliikemittauksien perusteella on johdettu jatkuvana vastemuuttujana käytetty todennäköisyysmuuttuja, joka kuvaa irtautumisen todennäköisyyttä pelokkaista kasvoista.

4.2 Tutkimuskysymykset ja -menetelmät

Tämän tutkielman tärkeimmät tutkimuskysymykset CoDaCoRe-menetelmän soveltamisessa olivat: (1) miten balanssien ja yhdelmien löytämät osajoukot eroavat toisistaan, (2) miten parametri λ vaikuttaa löydettyihin osajoukkoihin sekä (3) miten hyvin mallit ennustavat uusia havaintoja.

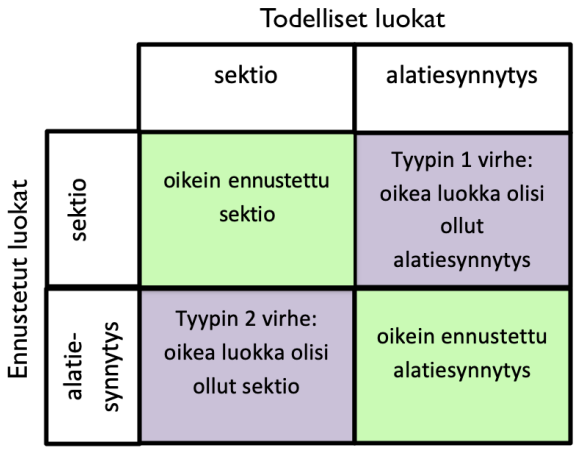
Tutkimuskysymyksiä lähestyttiin eri tavoin riippuen tarkastellusta vastemuuttujasta. Binääristä vastemuuttujaa tarkasteltaessa aineisto jaettiin aluksi opetus- ja testiaineistoon. Satunnainen jako tehtiin niin, että opetus- ja testiaineiston osuudet koko aineistosta olivat 70 % ja 30 %. Opetusaineistossa oli 305 havaintoa ja testiaineistossa oli 132. Opetusaineistosssa 256 (84 %) oli synnyttänyt alateitse ja 49 (16 %) sektiolla. Vastaavat osuudet testiaineistossa oli 107 (81 %) ja 25 (19 %). Testi- ja opetusaineisto muodostettiin, jotta voitiin arvioida opetusaineiston avulla muodostettujen osajoukkojen ennustuskykyä testiaineiston uusilla havainnoilla eli tarkastella tutkimuskysymystä (3). Samojen osajoukkojen avulla, jotka muodostettiin opetusaineiston avulla, tutkittiin myös kysymyksiä (1) ja (2).

Mallien hyvyyttä opetusaineistossa arvioitiin AUC-arvon avulla (*Area Under the ROC Curve*). AUC-arvo mittaa ns. ROC-käyrän alapuolisen alueen kokoa. ROC-käyrä on graafinen esitys binäärisen luokittelijan suorituskyvystä eri kynnysarvoilla. Tässä yhteydessä kynnysarvolla tarkoitetaan todennäköisyyskynnystä, jonka avulla binääriluokittelija luokittelee havainnot positiiviseen ja negatiiviseen luokkaan. Esimerkiksi jos asetettu todennäköisyyskynnys on 0.5 ja luokittelijan ennustama todennäköisyys, että havainto kuuluu positiiviseen luokkaan, on 0.6, havainto luokitellaan positiiviseen luokkaan ($0.6 > 0.5$). Jos todennäköisyyskynnykseksi olisikin asetettu 0.7, havainto luokiteltaisiin negatiiviseen luokkaan ($0.6 < 0.7$). AUC-arvo on suljetulla välillä $0.5 - 1$, jossa pienin arvo edustaa satunnaisen luokittelijan suorituskykyä ja suurin arvo vastaa täydellistä luokittelijaa eli luokittelijaa, joka ennustaa binäärisen vastemuuttujan arvon aina oikein. [33]

Tutkimuskysymyksen (3) kohdalla osajoukkojen ennustekykyä arvioitiin luokittelun tarkkuuden avulla, joka määritellään seuraavasti:

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}}, \tag{20}$$

jossa TP tarkoittaa oikein ennustettua sektiota, TN tarkoittaa oikein ennustettua alatiesynnytystä, FP tarkoitaa tyypin 1 virhettä ja FN tarkoittaa tyypin 2 virhettä. Tyypin 1 virhe tarkoittaa sitä, että ennusteen mukaan luokka oli sektio, vaikka todellisuudessa se oli alatiesynnytys. Tyypin 2 virhe tarkoittaa sitä, että ennusteen mukaan luokka oli alatiesynnytys, vaikka todellisuudessa se oli sektio. Kuvassa 15 on kuvattu tyypin 1 ja 2 virheitä ristiintaulukon avulla.



Kuva 15: *Binäärisen vasteen ennustamisen tarkkuuden tutkiminen ristiintaulukoinnin avulla.* Tyypin 1 virhe on määritelty niin, että ennusteen mukaan luokka oli sektio, vaikka todellisuudessa se olisi ollut alatiesynnytys. Tyypin 2 virhe on määritelty niin, että todellinen arvo olisi ollut sektio, vaikka ennuste luokitteli havainnon alatiesynnyttäneeksi.

Jatkuvan vastemuuttujan kohdalla tutkimuskysymyksiä (1) ja (2) tarkasteltiin koko aineiston avulla ($n = 55$). Mallien sopivuutta aineistoon tutkittiin selitysasteen (R^2) avulla. Mitä lähempänä selitysaste on arvoa yksi sen paremmin malli sopii aineistoon. Tutkimuskysymystä (3) tarkasteltaessa mallien ennustekykyä arvioitiin ristiinvalidoinnin avulla. Ristiinvalidoinnissa aineisto jaettiin viiteen yhtä suureen osaan k, jolloin yhdessä osassa oli 11 havaintoa. Jokaisella mahdollisella testi- ja opetusaineiston jaolla ennustettujen ja todellisten arvojen eroja mitattiin keskineliövirheen ($RMSE$) ja keskimääräisen absoluuttisen virheen (MAE) avulla. Lisäksi laskettiin Spearmanin korrelaatio, jonka avulla tutkittiin, miten vahvasti ennustetut ja todelliset arvot riippuivat toisistaan. Keskineliövirhe määritellään seuraavasti:

$$RMSE = \sqrt{\frac{1}{11}\sum_{i=1}^{11}(y_i - \hat{y}_i)^2},$$

jossa y_i on havainnon todellinen arvo ja $\hat{y}_i$ on ennustettu arvo. Vastaavilla merkinnöillä keskimääräinen absoluuttinen virhe on

$$MAE = \frac{1}{11}\sum_{i=1}^{11}|y_i - \hat{y}_i|.$$

Mikäli opetusaineisto ei löytänyt osajoukkoja, ennusteita ei voitu muodostaa, jolloin virheiden ja korrelaation suuruudeksi asetettiin puuttuva arvo NA.

Ennen osajoukkojen tunnistamista tehtiin muutamia valintoja. Ensinnäkin lukumääräaineistossa eli havaittujen mikrobien lukumäärissä esiintyi nollasoluja, joita CoDaCoRe-menetelmä ei salli. Tämän takia jokaiseen lukumääräaineiston soluun summattiin lukuarvo yksi (pseudoluku=1). Lisäksi siemenluvun arvoksi asetettiin nolla jokaiselle muodostetulle mallille. Siemenluvun avulla satunnaisuutta

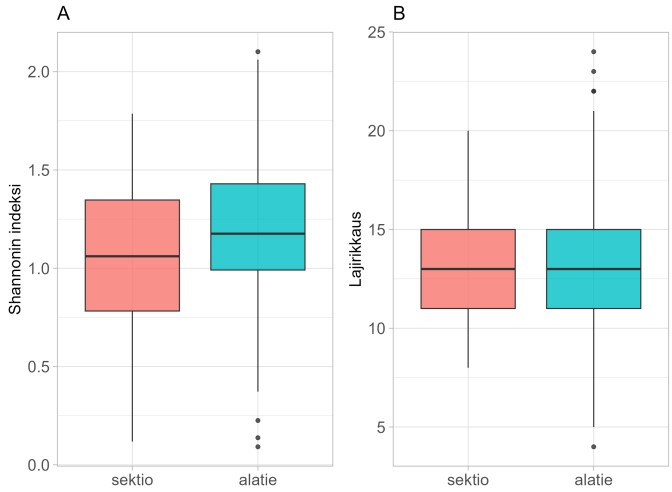
hyödyntävien funktioiden tulokset voidaan toistaa. CoDaCoRe-menetelmässä käytetään nopeimman laskeutumisen menetelmää, jonka löytämä ratkaisu on riippuvainen alkuarvosta (esimerkki 3.3). Siemenluvun valinnalla voi olla vaikutusta siihen, mistä alkuarvosta nopeimman laskeutumisen menetelmä lähtee liikkeelle. Lisäksi optimaalinen kynnysarvo $\hat{t}$ valitaan ristiinvalidoinnin avulla, jossa aineisto jaetaan osa-aineistoihin. Mikäli CoDaCoRe-menetelmä muodostaa osa-aineistot satunnaisesti, voisi olla mahdollista, että ristiinvalidoinnin mukaiset ennustevirheet vaihtelisivat siemenluvun mukaan.

Käytetty ohjelmisto. Analyysit tehtiin R-ohjelmistolla (versio 4.0.5) [34]. Aineistojen muokkaamiseen käytettiin hyödyksi kirjastoja `dplyr` [35], `tidyr` [36] ja `mia` [37]. Luvussa 2 hyödynnettiin kirjastoa `vegan` [16] alfa-monimuotoisuuden indeksien laskemiseen sekä kirjastoa `pheatmap` [17] lämpökartan tekoon. Soveltavassa osiossa käytettiin lisäksi kirjastoa `codacore` [38] CoDaCoRe-mallien sovittamiseen. Tulosten visualisointiin käytettiin kirjastoja `ggplot2` [39], `GGally` [40] ja `patchwork` [41].

5 CoDaCoRe-menetelmän soveltaminen aineistoon

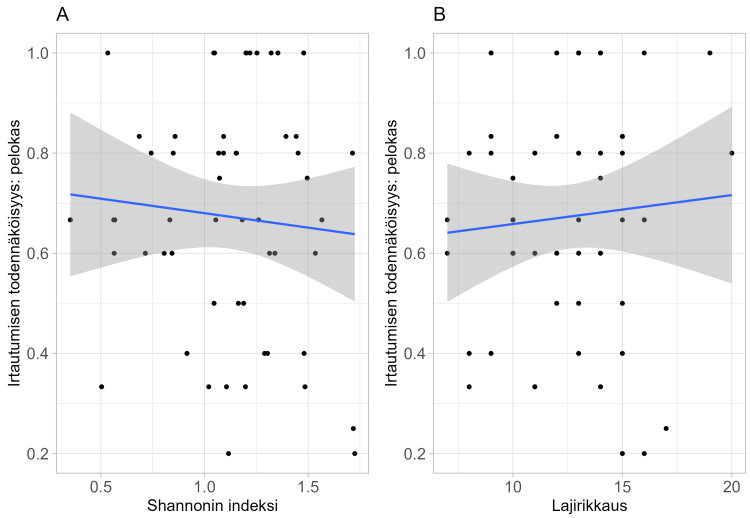
Tämän tutkielman empiirisessä osiossa sovellettiin CoDaCoRe-menetelmää luvussa 4 esiteltyyn oikeaan mikrobiomiaineistoon. Menetelmää käytettiin sekä binääriselle että jatkuvalle vastemuuttujalle. Binäärisenä vasteena oli synnytystapa (alatie vs. sektio). Jatkuvana vasteena oli muuttuja, joka kuvaa todennäköisyyttä, jolla lapsi irrottaa katseensa pelokkaasta kasvokuvasta.

Kuvaillaan aluksi mikrobiomiaineiston ja vasteiden välistä yhteyttä luvussa 2.3 esiteltyjen alfa-monimuotoisuuden indeksien avulla. Shannonin indeksin mukaan alatiesynnyttäneillä oli hieman korkeampi alfa-monimuotoisuus verrattuna sektiollla synnyttäneisiin (kuva 16). Lajirikkaudessa R_{obs} vastaava eroa ei ollut.



Kuva 16: *Alfa-monimuotoisuus on esitetty laatikkojanakuvaajana sekä sektio- että alatiesynnyttäneille.* Kuvassa A on Shannonin indeksi ja kuvassa B lajirikkaus R_{obs}. Shannonin indeksin mediaani oli korkeammalla alatiesynnyttäneillä (1.18) verrattuna sektiolla synnyttäneisiin (1.06). Lajirikkauden R_{obs} mukaan mediaanit olivat samat molemmilla synnytystavoilla; alatie (13) ja sektio (13). Alatiesynnyttäneillä oli muutamia poikkeuksellisen matalia ja korkeita arvoja sekä Shannonin indeksissä että lajirikkaudessa.

Shannonin indeksin mukaan alfa-monimuotoisuus oli alhainen, kun katseen irtautumisen todennäköisyys oli suuri (kuva 17:A). Toisaalta lajirikkaus R_{obs} kasvoi, kun irtautumisen todennäköisyys kasvoi (17:B).



Kuva 17: *Lapsen katseen irtautumisen todennäköisyyden ja alfa-monimuotoisuuden välinen yhteys hajontakuvioiden avulla.* Kuvassa A on esitetty Shannonin indeksi ja kuvassa B lajirikkaus R_{obs}. Sekä A- että B-kuvaan on sovitettu lineaarinen suora sekä sen pisteittäinen 95 %:n luottamusväli. Kuvassa A lineaarinen suora on laskeva ja kuvassa B nouseva.

5.1 Binäärinen vastemuuttuja

Tarkastellaan ensin synnytystapa-muuttujaa ja luvussa 4.2 esitettyjä tutkimuskysymyksiä (1) ja (2) eli miten balanssien ja yhdelmien mukaiset osajoukot erosivat ja miten parametri λ vaikutti löydettyihin osajoukkoihin. Taulukossa 7 on esitetty neljän eri CoDaCoRe-mallin löytämät balanssit. Kuhunkin malliin asetettiin oma parametrin λ arvo. Riippuen parametrin λ arvosta, mallit löysivät eri määrän balansseja (ks. sivu 26). Mallin 3 ja 4 ero on se, että mallissa 4 oli asetettu komento overlap = *false (F)*. Jos malli löytää useamman balanssin, tämä komento estää yhtä sukua esiintymästä useissa balansseissa. Eli esimerkiksi ne suvut, jotka ovat ensimmäisessä balanssissa, eivät voi olla jälkimmäisissä. Taulukossa 8 on esitetty yhdelmien mukaiset mallit (5 – 8), joissa parametrin λ arvoa muutettiin vastaavasti kuin taulukossa 7. Parametrin λ suurilla arvoilla mallit löysivät vain yhden yhdelmän.

Balanssit ja yhdelmät muodostuivat erilaisista osajoukoista verrattuna toisiinsa. Balansseihin perustuvissa osajoukoissa oli enemmän sukuja kuin yhdelmiin perustuvissa (taulukot 7, 8). Tämän tutkielman analyysit tukevat luvussa 2.4 esiteltyä väitettä, että balanssit saattaisivat löytää enemmän harvinaisia sukuja verrattuna yhdelmiin. Yhdelmien mukaisissa osajoukoissa esiintyi hyvin yleisiä sukuja, esimerkiksi Bacteroides- ja Veillonella-suvut. Balanssien mukaisissa osajoukoissa oli lisäksi harvinaisempia sukuja, esimerkiksi Sutterella-suku.

Parametrin λ arvo vaikutti sekä balanssien että yhdelmien mukaisiin osajoukkoihin. Parametrin λ alhaisella arvolla (0.1) osajoukoissa oli enemmän sukuja mukana verratuna parametrin arvoon 1. Sekä balanssien että yhdelmien korkein AUC-arvo oli parametrin λ arvolla 0.1 (balanssi: 0.81, yhdelmä: 0.74). Balanssit sopivat opetusaineistoon hieman paremmin verrattuna yhdelmiin.

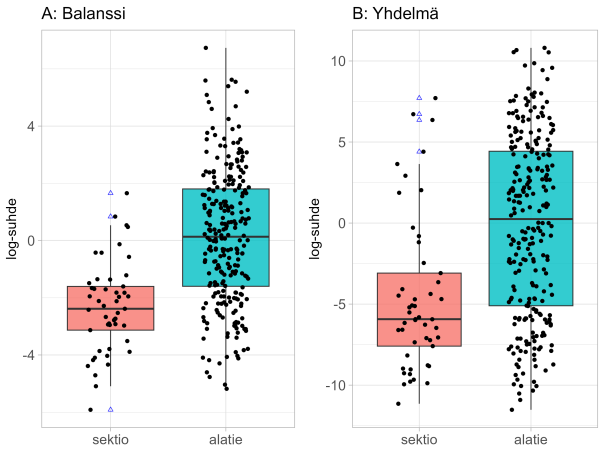
Taulukko 7: *CoDaCoRe-menetelmän mukaiset parhaimmat osajoukot $\hat{J}^+$ ja $\hat{J}^-$ balanssien mukaan binääriselle vasteelle (synnytystapa) parametrin λ eri arvoilla.* Jokaisen mallin balanssit on esitetty hierarkkisesti siten, että parhain balanssi on ensimmäisenä. Overlap-komennon avulla voidaan sallia saman suvun esiintyminen useissa löydetyissä balansseissa (overlap=T) tai estää tämä (overlap=F). Tunnistamattomat suvut on lyhennetty kirjaimilla UG.

Malli 1 $\lambda = 1$ overlap = T	**1. balanssi** $\hat{J^+}$ = {Bacteroides, Bittarella, Barnesiella, Parabacteroides, Sutterella} $\hat{J^-}$ = {Clostridiaceae:Clostridium, Veillonella, Enterobacteriaceae:UG, Lachnospiraceae:Ruminococcus, Citrobacter, Lachnospiraceae: Clostridium, Hungatella, Haemophilus, Clostridioides}
Malli 2 $\lambda = 0.6$ overlap = T	**1. balanssi** $\hat{J^+}$ = {Bacteroides, Bittarella, Barnesiella, Parabacteroides, Sutterella, Intestinibacter} $\hat{J^-}$ = {Clostridiaceae:Clostridium, Veillonella, Enterobacteriaceae:UG, Lachnospiraceae:Ruminococcus, Citrobacter, Enterococcaceae:UG, Lachnospiraceae: Clostridium, Hungatella, Haemophilus, Clostridioides, Cellulosilyticum}
Malli 3 $\lambda = 0.1$ overlap = T	**1. balanssi** $\hat{J^+}$ = {Bacteroides, Bittarella, Barnesiella, Parabacteroides, Sutterella, Acidaminococcus, Intestinibacter} $\hat{J^-}$ = {Clostridiaceae:Clostridium, Veillonella, Enterobacteriaceae:UG, Lachnospiraceae:Ruminococcus, Citrobacter, Enterococcaceae:UG, Lachnospiraceae:Clostridium, Hungatella, Haemophilus, Clostridioides, Cellulosilyticum} **2. balanssi** $\hat{J^+}$ = {Bacteroides} $\hat{J^-}$ = {Clostridioides} **3. balanssi** $\hat{J^+}$ = {Bacteroides, Bittarella, Barnesiella, Sutterella} $\hat{J^-}$ = {Clostridiaceae:Clostridium, Veillonella, Enterobacteriaceae: UG, Lachnospiraceae:Ruminococcus, Citrobacter, Lachnospiraceae: Clostridium, Haemophilus, Clostridioides}
Malli 4 $\lambda = 0.1$ overlap = F	**1. balanssi** $\hat{J^+}$ = {Bacteroides, Bittarella, Barnesiella, Parabacteroides, Sutterella, Acidaminococcus, Intestinibacter} $\hat{J^-}$ = {Clostridiaceae:Clostridium, Veillonella, Enterobacteriaceae:UG, Lachnospiraceae:Ruminococcus, Citrobacter, Enterococcaceae:UG, Lachnospiraceae:Clostridium, Hungatella, Haemophilus, Clostridioides, Cellulosilyticum} **2. balanssi** $\hat{J^+}$ = {Escherichia, Collinsella} $\hat{J^-}$ = {Flavonifractor}

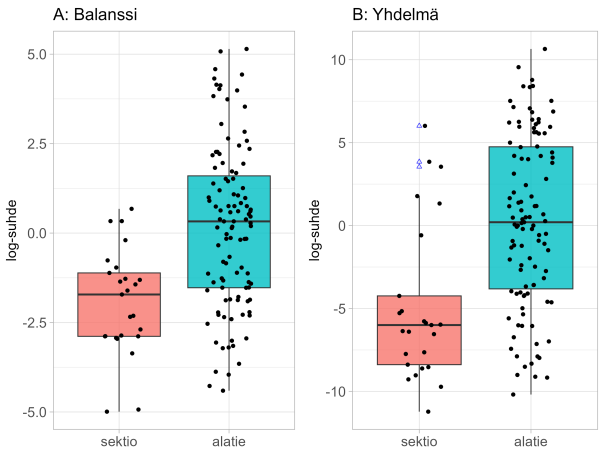
Taulukko 8: *CoDaCoRe-menetelmän mukaiset parhaimmat osajoukot $\hat{J}^+$ ja $\hat{J}^-$ yhdelmien mukaan binääriselle vasteelle (synnytystapa) parametrin λ eri arvoilla.* Jokaisen mallin yhdelmät on esitetty hierarkkisesti siten, että parhain yhdelmä on ensimmäisenä. Overlap-komennon avulla voidaan sallia saman suvun esiintyminen useissa löydetyissä yhdelmissä (overlap=T) tai estää tämä (overlap=F). Tunnistamaton suku on lyhennetty kirjaimilla UG.

Malli 5 $\lambda = 1$ overlap = T	**1. yhdelmä** $\hat{J}^+$ = {Bacteroides} $\hat{J}^-$ = {Veillonella}
Malli 6 $\lambda = 0.6$ overlap = T	**1. yhdelmä** $\hat{J}^+$ = {Bacteroides} $\hat{J}^-$ = {Veillonella}
Malli 7 $\lambda = 0.1$ overlap = T	**1. yhdelmä** $\hat{J}^+$ = {Bacteroides} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella} **2. yhdelmä** $\hat{J}^+$ = {Bacteroides} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella} **3. yhdelmä** $\hat{J}^+$ = {Bacteroides} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella} **4. yhdelmä** $\hat{J}^+$ = {Escherichia, Bacteroides, Bittarella} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella, Lachnospiraceae: Ruminococcus}
Malli 8 $\lambda = 0.1$ overlap = F	**1. yhdelmä** $\hat{J}^+$ = {Bacteroides} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella} **2. yhdelmä** $\hat{J}^+$ = {Escherichia, Bittarella} $\hat{J}^-$ = {Lachnospiraceae:Ruminococcus} **3. yhdelmä** $\hat{J}^+$ = {Collinsella, Barnesiella, Parabacteroides, Sutterella, Prevotella Acidaminococcus} $\hat{J}^-$ = {Bifidobacterium, Enterobacteriaceae:UG, Citrobacter, Enterococcaceae:UG, Lachnospiraceae:Clostridium, Hungatella, Lactobacillus, Haemophilus, Flavonifractor, Cellulosilyticum, Clostridioides} **4. yhdelmä** $\hat{J}^+$ = {Staphylococcus, Blautia, Enterococcus, Actinomyces, Alistipes, Intestinibacter} $\hat{J}^-$ = {Erysipelatoclostridium, Erysipelotrichaceae:Eubacterium}

R-ohjelmistossa CoDaCoRe-funktion parametrien oletusarvoina on $\lambda = 1$ sekä overlap = *true*. Näillä arvoilla löydettyjen osajoukkojen log-suhteet on esitetty kuvassa 18. Log-suhde on korkeammalla alatiesynnyttäneillä kuin sektiolla synnyttäneillä sekä balanssien että yhdelmien mukaan. Sama ilmiö toistuu testiaineistossa (kuva 19). Balanssien jakaumat ovat kapeampia verrattuna yhdelmien jakaumiin (kuvat 18, 19).



Kuva 18: *Opetusaineiston avulla piirretyt viivajanakuvaajat synnytystavoille.* Kuvassa A on esitetty y-akselilla balanssin mukainen log-suhde, kun on käytetty mallia 1 ($\lambda = 1$, taulukko 7). Kuvassa B taas on esitetty y-akselilla yhdelmän mukainen log-suhde, kun on käytetty mallia 5 ($\lambda = 1$, taulukko 8). Log-suhteet on piirretty viivajanakuvaajan avulla molemmille synnytystavoille. Kuvaajiin on piirretty havaintopisteet näkyviin. Kuvasssa A sektio-luokan mediaani on -2.4 ja alatiesynnyttäneiden mediaani on 0.1. Vastaavat arvot B-kuvassa on -5.9 ja 0.2. Balanssin jakauma on välillä $[-5.9, 6.7]$ kun taas yhdelmän jakauma on välillä $[-11.5, 10.8]$. Poikkeavien havaintojen viereen on piirretty sininen kolmio.



Kuva 19: *Testiaineiston avulla piirretyt viivajanakuvaajat synnytystavoille.* Kuvassa A on esitetty y-akselilla balanssin mukainen log-suhde, kun on käytetty mallia 1 ($\lambda = 1$, taulukko 7). Kuvassa B on esitetty yhdelmän mukainen log-suhde, kun on käytetty mallia 5 ($\lambda = 1$, taulukko 8). Alatiesynnyttäneiden mediaani on korkeammalla sekä balanssin että yhdelmän mukaan verrattuna sektion avulla synnyttäineisiin.

Seuraavaksi vastemuuttujalle (synnytystapa) sovitettiin logistinen regressiomalli, jossa selittävänä muuttujana oli mallin 1 (taulukko 7) mukainen balanssi tai vaihtoehtoisesti mallin 5 (taulukko 8) mukainen yhdelmä. Tarkastelu tehtiin sekä opetus- ja testiaineistolle. Opetusaineiston mukaan balanssin log-suhteen yhden yksikön nousu pienensi sektiosynnytyksen log-vastasuhdetta 43 prosenttia ($\beta = -0.57$, CI$=-0.77-\ -0.39$). Vastaava ilmiö oli nähtävissä myös testiaineistossa (taulukko 9). Yhdelmien mukaan tulos oli vastaava, mutta log-vetosuhteen kerroin β oli matalampi.

Taulukko 9: *Logistisen regression mukaiset estimaatit.* Sekä mallin 1 (taulukko 7) mukaiselle balanssilla että mallin 5 (taulukko 8) mukaiselle yhdelmällä sovitettiin logistinen regressio vastemuuttujalle synnytystapa (referenssiluokkana alatiesynnyttäneet). Regressiomalleista on esitetty estimaatit log-vetosuhteille β ja niiden 95 %:n luottamusvälit (CI) sekä opetus- että testiaineistolla laskettuna. Log-vetosuhteen avulla voidaan kuvata balanssin tai yhdelmän log-suhteen kasvun vaikutusta synnytystapaan (sektio vs. alatiesynnyttäneet).

		opetusaineisto		testiaineisto	
malli	muuttuja	β	95 % CI	β	95 % CI
1	vakio	-2.33	-2.86 – -1.88	-1.97	$-2.67-\ -1.39$
	balanssi	-0.57	-0.77 – -0.39	-0.53	-0.83 – -0.28
5	vakio	-2.05	-2.50 – -1.67	-1.98	-2.70 – -1.41
	yhdelmä	-0.16	-0.23 – -0.10	-0.21	-0.33 – -0.11

Uusien havaintojen ennustaminen. Opetusaineiston avulla muodostettujen osajoukkojen kykyä ennustaa uusia havaintoja arvioitiin testiaineiston avulla. CoDaCoRe-funktiossa voi valita, kuinka monella löydetyllä balanssilla tai yhdelmällä ennustetaan opetusaineistoa. Tässä tutkielmassa valittiin niin, että ennustettiin aina vain parhaimmalla balanssilla tai yhdelmällä. Ennustekyvyn hyvyys määritettiin tyypin 1 ja 2 virheiden sekä luokittelun tarkkuuden avulla, jotka määriteltiin tarkemmin luvussa 4.2.

Taulukossa 10 on esitetty luokittelun tarkkuus sekä tyypin 1 ja 2 virheet malleilla 1 – 8 (taulukot 7, 8). Lisäksi taulukossa on esitetty opetuaineiston AUC-arvo. Balansseilla on hieman korkeammat opetusaineiston AUC-arvot verrattuna yhdelmiin. Balanssit sopivat siis opetusaineistoon hieman paremmin. Yhdelmät taas ennustivat vastemuuttujaa hieman paremmin verrattuna balansseihin testiaineiston perusteella. Parametrin λ matalalla arvolla ei saavutettu parempia luokittelun tarkkuuksia, mutta opetusaineiston AUC-arvot olivat korkeammat. Tulos on oletettava, koska mitä enemmän osajoukoissa on lajeja, sitä paremmin ne sopivat aineistoon, mutta uusien havaintojen ennustaminen kärsii jossain vaiheessa ylisovittamisesta.

Taulukko 10: *Mallien sopivuus opetusaineistoon sekä ennustamisen tarkkuus testiaineistolla.* Taulukkoon on kerätty mallien 1 – 8 parametrit, mallissa käytetty log-suhde (B=balanssi, Y=yhdelmä), opetusaineiston AUC-arvo sekä testiaineiston avulla lasketut tyypin 1 ja 2 virheet (kuva 15) ja luokittelun tarkkuus (kaava (20)). Tyypin 1 virhe määriteltiin niin, että ennusteen mukaan synnytystapa oli sektio, vaikka todellisuudessa se oli alatiesynnytys. Tyypin 2 virhe määriteltiin niin, että havainnon todellinen arvo olisi ollut sektiosynnytys. Tyypin 1 ja 2 virheet on ilmoitettu lukumäärinä. Yhteensä testiaineistossa oli 132 havaintoa, joista 25 oli sektion avulla synnyttäneitä. Ennusteet on muodostettu käyttäen mallikohtaisesti parasta balanssia tai yhdelmää.

malli	log-suhde	λ	overlap	opetusaineiston AUC	tyypin 1 virhe	tyypin 2 virhe	luokittelun tarkkuus
1	B	1	T	0.802	22	5	0.796
2	B	0.6	T	0.810	22	8	0.773
3	B	0.1	T	0.812	21	7	0.788
4	B	0.1	F	0.812	21	7	0.788
5	Y	1	T	0.724	23	1	0.818
6	Y	0.6	T	0.724	23	1	0.818
7	Y	0.1	T	0.740	23	4	0.796
8	Y	0.1	F	0.740	23	4	0.796

5.1.1 Sensitiivisyysanalyysi

Sensitiivisyysanalyysinä tutkittiin, muuttuvatko löydetyt osajoukot, kun vaihdettaan asetettua siemenlukua. Tarkastelut tehtiin sekä balansseille että yhdelmille opetusaineiston avulla, kun parametrin λ arvo oli 1.

Yhdelmien mukaiset osajoukot olivat aina samat kaikilla testatuilla siemenluvuilla (taulukko 11). Balanssien mukaiset osajoukot taas vaihtelivat siemenluvun mukaan. Suurin joukko, josta balanssit muodostuivat, sisälsi 14 sukua ja pienin joukko muodostui vain neljästä suvusta. Mallien 1 ja $1^{(2912)}$ osajoukoissa ei ollut suurta eroa. Mallissa $1^{(2912)}$ ei esiintynyt Sutterella-sukua, joka oli mallissa 1. Sutterella-suvun pois jääminen ei juurikaan vaikuttanut viivajanakuvaajan muotoon (vrt. kuvat 18:A ja B1:A). Siemenluvun 140 mukaisissa kuvaajissa balanssien jakauma oli leveämmällä välillä verrattuna siemenluvun 2912 kuvaajiin (kuva B1). Kuitenkin kaikkien tutkittujen mallien mukaan alatiesynnyttäneiden mediaani oli korkeammalla kuin sektion avulla synnyttäneillä (kuvat 18, B1).

Sensitiivisyysanalyysien mukaisilla osajoukoilla arvioitiin myös ennustekyvyn hyvyyttä testiaineiston avulla. Tulokset on taulukoitu liitteen B taulukossa B1. Luokittelun tarkkuudet olivat lähes samat kaikilla kokeilluilla siemenluvuilla. Balanssien huonoin ennustuskyky oli mallilla $1^{(140)}$, jolloin luokittelun tarkkuus oli 0.78.

Taulukko 11: *Sensitiivisyysanalyysin tulokset.* Taulukossa on vertailtu, mitkä suvut esiintyivät malleissa 1, $1^{(2912)}$, $1^{(140)}$, 5, $5^{(2912)}$ ja $5^{(140)}$. Mallit 1, $1^{(2912)}$ ja $1^{(140)}$ on muodostettu balanssien avulla ja malleissa 5, $5^{(2912)}$ ja $5^{(140)}$ on käytetty yhdelmiä. Jokaisessa mallissa parametrin λ arvo oli yksi. Siemenluku on esitetty mallin yläindeksissä. Malli $1^{(140)}$ löysi kaksi suhdetta, joista paras on esitetty tässä. Bacteroides-suku esiintyi jokaisen mallin osajoukossa $\hat{J}^+$. Yhdelmien mukaiset mallit löysivät aina samat osajoukot.

	malli:					
$\hat{J}^+$	1	$1^{(2912)}$	$1^{(140)}$	5	$5^{(2912)}$	$5^{(140)}$
Bacteroides	x	x	x	x	x	x
Barnesiella	x	x				
Bittarella	x	x				
Parabacteroides	x	x				
Sutterella	x					

$\hat{J}^-$	1	$1^{(2912)}$	$1^{(140)}$	5	$5^{(2912)}$	$5^{(140)}$
Citrobacter	x	x				
Clostridiaceae:Clostridium	x	x				
Clostridioides	x	x	x			
Enterobacteriaceae:UG	x	x	x			
Haemophilus	x	x				
Hungatella	x	x				
Lachnospiraceae:Clostridium	x	x				
Lachnospiraceae:Ruminococcus	x	x	x			
Veillonella	x	x		x	x	x

5.2 Jatkuva vastemuuttuja

Sovelletaan seuraavaksi CoDaCoRe-menetelmää jatkuvaan vastemuuttujaan. Tarkastellaan ensin tutkimuskysymyksiä (1) ja (2) eli miten balanssien ja yhdelmien mukaiset osajoukot erosivat ja miten parametri λ vaikuttu niihin. Liitteen B taulukossa B2 on esitetty mallit A – D, joissa kaikissa käytettiin balansseja. Jokaiselle mallille asetettiin oma parametrin λ arvo (1, 0.6 tai 0.1). Kaikki mallit löysivät ainoastaan yhden balanssin. Mallien mukaiset osajoukot muodostuivat lähes samoista suvuista parametrin λ vaihtelusta huolimatta. Parametrin λ arvolla 1 (malli A) osajoukkojen suvut olivat

$$\hat{J}^+ = \{\text{Bifidobacterium, Lachnospiraceae:Clostridium, Haemophilus, Eggerthella}\}$$
$$\text{ja } \hat{J}^- = \{\text{Clostridiaceae:Clostridium, Bittarella, Sutterella}\}.$$

Mallien sopivuutta aineistoon arvioitiin selitysasteen (R^2) avulla. Suurin selitysaste oli parametrin λ ollessa 0.1 ($R^2 = 0.39$) ja pienin parametrin λ ollessa 1 ($R^2 = 0.32$).

Liitteen B taulukossa B3 on esitetty mallit E – H, joissa kaikissa käytettiin yhdelmiä. Riippuen parametrin λ arvosta mallit löysivät nollasta kolmeen yhdelmää. Malli E ($\lambda = 1$) ei löytänyt yhtäkään yhdelmää. Parametrin λ arvolla 0.6 (malli F) osajoukkojen suvut olivat

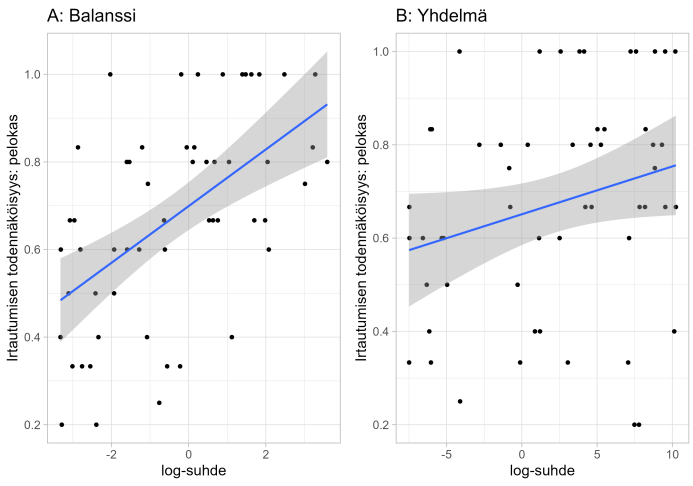
$$\hat{J}^{+} = \{\text{Bifidobacterium, Bacteroides}\} \text{ ja } \hat{J}^{-} = \{\text{Clostridiaceae:Clostridium}\}.$$

Myös yhdelmien mukaan parhain selitysaste oli parametrin λ arvolla 0.1 ($R^2 = 0.16$). Balanssien paras selitysaste oli kuitenkin korkeampi verrattuna yhdelmiin

Sekä mallin A (taulukko B2) että mallin F (taulukko B3) mukaisille log-suhteille muodostettiin lineaarinen malli, jonka vasteena oli todennäköisyysmuuttuja (taulukko 12). Balanssin yhden yksikön kasvu nosti todennäköisyysmuuttujan arvoa keskimäärin 0.06. Yhdelmien mukaan kasvu oli 0.01, mutta estimaatin luottamusväliin sisältyi nolla. Sekä balanssin että yhdelmän mukaan siis osajoukkojen lajien lukumäärien suhteen kasvu nosti katseen irtautumisen todennäköisyyttä (kuva 20)

Taulukko 12: *Lineaarisen regression mukaiset estimaatit.* Sekä mallin A (taulukko B2) mukaiselle balanssilla että mallin F (taulukko B3) mukaiselle yhdelmällä sovitettiin lineaarinen regressio todennäköisyysmuuttujan ollessa vasteena. Regressiomalleista on esitetty estimaatit β ja niiden 95%:n luottamusvälit (CI). Estimaatti β kertoo, kuinka paljon balanssin tai yhdelmän yhden yksikön muutos vaikuttaa todennäköisyysmuuttujan arvoon.

malli	muuttuja	β	95 % CI
A	vakio	0.70	0.64 – 0.75
	balanssi	0.06	0.04 – 0.09
F	vakio	0.65	0.58 – 0.72
	yhdelmä	0.01	0.00 – 0.02



Kuva 20: *Löydettyjen osajoukkojen mukaiset hajontakuviot.* Kuvassa A on piirretty mallia A ($\lambda = 1$) vastaava hajontakuvio. Kuvassa B on piirretty mallia F ($\lambda = 0.6$) vastaava hajontakuvio. Kuvien y-akselilla on vasteena käytetty todennäköisyysmuuttuja ja x-askelilla balanssi tai yhdelmä. Kuviin on sovitettu lineaarinen suora sekä sen pisteittäinen 95 %:n luottamusväli. Kun irtautumisen todennäköisyys kasvaa, myös log-suhde kasvaa.

Uusien havaintojen ennustaminen. Mallien ennustekykyä tutkittiin sekä balansseilla että yhdelmillä ristiinvalidoinnin avulla. Mallien ennustekyvyn hyvyyttä arvioitiin keskineliövirheen (*RMSE*), keskimääräisen absoluuttisen virheen (*MAE*) ja Spearmanin korrelaation avulla. Korrelaatiokertoimen avulla tutkittiin sitä, miten vahvasti ennustetut ja todelliset arvot riippuivat toisistaan.

Taulukoon 13 on koottu ristiinvalidoinnin tulokset kolmella eri parametrin λ arvoilla (1, 0.6, 0.1). Kun parametrin λ arvo oli 1, yhdelmät eivät löytäneet osajoukkoja. Balanssien parhain ennustuskyky oli parametrin λ arvolla 1. Yhdelmien paras ennustekyky oli silloin, kun parametrin λ arvo oli 0.6.

Taulukko 13: *Ristiinvalidoinnin tulokset balanssien ja yhdelmien mukaan.* Taulukkoon on listattu aineiston osa, k, joka on toiminut testiaineistona kullakin ristiinvalidoinnin kierroksella ja CoDaCoRe-mallissa käytetty parametrin λ arvo. Ristiinvalidoinnin tulosta arvioitiin keskineliövirheen (*RMSE*), keskimääräisen absoluuttisen virheen (*MAE*) ja korrelaation avulla. Sekä molemmista virheistä että korrelaatiosta on laskettu keskiarvo mallikohtaisesti. Jos testiaineisto ei löytänyt osajoukkoja, ennustettuja arvoja ei voitu laskea. Nämä tapaukset on merkitty taulukkoon kirjaimin *NA*.

			balanssi			**yhdelmä**		
	k	λ	RMSE	MAE	korrelaatio	RMSE	MAE	korrelaatio
1. malli	1	1	0.292	0.261	0.014	*NA*	*NA*	*NA*
	2	1	*NA*	*NA*	*NA*	*NA*	*NA*	*NA*
	3	1	0.223	0.196	0.232	*NA*	*NA*	*NA*
	4	1	0.216	0.176	0.382	*NA*	*NA*	*NA*
	5	1	0.238	0.176	0.192	*NA*	*NA*	*NA*
keskiarvo			**0.242**	**0.202**	**0.205**	***NA***	***NA***	***NA***
2. malli	1	0.6	0.291	0.254	0.014	0.263	0.235	0.148
	2	0.6	0.341	0.283	0.123	0.329	0.260	0.260
	3	0.6	0.239	0.211	-0.009	*NA*	*NA*	*NA*
	4	0.6	0.216	0.176	0.382	0.177	0.147	0.520
	5	0.6	0.222	0.163	0.164	0.182	0.138	0.075
keskiarvo			**0.262**	**0.218**	**0.135**	**0.238**	**0.195**	**0.251**
3. malli	1	0.1	0.295	0.254	-0.042	0.254	0.231	0.231
	2	0.1	0.342	0.276	-0.151	0.350	0.276	-0.137
	3	0.1	0.239	0.211	-0.009	*NA*	*NA*	*NA*
	4	0.1	0.203	0.162	0.262	0.184	0.152	0.460
	5	0.1	0.243	0.195	0.005	0.181	0.140	0.187
keskiarvo			**0.265**	**0.220**	**0.013**	**0.242**	**0.200**	**0.185**

5.2.1 Sensitiivisyysanalyysi

Myös jatkuvan vastemuuttujan tapauksessa sensitiivisyysanalyysinä katsottiin, miten löydetyt osajoukot muuttuivat, kun siemenlukua muutettiin. Koska yhdelmät eivät löytäneet osajoukkoja, kun parametrin λ arvo oli 1, sensitiivisyysanalyyseissa tutkittiin parametrin λ arvoa 0.6. Kaikilla kolmella siemenluvulla (0, 2912, 1414) balanssien löytämät osajoukot olivat samoja (taulukko 14). Osajoukoissa oli yhteensä kahdeksan eri sukua. Yhdelmien kohdalla osajoukoissa oli hieman vaihtelua. Pienin määrä sukuja, joista yhdelmät muodostuivat, oli kaksi ja suurimmillaan seitsemän.

Taulukko 14: *Sensitiivisyysanalyysin tulokset.* Taulukossa on vertailtu, mitkä suvut esiintyivät malleissa B, $B^{(2912)}$, $B^{(1414)}$, F, $F^{(2912)}$ ja $F^{(1414)}$. Mallit B, $B^{(2912)}$ ja $B^{(140)}$ on muodostettu balanssien kanssa. Malleissa F, $F^{(2912)}$ ja $F^{(1414)}$ on käytetty yhdelmiä. Mallien yläindeksi kertoo siemenluvun. Jokaisessa mallissa parametrin λ arvo on 0.6. Jokainen malli löysi vain yhden logaritmisen suhteen. Sekä Bifidobacterium- että Clostridiaccae:Clostridium-suvut esiintyivät jokaisessa mallissa. Balanssien mukaiset mallit muodostuivat aina samoista osajoukoista.

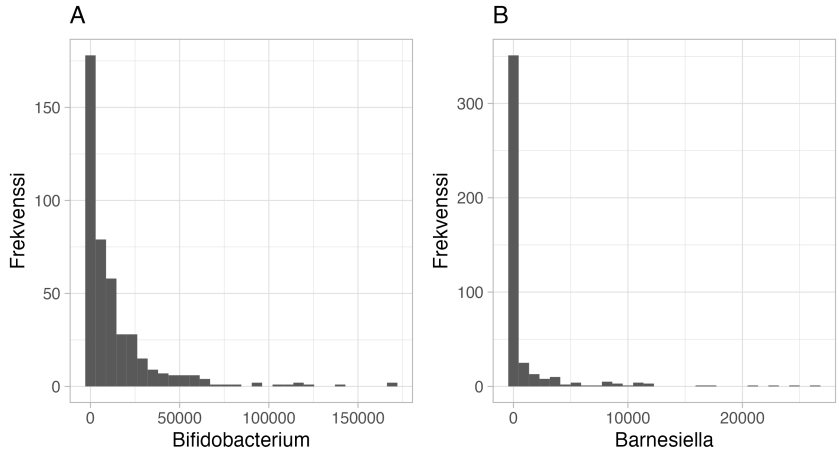
	malli:					
$\hat{J}^+$	B	$B^{(2912)}$	$B^{(1414)}$	F	$F^{(2912)}$	$F^{(1414)}$
Bacteroides				x		x
Bifidobacterium	x	x	x	x	x	x
Eggerthella	x	x	x			
Haemophilus	x	x	x			
Lachnospiraceae:Clostridium	x	x	x			
Lachnospiraceae:Ruminococcus						x

$\hat{J}^-$	B	$B^{(2912)}$	$B^{(1414)}$	F	$F^{(2912)}$	$F^{(1414)}$
Barnesiella						x
Bittarella	x	x	x			
Clostridiaceae:Clostridium	x	x	x	x	x	x
Enterococcaceae:UG						x
Erysipelatoclostridium	x	x	x			
Sutterella	x	x	x			
Veillonella						x

Taulukkoon 15 on koottu ne suvut, jotka esiintyivät binäärisen tai jatkuvan vastemuuttujan sensitiivisyysanalyyseissä (taulukot 11, 14). Löydetyissä osajoukoissa esiintyi sekä aineiston yleisimpiä sukuja että harvinaisempia. Kuvassa 21 on esitetty esimerkki yleisen suvun (Bifidobacterium) runsauden jakaumasta sekä harvaisemman suvun (Barnesiella) runsauden jakaumasta. Molempien lajien havaittujen lajirunsauksien jakauma oli vino, mikä on tyypillistä.

Taulukko 15: *Binäärisen ja jatkuvan vastemuuttujan sensitiivisyysanalyyseissä esiintyvät suvut.* Niiden sukujen nimet on tummennettu, jotka kuuluvat aineiston kymmenen yleisimmän suvun joukkoon. Taulukossa on ilmoitettu niiden solujen lukumäärä, joissa havaittu lajirunsautena on 1 (lisätty pseudoluku oli 1). Havaittujen lajirunsauksien mediaani ja kvartaaliväli (Q1, Q3) on ilmoitettu. Arvot on laskettu koko aineistosta ($n = 437$).

suku	solujen lkm, joissa arvo on 1	mediaani	(Q1, Q3)
Bacteroides	54	328	(10, 19014)
Barnesiella	278	1	(1, 41)
Bittarella	264	1	(1, 43)
Bifidobacterium	16	6080	(40, 17319)
Citrobacter	376	1	(1, 1)
Clostridiaceae:Clostridium	64	302	(13, 8628)
Clostridioides	394	1	(1, 1)
Eggerthella	302	1	(1, 15)
Enterobacteriaceae:UG	190	70	(1, 2574)
Erysipelatoclostridium	374	1	(1, 1)
Enterococcaceae:UG	194	12	(1, 269)
Haemophilus	240	1	(1, 48)
Hungatella	391	1	(1, 1)
Lachnospiraceae:Clostridium	366	1	(1, 1)
Lachnospiraceae:Ruminococcus	271	1	(1, 16)
Parabacteroides	385	1	(1, 1)
Sutterella	369	1	(1, 1)
Veillonella	19	1459	(68, 10911)



Kuva 21: *Kahden suvun havaittujen runsauksien jakauma on esitetty histogrammien avulla.* Kuvasssa A on esitetty Bifidobacterium-suvun havaitun lajirunsauden jakauma. Bifidobacterium-suku oli eniten edustettuna aineistossa. Kuvassa B on esitetty Barnesiella-suvun havaitun lajirunsauden jakauma. Barnesiella-suku oli harvinaisempi, mutta kuitenkin kahdenkymmenen yleisimmän suvun joukossa.

6 Yhteenveto ja pohdinta

Tämän tutkielman tavoitteena oli tutustua CoDaCoRe-menetelmään sekä teoreettisesti että empiirisen esimerkin avulla. CoDaCoRe-menetelmä on uusi ja tehokas menetelmä, jolla voidaan analysoida mikrobiomiaineistoja [3]. Aineistot ovat lukumääräaineistoja, joissa on ilmoitettu näytteissä havaittujen lajien runsaus. Mikrobiaineistojen tyypillisiä haasteita ovat suuri dimensio, nollasolut sekä lajien lukumäärien jakaumien vinous. Lisäksi aineistot voidaan tulkita kompositionaalisiksi. Aineistojen analysointiin tarkoitetut menetelmät kehittyvät erittäin nopeasti, mikä vaikeuttaa parhaimman menetelmän valitsemista.

Tutkielman alussa esiteltiin hyvin tyypillisiä analysointitapoja mikrobiomiaineistoille, alfa- ja beeta-monimuotoisuusindeksejä sekä DA-menetelmien ajatusta. Monimuotoisuusindeksien avulla voidaan kuvata mikrobiomiaineistojen koostumusta. Alfa-monimuotoisuus kuvaa yhden näytteen monimuotoisuutta kun taas beeta-monimuotoisuus kuvaa kahden näytteen välistä eroa. DA-analyysien avulla taas voidaan tutkia, eroaako jonkin yksittäisen lajin runsaus eri ryhmien välillä (esim. sairaat vs. terveet). Beeta-monimuotoisuus hyödyntää näytteen kaikkia lajeja, kun taas DA-menetelmät tarkastelevat lajeja yksittäin. CoDaCoRe-menetelmä taas tutkii pientä lajien osajoukkoa.

CoDaCoRe-menetelmä viittaa CoDa-menetelmiin eli menetelmiin, jotka ottavat huomioon aineiston kompositionaalisen luonteen. Mikrobiomiaineistot voidaan tulkita kompositionaaliksi näytteiden sekvensoinnin takia, koska sekvensoinnissa syntyvä satunnainen kirjastokoko eli yhden näytteen havaittujen lajien yhteenlaskettu lukumäärä ei vastaa biologisessa näytteessä olevien lajien yhteenlaskettua lukumäärää. CoDa-menetelmissä satunnainen kirjastokoko ei ole ongelma, koska menetelmä perustuu kussakin näytteessä havaittujen lajien runsauksien (mikrobilukumäärien) suhteisiin.

Mikrobiomiaineistoissa tyypillisesti lajeja on enemmän kuin näytteitä. CoDaCoRe-menetelmä pyrkii tehokkaasti pienentämään tutkittavien aineistojen dimensioita etsimällä vastemuuttujan kannalta tärkeimmät kaksi osajoukkoa, jotka sisältävät mahdollisimman vähän lajeja. Osajoukkojen runsauksien suhde voidaan määrittää joko balanssien tai yhdelmien avulla. Balanssien mukaisissa osajoukoissa saattaa olla enemmän harvinaisempia sukuja mukana verrattuna yhdelmiin. CoDaCoRe-menetelmän tuloksena on siis kaksi lajien osajoukkoa, jotka ennustavat valittua vastemuuttujaa (esim. sairaat vs. terveet kontrollit) parhaiten. Menetelmä hyödyntää nopeimman laskeutumisen menetelmää, jonka takia CoDaCoRe-menetelmä on laskennallisesti erittäin tehokas. Ennen kuin CoDaCoRe-menetelmää voidaan soveltaa mikrobiomiaineistoon, nollasolut tulee kuitenkin ottaa huomioon esimerkiksi lisäämällä jokaiseen soluun lukumäärä yksi.

CoDaCoRe-menetelmässä käytettävää ristiinvalidointia voidaan regularisoida parametrin λ avulla. Ristiinvalidoinnin avulla määritellään optimaalinen kynnysarvo t, jonka perusteella lajit valitaan osajoukkoihin. Parametrin λ avulla kontrolloidaan sitä, miten kaukana kynnysarvon mukainen ennustevirhe on parhaimmasta ennuste-

virheestä. Mikäli parametrin λ arvo on yksi, kynnysarvon mukaisen ennustevirheen tulee olla yhden standardipoikkeman päässä parhaimmasta ennustevirheestä. Kun parametri on lähellä arvoa 1, menetelmä etsii mahdollisimman harvat osajoukot. Parametrin ollessa lähellä arvoa nolla, löydetyt osajoukot sopivat paremmin aineistoon eli osajoukoissa on enemmän lajeja mukana.

Tässä tutkielmassa sovellettiin CoDaCoRe-menetelmää FinnBrain-syntymäkohorttitutkimuksessa kerättyyn mikrobiomiaineistoon. Empiirisessä osiossa tärkeimmät tutkimuskysymykset sekä binääriselle että jatkuvalle vastemuuttujalle olivat: (1) miten balanssien ja yhdelmien löytämät osajoukot eroavat toisistaan, (2) miten parametri λ vaikuttaa löydettyihin osajoukkoihin sekä (3) miten hyvin mallit ennustavat uusia havaintoja. Sensitiivisyysanalyysinä tutkittiin algoritmin tarvitseman siemenluvun muutoksen vaikutusta löydettyihin osajoukkoihin.

Tässä työssä mallien mukaiset osajoukot raportoitiin kertomalla, mitkä lajit olivat osajoukossa $\hat{J^+}$ ja mitkä $\hat{J^-}$. Balanssien ja yhdelmien mukaisissa kaavoissa osajoukko J^+ on osoittajassa ja $\hat{J^-}$ nimittäjässä. Ei kuitenkaan ole merkitystä, miten päin osajoukot kaavaan sijoitetaan. Tämä johtuu siitä, että $\log(\hat{J^+}/\hat{J^-})$ = -$\log(\hat{J^-}/\hat{J^+})$ eli jos osajoukkojen roolit ovat päinvastoin, log-suhteen etumerkki vaihtuu, mutta jako kahteen osajoukkoon on sama. Lisäksi CoDaCoRe-malli voi löytää useita balansseja (tai yhdelmiä), jotka yhdessä ennustavat vastemuuttujaa parhaiten. Tässä tutkielmassa toimittiin yksinkertaistamisen vuoksi niin, että mikäli malli löysi useita balansseja (tai yhdelmiä), vain ensimmäistä tutkittiin tarkemmin.

Binäärisenä vasteena tutkittiin synnytystapaa (alatie vs. sektio) ja jatkuvana muuttujana oli todennäköisyysmuuttuja, joka oli johdettu silmänliikemittauksista lapsen ollessa 30 kuukauden ikäinen. Todennäköisyys kuvasi katseen irtautumista pelokkaista kasvokuvista. Synnytystapaa tutkittaessa aineiston koko oli 437 ja todennäköisyysmuuttujan tapauksessa 55. Mikrobiomiaineistossa biologinen näyte oli ulostenäyte, joka oli kerätty lapsen ollessa 2.5 kuukauden ikäinen. CoDaCoRe-menetelmän tarkoituksena on ennustaa vastemuuttujaa. Ajallisesti synntystapa on mitattu ennen ulostenäytettä eli analyysien suunta oli ns. väärinpäin, koska ulostenäytteellä ennustettiin jo tapahtunutta tapahtumaa. Todennäköisyysmuuttuja oli mitattu ulostenäytteen jälkeen eli analyysien suunta oli ns. oikeinpäin. Analyysien yksi vahvuus oli soveltaa CoDaCoRe-menetelmää kahteen hyvin eri kokoiseen aineistoon.

Sekä jatkuvalle että binääriselle vastemuuttujalle sovitettiin useita CoDaCoRe-malleja parametrin λ eri arvoilla (1, 0.6, 0.1). Huomattiin, että mallit sopivat aineistoon hieman paremmin parametrin λ matalalla arvolla ($\lambda = 0.1$). Tulos päti sekä balansseille että yhdelmille. Jatkuvan vastemuuttujan kohdalla mallien sopivuutta koko aineistoon kuvattiin selitysasteen avulla. Myös näissä tarkasteluissa huomattiin, että sekä balanssien että yhdelmien paras selitysaste oli parametrin λ arvolla 0.1. Jatkuvan vastemuuttujan tapauksessa yhdelmien selitysasteet olivat kuitenkin huomattavasti huonommat verrattuna balansseihin. Yhdelmien paras selitysaste oli 0.16 ja balanssien paras selitysaste oli 0.39. Erityisesti binäärisen vastemuuttujan tapauksessa parametrin λ matalalla arvolla balanssien mukaiset osajoukot sisälsivät myös aineiston harvinaisempia sukuja. Molempien vasteiden tapauksessa löydettiin

siis osajoukot, jotka olivat vastemuuttujan kannalta tärkeimmät. Tutkittujen mallien mukaan alatiesynnyttäneillä osajoukkojen lajien runsauksien log-suhde oli korkeammalla verrattuna sektiolla synnyttäneisiin sekä balanssien että yhdelmien mukaan. Jatkuvaa vastetta tutkittaessa todennäköisyysmuuttuja kasvoi, kun osajoukkojen lajien runsauksien log-suhde kasvoi.

Mallien kykyä ennustaa uusia havaintoja arvioitiin kahdella eri lähestymistavalla. Binäärisen vasteen tapauksessa aineiston koko oli riittävän suuri opetus- ja testiaineiston muodostamiseen. Opetusaineistolla muodostettujen mallien ennustekykyä arvioitiin testiaineistolla luokittelun tarkkuuden avulla. Balansseja tutkittaessa luokittelun tarkkuus oli parhain (0.80), kun parametrin λ arvo oli 1. Kun $\lambda = 0.1$, luokittelun tarkkuus oli vain hieman matalampi (0.79). Yhdelmien mukainen tulos oli vastaava.

Jatkuvan vastemuuttujan tapauksessa balanssien ja yhdelmien ennustuskykyä tutkitiin ristiinvalidoinnin avulla. Ennustusteen hyvyyttä tutkittiin keskineliövirheen (RMSE), keskimääräisen absoluuttisen virheen (MAE) sekä Spearmanin korrelaation avulla. Parametrin $\lambda = 1$ arvolla yhtäkään yhdelmää ei löytynyt, joten ennusteita ei muodostettu. Balanssit kuitenkin löysivät osajoukot myös parametrin λ arvolla 1, jolloin ennustuskyky olikin parhain. Yhdelmien paras ennustekyky oli parametrin λ arvolla 0.6, joilloin balanssit suorituivat taas hieman huonommin.

Sensitiivisyysanalyysien mukaan löydetyt osajoukot olivat riippuvaisia asetetusta siemenluvusta. Ilmiö oli nähtävissä sekä binäärisellä että jatkuvalla vasteella. Binääristä vastemuuttujaa tarkasteltaessa yhdelmät löysivät aina samat osajoukot, mutta balanssien mukaiset osajoukot vaihtelivat. Suurin joukko, josta balanssit muodostuivat, sisälsi 14 sukua ja pienin joukko vain 4 sukua. Kaikkien mallien ennustekyvyt olivat kuitenkin lähes samat. Jatkuvalla vasteella tehdyissä sensitiivisyysanalyyseissa yhdelmien tulos muuttui siemenluvun mukaan. Suurin joukko, josta yhdelmät muodostuivat, oli seitsemän sukua, ja pienin taas sisälsi vain kaksi sukua. Balanssit muodostuivat aina samoista osajoukoista, joissa oli yhteensä kahdeksan sukua. Hyvin todennäköistä on, että CoDaCoRe-menetelmän käyttämä nopeimman laskeutumisen menetelmän alkuarvo on riippuvainen asetetusta siemenluvusta. Nopeimman laskeutumisen menetelmän löytämät optimiarvot ovat hyvin riippuvaisia lähtöarvosta eikä menetelmässä voida taata globaalia optimia.

Ennen analyyseja kaikkiin havaittuihin lajirunsauksiin lisättiin pseudolukuna arvo yksi. Toinen mielenkiintoinen sensitiivisyysanalyysi olisikin ollut tutkia, miten pseudoluvun vaikutus näkyy osajoukoissa. Tässä tutkielmassa tehdyissä analyyseissa huomattiin, että sekä balanssien että yhdelmien osajoukossa esiintyi harvinaisia sukuja. Pseudoluku näkyy siis jollain tavalla mallien löytämissä osajoukoissa. Esimerkiksi binäärisen vasteen osajoukoissa esiintyi Clostridioides-suku, jonka tapauksessa yhteensä 437 näytteestä oli 394 solua, joissa runsaus oli yksi (asetettu pseudoluku oli 1).

CoDaCoRe-menetelmän tarjoamia selviä biomarkkereita pidetään erittäin hyvänä asiana menetelmän esittelevässä artikkelissa [3]. Erityisesti binäärisellä vastemuut-

tujalla sensitiivisyysanalyysissä muodostetut mallit ennustivat vastetta hyvin eli tässä mielessä löydetyt biomarkkerit toimivat. Jos kuitenkin biomarkkereiden sisältämistä lajeista halutaan tehdä biologisia tulkintoja, CoDaCoRe-menetelmä ei ole paras menetelmä käytettäväksi, koska osajoukoissa esiintyvät riippuvat siemenluvusta. Siemenluvun vaikutusta oli pohdittu myös CoDaCoRe-menetelmän github-sivustolla [42]. Erään käyttäjän kommentissa oli pohdittu, että CoDaCoRe-mallin voisi tehdä esimerkiksi sata kertaa ilman siemenluvun asettamista ja katsoa, mitä lajeja osajoukoissa keskimäärin olisi. CoDaCoRe-menetelmän laskennallista tehokkuutta kuitenkin korostetaan ja mallin muodostaminen sata kertaa huonontaa tätä ominaisuutta. Mikrobiomiaineistoissa lajien lukumäärä on tavallisesti suurempi kuin näytteiden lukumäärä eli laskennallinen tehokkuus on tärkeää. Laskennallisen tehokkuuden parantuessa tulosten stabiilius ja oikeellisuus ei kuitenkaan saisi vaarantua. Jos halutaan tehdä biologisia tulkintojen löydetyistä osajoukoista, tulosten stabiilius on tärkeämpää verrattuna laskennalliseen tehokkuuteen.

Tässä tutkielmassa tutkittiin vain yhdessä aikapisteessä kerättyä biologista näytettä. FinnBrain-tutkimuksessa on kerätty ulostenäytteitä monista aikapisteistä. Tämän takia yksi mielenkiintoinen jatkotutkimus olisi pyrkiä hyödyntämään useita mittauspisteitä. Lisäksi tässä tutkielmassa ei tutkittu kovariaattien lisäämistä malleihin. Tämän pitäisi kuitenkin olla mahdollista CoDaCoRe-menetelmässä.

CoDaCoRe-menetelmä ei ole ainut menetelmä, jolla voidaan määrittää optimaalisia balansseja. Jatkotarkasteluna olisikin mielenkiintoista verrata CoDaCoRe-menetelmää esimerkiksi Selbal-menetelmään. Selbal-menetelmässä optimaalisen balanssin etsiminen aloitetaan kahdesta lajista eli parittaisesta logaritmisesta suhteesta. Tämän jälkeen menetelmä lisää tarkasteluun yhden lajin ja laskee, onko useamman lajin balanssi parempi verrattuna aikaisempaan. Optimaalisen balanssin määrittäminen tapahtuu siis eri tavalla verrattuna CoDaCoRe-menetelmään, mutta se muodostetaan aineiston perusteella optimaaliseksi kuten CoDaCoRe-menetelmässäkin. Tämän tutkielman pohjalta olisi mielenkiintoista tutkia esimerkiksi, ovatko Selbal-menetelmän mukaiset tulokset stabiilimpia tai onko se laskennallisesti hitaampi verrattuna CoDaCoRe-menetelmään. Biologisesta näkökulmasta olisi mielenkiintoista nähdä, löytävätkö CoDaCoRe- ja Selbal-menetelmät samanlaisia bakteerilajien osajoukkoja.

Viitteet

[1] Dirk Haller, editor. *The gut microbiome in health and disease*. Springer, Cham, Switzerland, 2018. ISBN 9783319905457 9783319905440.

[2] Yinglin Xia, Jun Sun, and Ding-Geng Chen. *Statistical Analysis of Microbiome Data with R*. ICSA Book Series in Statistics. Springer Singapore, Singapore, 2018. ISBN 9789811315336 9789811315343. doi: 10.1007/978-981-13-1534-3. URL `http://link.springer.com/10.1007/978-981-13-1534-3`.

[3] Elliott Gordon-Rodriguez, Thomas P. Quinn, and John P. Cunningham. Learning sparse log-ratios for high-throughput sequencing data. *Bioinformatics (Oxford, England)*, 38(1):157–163, December 2021. ISSN 1367-4811. doi: 10.1093/bioinformatics/btab645.

[4] Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41(Database issue):D590–596, January 2013. ISSN 1362-4962. doi: 10.1093/nar/gks1219.

[5] Jacob T. Nearing, André M. Comeau, and Morgan G. I. Langille. Identifying biases and their potential solutions in human microbiome studies. *Microbiome*, 9(1):113, May 2021. ISSN 2049-2618. doi: 10.1186/s40168-021-01059-0. URL `https://doi.org/10.1186/s40168-021-01059-0`.

[6] Jocelyn M. Choo, Lex EX Leong, and Geraint B. Rogers. Sample storage conditions significantly influence faecal microbiome profiles. *Scientific Reports*, 5(1):16350, November 2015. ISSN 2045-2322. doi: 10.1038/srep16350. URL `https://www.nature.com/articles/srep16350`.

[7] Michael R. McLaren, Jacob T. Nearing, Amy D. Willis, Karen G. Lloyd, and Benjamin J. Callahan. Implications of taxonomic bias for microbial differential-abundance analysis. preprint, Bioinformatics, August 2022. URL `http://biorxiv.org/lookup/doi/10.1101/2022.08.19.504330`.

[8] M. Luz Calle. Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, 17(1):e6, March 2019. ISSN 1598-866X. doi: 10.5808/GI.2019.17.1.e6.

[9] Gregory B. Gloor, Jean M. Macklaim, Vera Pawlowsky-Glahn, and Juan J. Egozcue. Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, 8, 2017. ISSN 1664-302X. URL `https://www.frontiersin.org/articles/10.3389/fmicb.2017.02224`.

[10] Gregory B. Gloor, Jia Rong Wu, Vera Pawlowsky-Glahn, and Juan José Egozcue. It's all relative: analyzing microbiome data as compositions. *Annals of Epidemiology*, 26(5):322–329, May 2016. ISSN 1873-2585. doi: 10.1016/j.annepidem.2016.03.003.

[11] Sophie Weiss, Zhenjiang Zech Xu, Shyamal Peddada, Amnon Amir, Kyle Bittinger, Antonio Gonzalez, Catherine Lozupone, Jesse R. Zaneveld, Yoshiki Vázquez-Baeza, Amanda Birmingham, Embriette R. Hyde, and Rob Knight. Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome*, 5(1):27, March 2017. ISSN 2049-2618. doi: 10.1186/s40168-017-0237-y.

[12] Huang Lin and Shyamal Das Peddada. Analysis of microbial compositions: a review of normalization and differential abundance analysis. *NPJ biofilms and microbiomes*, 6(1):60, December 2020. ISSN 2055-5008. doi: 10.1038/s41522-020-00160-w.

[13] Michael I. Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):550, 2014. ISSN 1474-760X. doi: 10.1186/s13059-014-0550-8.

[14] Paul J. McMurdie and Susan Holmes. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, 10(4):e1003531, April 2014. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1003531. URL `https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003531`.

[15] T. Bastiaanssen, T. Quinn, and A. Loughman. Treating Bugs as Features: A compositional guide to the statistical analysis of the microbiome-gut-brain axis. July 2022. URL `https://www.semanticscholar.org/paper/Treating-Bugs-as-Features%3A-A-compositional-guide-to-Bastiaanssen-Quinn/6ed7e47e540f5703c53186e50de7dd110c95b84a`.

[16] Jari Oksanen, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, Matt Barbour, Michael Bedward, Ben Bolker, Daniel Borcard, Gustavo Carvalho, Michael Chirico, Miquel De Caceres, Sebastien Durand, Heloisa Beatriz Antoniazi Evangelista, Rich FitzJohn, Michael Friendly, Brendan Furneaux, Geoffrey Hannigan, Mark O. Hill, Leo Lahti, Dan McGlinn, Marie-Helene Ouellette, Eduardo Ribeiro Cunha, Tyler Smith, Adrian Stier, Cajo J. F. Ter Braak, and James Weedon. vegan: Community Ecology Package, October 2022. URL `https://CRAN.R-project.org/package=vegan`.

[17] Raivo Kolde. pheatmap: Pretty Heatmaps, January 2019. URL `https://CRAN.R-project.org/package=pheatmap`.

[18] Anna-Katariina Aatsinki, Leo Lahti, Henna-Maria Uusitupa, Eveliina Munukka, Anniina Keskitalo, Saara Nolvi, Siobhain O'Mahony, Sami Pietilä, Laura L. Elo, Erkki Eerola, Hasse Karlsson, and Linnea Karlsson. Gut microbiota composition is associated with temperament traits in infants. *Brain, Behavior, and Immunity*, 80:849–858, August 2019. ISSN 1090-2139. doi: 10.1016/j.bbi.2019.05.035.

[19] Katerina V.-A. Johnson. Gut microbiome composition and diversity are related to human personality traits. *Human Microbiome Journal*, 15:None, March 2020. ISSN 2452-2317. doi: 10.1016/j.humic.2019.100069.

[20] Andrew D. Fernandes, Jean M. Macklaim, Thomas G. Linn, Gregor Reid, and Gregory B. Gloor. ANOVA-like differential expression (ALDEx) analysis for mixed population RNA-Seq. *PloS One*, 8(7):e67019, 2013. ISSN 1932-6203. doi: 10.1371/journal.pone.0067019.

[21] Huang Lin and Shyamal Das Peddada. Analysis of compositions of microbiomes with bias correction. *Nature Communications*, 11(1):3514, July 2020. ISSN 2041-1723. doi: 10.1038/s41467-020-17041-7.

[22] Jacob T. Nearing, Gavin M. Douglas, Molly G. Hayes, Jocelyn MacDonald, Dhwani K. Desai, Nicole Allward, Casey M. A. Jones, Robyn J. Wright, Akhilesh S. Dhanani, André M. Comeau, and Morgan G. I. Langille. Microbiome differential abundance methods produce different results across 38 datasets. *Nature Communications*, 13(1):342, January 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-28034-z. URL `https://www.nature.com/articles/s41467-022-28034-z`.

[23] Thomas P. Quinn, Elliott Gordon-Rodriguez, and Ionas Erb. A Critique of Differential Abundance Analysis, and Advocacy for an Alternative, June 2021. URL `http://arxiv.org/abs/2104.07266`. arXiv:2104.07266 [q-bio, stat].

[24] Michael Greenacre, Eric Grunsky, and John Bacon-Shone. A comparison of isometric and amalgamation logratio balances in compositional data analysis. *Computers & Geosciences*, 148:104621, March 2021. ISSN 0098-3004. doi: 10.1016/j.cageo.2020.104621. URL `https://www.sciencedirect.com/science/article/pii/S0098300420305999`.

[25] Michael Greenacre and Eric Grunsky. The isometric logratio transformation in compositional data analysis: a practical evaluation. 2018. doi: 10.13140/RG.2.2.10817.20322. URL `http://rgdoi.net/10.13140/RG.2.2.10817.20322`.

[26] Thomas P. Quinn and Ionas Erb. Amalgams: data-driven amalgamation for the dimensionality reduction of compositional data. *NAR genomics and bioinformatics*, 2(4):lqaa076, December 2020. ISSN 2631-9268. doi: 10.1093/nargab/lqaa076.

[27] J. Rivera-Pinto, J. J. Egozcue, V. Pawlowsky-Glahn, R. Paredes, M. Noguera-Julian, and M. L. Calle. Balances: a New Perspective for Microbiome Analysis. *mSystems*, 3(4):e00053–18, August 2018. ISSN 2379-5077. doi: 10.1128/mSystems.00053-18. URL `https://journals.asm.org/doi/10.1128/mSystems.00053-18`.

[28] Elliott Gordon-Rodriguez, Thomas P. Quinn, and John P. Cunningham. Supplementary Material for: Learning Sparse Log-Ratios for High-Throughput Sequencing Data. URL `https://academic.oup.com/bioinformatics/article/38/1/157/6366546`.

[29] Ethem Alpaydin. *Introduction to machine learning*. Adaptive computation and machine learning series. The MIT Press, Cambridge, Massachusetts, fourth edition edition, 2020. ISBN 9780262043793.

[30] Marko Mäkelä. *Luentomoniste: Optimointialgoritmit*. 2022.

[31] Linnea Karlsson, Mimmi Tolvanen, Noora M. Scheinin, Henna-Maria Uusitupa, Riikka Korja, Eeva Ekholm, Jetro J. Tuulari, Marjukka Pajulo, Minna Huotilainen, Tiina Paunio, Hasse Karlsson, and FinnBrain Birth Cohort Study Group. Cohort Profile: The FinnBrain Birth Cohort Study (FinnBrain). *International Journal of Epidemiology*, 47(1):15–16j, February 2018. ISSN 1464-3685. doi: 10.1093/ije/dyx173.

[32] Eeva-Leena Kataja, Linnea Karlsson, Jukka M. Leppänen, Juho Pelto, Tuomo Häikiö, Saara Nolvi, Henri Pesonen, Christine E. Parsons, Jukka Hyönä, and Hasse Karlsson. Maternal Depressive Symptoms During the Pre- and Postnatal Periods and Infant Attention to Emotional Faces. *Child Development*, 91(2), 2018. ISSN 0009-3920, 1467-8624. doi: 10.1111/cdev.13152. URL `https://onlinelibrary.wiley.com/doi/10.1111/cdev.13152`.

[33] Francisco Melo. Area under the ROC Curve. In Werner Dubitzky, Olaf Wolkenhauer, Kwang-Hyun Cho, and Hiroki Yokota, editors, *Encyclopedia of Systems Biology*, pages 38–39. Springer, New York, NY, 2013. ISBN 9781441998637. doi: 10.1007/978-1-4419-9863-7_209. URL `https://doi.org/10.1007/978-1-4419-9863-7_209`.

[34] R: The R Project for Statistical Computing. URL `https://www.r-project.org/`.

[35] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, Davis Vaughan, Posit Software, and PBC. dplyr: A Grammar of Data Manipulation, April 2023. URL `https://cran.r-project.org/web/packages/dplyr/index.html`.

[36] Hadley Wickham, Davis Vaughan, Maximilian Girlich, Kevin Ushey, Posit, and PBC. tidyr: Tidy Messy Data, January 2023. URL `https://cran.r-project.org/web/packages/tidyr/index.html`.

[37] Felix G. M. Ernst, Sudarshan A. Shetty, Tuomas Borman, Leo Lahti, Yang Cao, Nathan D. Olson, Levi Waldron, Marcel Ramos, Héctor Corrada Bravo, Jayaram Kancherla, and Domenick Braccia. mia: Microbiome analysis, 2023. URL `https://bioconductor.org/packages/mia/`.

[38] Elliott Gordon-Rodriguez and Thomas Quinn. codacore: Learning Sparse Log-Ratios for Compositional Data, August 2022. URL `https://CRAN.R-project.org/package=codacore`.

[39] Hadley Wickham, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, Dewey Dunnington, Posit, and PBC. ggplot2: Create Elegant Data Visualisations Using the Grammar

of Graphics, April 2023. URL https://cran.r-project.org/web/packages/ggplot2/index.html.

[40] Barret Schloerke, Di Cook, Joseph Larmarange, Francois Briatte, Moritz Marbach, Edwin Thoen, Amos Elberg, Ott Toomet, Jason Crowley, Heike Hofmann, and Hadley Wickham. GGally: Extension to 'ggplot2', June 2021. URL https://cran.r-project.org/web/packages/GGally/index.html.

[41] Thomas Lin Pedersen. patchwork: The Composer of Plots, August 2022. URL https://cran.r-project.org/web/packages/patchwork/index.html.

[42] GitHub. URL https://github.com/egr95/R-codacore/issues/13. luettu 2023-05-16.

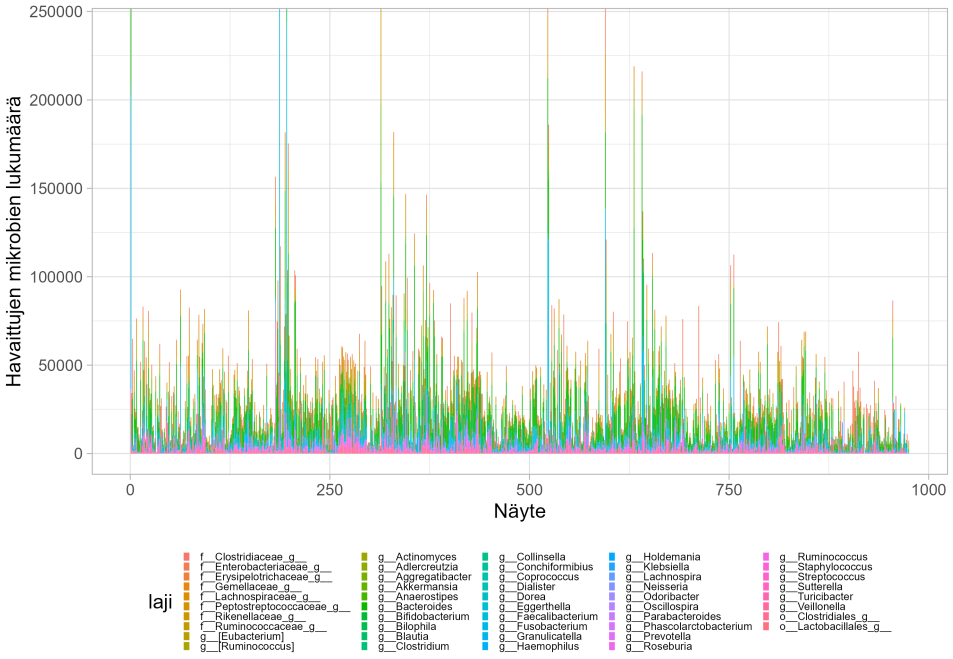
[43] Ruairi Robertson. 16S rRNA Gene Sequencing vs. Shotgun Metagenomic Sequencing. URL https://blog.microbiomeinsights.com/16s-rrna-sequencing-vs-shotgun-metagenomic-sequencing. luettu 2023-03-30.

Liitteet

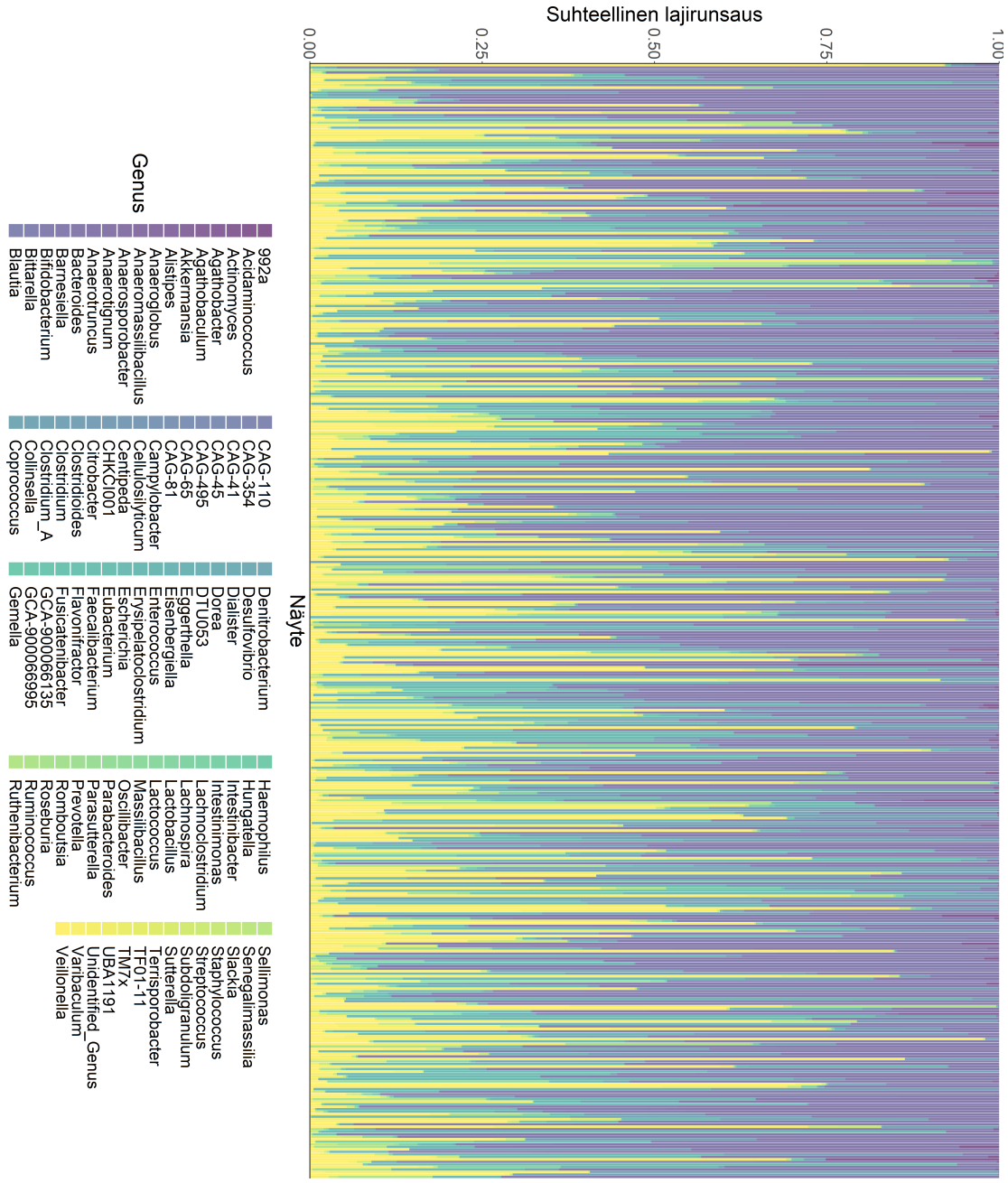
A Kuvia

Taulukko A1: *16S rRNA- ja shotgun-sekvensoinnin hyvät ja huonot puolet.* Hyvät puolet on ilmoitettu plusmerkillä ja huonot puolet miinusmerkillä. [2, 43]

16S rRNA-sekvensointi	**shotgun-sekvensointi**
+ Voidaan tutkia bakteereja ja arkkeja + 16S rRNA-geeni löytyy kaikista bakteereista ja arkeista + On saatavilla laajoja tietokantoja + Halvempi	+ Voidaan tutkia myös viruksia, aitotumallisia ja sieniä + Taksonomisen kompositionaalisuuden lisäksi kertoo myös funktionaalisuudesta + Havaitsee helpommin uusia ja harvinaisia viruksia
− Yliestimoi lajien runsautta − Havaitsee vain taksonomisen komposition −Vaikea käyttää uusien tai hyvin poikkeavien mikrobien (esim. virukset ja sienet) analysoimisessa − Ei standardoituja tilastollisia analysointitapoja	− Teknisesti vaikeampi ja kalliimpi − Lopputuloksena aineisto on iso ja monimutkainen, minkä seurauksena mm. tilastolliset analyysit ja laskenta vaikeutuvat



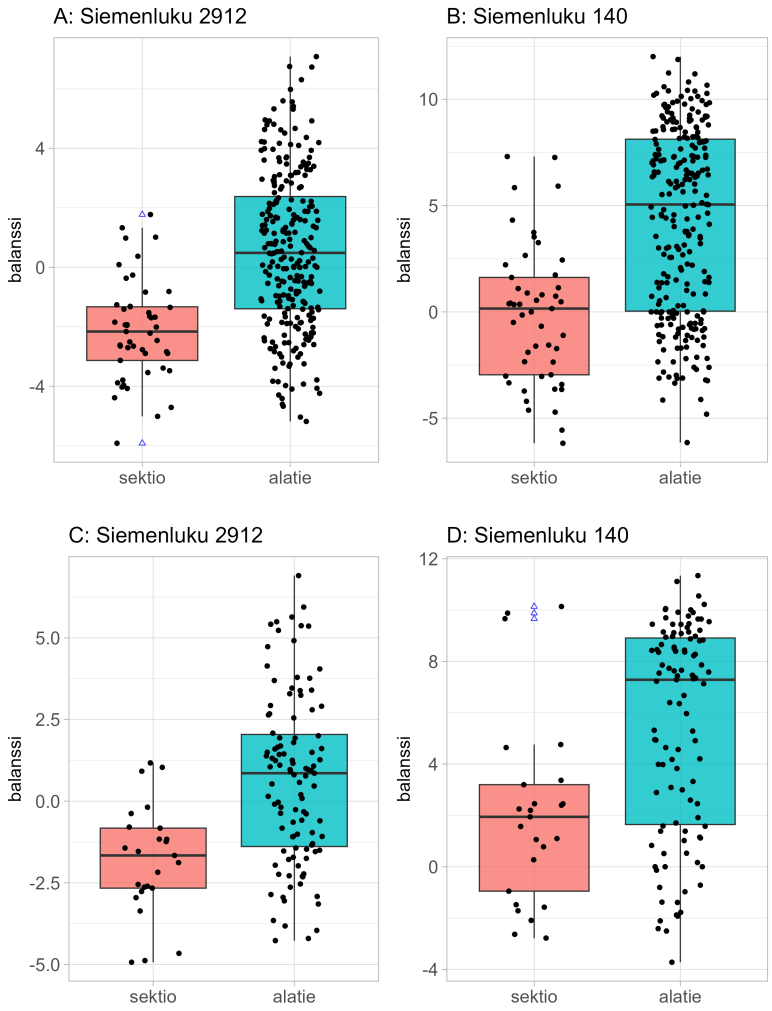
Kuva A1: Pylväsdiagrammi Crohnin sairautta kuvaavasta aineistosta. Näytteet on x-akselilla ja y-akselilla on lajien havaitut lukumäärät välillä [0, 25000]. Tämän esimerkin yhteydessä lajit ovat todellisuudessa sukuja. Luvun selvyyden vuoksi käytetään sanaa laji.



Kuva A2: Pinotussa pylväsdiagrammissa x-askelilla on eri näytteet ja y-akselilla suhteelliset lajirunsaudet. Kuvaan on merkitty eri väreillä eri suvut. Tunnistamattomat suvut on laskettu yhteen.

B CoDaCoRe-menetelmän tuloksia

Binäärinen vaste



Kuva B1: *Binäärisen vasteen (synnytystapa) sensitiivisyysanalyysien mukaiset kuvaajat (taulukko 11).* Kuvissa A ja C siemenluku on 2912. Kuva A on piirretty opetusaineiston avulla ja C testiaineiston avulla. Kuvissa B ja D siemenlukuna on 140. Kuva B on piirretty opetusaineiston avullla ja D testiaineiston avulla. Kaikkien kuvaajien mukaan alatiesynnyttäneiden mediaani oli korkeammalla verrattuna sektiolla synnyttäneisiin. Siemenlukujen 2912 ja 140 mukaiset mallit eroavat toisistaan. Siemenluvun 140 mukaisissa kuvaajissa balanssien jakauma on leveämmällä välillä verrattuna siemenluvun 2912 kuvaajiin.

Taulukko B1: *Mallien sopivuus opetusaineistoon sekä ennustamisen tarkkuus testiaineistolla sensitiivisyysanalyysin perusteella.* Taulukkoon on kerätty sensitiivisyysanalyysien mukaisten mallien opetusaineistoon perustuva AUC-arvo ja testiaineistoon perustuva luokittelun tarkkuus. Mallien $5^{(2912)}$ ja $5^{(140)}$ osajoukot muodostuivat samoista suvuista, joten niiden mukaiset arvot olivat samat.

malli	log-suhde	opetusaineiston AUC	luokittelun tarkkuus
$1^{(2912)}$	B	0.80	0.80
$1^{(140)}$	B	0.78	0.78
$5^{(2912)}$ ja $5^{(140)}$	Y	0.72	0.82

Jatkuva vaste

Taulukko B2: *CoDaCoRe-menetelmän mukaiset parhaimmat osajoukot $\hat{J}^+$ ja $\hat{J}^-$ balanssien mukaan jatkuvalle vasteelle, katseen irtautumisen todennäköisyys, parametrin λ eri arvoilla.* Jokainen malli löysi vain yhden balanssin. Jokaisen mallin selitysaste on ilmoitettu (R^2).

Malli A $\lambda = 1$ overlap = T $R^2 = 0.315$	**1. balanssi** $\hat{J}^+$ = {Bifidobacterium, Lachnospiraceae:Clostridium, Haemophilus, Eggerthella} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Bittarella, Sutterella}
Malli B $\lambda = 0.6$ overlap = T $R^2 = 0.368$	**1. balanssi** $\hat{J}^+$ = {Bifidobacterium, Lachnospiraceae:Clostridium, Haemophilus, Eggerthella} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Bittarella, Sutterella, Erysipelatoclostridium}
Malli C $\lambda = 0.1$ overlap = T $R^2 = 0.393$	**1. balanssi** $\hat{J}^+$ = {Bifidobacterium, Lachnospiraceae:Clostridium, Haemophilus, Blautia, Eggerthella} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Bittarella, Sutterella, Erysipelatoclostridium}
Malli D $\lambda = 0.1$ overlap = F $R^2 = 0.393$	**1. balanssi** $\hat{J}^+$ = {Bifidobacterium, Lachnospiraceae:Clostridium, Haemophilus, Blautia, Eggerthella} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Bittarella, Sutterella, Erysipelatoclostridium}

Taulukko B3: *CoDaCoRe-menetelmän mukaiset parhaimmat osajoukot $\hat{J}^+$ ja $\hat{J}^-$ yhdelmien mukaan jatkuvalle vasteelle, katseen irtautumisen todennäköisyys, parametrin λ eri arvoilla.* Jokaisen mallin yhdelmät on esitetty hierarkkisesti. Parhain yhdelmä on ensimmäisenä. Tunnistamaton suku on lyhennetty kirjaimilla UG. Mallien selitysaste on ilmoitettu (R^2). Mikäli malli löysi useamman yhdelmän, selitysasteen laskemiseen on käytetty vain ensimmäistä yhdelmää.

Malli E $\lambda = 1$ overlap =T $R^2 = NA$	Ei löydettyjä yhdelmiä
Malli F $\lambda = 0.6$ overlap = T $R^2 = 0.082$	**1. yhdelmä** $\hat{J}^+$ = {Bifidobacterium, Bacteroides} $\hat{J}^-$ = {Clostridiaceae:Clostridium}
Malli G $\lambda = 0.1$ overlap = T $R^2 = 0.156$	**1. yhdelmä** $\hat{J}^+$ = {Bifidobacterium, Bacteroides, Lachnospiraceae:Ruminococcus, Lachnospiraceae:Clostridium} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella, Enterococcaceae:UG, Barnesiella} **2. yhdelmä** $\hat{J}^+$ = {Bifidobacterium} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella}
Malli H $\lambda = 0.1$ overlap = F $R^2 = 0.156$	**1. yhdelmä** $\hat{J}^+$ = {Bifidobacterium, Bacteroides, Lachnospiraceae:Ruminococcus, Lachnospiraceae:Clostridium} $\hat{J}^-$ = {Clostridiaceae:Clostridium, Veillonella, Enterococcaceae:UG, Barnesiella} **2. yhdelmä** $\hat{J}^+$ = {Escherichia, Enterobacteriaceae:UG, Hungatella} $\hat{J}^-$ = {Bittarella, Sutterella, Erysipelatoclostridium} **3. yhdelmä** $\hat{J}^+$ = {Collinsella, Parabacteroides, Prevotella} $\hat{J}^-$ = {Flavonifractor}

C R-koodi

```
# kirjastot
library(ggplot2)
library(patchwork)
library(codacore)
library(ggplot2)
library(dplyr)
library(tidyr)
library(vegan)
library(GGally)
library(gmodels)
library(pheatmap)
library(data.table)

theme_set(theme_light(base_size = 15))

# Luku 2 ------------------------------
data("Crohn")
head(Crohn)[1:6, 1:4]
x <- Crohn[, -ncol(Crohn)]
y <- Crohn[, ncol(Crohn)]

Crohn$summa <- rowSums(x)
summary(Crohn$summa)

#pitkaan muotoon
x$id = c(1:975)
long <- x %>%
  pivot_longer(cols = everything(1:48),
               names_to="laji", values_to="value")

plot <- long %>% #filter(id <= 50) %>%
  ggplot(aes(x=id, y=value, fill=laji)) +
  geom_bar(stat="identity", position="stack") +
  coord_cartesian( ylim = c(0, 260000)) +
  labs(x='Nyte', y='Havaittujen mikrobien lukumaara')+
  theme(legend.position="bottom", legend.text=element_text(size=8),
        legend.key.height= unit(2, 'mm'),
        legend.key.width= unit(2, 'mm'))

#suhteellinen osuus
x <- Crohn[, -c(49:50)]
rel_ab <- t(apply(x, 1, function(i) i/sum(i)))
rel_ab <- data.frame(rel_ab)
rel_ab$id = c(1:975)

#pitkaan muotoon
long_rel <- rel_ab %>%
```

```
  pivot_longer(cols = everything(1:48),
               names_to="laji", values_to="value")

plot <- long_rel %>% #filter(id <= 50) %>%
  ggplot(aes(x=id, y=value, fill=laji)) +
  geom_bar(stat="identity", position="stack") +
  #coord_cartesian( ylim = c(0, 240000)) +
  labs(x='Nyte', y='Havaittujen mikrobien  lukumr  suhteutettuna
      toisiina')+
  theme(legend.position="bottom", legend.text=element_text(size=8),
        legend.key.height= unit(2, 'mm'),
        legend.key.width= unit(2, 'mm'))

#alfa-monimuotoisuus
Shannon <- vegan::diversity(x, "shannon")
Simpson <- vegan::diversity(x, "simpson")
Lajirikkaus <- specnumber(x)

alpha <- data.frame(cbind(Shannon, Simpson, Lajirikkaus))
alpha <- cbind(alpha, Crohn)

lab <- c("tapaus", "verrokki")
names(lab) <- c('CD', 'no')

p1 <- ggplot(alpha, aes(x = y, y = Shannon, fill=y)) +
  geom_boxplot(alpha = 0.80) +
  #geom_jitter(shape=16, position=position_jitter(0.2))+
  labs(title='Shannonin indeksi', y ='', x='')+
  scale_x_discrete(labels=lab)+
  theme(legend.position = "none",
        axis.text.x = element_text(size=17))

p2 <- ggplot(alpha, aes(x = y, y = Simpson, fill=y)) +
  geom_boxplot(alpha = 0.80) +
  #geom_jitter(shape=16, position=position_jitter(0.2))+
  labs(title='Simpsonin indeksi', x='', y='')+
  scale_x_discrete(labels=lab)+
  theme(legend.position = "none",
        axis.text.x = element_text(size=17))

p3 <- ggplot(alpha, aes(x = y, y = Lajirikkaus, fill=y)) +
  geom_boxplot(alpha = 0.80) +
  #geom_jitter(shape=16, position=position_jitter(0.2))+
  labs(title='Lajirikkaus', x='', y='')+
  scale_x_discrete(labels=lab)+
  theme(legend.position = "none",
        axis.text.x = element_text(size=17))

kuva <- p1 + p2 + p3 + plot_layout(ncol = 3)
```

```
#jakauma
scatter <- ggpairs(alpha[, 1:3],
                   upper = "blank")+
  theme_bw(base_size = 25)+
  theme(plot.background = element_rect(fill = "white"),
        panel.background = element_rect(fill = "white"))

# beeta-monimuotoisuus
# Heatmap
d <- data.table(
         nayte_1 = c(0, 0.2, 0.1, 0.23, 0.31, 0.14, 0.65, 0.7, 0.72, 0.55),
         nayte_2 = c(0.2, 0, 0.32, 0.11, 0.23, 0.05, 0.81, 0.89, 0.74,
             0.83),
         nayte_3 = c(0.1, 0.32, 0, 0.11, 0.21, 0.09, 0.61, 0.71, 0.8, 0.9),
         nayte_4 = c(0.23, 0.11, 0.11, 0, 0.3, 0.25, 0.6, 0.55, 0.78, 0.9),
         nayte_5 = c(0.31, 0.23, 0.21, 0.3, 0, 0.15, 0.7, 0.9, 0.75, 0.79),
         nayte_6 = c(0.14, 0.05, 0.09, 0.25, 0.15, 0, 0.58, 0.76, 0.89,
             0.65),
         nayte_7 = c(0.65, 0.81, 0.61, 0.6, 0.7, 0.58, 0, 0.21, 0.11,
             0.13),
         nayte_8 = c(0.7, 0.89, 0.71, 0.55, 0.9, 0.76, 0.21, 0, 0.04,
             0.14),
         nayte_9 = c(0.72, 0.74, 0.8, 0.78, 0.75, 0.89, 0.11, 0.04, 0,
             0.21),
         nayte_10 = c(0.55, 0.83, 0.9, 0.9, 0.79, 0.65, 0.13, 0.14, 0.21,
             0)
         )

row.names(d) <- colnames(d)
sairaus <- data.frame( Sairaus= rep(c("ei", "kyll"), c(6,4)))
row.names(sairaus) <- colnames(d)

ann_colors = list(Sairaus = c(ei = "cornflowerblue", kyll = "yellow1"))

pheatmap(d, annotation_col = sairaus, cluster_cols=F, cluster_rows=F,
    color = hcl.colors(50, "Magenta"),
        annotation_colors = ann_colors, angle_col=90, filename =
            "heat.png")

# Crohn aineiston etisyysmatriisi
BrayCurtis <- vegan::vegdist(rel_ab[1:4,1:48], method = "bray", upper=T)

# Luku 3 -----------------------------
#sigmoidifunktio
sigmoid <- function(x) {
  1 / (1 + exp(-x))
}
```

```
x <- seq(-20, 20, 0.1)
y <- sigmoid(x)
data <- as.data.frame(cbind(x,y))

p <- ggplot(data, aes(x = x, y = y))+
  geom_point()+
  labs(y='sigmoid(a)', x='a', title='A')

#relu-funktio
relu <- function(x){
  y <- NULL
  for(i in 1:length(x)){
    y[i] = max(x[i], 0)
  }
  return(y)
}

x <- seq(-20, 20, 0.1)
y <- relu(x)
data <- as.data.frame(cbind(x,y))

p1 <- ggplot(data, aes(x = x, y = y))+
  geom_point()+
  labs(y='ReLU(a)', x='a', title= 'B')

#Nopeimman laskeutumisen esimerkki: molemmille alkuarvoille
f <- function(x){0.5*x^4-0.5*x^3-3*x^2+2*x+4}
x <- seq(-3, 3, 0.001)
y <- f(x)
data <- data.frame(x=x, y=y)
df <- function(x){2*x^3-1.5*x^2-6*x+2}

a <- -3
x1 <- NULL
y1 <- NULL

for(i in 1:5){
  print('***')
  x1[i] = a
  y1[i] = f(a)
  print(cbind(a, f(a)))
  a = a -0.1*df(a)
}

s <- data.frame(x1, y1)

p1 <- ggplot(data, aes(x = x, y = y))+
  geom_point()+
  labs(title='A', y='f(x)')+
```

```
  geom_point(data=s, aes(x=x1, y=y1), colour="red", size=5)+
  theme(text=element_text(size=30))

# Luku 4 ----------------------------
#aineiston kuvailu: Leo Lahti ym. https://microbiome.github.io/OMA/
# binaarinen vaste
dm <- data_2.5kk[ , !is.na(data_2.5kk$delivery_mode)]
tse1 <- agglomerateByRank(dm, rank = "Genus")

# nimet paremmiksi
nimet <- tibble(g = rowData(tse1)[,6],
           f = rowData(tse1)[,5],
           name = rownames(tse1)) %>%
  rowwise() %>%
  mutate(rowname = case_when(str_detect(name, "Unidentified_Genus") == T ~
      paste(f, name, sep = ":"),
                             str_detect(name, "Unidentified_Genus") == F ~
                               name)) %>%
  mutate(rowname = case_when(str_detect(rowname, "Eubacterium") == T ~
      paste(f, name, sep = ":"),
                             str_detect(name, "Eubacterium") == F ~ rowname))
                               %>%
  mutate(rowname = case_when(str_detect(rowname, "Clostridium") == T ~
      paste(f, name, sep = ":"),
                             str_detect(name, "Clostridium") == F ~ rowname))
                               %>%
  mutate(rowname = case_when(str_detect(rowname, "Ruminococcus") == T ~
      paste(f, name, sep = ":"),
                             str_detect(name, "Ruminococcus") == F ~ rowname))

getTopTaxa(tse1, method="median", top=5)
rownames(tse1) <- nimet$rowname
d <- t(assay(tse1))
d <- d + 1
y <- as.factor(colData(tse1)$delivery_mode)
y

set.seed(124)
trainIndex <- sample(1:nrow(d), 0.7 * nrow(d))
dTrain <- d[trainIndex, ] #305
yTrain <- y[trainIndex]

dTest <- d[-trainIndex, ]
yTest <- y[-trainIndex]

# balanssit
lambda <- c(1, 0.6, 0.1, 0.1)
overlap <- c(T, T, T, F)
model <- NULL
```

```
log_suhteet <- NULL

for(i in 1:4){
tf$random$set_seed(0) #siemenluvun asettaminen
set.seed(0)

print('Uusi malli')
  model = codacore(
  dTrain,
  yTrain,
  logRatioType = 'balances',
  lambda = lambda[i],
  overlap = overlap[i]
)
print(model)

# ennustaminen
yHat = predict(model, dTest, logits = F, numLogRatios = 1)
cat("Test set AUC =",
    pROC::auc(pROC::roc(yTest, yHat, quiet = T)))

# Todennakisyyksien muuttaminen luokiksi
failure = yHat < 0.5
success = yHat >= 0.5
yHat[failure] = levels(y)[1]
yHat[success] = levels(y)[2]
print(confusionMatrix(as.factor(yTest), as.factor(yHat)))

cat("Classification accuracy on test set =",
    round(mean(yHat == yTest), 2))

if(i==1){
  log_suhteet = getLogRatios(model)
}
}

#kuvat tehty ggplot2-kirjastolla
#jatkuva vaste
d <- merge(md, data, by='ID_C', all=F)%>%
  column_to_rownames(., var='ID_C')
lajit <- d[,8:104]
lajit <- lajit + 1
pelko <- d[,4]

rmse_1 <- NULL
mae_1 <- NULL
rmse_06 <- NULL
mae_06 <- NULL
rmse_01 <- NULL
```

```
mae_01 <- NULL
cor_1 <- NULL
cor_06 <- NULL
cor_01 <- NULL

# cv folds=5
for(i in 1:5){
alku = 55/5*i-10
loppu = 55/5*i
testIndex <- c(alku:loppu)
print(testIndex)
dTrain <- lajit[-testIndex, ]
yTrain <- pelko[-testIndex]

dTest <- lajit[testIndex, ]
yTest <- pelko[testIndex]

# train the model
tf$random$set_seed(0)
set.seed(0)
model_1 = codacore(
  dTrain,
  yTrain,
  logRatioType = 'balances',
  lambda = 1,
  overlap = T
)

tf$random$set_seed(0)
set.seed(0)
model_06 = codacore(
  dTrain,
  yTrain,
  logRatioType = 'balances',
  lambda = 0.6,
  overlap = T
)

tf$random$set_seed(0)
set.seed(0)
model_01 = codacore(
  dTrain,
  yTrain,
  logRatioType = 'balances',
  lambda = 0.1,
  overlap = T
)

# ennustaminen
```

```
if(getNumLogRatios(model_1) == 0){
  rmse_1[i] <- NA
  mae_1[i] <- NA
  cor_1[i] <- NA
}
else{
#asLogits = TRUE kun jatkuva!
  yHat_1 = predict(model_1, dTest, asLogits = TRUE, numLogRatios = 1)

  rmse_1[i] = sqrt(mean((yTest - yHat_1)^2))
# print(rmse_1[i])
  mae_1[i] = sum(abs(yTest - yHat_1))/length(yTest)
# print(mae_1[i])
  cor_1[i] = cor(yTest, yHat_1, method = "spearman")
# print(cor_1)
}
if(getNumLogRatios(model_06) == 0){
  rmse_06[i] <- NA
  mae_06[i] <- NA
  cor_06[i] <- NA
}
else{
  yHat_06 = predict(model_06, dTest, asLogits = TRUE, numLogRatios = 1)

  rmse_06[i] = sqrt(mean((yTest - yHat_06)^2))
# print(rmse_06[i])
  mae_06[i] = sum(abs(yTest - yHat_06))/length(yTest)
# print(mae_06[i])
  cor_06[i] = cor(yTest, yHat_06, method = "spearman")
# print(cor_06)
}
if(getNumLogRatios(model_01) == 0){
  rmse_01[i] <- NA
  mae_01[i] <- NA
  cor_01[i] <- NA
}
else{
  yHat_01 = predict(model_01, dTest, asLogits = TRUE, numLogRatios = 1)

  rmse_01[i] = sqrt(mean((yTest - yHat_01)^2))
# print(rmse_01[i])
  mae_01[i] = sum(abs(yTest - yHat_01))/length(yTest)
# print(mae_01[i])
  cor_01[i] = cor(yTest, yHat_01, method = "spearman")
# print(cor_01)
}
}

mean(rmse_1, na.rm=T)
```

```
mean(mae_1, na.rm=T)
mean(rmse_06, na.rm=T)
mean(mae_06, na.rm=T)
mean(rmse_01, na.rm=T)
mean(mae_01, na.rm=T)
mean(cor_1, na.rm=T)
mean(cor_06, na.rm=T)
mean(cor_01, na.rm=T)


# yhdelmat vastaavasti kuin balanssit
# mallien muodostaminen koko aineistolla vastaavasti kuten binaarisella
    vasteella
# lineaarinen ja logistinen regressio muodostetuilla log-suhteilla.
```