Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Johann Saavaste

Isheemiliste südamehaiguste markerite leidmine Tartu Ülikooli Eesti Geenivaramu metaboloomika andmetele geneetilist algoritmi rakendades

Matemaatiline Statistika

Bakalaureusetöö (9 EAP)

Juhendajad:

PhD Jaanika Kronberg

PhD Krista Fischer

Tartu 2023

ISHEEMILISTE SÜDAMEHAIGUSTE MARKERITE LEIDMINE TARTU ÜLIKOOLI EESTI GEENIVARAMU METABOLOOMIKA ANDMETELE GENEETILIST ALGORITMI RAKENDADES

Bakalaureusetöö

Johann Saavaste

Lühikokkuvõte

Metaboloomika teadusharu uurib organismi ainevahetuse vahe- ja lõppsaadustena tekkinud väikesi molekule ehk metaboliite. Metaboliidi kontsentratsiooni kõrvalekalded võivad olla indikaatoriks haiguse tekkele. Igal aastal sureb maailmas enim inimesi isheemiliste südamehaiguste tagajärjel. Käesoleva bakalaureusetöö eesmärk on leida metaboliite, mis sobiksid isheemilisi südametõbesid kirjeldavateks markeriteks. Töös leitakse geneetilise algoritmiga kõige suurema seosega metaboliidid. Hinnatakse 2 logistilist regressioonimudelit, mille tunnustena võeti arvesse väljavalitud metaboliidid, sugu, vanus, kehamassiindeks ning elupaik.

CERCS teaduseriala: P160 Statistika, operatsioonianalüüs, programmeerimine, finants- ja kindlustusmatemaatika; B110 bioinformaatika, meditsiiniinformaatika, biomatemaatika, biomeetrika.

Märksõnad: Metaboloomika, geneetiline algoritm, lähim tsentroid, logistiline regressioon.

FINDING MARKERS FOR ISCHEMIC HEART DISEASES USING GENETIC ALGORITHM ON ESTONIAN GENOME CENTER'S DATA

Bachelor thesis

Johann Saavaste

Abstract

Metabolomics studies small molecules, also known as metabolites, which are intermediates or products of metabolism. Deviations of metabolite concentration could indicate development of a disease. Ischemic heart disease is the highest cause of death in the world each year. The aim of this bachelor's thesis is to find metabolites that would fit as descriptive markers of ischemic heart diseases. The metabolites with the largest association are found in the thesis using genetic algorithm. 2 logistic regression models are evaluated using chosen metabolites, sex, age, body mass index and place of residence as predictors.

CERCS research specialisation: P160 Statistics, operations research, programming, financial and actuarial mathematics; B110 bioinformatics, medical informatics, biomathematics, biometrics.

Key Words: Metabolomics, genetic algorithm, nearest centroid, logistic regression.

Sisukord

Sissejuhatus

Metaboloomika on teadusharu, mis keskendub metaboliitide uurimisele organismis. Metaboliidid tekivad metabolismi ehk organismi ainevahetuse vahe- või lõppsaadusena. Haiguste ennetamiseks ja raviks uuritakse organismi metaboliitide kontsentratsioone tuvastamaks haiguse tekkeprotsessi juures olulisi metaboolseid muutuseid. Kui fikseerida metaboliitide kontsentratsioonide normid, on võimalik normilt kõrvalekaldumiste korral tuvastada muutusi organismis ning varakult haigusele jaole jõuda.

Kõige enam inimesi sureb maailmas südame isheemiatõbede tagajärjel. Maailma Terviseorganisatsiooni andmetel suri 2019. aastal haigusesse umbes 9 miljonit inimest, mis moodustas 16% kõikidest samal aastal toimunud surmajuhtumitest (Maailma Terviseorganisatsioon, 2020). Südame isheemiatõvesse haigestutakse, kui südamearterite kitsenemise tagajärjel jõuab südamelihastesse vähem verd ja hapnikku.

Käesoleva bakalaureusetöö eesmärk on leida Tartu Ülikooli Eesti Geenivaramu andmete põhjal leida metaboliidid, mis potentsiaalselt sobiksid isheemilisi südamehaigusi kirjeldavateks markeriteks.

Töö käigus rakendatakse metaboliitide leidmiseks geneetilist algoritmi 3 erineva klassifikatsioonimeetodiga. Haiguse hindamiseks logistilise regressioonimudeliga võetakse lisaks leitud metaboliitidele arvesse geenidoonori sugu, vanust, kehamassiindeksit ning elukohta.

Töö kirjutati tekstitöötlusprogrammiga LaTeX ning statistiline analüüs viidi läbi rakendustarkvaras R.

1 Metaboliidid

Metaboliidid on väikese molekulmassiga molekulid, mis saadakse organismi ainevahetuse vahe- või lõppsaadusena. Metaboloomikas tuvastatakse süstemaatiliselt organismi metaboliite ning mõõdetakse nende kontsentratsiooni. Kaks peamist tehnoloogiat metaboliitide tuvastamiseks on tuumamagnetresonantsspektroskoopia (inglise keeles *nuclear magnetic resonance spectroscopy*, NMR) ning massspektromeetria (MS) (Idle ja Gonzalez, 2007).

Mõlemal tehnoloogial on omad eelised ja puudused. MS võimaldab meetodist sõltuvalt tuvastada üle 1000 metaboliidi, seevastu NMR-ga tuvastatavate metaboliitide arv jääb alla 100. MS on rahaliselt kulukam, kuna preparaadi ettevalmistus on keerukam ning analüüsiks kulub rohkem aega (Euroopa Bioinformaatika instituut, 2023).

Lipiidide alla kuuluvad erineva struktuuriga hüdrofoobsed biomolekulid nagu triglütseriidid ehk rasvad, fosfolipiidid ning steroidid. Triglütseriidid koosnevad glütseroolist ning 3 rasvhappest. Neis talletatakse organismi energiavaru ning lagundatakse rasvlahustuvaid vitamiine. Fosfolipiididel on üks rasvhappega side asendatud fosfaatgrupiga. Fosfaatgrupi hüdrofiilsuse tõttu leidub fosfolipiide rakumembraani koostises, kaitstes raku sisu teatud ehitusega molekulide eest. Steroidide hulgas on tähtsal kohal kolesterool, millest sünteesitakse teisi steroidhormoone näiteks östrogeen, testosteroon ja kortisool (Ahmed, Shah ja Ahmed, 2023).

Ksenobiootikumideks nimetatakse kemikaale, mida inimorganism ei sünteesi ja mis satuvad organismi väliskeskkonnast. Need jõuavad organismi üldjuhul toiduga. Peamiselt on tegu taimset päritolu metaboliitidega. Osa ksenobiootikume on tööstuslikul viisil sünteesitud ja reoveega levinud looduskeskkonda. Ksenobiootikumide hulka kuuluvad näiteks hügieenitarbed, ravimid, pestitsiidid ning tööstusjäätmed (Štefanac, Grgas ja Landeka Dragičević, 2021).

1.1 Südame isheemiatõved

RHK-10 süsteemi kuuluvaid I20-I25 haiguseid nimetatakse südame isheemiatõbedeks (edaspidi ISH). Sinna kuuluvad stenokardia ehk rinnaangiin, müokardiiniinfarkti erinevad juhud ning kroonilised südame isheemiatõved nagu ateroskleroootiline südamehaigus (Maailma Terviseorganisatsioon, 2016). ISH on põhjustatud südamearterite kitsenemisest, mille tõttu südamelihastesse jõuab vähem verd ning hapnikku. Rinnas tekkiv valu võib viia südameinfarktini (Ameerika Südameassotsiatsioon, 2022).

Südamearterid kitsenevad, kui arteri siseseintele ladestuvad rasvad ning kolesteroolijäägid, mis aja möödudes kõvastuvad. Varasem uuring on näidanud, et kõrgem kolesteroolijääkide kontsentratsioon suurendab ISH riski. Uuringus analüüsiti 74 000 Kopenhaageni elaniku kolesterooli taset veres, kellest 11 000 oli ISH diagnoositud ajavahemikus 1976–2010 (Varbo *et al.*, 2013).

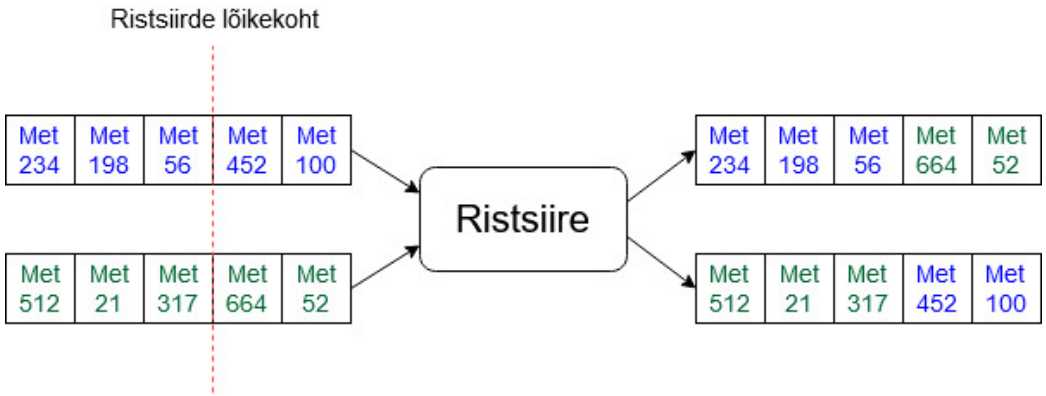
Oluliseks riskifaktoriks ISH esinemisele on leitud ka diabeedi olemasolu. Kopenhaageni linnaosa vähemalt 20-aastastest elanikest moodustati 13 000 pealine valim ning jagati teist tüüpi diabeedi järgi haigus-, soodumustega grupiks ning kontrollgrupiks. 20 aastat hiljem uuriti ISH esinemist valimis sh ka surmaga lõppenud juhtumid. Leiti, et teist tüüpi diabeetikutel on 2 kuni 3 korda suurem risk ISH-sse haigestuda. Diabeetikutest naiste suhteline haigestumise risk oli diabeedihaigete meeste omast oluliselt kõrgem (Almdal *et al.*, 2004).

2 Geneetiline algoritm

Geneetiline algoritm (edaspidi GA) on optimiseerimisalgoritm, mis tugineb Charles Darwini kirjeldatud looduslikule valikule. Sellega leitakse suure populatsiooniga kogumitest kõige parema sobivusega kombinatsioone ilma kõikide variantide ajamahuka läbivaatamiseta. Algoritmis kasutatavad mõisted on üle võetud bioloogiast ja geneetikast ning järgnevates lõikudes ei tasu neid seostada nimetatud valdkondadega (Mitchell, 1998).

Järgnevalt kirjeldatakse põlvkonna protsessi.

Kõigepealt luuakse juhusliku protsessi tulemusel metaboliitide kogumite ehk kromosoomide populatsioon. Defineeritakse sobivusfunktsioon ning arvutatakse iga kromosoomi sobivus. Seejärel valitakse kaks kõige suurema sobivusega kromosoomi. Valitud kromosoomidel teostatakse ristsiire – mõlemal jagatakse metaboliitide järjestused i-nda liikme järel kaheks osaks ning kromosoomid vahetavad omavahel ühe osa. Saadud järglastel vahetatakse osa metaboliite mõne teise vastu ehk toimub mutatsioon. Uue populatsiooniga arvutatakse taas iga kromosoomi sobivus (Katobh, Chauhan ja Kumar, 2021).



Joonis 1: Ristsiirde skeem

2.1 Rakendustarkvara R pakett GALGO

Alapeatükk on kirjutatud (Trevino ja Falciani, 2006) ning (Trevino ja Falciani, 2018) põhjal.

Programmeerimiskeeles R on GA rakendamiseks loodud pakett GALGO, millega on võimalik luua statistilisi mudeleid suuremõõtmelistele andmetele. GALGO-s luuakse *BigBang* objekt, milles rakendatakse GA protsessi mitme kromosoomi populatsiooniga ehk nišiga. Nišid võivad omavahel juhuslikult kromosoome vahetada. Kromosoomi sobivuse hindamiseks luuakse treeningandmestik, mille põhjal leitakse prognoos testandmestikule sõltuvalt klassifitseerimismeetodist. Prognoosi täpsuse põhjal arvutatakse kromosoomi sobivus. Kui leitakse kromosoom, mille sobivus ületab eesmärgile vastavat sobivust, siis see valitakse välja ning lõpetatakse uute põlvkondade loomine. Väljavalitud soovitud sobivusega kromosoomi nimetatakse lahendiks. Perioodi esimese põlvkonna ning lahendi leidmise vahel nimetatakse evolutsiooniks. Vaikimisi sisaldab evolutsioon 200 põlvkonda. Kui evolutsiooni kestel lahendini ei jõuta, siis valitakse välja viimase põlvkonna parim kromosoom.

Kui programm on jooksutanud läbi sadu evolutsioone, siis viiakse läbi ettesuunatud valik (inglise keeles *forward selection*). Selleks järjestatakse metaboliidid kahanevas järjekorras vastavalt sellele, mitmes kromosoomis see evolutsiooni lõpus sisaldus. Seejärel, alustades esialgu ainult järjekorras esimesest metaboliidist, luuakse mudel prognoosimaks haigusgruppi. Leitakse mudeli klassifitseerimise täpsus ning mudelisse lisatakse järjekorras järgmine metaboliit. Protsessi korratakse senikaua, kuni mudelisse on lisatud kõik metaboliidid. Parimaks mudeliks valitakse mudel, milles haigusgrupi klassifitseerimise täpsus on maksimaalne ning metaboliitide arv minimaalne.

2.2 Klassifitseerimismeetodid

Töös rakendatakse andmestikule geneetilist algoritmi programmeerismiskeeles R paketi GALGO abil 3 erineva klassifitseerimismeetodiga ja võrreldakse saadud tulemusi ning nende efektiivsust. Olgu treeningandmestik X, mis koosneb n paarist $(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$, kus $\vec{x}_i$ tähistab ühele vaatlusele vastavat m-mõõtmelist vektorit ning y_i vaatluse klassi. Olgu $y_i \in Y$ võimalike klasside hulk.

2.2.1 Lähim tsentroid

Alapeatükk on kirjutatud (Levner, 2005) põhjal.

Lähima tsentroidi (inglise keeles *nearest centroid*) meetodil arvutatakse treeningandmetes igale klassile selle vektorite aritmeetiline keskmine ehk tsentroid. Testandmete abil leitakse tundmatu klassiga vektori kaugused kõikidele klassidele vastavatest tsentroididest. Prognoositud klassiks valitakse vähima kaugusega tsentroidile vastav klass.

Klassile k vastav tsentroid leitakse valemiga

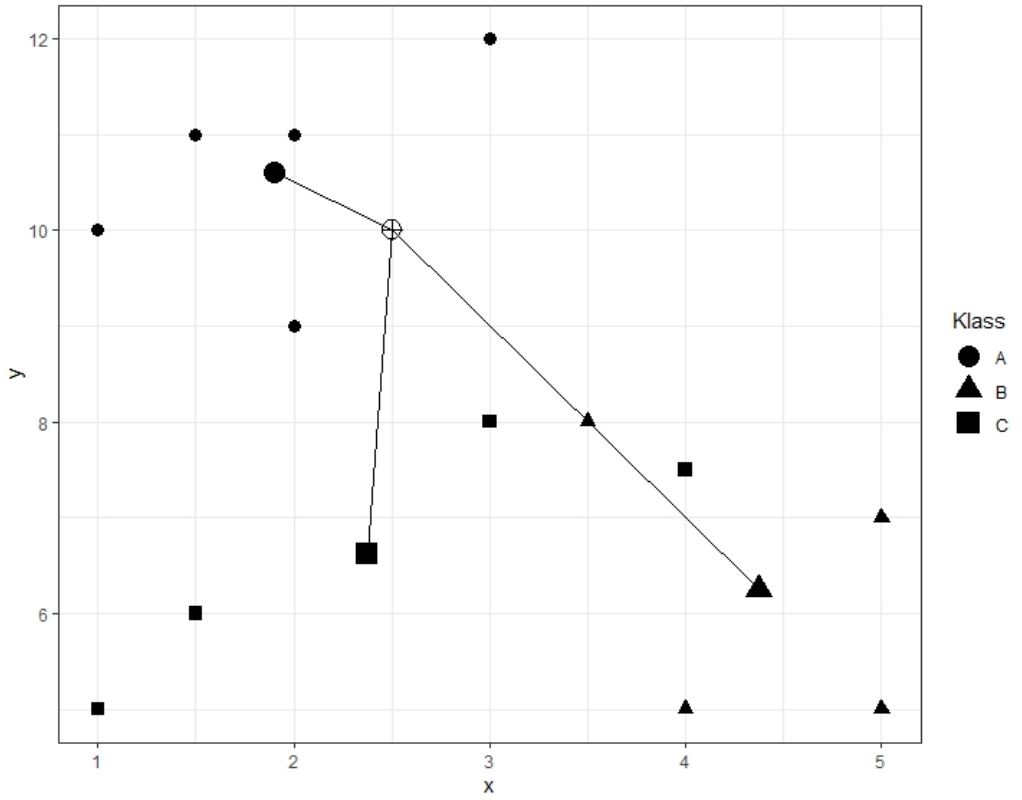
$$\vec{\mu}_k = \frac{1}{|D_k|} \sum_{y_i \in D_k} \vec{x_i},$$

kus D_k on klassile k vastavate andmete hulk.

Testandmetes vaatlusega $\vec{x}_j$ ning tundmatu klassiga y_j määratakse klass valemiga

$$\hat{y}_j = \arg\min_{\ell \in Y} \|\vec{\mu}_\ell - \vec{x}_j\| = \arg\min_{\ell \in Y} \sqrt{\sum_{i=1}^{m} (\mu_{\ell i} - x_{ji})^2}.$$

Joonisel 2 on kahemõõtmelised vaatlused kujutatud väiksemate kujunditena. Silmnähtavalt suurte kujunditega on kujutatud vaatluste klassidele vastavate tsentroidide asukohad. Kui hinnata tundmatu vaatluse klassi, mida joonisel on kujutatud ringjoonega ümbritsetud ristina, siis sellele kõige lähemal asub klassi A tsentroid.



Joonis 2: Lähim tsentroid kahemõõtmelises ruumis. Ringjoonega ümbritsetud ristiga tähistatud tundmatu klassiga vaatluse klassifitseerimisel valitakse klass A.

2.2.2 Tugivektor-masin

Tugivektor-masina (inglise keeles *support vector machine*, edaspidi TVM) meetodil paigutatakse treeningandmed m-mõõtmelisse ruumi punktidena. Sobiva tuumafunktsiooni (inglise keeles *kernel function*) korral asetsevad punktid nii, et klasse saab ruumis eraldada hüperpinnaga. Hüperpind valitakse selline, et mõlema klassi punktid asetseksid sellest maksimaalsel kaugusel. Testandmed paigutatakse valitud hüperpinnaga samasse ruumi ning klassi prognoosimisel vaadeldakse punkti asukohta hüperpinna suhtes (Cervantes *et al.*, 2020).

2.2.3 Juhuslik mets

Juhusliku metsa (inglise keeles *random forest*) meetodi korral luuakse treeningandmetest mitu juhuslikku tagasipanekuga valimit. Iga valimi põhjal rajatakse otsustuspuu, mis muutujate omadusi arvestades määrab sellele vastava klassi. Testandmetega klassi prognoosimisel valitakse maksimaalne klassi määranud otsustuspuude arv. Juhusliku metsa eeliseks on tema madal tundlikkus üleprognoosimisele (Belgiu ja Drăguţ, 2016).

2.3 Varasemad uuringud

Metaboloomika andmetest püütakse leida biomarkereid erinevatele haigustele. Näiteks on leitud GALGO abil biomarkerid teist tüüpi diabeedile. Uuringus analüüsiti 41–63-aastaste inimeste plasmaproovide mass-spektromeetria metaboloomika andmeid. 80 inimest jaotati võrdselt 4 gruppi, millest esimesse kuulusid diabeedi eelsoodumusega isikud, teise teist tüüpi diagnoosiga isikud, kolmandasse diabeetilise nefropaasi haigusega isikud ning neljandaks oli kontrollgrupp. Metaboliitidest biomarkeid prognoosivate mudelite koostamiseks kasutati järgnevaid klassifitseerimismeetodeid: k-lähimad naabrid, lähim tsentroid ning TVM meetod. Ettesuunatud valiku ning ristvalideerimise käigus leiti teist tüüpi diabeedi biomarkeriteks 5 metaboliiti (Morgan-Benita *et al.*, 2022). Lisaks on GALGO-t kasutatud rottide maksahaigustega seotud oluliste metaboliitide leidmiseks (Lin *et al.*, 2011).

GA on varasemalt kasutatud mitmetes protsessides, kus traditsioonilise algoritmiga jäädi hätta. Näiteks hinnati GA abil matemaatiline mudel glükoosi metabolismile, mille täpsus oli oluliselt parem traditsioonilisel meetodil saadud mudelist (Morbiducci, Andrea Tura ja Grigioni, 2005).

3 Logistiline regressioon

Peatükk on kirjutatud (Nick ja Campbell, 2007) põhjal.

Olgu uuritav tunnus Y binaarne tunnus, mille väärtused on kodeeritud järgmiselt: 1 – uuritav sündmus toimus, 0 – uuritav sündmus ei toimunud. Logistiline regressiooni meetodil mudeldatakse sündmuse toimumise šansi logaritmi. Mudeli üldkuju on järgmine:

$$\ln\left(\frac{P(Y=1 \mid X_1,\ldots,X_k)}{1-P(Y=1 \mid X_1,\ldots,X_k)}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k,$$

kus $X_1,\ldots,X_k$ tähistavad mudeli argumenttunnuseid, β_0 mudeli vabaliiget ning $\beta_1,\ldots,\beta_k$ argumenttunnuste kordajaid.

Mudeli valemist on võimalik avaldada valem sündmuse toimumise tõenäosuse arvutamiseks.

$$P(Y=1 \mid X_1,\ldots,X_k) = \frac{\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}{1+\exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}$$

4 Andmestik

Töös kasutatakse Eesti Geenivaramu aastatel 2002-2019 kogutud plasmaproovide mass-spektromeetria metaboloomika andmeid. Andmete kasutamine oli Eesti bioeetika ja inimuuringute nõukogu loa 234T-12 "*Omics for Health*" raames. Andmed väljastati Eesti Geenivaramust andmeväljastuse nr R26 alusel. Iga rida vastas ühele mõõdetud metaboliidi kontsentratsioonile. Veergudeks olid 999 doonorit. Metaboliidid on eristatud järjekorranumbri ning HMDB koodi kujul, mille kohta on võimalik teha päring Inimese Metaboliitide Andmebaasis (inglise keeles *the Human Metabolome Database*).

Töös vaadeldakse haigusgrupina doonoreid, kellel on pärast proovi andmist RHK-10 süsteemis diagnoositud vähemalt üks isheemilistest südamehaigustest. Töö andmebaasidega teostas juhendaja Jaanika Kronberg, kes andis autorile geenidoonorite klassifikatsiooni kujul 0/1/2, kus 0 on kontrollgrupp, 1 on isikud, kellel haigus tekkis pärast proovi andmist ja 2 on isikud, kellel oli haigus proovi andmise ajal diagnoositud. Doonoritest oli 69-l haigus diagnoositud peale proovi andmist, 138 doonoril enne proovi andmist ning ülejäänud 762 olid kontrollgrupis.

Esialgses andmestikus oli kokku mõõtmisi 1505 metaboliidi kohta. Moodustati 2 andmestikku, millest esimeses jäeti välja isikud, kel oli proovi andmisel haigus diagnoositud, ning teises isikud, kes haigestusid peale proovi andmist. Enam kui 1000 metaboliidil puudus vähemalt ühe isiku kohta mõõtmistulemus. Kuna TVM ja juhusliku metsa klassifitseerimismeetodid eeldavad, et andmed ei sisalda puuduvaid väärtusi, siis jäeti andmestikest välja metaboliidid, milles vähemalt 60% vaatlustest puudusid väärtused. Ülejäänud tühjad väljad asendati poolega vastava metaboliidi mõõtmistulemuste miinimumist. Andmestikus oli metaboliite, millel puudus HMDB kood, mistõttu eemaldati need samuti. Peale andmestiku puhastamist jäid alles 765 metaboliidi mõõtmistulemused.

Metaboliitidele lisaks kasutati andmeid doonorite tausta kohta nagu sugu ja vanus proovi andmise hetkel. Nende tunnuste põhjal leiti andmestikust igale isikule

haigusrühmast 2 temale vastavat kontrollgrupi liiget, millega saadud tulemusi töös eraldi vaadeldakse. Lisaks sisaldas andmestik teavet doonorite elupaiga (linn, väikelinn, maa) ja kehamassiindeksi kohta.

5 Klassifikatsioonimeetodite võrdlus

Haigusgrupina vaadeldi ainult isikuid, kellel diagnoositi ISH peale proovi andmist. Haiguse biomarkerite leidmiseks rakendati GA haigus- ja kontrollgrupi metaboloomika andmetele lähima tsentroidi, TVM ning juhusliku metsa klassifikatsioonimeetodiga. Protsess viidi läbi, kasutades rakendustarkvara R paketti GALGO (Trevino ja Falciani, 2018). Iga klassifikatsioonimeetodi peal prooviti rakendada samu parameetreid. Enamik parameetritest olid vaikeväärtustega, kuid sobivuseks määrati 80%, evolutsioonide arvuks 1000 ning kromosoomi suuruseks 10 metaboliiti. Tabelis 1 on välja toodud klassifikatsionimeetodite peamised erinevused töös kasutatavate andmete peal rakendamisel.

Tabel 1: Klassifikatsioonimeetodite võrdlus töös kasutatavate andmete põhjal

	Lähim tsentroid	**TVM**	**Juhuslik mets**
Implementatsioon	Programmeerimiskeeles C	Autori modifitseeritud R-i kood	Autori modifitseeritud R-i kood
Põlvkonna kestus	0,2 s	1 s	20–30 s
Gruppide kaalud	kaalud haigus- ja kontrollgrupi osakaalude põhjal	haigusgrupil rohkem kaalu	haigusgrupil rohkem kaalu
Parim sobivus evolutsiooni lõpus (kuni 200 põlvkonda)	80%	20–40% (sõltub tuumameetodist)	10–25%

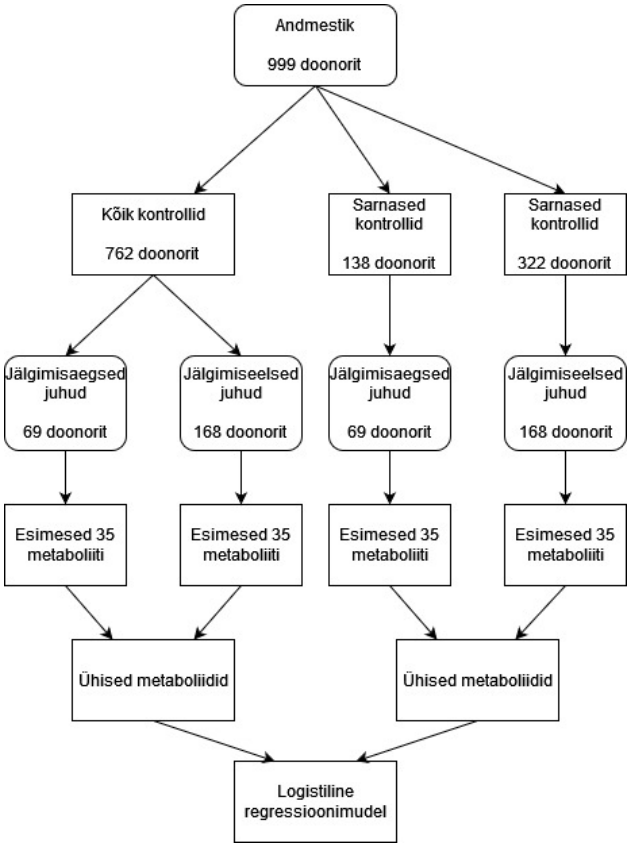
Kuna haigusgrupi osakaal andmestikus oli madal, siis igale klassifikatsioonimeetodile prooviti treening- ja testandmestiku loomisel arvestada ka haigusgrupi kaaluga. Vastasel juhul oleks prognoositud õigesti valdavalt kontrollgruppi. GALGO-s on lähima tsentroidi meetod kiiruse huvides implementeeritud programmeerimiskeeles

C, mistõttu autoril ei õnnestunud kirjutada koodi R-s, mis arvestaks haigus- ja kontrollgrupi kaale (Trevino ja Falciani, 2018). TVM ja juhusliku metsa meetodil õnnestus muuta koodi selliselt, et treening- ja testandmestikel oleks piisavalt haigusgrupi isikuid.

Paraku osutusid TVM ja juhusliku metsa meetodi rakendamine väga ajamahukaks. Suur ajakulu oli tingitud kromosoomide madalast sobivusest ning põlvkonna kestusest. Põlvkonna kestuses mängib rolli andmete analüüsiks kasutatav riistvara, milleks antud juhul oli autori isiklik sülearvuti. Kui üks põlvkond kestab TVM meetodil keskmiselt 1 sekundi, siis järgmise evolutsioonini jõutakse maksimaalselt 200 sekundiga. 1000 evolutsiooni korral oleks protsessi maksimaalne ajakulu olnud üle 50 tunni. Juhusliku metsa meetodil kestaks protsess kauemini, kuna põlvkonna ajakulu oli 20–30 sekundit. Protsessi kiirendamiseks prooviti alandada sobivust 50% peale ning vähendada evolutsioonis põlvkondade arvu 100 peale. Kuigi protsessi ajakulu vähenes, ei suurenud sellegipoolest lahendite arv. Lahendite saamiseks prooviti muuta tuumafunktsiooni parameetreid, kuid sellest hoolimata ei õnnestunud leida oluliselt rohkem lahendeid.

6 Tulemused

Töös vaadeldakse haigusgruppidena doonoreid, kel oli haigus diagnoositud peale proovi andmist ehk jälgimisaegseid juhte, ja doonoreid, kellel oli haigus diagnoositud proovi andmise ajaks ehk jälgimiseelseid juhte. Mõlemad haigusgrupid sisalduvad eraldi andmestikes. Analüüs koosneb kahest osast. Esimeses osas analüüsitakse 2 andmestikku, mis sisaldab kõiki kontrolle. Teises osas valitakse haigusgrupi isikutele sarnased kontrollid ning moodustatud 2 andmestikku analüüsitakse sarnaselt esimese osaga. Analüüsitavad metaboliidid olid kõikides andmestikes ühesugused. Analüüs teostati rakendustarkvaras R versiooniga 4.1.1, kasutades paketti GALGO versiooniga 1.4 (Trevino ja Falciani, 2018). Joonisel 3 on visualiseeritud töö käik.



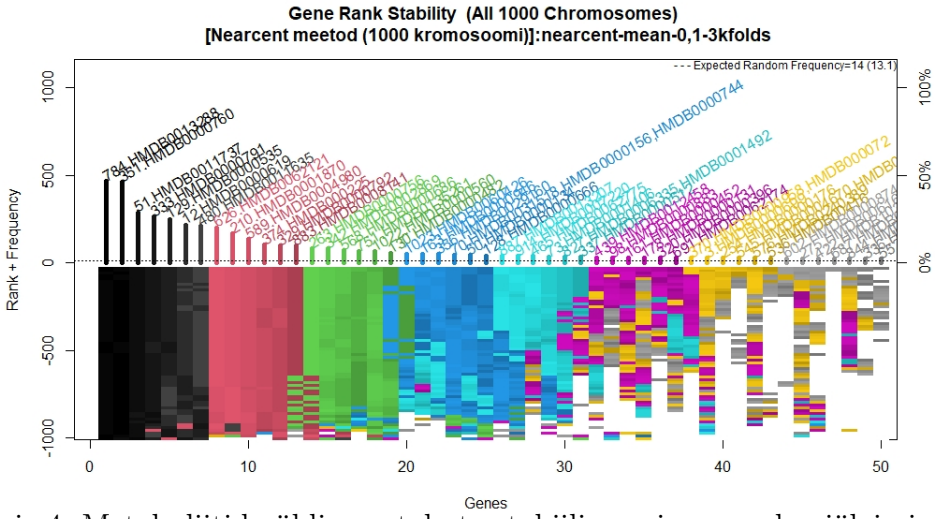
Joonis 3: Lihtsustatud analüüsi käik

Iga andmestiku tulemustega saab tutvuda lisa 1 tabelitest.

6.1 Haigusjuhtude võrdlus kõiki kontrollgrupi isikuid sisaldava andmestikuga

Mõlema andmestiku korral rakendati GALGO *BigBang* objektis ühesuguseid parameetreid, kus kromosoomi suuruseks määrati 10 metaboliiti, evolutsiooni pikkuseks 200 põlvkonda ning kokku viidi läbi 1000 evolutsiooni (Trevino ja Falciani, 2018). Ainukese parameetrina erines soovitud sobivus, kuna jälgimiseelsetel juhtudel ei õnnestunud enamiku evolutsioonide käigus saavutada sobivuseks 80%, mistõttu alandati seda 70% peale. Sellegipoolest jäeti jälgimisaegsete juhtude korral soovitud sobivus 80% peale.

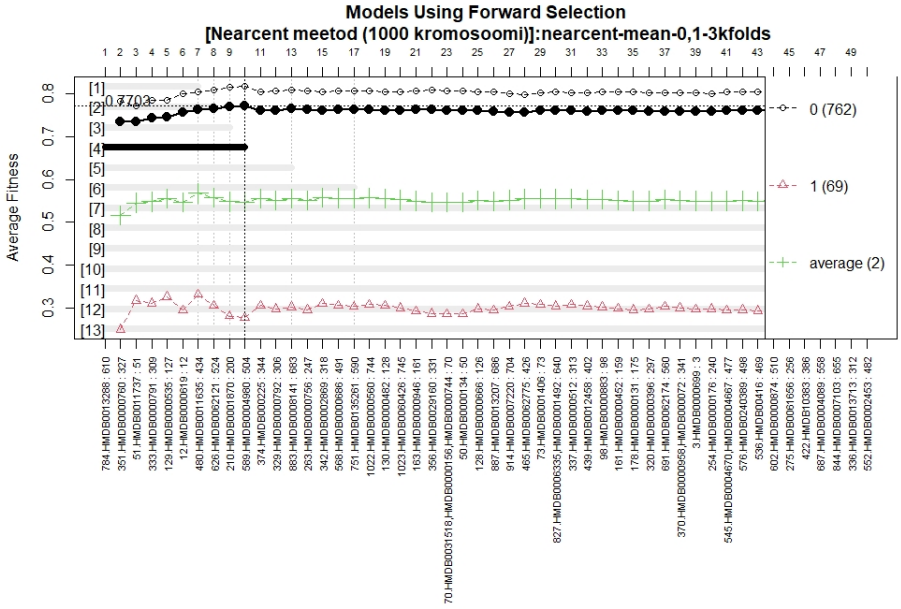
6.1.1 Jälgimisaegseid juhte prognoosivad metaboliidid



Joonis 4: Metaboliitide üldine astakute stabiilisus ning sagedus jälgimisaegseid juhte ja kõiki kontrolle sisaldava andmestiku põhjal

Tulemused andmestikus, kus haigusgruppi kuulusid jälgimisaegsete juhtudega isikud, on esitatud joonisel 4. Geenide nimetusena on kujutatud 50 kõige sagedamat metaboliiti kahanevas järjekorras, mis sisaldusid lahenditeks valitud kromosoomides. Vertikaalse telje ülemine osa kujutab metaboliitide sagedust. Värvidega on näidatud metaboliidi asukohta järjekorras eelneva evolutsiooni lõpus. Värvid kõige

alumises reas näitavad metaboliitide asukohta järjestuses peale esimest evolutsiooni. Mida ühevärvilisemalt on metaboliidid kujutatud, seda stabiilsemana on nende järjestus evolutsioonide käigus püsinud. (Trevino ja Falciani, 2018) On näha, et esimesel 10 metaboliidil on püsinud järjestus ligikaudu viimased 900 evolutsiooni muutumatuna. Kõige rohkem leidus lahendites 464 korda metaboliit HMDB0013288 ehk nonaüülkarnitiin, millele järgnes 462 korda esinenud HMDB0000760 ehk hüokoolhape (inglise keeles *Hyocholic acid*).

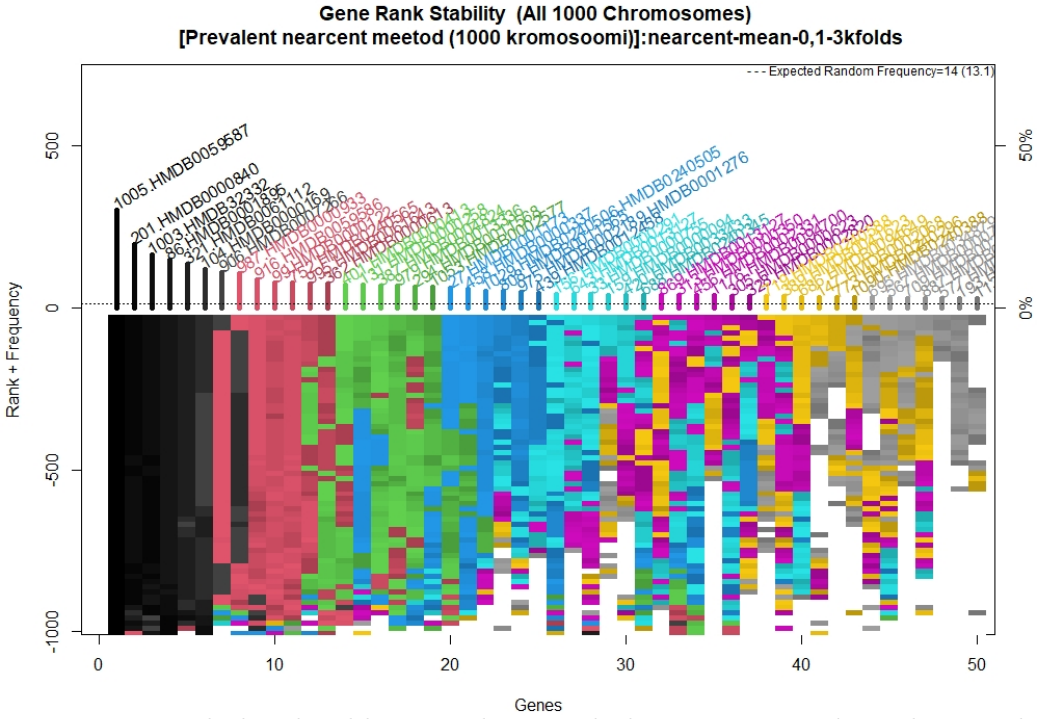


Joonis 5: Jälgimisaegsete juhtude metaboliitide järjestuse ettesuunatud valik

Joonisel 5 on abstsissteljel taas metaboliidid järjestatud sageduse järgi. Punaste kolmnurkadena on kujutatud jälgimisaegsete grupi ning musta ringjoonena kontrollgrupi klassifitseerimise täpsust. Mustade ringidena on kujutatud mudeli üldist täpsust ehk mõlema grupi õigesti klassifitseeritud vaatluste osakaalu kõikidest vaatlustest. Kõige parema täpsusega mudeli saaks GALGO hinnangul, kui mudelisse lisada kõik metaboliidid alates vasakult kuni katkendliku vertikaalse jooneni, kuna mustade ringidena tähistatud üldine täpsus on sel juhul kõige kõrgem (Trevino ja

Falciani, 2018). Antud juhul saaks kõige parema mudeli esimese 10 metaboliidiga. Kuna töö eesmärk on leida metaboliidid, mis prognoosivad hästi haigusgruppi, siis jooniselt 5 paistab, et kõige kõrgema täpsusega haigusgrupi prognoosiga mudeli saaks, kui mudelisse lisada esimesed 7 metaboliiti.

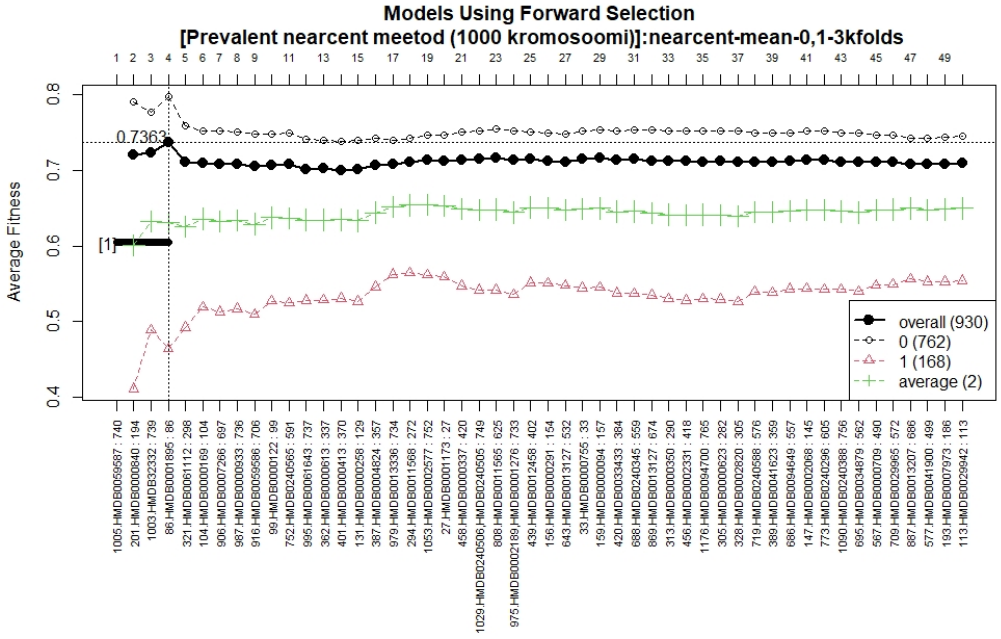
6.1.2 Jälgimiseelseid juhte prognoosivad metaboliidid



Joonis 6: Metaboliitide üldine astakute stabiilisus ning sagedus jälgimiseelsete juhte ja kõiki kontrolle sisasldava andmestiku põhjal

Joonisel 6 paistab, et esimese 10 metaboliidi järjestus on püsinud viimase 800 evolutsiooni jooksul enamvähem muutumatuna. Ainult sageduselt 7. ja 8. metaboliidid on vahepeal oma kohad vahetanud ning hiljem tagasi vanale kohale jäänud. Kõige rohkem valiti välja 300 korda metaboliit HMDB0059587 ehk perfluorooktaanhape (PFOA), millele järgnes 195 korda esinenud HMDB0000840 ehk salitsüluurhape.

Joonisel 7 on näha, et kõige parema üldise täpsusega mudeli saaks, kui mudeli argumentideks võtta esimesed 4 metaboliiti. Valitud mudeli üldine täpsus oleks



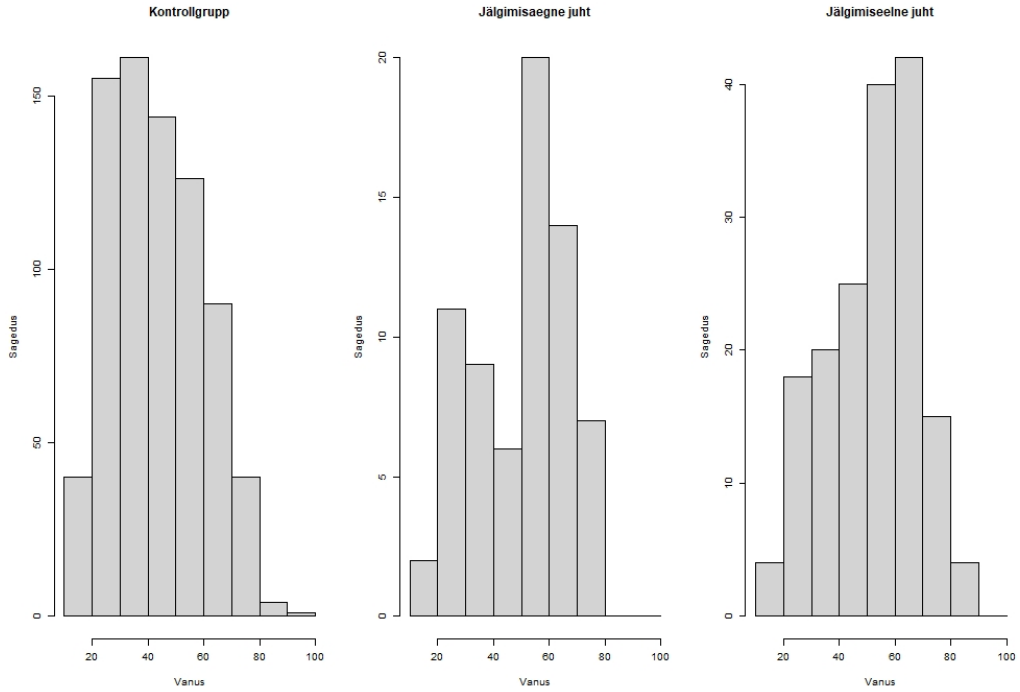
Joonis 7: Jälgimiseelsete juhtude metaboliitide järjestuse ettesuunatud valik

74%. Sellegipoolest on haigusgrupi prognoosimise täpsus selles mudelis madalam kui suuremal osal teistest mudelitest. Kõige parema täpsusega haigusgruppi prognoosiva mudeli saaks, kui võtta mudelisse esimesed 18 metaboliiti. Sellise mudeli haigusgrupi prognoosimise täpsus oleks joonise põhjal 56% ning üldine täpsus 71%, mis ei erine palju GALGO valitud mudeli täpsusest.

6.2 Haigusjuhtude võrdlus haigusgrupiga sarnaseid kontrolle sisaldava andmestikuga

Kontrollgrupiga võrreldes on haigusgrupp oluliselt väiksema osakaaluga. Seda eriti jälgimisaegsete juhtudega andmestikus, kus haigusgrupp moodustab alla 10% vaadeldavatest geenidoonoritest. Kui uurida gruppide vanuselist jaotust proovi andmise ajal joonisel 8, on näha, et nende jaotused ei ole ühesugused. Kontrollgruppi kuuluvad peamiselt alla 50-aastased. Seevastu ISH diagnoosiga isikud olid peami-

selt üle 50 aasta vanused. Haigusgrupi isikutele vastavate kontrollidega saaks teha paremaid järeldusi.



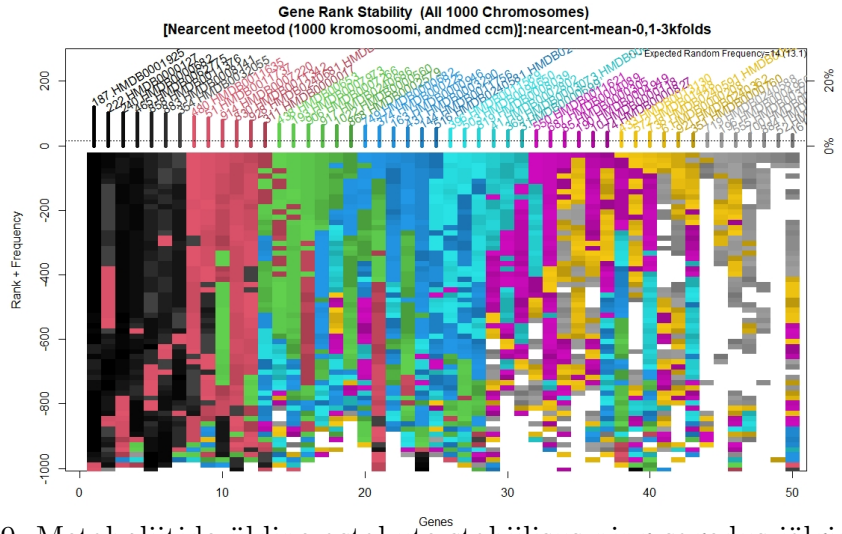
Joonis 8: Kontrollgrupi, jälgimisaegsete ja -eelsete juhtude vanuste jaotused proovi andmise ajal

Autori kirjutatud R-i skriptiga valiti igale haigusgrupi isikule kontrollgrupi isikute hulgast välja 2 temaga kõige sarnasemat. Kontroll pidi olema haigusgrupi isikuga samast soost ning mõlema vanusevahe ei tohtinud ületada 2 aastat. Kui ühele haigusgrupi isikule vastas rohkem kui 2 kontrollgrupi isikut, siis valiti nende hulgast 2 välja juhuslikult. Saadi 2 andmestikku, millest ühes olid jälgimisaegsete juhtudega isikud ning nendega sarnased kontrollid ja jälgimiseelsete juhtudega isikud ning nendega sarnased kontrollid. Kõigile 168 jälgimiseelsete juhtudega isikutele ei õnnestunud leida 2 sarnast kontrolli, mistõttu valiti välja vaid 322 kontrolli. Seega kontrollgrupi isikute osakaal jälgimisaegsete juhtudega andmestikus on suurem kui jälgimiseelsete juhtudega andmestikus.

Mõlemal andmestikul rakendati GALGO-ga GA protsessi ühesuguste parameetrite-

ga. Kokku tehti 1000 evolutsiooni, igas evolutsioonis maksimaalselt 200 põlvkonda. Iga kromosoom sisaldas 10 metaboliiti ning soovitud sobivuseks valiti 70%.

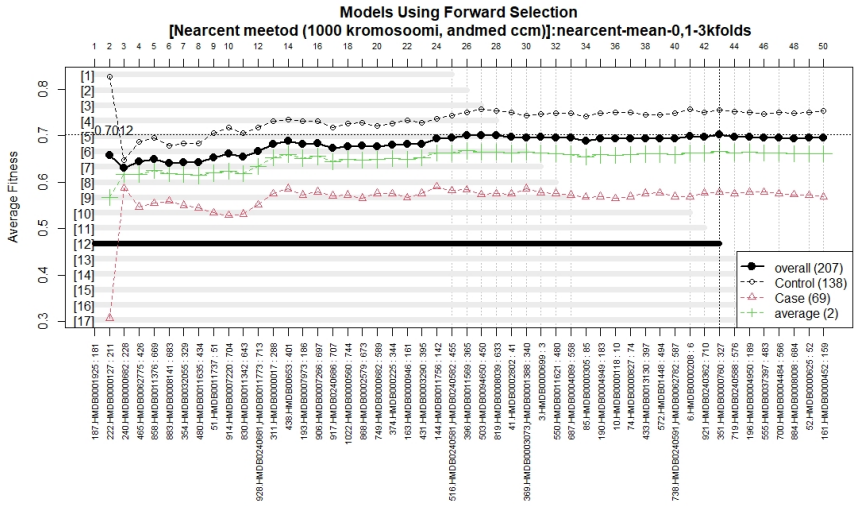
6.2.1 Jälgimisaegseid juhte prognoosivad metaboliidid



Joonis 9: Metaboliitide üldine astakute stabiilisus ning sagedus jälgimisaegseid juhte ning sarnaseid kontrolle sisaldava andmestiku põhjal

Joonisel 9 paistab, et esimese 10 metaboliidi järjestus on olnud kõigi 1000 evolutsiooni käigus pidevas muutumises. Graafikult paistab, et lõpliku järjestuseni on jõutud viimase 100 evolutsiooni käigus. Sellele viitab ka madal vahe metaboliitide sageduste arvude vahel. Kõige sagedamat metaboliiti HMDB0001925 ehk ibuprofeeni esines 120 korda, millele järgnes 102 korda esinenud HMDB0000127 ehk glükuroonhape.

Joonisel 10 on näha, et kõige parema üldise täpsusega 70% on mudel, millesse lisada esimesed 43 metaboliiti. Kõige parema täpsusega prognoosib haigusgruppi esimese 3 metaboliidiga mudel ligikaudu 60% täpsusega. Sellegipoolest mudeli üldine täpsus 63% on teiste mudelitega võrreldes madalaim. Jooniselt paistab, et 14 metaboliidiga mudel prognoosib haigusgruppi sarnase täpsusega, ent selle üldine täpsus on lähemal GALGO poolt valitud mudeli täpsusele.

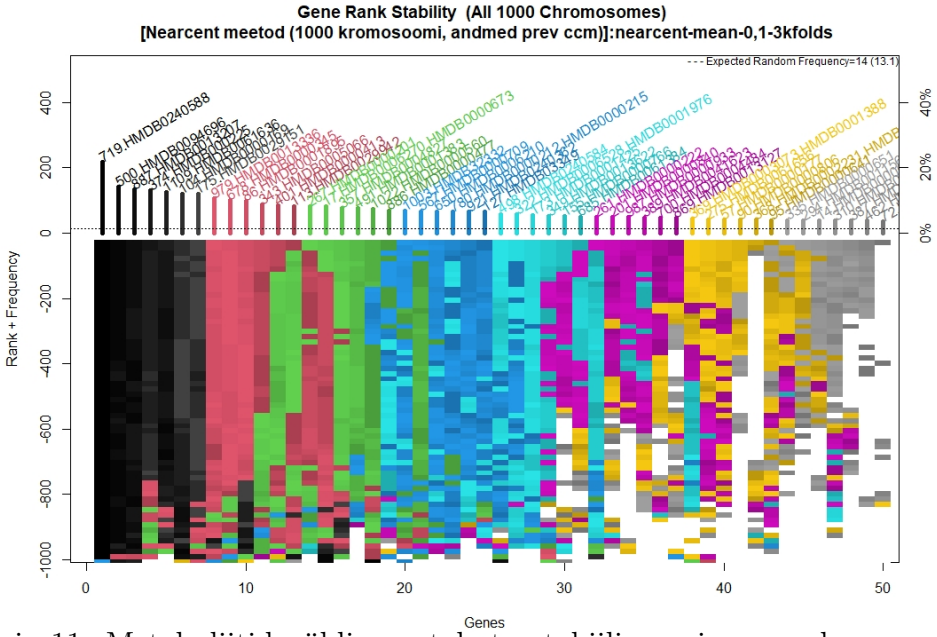


Joonis 10: Sarnaste kontrollidega jälgimisaegsete juhtude metaboliitide järjestuse ettesuunatud valik

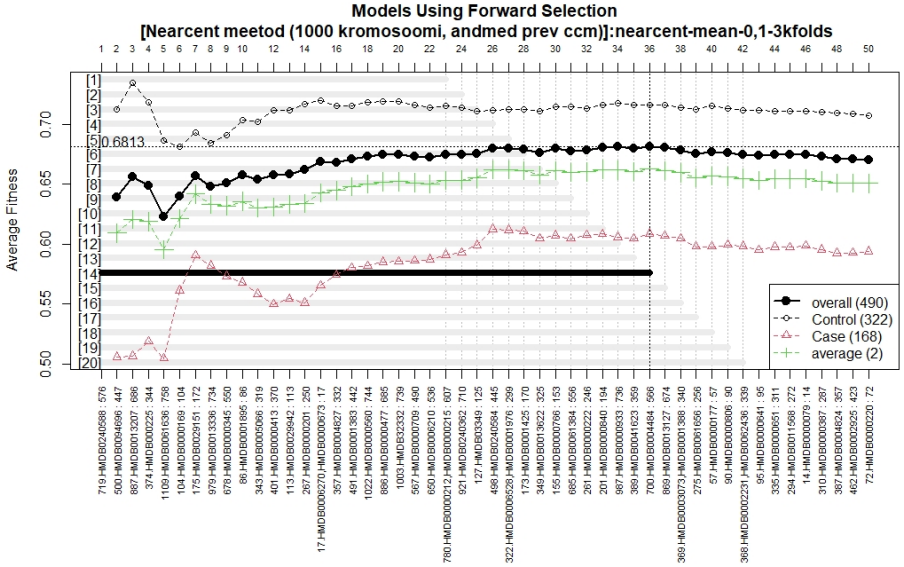
6.2.2 Jälgimiseelseid juhte prognoosivad metaboliidid

Joonisel 11 on näha, et esimese 10 metaboliidi järjestus on püsinud enam-vähem muutumatuna viimased 700 evolutsiooni. Kuigi need metaboliidid on mitmeid kordi vahepeal omavahel järjestust vahetanud, ei paista, et mõni metaboliit tagapool oleks esikümne hulka jõudnud. Kirjud värvid esimeste metaboliitide all viitavad sellele, et metaboliidid asusid esimese 200 evolutsiooni järel tagapool. Kõige sagedamini valiti 217 korda metaboliiti HMDB0240588 ehk müristoleoüülkarnitiin (inglise keeles *myristoleoylcarnitine*) ning teisel kohal on 141 korral esinenud HMDB0094696 ehk N-metüül-L-proliin (inglise keeles *N-methyl-L-proline*).

Kui vaadata joonist 12, siis on näha, et kõige parema üldise täpsusega 68% mudeli saab, kui mudeli argumentideks võtta esimesed 36 metaboliiti. Võttes mudeli argumentideks 26 metaboliiti, oleks mudeli haigusgrupi prognoosi täpsus natuke suurem ning üldine täpsus sarnane GALGO valitud mudeliga võrreldes. Sarnanesed täpsused oleksid ka juhul, kui võtta mudeli argumentideks esimesed 27 metaboliiti.



Joonis 11: Metaboliitide üldine astakute stabiilisus ning sagedus sarnaste kontrollidega jälgimiseelsete juhtude põhjal



Joonis 12: Sarnaste kontrollidega jälgimiseelsete juhtude metaboliitide järjestuse ettesuunatud valik

6.3 Logistiline regressioonimudel

Autor jooksutas GALGO-ga andmeid mitu korda läbi ning sooritas saadud metaboliitide järjestustele ettesuunatuid valikuid. Ettesuunatud valikutes tulid parima täpsusega mudelid metaboliitide arvuga 1 kuni 43. Optimaalseks metaboliitide arvuks mudelisse valiti 35, mille põhjal leiti sõltuvalt kontrollgrupi isikute arvust esimese 35 metaboliidi hulgast ühised metaboliidid jälgimisaegsete ja -eelsete juhtude vahel. Kokku saadi 4 metaboliiti, millest 1 leiti kõikide kontrollide põhjal ja ülejäänud sarnaste kontrollidega andmestikust. Metaboliitide nimetused ja järjekorranumbrid vastava haigusgrupi andmestiku kohta on välja toodud tabelis 2.

Tabel 2: Jälgimisaegsete ja -eelsete juhtude ühised metaboliidid esimese 35 hulgast

HMDB kood	**Ingliskeelne nimetus**	**Järjekorranumber järjestuses**	
		Jälgimisaegne	**Jälgimiseelne**
HMDB0012458	*7alpha-Hydroxy-3-oxo-4-cholestenoate*	32	25
HMDB0002250	*Dodecanoylcarnitine*	11	4
HMDB0000560	*Goshuyic acid*	18	18
HMDB0013207	*9-Hexadecenoylcarnitine*	26	3

Koos isiku taustinfoga nagu sugu, vanus, kehamassiindeks ning elukoha tüüp lisati saadud metaboliidid logistilise regressioonimudelitesse, mis prognoosivad jälgimisaegseid ja -eelseid juhte. Mudel hinnati andmestikuga, mis sisaldas kõiki kontrollgrupi isikuid. Kui tunnuse p-väärtus ületas olulisuse nivood $\alpha = 0{,}05$, jäeti see mudelist välja. Tunnuseid jäeti ükshaaval välja, alustades kõige suurema p-väärtusega tunnusest.

Jälgimisaegsete juhtude mudeldamisel osutusid statistiliselt oluliseks vaid 2 tun-

nust: isiku vanus ja metaboliidi dodekanoüülkarnitiini (inglise keeles *Dodecanoylcarnitine*, *Dodecan*) kontsentratsioon. Tabelis 3 on välja toodud lõpliku mudeli parameetrite hinnangud koos standardvigadega ja p-väärtustega.

Tabel 3: Jälgimisaegse juhu lõplik mudel

Tunnus	**Parameetri hinnang**	**Standardviga**	**P-väärtus**
Vabaliige	-3,60	0,41	$2 \cdot 10^{-16}$
Vanus	0,02	0,0078	0,012
Dodecan	0,23	0,12	0,045

Mudelis on mõlema tunnuse parameetrite hinnangud positiivse väärtusega, mis tähendab, et suurema väärtuse korral suureneb šanss peale proovi andmist ISH-sse haigestuda. Olgu vaatluse all 2 sama vanusega isikut, kellest ühel on dodekanoüülkarnitiini kontsentratsioon 1 ühiku võrra suurem kui teisel. Mudeli põhjal oleks suurema dodekanoüülkarnitiini kontsentratsiooniga isikul 1,26 korda suurem šanss haigestuda kui teisel.

Tabel 4: Jälgimiseelse juhu lõplik mudel

Tunnus	**Parameetri hinnang**	**Standardviga**	**P-väärtus**
Vabaliige	-3,39	0,51	$1{,}31 \cdot 10^{-14}$
Naine	-0,051	0,18	0,0058
Kehamassiindeks	0,04	0,016	0,011
Vanus	0,03	0,0057	$7{,}7 \cdot 10^{-6}$
Gosh. acid	0,3	0,086	0,00047

Jälgimiseelsete juhtude lõplikusse mudelisse jäi alles 4 statistiliselt olulist tunnust: sugu, kehamassiindeks, vanus ning metaboliidi (Z,Z)-5,8-tetradekadieenhappe (inglise keeles *Goshuyic acid*) kontsentratsioon. Soo puhul on mudelis baastunnuseks meessugu. Tabelis 4 on välja toodud mudeli parameetrite hinnangud, standardvead ning p-väärtused.

Mudelis on tunnuste parameetri hinnangud positiivsed, välja arvatud sool. On näha, et kõrgema kehamassiindeksiga, vanusega ja *Gosh. acid* kontsentratsiooniga meestel on suurem šanss olla varem diagnoositud ISH-ga proovi andmisel. Olgu vaatluse all 2 sama kehamassiindeksiga ning vanusega meest, kellest ühel on ühiku võrra suurem metaboliidi *Gosh. acid* kontsentratsioon kui teisel. Mudeli põhjal on suurema metaboliidiga isikul 1,35 korda suurem šanss olla varasema ISH diagnoosiga kui kui teisel.

7 Arutelu

Käesoleva töö tulemustena leitud metaboliite on kavas uurida lähemalt edaspidistes uuringutes, et leida südame isheemiatõbede tekkeprotsessidega seotud metaboliite.

Töös tutvustati geneetilist algoritmi ja rakendati seda Tartu Ülikooli Eesti Geenivaramu andmestikule. Tulemusteks saadi 4 andmestiku põhjal 4 erinevat metaboliitide järjestust vastavalt sellele, mitmes parima sobivusega kromosoomis metaboliit sisaldus. Kui igast järjestusest võtta esimesed 35 metaboliiti ehk kõige sagedamad, siis leiduvad vaid üksikud metaboliidid, mis sisalduvad vähemalt 2 järjestuses. Neist mudelisse lisati 4 metaboliiti, mis kattusid andmestikega, millel olid ühised kontrollid. Hinnati 2 logistilist regressioonimudelit. Kummagi mudeli tunnuste hulgas oli 1 metaboliit statistiliselt oluline.

Statiliselt oluliste metaboliitide hulka kuulus dodekanoüülkarnitiin. See atsüülkarnitiinide hulka lipiidide grupis, mille peamisteks ülesanneteks on transportida orgaanilisi happeid ja rasvhappeid mitokondrisse energia tootmiseks (Inimese Metaboliitide Andmebaas, 2023a). Kõrgema kontsentratsiooniga atsüülkarnitiinide metaboliite on seostatud teist tüüpi diabeetikutest südame- ja veresoonkonnahaiguse diagnoosiga patsientidel (Zhao *et al.*, 2020) ning südame vasaku vatsakese puudulikkusega (Zordoky *et al.*, 2015).

Teine statistiliselt oluline metaboliit oli (Z,Z)-5,8-tetradekadieenhappe, mis on pika ahelaga rasvhape ja asub peamiselt rakumembraanis. See tekib vahesaadusena küllastumata rasvhapete oksüdeerumisel ning sisaldub rapsiõlis (Inimese Metaboliitide Andmebaas, 2023b).

Kõige sagedamate metaboliitide hulgas leidub enim lipiide, kuna andmestikus üle poole metaboliitidest on lipiidid. Neist enamik on rasvhapped, kuid eksisteerivad ka mõned steroidid.

Ksenobiootikume ülejäänud metaboliidi tüüpidega võrreldes esineb samuti teistest rohkem. Näiteks kõikide kontrollidega jälgimiseelsete juhtude andmestikus kuulus

esimese 4 metaboliidi hulka 3 ksenobiootikumi. Nende hulgas esikohal paiknev perfluorooktaanhape (PFOA) on väga levinud kemikaal, kuna organism ei sünteesi ega lagunda seda. PFOA-d peetakse mürgiseks ning on leitud seoseid kolesterooliga. (Steenland, Fletcher ja Savitz, 2010). On leitud ka, et PFOA tase on kõrgem vanemaealiste inimeste hulgas (Steenland *et al.*, 2009).

Jälgimisaegsete juhtude puhul töö autor ei olnud teadlik, mitu aastat peale proovi andmist haigus diagnoositi. Kui haigus avaldus 10 aastat hiljem, siis vanade metaboloomika andmete põhjal ei pruugi seda prognoosida. Ideaaltingimustes eeldatakse, et isik ei ole vahepeal oma käitumisharjumusi muutnud. Erinevalt haiglas läbiviidavatest sarnastest uuringutest, kus on kontrollitud keskkond, geenidoonoritele sääraseid korraldusi ei määrata. Seega võib haigus tekkida muu elustiili käigus, mille kohta andmed puuduvad.

Tulemusteni jõuti ainult ühe klassifikatsioonimeetodiga, kuigi plaanis oli võrrelda 3 meetodi tulemusi. Peamine takistus teistel meetoditel oli madal sobivus, mis põlvkondade jooksul tõusis vähehaaval. Et jõuda soovitud sobivuseni, oli vaja läbida rohkem põlvkondi, mistõttu kuluks protsessiks rohkem aega. Autori isikliku rüperaaliga võib selleks kuluda nädalaid. Kui tulevikus andmestik uuesti TVM ja juhusliku metsa klassifikatsioonimeetodiga GALGO-s läbi jooksutada, siis tarvitseb seda teha võimsama riistvaraga.

Töös leitud markeritest ei saa teha põhjuslikke järeldusi, kuna ajapuuduse tõttu ei olnud võimalik metaboliite täpsemalt uurida. Need tuleks tulevikus valideerida iseseisva andmestikuga. Geenivaramul on olemas veel metaboloomika andmestikke, millega saab tulemusi valideerida.

Kokkuvõte

Bakalaureusetöö eesmärk oli Tartu Ülikooli Geenivaramu metaboloomika andmete põhjal leida metaboliidid, mis sobiksid potentsiaalseteks südame isheemiatõbesid kirjeldavateks markeriteks.

Tulemuste leidmiseks rakendati andmetele geneetilist algoritmi. Töös üritati võrrelda 3 klassifikatsioonimeetodil saadud tulemusi. Tulemused saadi lähima tsentroidi meetodil. Paraku protsessiks nõutava mahuka ajakulu tõttu tugivektor-masina ning juhusliku metsa meetodil tulemusteni ei jõutud.

Südame isheemiatõvega diagnoositud geenidoonorid jagati kahte gruppi: jälgimisaegsed juhud ning jälgimiseelsed juhud. Kummagi haigusgrupiga eraldi moodustati andmestikud, millest esimesse kuulusid kõik kontrollgruppi kuulunud geenidoonorid ning teise haigusgrupi isikutega sarnased kontrollid. Kõikidel andmestikel rakendati geneetilist algoritmi lähima tsentroidi klassifikatsioonimeetodiga.

Iga andmestikuga saadud tulemustest valiti välja esimesed 35 metaboliiti. Nende hulgast saadi 4 metaboliiti, mis kuulusid võrdse arvu kontrollidega andmestike ühisossa. Töös hinnati 2 logistilist regressioonimudelit prognoosimaks haigusgruppe, kus lisaks saadud metaboliitidele kasutati tunnustena doonori sugu, vanust proovi andmise hetkel, kehamassiindeksit ning elupaika. Mõlema mudeli peale sai kokku 2 statistiliselt olulist metaboliiti: dodekanoüülkarnitiin ja (Z,Z)-5,8-tetradekadieenhappe. Varasemad uuringud on leidnud seoseid dodekanoüülkarnitiiniga ja südamega seotud haigustega.

Kasutatud allikad

Ahmed, S., P. Shah ja O. Ahmed (2023). *StatPearls [Internet]. Biochemistry, Lipids*. Treasure Island (FL): StatPearls Publishing. URL: https://www.ncbi.nlm.nih.gov/books/NBK525952/ (vaadatud 06.05.2023).

Almdal, Thomas, Henrik Scharling, Jan Skov Jensen ja Henrik Vestergaard (2004). "The independent effect of type 2 diabetes mellitus on ischemic heart disease, stroke, and death: a population-based study of 13 000 men and women with 20 years of follow-up". *Archives of internal medicine* 164.13, lk. 1422–1426.

Ameerika Südameassotsiatsioon (2022). *Silent Ischemia and Ischemic Heart Disease*. URL: https://www.heart.org/en/health-topics/heart-attack/about-heart-attacks/silent-ischemia-and-ischemic-heart-disease (vaadatud 17.04.2023).

Belgiu, Mariana ja Lucian Drăguţ (2016). "Random forest in remote sensing: A review of applications and future directions". *ISPRS Journal of Photogrammetry and Remote Sensing* 114, lk. 24–31. ISSN: 0924-2716. DOI: https://doi.org/10.1016/j.isprsjprs.2016.01.011. URL: https://www.sciencedirect.com/science/article/pii/S0924271616000265 (vaadatud 28.03.2023).

Cervantes, Jair, Farid Garcia-Lamont, Lisbeth Rodríguez-Mazahua ja Asdrubal Lopez (2020). "A comprehensive survey on support vector machine classification: Applications, challenges and trends". *Neurocomputing* 408, lk. 189–215. ISSN: 0925-2312. DOI: https://doi.org/10.1016/j.neucom.2019.10.118. URL: https://www.sciencedirect.com/science/article/pii/S0925231220307153 (vaadatud 28.03.2023).

Euroopa Bioinformaatika instituut (2023). *Comparison of NMR and MS*. URL: https://www.ebi.ac.uk/training/online/courses/metabolomics-

introduction/designing-a-metabolomics-study/comparison-of-nmr-and-ms/ (vaadatud 21.04.2023).

Idle, Jeffrey R ja Frank J Gonzalez (2007). "Metabolomics". *Cell metabolism* 6.5, lk. 348–351.

Inimese Metaboliitide Andmebaas (2023a). *Showing metabocard for Dodecanoylcarnitine (HMDB0002250)*. URL: https://hmdb.ca/metabolites/HMDB0002250 (vaadatud 09.05.2023).

– (2023b). *Showing metabocard for Goshuyic acid (HMDB0000560)*. URL: https://hmdb.ca/metabolites/HMDB0000560 (vaadatud 09.05.2023).

Katobh, S., S. S. Chauhan ja V. Kumar (2021). "A review on genetic algorithm: past, present, and future". *Multimedia Tools and Applications* 80.5, lk. 8091–8126. URL: https://link.springer.com/article/10.1007/s11042-020-10139-6 (vaadatud 27.03.2023).

Levner, Ilya (2005). "Feature selection and nearest centroid classification for protein mass spectrometry". *BMC bioinformatics* 6.1, lk. 1–14.

Lin, Xiaohui, Quancai Wang, Peiyuan Yin, Liang Tang, Yexiong Tan, Hong Li, Kang Yan ja Guowang Xu (2011). "A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection". *Metabolomics* 7, lk. 549–558.

Maailma Terviseorganisatsioon (2016). *International Statistical Classification of Diseases and Related Health Problems 10th Revision (ICD-10)*. URL: https://icd.who.int/browse10/2016/en#/I20-I25 (vaadatud 28.03.2023).

Maailma Terviseorganisatsioon (2020). *The top 10 causes of death*. URL: https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death (vaadatud 08.05.2023).

Mitchell, Melanie (1998). *An introduction to genetic algorithms*. MIT press.

Morbiducci, Umberto, Andrea Tura ja Mauro Grigioni (2005). "Genetic algorithms for parameter estimation in mathematical modeling of glucose metabolism". *Computers in Biology and Medicine* 35.10, lk. 862–874. ISSN: 0010-4825. DOI: https://doi.org/10.1016/j.compbiomed.2004.07.005. URL: https://www.sciencedirect.com/science/article/pii/S0010482504000976.

Morgan-Benita, Jorge, Ana G Sánchez-Reyna, Carlos H Espino-Salinas, Juan José Oropeza-Valdez, Huizilopoztli Luna-García, Carlos E Galván-Tejada, Jorge I Galván-Tejada, Hamurabi Gamboa-Rosales, Jose Antonio Enciso-Moreno ja José Celaya-Padilla (2022). "Metabolomic Selection in the Progression of Type 2 Diabetes Mellitus: A Genetic Algorithm Approach". *Diagnostics* 12.11, lk. 2803.

Nick, Todd G ja Kathleen M Campbell (2007). "Logistic regression". *Topics in biostatistics*, lk. 273–301.

Steenland, Kyle, Tony Fletcher ja David A Savitz (2010). "Epidemiologic evidence on the health effects of perfluorooctanoic acid (PFOA)". *Environmental health perspectives* 118.8, lk. 1100–1108.

Steenland, Kyle, Chuangfang Jin, Jessica MacNeil, Cathy Lally, Alan Ducatman, Veronica Vieira ja Tony Fletcher (2009). "Predictors of PFOA levels in a community surrounding a chemical plant". *Environmental health perspectives* 117.7, lk. 1083–1088.

Štefanac, Tea, Dijana Grgas ja Tibela Landeka Dragičević (oktoober 2021). "Xenobiotics-division and methods of detection: A review". en. *J. Xenobiot.* 11.4, lk. 130–141.

Trevino, V. ja F. Falciani (2006). "GALGO: an R package for multivariate variable selection using genetic algorithms". *Journal of the American Statistical Association* 22.9, lk. 1154–1156. URL: https://www.researchgate.net/publication/7266382_GALGO_An_R_package_for_multivariate_variable_selection_using_genetic_algorithms (vaadatud 26.03.2023).

Trevino, Victor ja Francesco Falciani (2018). *galgo: Genetic Algorithms for Multivariate Statistical Models from Large-Scale Functional Genomics Data*. R package version 1.4. URL: https://github.com/vtrevino/GALGO.

Varbo, Anette, Marianne Benn, Anne Tybjærg-Hansen, Anders B Jørgensen, Ruth Frikke-Schmidt ja Børge G Nordestgaard (2013). "Remnant cholesterol as a causal risk factor for ischemic heart disease". *Journal of the American College of Cardiology* 61.4, lk. 427–436.

Zhao, Shuo, Xiao-Fei Feng, Ting Huang, Hui-Huan Luo, Jian-Xin Chen, Jia Zeng, Muyu Gu, Jing Li, Xiao-Yu Sun, Dan Sun, Xilin Yang, Zhong-Ze Fang ja Yun-Feng Cao (mai 2020). "The association between acylcarnitine metabolites and cardiovascular disease in Chinese patients with type 2 diabetes mellitus". en. *Front. Endocrinol. (Lausanne)* 11, lk. 212.

Zordoky, Beshay N, Miranda M Sung, Justin Ezekowitz, Rupasri Mandal, Beomsoo Han, Trent C Bjorndahl, Souhaila Bouatra, Todd Anderson, Gavin Y Oudit, David S Wishart, Jason R B Dyck ja Alberta HEART (mai 2015). "Metabolomic fingerprint of heart failure with preserved ejection fraction". en. *PLoS One* 10.5, e0124844.

Lisa 1. GALGO protsessiga leitud kõige sagedamad metaboliidid

Tabel 5: Kõige sagedamad 35 metaboliiti kõikide kontrollidega jälgimisaegsetel juhtudel

Jrk	**HMDB kood**	**Metaboliidi ingliskeelne nimetus**
1	HMDB0013288	*Nonanoylcarnitine*
2	HMDB0000760	*Hyocholic acid*
3	HMDB0011737	*gamma-Glutamylglutamic acid*
4	HMDB0000791	*Octanoylcarnitine*
5	HMDB0000535	*Caproic acid*
6	HMDB0000619	*Cholic acid*
7	HMDB0011635	*p-Cresol sulfate*
8	HMDB0062121	*Dihydroferulic acid*
9	HMDB0001870	*Benzoic acid*
10	HMDB0004980	*cis-4-Decenoic acid*
11	HMDB0002250	*Dodecanoylcarnitine*
12	HMDB0000792	*Sebacic acid*
13	HMDB0008141	PC(18:2(9Z,12Z)/18:3(9Z,12Z,15Z))
14	HMDB0000756	*Hexanoylcarnitine*
15	HMDB0002869	*Campesterol*

16	HMDB0000686	*Isoursodeoxycholic acid*
17	HMDB0135261	*Propyl 4-hydroxybenzoate sulfate*
18	HMDB0000560	*Goshuyic acid*
19	HMDB0000482	*Caprylic acid*
20	HMDB0060426	*8-Methoxykynurenate*
21	HMDB0000946	*Ursodeoxycholic acid*
22	HMDB0029160	*gamma-Glutamyltryptophan*
23	HMDB0031518 või HMDB0000156	*D-Malic acid* või *Malic acid*
24	HMDB0000134	*Fumaric acid*
25	HMDB0000666	*Heptanoic acid*
26	HMDB0013207	*9-Hexadecenoylcarnitine*
27	HMDB0007220	DG(18:1(9Z)/18:3(6Z,9Z,12Z)/0:0)
28	HMDB0062775	*4-Vinylphenol sulfate*
29	HMDB0001406	*Niacinamide*
30	HMDB0006335 või HMDB0001492	*beta-Tocopherol* või *gamma-Tocopherol*
31	HMDB0000512	*N-Acetyl-L-phenylalanine*
32	HMDB0012458	*7alpha-Hydroxy-3-oxo-4-cholestenoate*
33	HMDB0000883	*L-Valine*
34	HMDB0000452	*L-alpha-Aminobutyric acid*

35	HMDB0000131	*Glycerol*

Tabel 6: Kõige sagedamad 35 metaboliiti kõikide kontrollidega jälgimiseelsetel juhtudel

Jrk	**HMDB kood**	**Metaboliidi ingliskeelne nimetus**
1	HMDB0059587	*Perfluorooctanoic acid (PFOA)*
2	HMDB0000840	*Salicyluric acid*
3	HMDB32332	*Hydroxylated lecithin*
4	HMDB0001895	*Salicylic acid*
5	HMDB0061112	*3-Carboxy-4-methyl-5-propyl-2-furanpropionic acid*
6	HMDB0000169	*D-Mannose*
7	HMDB0007266	DG(18:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z,16Z,19Z)/0:0)
8	HMDB0000933	*Traumatic acid*
9	HMDB0059586	*Perfluorooctanesulfonate (PFOS)*
10	HMDB0000122	*D-Glucose*
11	HMDB0240565	*Umbelliferone sulfate*
12	HMDB0061643	*3-carboxy-4-methyl-5-pentyl-2-furanpropanoic acid*
13	HMDB0000613	*Erythronic acid*

14	HMDB0000413	*3-Hydroxydodecanedioic acid*
15	HMDB0000258	*Sucrose*
16	HMDB0004824	*N2,N2-Dimethylguanosine*
17	HMDB0013336	*3-Hydroxyhexadecanoylcarnitine*
18	HMDB0011568	MG(18:2(9Z,12Z)/0:0/0:0)
19	HMDB0002577	*Cholic acid glucuronide*
20	HMDB0001173	*5'-Methylthioadenosine*
21	HMDB0000337	*(S)-3,4-Dihydroxybutyric acid*
22	HMDB0240506 või HMDB0240505	*Enterolactone 3"-sulfate* või *Enterolactone 3'-sulfate*
23	HMDB0011565	MG(16:1(9Z)/0:0/0:0)
24	HMDB0002189 või HMDB0001276	*N8-Acetylspermidine* või *N1-Acetylspermidine*
25	HMDB0012458	*7alpha-Hydroxy-3-oxo-4-cholestenoate*
26	HMDB0000291	*Vanillylmandelic acid*
27	HMDB0013127	*3-Hydroxybutyrylcarnitine*
28	HMDB0000755	*Hydroxyphenyllactic acid*
29	HMDB0000094	*Citric acid*
30	HMDB0033433	*(S)-Homostachydrine*
31	HMDB0240345	*N2,N5-Diacetylornithine*
32	HMDB0013127	*3-Hydroxybutyrylcarnitine*

33	HMDB0000350	*3-Hydroxysebacic acid*
34	HMDB0002331	*Imidazoleacetic acid riboside*
35	HMDB0094700	*Perfluorohexanesulfonate (PFHxS)*

Tabel 7: Kõige sagedamad 35 metaboliiti sarnaste kontrollidega jälgimisaegsetel juhtudel

Jrk	**HMDB kood**	**Metaboliidi ingliskeelne nimetus**
1	HMDB0001925	*Ibuprofen*
2	HMDB0000127	*D-Glucuronic acid*
3	HMDB0000682	*Indoxyl sulfate*
4	HMDB0062775	*4-Vinylphenol sulfate*
5	HMDB0011376	PE(P-18:0/18:2(9Z,12Z))
6	HMDB0008141	PC(18:2(9Z,12Z)/18:3(9Z,12Z,15Z))
7	HMDB0032055	*N-Acetylhistidine*
8	HMDB0011635	*p-Cresol sulfate*
9	HMDB0011737	*gamma-Glutamylglutamic acid*
10	HMDB0007220	DG(18:1(9Z)/18:3(6Z,9Z,12Z)/0:0)
11	HMDB0011342	PE(P-16:0/18:1(9Z))
12	HMDB0240681 või HMDB0011773	Cer(d16:1/16:0) või Cer(d18:1/14:0)
13	HMDB0000017	*4-Pyridoxic acid*

14	HMDB00653	*Cholesterol sulfate*
15	HMDB0007973	PC(16:0/18:2(9Z,12Z))
16	HMDB0007266	DG(18:2(9Z,12Z)/22:6(4Z,7Z,10Z,13Z, 16Z,19Z)/0:0)
17	HMDB0240686	Cer(d18:2(4E,14Z)/16:0)
18	HMDB0000560	*Goshuyic acid*
19	HMDB0002579	*Glycochenodeoxycholic acid 3-glucuronide*
20	HMDB0000682	*Indoxyl sulfate*
21	HMDB0002250	*Dodecanoylcarnitine*
22	HMDB0000946	*Ursodeoxycholic acid*
23	HMDB0003290	*Gulonic acid*
24	HMDB0011756	*N-Acetylleucine*
25	HMDB0240581 või HMDB0240582	*5alpha-Pregnan-3beta,20beta-diol disulfate* või *5alpha-Pregnan-3alpha,20beta-diol disulfate*
26	HMDB0011569	MG(18:3(6Z,9Z,12Z)/0:0/0:0)
27	HMDB0094650	*5alpha-Pregnan-3beta,20alpha-diol disulfate*
28	HMDB0008039	PC(18:0/18:2(9Z,12Z))
29	HMDB0002802	*Cortisone*

30	HMDB0003073 või HMDB0001388	*gamma-Linolenic acid* või *alpha-Linolenic acid*
31	HMDB0000699	*1-Methylnicotinamide*
32	HMDB0011621	*Cinnamoylglycine*
33	HMDB0004089	*Formylanthranilic acid*
34	HMDB0000305	*Vitamin A*
35	HMDB0004949	Cer(d18:1/16:0)

Tabel 8: Kõige sagedamad 35 metaboliiti sarnaste kontrollidega jälgimiseelsetel juhtudel

Jrk	**HMDB kood**	**Metaboliidi ingliskeelne nimetus**
1	HMDB0240588	*Myristoleoylcarnitine*
2	HMDB0094696	*N-Methyl-L-proline*
3	HMDB0013207	*9-Hexadecenoylcarnitine*
4	HMDB0002250	*Dodecanoylcarnitine*
5	HMDB0061636	*3-hydroxydecanoyl carnitine*
6	HMDB0000169	*D-Mannose*
7	HMDB0029151	*gamma-Glutamylhistidine*
8	HMDB0013336	*3-Hydroxyhexadecanoylcarnitine*
9	HMDB0000345	*3-Hydroxyadipic acid*
10	HMDB0001895	*Salicylic acid*

11	HMDB0005066	*Tetradecanoylcarnitine*
12	HMDB0000413	*3-Hydroxydodecanedioic acid*
13	HMDB0029942	*D-Arabinose*
14	HMDB0000201	*L-Acetylcarnitine*
15	HMDB0006270 või HMDB0000673	*Linoelaidic acid* või *Linoleic acid*
16	HMDB0004827	*Proline betaine*
17	HMDB0001383	*Sphinganine 1-phosphate*
18	HMDB0000560	*Goshuyic acid*
19	HMDB0000477	*7Z,10Z-Hexadecadienoic acid*
20	HMDB32332	*Hydroxylated lecithin*
21	HMDB0000709	*L-Cysteinylglycine disulfide*
22	HMDB0006210	*Heptadecanoyl carnitine*
23	HMDB0000212 või HMDB0000215	*N-Acetylgalactosamine* või *N-Acetyl-D-glucosamine*
24	HMDB0240362	*N-stearoyl-sphingadienine*
25	HMDB03349	*L-Dihydroorotic acid*
26	HMDB0240584	*(4E,15E)-Bilirubin*
27	HMDB0006528 või HMDB0001976	*Docosapentaenoic acid (22n-3)* või *Docosapentaenoic acid (22n-6)*
28	HMDB0001425	*Estrone sulfate*

29	HMDB0013622	*10Z-Nonadecenoic acid*
30	HMDB0000766	*N-Acetyl-L-alanine*
31	HMDB0061384	*N-(3-acetamidopropyl)pyrrolidin-2-one*
32	HMDB0000222	*Palmitoylcarnitine*
33	HMDB0000840	*Salicyluric acid*
34	HMDB0000933	*Traumatic acid*
35	HMDB0041623	*N6-Carbamoyl-L-threonyladenosine*

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Johann Saavaste,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose "Isheemiliste südamehaiguste markerite leidmine Tartu Ülikooli Eesti Geenivaramu metaboloomika andmetele geneetilist algoritmi rakendades", mille juhendajad on Jaanika Kronberg ja Krista Fischer, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

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

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

Johann Saavaste

09.05.2023