TARTU ÜLIKOOL
LOODUS- JA TEHNOLOOGIATEADUSKOND
MOLEKULAAR- JA RAKUBIOLOOGIA INSTITUUT
BIOINFORMAATIKA ÕPPETOOL

Maarja Lepamets

Oligomeeridel põhinevate bioinformaatiliste meetodite kasutamine bakterite määramiseks sekveneerimislugemitest

Bakalaureusetöö

Juhendaja teadur Lauris Kaplinski

TARTU, 2014

Sisukord

Kasutatud lühendid

AAI	*amino acid identity*, aminohapete identsus
ANI	*average nucleotide identity*, keskmine nukleotiidide identsus
BLAST	*Basic Local Alignment Search Tool*, bioinformaatiline tööriist lokaalsete joonduste otsimiseks
BSO	*bacteria-specific oligomer*, bakterispetsiifiline oligomeer
CAS	*compare-and-swap*, operatsioon muutuja väärtuse võrdlemiseks ja vahetamiseks mitmelõimelistes programmides
DDH	DNA-DNA hübridisatsioon
FAME	*fatty acid methyl ester*, rasvhappe metüülester
FTIR	*Fourier Transformation Infrared*, infrapunasel valgusel põhinev spektroskoopia meetod
GiB	gibibait (1 GiB $= 1024^3$ baiti)
HGT	*horizontal gene transfer*, horisontaalne geeniülekanne
ICP-FT MS	*Ion Cyclotron Fourier Transform mass spectrometry*, massispektromeetria meetod, mis põhineb osakeste tiirlemissageduse mõõtmisel
MALDI-TOF MS	*Matrix-Assisted-Laser Desorption/Ionization Time of Flight mass spectrometry*, massispektromeetria meetod, mis põhineb osakeste lennuaja mõõtmisel
MiB	mebibait (1 MiB $= 1024^2$ baiti)
MLSA	*Multilocus Sequence Analysis*, geenide järjestustel põhinev bakterite määramise meetod
MLST	*Multilocus Sequence Typing*, geenide alleelidel põhinev bakterite määramise meetod
MS	massispektromeetria
NGS	*next generation sequencing*, teise põlvkonna sekveneerimine
RBR	*relative binding ratio*, suhteline paardumise suhe

Sissejuhatus

Baktereid on vaja määrata paljudes valdkondades nagu diagnostika, toiduainetööstus ja keskkonnakaitse. Määramismeetodid jagunevad fenotüübil ja genotüübil põhinevateks. Viimaste aastakümnete jooksul on hakatud baktereid klassifitseerima nende DNA järjestuste alusel. Seetõttu kasutatakse ka bakterite identifitseerimisel eelkõige genoomset informatsiooni. Oluliseks eeliseks on samuti see, et genotüübil põhinevate meetoditega saab määrata liike tervest bakterite kooslusest ja ilma puhaskultuure kasvatamata.

Üheks perspektiivseks DNA järjestustel põhinevaks määramismeetodite klassiks on teise põlvkonna sekveneerimisandmete analüüs. Levinud variantideks on võrrelda tervete bakterigenoomide või kindlate geenide homoloogiat erinevate taksonite vahel. Samuti võib kasutada lühikesi bakterispetsiifilisi markerjärjestusi kogu genoomi ulatusest. Kaks esimest meetodit eeldavad sekveneerimislugemite kokkupanemist pikemateks järjestusteks ning on seega aeganõudvamad. Mitmetes rakendustes on aga bakterite määramise kiirus elulise tähtsusega. Lühikestel DNA järjestustel ehk oligomeeridel põhinevaid meetodeid võib rakendada otse lugemitele, mistõttu on tulemusi võimalik saada tunduvalt kiiremini. Seni pole aga loodud piisavalt head bioinformaatilist algoritmi, mis lugemites leiduvate oligomeeride abil sekveneeritud baktereid suudaks määrata.

Käesoleva bakalaureusetöö üldeesmärk on tutvustada bakterite määramise problemaatikat ja levinumaid metoodikaid. Sealjuures on rõhuasetus pandud bakterispetsiifilistel oligomeeridel baseeruvatele bioinformaatilistele meetoditele ja algoritmidele. Töö praktilise osa eesmärk on luua ja kirjeldada uut oligomeeridel põhinevat bioinformaatilist tööriista. Samuti hinnata loodud programmipaketi potentsiaali sekveneerimislugemitest bakterite määramiseks. Teoreetilise osa viimases peatükis on antud ka lühike ülevaade algoritmide keerukusest ja otsingualgoritmidest.

Märksõnad: *bakterite määramine, sekveneerimisandmete analüüs, bioinformaatilised algoritmid, oligomeerid, k-meerid*

1 Kirjanduse ülevaade

1.1 Bakteriliikide määramise problemaatika

Baktereid leidub kõikides Maa biosfääri osades. Nad elavad nii veekogudes, mullas, sügavamates pinnakihtides kui ka maapealses biomassis, kaasa arvatud loomade ja inimeste organismides. Prokarüootide rakkude koguarvuks on hinnatud $4-6\times10^{30}$ rakku ning ligikaudu 3.5×10^{11} kuni 5.5×10^{11} tonni kogu maakeral leiduvast süsinikust kuulub bakterite ja arhede koostisesse (Whitman *et al.*, 1998). Bakterid osalevad paljudes olulistes bioloogilistes protsessides nagu süsiniku- ja lämmastikuringe, keskkonnast jääkproduktide lagundamine või haiguste tekitamine.

Bakteritel ei ole üheselt ja selgelt eristuvaid liike, sest neil puudub suguline paljunemine. Nad evolutsioneeruvad väga kiiresti ja seega on nende varieeruvus ühe taksonoomilise grupi piires suur. Samas leidub baktereid, kes paiknevad evolutsiooniliselt üksteisest kaugel, kuid on mõne tunnuse või markergeeni järjestuse poolest peaaegu identsed. Bakterite klassifitseerimise muudab keeruliseks ka asjaolu, et evolutsiooniliselt kaugetest taksonoomilistest rühmadest pärit isendid võivad vahetada geneetilist materjali horisontaalse geeniülekande (HGT, *horizontal gene transfer*) käigus. On näidatud, et 30% kuni 50% kõigist bakteritest on vähemalt ühe valgudomeeni kodeeriva genoomipiirkonna omandanud HGT käigus (Choi ja Kim, 2007). Samuti paiknevad osad fenotüüpi määravad geenid mõnikord plasmiididel ning võivad seega kergesti kaduma minna.

Kokkuleppelisi bakteriliikide definitsioone on mitmeid. Erinevate bakterite määramise metoodikate põhjal on loetud samasse liiki kuuluvateks bakteritüvedeks neid, millel on sarnane fenotüüp, mille G + C sisaldused erinevad alla 5%, mille genoomsed DNA-d hübridiseeruvad vähemalt 70% ulatuses, mille aminohapete identsus (AAI) ja keskmine nukleotiidide identsus (ANI) on üle 95% ning mille 16S rRNA geenid on rohkem kui 98% ulatuses identsed (Thompson *et al.*, 2013).

1.1.1 Diagnostika ja epidemioloogia

Bakterid võivad põhjustada inimestel ja loomadel mitmeid haigusi. Raskematest haigustest on bakteriaalsed näiteks tuberkuloos, salmonelloos, teetanus ja difteeria. Bakteriaalse päritoluga võivad olla ka näiteks 2. tüübi diabeet (Moreno-Indias *et al.*, 2014), ülekaalulisus (Ley, 2010) ja hambakaaries (Benítez-Páez *et al.*, 2014). Muutused inimese loomulikus mikrofloora või patogeenide poolt põhjustatud põletikud võivad tekitada soole-, naha-, kopsu-, maksa-

ja rinnavähki (Schwabe ja Jobin, 2013).

Kliiniline mikrobioloogia jaguneb diagnostikaks ja epidemioloogiaks. Diagnostika ülesandeks on määrata konkreetsele patsiendile bakteriaalse haiguse korral õige diagnoos. Epidemioloogia ülesanne on tuvastada ja võimalusel ennetada haiguste levikut.

Bakteriaalsete haiguste diagnoosimiseks võetakse patsiendilt proov, millest määratakse patsiendi organismis leiduvad bakterid. Traditsioonilised sammud selleks on isolaadi külvamine söötmele, üleskasvanud bakterite liikide määramine ning bakterite patogeense mõju ja ravimitele vastuvõtlikkuse tuvastamine. Olenevalt bakteriliigist võib kogu protsess aega võtta paar päeva kuni mitu kuud. Tihti aga on patsiendi jaoks elulise tähtsusega alustada õiget ravi võimalikult varakult. Seega on üheks diagnostika arenduse olulisimaks suunaks uute kiiremate meetodite väljatöötamine. Epidemioloogia seisukohalt on oluline teada ka haigustekitaja päritolu ja sugulust teiste tüvedega. (Didelot *et al.*, 2012)

Kliinilises mikrobioloogias kasutatavad bakterite määramise metoodikad peavad võimaldama määramist perekonna, liigi või tüve tasandil.

1.1.2 Ökoloogia ja keskkonnakaitse

Baktereid leidub kõigis Maa ökosüsteemides. Nende liike on määratud ookeanidest (DeLong, 2005), mullast (Daniel, 2005), kuumaveeallikatest (Ward *et al.*, 1998), polaarjäästt (Christner *et al.*, 2003) ja mujalt.

Keskkonna mikrobioloogiliste proovide analüüs võimaldab teadlastel saada informatsiooni keskkonnatingimuste ja saastatuse kohta. Näiteks on täheldatud, et mulla funktsionaalne mikrobioloogiline koostis varieerub sõltuvalt CO_2 ja O_3 kontsentratsioonist atmosfääris (He *et al.*, 2014). Levinud meetodiks on ka *Escherichia coli* tüvede määramisega testida joogivee puhtust (Edberg *et al.*, 2000).

Mullas elavad bakterid ja samas piirkonnas kasvavad taimed mõjutavad vastastikku üksteist. Näiteks elavad liblikõielised sümbioosis mõnede α- ja β-proteobakteritega, kes aitavad taimedel juurte kaudu omastada mullast õhulämmastikku (van Rhijn ja Vanderleyden, 1995). Pini *et al.* (2012) on teadaolevalt esimesed, kes on uurinud kogu ühe taimeliigiga (liblikõieline põllukultuur *Medicago sativa*) seotud mikrobioomi, võttes mikrobioloogilisi proove mullast ja taime kudedest ning määrates saadud proovidest bakteriliike.

Keskkonnakaitse jaoks otsitakse bakteriliike, keda saaks kasutada keskkonnasaaste eemaldamiseks. Sellist protsessi nimetatakse bioremedatsiooniks. Eelkõige on uuritud bakterite rolli õlireostuste neutraliseerimisel. Acosta-González *et al.* (2013) on määranud bakte-

reid õlireostuse piirkonnast ning leidnud sealt γ- ja δ-proteobakterite esindajaid, kes suudavad lagundada naftaleeni ja teisi süsivesinikke.

Kahjuks pole suur osa keskkonnast eraldatud baktereid laboritingimustes kultiveeritavad. Seega on meetodid nende määramiseks piiratud. Enamasti saab selliseid baktereid määrata vaid keskkonnaproovidest DNA eraldamisega (Logue *et al.*, 2008).

1.1.3 Toiduainetööstus

Põhilisteks toidust põhjustatud haiguste tekitajateks on toiduainetes leiduvad bakterid. Tuntumad neist on perekonna *Salmonella* esindajad, keda leidub linnulihas, munades ja piimatoodetes ning kes põhjustavad salmonelloosi, vibrioosi tekitava *Vibrio* perekonna esindajad, keda leidub kalas ja *Clostridium botulinum*, kes mõnikord paljuneb konserveeritud toidus ning põhjustab botulismi.

Inimeste haigestumise vältimiseks määratakse toiduainetööstuses toidu koostisosadest patogeene. Määramise teeb keeruliseks toidu heterogeensus. Võetud proov võib sisaldada ainult üksikuid haiguspõhjustaja rakke, mis ei pruugi paljude meetoditega detekteeritavad olla. Samuti sisaldavad mitmed toiduained (näiteks toores liha) hulgaliselt neutraalseid baktereid, mis patogeenide määramise seisukohalt on ainult müra. Kuna tootmisprotsessid on tänapäeval väga kiired, siis on oluline, et ka bakterite määramist toiduainetest saaks teha võimalikult lühikese ajaga. (Hoorfar, 2011)

1.2 Bakterite määramise meetodid

1.2.1 Fenotüübil põhinevad meetodid

Traditsiooniliselt on bakteriperekondi ja -liike määratud nende fenotüübi põhjal. Üheks levinud fenotüübiliseks bakterite määramise meetodiks on kirjeldada bakterikolooniate morfoloogiat ja kasvukiirust (Bochner, 2009; Kämpfer ja Glaeser, 2012). Mikroskoobi abil saab määrata veel bakteriraku suurust ja kuju, viburi olemasolu ning võimet moodustada spoore (Moore *et al.*, 2010). Nimetatud tunnused erinevad bakteritel perekonniti või liigiti ning neid saab kasutada bakterite määramiseks.

Baktereid on fenotüübiliselt võimalik määrata nende ainevahetusse kaasatavate metaboliitide järgi. Selleks on loodud spetsiaalsed fenotüüpi määravad mikrokiibid. Igas kiibi lahtris on erinevat süsiniku-, lämmastiku-, väävli- või fosforiühendit sisaldav sööde. Kiibil detekteeritakse vastavate ühendite kaasamist bakteri ainevahetusse. Sarnaselt saab mikrokiibiga määrata, milliste antibiootikumide suhtes on uuritaval bakteril resistentsus. Taoline meetod võimaldab määrata bakterit tüve täpsusega, kui uuritava tüve profiil on varem kindlaks tehtud (Bochner, 2003). Sobivatel substraatidel ja ensüümide aktiivsusel põhinevaid bakterite liikide määramise teste on loonud näiteks firmad bioMérieux ja Biolog.

Bakterite fenotüübiliste tunnuste alla kuuluvad ka rakukesta, rakumembraani ja tsütoplasma keemiline koostis ning struktuur. Erineva rakukesta ja -membraani ehitusega baktereid saab eristada rakkude värvimisega näiteks Grami järgi. Keemiliselt erinevad bakterid peptidoglükaani struktuuri ning rasvhapete, polaarsete lipiidide, kinoonide, pigmentide ja polüamiinide konsentratsiooni poolest rakus (Tindall *et al.*, 2010). Laialdaselt kasutatakse bakterite eristamiseks nende rasvhappe metüülestrite (FAME) profiile (Kunitsky *et al.*, 2006; Ritchie *et al.*, 2000). FAME profiilid on stabiilsed ja hästi konserveerunud ning erinevad bakteriperekonniti. Bakterite identifitseerimisega tegelev firma MIDI on loonud kommertsiaalse süsteemi Sherlock® MIS (*Microbial ID System*), mis baseerub FAME profiilidel. Mõndasid bakteritüvesid saab määrata nende rakkude pinnal asuvate antigeenide põhjal. Selleks tuleb disainida vastava antigeeniga seonduv antikeha (Duval *et al.*, 2014).

Hiljuti on hakatud bakteriliikide ja -tüvede määramiseks kasutama massispektromeetriat (MS). Üks enimkasutatavatest meetoditest on MALDI-TOF MS (*Matrix-Assisted-Laser Desorption/Ionization Time of Flight mass spectrometry*). MALDI-TOF töö põhimõte seisneb ioniseeritud molekulide juhtimises läbi laetud elektrivälja ja nende lennuaja detekteerimises (Dingle ja Butler-Wu, 2013). Saadud aeg sõltub iooni massi ja laengu suhtest, mis väljastatakse bakterispetsiifilise mass-spektrogrammina. MALDI-TOF MS meetod on loodud eelkõige suur-

te molekulide ja molekulikomplekside detekteerimiseks (Karas ja Hillenkamp, 1988). Sarnaselt kasutatakse ka ICP-FT MS (*High-Field Ion Cyclotron Fourier Transform mass spectrometry*) meetodit, mis on eelistatud väiksemate molekulide, näiteks erinevate metaboliitide detekteerimiseks (Rosselló-Móra, 2012). ICP-FT MS meetod leiab ioonide massi ja laengu suhte, detekteerides osakeste tiirlemissagedusi tsüklotronis (Marshall *et al.*, 1998). MS meetodeid on bakterite määramiseks kasutanud näiteks Antón *et al.* (2013).

Bakterite määramiseks on kasutatud ka erinevaid spektroskoopia meetodeid, näiteks FTIR (*Fourier Transformation Infrared*) spektroskoopia (Maity *et al.*, 2013) ja Ramani spektroskoopia (Meisel *et al.*, 2014). Mõlemal juhul registreeritakse valguse intensiivsus erinevatel valguse lainepikkustel. Saadud spektri kuju oleneb bakterirakkude molekulaarsest koostisest.

Fenotüübil põhinevatel bakterite määramise meetoditel on kõigil üks ühine puudus. Nimelt ei sõltu nende meetodite tulemused mitte ainult bakterite geneetilisest materjalist, vaid on tugevas seoses ka nende kasvukeskkonnaga. Seetõttu ei pruugi mingites kasvutingimustes tunnuse mitteavaldumine tähendada vastavate geenide puudumist bakteri genoomist (Kämpfer ja Glaeser, 2012). Samuti vajavad fenotüübil põhinevad meetodid rohkelt materjali ning eeldavad enamasti puhaskultuuride kasvatamist. Bakterirakkude ja -kolooniate morfoloogilised tunnused ning bakterite keemiline koostis pole tihti liikide ning tüvede eristamiseks piisavalt varieeruvad. Seega saab fenotüübil põhinevate meetoditega määrata baktereid vaid perekonna täpsustega. Erandiks on massispektromeetria- ja spektroskoopiameetodid, mis võimaldavad mitmeid baktereid määrata ka liigi, alamliigi ja tüve täpsusega. Lisaks on nad ühed odavamad määramismeetodid, kui välja arvata aparatuuri maksumus. Kahjuks aga on hetkel nende meetodite laialdasemaks kasutamiseks vajalikud andmebaasid veel liiga puudulikud.

Olenemata mitmetest puudustest, kasutatakse fenotüübil põhinevaid bakterite määramise meetodeid küllaltki tihti. Seda eelkõige põhjusel, et nende tulemusi on kergem analüüsida. Nad annavad informatsiooni bakteris toimuva geeniekspressiooni ja metabolismiradade kohta, mida ainult genoomse järjestuse alusel oleks tunduvalt keerulisem ennustada (Emerson *et al.*, 2008).

1.2.2 Genotüübil põhinevad meetodid

Kaasaegne bakterite taksonoomia põhineb suuresti liikide ja tüvede genoomide võrdlustest saadud andmetel. Geneetiline materjal ei sõltu bakteri elukeskkonnast ning on seega palju kindlamaks aluseks liikide määramisel. Mitmed genotüübil põhinevatest meetoditest võimaldavad üheaegselt analüüsida paljude ühes koosluses elavate liikide ja tüvede genoomseid järjestusi. Sellist analüüsivaldkonda nimetatakse metagenoomikaks.

Üks esimesi genoomi iseloomustavaid parameetreid, mis bakterite süstemaatikasse kaasati, oli G + C nukleotiidide sisaldus. On tuvastatud, et ühe bakteriliigi piires ei varieeru G + C sisaldus rohkem kui 5% (Rosselló-Móra ja Amann, 2001). Seda fakti kasutatakse bakteriliikide eristamisel.

Eri bakteriliikide ja -tüvede määramiseks saab kasutada erinevaid bakteri „sõrmejälje" võtmise meetodeid (*fingerprinting*) (Emerson *et al.*, 2008). Enamik neist meetoditest baseeruvad PCR-l. Näiteks disainitakse praimerid bakterigenoomides leiduvatele kordustele ning võrreldakse erinevatest genoomidest saadud PCR-i produkte (rep-PCR). Samuti võrreldakse sama praimeripaariga saadud produktide pikkuseid bakteriti. Oluliseks meetodiks on ka restriktsioonanalüüs, mille käigus võrreldakse restriktaasidega töödeldud DNA järjestuste pikkusi. Baktereid saab määrata ka liigi- või tüvespetsiifiliste PCR praimerite abil (Balboa *et al.*, 2011).

DDH meetodit peetakse liikide määramise „kuldseks standardiks". Meetod võimaldab bakterite genoome võrrelda paarikaupa (Rosselló-Móra ja Amann, 2001). Esmalt viiakse läbi denatureerimine, mille käigus lahutatakse mõlema bakteri DNA kaksikheeliksite ahelad. Seejärel lastakse tekkinud üksikahelatel uuesti hübridiseeruda. Lähedased bakteriliigid sisaldavad palju homoloogilisi piirkondi ning seetõttu tekivad hübriidsed DNA kaksikahelad ehk heterodupleksid. Arvutatakse saadud hübriidsete paardumiste protsent (RBR, *relative binding ratio*). Mõned DDH-l põhinevad meetodid esitavad RBR-i asemel homo- ja heterodupleksite sulamistemperatuuride vahe (ΔT_m). Kokkuleppeliselt määratakse bakterid samasse liiki kuuluvateks, kui RBR ületab 70% või ΔT_m on väiksem kui 5°C (Wayne *et al.*, 1987). DDH metoodikat soovitatakse kasutada sama bakteriliigi erinevate tüvede määramiseks ning juhul, kui kahe uuritava bakteri 16S rRNA geeni järjestuste sarnasus on suurem kui 97% (meetodi kirjeldus allpool) (Tindall *et al.*, 2010). DDH meetodi põhilised puudused on töömahukus ja suur ajakulu (Gevers *et al.*, 2005). Lisaks ei võimalda metoodika kumulatiivse andmebaasi loomist.

DDH meetodite edasiarendusena on pakutud fluorestseeruva SYBR Green I värvi kasutamist, mis seondub kaheahelalise DNA-ga ning võimaldab selle kogust spektrofotomeetriliselt mõõta (Gonzalez ja Saiz-Jimenez, 2002). Loveland-Curtze *et al.* (2011) on kirjeldanud metoodikat, kus DNA ahelate denatureerumist ja uuesti paardumist vaadeldakse reaalaja PCR-i süsteemis ning mõõdetakse fluorestsentsi taset iga 7 sekundi järel. Erinevalt teistest DDH meetoditest piisab fluoromeetrilise DDH puhul ka väiksest uuritava DNA hulgast (Loveland-Curtze *et al.*, 2011). Paralleliseerimaks DDH-l põhinevaid katseid on loodud mikrokiibid, mis võimaldavad uuritavat proovi testida korraga paljude erinevate bakterite täisgenoomide vastu (Wu *et al.*, 2008).

Üheks levinumaks fülogeneetiliseks markeriks, mida bakterite määramiseks kasutatak-

se, on 16S rRNA geen. Üldiselt eeldatakse, et bakterite rRNA geenide järjestuste varieeruvus peegeldab nende bakterite omavahelist evolutsioonilist kaugust. 16S rRNA geeni järjestuse teadasaamiseks amplifitseeritakse seda universaalsete PCR-i praimeritega ning seejärel sekveneeritakse Sangeri meetodil. Ribosomaalsete RNA järjestuste jaoks on loodud andmebaase ning bioinformaatilisi tööriistu, mis hõlbustavad bakteriliikide määramist (Chun *et al.*, 2007). 16S rRNA geeni kasutamist bakterite määramiseks piirab asjaolu, et mõnel juhul on rRNA geenide liigisisene varieeruvus kuni 5%. Samuti on juhtumeid, kus erinevatest liikidest bakterite rRNA geenid on üle 99% ulatuses homoloogsed (Mende *et al.*, 2013).

Tänu teise põlvkonna sekveneerimismeetodite laialdasele levikule on loodud mitmeid meetodeid, mis võimaldavad bakterite täisgenoome ja geenide järjestusi kasutades liike ja tüvesid määrata *in silico*. Traditsioonilisi genoomide hübridiseerumisel põhinevaid metoodikaid on samuti võimalik jäljendada arvutis (Auch *et al.*, 2010). DDH asendusena on pakutud veel keskmise nukleotiidide identsuse (ANI, *average nucleotide identity*) ja aminohapete identsuse (AAI, *amino acid identity*) ning muude sarnaste indeksite arvutamist. On leitud, et ANI indeksi väärtus 94% vastab DDH RBR-i 70%-le (Konstantinidis ja Tiedje, 2005). Richter ja Rosselló-Móra (2009) on ANI indeksit soovitanud kombineerida bakteriliigi koodonkasutust iseloomustavate parameetritega. Kuigi ANI arvutamine on lihtsam ja kiirem kui DDH, on ta suuremõõtmeliste uurimuste korral siiski küllaltki arvutusmahukas (Mende *et al.*, 2013).

Alternatiivseks meetodiks on kasutada bakterite määramiseks mitmete markergeenide järjestusi (MLSA, *Multilocus Sequence Analysis*). Geneetilisteks markeriteks sobivad geenid, mis on enamikele bakteritele universaalsed, leiduvad genoomides ühes koopias (Ciccarelli *et al.*, 2006) ning millega ei toimu horisontaalset geeniülekannet (Sorek *et al.*, 2007). Mende *et al.* (2013) on tuvastanud 40 nendele tingimustele vastavat valku kodeerivat geeni, mille järjestuste alusel saab uuritavaid baktereid klasterdada taksonoomilistesse ühikutesse. MLSA meetodi lihtsustamiseks liidetakse enamasti kõik uuritavad geenijärjestused kokku ning analüüsitakse saadud pikki järjestusi (Gevers *et al.*, 2005).

Kui MLSA metoodika kasutamisel grupeeritakse genoome geenide nukleotiidsete järjestuste alusel, siis MLST (*Multilocus Sequence Typing*) meetodi puhul vaadatakse genoomis esinevate geenialleelide erinevaid kombinatsioone. Sellist lähenemist põhjendatakse asjaoluga, et horisontaalne geeniülekanne ei põhjusta mitte punktmutatsioone, vaid pikemaid polümorfisme (Didelot ja Maiden, 2010). MLST meetodis kasutatakse enamasti „koduhoidja"-geene (*housekeeping genes*), mis vastutavad põhiliste bakteri elutegevuseks vajalike funktsioonide eest. Saadud alleelikombinatsioonide põhjal saab baktereid grupeerida taksonoomilistesse ühikutesse. „Koduhoidja"-geenide suure varieeruvuse tõttu saab MLST metoodikat kasutada vaid lähedaste

bakteritüvede eristamiseks (Maiden *et al.*, 2013).

Bakterite määramiseks metagenoomika andmestikest geenijärjestuste alusel on kirjutatud mitmeid bioinformaatilisi tööriistu. Mõned näited on MetaPhlAn (Segata *et al.*, 2012) ja WebCARMA (Gerlach *et al.*, 2009), mis baseeruvad järjestuste homoloogiaotsingul andmebaasi vastu. NBC (Rosen *et al.*, 2011) ja PhyloPythiaS (McHardy *et al.*, 2007) klassifitseerivad järjestusi nende nukleotiidse koostise alusel.

MLSA ning MLST meetodid piirduvad vaid väikese arvu geenide piirkonnas paiknevate järjestuste võrdlemisega. Juba ainuüksi see kitsendus jätab suure hulga järjestusepõhist informatsiooni vaatluse alt välja. Seetõttu on kirjeldatud bakterite määramise meetodeid, mis analüüsivad nukleotiidseid järjestusi kogu genoomi ulatuses. Nimetatud meetodid baseeruvad oligomeeridel, mis on defineeritud kui mõne kuni mõnekümne nukleotiidi pikkused DNA järjestused. Juba Karlin ja Burges (1995) täheldasid, et baktereid saab määrata nende genoomis leiduvate dinukleotiidide jaotuse alusel. 3- kuni 5-nukleotiidseid oligomeere kasutatakse erinevate bakterite koodonkasutuse iseloomustamiseks (Teeling *et al.*, 2004). Wood ja Salzberg (2014) ning Hasman *et al.* (2014) on pakkunud välja metoodikad, mis põhinevad pikemate oligomeeride jaotustel. Need meetodid on efektiivsed eeldusel, et nimetatud jaotused erinevad bakteriliigiti märkimisväärselt. Teine võimalus on leida iga bakteriliigi või -tüve jaoks spetsiifilised oligomeerid. Metagenoomika proovidest bakterite määramiseks võib tüvespetsiifilisi oligomeere otsida otse sekveneerimislugemistest (Tu *et al.*, 2014a) või disainida neist proovid bakterite määramiseks mõeldud mikrokiibile (Tu *et al.*, 2013).

1.3 Bakterispetsiifilistel oligomeeridel põhinevate määramismeetodite bioinformaatiline problemaatika

Bioinformaatiliste rakenduste kontekstis nimetatakse pikkusega k oligomeere enamasti k-meerideks. Näiteks võib 20 nukleotiidi pika oligomeeri kohta öelda 20-meer. Aeg-ajalt nimetatakse k-meere ka sõnadeks pikkusega k. Terve genoomse järjestuse saab jaotada üksteisega ülekattes paiknevateks k-meerideks või sõnadeks. Selliselt saadud k-meere saab kasutada bakteritüvede, -liikide või -perekondade määramiseks. Selleks võib iga bakteri puhul vaadata tema genoomis paiknevate k-meeride jaotust (Hasman *et al.*, 2014) või leida bakteritele genoomispetsiifilisi k-meere (Tu *et al.*, 2014a).

Bakterispetsiifilisi k-meere saab kasutada mitmes erinevas bakterite määramise metoodikas. Ka spetsiifiliste PCR-i praimerite ja mikrokiibi proovide disain taandub eelkõige bakteriperekonnale, -liigile või -tüvele iseloomulike k-meeride otsimisele. Lisaks saab kõiki või valitud alamhulka bakterispetsiifilistest k-meeridest kasutada markerjärjestustena. Otsides neid markereid teise põlvkonna sekveneerimislugemitest, saab määrata sekveneeritud proovi bakterilist koosseisu.

Kõik spetsiifilistel k-meeridel põhinevad bakterite määramise meetodid baseeruvad kahel omavahel seotud eeldusel. Esimeseks eelduseks on, et teatud pikkusest pikemate sõnade hulk genoomis on tunduvalt väiksem kui sama pikkusega kõikvõimalike sõnade hulk. Näiteks on erinevaid võimalikke 16-meere $4^{16} \approx 4.3$ miljardit, mis on keskmiselt ligikaudu 1000 korda suurem, kui bakteri genoomi pikkus. Vaadates 16 nukleotiidist pikemaid k-meere, suureneb saadud vahe veelgi. Lisaks sisaldab bakterigenoom korduvaid järjestusi (Delihas, 2008), mistõttu on genoomis leiduvate unikaalsete k-meeride arv väiksem kui genoomi enda pikkus. Teine eeldus, millega genoomispetsiifilistel k-meeridel põhinevad meetodid arvestavad, on see, et erinevate bakterite genoomide järjestused on suures osas väga varieeruvad (Reva ja Tümmler, 2004). Seega sisaldab iga bakteritüvi erinevat alamhulka kõikvõimalike k-meeride hulgast ning iga tüve jaoks on võimalik leida talle spetsiifilised k-meerid.

1.3.1 Bakterispetsiifiliste PCR-i praimerite disain

Genoomispetsiifilisi PCR-i praimereid disainitakse mõne konkreetse geeni ümbrusesse. Tihti kasutatakse selleks 16S rRNA geeni (Osorio *et al.*, 1999; Trkov ja Avgustin, 2003), kuid sobivad on ka teised bakteri elutegevuseks hädavajalikud geenid (Nhung *et al.*, 2007). Kui uuritava bakteriliigi või -perekonna esindajate genoomid sisaldavad mõnda ainult neile iseloomulikku geenijärjestust, sünteesitakse praimerid selle geeni ümbrusesse (Zhai *et al.*, 2014). Praimerite

disainiks vajamineva geenijärjestuse võib saada nukleotiidide andmebaasist või ise universaalsete praimeritega geeni amplifitseerides ja Sangeri meetodil sekveneerides.

Spetsiifilisi PCR-i praimereid saab disainida bakterite erinevatele taksonoomilistele tasemetele. Ühele bakteriperekonnale spetsiifiliste praimerite disainil valitakse võimaluse korral märklauaks mõni uuritavale perekonnale spetsiifiline geen ning potentsiaalseteks praimerite seondumiskohtadeks piirkonnad, mis on homoloogsed kõigile sinna perekonda kuuluvatele liikidele. Kui soovitakse leida praimerijärjestusi, mille abil saaks eristada ühte bakteriliiki või -tüve teistest samasse perekonda kuuluvatest bakteritest, valitakse seondumiskohtadeks genoomipiirkonnad, mis ei oma teiste bakterite genoomidega suurt homoloogiat.

Mitme bakteriliigi määramiseks ühest ja samast proovist võib kasutada terveid praimerite süsteeme, milles iga praimeripaar on erinevale liigile spetsiifiline (Batra *et al.*, 2013). Liigispetsiifilised võivad olla ka vaid mõned kõigist praimeritest. Heaks näiteks on Nhung *et al.* (2007) disainitud süsteem viie erineva *Vibrio* liigi detekteerimiseks. Kokku kasutavad nad kuut praimerit, millest üks on universaalne kõigile viiele *Vibrio* liigile ning ülejäänud on liigispetsiifilised. Iga uuritava liigi kohta saadav produkt on erineva pikkusega.

Praimerite disainil on oluline silmas pidada, et praimerina ei saa kasutada igat sobiva pikkusega (praimerite puhul vahemik 16 kuni 28 nukleotiidi) liigispetsiifilist oligomeeri. Amplifikatsiooni edukaks toimumiseks tuleb arvestada ka praimerite G + C sisaldust, 3' otsa järjestust, sulamistemperatuuri ja soovitud PCR-i produkti parameetreid (Chuang *et al.*, 2013).

Praimerite disainiks on loodud mitmeid programme, millest enimkasutatav on Primer3 (Untergrasser *et al.*, 2012). Primer3 võtab sisendiks DNA järjestuse, millele soovitakse praimereid disainida. Sisendjärjestuses võivad need nukleotiidid, mida praimeri koosseisu ei soovita, olla maskeeritud (asendatud tähega N). Seega saab kasutaja praimerite seondumiskohtadeks valida just need piirkonnad, mis on uuritavale bakterile spetsiifilised. Järjestuse maskeerimiseks saab kasutada *k*-meeridel põhinevat paketti GenomeMasker (Andreson *et al.*, 2006). Praimerite spetsiifilisuse testimiseks kasutatakse enamasti programme BLAST (*Basic Local Alignment Search Tool*) (Altschul *et al.*, 1997) või MEGABLAST (Zhang *et al.*, 2000), millega saab otsida potentsiaalseid alternatiivseid seondumiskohti. Samuti on kirjutatud programme PCR-i amplifikatsiooni simuleerimiseks *in silico* (Andreson *et al.*, 2006).

PCR-iga saab ühest proovist samaaegselt tuvastada vaid loetud arvu baktereid. See metoodika sobib, kui soovitakse detekteerida, kas proov sisaldab ühte või mõnda konkreetset bakteriliiki (näiteks diagnostikas patogeeni), kuid sellega ei saa analüüsida terveid bakterite kooslusid. Samuti ei võimalda PCR tuvastada veel avastamata bakteriliike. Oluliseks piiranguks on ka nõuded PCR-i praimerite järjestustele. PCR-i õnnestumine sõltub praimeri seondumiskoha

stabiilsusest. Kui seondumise piirkonnas tekib mutatsioon, ei pruugi disainitud praimer enam töötada.

1.3.2 Bakterispetsiifiliste mikrokiibiproovide disain

Mikrokiibid võimaldavad iga bakteri kohta kasutada rohkem kui vaid paari spetsiifilist oligomeeri. Seega pole meetod nii tundlik üksikutele mutatsioonidele. Samuti saab korraga määrata suurema arvu baktereid. Sellest hoolimata jäävad PCR-i praimerite järjestustele seatud kitsendused kehtima ka kiibiproovide korral. Proovide töökindlus sõltub sarnaselt praimeritele sulamistemperatuurist, G + C sisaldusest ning proovi pikkusest. Vältida tuleks proove, mis sisaldavad pikemaid ühenukleotiidseid korduseid (Mulle *et al.*, 2010). Kiibiproovidena kasutatakse oligomeere pikkuses 20 kuni 70 nukleotiidi (Liu *et al.*, 2010).

Sarnaselt PCR-i praimeritele disainitakse mikrokiipide proovid enamasti genoomi kodeerivatele aladele. Peamiseks põhjuseks on see, et kodeerivad alad on rohkem konserveerunud eri bakteritüvede vahel (Liu *et al.*, 2010). Mikrokiipe on loodud nii paari üksiku kui ka üle 300 erineva bakteriliigi detekteerimiseks. Esimeste hulka kuuluvad kiibid patogeenide tuvastamiseks peremeesorganismist või toidust (Goji *et al.*, 2012; González *et al.*, 2004). Paljude erinevate liikide proove sisaldava mikrokiibi näitena võib tuua HuMiChip kiibi, mis on disainitud inimese mikrobioomi koosluse määramiseks (Tu *et al.*, 2014b).

Sobivaid proove saab disainida praimeridisaini programmidega, näiteks Primer3-ga (Untergrasser *et al.*, 2012). Eraldi kiibiproovide disainiks on kirjutatud ka erinevaid kõrgemataseme programme, mis kasutavad Primer3-e ja/või BLAST-i (Altschul *et al.*, 1997) funktsionaalsust. Sellised programmid on näiteks PRIMERGENS (Xu *et al.*, 2002), OligoComm (Li *et al.*, 2005) ja OligoArray (Rouillard *et al.*, 2003). Tu *et al.* (2013) on loonud *k*-meeridel põhineva metoodika mikrokiibi proovide disainiks.

1.3.3 Bakterispetsiifilistel markerjärjestustel põhinevad tehnoloogiad

Bioinformaatiliste tööriistade abil on võimalik leida bakterispetsiifilisi *k*-meere kogu genoomi ulatusest. Leides sellised *k*-meerid kõigile sekveneeritud bakterigenoomidele, saab luua andmebaasi, mille abil saaks määrata teadaolevate või nendega päritolult lähedaste bakterite liike. Vajaliku andmebaasi loomiseks tuleb esmalt otsitavatele *k*-meeridele määrata optimaalne pikkus. Kui kasutada liiga lühikesi *k*-meere, siis võib juhtuda, et mõne bakteritüve jaoks spetsiifilisi *k*-meere ei leidu (Tu *et al.*, 2014a). Liiga pikkade *k*-meeride korral leitakse küll palju genoomispetsiifilisi sõnu, kuid selle võrra suureneb ka loodava andmebaasi maht. Kui kasutada

sobiva pikkusega *k*-meere, saab koostatud andmebaasi kasutada ka bakterispetsiifiliste PCR-i praimerite või kiibiproovide disainil (Tu *et al.*, 2013).

Bakteri genoomispetsiifiliste *k*-meeride leidmisel tuleb kõigepealt luua nimekiri kõigist mittespetsiifilistest *k*-meeridest. Sinna nimekirja kuuluvad *k*-meerid, mis sisalduvad teistes teadaolevate bakterite genoomides ning mida seega ei saa kasutada uuritava bakteri määramiseks. Mõistlik on lisada mittespetsiifiliste *k*-meeride nimekirja ka kõik inimese genoomis leiduvad *k*-meerid, sest uuritavate bakterite proovid võivad laboritingimustes kergesti inimese DNA-ga saastuda. Kui bakterid on võetud mõnest teisest organismist, siis peaks mittespetsiifiliste *k*-meeride nimekiri sisaldama ka peremeesorganismi genoomsest järjestusest pärit *k*-meere. Uuritava bakteri genoomile spetsiifilisi *k*-meere saab leida, võrreldes kõiki selle bakteri *k*-meere mittespetsiifiliste *k*-meeride nimekirjaga ning märkides üles need *k*-meerid, mida viimases ei leidu. (Tu *et al.*, 2014a)

Saadud bakterispetsiifiliste *k*-meeride andmebaasi saab kasutada teise põlvkonna sekveneerimisandmetest bakterite liigilise koostise määramiseks. Selleks võib libiseva aknaga liikuda mööda sekveneerimislugemeid ning märkida üles kõik need leitud *k*-meerid, mis on mõnele liigile spetsiifilised. Lõpuks saab määrata, milliste bakteriliikide *k*-meeride osakaal uuritavas proovis oli statistiliselt oluline.

Markerjärjestuste kasutamine bakterite määramiseks on oligomeeridel põhinevatest metoodikatest kõige stabiilsem, sest see pole eriti tundlik üksikutes genoomipiirkondades toimuvate mutatsioonide või sekveneerimisvigade suhtes. Kogu analüüs toimub *in silico* ning on seega kiirem ning saadud tulemused usaldusväärsemad. Samuti on kerge katsetulemusi korrata ning kord sekveneeritud proove saab analüüsida ka tagantjärele.

1.4 Bioinformaatilised algoritmid ja programmid

1.4.1 Nukleotiidide kodeerimine binaarkujule

Nukleotiidseid järjestusi saab digitaalsele kujule kodeerida mitmel moel. Inimesele on kõige arusaadavam talletada iga nukleotiid ühebaidise tähemärgina: A, C, G või T (RNA-s U). Taoline tähistus on tekstifaili trükitult lihtsasti loetav. Arvuti jaoks on sellise tähistuse korral iga ühebaidine tähemärk kodeeritud 8-bitiseks märgita täisarvuks. Iga bitt võrdub ühega kahendsüsteemi sümbolitest 0 või 1. Kaheksabitine täisarv võimaldab esitada kuni $2^8 = 256$ erinevat sümbolit. Et nukleotiide on ühes järjestuses ainult neli erinevat, kulutab taoline kodeering asjatult mälu.

Nukleotiidsete järjestustega manipuleerivate programmide töös kodeeritakse üks nukleotiid enamasti kahebitiseks järjestuseks. Nukleotiidi A tähistatakse bittidega 00, nukleotiidi C bittidega 01, nukleotiidi G bittidega 10 ja nukleotiidi T (U) bittidega 11. Sarnaselt on võimalik arvudena esitada terveid oligomeere. Näiteks järjestust GGTACC saab kodeerida 12 bitiks 101011000101. Saadud bittide järjestus kujutab kahendsüsteemi täisarvu ning seda saab omakorda esitada täisarvuna kümnendsüsteemis. Eespool mainitud järjestusele vastab kümnendsüsteemis arv 2757.

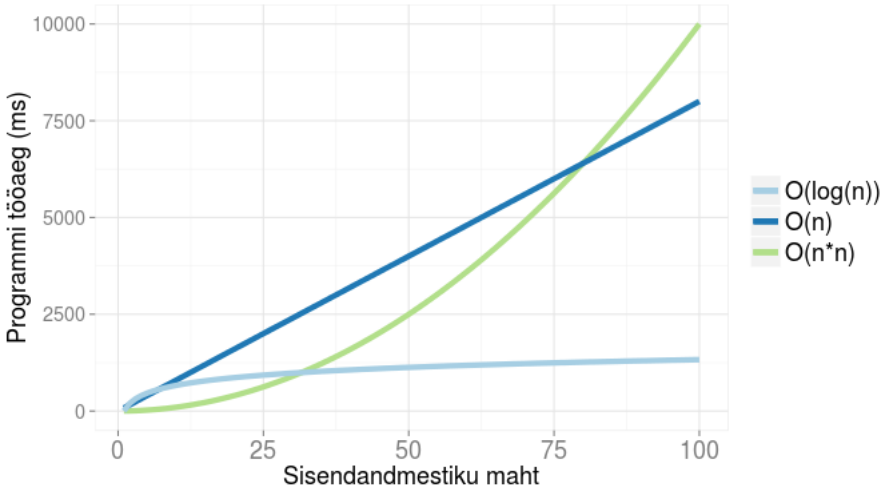
1.4.2 Algoritmide keerukus

Algoritmide võrdlemiseks ja paremuse järgi reastamiseks kasutatakse nende ajalist ja ruumilist keerukust. Ajaline keerukus väljendab seda, millises proportsioonis pikeneb algoritmi tööaeg sisendandmemahu lineaarsel suurenemisel. Ruumiline keerukus näitab, kuidas mõjutab sisendandmemahu suurendamine algoritmi mälukasutust. Keerukust tähistatakse tähega O.

Olgu meil sisendandmestik suurusega n kirjet ning kaks algoritmi, mis selle andmestiku töötlemisel teevad vastavalt $f(n)$ ja $g(n)$ elementaaroperatsiooni. Keerukuse matemaatiline definitsioon ütleb, et funktsiooni $f(n)$ keerukus on võrdne või väiksem kui funktsiooni $g(n)$ keerukus (kirjutatakse $f(n) \sim O(g(n))$), kui leidub selline andmemahu suurus n_0 ja positiivne reaalarv c, et iga $n \geq n_0$ korral kehtib

$$f(n) \leq c * g(n).$$

Harilikult kasutatakse $O(g(n))$ kohal kokkuleppelisi keerukusklasse. Keerukusklasse saab järjestada. Levinumate keerukusklasside järjestus, alustades väikseimast keerukusest, on järgmine: $O(1)$ (konstantne), $O(\log n)$ (logaritmiline), $O(n)$ (lineaarne), $O(n \log n)$, $O(n^a)$ (polünomiaalne), $O(a^n)$ (eksponentsiaalne), $O(n!)$ (faktoriaalne). Oluline on märkida, et väikeste



Joonis 1: Kolme hüpoteetilise programmi tööaegade võrdlus: *Kui lineaarse (tumesinine) või logaritmilise (helesinine) keerukusega algoritmide üks elementaaroperatsioon on palju aeglasem kui ruutkeerukusega (roheline) algoritmi elementaaroperatsioon, siis väikestel andmemahtudel võib ruutkeerukusega algoritm töötada palju kiiremini (nt. andmemahtude vahemikus* 1 *kuni* 25*). Andmemahtude kasvamisel jääb ruutkeerukuga funktsiooni töökiirus aga väiksemate keerukustega algoritmide töökiirustele alla.*

andmemahtude juures võib suurema ajalise keerukusega funktsioon olla kiirem kui väiksema ajalise keerukusega funktsioon, kuid andmemahu kasvades tagab väiksem keerukus alati suurema ajalise võidu (vt. joonis 1). Tihti eristatakse algoritmidel keskmist ja halvima juhu keerukust.

Keerukusklasside olemust illustreerib järgmine näide: Olgu meil algoritm, mis teeb n suurusega sisendi puhul $f(n)$ elementaaroperatsiooni ja kehtigu

$$f(n) \sim O(n^2).$$

Öeldakse, et vaadeldava algoritmi keerukus on piiratud ruutkeerukusega. Kui taolise algoritmi sisendandmestikku suurendada kümme korda, suureneb algoritmi elementaaroperatsioonide arv ja seega ka tööaeg sada korda.

1.4.3 Otsingualgoritmid

Otsingualgoritmid on üks laialdasemalt kasutatavaid ja paremini läbiuuritud algoritmide klasse. Neid kasutatakse olemasolevast andmete nimekirjast, massiivist või tabelist ühe konkreetse objekti või kirje otsimiseks. DNA järjestuste analüüsil on andmeteks näiteks nimekiri kõigist uuritava organismi genoomis leiduvatest k-meeridest. Olenevalt uurimisküsimusest võib iga kirje sisaldada ka infot k-meeri asukohtadest ja esinemissagedusest genoomis. Genoomset

järjestust ennast võib samuti pidada k-meeride massiiviks, kus k-meerid on kirjutatud järjestikku ja osaliselt ülekattes. Potentsiaalsed uurimisküsimused on näiteks, kas uuritav genoom sisaldab ühte või teist k-meeri või millistes genoomipositsioonides huvipakkuv k-meer paikneb. Mõlemale vastamiseks tuleb, olenemata k-meere sisaldava andmestruktuuri tüübist, kasutada otsingualgoritme.

Kõige triviaalsem otsingualgoritm seisneb otsitava elemendi järjest võrdlemises kõigi andmemassiivis leiduvate elementidega. Selline algoritm sõltub andmemassiivi pikkusest lineaarselt ehk algoritmi keerukus on $O(n)$. Kirjeldatud lineaarset otsingualgoritmi on väga lihtne implementeerida ning andmete massiiv ei vaja mingit täiendavat töötlust. Seetõttu on lineaarne otsing mõistlik, kui ühte ja sama andmestikku ei kasutata korduvalt ning otsitavate elementide arv on väike.

Kui sama andmestikku soovitakse palju kordi otsimiseks kasutada, on targem implementeerida väiksema keerukusega otsingualgoritm. Üheks kõige tuntumaks otsingualgoritmiks on kahendotsing (*binary search*). Kahendotsingu halvima juhu keerukus on $O(\log n)$. Kahendotsingu kasutamiseks tuleb andmete massiiv esmalt järjestada. Selleks peab leiduma funktsioon, mis ütleb iga kahe erineva massiivielemendi kohta, kumb neist on massiivis eespool. Universaalsed sorteerimisalgoritmid (näiteks kiirsort) on keskmise juhu keerukusega $O(n \log n)$. Täisarve saab aga bitiblokkide kaupa sorteerida ka lineaarse keerukusega (*radix-sort*). Peatükis 1.4.1 kirjeldatud meetodil on igat k-meeri võimalik esitada täisarvuna. Seega saab k-meere järjestada neile vastavate täisarvude alusel. Kuigi sorteerimine on kahendotsingust keerukam algoritm, tuleb seda teha vaid üks kord andmebaasi ettevalmistamisel. Hiljem saab samast sorteeritud massiivist elemente korduvalt otsida.

Kahendotsingu korral võrreldakse otsitavat elementi esmalt andmestiku keskmise elemendiga. Kuna andmestik on järjestatud, siis on võimalikud kolm varianti: otsitav element võrdub keskmise elemendiga, asub keskmisest elemendist massiivis eespool (on väiksem) või asub keskmisest elemendist massiivis tagapool (on suurem). Esimesel juhul otsing lõpetatakse ning algoritm tagastab elemendi asukoha massiivis. Teistel juhtudel korratakse otsingusammu, nüüd aga juba ainult poolega algsest massiivist. Seega väheneb andmestiku hulk, millest elementi otsitakse, iga sammuga kaks korda. Otsingusammu korratakse otsitava elemendi leidmiseni või kuni alammassiiv, millest elementi otsitakse, ei sisalda rohkem elemente.

Kui on vaja teha suurtest andmetabelitest väga palju otsinguid, võib implementeerida konstantse keerukusega otsingualgoritmi. Sellised algoritmid eeldavad enamasti andmete hoidmist keerulisemates andmestruktuurides nagu paisktabelid või puud. Mõlemal juhul on võimalik implementeerida algoritm, mis leiab suvalise elemendi struktuurist fikseeritud arvu

elementaaroperatsioonidega (keerukus $O(1)$). Kuigi otsing nendest andmestikest on väga kiire, võib vajaliku andmestruktuuri loomine ja andmetega täitmine osutuda ajakulukaks. Seega on nende kasutamine õigustatud eelkõige siis, kui saadud andmestruktuure hoitakse staatilistena ning ei muudeta oluliselt kasutamisperioodi vältel (näiteks ei ühendata kahte andmestruktuuri omavahel). Nimetatud struktuure on tihti keeruline failidesse kirjutada. Samuti on nad enamasti suured, sest lisaks elementidele on vaja säilitada veel informatsiooni elementidevaheliste seoste kohta.

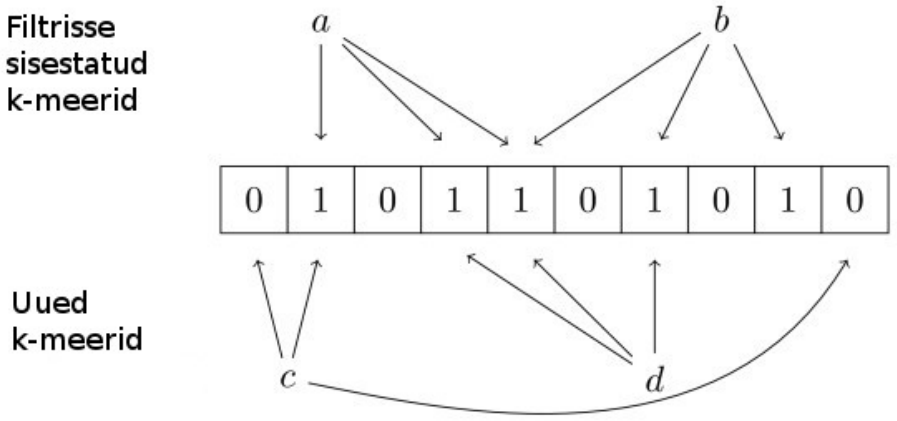
1.4.4 Algoritmid *k*-meeride loendamiseks genoomist või sekveneerimislugemitest

Uuritava järjestuse (genoom või sekveneerimislugemid) kõigi *k*-meeride esinemise sageduste leidmist nimetatakse *k*-meeride loendamiseks. Näiteks sisaldab järjestus AGTAGTAGT nelja ülekattes 6-meeri. Unikaalseid 6-meere on toodud järjestuses kolm: AGTAGT, GTAGTA ja TAGTAG. Neist esimese sagedus on võrdne kahega ning ülejäänute sagedused on võrdsed ühega. Sellist loendamist kasutatakse genoomide kokkupanemisel (Miller *et al.*, 2008), vigaste lugemite parandamisel (Kelley *et al.*, 2010), järjestuste joondamisel (Edgar, 2004), liikide tuvastamisel (Hasman *et al.*, 2014) ja mujal.

K-meeride loendamiseks genoomsetest järjestustest on loodud mitmeid algoritme ja bioinformaatilisi tööriistu. Enamasti on nende väljundiks sõnastikulaadne andmestruktuur või tabel, kus igale genoomist leitud *k*-meerile vastab tema sagedus. Sõnade naiivselt järjest lisamine sellisesse tabelisse on väga ajakulukas. Seetõttu on välja mõeldud mitmeid *k*-meeride loendamist kiirendavaid algoritme.

Celera paketti (Miller *et al.*, 2008) kuuluv *k*-meeride lugeja Meryl jaotab *k*-meerid nende järjestuste põhjal esmalt 2^{24} alamhulgaks, mis hiljem grupisiseselt sorteeritakse. Et kõiki alamhulki hoitakse korraga mälus, siis kulutab ülaltoodud algoritm suurte genoomide korral sadu gigabaite operatiivmälu.

Tallymer (Kurtz *et al.*, 2008) kasutab oma algoritmis sufiksite ehk järelliidete massiivi (*suffix array*). Järjestuse sufiksiks nimetatakse igat tema alamjärjestust, mis lõppeb järjestuse viimase nukleotiidiga, ja prefiksiks igat alamjärjestust, mis algab järjestuse esimese nukleotiidiga. Algoritmi töö käigus järjestatakse kõik uuritava järjestuse sufiksid leksikograafiliselt ning nende järjekorranumbrid salvestatakse sufiksite massiivi. Lisaks talletatakse iga sufiksi kohta ka tema ja talle eelneva sufiksi pikima ühise prefiksi pikkus. Kasutades alamjärjestuste sisalduvust üksteises kui eellane-järglane suhet, moodustatakse puustruktuur, mille juur sisaldab kõiki alamjärjestusi (on algne järjestus ise) ja lehttippudes olevad alamjärjestused sisalduvad kõigis



Joonis 2: K-meeride lisamine Bloomi filtrisse: *Iga k-meeri lisamisel kontrollitakse sellele k-meerile vastavaid bitte filtris. Juba olemasolevate k-meeride (a, b) korral on kõik vastavad bitiväärtused võrdsed* 1*-ga. Uute k-meeride (c, d) korral peaks mõni bitiväärtustest olema* 0*, kuid aeg-ajalt leitakse mõni valepositiivne (d). Bloomi filtri abil saab leida k-meere, mis leiduvad uuritavas järjestuses rohkem kui ühe korra. (Melsted ja Pritchard, 2011)*

neile eelnevates järjestustes. Saadud struktuurist saab üheselt lugeda iga alamjärjestuse (ehk *k*-meeri) sageduse algses järjestuses.

K-meeride loendamise kiirust saab tõsta ka algoritmi paralleliseerimisega mitme protsessoriga arvutis. Paralleliseerimine tekitab probleeme, kui ühe programmi mitu lõime (paralleelselt töötavad programmi osad) üritavad samaaegselt muuta sama mälupiirkonda. Seepärast kaasneb paralleliseerimisega harilikult ühise mäluosa lukustamine korraga vaid ühe lõime jaoks. Kui see mälupiirkond on aga keskne programmi töös (näiteks andmestruktuur, kuhu pidevalt sisestatakse *k*-meere ning muudetakse nende esinemissagedusi), võib taoline lukustamine kaotada paralleelse algoritmi kõik eelised. Mitmelõimeline *k*-meeride loendamise programm Jellyfish (Marçais ja Kingsford, 2011) kasutab lukustamise asemel CAS (*compare-and-swap*) operatsiooni, millega saab andmeid mingis mälupiirkonnas muuta, ilma et ükski teine lõim seda samaaegselt teha saaks.

BFCounter (Melsted ja Pritchard, 2011) on heuristiline programm, mis loendab *k*-meere, mida esineb järjestuses rohkem kui üks kord. Ta kasutab andmestruktuurina Bloomi filtrit. Bloomi filter on bitivektor, mille iga bitt on algselt väärtusega null. Lisaks defineeritakse funktsioonid $h_1, ..., h_n$, millest igaüks seab vaadeldava *k*-meeriga vastavusse ühe positsiooni selles vektoris. Sisestades *k*-meeri Bloomi filtrisse, seatakse kõikide nende funktsioonide poolt määratud positsioonide bitid ühtedeks. Et enamikele *k*-meeridele vastab erinev kombinatsioon bitivektori positsioone, saab iga *k*-meeri kohta otsustada, kas ta juba sisaldub Bloomi filtris või mitte (vt.

joonis 2). Kasutades taolist filtrit, saab sõnastiku struktuuri sisestada ainult need *k*-meerid, mida on juba korra varem nähtud. Valepositiivsete arv on Bloomi filtri korral väike.

KMC (*k-mer counter*) (Deorowicz *et al.*, 2013) algoritm on mõeldud sekveneerimislugemitest *k*-meeride sageduste leidmiseks. KMC kirjutab *k*-meerid nende prefiksite alusel ajutistesse failidesse. Taoline lähenemine võimaldab piirata oluliselt operatiivmälu kasutust ka suurte andmemahtude töötlemisel. Failide sisud sorteeritakse $O(n)$ keerukusega *radix-sort* meetodil, ja ühendatakse. Saadud sõnastikust eemaldatakse *k*-meerid, mille sagedus on väiksem kui kasutaja määratud piirmäär. Lõplikust andmestruktuurist saab *k*-meere otsida kahendotsingu meetodil.

Teine ajutisi faile kirjutav programm on DSK (*disc streaming of k-mers*) (Rizk *et al.*, 2013). DSK kasutab paiskfunktiooni $h(m)$, kus m on parasjagu uuritav *k*-meer. Funktsioon $h(m)$ määrab, millisesse faili m kirjutatakse. Teise sammuna loetakse kõik ühes failis paiknevad *k*-meerid paisktabelisse, mille abil loetakse kokku iga *k*-meeri sagedus.

1.4.5 *K*-meeri põhised algoritmid bakterite määramiseks metagenoomika andmetest

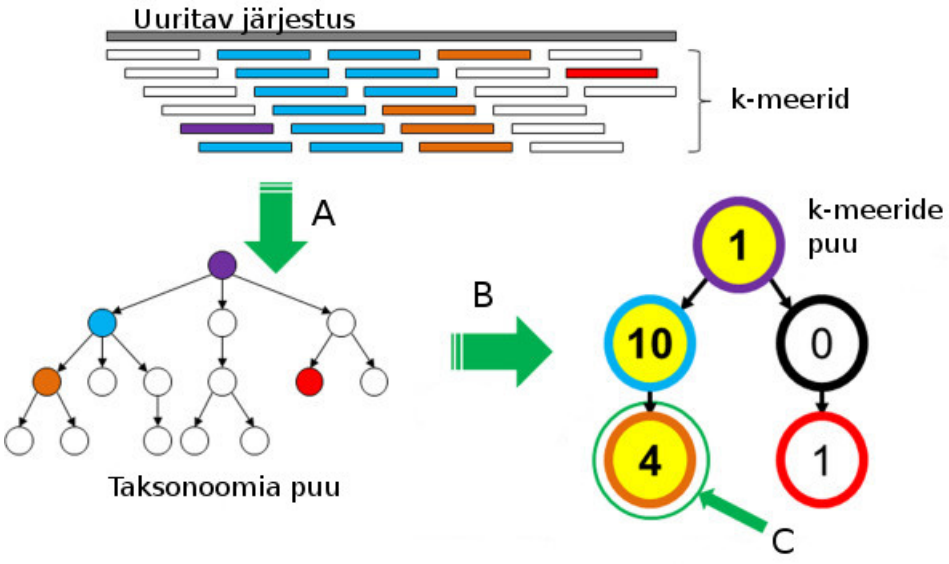
Viimase aasta jooksul on hakatud kirjeldama metoodikaid, mis võimaldaksid *k*-meeride põhjal tuvastada metagenoomika andmetest bakterite perekondi, liike või tüvesid. Nende meetodite realiseerimiseks on esmalt tarvilik luua andmebaas, mis sisaldaks teadaolevate bakterite kohta kõiki nende genoomides leiduvaid *k*-meere või *k*-meeride spetsiifilist alamhulka. Lisaks on vaja algoritmi andmebaasist kirjete otsimiseks ning teooriat, mille põhjal leitud *k*-meeride abil bakterit määrata.

Ühte *k*-meeridel põhinevat bakterite määramise meetodit on kirjeldanud Hasman *et al.* (2014). See meetod põhineb andmebaasil, mille iga kirje sisaldab kahte välja: *k*-meer ja nimekiri teadaolevatest bakterigenoomidest, mis seda *k*-meeri sisaldavad. Andmebaasi mahu kokkuhoidmiseks võib talletada ainult neid *k*-meere, mis algavad kasutaja poolt määratud nukleotiidide kombinatsiooniga. Bakterite määramiseks otsitakse andmebaasist tema genoomis või sekveneerimislugemites leiduvaid *k*-meere. Iga andmebaasis leiduva bakterigenoomi kohta salvestatakse nende *k*-meeride arv, mis leiduvad nii temas kui uuritavas bakteris. Nende arvude alusel leiab algoritm uuritava bakteri genoomiga kõige sarnasema teadaoleva genoomi.

Genoomispetsiifiliste markerjärjestuste leidmiseks bakterite genoomist võib kasutada meetodit GSMer (*genome-specified marker*) (Tu *et al.*, 2014a). Esimese sammuna moodustab GSMer *k*-meeride tabeli, mis sisaldab kõiki inimese genoomis leiduvaid *k*-meere ning *k*-meere, mis leiduvad rohkem kui ühes teadaolevas bakterigenoomis. *K*-meeride loendamiseks kasutab GSMer programmi Meryl (Miller *et al.*, 2008). Loetakse *k*-meere, mille pikkus on

vahemikus 18 kuni 20 nukleotiidi. Seejärel jagatakse bakterigenoom, mille unikaalseid markereid soovitakse leida, 50-nukleotiidseteks fragmentideks (genoomis pikkusega L on selliseid fragmente $L - 50$) ja üritatakse nendele asetada k-meere varem loodud tabelist. Kui mingil 50-nukleotiidsel fragmendil on homoloogia kas või ühe k-meeriga, siis see fragment loetakse uuritavale bakterile mittespetsiifiliseks ja eemaldatakse vaatluse alt. Viimase sammuna eemaldatakse MEGABLAST programmi (Zhang *et al.*, 2000) abil 50-nukleotiidsete fragmentide hulgast ka need, mis annavad mõne teise bakteri või inimese genoomiga suurema kui 85% identsusega joonduse. Allesjäänud 50-nukleotiidsed fragmendid loetakse uuritavale bakterile spetsiifilisteks markerjärjestusteks. Sarnaselt saab genoomispetsiifilisi markereid leida kõigile teadaolevatele bakteritele. Saadud markerjärjestusi saab MEGABLAST-iga otsida metagenoomika andmestikest ning saadud vastete põhjal kirjeldada metagenoomika proovide mikrobioloogilist koosseisu. Lisaks võib alles jäänud 50-nukleotiidsetel markerjärjestustel analüüsida nende G + C sisaldust, sulamistemperatuuri ja sekundaarstruktuuride moodustumist. Saadud informatsiooni põhjal on võimalik disainida bakterite määramiseks sobiva mikrokiibi (Tu *et al.*, 2013).

Kraken (Wood ja Salzberg, 2014) kaasab bakterite määramisse peale k-meeride ka taksonoomilist informatsiooni. Sarnaselt kahele eespool kirjeldatud metoodikale, vajab ka tema eelnevalt koostatud andmebaasi, mis sisaldaks kõikides teadaolevates bakterigenoomides leiduvaid



Joonis 3: Krakeni tööpõhimõte: *Esmalt loendatakse uuritavast järjetusest kõik k-meerid ning leitakse igale k-meerile teda sisaldavate bakterite viimane ühine eellane (A). Saadud taksonoomilist puud kasutatakse uuritavate bakterite määramiseks, sisestades iga puu harukoha kohta, mitu selle eellase genoomis leiduvat k-meeri leiti (B). Viimaks leitakse moodustatud puustruktuurist suurima k-meeride kogusummaga tee (C). (Wood ja Salzberg, 2014)*

k-meere. *K*-meeride loendamiseks kasutab ta programmi Jellyfish (Marçais ja Kingsford, 2011), kuid lisab omalt poolt igale leitud *k*-meerile identifikaatori, mis tähistab kõigi seda *k*-meeri sisaldavate bakterite viimast ühist eellast (vt. joonis 3A). Andmebaasist otsimise hõlbustamiseks on igast *k*-meerist leitud tema leksikograafiliselt väikseim *M*-nukleotiidne fragment ning andmebaas on järjestatud nende fragmentide alusel. *K*-meeri otsimiseks andmebaasist leitakse selle *k*-meeri leksikograafiliselt vähim *M*-nukleotiidne fragment, leitakse indeksite failist sama fragmenti sisaldavate *k*-meeride positsioonid andmebaasis ning sooritatakse selles positsioonide vahemikus kahendotsing.

Krakeni abil bakteri määramiseks loetakse tema genoomist või sekveneerimislugemitest *k*-meere ning leitakse iga *k*-meeri kohta andmebaasist seda sisaldavate bakterite viimane ühine eellane. Lähtudes saadud informatsioonist koostatakse puu, mille iga harukoht tähistab ühte tuvastatud viimast ühist eellast ning sisaldab nende uuritavas genoomis leiduvate *k*-meeride arvu, mis sellele eellasele viitavad (vt. joonis 3B). Pärast puu konstrueerimist liidetakse harukohtades olevate *k*-meeride arvud kõikvõimalikke teid pidi (puu juurest lehttipuni) kokku. Leitakse tee, mille harukohtade summa on maksimaalne. Selliselt leitud tee lehttipus paiknev liik või tüvi peaks uuritava bakteriga päritolult kõige enam sarnanema (vt. joonis 3C). Sama tee harukohad peaks tähistama uuritava bakteri eellaseid. Seega võimaldab leitud tee iseloomustada uuritava bakteri liigilist ja perekondlikku kuuluvust. Kui maksimaalse summaga teid on rohkem kui üks, leitakse nende viimane ühine harukoht ning iseloomustatakse uuritavat bakterit sellest lähtuvalt. (Wood ja Salzberg, 2014)

2 Praktiline töö

2.1 Töö eesmärgid

Tänapäeval kasutatakse bakterite määramiseks eelkõige genoomset informatsiooni. Sagedaseks meetodiks on paljundada PCR-iga ja sekveneerida uuritava bakteri ühe või mitme markergeeni järjestusi. Seejärel saab leitud järjestustele otsida vasteid bakterite geene sisaldavast andmebaasist ning tuvastada seeläbi uuritava bakteri liiki ja päritolu. Teise põlvkonna sekveneerimismeetodite abil saadud andmed võimaldavad aga üksikute geene sisaldavate genoomipiirkondade asemel võrrelda lühikesi unikaalseid järjestusi kogu genoomi ulatusest.

Teadaolevalt on siiani kirjeldatud vaid ühte metoodikat, mis võimaldaks baktereid määrata igale bakteriliigile või -tüvele spetsiifiliste genoomipiirkondade abil (Tu *et al.*, 2014a). Nimetatud töö põhineb spetsiifiliste 50-nukleotiidsete markerjärjestuste leidmisel, kasutades bakterigenoomides esinevaid *k*-meere. Bakterite genoomid on lühikesed ning sisaldavad vähem korduseid kui eukarüootide genoomid. Seega saaks kasutada ka lühemaid spetsiifilisi järjestusi kui 50 nukleotiidi. See võimaldaks kokku hoida andmebaasi loomiseks kuluvat aega ning vajaduse korral ka andmebaasi mahtu. Samuti oleks lühemaid järjestusi kasutav metoodika vähem tundlik mutatsioonidele ja sekveneerimisvigadele.

Käesoleva töö eesmärgiks on luua *k*-meeridel põhinev bioinformaatiline tööriist, mis võimaldaks analüüsida nukleiinhappejärjestustes sisalduvate *k*-meeride sagedusi. Muuhulgas peaks nimetatud tööriist võimaldama ka leida kuni 32-nukleotiidseid bakterigenoomidele spetsiifilisi järjestusi. Töö alameesmärgid on püstitatud järgmiselt:

1. Kirjutada tarkvara *k*-meeride loendamiseks genoomist/sekveneerimislugemitest, mis töötaks mõistliku kiirusega nii bakterite kui ka eukarüootide andmestikel ja võimaldaks kasutada sõnapikkust kuni 32 nukleotiidi.

2. Kirjutada tarkvara hulgateoreetiliste operatsioonide sooritamiseks *k*-meeride tabelitel.

3. Testida kirjutatud tarkvara bakterispetsiifiliste *k*-meeride tuvastamiseks teise põlvkonna sekveneerimislugemitest.

2.2 Materjalid ja metoodika

2.2.1 Kasutatud DNA järjestused ja riistvara

Töös kasutatud bakteriliikide DNA järjestused laaditi alla NCBI Nucleotide andmebaasist (30.04.2014 seisuga). Kokku kasutati 6174 bakteriliiki 1315 erinevast perekonnast. Liigid, mida kasutati, võeti NCBI Taxonomy andmebaasist (30.04.2014 seisuga). Alla laaditud järjestuste hulka kuulusid nii täisgenoomid kui ka üksikute geenide järjestused. Kokku laaditi alla ligikaudu 140 GiB (gibibait) mahus DNA järjestusi. Teise põlvkonna sekveneerimislugemitest bakteriliikide spetsiifiliste *k*-meeride leidmiseks kasutati projekti SARMB11183T raames Illumina HiSeq 2000 tehnoloogiaga sekveneeritud 96 tüve paarislugemeid. On teada, et iga andmestik on saadud kas *Escherichia coli* või *Klebsiella pneumoniae* bakteri sekveneerimisel. Sisendandmete mahu vähendamiseks võeti käesolevas töös arvesse igast paarislugemist ainult üks fragment.

Kogu arvutuslik töö tehti Tartu Ülikooli bioinformaatika õppetooli CentOS 5.10 Linuxi serveris, millel on 32 tuuma taktsagedusega 2.27 GHz ja 512 GiB operatiivmälu.

2.2.2 Programmide pakett nukleiinhappe järjestuste analüüsimiseks *k*-meeride põhjal

Loodud bioinformaatiline tööriist nukleiinhappe järjestuste analüüsimiseks koosneb kolmest eraldiseisvast alamprogrammist. *glistmaker* loendab etteantud FastA formaadis nukleiinhappe järjestustest *k*-meere ja arvutab nende sagedusi ning väljastab saadud tulemused binaarse tabelina. *glistquery* on mõeldud *glistmaker*'i loodud tabelist kasutaja etteantud *k*-meeride otsimiseks. *glistcompare* teostab hulgateoreetilisi operatsioone (ühisosa, ühend, vahe) kahe talle sisendiks antud *glistmaker*'i loodud tabeliga. Kogu pakett on kirjutatud laialt kasutatavas madala taseme süsteemprogrammeerimise keeles C (Kernighan ja Ritchie, 1988), mis võimaldab programmeerijal kontrollida kogu mäluhaldust ning on seega ideaalne suurte mälumahtudega töötamiseks. Programmid on loodud operatsioonisüsteemis Linux ning on käivitatavad käsurealt.

2.2.3 *K*-meeride tabel

Kõigi kolme programmi töös kasutatakse keskse andmestruktuurina *k*-meeride tabelit. Iga tabel koosneb päisest ja kirjetest. Päises on kirjas tabelis leiduvate *k*-meeride pikkus ja arv. Iga kirje sisaldab binaarkujule kodeeritud *k*-meeri ning tema sagedust. Tabeli kirjed on mitte-redundantsed ning järjestatud *k*-meeride alusel (vaata peatükki 1.4.3). Kirjete arvu vähendamiseks sisaldab tabel iga teineteisega pööratult komplementaarse *k*-meeri kohta ainult ühte kirjet (täpsemalt peatükis 2.2.4).

glistmaker võimaldab luua maksimaalselt 32-meere sisaldavaid tabeleid. Olenemata pikkusest on iga *k*-meer kodeeritud 64-bitiseks märgita täisarvuks (vaata peatükki 1.4.1). Lühemate *k*-meeride kõik kasutamata bitid on võrdsustatud nulliga. Oluline on märkida, et kui *k*-meeri pikkus pole fikseeritud, siis võib sama 64-bitine märgita täisarv tähistada 32 erinevat *k*-meeri (sest 00 kodeerib nukleotiidi A). Seetõttu peab igas *k*-meeride tabelis olema *k*-meeride pikkus üheselt määratud. Soovides kasutada teise pikkusega *k*-meere, tuleb luua uus tabel.

Iga *k*-meeri sagedus on tabelisse sisestatud kui 32-bitine märgita täisarv. Seega on tabeli iga kirje suurus 96 bitti. Tabeli suurus sõltub sellest, kui palju on selles nukleiinhappejärjestuses, mille põhjal tabel loodi, unikaalseid *k*-meere.

2.2.4 *glistmaker* – *k*-meeride lugemine nukleiinhappejärjestustest

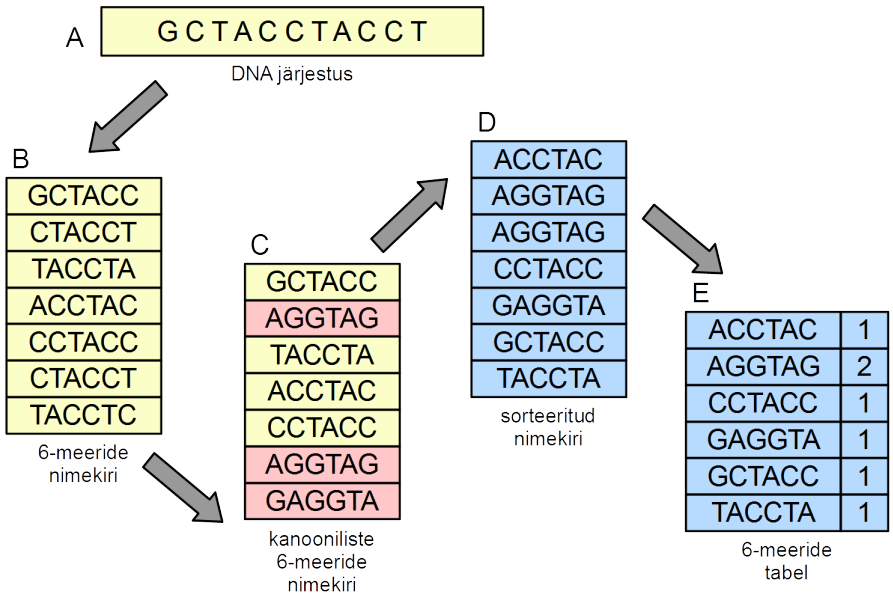
Antud töös loodud programm *glistmaker* on paketi GenomeMasker (Andreson *et al.*, 2006) *k*-meeride loendamise programmi uuem versioon. Nende algoritmiline ülesehitus on analoogne, kuid uus *glistmaker* kasutab kiiremat *k*-meeride sorteerimist ning väljastab kõikide *k*-meeride sagedused, mitte ainult üleesindatud *k*-meeride nimekirja. Samuti võimaldab uuem versioon loendada pikemaid *k*-meere.

glistmaker loendab FastA formaadis nukleiinhappejärjestusest *k*-meere libiseva aknaga. Akna laius on määratud *k*-meeri pikkusega ning võib olla kuni 32 nukleotiidi. Igal sammul on aknas üks *k*-meer, mis kodeeritakse täisarvuks nii, nagu kirjeldatud peatükis 1.4.1. Lisaks leitakse hetkel aknas oleva *k*-meeri pööratud komplementaarne järjestus, mis kodeeritakse samuti täisarvuks. Nendest kahest täisarvust säilitatakse väiksem. Sellist tähistust nimetatakse *k*-meeri kanooniliseks kujuks. Järjestust mööda liikudes salvestatakse kõikide leitud *k*-meeride kanoonilised kujud ajutiselt selleks ettenähtud nimekirja (vt. joonised 4A – 4C).

Kui loetavas nukleiinhappejärjestuses on korduseid, sisaldab saadud ajutine tabel mitmeid *k*-meere rohkem kui ühe korra. Kõige efektiivsem viis iga *k*-meeri sageduse arvutamiseks on *k*-meerid esmalt järjestada. Selleks kasutab *glistmaker* modifitseeritud varianti lineaarse keerukusega sorteerimismeetodist *radix-sort* (Duvanenko, 2009). Erinevalt harilikust *radix-sort* meetodist ei vaja modifitseeritud variant järjestamiseks lisamälu ning kasutab väikeste andmemahtude korral *insertion sort* meetodit.

Sorteerimisjärgselt paiknevad mitmes korduses esinevad *k*-meerid tabelis kõrvuti (vt. joonis 4D). *glistmaker* loeb tabeli veelkord läbi ning salvestab iga *k*-meeri esinemissageduse (vt. joonis 4E). Viimase sammuna kirjutatakse iga tabelis esinenud unikaalne *k*-meer koos sagedusega faili. Failile lisatakse ka päis, mis sisaldab informatsiooni tabelis leiduvate *k*-meeride kohta.

Eraldi käsurea parameetriga on võimalik määrata, et faili kirjutataks ainult need *k*-meerid, mille sagedus on suurem kui kasutaja määratud piirmäär.



Joonis 4: glistmaker'i tööpõhimõte* 6*-meeride näitel. *DNA järjestusest (A) leitakse järjest kõik k-meerid (B). Vajadusel teisendatakse k-meere nii, et nad oleksid kanoonilisel kujul (C). Seejärel k-meerid järjestatakse (D) ning leitakse iga k-meeri sagedus (E). Saadud tabel kirjutatakse binaarsel kujul faili.*

2.2.5 *glistquery* – *k*-meeride otsimine tabelist

glistquery võtab sisendiks *glistmaker*'i koostatud tabeli faili ning otsib sealt kasutaja määratud *k*-meere, kasutades kahendotsingut. Programmi töö tulemusena trükitakse standardväljundisse (vaikimisi ekraanile) otsitav *k*-meer koos tema tabelist leitud sagedusega. Kui tabelis *k*-meeri ei leidu, määratakse tema sageduseks 0. Kui otsitava *k*-meeri pikkus ei võrdu tabelis olevate *k*-meeride pikkustega, annab programm veateate. Kasutaja saab *k*-meere *glistquery*'le ette anda ühekaupa või terve nimekirja korraga. Kui otsitavat *k*-meeri selleks ettenähtud käsurea parameetriga ei täpsustata, trükib *glistquery* kogu tabeli.

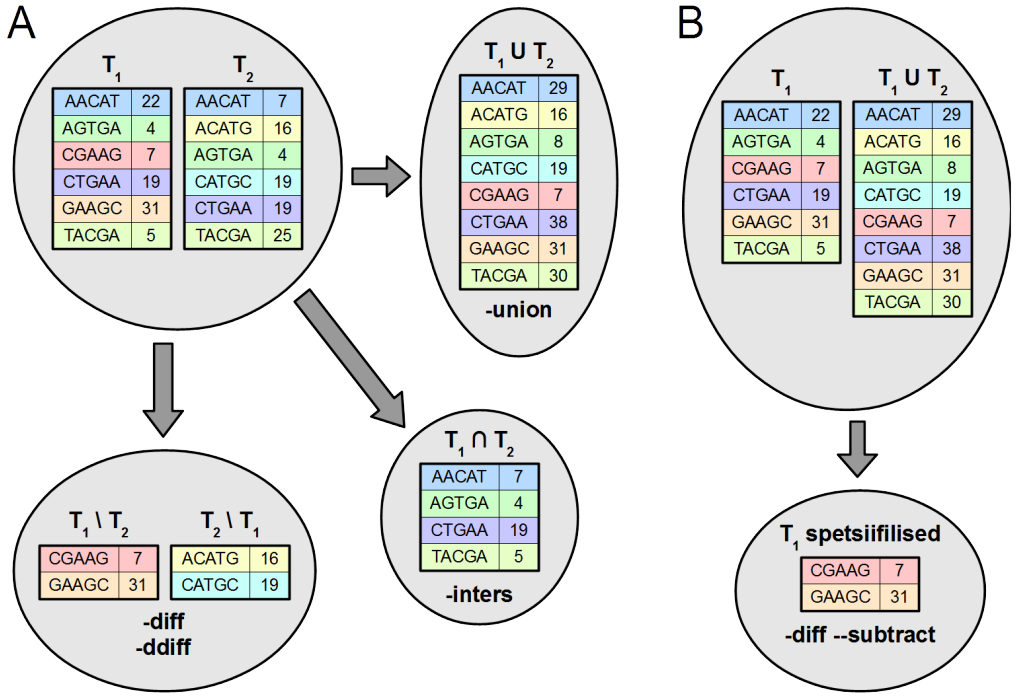
glistquery võimaldab sooritada otsinguid vigadega. Kasutaja saab määrata, mitut viga ta *k*-meeri otsingul lubab. *glistquery* genereerib otsitavast *k*-meerist kõik nii mitme veaga variandid, otsib tabelist kõigi variantide sagedused ning tagastab sageduste summa. Eraldi käsurea parameetriga on võimalik lasta *glistquery*'l trükkida summaarse sageduse asemel kõik vigadega variandid eraldi.

2.2.6 *glistcompare* – hulgateoreetilised operatsioonid *k*-meeride tabelitega

Hulgateoreetilised operatsioonid on antud kontekstis defineeritud kui binaarsed tehted hulkadega. Levinumad neist on hulkade ühend (operatsioon ”ja”, disjunktsioon), ühisosa (operatsioon ”või”, konjunktsioon) ja vahe. Kahe hulga ühend on hulk, mis sisaldab kõiki mõlema hulga elemente. Kahe hulga ühisosa on hulk, mis sisaldab ainult neid elemente, mis esinevad mõlemas hulgas. Kahe hulga vahe on hulk, mis sisaldab elemente, mis leiduvad esimeses hulgas, kuid mitte teises hulgas.

glistcompare sooritab hulgateoreetilisi tehteid kahe *glistmaker*'i loodud tabeliga. Sooritatava tehte määrab kasutaja käsurea parameetriga. Võimalikud variandid on leida kahe tabeli ühend (käsurea parameeter: *-union*), ühisosa (*-intrs*), vahe (*-diff*) ja „topelt-vahe”(*-ddiff*), mis on defineeritud kui kahe tabeli vahe mõlemat pidi (vt. joonis 5A). Kõik hulgateoreetilised tehted on lineaarse keerukusega operatsioonid.

glistcompare loeb kõigi tehete sooritamisel paralleelselt kahte tabelit. Iga sammuga liigutakse edasi kas mööda ühte, mööda teist või mööda mõlemat tabelit. Tähistagu M_i *k*-meere



Joonis 5: glistcompare'i funktsioonid tabelitega T_1 ja T_2. *A: Kahe tabeli ühend ($T_1 \cup T_2$) sisaldab kõiki kirjeid mõlemast tabelist, ühisosa ($T_1 \cap T_2$) sisaldab kirjeid, mis leiduvad mõlemas tabelis ning vahe ($T_1 \setminus T_2$ ja $T_2 \setminus T_1$) sisaldab kirjeid, mis leiduvad esimeses, kuid mitte teises tabelis. B: Tabeli T_1 spetsiifiliste kirjete leidmine juhul, kui T_1 on teise tabeli alamhulk. Kõigi tulemuste tabelite alla on märgitud vajalikud glistcompare'i käsurea parameetrid.*

ühes tabelis ja N_j k-meere teises tabelis, kus i ja j on vastavate kirjete indeksid. Olgu meil mingi fikseeritud i ja j korral $M_i < N_j$. Sellisel juhul suurendatakse i väärtust (liigutakse edasi mööda esimest tabelit), kuni võrratus kehtib. Kui $M_i > N_j$, suurendatakse j väärtust (liigutakse edasi mööda teist tabelit) ning kui $M_i = N_j$, suurendatakse mõlema indeksi väärtust (liigutakse edasi mööda mõlemat tabelit). Taoline liikumisviis tagab selle, et k-meeride sisestamisel uude tabelisse on nad endiselt järjestatud.

Kõigi *glistcompare*'i operatsioonide tulemusena tekivad uued tabelite failid. Ühendi puhul kirjutatakse uude faili ainult need k-meerid, mis leiduvad mõlemas algses tabelis. Uues tabelis määratakse k-meeri sageduseks esialgsete sageduste summa. Ühisosa puhul kirjutatakse faili kõik mõlemas tabelis leiduvad k-meerid ning nende sagedusteks määratakse väiksem kummagi lähtetabeli sagedustest.

Kahe tabeli vahe leidmisel võtab *glistcompare* arvesse valikulisi lisaparameetreid. Soovides tuvastada k-meere, mis leiduvad ühes tabelis, aga mitte teises, saab kasutaja määrata, mitmest veast alates loetakse teises tabelis olev k-meer erinevaks. Näiteks kui kasutaja määrab minimaalseks vigade arvuks 2, kirjutatakse uude tabelisse esimesest tabelist ainult need k-meerid, millel on iga teises tabelis oleva k-meeriga vähemalt kolmenukleotiidne erinevus. Selleks kasutab *glistcompare glistquery* vigade generaatori ja k-meeride otsimise funktsionaalsust. Mõnede praktiliste rakenduste hõlbustamiseks on *glistcompare*'i vahe leidmisele lisatud valikuline parameeter *--subtract*. Selle parameetriga ütleb kasutaja programmile, et esimene tabel on teise tabeli alamhulk. Sellisel juhul lahutatakse esimese sammuna teisest tabelist esimese tabeli kirjete sagedused maha ning seejärel väljastatakse ainult esimeses tabelis leiduvad k-meerid (vt. joonis 5B).

2.2.7 Programmide tööaja mõõtmine

Programmide tööaegu mõõdeti Linuxi süsteemse „time" programmiga. Käesolevas töös esitatud tööajad on arvutatud kui programmide kasutaja- ja süsteemiaegade summa. Reaalne tööaeg sõltub lisaks veel arvuti koormusest.

Programmide tööaegade mõõtmiseks kasutati k-meere pikkusega 16 ja 32. Selleks loendati *glistmaker*'iga k-meere bakteritüvede *Mycoplasma genitalium G37*, *Escherichia coli O157:H7*, *Myxococcus xanthus DK:1622* ja inimese (*Homo sapiens*, GRCh37) genoomidest. *glistquery* tööaja leidmiseks genereeriti miljon erinevat 32-meeri järjestust. *glistcompare*'i ja *glistquery* programmide kiiruseid mõõdeti *glistmaker*'iga loodud 32-meeride tabelitel.

2.2.8 Bakterispetsiifiliste *k*-meeride määramine

6174 bakteriliigi DNA järjestustest tehti *glistmaker*'i abil 32-meeride tabelid. Saadud tabelid pandi kokku suureks ühendiks. Selleks kirjutati programmeerimiskeeles Perl lühike programmijupp, mis kasutab *glistcompare*'i ühendi loomise funktsionaalsust ning liidab tabeleid rekursiivselt paarikaupa kokku, kuni alles jääb vaid üks suur tabel.

Genoomispetsiifilised 32-meerid leiti *Escherichia coli* ja *Klebsiella pneumoniae* liikidele. Selleks võrreldi kummagi liigi *glistmaker*'i abil loodud 32-meeride tabelit kõikide bakterite 32-meeridest kokku pandud ühendiga. Kasutati *glistcompare*'i kahe tabeli vahe leidmise funktsiooni lisaparameetriga *--subtract*, et programm ei arvestaks ühendis olevaid vastavalt *E. coli* või *K. pneumoniae* enda 32-meere.

2.2.9 Bakterispetsiifiliste *k*-meeride otsimine sekveneerimislugemitest

Bakterispetsiifilisi *k*-meere otsiti 96 teise põlvkonna sekveneerimisandmestikust. Iga andmestiku kohta oli teada, et ta sisaldab kas *E. coli* või *K. pneumoniae* liiki bakterit. Liigi määramiseks kasutati kummagi liigi spetsiifiliste 32-meeride tabeleid.

Esmalt tehti *glistmaker*'i abil kõigist sekveneerimisandmestikest eraldi 32-meeride tabelid. Seejärel leiti saadud 32-meeride ühisosa nii *E. coli* kui ka *K. pneumoniae* spetsiifiliste 32-meeridega, kasutades programmi *glistcompare* parameetriga *-intrs*. Töö tulemusena esitati mõlema liigi spetsiifiliste *k*-meeride arvud sekveneerimisandmestikes.

2.3 Tulemused

2.3.1 Programmide tööajad ja väljundandmestike mahud

Kirjutati bioinformaatiline tööriist, mis loendab *k*-meere nukleotiidsetest järjestustest ning töötleb saadud *k*-meeride tabeleid. Tööriist koosneb kolmest programmist: *glistmaker*, *glistquery* ja *glistcompare*. Programmide töö efektiivsuse hindamiseks mõõdeti nende tööaegu erinevate suurustega sisendandmestikel. *glistmaker*'i testimiseks kasutati nelja genoomset järjestust. Nende hulgas oli kolm eri suurusega bakteri genoomi ja inimese genoom. Sõnapikkuseks valiti 32. Tabelis 1 on esitatud *glistmaker*'i tööaeg ja väljundi suurus kõigi nelja genoomi kohta. Ka suurte bakterigenoomide puhul jäi programmi tööaeg alla kahe sekundi. Kõigist inimese kromosoomidest kokku loendas *glistmaker* 32-meere veidi alla 20 minuti. Inimese puhul mõõdeti ka 16-meeride loendamise tööaega ning saadud tulemus jäi alla 12 minuti. Kõigist NCBI andmebaasist võetud 6174 bakteriliigi DNA järjestustest 32-meeride loendamiseks kulus kokku ligikaudu kuus tundi.

Kuna *glistcompare*'i ühendi, ühisosa ja ilma vigadeta vahe leidmise operatsioonid on programmi keerukuse ja tööaja mõttes analoogsed, mõõdeti aega ainult ühendi moodustamisel. Selleks kasutati *E. coli* ja *M. xanthus*'e genoomidest tehtud 32-meeride tabeleid. Kahe tabeli ühendi moodustamine võttis aega 1.5 sekundit. Samuti pandi suureks ühendiks kokku kõigi NCBI andmebaasist võetud bakterite 32-meerid. *glistcompare*'i tööajaks tuli ligikaudu 9 tundi. Saadud ühendi suurus oli 224 GiB ja see sisaldas kokku ligikaudu 20 miljardit 32-meeri.

glistquery töökiirust mõõdeti *E. coli* 32-meeride tabelil ja kõikide bakterite 32-meeride ühendil. Sõnade otsimise kiirusteks saadi vastavalt 27 ja 1.67 miljonit sõna minutis.

Tabel 1: *glistmaker'i tööajad ja tulemuste mahud kolme bakteri (Mycoplasma genitalium G37, Escherichia coli O157:H7, Myxococcus xanthus) ja inimese (Homo sapiens) genoomi puhul. Bakterite genoomidest loendati* 32*-meere, inimese genoomist* 32*- ja* 16*-meere. Mbp – miljonit aluspaari, MiB – mebibait (*1 *MiB* $= 1024^2$ *baiti).*

organism	genoomi suurus	tabeli koostamise aeg	tabeli maht	*k*-meere tabelis
M. genitalium	0.59 Mbp	0.023 sek	6.6 MiB	$\sim$ 0.57 mln
E. coli	5.65 Mbp	1.009 sek	60 MiB	$\sim$ 5.2 mln
M. xanthus	9.27 Mbp	1.801 sek	104 MiB	$\sim$ 9.1 mln
H. sapiens ($k = 32$)	3146.81 Mbp	19 min, 13.323 sek	28650 MiB	$\sim$ 2.5 mld
H. sapiens ($k = 16$)	3146.81 Mbp	11 min, 36.736 sek	9132 MiB	$\sim$ 0.8 mld

2.3.2 Bakterispetsiifiliste k-meeride otsimine

Kirjutatud bioinformaatilist tööriista testiti bakteriliikide määramise kontekstis. Selleks leiti spetsiifilised 32-meerid kahele bakteriliigile: *Escherichia coli* ja *Klebsiella pneumoniae*. Spetsiifilisi sõnu leiti vastavalt *E. coli*'le 712221 ja *K. pneumoniae*'le 13450672. Mõlemal andmestikul töötas *glistcompare* ligikaudu 7 minutit.

Saadud spetsiifilisi 32-meere otsiti 96 sekveneerimisandmestikust. Kõigist andmestikest 39-s leidus *E. coli* spetsiifilisi sõnu rohkem kui *K. pneumoniae* spetsiifilisi sõnu. Ülejäänud 57 andmestikus domineerisid *K. pneumoniae* spetsiifilised sõnad. Kaheteistkümnel juhul 57-st oli aga kõigist *E. coli* spetsiifilistest sõnadest esindatud suurem osakaal kui kõigist *K. pneumoniae* spetsiifilistest sõnadest. Ühe bakteriliigi spetsiifiliste sõnade arv ületas teise bakteriliigi spetsiifiliste sõnade arvu rohkem kui 10 korda 61% andmestikes, rohkem kui 5 korda 77% andmestikest ja rohkem kui 3 korda 83% andmestikest.

Kahekümne viies andmestikus, kus *E. coli* spetsiifilisi sõnu oli rohkem, leiti enam kui 40% kõigist tuvastatud *E. coli* spetsiifilistest sõnadest. Mitte ühegi andmestiku puhul polnud antud osakaal suurem kui 50%. Andmestikest, kus *K. pneumoniae* spetsiifilised sõnad olid üleesindatud, oli vastavatest osakaaludest suurim kõigest 4.6%. Tulemused kõigi sekveneerimisandmestike kohta on toodud töö lisades.

2.4 Arutelu

Käesoleva töö põhiliseks eesmärgiks oli kirjutada programmide pakett, mis looks *k*-meeride sagedustabeleid etteantud nukleotiidsete järjestuste alusel ning teostaks saadud tabelitega lihtsamaid hulgateoreetilisi operatsioone. Selleks kirjutati kolm eraldiseisvat programmi: *glistmaker*, *glistquery* ja *glistcompare*. Programmide kirjutamiseks valiti programmeerimiskeel C, mis on disainitud eelkõige arvuti süsteemsete programmide (näiteks operatsioonisüsteemid) kirjutamiseks ja sobib seega väga suurte andmemahtude töötlemiseks. Valik oli õigustatud, sest juba antud töö kontekstis tuli programmidel töötada andmestikega, mille suurused ületasid 200 GiB.

Kirjutatud programme testiti bakterite ja inimese genoomi peal. Kasutades *glistmaker*'it loodi genoomis leiduvate 32-meeride tabelid. Mõlemal juhul töötasid programmid mõistliku ajaga. On oluline märkida, et lühemate *k*-meeride korral töötab programm kiiremini kui pikemate *k*-meeride korral. *glistquery* tööaeg sõltub selle tabeli suurusest, millest *k*-meere otsitakse. Ühe bakteri ja kõigi bakterite *k*-meeride tabelist sõna otsimise kiirus erineb ligi 16 korda. Saadud tulemus on loogiline, sest tabelite mahud erinevad ligikaudu 4000 korda ning otsingualgoritm on logaritmilise keerukusega. Hea näitaja on ka see, et nii *glistmaker* kui ka *glistcompare* suudab ligikaudu 140 GiB nukleotiidseid järjestusi töödelda kiiremini kui poole ööpäevaga.

Kogu meetodi miinuseks on väljundandmestike suur maht, kuid seda pole võimalik vähendada ilma, et mingi osa andmetest kaduma ei läheks. Samuti nõuab hetkel töötav *glistmaker*'i implementatsioon operatiivmälu mahus, mis on võrdeline ühes sisendandmestikus leiduvate *k*-meeride koguarvuga. Operatiivmälu kasutust saab vähendada, kui implementeerida *k*-meeride lugemine fikseeritud mahuga ajutistesse tabelitesse ning saadud tabeleid jooksvalt ühendada. Suure töökiiruse säilitamiseks tuleks sel juhul *glistmaker* muuta mitmelõimeliseks. Suurtest ja paljude kordustega genoomidest *k*-meeride loendamise efektiivsemaks muutmiseks võib hetkel kasutusel olevad *k*-meeride tabelid asendada *k*-meere sisaldava puustruktuuride (*radix trie*). Taoline struktuur võimaldaks ka *glistquery*'l teha *k*-meeride otsinguid konstantse keerukusega.

Teiseks töös püstitatud ülesandeks oli testida, kas loodud tööriista võiks potentsiaalselt kasutada bakterite määramiseks teise põlvkonna sekveneerimisandmestikest. Selleks leiti esmalt *Escherichia coli* ja *Klebsiella pneumoniae* spetsiifilised 32-meerid. *K*-meeri pikkuse valik pole antud töö kontekstis oluline seni, kuni valitud pikkuse juures leidub piisavalt spetsiifilisi *k*-meere. *E. coli*'le leiti 712221 ja *K. pneumoniae*'le 13450672 liigispetsiifilist 32-meeri. See tähendab, et leitud unikaalsete *k*-meeride arvud erinevad nende kahe liigi puhul peaaegu 20

korda, kuigi genoomide suurused on samas suurusjärgus. Selle üheks võimalikuks põhjuseks on asjaolu, et NCBI andmebaasis võivad *E. coli* ja *K. pneumoniae* liikide alla koondatud bakterite üksused olla oluliselt erineva liigisisese heterogeensuse või fülogeneetilise kaugusega.

Töös kasutati 96 sekveneerimisandmestikku, millest kõik sisaldasid kas *E. coli* või *K. pneumoniae* lugemeid. Ülesandeks oli katsetada, kas andmestikest leitud kummagi bakteri spetsiifiliste 32-meeride arvude põhjal on võimalik ennustada, kumma bakteriliigiga on tegemist. Tulemustest selgub, et kuigi enamike andmestike korral on 32-meeride arv selgelt kas ühe või teise liigi kasuks, on nimetatud primitiivsel ennustamismeetodil siiski väga tugevaid puuduseid.

Oluline on määrata, kas bakteriliigi detekteerimisel tuleks kasutada leitud spetsiifiliste k-meeride absoluutarvu või osakaalu kõigist antud liigile spetsiifilistest k-meeridest. Kuigi osakaalud on ülevaatlikumad, sest ei sõltu genoomi suurusest ega spetsiifiliste k-meeride arvust, siis, võrreldes kahte väga erineva liigisisese heterogeensusega bakteriliiki, võib absoluutarvu kaasamine analüüsi anda õigema tulemuse. Töös saadud tulemustest on näha, et olenevalt sellest, kas kasutatakse absoluutarvu või osakaalu, klassifitseeritaks erinevalt 12 andmestikku.

Teiseks oluliseks küsimuseks on otsustada, millisest tuvastatud spetsiifiliste k-meeride arvust või osakaalust alates üldse saab bakterit määrata. Selleks on vaja luua sobiv statistiline test, mis võtab arvesse taksoni varieeruvust. Madalamatel taksonitel nagu liik või tüvi peaks leitama suurem osakaal kõigist spetsiifilistest k-meeridest. Perekonna tasandil bakteri määramiseks piisab väiksemast osakaalust. Antud töö tulemustest on näha, et *E. coli* leitud spetsiifiliste 32-meeride osakaal jääb enamasti 50% lähedusse, mis võib olla piisavalt suur osakaal liigi määramiseks, arvestades *E. coli* tüvede varieeruvust ja rohkust. *K. pneumoniae* puhul on aga ka suurim leitud osakaaludest ligi 10 korda väiksem. See on ilmselt tingitud töös kasutatud *K. pneumoniae* nukleotiidsete järjestuste suurest heterogeensusest. Lisaks võib mõlemal juhul osakaalu langetada sekveneerimisvigade esinemine liigispetsiifiliste k-meeride piirkondades. Potentsiaalselt on võimalik leitud k-meeride osakaale suurendada, leides optimaalse k-meeri pikkuse. Kõrgemate taksonoomiliste üksuste puhul võib kaaluda ka kõigi tüvede spetsiifiliste k-meeride ühisosa leidmist.

Käesoleva töö põhjal võib öelda, et k-meeridel põhinev metoodika ja loodud bioinformaatiline tööriist on perspektiivsed võimalused tulevikus bakterite määramiseks. Selleks tuleb aga esmalt leida meetodi jaoks optimaalsed parameetrid ning luua sobivad statistilised analüüsimeetodid, mis arvestaksid bakterite taksonisisest varieeruvust ning võimaldaks baktereid määrata ka metagenoomika lugemitest. Meetodeid, mis lubavad baktereid detekteerida teise põlvkonna sekveneerimislugemitest, arendatakse eelkõige lootuses, et tulevikus on neid võimalik liita sekveneerimiskonveieritesse.

Kokkuvõte

Bakterite määramine on oluline mitmetes kliinilistes ja tööstuslikes protsessides nagu haiguste diagnoosimine või toiduainete kvaliteedi testimine. Tihti on vajalik proovidest baktereid identifitseerida võimalikult kiiresti. Üheks sobivaks meetodiks on teise põlvkonna sekveneerimislugemitest bakterispetsiifiliste oligomeeride otsimine ja nende põhjal bakterite määramine.

Käesoleva bakalaureusetöö esimeseks eesmärgiks oli luua bioinformaatiline tööriist, mis loendaks nukleotiidsest järjestusest oligomeere, moodustaks leitud oligomeeridest sagedustabeli ja teostaks saadud tabelitega hulgateoreetilisi tehteid. Töö tulemusena valmis programmide pakett, mis vastab ülaltoodud nõuetele ning töötab mõistliku ajaga nii bakterite kui ka eukarüootide genoomidel.

Töö teiseks eesmärgiks oli testida loodud programmi sobivust bakterispetsiifiliste oligomeeride leidmiseks teise põlvkonna sekveneerimislugemitest. Selleks otsiti sekveneerimisandmestikest *Escherichia coli* ja *Klebsiella pneumoniae* spetsiifilisi oligomeere. Kõigist andmestikest 83%-l ületas ühe bakteriliigi leitud spetsiifiliste oligomeeride arv teise omi rohkem kui 3 korda. Seega on põhjust uskuda, et tööriist sobib bakterite määramiseks, kuid vajab veel parameetrite optimiseerimist ja sobivat statistilist testi, mille abil otsustada iga bakteriliigi esindatuse üle uuritavas proovis.

Oligomer-based bioinformatic methods for identifying bacteria from sequencing reads

Maarja Lepamets

Summary

Different methods for identification of bacterial genera, species or strains are carried out daily in diagnostics and food processing. In many cases, it is necessary to gain an accurate insight to the inhabitants of a bacterial community as fast as possible. Thus, it is essential to find new methods for bacteria identification that do not need prerequisites as growing pure cultures or assembling the next-generation sequencing (NGS) reads. One of the options would be to identify short bacteria-specific oligomers (BSO) straight from the NGS reads. However, there are no good enough bioinformatics tools developed for such analysis.

The aim of this thesis was to build a software package that would be able to count the oligomers from a given nucleotide sequence and perform set-theoretic operations (union, intersection, complement) on tables of those oligomers. Three programs named *glistmaker*, *glistquery* and *glistcompare* were implemented that satisfy the conditions set and run reasonably fast on bacterial genomes as well as on human genome.

The next step was to test whether the software could be of use for identifying BSO from the NGS datasets. For this, BSO were found for *Escherichia coli* and *Klebsiella pneumoniae* species. Those BSO were searched for from 96 different NGS datasets. In 88% of the datasets the BSO of one bacteria over-represented the BSO of the other bacteria more than 3-fold. In 61% of the datasets the over-representation was larger than 10-fold. Therefore, it can be believed that the tool can eventually be used for identification of bacteria.

Still, before the software can be tested for actual bacteria identification one has to develop a statistical method which says whether the number of BSO per dataset is significant for a given taxon. This method must also consider the heterogeneity of a taxon. Furthermore, the oligomer length must be tuned to minimize the software's dependence on mutations and sequencing errors.

Kirjanduse loetelu

Acosta-González, A., Rosselló-Móra, R., Marqués, S. (2013). Characterization of the anaerobic microbial community in oil-polluted subtidal sediments: aromatic biodegradation potential after the Prestige oil spill. Environ Microbiol. 15: 77 – 92.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389 – 3402.

Andreson, R., Reppo, E., Kaplinski, L., Remm, M. (2006). GENOMEMASKER package for designing unique genomic PCR primers. BMC Bioinformatics. 7: 172.

Antón, J., Lucio, M., Peña, A., Cifuentes, A., Brito-Echeverría, J., Moritz, F., Tziotis, D., López, C., Urdiain, M., Schmitt-Kopplin, P., Rosselló-Móra, R. (2013). High metabolomic microdiversity within co-occurring isolates of the extremely halophilic bacterium *Salinibacter ruber*. PLoS One. 8: e64701.

Auch, A. F., von Jan, M., Klenk, H. P., Göker, M. (2010). Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. Stand Genomic Sci. 2: 117 – 134.

Balboa, S., Doce, A., Diéguez, A. L., Romalde, J. L. (2011). Evaluation of different species-specific PCR protocols for the detection of *Vibrio tapetis*. J Invertebr Pathol. 108: 85 – 91.

Batra, S. A., Krupanidhi, S., Tuteja, U. (2013). A sensitive and specific multiplex PCR assay for simultaneous detection of *Bacillus anthracis*, *Yersinia pestis*, *Burkholderia pseudomallei* and *Brucella* species. Indian J Med Res. 138: 111 – 116.

Benítez-Páez A., Belda-Ferre, P., Simón-Soro, A., Mira, A. (2014). Microbiota diversity and gene expression dynamics in human oral biofilms. BMC Genomics. [Epub ahead of print].

Bochner, B. R. (2003). New technologies to assess genotype-phenotype relationships. Nat Rev Genet. 4: 309 – 14.

Bochner, B. R. (2009). Global phenotypic characterization of bacteria. FEMS Microbiol Rev. 33: 191 – 205.

Choi, I. G., Kim, S. H. (2007). Global extent of horizontal gene transfer. Proc Natl Acad Sci U S A. 104: 4489 – 4494.

Christner, B. C., Kvitko, B. H., Reeve, J. N. (2003). Molecular identification of bacteria and Eukarya inhabiting an Antarctic cryoconite hole. Extremophiles. 7: 177 – 183.

Chuang, L. Y., Cheng, Y. H., Yang, C. H. (2013). Specific primer design for the polymerase chain reaction. Biotechnol Lett. 35: 1541 – 1549.

Chun, J., Lee, J. H., Jung, Y., Kim, M., Kim, S., Kim, B. K., Lim, Y. W. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. Int J Syst Evol Microbiol. 57: 2259 – 2261.

Ciccarelli, F. D., Doerks, T., von Mering, C., Creevey, C. J., Snel, B., Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science. 311: 1283 – 1287.

Daniel, R. (2005). The metagenomics of soil. Nat Rev Microbiol. 3: 470 – 478.

Delihas, N. (2008). Small mobile sequences in bacteria display diverse structure/function motifs. Mol Microbiol. 67: 475 – 481.

DeLong, E. F. (2005). Microbial community genomics in the ocean. Nat Rev Microbiol. 3: 459 – 469.

Deorowicz, S., Debudaj-Grabysz, A., Grabowski, S. (2013). Disk-based *k*-mer counting on a PC. BMC Bioinformatics 14: 160.

Didelot, X., Maiden, M. C. (2010). Impact of recombination on bacterial evolution. Trends Microbiol. 18: 315 – 322.

Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E., Crook, D. W. (2012). Transforming clinical microbiology with bacterial genome sequencing. Nat Rev Genet. 13: 601 – 612.

Dingle, T. C., Butler-Wu, S. M. (2013). MALDI-TOF mass spectrometry for microorganism identification. Clin Lab Med. 33: 589 – 609.

Duval, B. D., Elrod, M. G., Gee, J. E., Chantratita, N., Tandhavanant, S., Limmathurotsakul, D., Hoffmaster, A. R. (2014). Evaluation of a Latex Agglutination Assay for the Identification of *Burkholderia pseudomallei* and *Burkholderia mallei*. Am J Trop Med Hyg. [Epub ahead of print].

Duvanenko, V. J. (2009). In-place Hybrid N-bit-Radix Sort. Algorithm Improvement through Performance Measurement: Part 3. Dr. Dobb's Journal. nov. 9, 2009.

Edberg, S. C., Rice, E. W., Karlin, R. J., Allen, M. J. (2000). *Escherichia coli*: the best biological drinking water indicator for public health protection. Symp Ser Soc Appl Microbiol. 106S – 116S.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32: 1792 – 1797.

Emerson, D., Agulto, L., Liu, H., Liu, L. (2008). Identifying and characterizing bacteria in an era of genomics and proteomics. BioScience 58: 925 – 936.

Gerlach, W., Jünemann, S., Tille, F., Goesmann, A., Stoye, J. (2009). WebCARMA: a web application for the functional and taxonomic classification of unassembled metagenomic reads. BMC Bioinformatics. 10: 430.

Gevers, D., Cohan, F. M., Lawrence, J. G., Spratt, B. G., Coenye, T., Feil, E. J., Stackebrandt, E., Van de Peer, Y., Vandamme, P., Thompson, F. L., Swings, J. (2005). Re-evaluating prokaryotic species. Nat Rev Microbiol. 3: 733 – 739.

Goji, N., Macmillan, T., Amoako, K. K. (2012). A new generation microarray for the simultaneous detection and identification of *Yersinia pestis* and *Bacillus anthracis* in food. J Pathog. 2012: 627036.

Gonzalez, J. M., Saiz-Jimenez, C. (2002). A fluorimetric method for the estimation of G+C mol% content in microorganisms by thermal denaturation temperature. Environ Microbiol. 4: 770 – 773.

González, S. F., Krug, M. J., Nielsen, M. E., Santos, Y., Call, D. R. (2004). Simultaneous detection of marine fish pathogens by using multiplex PCR and a DNA microarray. J Clin Microbiol. 42: 1414 – 1419.

Hasman, H., Saputra, D., Sicheritz-Ponten, T., Lund, O., Svendsen, C. A., Frimodt-Møller, N., Aarestrup, F. M. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. J Clin Microbiol. 52: 139 – 46.

He, Z., Xiong, J., Kent, A. D., Deng, Y., Xue, K., Wang, G., Wu, L., van Nostrand, J. D., Zhou, J. (2014). Distinct responses of soil microbial communities to elevated CO2 and O3 in a soybean agro-ecosystem. ISME J. 8: 714 – 726.

Hoorfar, J. 2011. Rapid detection, characterization, and enumeration of foodborne pathogens. Doctoral thesis. APMIS Suppl. 119: 1 – 24.

Karas, M., Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal Chem. 60: 2299 – 2301.

Karlin, S., Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. Trends Genet. 11: 283 – 290.

Kelley, D. R., Schatz, M. C., Salzberg, S. L. (2010). Quake: quality-aware detection and correction of sequencing errors. Genome Biol. 11: R116.

Kernighan, B. W., Ritchie, D. M. 1988. The C programming language. 2nd ed., Prentice Hall PTR, Englewood Cliffs, New Jersey.

Konstantinidis, K. T., Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 102: 2567 – 2572.

Kunitsky, C., Osterhout, G., Sasser, M. 2006. Identification of microorganisms using fatty acid methyl ester (FAME) analysis and the MIDI Sherlock Microbial Identification System, p. 1 – 18. In Encyclopedia of Rapid Microbiological Methods. 3 ed. MIDI, Inc. Newark, DE, USA.

Kurtz, S., Narechania, A., Stein, J. C., Ware, D. (2008). A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes. BMC Genomics. 9: 517.

Kämpfer, P., Glaeser, S. P. (2012). Prokaryotic taxonomy in the sequencing era–the polyphasic approach revisited. Environ Microbiol. 14: 291 – 317.

Ley, R. E. (2010). Obesity and the human microbiome. Curr Opin Gastroenterol. 26: 5 – 11.

Li, X., He, Z., Zhou, J. (2005). Selection of optimal oligonucleotide probes for microarrays using multiple criteria, global alignment and parameter estimation. Nucleic Acids Res. 33: 6114 – 6123.

Liu, H., Bebu, I., Li, X. (2010). Microarray probes and probe sets. Front Biosci (Elite Ed). 2: 325 – 338.

Logue, J. B., Bürgmann, H., Robinson, C. T. (2008). Progress in the ecological genetics and biodiversity of freshwater bacteria. BioScience 58: 103 – 113.

Loveland-Curtze, J., Miteva, V. I., Brenchley, J. E. (2011). Evaluation of a new fluorimetric DNA-DNA hybridization method. Can J Microbiol. 57: 250 – 255.

Maiden, M. C., van Rensburg, M. J., Bray, J. E., Earle, S. G., Ford, S. A., Jolley, K. A., McCarthy, N. D. (2013). MLST revisited: the gene-by-gene approach to bacterial genomics. Nat Rev Microbiol. 11: 728 – 736.

Maity, J. P., Kar, S., Lin, C. M., Chen, C. Y., Chang, Y. F., Jean, J. S., Kulp, T. R. (2013). Identification and discrimination of bacteria using Fourier transform infrared spectroscopy. Spectrochim Acta A Mol Biomol Spectrosc. 116: 478 – 84.

Marçais, G., Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics 27: 764 – 770.

Marshall, A. G., Hendrickson, C. L., Jackson, G. S. (1998). Fourier transform ion cyclotron resonance mass spectrometry: a primer. Mass Spectrom Rev. 17: 1 – 35.

McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P., Rigoutsos, I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 4: 63 – 72.

Meisel, S., Stöckel, S., Rösch, P., Popp, J. (2014). Identification of meat-associated pathogens via Raman microspectroscopy. Food Microbiol. 38: 36 – 43.

Melsted, P., Pritchard, J. K. (2011). Efficient counting of k-mers in DNA sequences using a bloom filter. BMC Bioinformatics. 12: 333.

Mende, D. R., Sunagawa, S., Zeller, G., Bork, P. (2013). Accurate and universal delineation of prokaryotic species. Nat Methods. 10: 881 – 884.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. Bioinformatics 24: 2818 – 2824.

Moore, E. R., Mihaylova, S. A., Vandamme, P., Krichevsky, M. I., Dijkshoorn, L. (2010). Microbial systematics and taxonomy: relevance for a microbial commons. Res Microbiol. 161: 430 – 438.

Moreno-Indias, I., Cardona, F., Tinahones, F. J., Queipo-Ortuño, M. I. (2014). Impact of the gut microbiota on the development of obesity and type 2 diabetes mellitus. Front Microbiol. 5: 190.

Mulle, J. G., Patel, V. C., Warren, S. T., Hegde, M. R., Cutler, D. J., Zwick, M. E.(2010). Empirical evaluation of oligonucleotide probe selection for DNA microarrays. PLoS One. 5: e9921.

Nhung, P. H., Ohkusu, K., Miyasaka, J., Sun, X. S., Ezaki, T. (2007). Rapid and specific identification of 5 human pathogenic *Vibrio* species by multiplex polymerase chain reaction targeted to *dnaJ* gene. Diagn Microbiol Infect Dis. 59: 271 – 275.

Osorio, C. R., Collins, M. D., Toranzo, A. E., Barja, J. *L., Romalde, J. L. (1999). 16S rRNA gene sequence analysis of* Photobacterium damselae and nested PCR method for rapid detection of the causative agent of fish pasteurellosis. Appl Environ Microbiol. 65: 2942 – 2946.

Pini, F., Frascella, A., Santopolo, L., Bazzicalupo, M., Biondi, E. G., Scotti, C., Mengoni, A. (2012). Exploring the plant-associated bacterial communities in *Medicago sativa L.* BMC Microbiol. 12: 78.

Reva, O. N., Tümmler, B. (2004). Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. BMC Bioinformatics. 5: 90.

van Rhijn, P., Vanderleyden, J. (1995). The Rhizobium-plant symbiosis. Microbiol Rev. 59: 124 – 142.

Richter, M., Rosselló-Móra, R. (2009). Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106: 19126 – 19131.

Ritchie, N. J., Schutter, M. E., Dick, R. P., Myrold, D. D. (2000). Use of length heterogeneity PCR and fatty acid methyl ester profiles to characterize microbial communities in soil. Appl Environ Microbiol. 66: 1668 – 1675.

Rizk, G., Lavenier, D., Chikhi, R. (2013). DSK: k-mer counting with very low memory usage. Bioinformatics 29: 652 – 653.

Rosen, G. L., Reichenberger, E. R., Rosenfeld, A. M. (2011). NBC: the Naive Bayes Classification tool webserver for taxonomic classification of metagenomic reads. Bioinformatics. 27: 127 – 129.

Rosselló-Móra, R., Amann, R. (2001). The species concept for prokaryotes. FEMS Microbiol Rev. 25: 39 – 67.

Rosselló-Móra, R. (2012). Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. Environ Microbiol. 14: 318 – 334.

Rouillard, J. M., Zuker, M. and Gulari, E. (2003). OligoArray 2.0: Design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res. 31: 3057 – 3062.

Schwabe, R. F., Jobin, C. (2013). The microbiome and cancer. Nat Rev Cancer. 13: 800 – 812.

Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O., Huttenhower, C. (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. Nat Methods. 9: 811 – 814.

Sorek, R., Zhu, Y., Creevey, C. J., Francino, M. P., Bork, P., Rubin, E. M. (2007). Genome-wide experimental determination of barriers to horizontal gene transfer. Science. 318: 1449 – 1452.

Zhai, L., Yu, Q., Bie, X., Lu, Z., Lv, F., Zhang, C., Kong, X., Zhao, H. (2014). Development of a PCR test system for specific detection of *Salmonella Paratyphi B* in foods. FEMS Microbiol Lett. [Epub ahead of print].

Zhang, Z., Schwartz, S., Wagner, L., Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J Comput Biol. 7: 203 – 214.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R., Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. Environ Microbiol. 6: 938 – 947.

Thompson, C. C., Chimetto, L., Edwards, R. A., Swings, J., Stackebrandt, E., Thompson, F. L. (2013). Microbial genomic taxonomy. BMC Genomics. 14: 913.

Tindall, B. J., Rosselló-Móra, R., Busse, H. J., Ludwig, W., Kämpfer, P. (2010). Notes on the characterization of prokaryote strains for taxonomic purposes. Int J Syst Evol Microbiol. 60: 249 – 266.

Trkov, M., Avgustin, G. (2003). An improved 16S rRNA based PCR method for the specific detection of *Salmonella enterica*. Int J Food Microbiol. 80: 67 – 75.

Tu, Q., He, Z., Deng, Y., Zhou, J. (2013). Strain/species-specific probe design for microbial identification microarrays. Appl Environ Microbiol. 79: 5085 – 5088.

Tu, Q., He, Z., Zhou, J. (2014). Strain/species identification in metagenomes using genome-specific markers. Nucleic Acids Res. 42: e67.

Tu, Q., He, Z., Li, Y., Chen, Y., Deng, Y., Lin, L., Hemme, C. L., Yuan, T., Van Nostrand, J. D., Wu, L., Zhou, X., Shi, W., Li, L., Xu, J., Zhou, J. (2014). Development of HuMiChip for functional profiling of human microbiomes. PLoS One. 9: e90546.

Untergrasser, A., Cutcutache, I., Kõressaar, T., Ye, J., Faircloth, B. C., Remm, M., Rozen, S. G. (2012). Primer3 - new capabilities and interfaces. Nucleic Acids Research 40: e115.

Ward, D. M., Ferris, M. J., Nold, S. C., Bateson, M. M. (1998). A natural view of microbial biodiversity within hot spring cyanobacterial mat communities. Microbiol Mol Biol Rev. 62: 1353 – 1370.

Wayne, L., Brenner, D. J., Colwell, R. R., Grimont, P. A. D., Kandler, O. *et al.* (1987). Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. Int J Syst Bacteriol. 37: 463 – 464.

Whitman, W. B., Coleman, D. C., Wiebe, W. J. (1998). Prokaryotes: the unseen majority. Proc Natl Acad Sci U S A. 95: 6578 – 6583.

Wood, D. E., Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biol. [Epub ahead of print].

Wu, L., Liu, X., Fields, M. W., Thompson, D. K., Bagwell, C. E., Tiedje, J. M., Hazen, T. C., Zhou, J. (2008). Microarray-based whole-genome hybridization as a tool for determining procaryotic species relatedness. ISME J. 2: 642 – 55.

Xu, D., Li, G., Wu, L., Zhou, J., Xu, Y. (2002). PRIMEGENS: robust and efficient design of gene-specific probes for microarray analysis. Bioinformatics. 18: 1432 – 1437.

Kasutatud veebiaadressid

NCBI Nucleotide andmebaas. (30. aprill, 2014). http://www.ncbi.nlm.nih.gov/nuccore/

NCBI Taxonomy andmebaas. (30. aprill, 2014). http://www.ncbi.nlm.nih.gov/taxonomy

Perl 5 version 18.2 documentation. http://perldoc.perl.org/

Lisad

Lisa A: *Teise põlvkonna sekveneerimisandmestikest leitud Escherichia coli ja Klebsiella pneumoniae spetsiifiliste k-meeride arvud ja osakaalud kõikide vastava liigi spetsiifiliste k-meeride arvust. Kokku on analüüsitud* 96 *andmestikku.*

Andmestik	*E. coli* spetsiifilised *k*-meerid		*K. pneumoniae* spetsiifilised *k*-meerid	
	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal
EEITKB301a	343405	0.4822	34075	0.0025
EEITKB301	343488	0.4823	37765	0.0028
EEITKB303	144107	0.2023	60521	0.0045
EEITKB304	42577	0.0598	37215	0.0028
EEITKB307	342817	0.4813	38672	0.0029
EEITKB309	36685	0.0515	67806	0.005
EEITKB319	337299	0.4736	46010	0.0034
EEITKB325	19636	0.0276	35476	0.0026
EEITKB327a	67647	0.095	38071	0.0028
EEITKB327	66505	0.0934	46031	0.0034
EEITKB334	341963	0.4801	30849	0.0023
EEITKB336	341179	0.479	36070	0.0027
EEITKB338	339445	0.4766	31772	0.0024
EEITKB340	341728	0.4798	32700	0.0024
EEITKB342	340585	0.4782	35873	0.0027
EEITKB344a	343419	0.4822	29144	0.0022
EEITKB344	334197	0.4692	32416	0.0024
EEIVKB10	6900	0.0097	529839	0.0394
EEIVKB11	11818	0.0166	455453	0.0339
EEIVKB12	8146	0.0114	466865	0.0347
EEIVKB13	18209	0.0256	535439	0.0398
EEIVKB14	11056	0.0155	539538	0.0401
EEIVKB15	11596	0.0163	463169	0.0344
EEIVKB16	5292	0.0074	357351	0.0266

Andmestik	*E. coli* spetsiifilised *k*-meerid		*K. pneumoniae* spetsiifilised *k*-meerid	
	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal
EEIVKB17	10685	0.015	381089	0.0283
EEIVKB18	10421	0.0146	467774	0.0348
EEIVKB19	19363	0.0272	568884	0.0423
EEIVKB1a	16569	0.0233	542980	0.0404
EEIVKB1	14741	0.0207	524444	0.039
EEIVKB20a	344524	0.4837	32474	0.0024
EEIVKB20	343969	0.483	28937	0.0022
EEIVKB23	11669	0.0164	480948	0.0358
EEIVKB24a	10212	0.0143	472446	0.0351
EEIVKB24	23471	0.033	484923	0.0361
EEIVKB25	14680	0.0206	388241	0.0289
EEIVKB26	17100	0.024	404631	0.0301
EEIVKB27	10030	0.0141	543295	0.0404
EEIVKB2	22151	0.0311	451916	0.0336
EEIVKB30	18657	0.0262	505589	0.0376
EEIVKB31	7626	0.0107	364770	0.0271
EEIVKB32	11696	0.0164	532983	0.0396
EEIVKB3	18237	0.0256	525013	0.039
EEIVKB49	339513	0.4767	44369	0.0033
EEIVKB4	343166	0.4818	61233	0.0046
EEIVKB5	19106	0.0268	458267	0.0341
EEIVKB65	81299	0.1141	31891	0.0024
EEIVKB66	334537	0.4697	36992	0.0028
EEIVKB6	29119	0.0409	459147	0.0341
EEIVKB7	13210	0.0185	461213	0.0343
EEIVKB9	22011	0.0309	360894	0.0268
EELTKB210	344711	0.484	35040	0.0026
EELTKB212	343784	0.4827	36647	0.0027
EELTKB214	343253	0.4819	41760	0.0031

Andmestik	*E. coli* spetsiifilised *k*-meerid		*K. pneumoniae* spetsiifilised *k*-meerid	
	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal
EELTKB215	31853	0.0447	43902	0.0033
EELTKB219	40222	0.0565	34237	0.0025
EELTKB222	343776	0.4827	40022	0.003
EELTKB224	10519	0.0148	20276	0.0015
EELTKB225	142678	0.2003	39984	0.003
EELTKB227	17260	0.0242	36409	0.0027
EELTKB229	340486	0.4781	30511	0.0023
EELTKB230	22001	0.0309	38626	0.0029
EELTKB231	129812	0.1823	34147	0.0025
EELTKB233	122159	0.1715	25655	0.0019
EELTKB237	341794	0.4799	54227	0.004
EELTKB238	334860	0.4702	18506	0.0014
EELTKB241	329405	0.4625	20553	0.0015
EELTKB242	329185	0.4622	18109	0.0013
EELTKB244	59018	0.0829	16047	0.0012
EELTKB247	13181	0.0185	81973	0.0061
EELTKB248	135386	0.1901	19521	0.0015
EELTKB250	9442	0.0133	31358	0.0023
EELTKB251	10456	0.0147	19356	0.0014
EELTKB252	140268	0.1969	28574	0.0021
EELTKB253	40591	0.057	41736	0.0031
EELTKB254a	62523	0.0878	48007	0.0036
EELTKB254	61347	0.0861	40209	0.003
RUSPBB132a	11449	0.0161	618425	0.046
RUSPBB132	11632	0.0163	609773	0.0453
RUSPBB134	5855	0.0082	589254	0.0438
RUSPBB173	10626	0.0149	592751	0.0441
RUSPBB174	6736	0.0095	591684	0.044
RUSPBB180a	5873	0.0082	617227	0.0459

Andmestik	*E. coli* spetsiifilised *k*-meerid		*K. pneumoniae* spetsiifilised *k*-meerid	
	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal	leitud *k*-meeride arv	leitud spetsiifiliste *k*-meeride osakaal
RUSPBB180	12398	0.0174	610927	0.0454
RUSPBB260	8572	0.012	591331	0.044
RUSPBB265a	7129	0.01	574267	0.0427
RUSPBB265	18294	0.0257	594204	0.0442
RUSPBB266	10920	0.0153	611823	0.0455
RUSPBB269	11564	0.0162	609462	0.0453
RUSPBB279	5942	0.0083	610478	0.0454
RUSPBB280	10198	0.0143	608840	0.0453
RUSPBB281	5829	0.0082	612468	0.0455
RUSPBB323	10612	0.0149	592660	0.0441
RUSPBB324	6873	0.0097	593398	0.0441
RUSPBB73	8075	0.0113	607143	0.0451
RUSPBB78	17696	0.0248	618110	0.046
RUSPBB84	9781	0.0137	601735	0.0447

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Maarja Lepamets** (sünnikuupäev: 23. detsember 1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Oligomeeridel põhinevate bioinformaatiliste meetodite kasutamine bakterite määramiseks sekveneerimislugemitest,

mille juhendaja on **Lauris Kaplinski**,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 25. mai 2014.