

Tartu Ülikool
Botaanika- ja ökoloogia instituut

Kessy Abarenkov

**Seeneliikide määramine DNA triipkoodi meetodil ja tulemuste
rakendamine biogeograafilises uurimustöös**

Magistritöö

Juhendaja: prof. Urmas Kõljalg
Kaasjuhendaja: prof. Mairo Remm

Tartu 2006

Sisukord

1. Sissejuhatus.....	3
1.1 Liikide määramine.....	3
1.2 Liikide eristamise täpsus.....	3
1.3 Triipkoodistamisel kasutatavad DNA järjestused.....	5
1.4 Seente perekonnad <i>Tomentella</i> ja <i>Thelephora</i>	6
2. Materjal ja meetodika.....	7
2.1 Andmestiku iseloomustus.....	7
2.2 Fülogeneetilised analüüsid.....	8
2.3 'Pairwise' distantide analüüs.....	8
2.4 Liikide nimetamise põhimõtted.....	10
2.5 Liikide eristamine.....	10
3. Tulemused.....	11
4. Arutelu.....	16
4.1 Perekondade <i>Thelephora</i> ja <i>Tomentella</i> analüüs.....	16
4.2 Kasutatud meetodika analüüs.....	18
Kokkuvõte.....	20
Summary.....	21
Tänuõnad.....	22
Kasutatud kirjandus.....	23

Lisad

1. Sissejuhatus

1.1 Liikide määramine

Liikide määramiseks on traditsiooniliselt kasutatud fenotüübilisi tunnuseid. Hebert jt. (2003) on välja toonud mõningad sellise tavapärase lähenemise puudused: 1) liikide identifitseerimiseks kasutatavate tunnuste fenotüübiline plastilisus ja geneetiline varieeruvus võivad viia vale määramiseni; 2) kõrvale jäetakse morfoloogiliselt krüptiliste liikide esinemise võimalus; 3) vaadeldavad tunnused võivad olla seotud kindla elustaadiumi või sooga, mistõttu osad isendid jäävad määramata; 4) täpne määramine nõuab eksperdi oskusi ja kogemusi. Eelnevast lähtudes on välja pakutud, et lisaks morfoloogilistele tunnustele võiks liikide määramisel olulist rolli mängida ka DNA järjestused, mis võimaldaks antud ülesannet täita lihtsamini ja kiiremini (Hebert jt. 2003, Tautz jt. 2003). Kui varasematel aastatel kasutati DNA-l põhinevaid meetodeid peamiselt liikide määramiseks organismirühmades, millel morfoloogiliste tunnuste võrdlemine oli tihti väheinformatiivne (viirused, bakterid), siis järjest rohkem ilmub kõrgemaid organisme käsitlevaid uurimustöid, kus on rakendatud DNA-l põhinevaid identifitseerimissüsteeme (Druzhinina jt. 2005, Kress jt. 2005). Liikide määramiseks kasutatavaid DNA järjestusi on hakatud kutsuma triipkoodideks (Hebert jt. 2003).

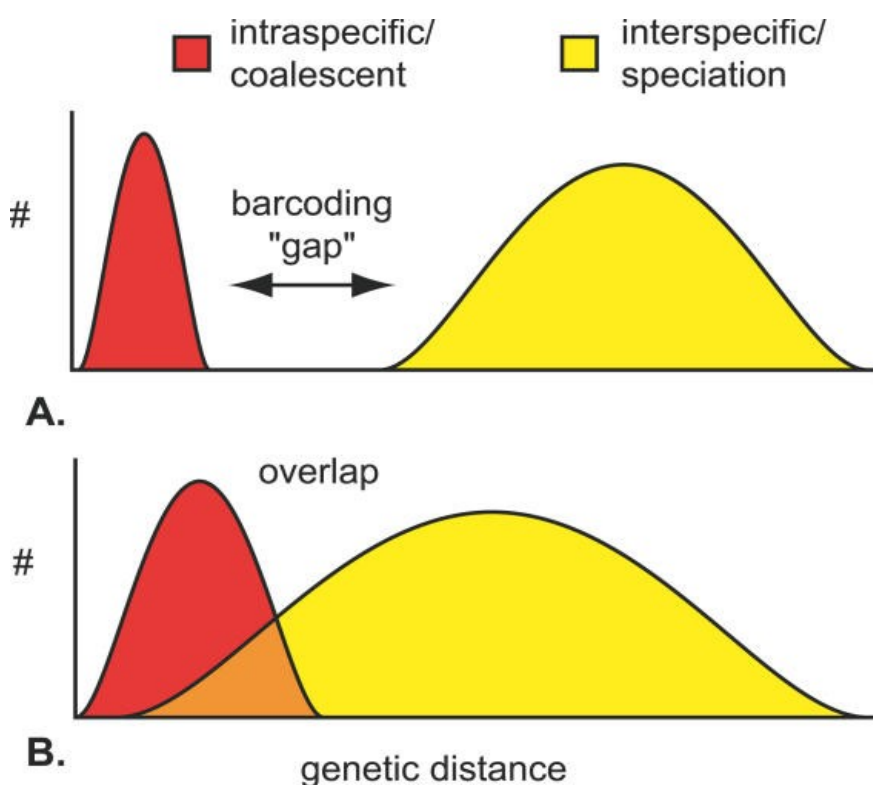
DNA triipkoodi meetod on pälvinud nii poolehoidu kui skeptilist käsitlemist. Poolehoidjate nägemuse järgi on lühikesi DNA fragmente võimalik kasutada liikide identifitseerimiseks tõstes selle kiirust, objektiivsust ja efektiivsust (Meyer ja Paulay 2005). Täpne määramine nõuaks muidugi põhjalikku molekulaarset andmebaasi, mille vastu tundmatut eksemplari võrrelda. Kuid kuna suurem osa bioloogilisest mitmekesisusest on kirjeldamata (arvatavast 10 – 15 miljonist maal eksisteerivast liigist on hetkel kirjeldatud 1,7 milj. (Savolainen jt. 2005)), on välja pakutud, et DNA triipkoodistamise teiseks eesmärgiks võiks olla kaasa aitamine uute liikide avastamisele. Vastased leiavad, et triipkoodistamine oleks samm tagasi ning hävitaks traditsioonilise süstemaatika muutes taksonoomia tüpoloogiaks. Lisaks kardetakse, et liikide määramine pelgalt geneetilise lahknevuse hulka kasutades võib viia ebaõige tulemuseni. Triipkoodi meetodi plusse ja miinuseid on analüüsinud Moritz ja Cicero (2004) ning Marshall (2005).

Põhiküsimused seoses DNA triipkoodidega on: 1) kas lühikesi DNA järjestusi kasutades on võimalik eristada nii suurt hulka liike? 2) kas antud meetod võimaldab eristada ka lähedasi või kiirelt evolutsioneeruvaid liike? 3) millised DNA järjestused oleksid erinevate taksonite eristamiseks kõige sobilikumad (Savolainen jt. 2005)?

1.2 Liikide eristamise täpsus

Keskseks probleemiks triipkoodistamise juures on täpsus, millega ühe geeni järjestuse põhjal saab liike identifitseerida. Täpsus oleneb suuresti sellest, kui suur on kindla markeri suhtes

liigisisene varieeruvus ja kui suur liikidevaheline erinevus, ning kui hästi nad omavahel eristuvad. Mida suurem on liigisisese geneetilise varieeruvuse ja sõsarliikide vahelise erinevuse kattuvus, seda halvemini sobib antud marker triipkoodistamiseks. Ideaalsel juhul ühildumine puuduks, nagu on näidatud joonisel 1 A. Algselt erinevate andmestike testimisel leiti, et liigisisese varieeruvuse ja liikidevaheliste erinevuste vahel on tühimik, mis oli siiski tingitud ebapiisavast andmete hulgast (1-2 järjestust liigi kohta ning proovide päritolu väike geograafiline ulatus, põhjustades väikese liigisisese varieerumise) (Hebert jt. 2004). Reaalselt arvatakse kahe jaotuse vahel olevat alati mõningane kattumine (Joonis 1 B), mis võib olla tingitud mitmest erinevast nähtusest. Üheks põhjuseks võib olla näiteks geneetiliselt polü- ja parafüleetiliste liikide esinemine, samuti olukord, kus liigisisene varieeruvus mõnes puu harus (liigis) on suurem kui liikidevahelised erinevused teistes puu osades (liikides). Esimesel juhul ei sobi antud marker suure tõenäosusega nende liikide omavaheliseks eristamiseks, teise näite puhul võib probleem tekkida mittetäielikult esindatud gruppides, kus eelnevalt defineeritud rühmadest väljajääva tundmatu eksemplari staatust on raske hinnata (Meyer ja Pauley 2005).



Joonis 1. Skeem triipkoodistamise kohta. Liigisisene varieeruvus on näidatud tumeda, liikidevaheline erinevus heleda taustaga. (A) ideaalne variant, jaotused ei kattu. (B) alternatiivne versioon jaotuste kattumisega. (Meyer ja Pauley 2005)

On välja pakutud, et tundmatute DNA järjestuste määramine mittetäielikult esindatud rühmade fülogeneesis võiks toimuda läviväärtuste kaudu, mis eristavad liigisiseseid varieeruvusi ja

liikidevahelisi erinevusi. Seega, kui tundmatu järjestus erineb teatud määratud järjestusest vähem kui kokkulepitud läviväärtus, on tegemist sama liigiga, kui tema erinevus ületab aga kindlaksmääratud erinevusi teiste liikidega, on tegemist uue taksoniga. Selline meetod annab aga võimaluse vale-positiivsete ja vale-negatiivsete tekkele. Vale-positiivseteks osutuvad need, mis määratakse gruppi, mille liigisisene varieeruvus on väiksem kui etteantud läviväärtus. Vale-negatiivsed on aga eksemplarid, mis peaksid kuuluma liiki, mille liigisisene varieeruvus ületab läviväärtust. Sellise läviväärtuste kasutamise puhul on liikide määramise täpsus suuresti sõltuvuses joonisel 1 näidatud liigisisese varieeruvuse ja liikidevahelise erinevuse jaotuste kattumisest. Läviväärtuste kasutamise võimalikkust ja liikide määramise täpsuse sõltuvust olemasoleva andmestiku põhjalikkusest on uurinud Meyer ja Paulay (2005), kes leidsid, et põhjalikult uuritud rühmade puhul annab triipkoodistamine hea tulemuse, mistõttu tuleks suuremat rõhku panna põhjalikult läbiuuritud võrdlevate andmebaaside loomisele. Hoiduda tuleks aga läviväärtuste kasutamisest liikide identifitseerimisel, kuna enamike suuremate klaadide puhul liigisisese varieeruvuse ja liikidevaheliste erinevuste jaotused kattuvad.

1.3 Triipkoodistamisel kasutatavad DNA järjestused

Triipkoodiks soovitatakse eri organismirühmadel erinevaid DNA piirkondi. Taimedel kasutatakse näiteks tuuma ribosomaalse tsistroni ITS (*internal transcribed spacer*) piirkonda ning plastiidide DNA järjestusi (Kress jt. 2005), loomadel on seevastu häid tulemusi andnud tsütokroom C oksidaasi I geen (COI), mis on lihtsalt tuvastatav ja omab head eristatavust (Hebert jt. 2004, Meyer ja Paulay 2005).

Seente identifitseerimiseks liigi tasemel on seni kasutatud peamiselt ITS piirkonda (Pritsch jt. 2000, Kõljalg jt. 2000, Druzhinina jt. 2005, Kõljalg jt. 2005, O'Brien jt. 2005).

Selleks, et testida triipkoodi meetodi sobivust kindlate DNA järjestuste puhul, on tarvis põhjalikku võrdlevat andmebaasi juba määratud järjestustega, mis oleks korrektselt nimetatud. Nimelt on seente puhul suur hulk järjestusi avalikes andmebaasides puudulikult määratud, põhjuseks nt. vead sekveneerimisel, vananenud määrang vm. Vilgalys (2003) on välja pakkunud, et kuni 20% geenipankades olevatest järjestustest võivad olla valesti määratud. Seda tõestas ka UNITE (<http://unite.zbi.ee>) andmebaasis olevate sekvenside jooksutamise *blast* programmi kasutades geenipanga (NCBI) järjestuste vastu, kus üle 20% sekvensidest andis suurima skoori teise liiginimega järjestusega. Vastuseks on alustatud uute andmebaaside loomisega, milles olevad sekvensid on identifitseeritud ekspertide poolt ning valede määrangute esinemise tõenäosus on viidud miinimumini (Kõljalg jt. 2005).

1.4 Seente perekonnad *Tomentella* ja *Thelephora*

Antud töös kasutatakse loodava bioinformaatilise süsteemi testimiseks perekondadesse *Tomentella* Pat. ja *Thelephora* Fr. kuuluvate seeneliikide ITS piirkondade nukleotiidsed järjestusi. Mõlemad perekonnad on globaalselt levinud, sh on mitmeid globaalselt levivaid liike nagu näiteks *Thelephora terrestris* Fr. (Corner 1968, Larsen 1968 ja 1974, Stalpers 1993, Kõljalg 1996). Perekonnas *Thelephora* on kirjeldatud üle 50 liigi ja paljud neist on kirjeldatud ainult ühe eksemplari põhjal (Corner 1968, Stalpers 1993). Perekonnas *Tomentella* on kirjeldatud ligikaudu 100 liiki ja ka siin on palju, eriti lõunapoolkeralt ühe eksemplari põhjal kirjeldatud liike (Larsen 1968 ja 1974, Stalpers 1993, Kõljalg 1996). Kuni viimase ajani on neid perekondi käsitletud ühte sugukonda Thelephoraceae (lehternahkiselised) kuuluvatena, aga mitte sõsar perekondadena (Corner 1968, Larsen 1968, Stalpers 1993). Morfoloogiliste (Kõljalg 1996) ja molekulaarsete (Kõljalg jt. 2000) tunnuste põhjal on nüüdseks selgunud, et tegemist on siiski sõsar perekondadega. Viimaste uurimuste põhjal saab väita, et tegemist on teineteise suhtes parafüleetiliste taksonitega ning üks osa *Tomentella* liike tuleb üle viia perekonda *Thelephora* (Kõljalg, suulised andmed). Seega on mõlema perekonna liikide kaasamine samasse analüüsi igati põhjendatud.

Perekonna *Thelephora* liikide võime moodustada puujuurtel seenjuuri on üldteada juba väga pikka aega (Weir 1921, Marx ja Bryan 1970). Perekonna *Tomentella* liike on aga traditsiooniliselt peetud surnud puidu lagundajateks, sest nende viljakehad viljuvad peamiselt pinnasel asuva surnud puidu alaküljel (Larsen 1968, Stalpers 1993). Suhteliselt hiljuti avastati aga järk-järgult, et ka perekonna *Tomentella* liigid moodustavad taimejuurtel seenjuuri. Kõigepealt püstitati hüpotees, et puudel esinevad teatud pruuni värvi seenjuured võivad olla moodustatud perekonna *Tomentella* liikide poolt (Danielson jt. 1984, Danielson ja Pruden 1989). Mõned aastad hiljem sünteesiti *in vitro* tingimustes seenjuured hariliku männi ja liigi *Tomentella crinalis* (Fr.) M.J.Larsen vahel (Kõljalg 1992). Hiljem kirjeldas Agerer (1996) morfoloogiliste ja anatoomiliste tunnuste põhjal hariliku männi ja *Tomentella albomarginata* (Bourdot & Galzin) M.J.Larsen vahelist ektomükoriisat. Eelmise sajandi üheksakümnendatel aastatel hakati seenjuurte mükobiondi määramiseks kasutama molekulaarseid meetodeid. See avas tee taimejuurte seensümbiontide kiirele ja küllaltki täpsele määramisele. Õige pea avastati, et perekonna *Tomentella* liigid mitte ainult ei moodusta puujuurtel ektomükoriisat, vaid on ühed kõige sagedamad mükoriisa moodustajad (Gardes ja Bruns 1996, Horton ja Bruns 1998, Baar jt. 1999, Kõljalg jt 2000, Kõljalg jt. 2001, Tedersoo jt. 2003, Tedersoo jt. 2006). Samuti leiti, et perekonna *Tomentella* liigid on orhideede mõningates uuritud rühmades ühed sagedamini esinevad juursümbiondid (Taylor ja Bruns 1997, McKendrik jt. 2000).

Prof. U. Kõljala käsutuses olevad avaldamata ITS regioonide nukleotiidsed järjestused ja avalikes geenipankades (EMBL, NCBI, DDBJ) olevad järjestused näitavad, et perekondade

Thelephora ja *Tomentella* liigid esinevad väga erinevates kooslustes kõikjal maailmas ning tõenäoliselt on üks olulisemaid ekto- ja orhidoidset mükoriisat moodustavaid taksoneid. Olemasolevate DNA järjestuste suur varieeruvus viitab samuti selle rühma väga kõrgele liigirikkusele. Seetõttu on perekondade *Thelephora* ja *Tomentella* liikide kiire, täpne ja võimalikult automatiseeritud määramine molekulaarsete meetodite, so ITS järjestuste baasil väga oluline ökoloogidele, biogeograafidele, taimede ja seente koevolutsiiooni uurijatele jt.

Antud töö eesmärgiks oli välja töötada esialgne metoodika, mis võimaldaks: 1) alla laadida geenipankades vabalt kättesaadavad valitud seenetaksoni ITS järjestused; 2) liita need olemasolevate avaldamata ITS järjestustega; 3) liidetud järjestused automaatselt aligneerida ehk järjendada; 4) jagada järjendatud sekventsids fülogeneetilise analüüsi ja järjestuste vahelise 'pairwise' distantssi alusel liikideks kasutades selleks kindlat läviväärtust; 5) leida ja liita järjestuste metaandmed (DNA päritolu, leiukoht, peremeestaim jne.) sekventsides andmebaasile; 6) saadud tulemuste põhjal viia läbi liikide biogeograafia ning taimede ja seente koevolutsiiooni alaseid analüüse.

2. Materjal ja metoodika

2.1 Andmestiku iseloomustus

Töös on kasutatud perekondadesse *Tomentella* ja *Thelephora* kuuluvate liikide eksemplaridelt eraldatud ribosomaalse DNA ITS1-5.8S-ITS2 järjestusi. Erandina kuulus analüüsi ka üks liigi *Pseudotomentella ochracea* Kõljalg & E. Larsson järjestus. Esimene osa andmestikust pärineb avalikest geenipankadest (EMBL, NCBI ja DDBJ), kusjuures kõik 300 kuni 1500 nukleotiidi pikkused seltsi Thelephorales (lehternahkiselaadsed) kuuluvad ITS järjestused laaditi alla NCBI kodulehelt seisuga 03.03.2006 (otsingu tingimused: "thelephorales[ORGN] and 300:1500[SLEN] and (ITS1 OR ITS2 or 5.8S or transcribed spacer or spacers)"). Seda võimaldab fakt, et kõik kolm eelnimetatud geenipanka vahetavad omavahel kord päevas andmeid, so uusi DNA järjestusi. Geenipangast pärinevatest seltsi Thelephorales ITS sekventsides eraldati perekondadesse *Thelephora* ja *Tomentella* kuuluvad järjestused kasutades UNITE andmebaasi programme *blastn* ja *galaxieBLAST* (<http://unite.ut.ee>, Kõljalg jt. 2005). Teise osa moodustavad prof. U. Kõljala käsutuses olevad perekondadesse *Thelephora* ja *Tomentella* kuuluvate liikide avaldatud ja avaldamata ITS järjestused. Kõik saadud sekventsids viidi erast.ut.ee serveris asuvasse mysql-põhisesse relatsioonilisse andmebaasi tabelitesse MAIN ja MAIN_2 (vt. lisa 1). Lisaks DNA järjestusele kanti andmebaasi nende olemasolu korral ka erinevad metaandmed: eksemplari leiukoht, päritolu (viljakehast, ektomükoriisast, pinnaseproovist jne.), peremeestaim ja viited kirjandusele. Enamikel juhtudel pärinevad metaandmed kas artiklitest või otse autoritelt.

Peremeestaime liikide asend eluslooduse süsteemis määrati NCBI taksonoomia brauserit (<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/>) kasutades. Eksemplaride leiukohtad kodeeriti vastavalt TDWG taimede geograafilise leviku kaardistamise standardile (Hollis ja Brummitt 1992). Metaandmete, sekvensi koodide ja NCBI blasti kasutamise abil tehti kindlaks ja märgiti ära kahe liidetud andmestiku kattuvad kirjed, mis võimaldas neid mitte kaasata järgnevasse analüüsisse rohkem kui ühes korduses.

2.2 Fülogeneetilised analüüsid

ITS järjestused aligneeriti programmiga *MAFFT ver. 5.743* (Katoh jt. 2002) FFT-NS-2 meetodil, ning korrigeeriti vajadusel käsitsi kasutades *SeaView* (Galtier jt. 1996) programmi. Nn. gap alad aligneeringu alguses ja lõpus kustutati. Lõplik maatriks sisaldas 695 järjestust ja 793 tunnust. Evolutsioonimudeli valimiseks kasutati *Modeltest ver. 3.7* tarkvara, mis leidis kaks sobivat mudelit. Posada (2004) on välja toonud AIC (Akaike information criterion) eelised hLRT (hierarchical likelihood ratio tests) ees soovitades mudeli valikul kasutada esimest meetodit. Näiteks on AIC vastupidiselt hLRT-le võimeline samaaegselt võrdlema erinevaid omavahel seotud ja seoseta mudeleid. Edasiseks analüüsiks valitigi AIC poolt välja pakutud transversiooniline mudel (TVM) koos muutumatute saitide osakaalu (+I) ja gamma jaotusega (+G). Nimetatud mudelit kasutati lähimsideme (ingl. k. neighbour joining, NJ) distantide analüüsi tegemisel programmiga *PAUP 4.0d81* (Swofford, 2003).

Kasutades *MacClade ver. 4.06* (Maddison ja Maddison 2003) ja *Mesquite ver. 1.06* (Maddison ja Maddison 2005) tarkvara lisati maatriksisse ka sekvenside päritolu, leiukohta ja peremeestaime iseloomustavad tunnused, neist kahte esimest iseloomustavad jaotused on ära toodud tabelis 1. Suurem osa sekvensidest pärinesid viljakehadest ja ektomükoriisast, mis olid enamikel juhtudel korjatud kas Euroopast või Põhja–Ameerikast.

Tabel 1. Sekvenside päritolu ja leiukohta kirjeldavad jaotused.

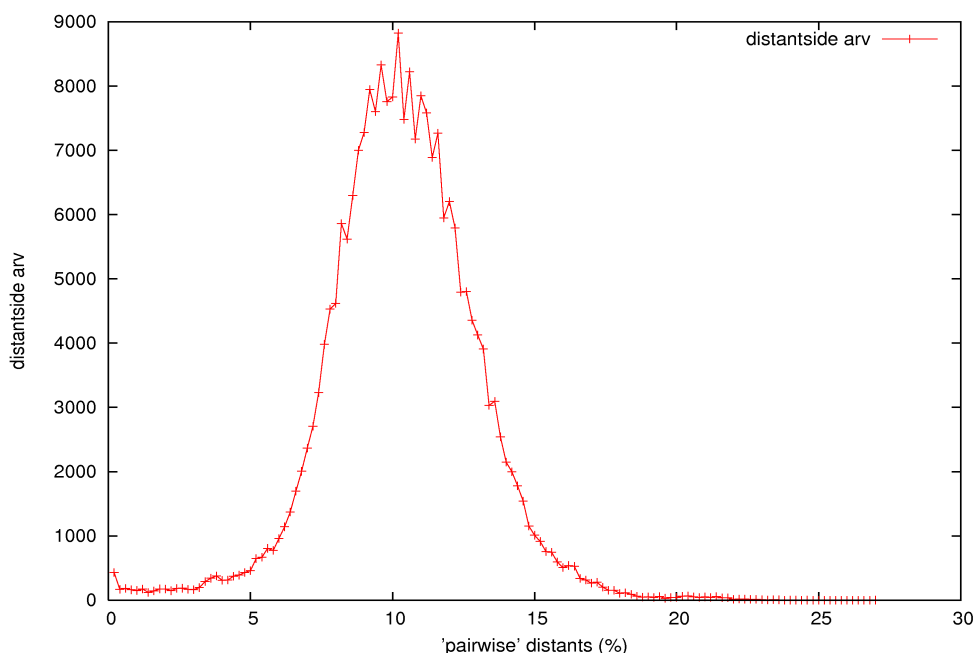
DNA päritolu	sekvenside arv	Leiukoht	sekvenside arv
teadmata	1	teadmata	19
viljakehadest	229	Euroopa	402
ektomükoriisast	341	Aasia, parasvööde	10
orhoidsest mükoriisast	72	Aasia, troopiline	2
lagunenud puidust	3	Australaasia	28
pinnasest	34	Põhja – Ameerika	202
mükoriisast	11	Lõuna – Ameerika	33
keskkonnaproovist	5		

2.3 'Pairwise' distantide analüüs

Aligneeritud järjestuste 'pairwise' distantide arvutamiseks kasutati *EMBOSS* tarkvara (Rice

jt. 2000) programmi *distmat*, mille erinevatest võimalikest arvutusmeetoditest valiti korrigeerimata distantsid (ingl. k. uncorrected distances), kuna kõik nukleotiidsetele järjestustele mõeldud meetodid andsid järjestuste grupeerimise seisukohalt sama lõpptulemuse.

Saadud 695 X 695 maatriksi analüüsimiseks kirjutati programmeerimiskeeles *perl* (<http://www.perl.org/>) skriptid, et hõlbustada suure arvude hulga ümberkäimist (näiteks on Microsoft Office paketi Excel 2000 programmis maksimaalne töölehe suurus 65 536 rida 256 veerus, kuid 695 X 695 maatriksis on 241 860 arvu). Perli skript “*dist_final.pl*” (Lisa 2) koostab etteantud maatriksi alusel distantside sagedustabeli 0,2 protsendilise klassipikkuse suurenemisega. Joonis 2 näitab lõpliku maatriksi 'pairwise' distantside sageduste jaotust, milles kahe järjestuse maksimaalne erinevus oli 26,96% ning 90% distantsidest kuulus erinevuste vahemikku 6,16 – 14,39%.



Joonis 2. Analüüsitud 695 ITS järjestuse omavaheliste 'pairwise' distantside esinemissagedus kuni 27%-ni.

Kasutades distantside maatriksit leiti järjestuste grupid, millesse kuuluvad sekvensid olid ühendatud distantside erinevusega kuni 3%, mis võiks olla konservatiivseks piiriks eristamiseks kandseente hulka kuuluvaid liike üksteisest (O'Brien jt. 2005, Tedersoo jt. 2003). Selleks kirjutati perli skript “*3_percent_final.pl*” (Lisa 3), mis etteantud distantside maatriksi alusel leiab, ja võimaluse korral ühendab järjestuste paarid 'pairwise' distantside erinevusega kuni 3%. Skripti väljundiks on *Mesquite* programmi jaoks *nexus* formaadis failile lisatavad plokid, mis (1) määravad ära fülogeneesipuul gruppide illustreerimiseks kasutatavad värvikoodid, ning (2) näitavad ära kindlatesse gruppidesse kuuluvad järjestused. Antud etapis identifitseeriti ja eemaldati analüüsist 23

sekvensi, mis käitusid 'pairwise' distantide arvutamisel 'jokkeritena' andes sarnasuse üle 97% paljude kaugete järjestustega. Need olid lühikesed sekvensid, millel puudus suurem järjestus ühest või ka mõlemast ITS piirkonnast. Lõplik maatriks sisaldas 695 järjestust 793 DNA tunnuse ja 5 peremeestaime ja sekvensi päritolu näitava tunnusega.

Sõsarliikide liigisisest varieeruvust ja liikidevahelist erinevust demonstreerivate graafikute jaoks valiti järjestuste grupid, mis asetsesid fülogeneesipuul (Lisa 4) lähestikku või olid 'pairwise' distantide 3 % piiri rakendamisel kokku liidetud (*Thelephora terrestris* ja *Tomentella sublilacina*³, *T. stuposa*² ja *T. stuposa*³ ning *T. atramentaria*¹ ja *T. badia*³) ning sisaldasid iga liigi kohta vähemalt 8 sekvensi. Graafikute joonistamiseks kasutati programmi *Gnuplot ver. 4.0*, jooned punktide ühendamiseks on vaid illustratiivsetel eesmärkidel.

2.4 Liikide nimetamise põhimõtted

Juhul, kus analüüsi põhjal ühte liiki kuuluvate sekvenside seas oli viljakeha(de)st pärit DNA, kasutati liigi tähistamiseks vastava viljakeha määrangut, so liiginime. Liigi epiteedile lisati numeratsioon juhtudel kui samaks morfoliigiks määratud viljakehade ITS järjestused kuulusid analüüsi põhjal erinevatesse nn. krüptilistesse liikidesse. Kui analüüsi põhjal saadud ühte liiki kuulusid ainult keskkonnaproovidest pärit ITS järjestused või kui sinna kuuluvat viljakeha ei õnnestunud liigini määrata kasutati liigi tähistamiseks perekonnanimele järgnevat numeratsiooni kujul "sp1", "sp2" jne. Kuna perekonnad *Thelephora* ja *Tomentella* on üksteise suhtes parafüleetilised, st asetsevad fülogeneesi puul läbisegi, siis kasutatakse nende puhul ühtset numeratsiooni. Töös kasutatud viljakehad määras U. Kõljalg.

2.5 Liikide eristamine

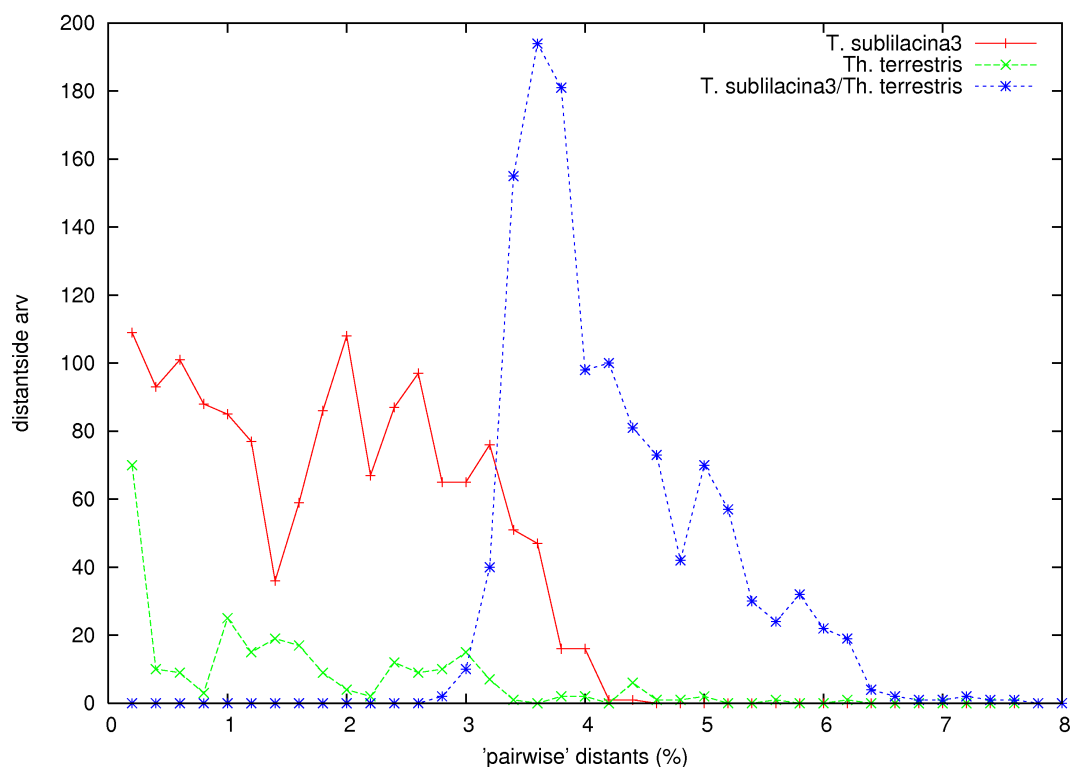
Nagu öeldud, kasutati liikide eristamiseks ITS järjestuste abil nende erinevuse/sarnasuse koefitsenti protsentides. Varasemates seente koosluste uuringutes on kaks ITS järjestust arvatud ühte ja samasse liiki kui nendevaheline erinevus on alla 3% (Tedersoo jt. 2003; O'Brien 2005). Sellist erinevuse protsenti peetakse konservatiivseks lähenemiseks, mis annab pigem väiksema liikide arvu. Ka antud uurimistöös selgus, et mõned morfoloogiliselt hästi eristatavad liigid on omavahel ITS järjestuste osas sarnasemad kui 97% ja tuleks antud meetodika alusel üheks liigiks liita. Viimasel juhul näidatakse fülogeneesi puul (Lisa 4) eri värvidega liigigruppe, mis moodustuvad 3% erinevuse rakendamisel ning ühe grupi sees asuvad nn. head morfoloogilised liigid eristatakse liiginimedega. Kokku oli kolm juhtu, kus 3% erinevuse rakendamisel saadud liigid sai morfoloogiliste tunnuste alusel kaheks eri liigiks jagada.

3. Tulemused

Töös kasutatud perekondadesse *Thelephora* ja *Tomentella* (ning üks *Pseudotomentella* liik) kuuluvad 695 ITS järjestust jagunesid 'pairwise' distantide 3 %-list erinevust rakendades 247 liigiks. Eristatud liigid on ära toodud Lisa 4. asuval kladogrammil. Kui ühte liiki kuuluvaks määrati rohkem kui üks sekvents, on nende järjestuste nimed tähistatud ühe ja sama värviga. Ainult ühe järjestusega liigid on märgitud musta värviga. Viimaseid oli 140 liiki 247-st.

Rakendatud 3% lävemeetod ei suutnud eristada kuut morfoloogiliselt hästi eristuvat liiki moodustades neist kolm liiki. Ühte liiki arvatud liigipaarideks olid *Tomentella galzini*1 ja *T. viridula*, *T. atramentaria*1 ja *T. badia*3 ning *T. sublilacina*3 ja *Thelephora terrestris* (nimekujud vastavad kladogrammil olevatele nimedele). Kasutades morfoloogilisi ja anatoomilisi tunnuseid on need liigipaarid Lisa 4. asuvaltel joonistel ja kladogrammil eristatud omaette liikideks.

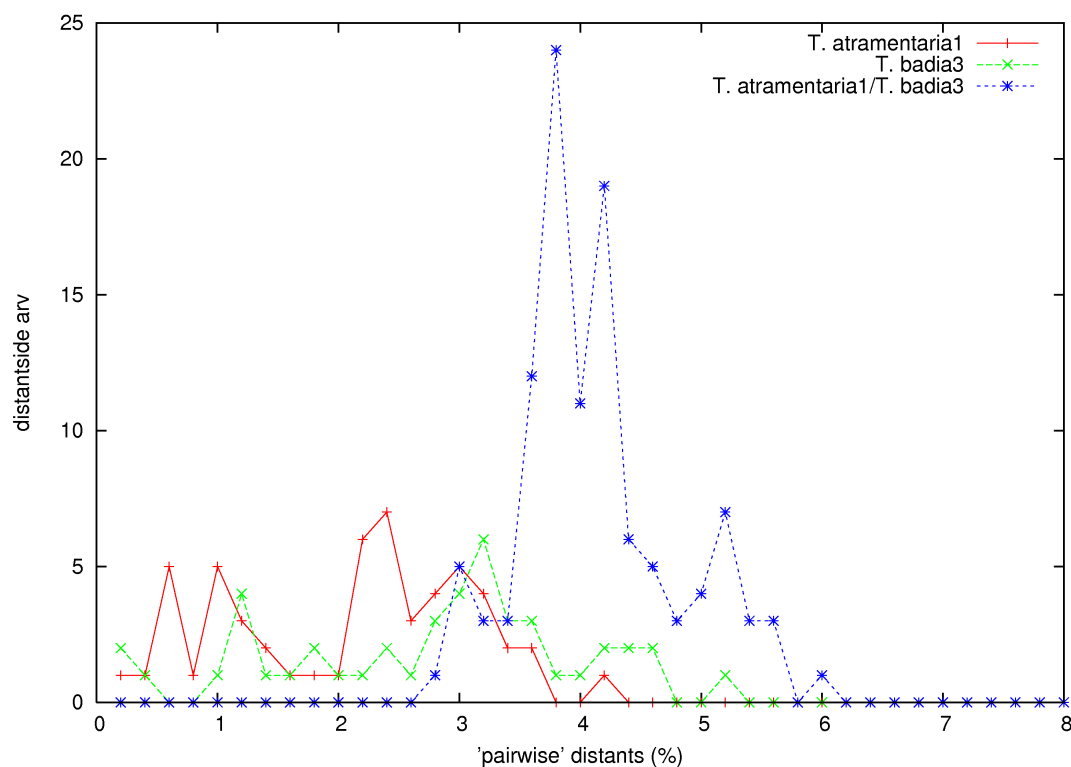
Kahe sõsarliigi, *Thelephora terrestris* ja *Tomentella sublilacina*3, liigisiseseid varieeruvused on suuremad kui on nende liikide vaheline erinevus. Jooniselt 3 võib välja lugeda, et piir, mille juures kaks liiki ühinevad, on 2,6 ja 3 % vahepeal, ning selle piiri sisse mahub vaid 1-2 järjestust.



Joonis 3. *Tomentella sublilacina*3 (tähistatud ristikestega) ja *Thelephora terrestris* (tähistatud x-dega) liigisiseseid 'pairwise' distantideid. Täpnidega märgitud joon tähistab liikidevahelisi 'pairwise' distantse. Viimasel juhul on näha, et liigi eristamise läve nihutamine 2,6% juurde lahutaks need kaks liiki eri taksonitesse.

Need kaks liiki eristuvad teineteisest väga hästi morfoloogiliste tunnuste põhjal. *Thelephora terrestris* moodustab maapinnal väikeste kübaratega viljakehi, *Tomentella sublilacina* seevastu moodustab alati mulla ülemistes horisontides substraadile (surnud puit, lehed jne) kinnituvaid liibuvaid viljakehi (vt. fotosid Lisa 4.). Erinev on ka nende kahe liigi eoste kuju ja ornamentatsiooni tüüp (Kõljalg 1996).

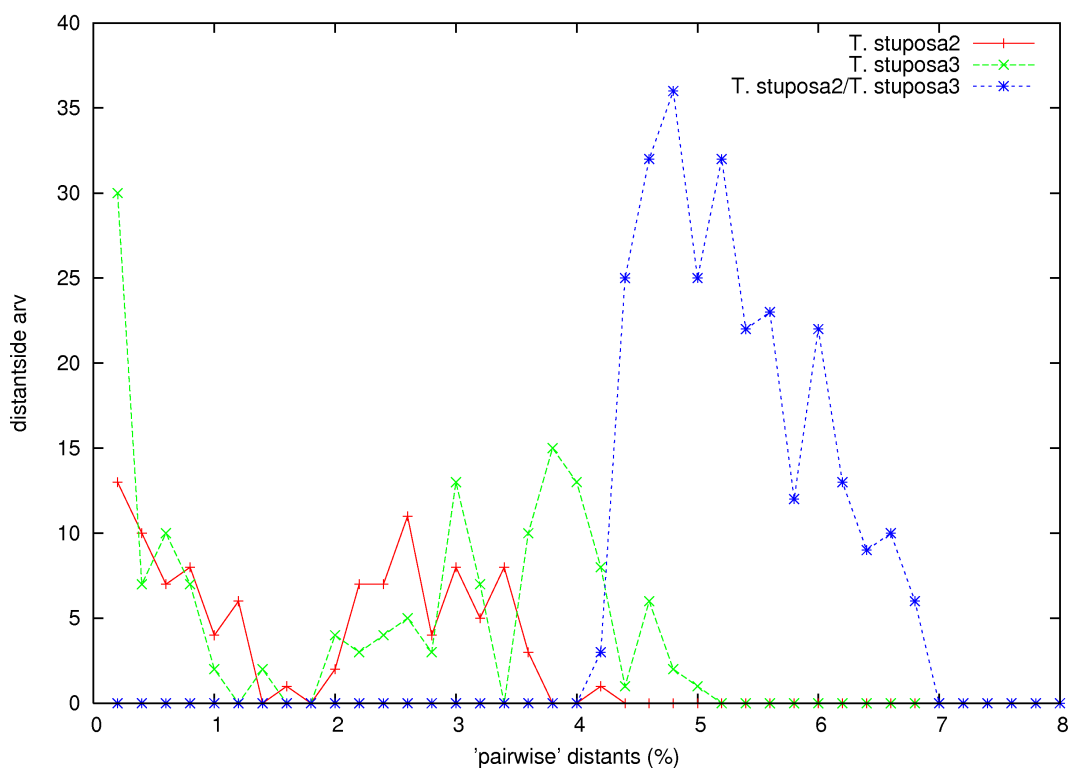
Sõsarliikide *Tomentella atramentaria*1 ja *T. badia*3 puhul, mis samuti 3 % rakendamisel kuuluvad ühte taksonisse, on eelmise paariga sarnane distantside jaotus (Joonis 4). Kõrge liigisisene varieeruvus *T. badia*3 sekventsides hulgas on põhjustatud nende erinevast geograafilisest päritolust. Põhja-Ameerikast ja Euroopast pärinevad järjestused erinevad kuni 5,2%. Anatoomiliselt on kaks liiki hästi eristuvad - *T. atramentaria* eksemplaridel puuduvad rakuvaheseintel pandlad ja eosed on märksa suuremad ning korrapärasema kujuga kui *T. badia* eosed (vt. jooniseid Lisa 4., Kõljalg 1996).



Joonis 4. *T. atramentaria*1 (tähistatud ristikestega) ja *T. badia*3 (tähistatud x-dega) liigisisene 'pairwise' distantid. Tärnidega märgitud joon tähistab liikidevahelisi 'pairwise' distantse. Viimasel juhul on näha, et liigi eristamise läve nihutamine 2,6% juurde lahutaks need kaks liiki eri taksonitesse.

Analüüsil 3% läve kasutades üheks taksoniks arvatud liikide *Tomentella galzini*1 ja *T. viridula* puhul ei esitata 'pairwise' distantside diagrammi vähesel sekventsides arvu tõttu (vastavalt 11 ja 8 ITS järjestust, vt. Lisa 4. asuvat kladogrammi). Mõlemad liigid eristuvad väga hästi

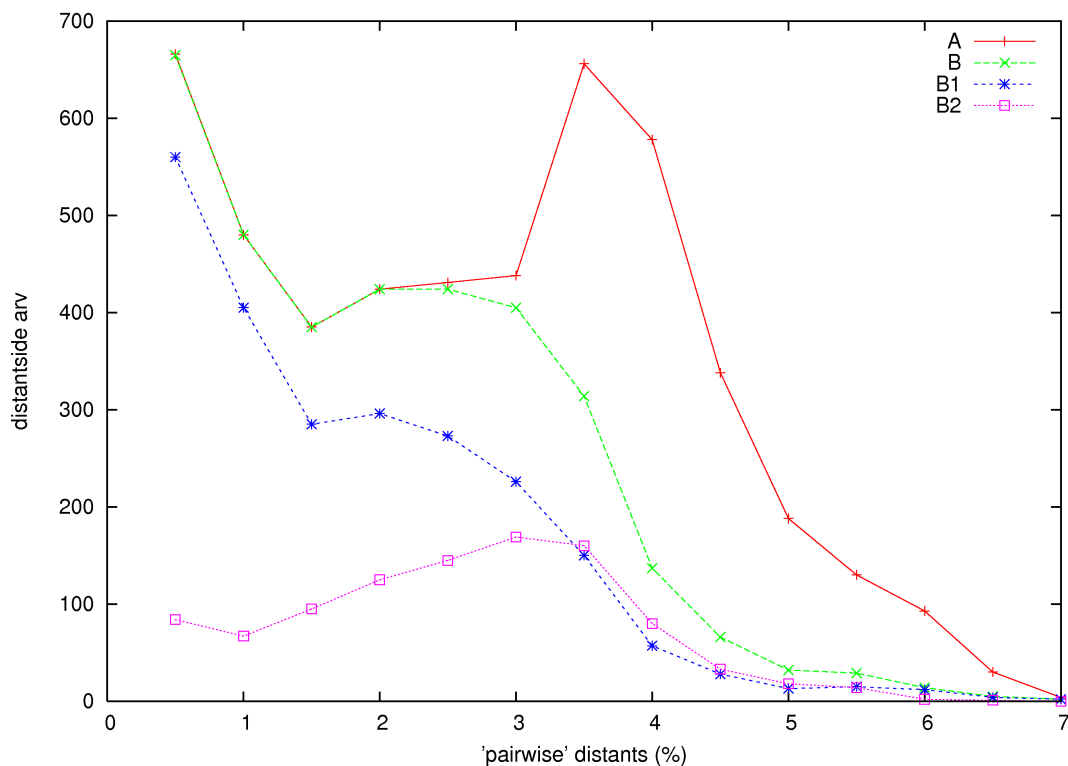
anatomiliste tunnuste põhjal - liigil *T. galzinii* arenevad hümeeniumis aheneva tipuga tsüstiidid ja liigil *T. viridula* puhetunud tipuga tsüstiidid (vt. joonised Lisa 4.; Kõljalg 1996).



Joonis 5. *T. stuposa2* (tähistatud ristikestega) ja *T. stuposa3* (tähistatud x-dega) liigisiseseid 'pairwise' distantseid. Tärnidega märgitud joon tähistab liikidevahelisi 'pairwise' distantse. Viimasel juhul on näha, et liigi eristamise läve tõstmise 4% juurde jätkaks need morfoloogiliselt väga sarnased taksonid ikkagi eri liikidesse.

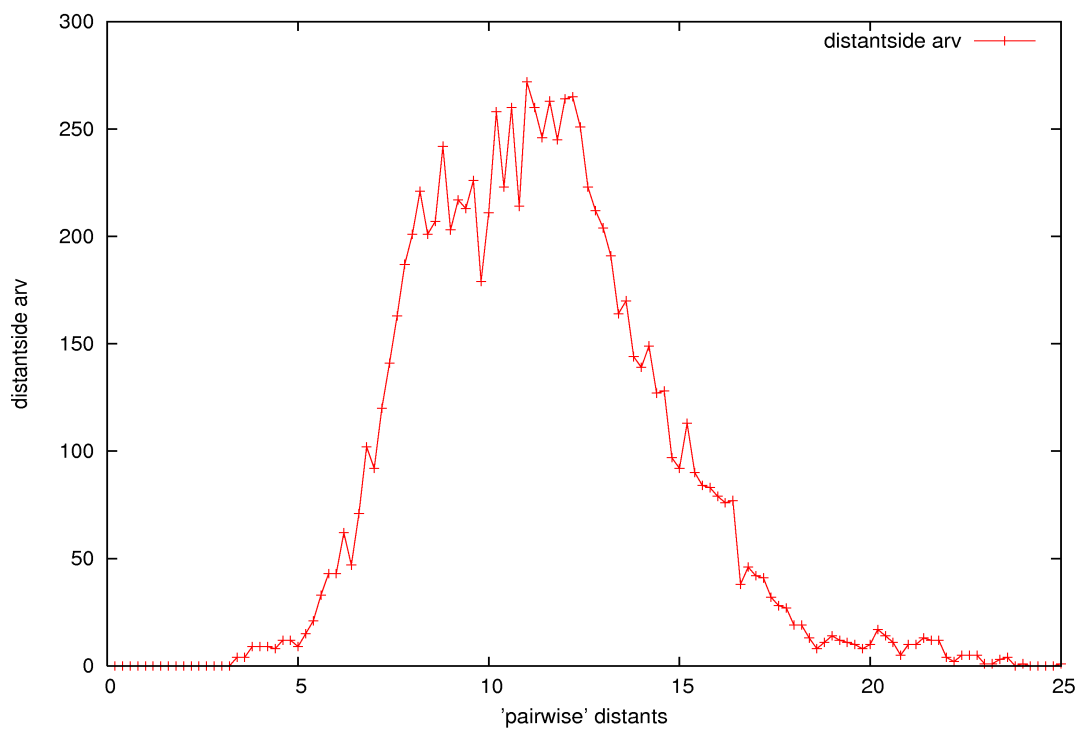
Eelmistest liigipaaridest erinev on liigi *T. stuposa* jagunemine 3% läve kasutamise korral morfoloogiliselt raskesti või mitteeristuvateks liikideks. Joonisel 5. on toodud kahe saadud liigi, *T. stuposa2* ja *T. stuposa3*, 'pairwise' distantide erinevused. Jooniselt on näha, et isegi läve tõstmise 4% juurde jätkaks need kaks taksonit eri liikidesse.

Saadud tulemused näitavad, et samalt või eri kontinendilt pärit ITS järjestuste liigisisene varieeruvus on väga erinev (Joonis 6). Samalt kontinendilt pärit ITS järjestuste liigisisene varieeruvus jääb rõhuvas enamuses 0,5 - 3% vahele näidates distantide arvu langust suurema erinevuse suunas. Eri kontinentidelt pärit järjestuste liigisisene varieeruvus jääb valdavalt 1,5 – 3,5% vahele aga näitab selget distantide arvu tõusu suurema erinevuse suunas. Läviväärtusega 3% saadud taksonite liigisisene varieeruvus näitab 3 – 4% vahel suurt distantide arvu tõusu (Joonis 6 A). See on tingitud ühte taksonisse arvatud liikide *T. sublilacina* ja *Thelephora terrestris* järjestuste suurest arvust.



Joonis 6. Analüüsil saadud liikide ITS regioonide liigisisene varieeruvus 'pairwise' distantide põhjal (ühe eksemplariga liigid välja jäetud). - A: Läviväärtusega 3% saadud taksonite liigisisene varieeruvus. - B: Liigisisene varieeruvus kui 3% läviväärtusega saadud kolm taksonit jagatud fenotüübiliste tunnuste alusel kuueks liigiks (*Tomentella atramentaria*1 ja *T. badia*3; *T. galzini*1 ja *T. viridula*; *T. sublilacina*3 ja *Thelephora terrestris*). - B1: Liigisisene varieeruvus samalt kontinendilt pärit eksemplaride põhjal. - B2: Liigisisene varieeruvus eri kontinentidelt pärit eksemplaride põhjal.

Ühe järjestusega liike saadi kokku 140, mis on enam kui pool kogu eristatud liikide arvust 247. Joonisel 7. on näidatud nende liikide vaheliste 'pairwise' distantide diagramm. Rõhuv osa 'pairwise' distantse on üle 5% erinevusega. Seetõttu saab väita, et enamik üksikute järjestustega liike on "head" liigid ja tõenäoliselt säilivad ka uute ITS järjestuste lisandumisel andmematriksisse. Samas on oluline märkida, et selle hüpoteesi kinnituseks on vaja uurida lähedaste liikide, sh sõsarliikide vahelisi 'pairwise' distantse. See annaks korrektsema pildi üksikute järjestuste põhjal eristatud liikide "headusest". Vajadus sellise analüüsi järele tekkis alles töö kirjutamise lõppfaasis ja seetõttu ei ole seda analüüsi töösse lisatud.



Joonis 7. Analüüsil saadud 140 üksikust ITS järjestusest koosneva liigi omavaheliste distantide jaotus.

4. Arutelu

4.1 Perekondade *Thelephora* ja *Tomentella* analüüs

Töös kasutatud perekondadesse *Thelephora* ja *Tomentella* kuuluvad 694 ITS järjestust ning üks *Pseudotomentella* sekvents jagunesid 'pairwise' distantide 3 %-list erinevust rakendades 247 liigiks, millest 140 põhinevad ainult ühel järjestusel. Seetõttu on võimalik, et osa neist liikidest tuleb tulevikus ühendada, kuna suurema järjestuste arvuga liikidel ületas liigisisene varieeruvus tihti 5% piiri. Nimelt kuuluvad kaks ITS järjestust, mis erinevad omavahel 3,5 % ja millel puudub 'ühendav' sekvents antud süsteemi raames eri liikidesse. Uute andmete, so järjestuste lisamisel võib aga leiduda neid 'ühendav' järjestust, mis olemasoleva enam kui 3% vahemiku kaotab ning seetõttu liidab kõik kolm järjestust üheks liigiks. Joonis 7. põhjal saab oletada, et enamikus üksikute järjestuste baasil eristatud liike on siiski "head" liigid, kuna nendevaheline erinevus on valdavalt üle 5%. Seda toetab ka fakt, et suur osa üksikute järjestuste põhjal eristatud liike pärinevad Põhja- ja Lõuna-Ameerikast ning Austraaliast. Kuna nendest regioonidest on praeguses andmemaatriksis suhteliselt vähem andmeid, siis on ka tõenäoline, et iga uus lisatav ITS järjestus võib esindada uut liiki. Tulemustes on juba mainitud, et selle seisukoha toetuseks tuleb põhjalikumalt uurida eelkõige sõsarliikide vahelisi distantse. Kui need jäävad valdavalt 5% lähistele, on tõenäosus nende ühendamiseks tulevikus samasse liiki märksa suurem kui Joonis 7 seda eeldaks.

Kõige suurema arvu sekventsidega on selles töös esindatud kaks morfoliiki, so *Thelephora terrestris* ja *Tomentella sublilacina*³. Selle põhjuseks on esiteks see, et tegemist on vähemalt põhjapoolkeral väga sageli esinevate liikidega ning teiseks see, et prof. U. Kõljala töörühm tunneb nende ja lähedaste taksonite ökoloogia ja biogeograafia vastu suurt huvi. Seetõttu on nende taksonite esindajaid ka valikuliselt rohkem sekveneeritud. *Thelephora terrestris* ja *Tomentella sublilacina*³ ühendamise 3% lävemeetodi põhjal võib olla tingitud mitmest põhjusest. Esiteks sekventside suur arv, mis pärinevad erinevatelt kontinentidelt. Seetõttu on suurem võimalus, et on leitud eksemplar, mis tunnuste poolest on sarnane mõlemale liigile. Teiseks on võimalik, et need kaks sõsarliiki on lahknenu suhteliselt hiljuti ning kolmandaks on võimalik, et mingil põhjusel on mutatsioonide teke ITS regioonides nende liikide puhul olnud aeglasem kui enamikul *Thelephora* ja *Tomentella* liikidel. Tõenäoline on toodud põhjuste koosinemine. Nende kahe liigi baasil saab püstitada hüpoteesi, et järjestuste arvu kasvades tuleb liigipiiriks olevat erinevust vähendada, kuna on suurem tõenäosus vahevormide leidmiseks. On väga oluline märkida, et kahe liigi, so *Thelephora terrestris* ja *Tomentella sublilacina*³ näide annab suure toetuse ideele eristada seeneliike DNA triipkoodi meetodil. Vaatamata suurele hulgale ITS sekventsidele mitmelt kontinendilt on kahe sõsarliigi vaheline nn. gap tervelt 2,6% ja seda tingimustes kus liigisisene varieeruvus ületab 5% piiri.

Tehtud analüüsid näitavad ka seda, et tõenäoliselt ei ole seeneliikide eristamine ühtse

protsendi alusel otstarbekas. Näiteks võisid sama perekonna liigid lahkneha, so tekkida eri aegadel ning mutatsioonide teke ITS järjestustes võib eri liikidel (indiviididel) kulgeda erineva tempoga. Seetõttu on tõenäoliselt otstarbekas luua eri liikidele või liigirühmadele nende eristamiseks erinev sarnasuse/erinevuse koefitsent. Seene viljakeha morfoloogiliste ja anatoomiliste tunnuste põhjal tundub 3% liigi eristamise lävena pigem morfoliike kokku panema kui lahutama (U. Kõljalg, suulised andmed).

Info peremeestaime kohta oli olemas 53%, so 367-l eksemplaril. See on liiga väike arv, et peremehe eelistuse kohta põhjalikumaid järeldusi teha. Suhe okaspuu- ja katteseemnetaimede vahel oli 1/5-le ning nad paiknesid fülogeneesipuul enamasti läbisegi. Siiski võis eristada ka väiksemaid klaade (7-8 järjestust) nt. *Tomentella stuposa*² ja *T. sublilacina*³ hulgas, millel peremeheks vaid okaspuu, kusjuures eksemplarid pärinesid rohkem kui ühelt kontinendilt. Saadud andmestiku põhjal saab oletada, et eristatud seeneliikidel puudub kitsam spetsialiseerumine ühele peremeestaime taksonile. Nagu öeldud võib aga täheldada liigisisest grupeerumist vastavalt peremeestaime taksonile. Sama fenomeni on täheldatud ka lähedases perekonnas *Tomentellopsis* kus liik *T. submollis* jagunes ITS järjestuste põhjal kolmeks klaadiks, millest üks esines ainult koos katteseemne- ja teine koos paljasseemnetaimedega ning kolmandasse klaadi kuuluvad eksemplarid omasid peremeestaime mõlemast taksonist (Kõljalg jt. 2002).

Joonisel 6. toodud samalt ja eri kontinendilt pärit ITS järjestuste liigisisese varieeruvuse erinevused on tõenäoliselt tingitud geograafilisest isolatsioonist ning näitavad hiljutise geenitriivi puudumist. Arvestades järjestuste suhteliselt väikest arvu (695 järjestust 247 liigi kohta) tuleb sellesse tulemusse siiski kriitiliselt suhtuda. Selliste järelduste tegemiseks on vaja laiaulatuslikke uuringuid, seda eriti lõunapoolkeral, kust on siiani küllaltki vähe andmeid, so ITS järjestusi. Tabel 1 sisaldab ITS järjestuste geograafilist päritolu, mis näitab, et eriti puudulikud on andmed troopilisest ja parasvöötme Aasiast (vastavalt 2 ja 10 järjestust), Australaasiast (28 järjestust) ning Lõuna-Ameerikast (33 järjestust). Aafrikast puuduvad andmed aga täielikult. Samas on kõigist neist regioonidest teada ektomükoriisat moodustavate peremeestaime väga suur liigirohkus - troopilises Aasias sugukonnad Dipterocarpaceae ja Fagaceae, Australaasias Myrtaceae ja Nothofagaceae, Lõuna-Ameerikas Nothofagaceae, Nyctaginaceae ja Dipterocarpaceae, Aafrikas Caesalpinioideae ja Uapacaceae. Lähtudes olemasolevast esialgselt informatsioonist (U. Kõljalg, suulised andmed) on tõenäoliselt ka nende peremeestaime ühed tavalisemad juursümbiondid perekondadesse *Thelephora* ja *Tomentella* kuuluvad liigid.

Analüüside tulemusena saadud liikide arvu põhjal ning eelmises paragrahvis mainitud andmete puudulikkust arvestades on spekulatiivne tuletada ligikaudne hetkel eksisteerivate *Thelephora* ja *Tomentella* liikide arv. Selline arvutus väljus antud töö raamest, kuid esialgselt võib ennustada, et liikide arv jääb ka tagasihoidlike arvutuste kohaselt 500 ja 1000 vahele.

4.2 Kasutatud metoodika analüüs

Töös kasutatud metoodika (andmete säilitamine andmebaasis, aligneerimine *mafft* programmiga, perli skriptide loomine) ja läbiviidud analüüsid ('pairwise' distantide põhjal gruppide arvutamine, distantide sõltuvus eksemplaride leiukohast) on enamalt jaolt valitud suurt andmete hulka silmas pidades. Siinkohal tooksin välja mõned probleemid, millega andmete kogumisel ja töötlemisel kokku tuli puutuda: 1) geenipangast saadud info ja metaandmed sekventsides kohti olid tihti puudulikud, mistõttu tuli vajalik informatsioon leida artiklitest, milles sekventsid olid avaldatud. H. Nilsson Göteborgi Ülikoolist (suulised andmed) on näidanud, et üle 99% geenipankades olevatest seente ITS järjestustest ei ole kunagi uuendatud, 40% sekventsides ei ole seotud ühegi publikatsiooniga ning 6%-l ülejäänutest on artikli puhul märgitud *submitted* või *in press*, millest 56% on üle 4 aastat vanad (mõned isegi kuni 14 aastat). Seetõttu tuli eraldiseisvalt üles otsida vähemalt pooled neist artiklitest, mida andmete kirjeldamiseks tarvis läks. Mitmel juhul polnud peremeestaime või isegi sekventsides geenipanga numbreid kirjas ka artiklis endas. Käesolevas töös sai annotatsiooni puudulikkus tõsiseks takistuseks näiteks peremeestaime eelistuse uurimisel. 2) Kasutatud andmestik oli mittetäielik, s.t. paljud liigid olid esindatud vaid ühe-kahe eksemplariga, mis suure tõenäosusega korjatud kas Euroopast või Põhja-Ameerikast. Seetõttu on võimalik, et töös esitatud graafikud ei pruugi kirjeldada reaalset olukorda (nt. joonis 6, mis kirjeldab 'pairwise' distantide ja leiukoha vahelist sõltuvust). Siiski oli ka hästi esindatud liike, nagu *Tomentella stuposa*, *T. sublilacina* ja *Thelephora terrestris*, mida kasutati näidisrühmadena liigisisese varieeruvuse ja liikidevahelise erinevuse demonstreerimiseks. 3) Gruppide moodustamise skripti puhul on oluline teada, et 3%-line maksimaalne erinevus ei näita seda protsenti, mis on absoluutseks piiriks teatud gruppi kuulumisel, s.t. grupisisene varieeruvus ei pea olema alla 3%, küll aga peavad gruppi kuuluvad järjestused olema üksteisega ühendatud nn. 3 % lingi kaudu. Oli olemas ka teine võimalus gruppide moodustamiseks – nimelt ühendada omavahel need sekventsides, mille erinevus kõigi teistega jääb alla 3 %, ning mis annaks väljundiks väiksemad grupid mõnede liikmete kordumisega. Sel juhul tuleks gruppide esitamiseks kasutada võrgustikke ning lõpptulemus oleks sama mis käesolevas töös – distantide erinevused ja liikmete arv grupis ei muutuks. Selline võrgustike kasutamine võiks olla järgmiseks eesmärgiks, annaks see ju võimaluse täpsemalt sekventsides omavahelisi suhteid vaadelda, samuti kaasata analüüsi metaandmed. Fülogeneesipuid võrgustikena esitavaid programme on loodud juba eelmise sajandi 90ndate algusest, hea ülevaate neist annavad Posada (2001) ja Morrison (2005). Põhiliseks probleemiks praeguste võrgustike joonistamise programmide puhul (*TCS* (Clement jt. 2000), *Network* (Bandelt jt. 1999), *SplitsTree4* (Huson 1998)) on see, et nad on enamasti mõeldud populatsiooni tasemel evolutsiooniliste sündmuste nagu hübriidsatsioon, rekombinatsioon, lateraalne geeniülekanne uurimiseks, ning võimalike tunnuste konfliktide analüüsimiseks. Nimetatud programmide rakendamine liigipiiride

leidmiseks ei ole tulemusrikas, mida kinnitas nende katsetamine kahe sõsarliigi vaheliste seoste leidmiseks.

Võrgustike asemel on antud töös kasutatud hiljuti avaldatud programmi *Mesquite*, mille eelistena võiks välja tuua tema modulaarsuse ja käsitlemise lihtsuse. Modulaarsus tähendab, et programm on üles ehitatud erinevateks analüüsideks erinevaid mooduleid kasutades, neid võivad kasutajad ise juurde kirjutada ja sel viisil vastavalt vajadusele programmi täiendada. Käesolevas töös uut moodulit ei kirjutatud, selle asemel modifitseeriti sisendiks olevat *nexus* formaadis faili lisades sellele kaks plokki, mis võimaldasid eri gruppidesse kuuluvaid järjestusi fülogeneesipuul kindla värvitooniga näidata. Tulevikueesmärgiks oleks võimalikult paljud käesolevas töös kasutatud meetodika etapid (aligneerimine, distantside maatriksi arvutamine, metaandmete analüüs, gruppide loomine, sõsarliikide vaheliste erinevuste graafikute joonistamine) automatiseerida, seejuures tuleks kindlasti läbi mõelda, kui otstarbekas oleks *Mesquite* programmile mooduli(te) kirjutamine.

Kokkuvõte

Töö tutvustab DNA triipkoodi mõistet ja selle kasutamise võimalikkust seeneliikide määramiseks. Töö esimeseks eesmärgiks oli välja töötada bioinformaatiline süsteem leidmaks seltsi Thelephorales (lehternahkiselaadsed) kahte perekonda, *Tomentella* ja *Thelephora* kuuluvate liikide ITS (*internal transcribed spacer*) nukleotiidsete järjestuste põhjal nende sarnasuse/erinevuse koefitsient, mida võiks kasutada tundmatute eksemplaride määramiseks kindlasse gruppi või liiki. Töö teiseks eesmärgiks oli koguda ja võimaluse korral analüüsida eksemplaride metaandmeid, nagu nende geograafiline leiukoht ja peremeestaime eelistus. Selleks kasutati 695 avalikest andmebaasidest (NCBI, EMBL, DDBJ) ja prof. U. Kõljala isiklikust kogust pärinevat nimetatud seenerühma ITS järjestust. Leiti, et seeneliikide eristamine DNA triipkoodi meetodil on teatud määral võimalik. Hoiduda tuleks siiski kindlaksmääratud läviväärtuste kasutamisest, kuna erinevad sõsarliigid eristuvad üksteisest erineva sarnasuse/erinevuse koefitsiendi juures. Leiti ka, et seeneliikide liigisisene varieeruvus on tihti tunduvalt suurem (ulatudes kuni 5,2%-ni) kui on sõsarliikide vaheline erinevus (mõnedel juhtudel alla 3%). Geograafiliselt üksteisest kaugemal (eri kontinentidel) olevate eksemplaride vahel oli keskmiselt suurem järjestuste erinevus. Kokku saadi 695 järjestuse baasil 247 liiki, millest 140 olid esindatud vaid ühe eksemplariga.

Summary

This study introduces the basic idea of DNA barcoding and its possible implementation to identify fungal species. The aims of this study were: 1) to evolve a bioinformatical system for finding similarity/dissimilarity coefficient based on ITS (*internal transcribed spacer*) nucleotide sequences of fungal species belonging to genera *Tomentella* and *Thelephora* which could then be used to identify unknown specimens to certain group or species; 2) to collect and analyse metadata of these specimens, for example their collecting area and host specificity. For this purpose 695 ITS sequences of *Tomentella* and *Thelephora* were analysed originating from public DNA databases (NCBI, EMBL, DDBJ) and from private collection of prof. U. Kõljalg. It is possible to a certain extent to identify fungal species using DNA barcoding method. One should beware of using threshold values though because different sister species differentiate from each other on varying similarity/dissimilarity coefficient. In many cases the intraspecific variability exceeds interspecific difference and geographically distant specimens have in average higher dissimilarity percent than those collected from the same continent. Based on 695 sequences 247 species were differentiated, 140 of them represented by only 1 specimen.

Tänuõnad

Tahan tänada oma juhendajat prof. Urmas Kõljalga suure abi ja nõuannete eest töö tegemisel. Samuti avaldan tänu kaasjuhendaja prof. Maidu Remmile, Henrik Nilssonile Göteborgi Ülikoolist statistiliste andmete eest ning kõigile mükoloogia õppetooli inimestele, kes töö ülevaatamisel ja illustreerimisel kaasa aitasid.

Kasutatud kirjandus

- Agerer R (1996) Ectomycorrhizae of *Tomentella albomarginata* (Thelephoraceae) on Scots pine. *Mycorrhiza*, **6**, 1-7.
- Baar J, Horton TR, Kretzer AM, Bruns TD (1999) Mycorrhizal colonisation of *Pinus muricata* from resistant propagules after a stand replacing wildfire. *New Phytologist*, **143**, 409-418.
- Bandelt H-J, Forster P, Röhl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37-48.
- Clement M, Posada D, Crandall K (2000) TCS: a computer program to estimate gene genealogies. *Molecular Ecology*, **9** (10), 1657-1660.
- Corner EJH (1968) A monograph of *Thelephora* (Basidiomycetes). *Beihefte zur Nova Hedwigia*, **27**, 1-110.
- Danielson RM, Zak JC, Parkinson D (1984) Mycorrhizal inoculum in a peat deposit formed under a white spruce stand in Alberta. *Canadian Journal of Botany*, **63**, 2557-2560.
- Danielson RM, Pruden M (1989) The ectomycorrhizal status of urban spruce. *Mycologia*, **81**, 335-341.
- Druzhinina IS, Kopchinskiy AG, Komoń M, Bissett J, Szakacs G, Kubicek CP (2005) An oligonucleotide barcode for species identification in *Trichoderma* and *Hypocrea*. *Fungal Genetics and Biology*, **42**, 813-828.
- Galtier N, Gouy M, Gautier C (1996) SeaView and Phylo_win, two graphic tools for sequence alignment and molecular phylogeny. *Computer Applications in the Biosciences*, **12**, 543-548.
- Gardes M, Bruns TD (1996) Community structure of ectomycorrhizal fungi in a *Pinus muricata* forest: above- and below-ground views. *Canadian Journal of Botany*, **74**, 1572-1583.
- Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *The Royal Society*, **270**, 313-321.

Hebert PDN, Stoeckle MY, Zemplak TS, Francis CM (2004) Identification of birds through DNA Barcodes. *PloS Biology*, **2** (10), 1657-1663.

Hollis S, Brummitt RK (1992) World geographical scheme for recording plant distributions. Plant Taxonomic Database Standards No. 2, Hunt Institute for Botanical Documentation, Pittsburgh. (Current electronic version) TDWG Standard.

Horton T, Bruns TD (1998) Multiple-host fungi are the most frequent and abundant ectomycorrhizal types in a mixed stand of Douglas fir (*Pseudotsuga menziesii*) and bishop pine (*Pinus muricata*). *New Phytologist*, **139**, 331-339.

Huson D (1998) SplitsTree: a program for analyzing and visualizing evolutionary data. *Bioinformatics*, **14**, 68-73.

Katoh K, Misawa K, Kuma K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research*, **30**, 3059-3066.

Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences*, **102** (23), 8369-8374.

Kõljalg U (1992) Mycorrhiza formed by basidiospores of *Tomentella crinalis* on *Pinus sylvestris*. *Mycological Research*, **96**, 215-220.

Kõljalg U (1996) *Tomentella* (Basidiomycota) and related genera in Temperate Eurasia. Synopsis Fungorum 9. Fungiflora, Oslo.

Kõljalg U, Dahlberg A, Taylor AFS, Larsson E, Hallneberg N, Stenlid J, Larsson K-H, Fransson PM, Karen O, Johnsson L (2000) Diversity and abundance of resupinate theleporoid fungi as ectomycorrhizal symbionts in Swedish boreal forests. *Molecular Ecology*, **9**, 1985-1996.

Kõljalg U, Jakucs E, Bóka K, Agerer R (2001) Three ectomycorrhizae with cystidia formed by different *Tomentella* species as revealed by rDNA ITS sequences and anatomical characteristics. *Folia Cryptogamica Estonica*, **38**, 27-39.

Kõljalg U, Tammi H, Timonen S, Agerer R, Sen R (2002) ITS rDNA sequence-based positioning of Pink-type ectomycorrhizas and *Tomentellopsis* species from boreal and temperate forests *Mycological Progress*, **1**, 81-92.

Kõljalg U, Larsson K-H, Abarenkov K, Nilsson RH, Alexander IJ, Eberhardt U, Erland S, Hoiland K, Kjoller R, Larsson E, Pennanen T, Sen R, Taylor AFS, Tedersoo L, Vralstad T, Ursing BM (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *New Phytologist*, **166**, 1063-1068.

Larsen MJ (1968) Tomentelloid fungi of North America. Technical Publication 93, State University College of Forestry at Syracuse University.

Larsen MJ (1974) A contribution to the taxonomy of the genus *Tomentella*. *Mycologia Memoirs*, **4**, 1-145.

Maddison DR, Maddison WP (2003) MacClade 4: Analysis of phylogeny and character evolution. Version 4.06. Sinauer Associates, Sunderland, Massachusetts.

Maddison WP, Maddison DR (2003) Mesquite: a modular system for evolutionary analysis. Version 1.0 <http://mesquiteproject.org>

Marshall E (2005) Taxonomy – Will DNA bar codes breathe life into classification? *Science*, **307**, 1037.

Marx DH, Brian WC (1970) Pure culture synthesis of ectomycorrhizae by *Thelephora terrestris* and *Pisolithus tinctorius* on different conifer hosts. *Canadian Journal of Botany*, **47**, 639-643.

McKendrik SL, Leake JR, Taylor DL, Read DJ (2000) Symbiotic germination and development of myco-heterotrophic plants in nature: ontogeny of *Corallorhiza trifida* and characterisation of its mycorrhizal fungi. *New Phytologist*, **145**, 523-537.

Meyer CP, Paulay G (2005) DNA barcoding: Error rates based on comprehensive sampling. *PloS Biology*, **3** (12), 2229-2238.

Moritz C, Cicero C (2004) DNA Barcoding: Promise and Pitfalls. *PloS Biology*, **2** (10), 1529-1531.

- Morrison DA (2005) Networks in phylogenetic analysis: new tools for population biology. *International Journal for Parasitology*, **35**, 567-582.
- O'Brien HE, Parrent JL, Jackson JA, Moncalvo JM, Vilgalys P (2005) Fungal community analysis by large-scale sequencing of environmental samples. *Applied and Environmental Microbiology*, **71** (9), 5544-5550.
- Posada D, Crandall KA (1998) Modeltest: testing the model of DNA substitution. *Bioinformatics*, **14** (9), 817-818.
- Posada D, Buckley TR (2004) Model selection and model averaging in phylogenetics: advantages of the AIC and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, **53**, 793-808.
- Pritsch K, Munch J-C, Buscot F (2000) Identification and differentiation of mycorrhizal isolates of black alder by sequence analysis of the ITS region. *Mycorrhiza*, **10**, 87-93.
- Rice P, Longden I, Bleasby A (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, **16** (6), 276-277.
- Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaediae of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society*, **360**, 1805-1811.
- Stalpers JA (1993) The Aphyllphoraceous fungi I. Keys to the species of the Thelephorales. *Studies in Mycology*, **35**, 1-168.
- Summerbell RC, Lévesque CA, Seifert KA, Bovers M, Fell JW, Diaz MR, Boekhout T, de Hoog GS, Stalpers J, Crous PW (2005) Microcoding: the second step in DNA barcoding. *Philosophical Transactions of the Royal Society*, **360**, 1897-1903.
- Swofford DL (2003) PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. Sinauer Associates, Sunderland, Massachusetts.
- Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy.

Trends in Ecology and Evolution, **18**, 70-74.

Taylor DL, Bruns TD (1997) Independent, specialized invasions of ectomycorrhizal mutualism by two nonphotosynthetic orchids. *Proceedings of the National Academy of Sciences of the USA*, **94**, 4510-4515.

Tedersoo L, Hallenberg N, Larsson KH, Kõljalg U (2003) Fine scale distribution of ectomycorrhizal fungi and roots across substrate layers including coarse woody debris in a mixed forest. *New Phytologist*, **159**, 153-165.

Tedersoo L, Suvi T, Larsson E, Kõljalg U (2006) Diversity and community structure of ectomycorrhizal fungi in a wooded meadow. *Mycological Research*, in press.

Vilgalys R (2003) Taxonomic misidentification in public DNA databases. *New Phytologist*, **160**, 4-5.

Weir JR (1921) *Thelephora terrestris*, *T. fimbriata* and *T. caryophyllea* on forest tree seedlings. *Phytopathology*, **11**, 141-144.

Andmebaasid:

National Center for Biotechnology Information (NCBI). [<http://www.ncbi.nlm.nih.gov/>]

EMBL Nucleotide Sequence Database (EMBL). [<http://www.ebi.ac.uk/embl/>]

DNA Data Bank of Japan (DDBJ). [<http://www.ddbj.nig.ac.jp/>]

UNITE. [<http://unite.ut.ee/>]

LISAD

Lisa 1. Andmestiku kogumiseks kasutatud andmebaasi tabelite struktuur.

tabel 'MAIN' (geenipangast pärinevad järjestused)	
id	automaatselt suurenev unikaalne id
seq_id	sekventsi unikaalne kood analüüside jaoks
acc_number	sekventsi kood geenipangas
species	liiginimi, kui on määratud
sequence	DNA järjestus
literature_id	viitava artikli kood (tabel 'literature')
DNA_source	DNA päritolu kood (tabel 'dna_source')
Host_name	peremeestaime nimi, kui olemas
Host_id	peremeestaime kood (tabel 'host')
locality	leiukoht, kui on olemas
L1_code	leiukoha L1 kood (tabel 'tblLevel1')
remarks	märkused sekventsi kohta
MAIN_2_id	tabeli MAIN_2 seq_id, kui on tegemist topeltkirjega
date_modified	viimase muutmise kuupäev

tabel 'MAIN_2' (U. Kõljalalt saadud järjestused)	
id	automaatselt suurenev unikaalne id
seq_id	sekventsi unikaalne kood analüüside jaoks
acc_number	sekventsi kood autori failis
locality	leiukoht, kui on olemas
L1_code	leiukoha L1 kood (tabel 'tblLevel1')
species	liiginimi, kui on määratud
sequence	DNA järjestus
literature_id	viitava artikli kood (tabel 'literature')
DNA_source	DNA päritolu kood (tabel 'dna_source')
Host_name	peremeestaime nimi, kui olemas
Host_id	peremeestaime kood (tabel 'host')
remarks	märkused sekventsi kohta
date_modified	viimase muutmise kuupäev

tabel 'dna_source' (DNA päritolu)	
id	unikaalne kood
Source_description	päritolu kirjeldus

tabel 'host' (peremeestaimed)	
id	unikaalne kood
host_name	peremeestaime nimi
classification	organismi taksonoomiline kuulumus geenipanga 'taxonomy browser' järgi
class	klassi kuulumus
subclass	alamklassi kuulumus
order	seltsi kuulumus
family	sugukonda kuulumus

tabel 'literature' (kirjandus)	
id	unikaalne kood
article	artikli pealkiri, autorid, ilmumisandmed
status	kas olemas pdf formaadis

tabel 'tblLevel1' (TDWG standardi järgi)	
L1_code	unikaalne kood
L1_continent	L1 kontinendi nimi

Lisa 2. Perli skript 'dist_final.pl', mis teeb etteantud 'pairwise' distantside maatriksi põhjal distantside sagedustabeli.

```
#####  
# skript: "pairwise" distantside sagedustabeli tegemiseks #  
# skripti jooksutamine: perl dist_final.pl "sisend_fail" "piir%" > "väljund_fail" #  
# autor: Kessy Abarenkov #  
# sisend: "pairwise" distantside maatriks, käsurealt #  
# väljund: "pairwise" distantside sagedustabel: vahemiku ülemine piir t\ distantside sagedus #  
# viimati muudetud: 11.05.2006 #  
#####
```

```
# !/usr/bin/perl -w
```

```
use strict;
```

```
##### muutujate defineerimine:
```

```
my $infile = $ARGV[0];
```

```
my $tab_nr = $ARGV[1];
```

```
my $dist_file = "dists_out.txt";
```

```
my @matrix = ();
```

```
my @dists = ();
```

```
my @maatriks = ();
```

```
my $arv;
```

```
my $sekv_arv;
```

```
open (FILE, $infile);
```

```
open (DISTFILE, ">$dist_file");
```

```
@matrix = <FILE>;
```

```
##### alates sisendfaili 8ndast reast kirjutab kõik distantseid vahefaili "dists_out.txt":
```

```
for (my $k=8; $k<scalar@matrix; $k++) {
```

```
    @dists = split(" ", $matrix[$k]);
```

```
    for (my $i=0; $i<(scalar@dists-2); $i++) {
```

```
        print DISTFILE "$dists[$i]\n";
```

```
    }
```

```
}
```

```

close DISTFILE;
close FILE;
$stab_nr = 5*$stab_nr;
$sekv_arv = (scalar@matrix)-8;

##### täidab maatriksi 0 väärtustega:
for (my $i=0; $i<$stab_nr; $i++) {
    $maatriks[$i] = 0;
}
open (FILE2, $dist_file);

##### võtab vahefailist distantsid ja suurendab maatriksit/jaotustabelit vastavalt distantsi
##### väärtusele:
while (<FILE2>) {
    unless ($_ eq "\n") {
        for (my $i=0; $i<$stab_nr; $i++) {
            if (($_ >= $i*0.2) && ($_ < ($i+1)*0.2)) {
                $maatriks[$i]++;
            }
        }
    }
}

##### prindib väljundisse sagedustabeli:
for (my $k=0; $k < $stab_nr; $k++) {
    if ($k=='0') {
        $sarv = ($k+1)*0.2;
        print "$sarv\t" . ($maatriks[$k]-$sekv_arv) . "\n";
    } else {
        $sarv = ($k+1)*0.2;
        print "$sarv\t$maatriks[$k]\n";
    }
}
close FILE2;
unlink($dist_file);
exit;

```


Lisa 3. Perli skript “3_percent_final.pl”, mis grupeerib järjestused etteantud 'pairwise' distantide maatriksi põhjal andes väljundiks kaks *nexus* formaadis failile lisatavat ploki *Mesquite* programmis kasutamiseks.

```
#####  
# skript: Mesquit programmi jaoks nexus faili gruppide moodustamise ploki loomiseks      #  
# skripti jooksutamine: perl 3_percent_final.pl "sisend_fail" > "väljund_fail"          #  
# autor: Kessy Abarenkov                                                                #  
# sisend: "pairwise" distantide maatriks, käsurealt                                    #  
# väljund: nexus failile lisatav plokk                                                #  
# viimati muudetud: 12.05.2006                                                         #  
#####
```

```
#!/usr/bin/perl -w
```

```
use strict;
```

```
##### muutujate defineerimine:
```

```
my $infile = $ARGV[0];
```

```
my @maatriks;
```

```
my ($name, $num);
```

```
my %names = ();
```

```
my @matrix = ();
```

```
my @dists = ();
```

```
my $group_file = "group_file.txt";
```

```
my @groups = ();
```

```
my @groups3 = ();
```

```
my $row;
```

```
my @pairs = ();
```

```
my @pairs2 = ();
```

```
my @rows = ();
```

```
my $scl;
```

```
my @arr = ();
```

```
my @arr2 = ();
```

```
my @arr4 = ();
```

```
my ($count, $count3);
```

```
my %seen = ();
```

```
my @uniq = ();
```

```
open (FILE, $infile);
```

```
@matrix = <FILE>;
```

```
##### maatriksi täitmine:
```

```
for (my $k=7; $k<scalar@matrix; $k++) {
```

```
    my @dists = split("\t", $matrix[$k]);
```

```
    for (my $i=0; $i<scalar@dists; $i++) {
```

```
        chomp $dists[$i];
```

```
        $maatriks[$k-7][$i] = $dists[$i];
```

```
    }
```

```
}
```

```
##### sekvenside nimedega paistabeli loomine:
```

```
for (my $k=1; $k<scalar@maatriks; $k++) {
```

```
    ($name, $num) = split(' ', $maatriks[$k][(scalar@maatriks)+1]);
```

```
    $names{ $num } = $name;
```

```
    $maatriks[$k][(scalar@maatriks)+1] = $num;
```

```
}
```

```
##### prindib vahefaili "group_file.txt" sekvenside paarid, mille omavaheline sarnasus on
```

```
##### väiksem kui 3%:
```

```
open (GROUP_FILE, ">$group_file");
```

```
for (my $k=1; $k<scalar@maatriks; $k++) {
```

```
    for (my $i=0; $i<(scalar@maatriks); $i++) {
```

```
        if ($maatriks[$k][$i] eq "") { # kui maatriksis on tühi element:
```

```
            print "";
```

```
        } elsif ($maatriks[$k][$i] <= '3.00') { # kui sarnasus on väiksem/võrdne kui 3%:
```

```
            if ($maatriks[0][$i] == $maatriks[$k][(scalar@maatriks)+1]) { # kui tegemist on võrdlusega  
                # iseendaga:
```

```
                print "";
```

```
            } else {
```

```
                print GROUP_FILE "$maatriks[0][$i] : $maatriks[$k][(scalar@maatriks)+1]
```

```
$maatriks[$k][$i]\n";
    }
} else { print ""; }
}
}
close GROUP_FILE;
```

tekitab sekventsidi grupid:

```
open (GR_FILE, 'group_file.txt');
```

```
while (<GR_FILE>) {
    @pairs = split(' ', $_);
    $_ = '!' . join(':', @arr) . '!';
    if (/:$pairs[2]:/) {
        print "";
    } else {
        push (@arr, $pairs[2]);
    }
}
close GR_FILE;
```

deklareerib maatriksis gruppide esimesed elemendid:

```
for (my $i=0; $i<scalar@arr; $i++) {
    $groups[($arr[$i])[0] = $arr[$i];
}
}
```

käib vahefaili uuesti läbi teiste elementide saamiseks:

```
open (GR_FILE, 'group_file.txt');
while (<GR_FILE>) {
    @pairs2 = split(' ', $_);
    for $row (@groups) {
        if (defined(@$row)) {
            @rows = @$row;
            $_ = '!' . join(':', @rows) . '!';
        }
    }
}
```

```

if ((/:$pairs2[0]:/) && (/:$pairs2[2]:/)) { # kui mõlemad paari liikmed on maatriksis,
    print ""; # ei tee midagi:
} elsif ((/:$pairs2[0]:/) || (/:$pairs2[2]:/)) { # kui üks paari liikmetest on maatriksis olemas,
    if (/:$pairs2[0]:/) {
        push(@$row, $pairs2[2]); # lisab ka teise:
    } else {
        push (@$row, $pairs2[0]);
    }
} else { # kui kumbagi elementi maatriksis ei ole,
    print ""; # ei tee midagi:
}
}
}
}
}
close GR_FILE;

```

sorteerib maatriksi:

```

for $row ( @groups ) {
    if (defined(@$row)) {
        @$row = sort {$a <=> $b} @$row;
    } else {
        print "";
    }
}
}

```

moodustab uue maatriksi, et kaotada defineerimata read:

```

$count=0;
for $row ( @groups ) {
    if (defined(@$row)) {
        @arr2 = @$row;
        push(@{$groups3[$count]}, @arr2);
        $count++;
    }
}
}

```

võrdleb gruppe omavahel, ühendab nad, kui nad sisaldavad ühiseid elemente:

```
for my $i ( 0 .. $#groups3-1) {
  for my $k ( $i+1 .. $#groups3) {
    if (compare_unite_arrays(\@{$groups3[$i]}, \@{$groups3[$k]}) == 1) {
      @arr4 = @{$groups3[$k]};
      push (@{$groups3[$i]}, @arr4);
      undef(@{$groups3[$k]});
    } else {
      print "";
    }
  }
}
```

eemaldab maatriksist korduvad elemendid:

```
for (my $i=0; $i<scalar@groups3; $i++) {
  %seen=();
  foreach my $item(@{$groups3[$i]}) {
    push (@{$uniq[$i]}, $item) unless $seen{$item}++;
  }
}
```

defineerib nexus faili ploki jaoks värvide RGB koodid:

```
my @color_codes = ('1.0 0.0 0.0', '0.0 1.0 0.0', '1.0 0.81960784 0.70196078', '0.96862745
0.835294150 1.0', '0.70196078 0.76078431 1.0', '0.0 0.658823551 0.0', '0.921568651 1.0 0.8', '0.0
0.0 0.658823551', '0.921568651 0.8 1.0', '1.0 0.658823551 0.0', '1.0 0.32941176 0.0', '1.0
0.894117653 0.866666655', '0.0 1.0 0.32941176', '0.0 0.658823551 1.0', '0.0 0.32941176 1.0', '1.0
0.0 0.658823551', '1.0 0.30196078 0.580392154', '1.0 0.658823551 0.658823551', '1.0 0.56862745
0.56862745', '0.658823551 1.0 0.658823551', '0.572549050 1.0 0.466666655', '0.866666655
0.866666655 1.0', '0.32941176 0.32941176 1.0', '0.658823551 0.658823551 0.0', '1.0 0.90196078
0.835294150', '0.658823551 0.0 0.658823551', '0.96078431 0.90196078 1.0', '0.0 0.658823551
0.658823551', '0.81960784 0.70196078 1.0', '1.0 0.658823551 0.32941176', '1.0 0.32941176
0.658823551', '1.0 0.32941176 0.0', '1.0 0.0 0.32941176', '0.658823551 1.0 0.32941176',
'0.32941176 1.0 0.0', '0.658823551 0.32941176 1.0', '0.32941176 0.0 1.0', '1.0 0.8 0.4', '1.0 0.4 0.8',
'1.0 0.70196078 0.2', '1.0 0.2 0.70196078', '0.8 1.0 0.4', '0.4 1.0 0.8', '0.70196078 1.0 0.2', '0.2 1.0
0.70196078', '0.8 0.4 1.0', '0.4 0.8 1.0', '0.2 0.70196078 1.0', '0.70196078 0.2 1.0', '0.73333333 1.0
0.666666655', '0.0 0.2 0.70196078', '0.82745098 0.56862745 1.0', '0.70196078 0.0 0.2', '1.0
```



```
    }  
}  
print "\n\nEND;\n";  
print "\nBEGIN SETS;\n";  
print "TAXPARTITION * UNTITLED = ";
```

```
$counter2 = 1;  
for $row ( @uniq ) {  
    if (defined ( @$row )) {  
        @$row = sort { $a <=> $b } @$row;  
        $tax = $tax . "$counter2 : @$row, ";  
        $counter2++;  
    } else {  
        print "";  
    }  
}
```

```
$tax = substr($tax, 0, -2);  
print "$tax;\n";  
print "\nEND;\n\n";  
close FILE;  
unlink("$group_file");  
exit;
```

```
##### subroutines #####
```

```
sub compare_unite_arrays {  
    my ($first, $second) = @_;  
    for (my $i = 0; $i < @$first; $i++) {  
        for (my $k=0; $k<@$second; $k++) {  
            return 1 if $first->[$i] eq $second->[$k];  
        }  
    }  
    return 0;  
}
```