

Mediaanvõrgud fülogeneesi kujutamisel

Mait Metspalu

Inimese mitokondriaalse DNA (mtDNA) ning Y-kromosoomi mitterekombineeruva ala (NRY) järjestuste varieeruvusuuringud on viimase paarikümne aasta jooksul muutunud valdavaks vahendiks inimese demograafilise ajaloo ning fülogeograafia uurimisel. Tavapäraste puukonstrueerimis-meetoditega on mtDNA andmete interpreteerimine raskendatud: Vaid 56 haplotüübiga mtDNA andmestiku puhul ulatub võrdselt parsimoonsete (MP) puude (säätupuude) arv üle miljardi (Excoffier & Smouse 1994). Probleemi põhjustavad põhiliselt kaks tegurit.

- (1) **küllastumine** – asendusi võib ühes positsioonis toimuda korduvalt, kuid vaatleja “näeb” vaid viimase tulemust. Kui viimane asendus on viinud tunnuse tagasi algsesse seisu, jäävad vahepealsed muutused märkamatuks.
- (2) **paralleelsed asendused** – sama asendus samas positsioonis võib realiseeruda paralleelselt erinevates (tegeliku evolutsiooni)puu harudes.

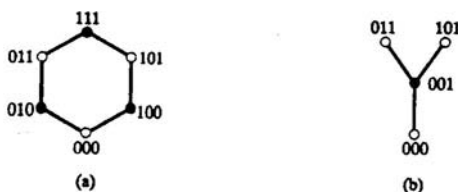
Mõlemad tegurid võimenduvad, kui analüüsitavate informatiivsete positsioonide arv ei ole kuigi suur. Sellest järeldub, et ühe juhuslikult valitud (MP) puu esitamine võib osutuda väga eksitavaks, samas kui konsensuspuu on seda infovaesem, mida suurem on konsensus.

Andmete “ausama” esitamise tarbeks hakati kasutama mediaanvõrke, mis esimeses lähenduses sisaldavad (kujutavad) kõiki antud andmetest tulenevaid kõige parsimoonsemaid MP puid.

Siinkohal on võrkude kasutamine fülogeneesi kujutamisel tingitud andmete iseärasusest, mitte fülogeneesi enda võrgulisest struktuurist. Võrku on siin vaja vaid kui vahendit “õige” puu otsingul – kus retikulatsioonid viitavad paralleelsetele asendustele, küllastunud positsioonidele, aga ka sekveneerimisvigadele.

Mediaanvõrk (median network) MN (Bandelt 1994; Bandelt et al. 1995)

- Andmed viiakse binaarsele kujule
- Andmed minimiseeritakse:
 - i) sarnased haplotüübid koondatakse ning registreeritakse nende sagedus;
 - ii) tunnused, mis on kogu andmestikus samas olekus, kustutatakse;
 - iii) tunnused, mis jagavad andmestiku kaheks grupiks sarnaselt, liidetakse üheks tunnuseks, liites ka nende kaalud.
- Haplotüübid on seega esitatud 0-1 vektoritena, mille pikkus on võrdne tunnuste arvuga.



Joonis 1. (Bandelt et al. 1995). Kaks ühendatud “ühesammu” võrku (*one-step connected networks*), mis on konstrueeritud olemasolevate vektorite 000, 011 ja 101 (avatud ringid) põhjal. (a) vektorite paarid on ühendatud uue vektori lisamise teel; (b) p olemasolevate vektorite tripletile on leitud neid ühendav konsensusvektor e. mediaanvektor, mis viib parsimoonsema vektorite ühendamiseni.

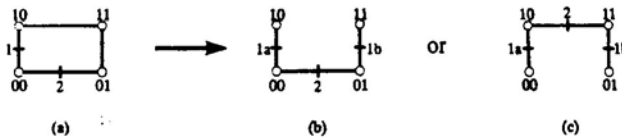
- Mediaanvõrk konstrueeritakse olemasolevate vektorite arvu suurendamise teel, lisades iga vektorite tripleti mediaanvektor (konsensusvektori), seni kuni kõik võimalikud mediaanvektorid on lisatud ja kõik vektorid on ühe sammu kaudu ühendatud (*one-step connected*).

Redutseeritud mediaanvõrk (reduced median network) RMN (Bandelt 1994; Bandelt et al. 1995)

Mediaanvõrgu redutseerimisel asendatakse hüpoteetiliselt paralleelse asenduse läbinud tunnus kahe eraldi tunnusega (*character splitting* –

tunnuse jagamine). Tulemuseks saadakse suurem andmemaatriks, millest samas konstrueeritakse vähem retikulatsioone sisaldav mediaanvõrk.

Tunnuse jagamisel võetakse arvesse tunnusele omistatud kaal ning vastavate haplotüüpide sagedus populatsioonis. Kokkuvõttes saadakse suurem andmemaatriks, mis samas sisaldab vähem homoplaasiaid ning seega on vastavas mediaanvõrgus vähem retikulatsioone. Suuremate andmehulkade puhul on tulemuseks “puulaadsem” mediaanvõrk (kuivõrd väikeste reaalsete andmehulkade puhul RMN ja MN tavaliselt ei erine). RMN-i puhul pole formaalset garantiid, et ta sisaldab kõiki kõige parsimoonsemaid puid, kuid praktikas tundub see siiski nii olevat.



Joonis 2 (Bandelt et al. 1995).

Nelja vektori mediaanvõrk (a) redutseeritakse kui tunnusel 1 on tunnusest 2 2 korda* madalam kaal läbi tunnuse jagamise (1 -> 1a, 1b). Tekkinud kahest fülogeneetilisest rajast (b) ja (c) on eelistatum (b) kui vektorid (haplotüübid) 00 ja 01 on populatsioonis sagedasemad kui vastavalt 10 ja 11.

* RM puhul on parameeter r (*reduction threshold*) vähimisi $r=2$. r määrab kui mitu korda peab fülogeneetiline rada toetavate tunnuste kaalude summa olema suurem rada mitte toetava tunnuse kaalust, et postuleerida paralleelne asendus viimases. Seega, mida väiksem on r , seda redutseeritum ja tõenäoselt ka elukaugem on võrk.

Uusim täiendus mediaanvõrkude algoritmide perre on **ahne mediaanvõrk** (*Greedy median network*) GRM (Bandelt et al. 2000). See on edasiarendus RM-ist, kus on täiendatud eelkõige redutseerimise küsimusi.

Ühendatud mediaanide võrk (*median joining network*) MJ (Bandelt et al. 1999)

Ühendatud mediaanide võrgu konstrueerimisel kombineeritakse Kruskali minimaalkauguse puu algoritmi ja Farrise heuristilist maksimaalse parsimoonisuse algoritmi, kus puule lisatakse mediaanvektoreid ükshaaval.

Erinevalt viimasesest ei lahendata MJ puhul “sõlmi”. Erinevalt MN ja RMN-st on MJ kasutatav ka mitmeseisundiliste tunnuste puhul.

MJ alustab minimaalkauguse puudest ja hakkab vektorite triplettidele lisama mediaanvektoreid (konsensusvektoreid). Et mitte uppuda täis-mediaanvõrku (selle keerukusse suuremate andmemaatriksite korral) valitakse MJ puhul vektorite triplete, mida mediaanvektoriga ühendada. Valitud saamiseks peab üks vektor tripletist olema ühendatud kahe ülejäänud vektoriga.

Kokkuvõte

Võrdselt parsimoonsete puude suur hulk puulaadsete algoritmide kasutamisel on stimuleerinud mediaanvõrkude algoritmide arendamist. Kasutajasõbralik arvutiprogramm 1999. aastast ja selle edasiarendused on soodustanud mediaanvõrkude laialdast kasutamist fülogeneesi rekonstrueerimisel, saades populaarseks inimese mtDNA fülogeneesi kujutamisel. Siiski tuleb nentida, et paljude kasutajate (ka siinkirjutaja) jaoks on selle programmi näol tegu “musta kastiga”, mille puhul andmetöötlus on toimunud algoritmidesse sügavamalt süvenemata.

Bandelt, H.-J. 1994. Phylogenetic networks. *Verh. Naturwiss. Ver. Hamburg* 34:51-71.

Bandelt, H.-J., P. Forster, B. C. Sykes, and M. B. Richards. 1995. Mitochondrial portraits of human populations using median networks. *Genetics* 141:743-53.

Bandelt, H.-J., V. Macaulay, and M. Richards. 2000. Median networks: speedy construction and greedy reduction, one simulation, and two case studies from human mtDNA. *Mol Phylogenet Evol* 16:8-28.

Excoffier, L., and P. E. Smouse. 1994. Using allele frequencies and geographic subdivision to reconstruct gene trees within a species: molecular variance parsimony. *Genetics* 136:343-59.