Minimirekombinaatiohaplotyypitys sukupuutietoa käyttäen

Jaana Heino
Jaana.Heino@cs.helsinki.fi

Tiedon louhinta biomolekyyliaineistoista
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C–2003–52, s. 100–110, marraskuu 2003

Tiivistelmä

Ihmisellä on kutakin kromosomia - ja siten kutakin kohtaa DNA:ssa, esimerkiksi geneettistä markkeria - kaksi kappaletta. Tavalliset tavat tuottaa DNA-markkeridataa yksilöistä antavat tiedon siitä mitkä kaksi alleelia ihmisellä on kussakin markkerissa, mutta eivät tietoa siitä mikä markkeri on peritty yksilön isältä ja mikä äidiltä. Annettuna sukupuu ja genotyyppidataa yksilöistä, haplotyypityksellä tarkoitetaan sen selvittämistä, miltä yksilön sukupuussa olevalta esivanhemmalta mikäkin tämän perimä alleeli on alunperin peräisin.

Minimirekombinaatiohaplotyypityksessä tavoitteena on löytää sellainen haplotyypitys, joka vaatii mahdollisimaan vähän rekombinaatioita eli tietyn yksilön kahden kromosomin sekoittumista keskenään ennen kuin yksilön lapsi perii niistä jommankumman.

Palautuksella tunnetusti NP-kovaan *exact cover by 3-sets*-ongelmaan voidaan osoittaa, että minimirekombinaatiohaplotyypitys on NP-kova ongelma. Joitakin erikoistapauksia ongelmasta kyetään kuitenkin ratkaisemaan joko eksaktisti tai stokastisin menetelmin.

1 Johdanto

Ihmisen perimä on varastoituna kunkin solun tumassa sijaitseviin kromosomeihin. Ihmisellä on kutakin kromosomia - ja siten kutakin kohtaa DNA:ssa, kuten geeniä, geneettistä markkeria, ja niin edelleen - kaksi kappaletta. Tavalliset tavat tuottaa genotyyppidataa yksilöistä antavat tiedon siitä mitkä kaksi versiota tietystä markkerista (eli mitkä kaksi *markkerialleelia*) ihmisellä on kussakin markkerikohdassa eli *lokuksessa*, mutta eivät tietoa siitä mitkä alleelit eri lokuksissa ovat sijaitsevat samassa kromosomissa.

Koska kuitenkin ihminen perii kromosominsa kokonaisuuksina vanhemmiltaan, toisen isältä ja toisen äidiltä, on perinnöllisyyden tutkimisen kannalta oleellista tietää mitkä markkeerialleelit ovat periytyneet yhdessä. Genotyyppidatan jakamista kahteen osaan, äidiltä perittyihin ja isältä perittyihin alleeleihin, sanotaan haplotyypitykseksi, ja se on yleensä ensimmäinen työvaihe esimerkiksi juuri perinnöllisten tautien analysoinnissa.

Tässä tutkielmassa käsitellään haplotyypitystä, jossa tutkittavat yksilöt ovat sukua toisilleen, ja käytössä on kunkin yksilön genotyypin lisäksi tieto näistä sukulaisuussuhteista (sukupuu). Määritellään (luvussa 2) eräs haplotyypitysongelma, nimittäin vähäisimpien rekombinaatioiden haplotyyppikonfiguraation etsiminen, ja todistetaan (luvussa 3), että tämä ongelma on NP-täydellinen. Lopuksi (luvussa 4) hyvin lyhyesti käsitellään tämän seurauksia ja ongelman ratkaisumahdollisuuksia.

Tässä työssä käsiteltävä ongelma on vain yksi mahdollinen määritelmä haplotyypitykselle. Ensinnäkin on mahdollista määritellä etsittävä haplotyypitys muutenkin kuin rekombinaatioiden vähäisyyden kautta, ja toiseksi on mahdollista tietyin oletuksin tehdä haplotyypitystä ilman sukupuutietoa. Näitä ongelmia ei kuitenkaan käsitellä tässä.

2 Määritelmiä

2.1 Markkerit, genotyypit, haplotyypit

Ihmisen perimä on jakaantunut 22 somaattiseen kromosomipariin sekä kahteen sukupuolikromosomiin. Kunkin somaattisen parin kummassakin kromosomissa sijaitsevat samat *geenit* eli proteiineja koodaavat osat samassa järjestyksessä, samanlaisin välialuein. Ihmisellä on siis joitakin sukupuolikromosomien geenejä lukuunottamatta kaksi kopiota kustakin kohdasta DNA:ssa. Rajoitumme seuraavissa määritelmissä [without loss of generality] tarkkailemaan vain yhtä kromosomiparia.

Tiettyä kohtaa kromosomissa kutsutaan *lokukseksi. Markkeri* on pari, jossa tiettyyn lokukseen liittyy tietty joukko mahdollisia tiloja siten, että laboratoriossa voidaan kokeellisesti määrittää DNA-juosteesta ko. markkerin tila. Kunkin markkerin mahdollisia tiloja kutsutaan ko. markkerin *alleelien joukoksi.* Kullakin yksilöllä on siis kussakin markkerissa kaksi alleelia. Tietyn markkerin alleeleja on tapana merkitä järjestysnumeroin, mutta on huomattava, että numerointi on yleensä täysin mielivaltainen.

Tavallisimmat markkerianalyysityypit ovat niin sanotut yhden nukleotidin polymorfismit (*single nucleotide polymorphisms*, SNPs), joissa alleelien joukko kaikille markkereille on $\{0, 1\}$, ja mikrosatelliittimarkkerit, joissa markkerien joukko tyypillisesti on suurempi.

Olkoon $M = \{m_1, ..., m_n\}$ joukko markkereita ja $A(m_i) = \{1, ..., k_i\}, k_i \in N_+$ kunkin markkerin alleelien joukko. Sanomme, että M on *geneettinen kartta.* Yksilön *genotyyppi* geenikartan M suhteen on järjestämättömien parien jono

$$g = \langle (g_1[1], g_2[1]), ..., (g_1[n], g_2[n]) \rangle$$

missä $g_1[i], g_2[i] \in A(m_i)$ ovat yksilön alleelit ko. markkerissa. Jos $g_1[i] = g_2[i]$, yksilö on markkerin m_i suhteen *homotsygootti*, muuten *heterotsygootti.* Jatkossa oletamme, että markkerit on indeksoitu sen mukaan, missä järjestyksessä ne sijaitsevat kromosomissa.

Haplotyyppi kartalla M on jono alleeleja $h = \langle h[1], ..., h[n] \rangle$ s.e. $h[i] \in A(m_i)$. Haplotyyppipari (h, h') on *konsistentti* genotyypin g kanssa jos ja vain jos kaikille markkereille $i \in 1, ..., n$ pätee

$$((h[i] = g_1[i] \wedge h'[i] = g_2[i]) \vee (h[i] = g_2[i] \wedge h'[i] = g_1[i]))$$

.

Helposti nähdään, että yhtä haplotyyppiparia vastaa vain yksi mahdollinen konsistentti genotyyppi, kun taas yhtä genotyyppiä voi vastata useita mahdollisia konsistentteja haplotyyppipareja. Helposti nähdään myös, että annettuna genotyyppi g ja haplotyyppi h, on olemassa tasan yksi haplotyyppi h' s.e. (h, h') on konsistentti g:n kanssa.

Esimerkki. Olkoon $A(m_i) = \{0, 1\}$ kaikille markkereille ja $n = 3$. Eräs mahdollinen genotyyppi tällaisella kartalla on $g = \langle (0, 1), (1, 1), (0, 1) \rangle$. Eräs esimerkki mahdollisesta g:n kanssa konsistentista haplotyyppiparista on $(\langle 0, 1, 0 \rangle, \langle 1, 1, 1 \rangle)$. Toinen esimerkki on $(\langle 1, 1, 0 \rangle, \langle 0, 1, 1 \rangle)$.

Yksilön haplotyypityksen kartan M suhteen muodostaa se haplotyyppipari, joka saadaan ottamalla kumpaankin haplotyyppiin ne markkerialleelit, jotka sijaitsevat samassa kromosomissa. Yksilö siis perii yhden haplotyypin isältään ja toisen äidiltään. On mahdollista analysoida yksilön perimä jo laboratoriossa haplotyyppitasolla, mutta menetelmät ovat kalliita [DBG$^+$01]. Siksi saatavilla oleva data yleensä koostuu genotyypeistä, jotka on laskennallisesti jaettava haplotyypeiksi, jos haplotyyppitietoa tarvitaan.

2.2 Sukupuut

Sukupuu on suunnattu syklitön verkko $G = (V, E)$ s.e. $V = M \cup F \cup L$, missä M on miessolmujen joukko, F on naissolmujen joukko, $M \cup F$ yksilösolmujen joukko ja L lisääntymissolmujen joukko, ja $E = \{(u, v) : u \in M \cup F$ ja $v \in L$ tai päinvastoin$\}$. Lisäksi vaaditaan, 1) että jokaisen lisääntymissolmun sisäätulevien kaarten määrä on tasan kaksi s.e. toinen kaarista tulee mies- ja toinen naissolmusta, ja 2) jokaisen yksilösolmun sisääntulevien kaarten määrä on korkeintaan 1 ja yksilösolmuihin tulevat kaaret lähtevät aina lisääntymissolmuista. (Kuva 1.) [LJ03]

Kukin lisääntymissolmu määrittelee *ydinperheen* seuraavasti. Määritelmän mukaan jokaiseen solmuun $l \in L$ liittyy yksi solmu $M_l \in M$, yksi solmu $F_l \in F$ ja yksi tai useampia solmuja $C_l \in M \cup F$. Syklittömyydestä seuraa, että $C_l \neq M_l, C_l \neq M_l$, aina. Solmut M_l, F_l ovat solmujen C_l *vanhemmat* (isä ja äiti, samassa järjestyksessä) ja C_l ovat vanhempiensa *lapset*. Niitä yksilösolmuja, joilla ei ole graafissa vanhempia, sanotaan *perustajiksi* (engl. *founders*). (Kuva 1.) [LJ03]

Genotyypitetty sukupuu on sukupuu, jossa johonkin tai kaikkiin yksilösolmuihin liittyy tieto niiden genotyypeistä (jonkin kartan M suhteen). Yksilösolmu on tällaisessa puussa genotyypitetty, jos nimenomaan siihen liittyy genotyyppitieto.

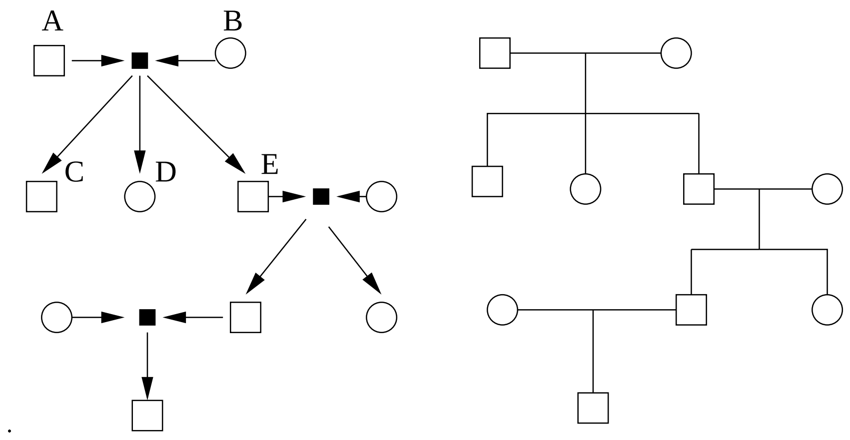
Mendelin periaate sanoo, että jokaisessa lapsen markkerilokuksessa toisen alleelin pitää olla peräisin äidiltä ja toisen isältä. Toisin sanoen jos oletetaan, että sukupuussa ei ole tapahtunut mutaatioita, niin kullekin markkerille m_i kaikissa puun tripleteissä täytyy päteä, että lapsen genotyyppi ko. markkerin suhteen on muodostettavissa ottamalla toinen äidin alleeleista ja toinen isän alleeleista ko. markkerissa.

Tarkastellaan seuraavaksi yhtä sellaista genotyypitettyä yksilösolmua, joka ei ole puun perustaja. Jokaiseen markkeriin ko. yksilön genotyypissä voidaan liittää tieto siitä, kumpi markkerialleeli on peritty yksilön isältä ja kumpi äidiltä. Kiinnitetään mielivaltaisesti yksilön genotyypin alleelien järjestys kussakin markkerissa; sovitaan esimerkiksi, että $g_1[i]$ on aina se alleeli, jonka järjestysnumero on pienin.

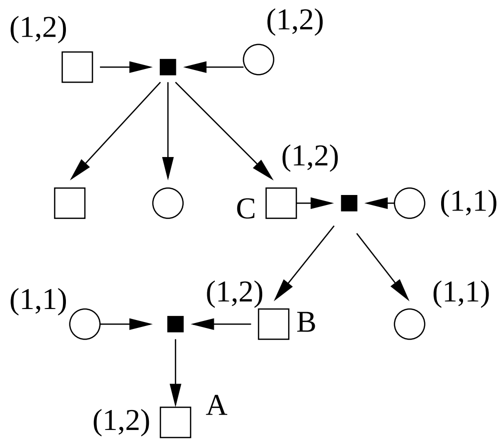
Nyt voidaan merkitä tietoa alleelin alkuperästä $PS^A(i)$ (missä i on markkerin indeksi, kuten edellä ja A yksilön tunnus; PS tulee sanoista *parental source*) ja määritellä $PS^A(i) = 0$, jos $g_1[i]$ on peritty isältä, 1 jos se on peritty äidiltä, ja -1 jos tietoa ei ole. Sovitaan vielä, että mikäli yksilö on homotsygootti markkerin i suhteen, $PS^A(i) = 0$ aina. [LJ03] Koska alleelit ovat identtiset, näin voidaan tehdä ilman että se häiritsee haplotyypitystä; määrittely vastaa tilannetta jossa alleelien järjestystä kiinnittäessä olisi sattumalta arvattu oikea järjestys.

Yksilö A on *määritetty* markkerin i suhteen, jos $PS^A(i) \geq 0$. Mikäli yksilön vanhempien vanhemmat ovat myös sukupuussa, tällöin voidaan määritellä vielä $GS^A(i, k)$ (*grandparental source*, $k \in \{1, 2\}$) ko. markkerin kummallekin alleelille kuvaamaan sitä, onko alleeli peräisin yksilön asianomaisen vanhemman äidiltä vai isältä. Samoin kuin PS, GS saa arvoja joukosta $\{-1, 0, 1\}$. [LJ03]

Sukupuun haplotyyppikonfiguraatio määrittelee ei-negatiivisen PS-arvon jokaisen genotyypitetyn solmun jokaiselle markkerille, ja ei-negatiivisen GS-arvon jokaiselle vastaavalle alleelille niin, että Mendelin periaatetta ei rikota. [LJ03]



Kuva 1: Esimerkki sukupuusta. Tyhjät neliöt kuvaavat miessolmuja, ympyrät naissolmuja, ja täytetyt neliöt lisääntymissolmuja. Solmut A ja B ovat solmujen C, D, ja E vanhemmat. Vasemmalla täsmälleen määritelmän mukaan piirretty puu, oikealla sama puu tavallisemmin käytetyllä notaatiolla.

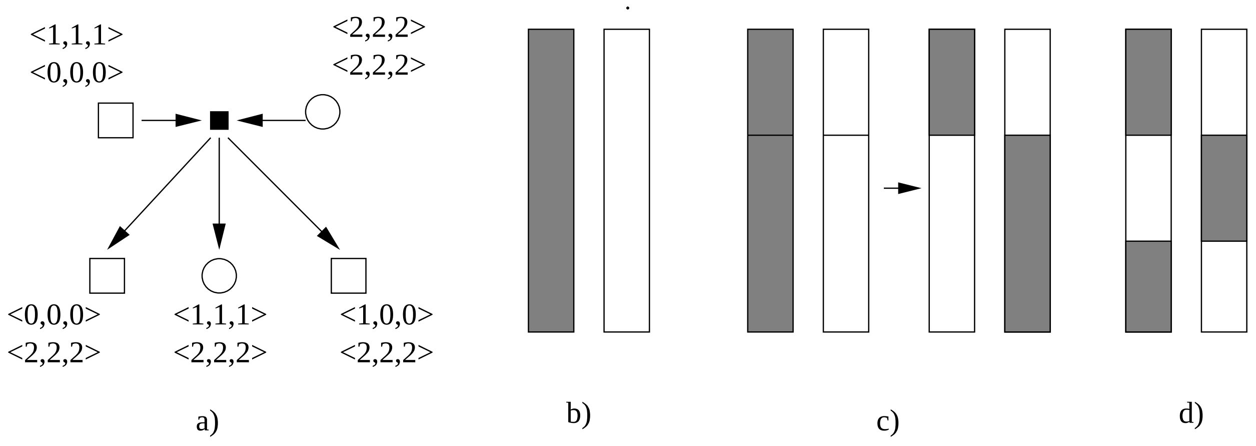


Kuva 2: Eräs genotyypitetty sukupuu. Ks. esimerkin selitys tekstissä.

On selvää, että sukupuun haplotyyppikonfiguraatio määrittelee samalla haplotyypityksen kullekin genotyypitetylle yksilölle; haplotyypit ovat ne alleelit, jotka yksilö on perinyt tietyltä vanhemmaltaan.

Esimerkki. Tarkastellaan kuvan 2 sukupuuta, johon on merkitty kullekin yksilölle yhden markkerin genotyyppi (mahdolliset alleelit $\{1,2\}$). Selvästi yksilön A on täytynyt saada alleeli 1 äidiltään ja alleeli 2 isältään. PS^A on siis 1. Samoin yksilön B on täytynyt saada alleeli 2 isältään, sillä B:n äidillä ei ole alleelia 2. Siis PS^B on myös 1. A:n alleeli 2 on peräisin isänisältä; siis $GS^A(2) = 0$. Yksilö C sen sijaan on epäinformatiivinen ko. alleelin suhteen: sukupuun perusteella ei ole mahdollista sanoa kummalta vanhemmaltaan hän on perinyt kummankin alleelinsa. Siis $PS^C = -1$. Mahdollisia Mendelin periaatetta noudattavia haplotyyppikonfiguraatioita on siis tässä puussa useita.

On syytä huomata, että mikäli markkereissa on tapahtunut mutaatiota, Mendelin periaatetta noudattavaa sukupuuta ei välttämättä ole, ja että puun perustajille arvot voidaan määritellä mielivaltaisesti periaatteen rikkoutumatta.



Kuva 3: a) Sukupuu kolmen markkerin suhteen haplotyypitetystä perheestä. Perheen kaksi ensimmäistä lasta ovat perineet isänsä jommankumman haplotyypin sellaisenaan. Kolmannen lapsen kohdalla on tapahtunut rekombinaatio. Huomaa, että on mahdotonta sanoa, onko jonkun lapsen kohdalla äidin haplotyypissä tapahtunut rekombinaatiota, sillä tämän kaksi haplotyyppiä ovat identtiset. b) Yleensä sukusolujen muodostuessa tietty alue kromosomissa päätyy sellaisenaan kahteen sukusoluun. c) Toisinaan kuitenkin kaksi kopiota kromosomista risteää jossain kohdassa ja tällöin voi käydä niin, että ne irtoavat risteämiskohdasta niin, että materiaalia vaihtuu. Tätä sanotaan rekombinaatioksi. Tarkat biologiset yksityiskohdat eivät ole tärkeitä, mutta on tärkeää ymmärtää, että risteämiskohdan molemmin puolin kaikki materiaali vaihtaa paikkaa. Kohdan d) tilanne vaatii siis kaksi rekombinaatiota ja jos oletetaan, että rekombinaatiot ovat riippumattomia, on epätodennäköisempi kuin kohdan c) tapahtuma.

2.3 Rekombinaatio

Mikäli kromosomit periytyisivät aina samanlaisina, yhteen kromosomiin rajoittuvalla kartalla M yksilön yhden haplotyypin kaikkien alleelien GS-arvot olisivat samat. Näin ei kuitenkaan aina ole, sillä sukusolujen syntymisen aikana on mahdollista, että kromosomiparin kromosomit vaihtavat materiaalia keskenään; tätä sanotaan *rekombinaatioksi*. Vaikka siis yksilö perii tasan yhden kromosomin äidiltään, tästä kromosomista osa voi olla peräisin äidinäidiltä ja osa äidinisältä; toisin sanoen yksilön äidiltään perimä kromosomi voi olla rekombinaatio äidin kahdesta kromosomista.

Mikäli yksilön tiettyyn haplotyyppiin kuuluvissa alleeleissa $h[i]$ ja $h[i+1]$ toisen GS-arvo on 0 ja toisen 1, sanotaan, että markkerien m_i ja m_{i+1} välillä on tapahtunut rekombinaatio. Yleensä oletetaan, että rekombinaatiot eivät tapahdu markkerien sisällä, vaikka se tosiasiassa onkin mahdollista esimerkiksi mikrosatelliittien tapauksessa.

Esimerkki. Olkoon yksilön A haplotyypit kolmen eri markkerin suhteen $(\langle 0,0,0\rangle, \langle 1,1,1\rangle)$; hän on siis heterotsygootti kaikissa markkereissa. Jos rekombinaatiota ei tapahtuisi, yksilön A lapset perisivät häneltä aina jommankumman haplotyypin. Jos kuitenkin markkerien 1 ja 2 välillä tapahtuu rekombinaatio sukusolujen muodostuessa, lapsi perii jommankumman haplotyypeistä $(\langle 0,1,1\rangle, \langle 1,0,0\rangle)$. Jos sekä markkerien 1 ja 2 että markkerien 2 ja 3 välillä tapahtuu rekombinaatio, lapsi perii jommankumman haplotyypeistä $(\langle 0,1,0\rangle, \langle 1,0,1\rangle)$. Ks. myös kuva 3.

Rekombinaatiot ihmisessä ovat melko harvinaisia [GSN$^+$02]. On siis järkeenkäypää ajatella, että kaikista mahdollisista haplotyyppikonfiguraatioista ne, jotka vaativat vähemmän rekombinaatioita ovat todennäköisemmin oikeita.

Vähäisimmän rekombinaation haplotyyppikonfiguraatio-ongelma (*Mimimum Recombinant Haplotype Configuration*, MRHC) määritellään seuraavasti: annettuna genotyypitetty sukupuu, löydettävä haplotyyppikonfiguraatio joka vaatii pienimmän mahdollisen määrän rekombinaatioita. [LJ03]

3 Ongelman NP-täydellisyys

Seuraavaksi osoitetaan, seuraten Jing Lin ja Tao Jiangin todistusta [LJ03], että MRHC on NP-kova, palauttamalla se välivaiheiden kautta *exact cover by 3-sets*-ongelmaan, joka on tunnetusti NP-kova.

Exact cover by 3-sets (XC3). Annettuna joukko U s.e. $|U| = 3m$ (missä m on jokin positiivinen kokonaisluku) ja kokoelma U:n osajoukkoja $S = \{S_1, ..., S_n\}$ s.e. $|S_i| = 3$ kaikille $i = 1, ...n$, pääteltävä onko S:ssä m kpl erillisiä joukkoja siten, että niiden yhdiste on U.

XC3 on NP-kova, ja myös sen muunnelma 3XC3, jossa kukin U:n alkio esiintyy korkeintaan kolmessa S:n joukossa, on NP-kova [LJ03]. Li ja Jiang osoittavat, että ongelma on edelleen NP-kova, jos oletetaan, että kukin U:n elementti esiintyy *täsmälleen* kolmessa S:n joukossa. MHRC:n NP-kovuus perustuu palautukseen tähän "3XCX3"-ongelmaan.

Väite. MRHC on NP-kova.

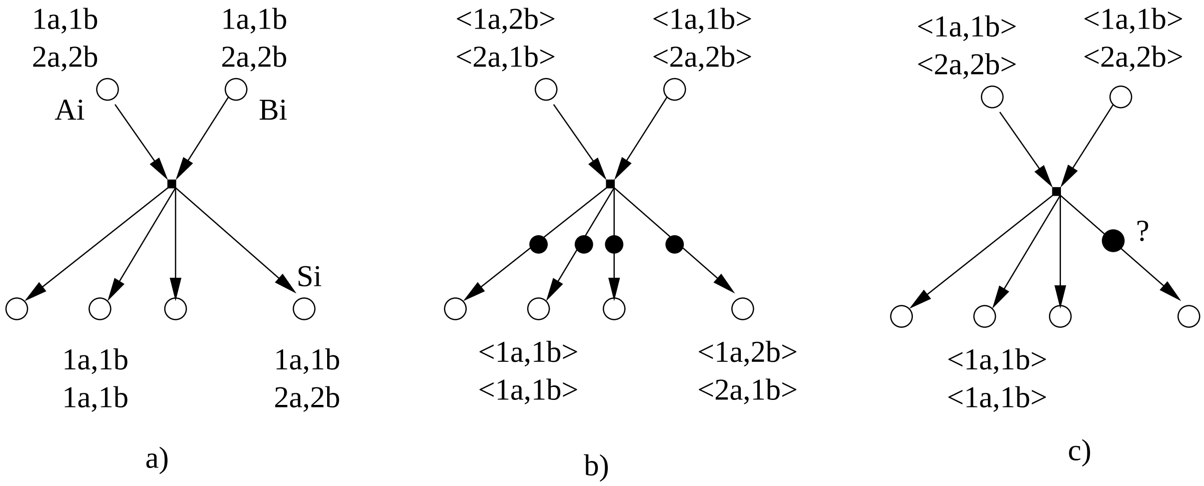
Havaintoja. Selvästi riittää osoittaa, että tapaus, jossa sukupuun kartassa M on vain kaksi markkerilokusta, on NP-kova, sillä algoritmia joka ratkaisee useamman markkerin tapauksen voitaisiin käyttää kahden markkerin tapauksen ratkaisemiseen ottamalla mukaan "ylimääräisiä" markkereita.

Samoin voidaan olettaa, että kussakin markkerilokuksessa esiintyy vain kahta alleelia; vaikka jotkut useampia alleeleja sisältävät erikoistapaukset olisivatkin ratkaistavissa pienemmässä ajassa, ei rajoite, jossa sukupuussa täytyy esiintyä vähintään m:ää eri alleelia tietyistä lokuksista, ole todellisen elämän tilanteessa mielekäs - erityisesti SNP-markkerien yleistyessä bialleelinen tapaus on nykyään enemmin sääntö kuin poikkeus.

Todistuksen idea. Rajoitutaan siis tarkastelemaan kahta markkeria joissa kummassakin on kaksi mahdollista alleelia. Merkitään selvyyden vuoksi toisen markkerin alleeleja $\{1_a, 2_a\}$ ja toisen $\{1_b, 2_b\}$. Tavoitteena on osoittaa, että annettuna mielivaltainen 3XCX3-ongelman instanssi, on mahdollista konstruoida polynomiaalisessa ajassa genotyypattu sukupuu, jonka vähimpien rekombinaatioiden haplotyypitys ratkaisee annetun 3XCX3-ongelman.

Konstruktio aloitetaan muodostamalla yksilösolmu kullekin joukolle S_i. Näitä joukkoja/solmuja on siis yhtä monta kuin alkioita 3XCX3-ongelman perusjoukossa, ja kukin alkio kuuluu tasan kolmeen joukkoon. Jokaisen yksilön S_i genotyyppi on $\langle(1_a, 2_a), (1_b, 2_b)\rangle$, ts. he ovat heterotsygootteja molempien markkerien suhteen; tällöin S_i:n haplotyypit ovat joko $(\langle 1_a, 1_b\rangle, \langle 2_a, 2_b\rangle) = H$ tai $(\langle 1_a, 2_b\rangle, \langle 2_a, 1_b\rangle) = H'$. (Merkinnät H ja H' tarvitaan myöhemmin lyhyyden vuoksi.)

Nyt jos voidaan muodostaa sukupuu, jossa ottamalla ne S_i joiden haplotyypitys minimirekombinaatiokonfiguraatiossa on H', saadaan suoraan 3XCX3-ongelman ratkaisu, on väite osoitettu.



Kuva 4: a) Eräs solmu S_i vanhempineen ja sisaruksineen ja kunkin genotyypit (S_i:n sisaruksilla sama). b) Sisaruksien haplotyypitykselle on vain yksi vaihtoehto; heidän on täytynyt periä molemmilta vanhemmiltaan haplotyyppi $\langle 1a, 1b\rangle$. Jos siis toinen vanhemmista on tyyppiä H' on jokainen sisarus rekombinantti (ja jos vain toinen vanhemmista on tätä tyyppiä, myös S_i on). c) Jos taas kumpikin vanhemmista on tyyppiä H, ei välttämättä kukaan lapsista ole rekombinantti - ja jos S_i onkin, rekombinaatioiden määrä on silti vähäisempi kuin kohdassa b). Näin valitsemalla lasten genotyypit sopivasti on pakotetettu tietty haplotyypitys vanhemmille.

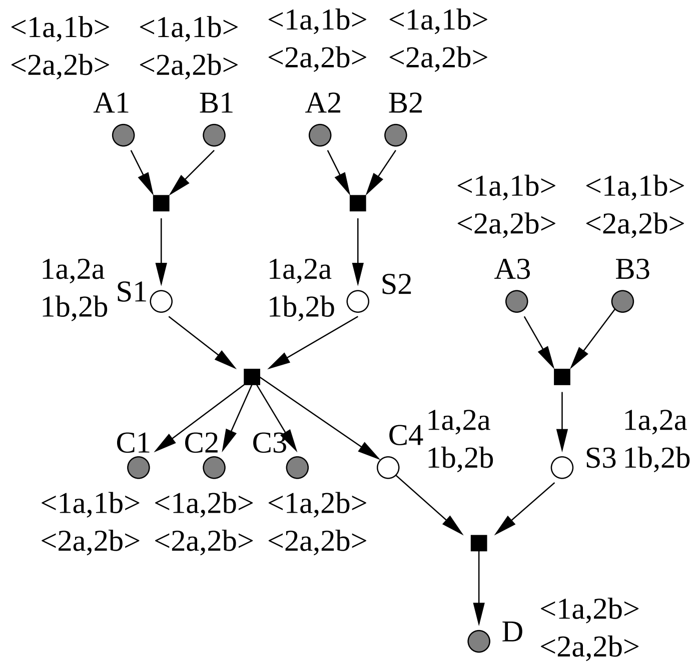
Sukupuun konstruktio. Kullekin yksilölle S_i luodaan vanhemmat A_i ja B_i, jotka myös ovat heterotsygootteja molempien markkereiden suhteen. Antamalla A_i:lle ja B_i:lle tietty vakiomäärä lapsia, joiden genotyyppi on $\langle(1_a, 1_a), (1_b, 1_b)\rangle$, voidaan "pakottaa" minimirekombinaatiopuu sellaiseksi, että A_i:n ja B_i:n haplotyypit ovat $(\langle 1_a, 1_b\rangle, \langle 2_a, 2_b\rangle)$. Tällöin S_i:n haplotyypit voivat olla $(\langle 1_a, 2_b\rangle, \langle 2_a, 1_b\rangle)$ vain, jos S_i itse on rekombinantti. (Kuva 4.)

Oletetaan nyt, että U:n tietty elementti esiintyy joukoissa S_1, S_2 ja S_3. Tavoitteena on seuraavaksi generoida sellainen pieni sukupuu, jossa esiintyvät nämä kolme solmua ja jonka minimirekombinaatiotapauksessa vain yhden niistä haplotyypit ovat H' ja muiden kahden ovat H.

Tässä pienessä sukupuussa S_1 ja S_2 lisääntyvät keskenään saaden lapset $C_1, C_2, C_3, C4$. Samaan tapaan kuin yllä solmun S_i vanhempien tapauksessa voidaan nämä lapset pakottaa tiettyihin haplotyyppikonfiguraatioihin minimirekombinaatiopuussa luomalla heille sellaiset puoliso ja lapset, että muilla kuin tietyillä haplotyypeillä puuhun syntyisi useita rekombinaatiota. Tässä tapauksessa pakotetaan haplotyypit olemaan C_1:llä $(\langle 1_a, 2_b\rangle, \langle 2_a, 2_b\rangle)$ ja C_2:lla ja C_3:lla $(\langle 1_a, 2_b\rangle, \langle 2_a, 2_b\rangle)$. C_4:lle määritellään genotyyppi $\langle(1_a, 2_a), (1_b, 2_b)\rangle$, mutta ei kiinnitetä haplotyyppejä.

Seuraavaksi lisätään puuhun S_3 siten, että se lisääntyy C_4:n kanssa saaden jälkeläisen D. D pakotetaan haplotyyppeihin $(\langle 1_a, 2_b\rangle, \langle 2_a, 2_b\rangle)$.

Tämä syntynyt sukupuu (kuva 5) on sellainen, että minimirekombinaatiotilanteessa vain yksi solmuista S_1, S_2, S_3 voi olla haplotyyppikonfiguraatiossa H'. Tämä voidaan helposti nähdä käymällä läpi kaikki vaihtoehdot. Huomataan ensin, että S_1:lle ja S_2:lle on olemassa neljä haplotyyppivaihtoehtoa: joko molemmat ovat haplotyyppejä H, molemmat ovat haplotyyppejä H' tai ne edustavat eri haplotyyppejä; puun topologiasta seuraa, että järjestyksestä riippumatta viimeiset tapaukset ovat identtiset, eli erilaisia tapauksia on yhteensä kolme. S_3 voi samoin olla H tai H'; yhteensä saadaan siis kuusi erilaista tapausta.



Kuva 5: Solmuista S_1,S_2,S_3 muodostettu apupuu. Ympyrät kuvaavat yksilösolmuja, neliöt lisääntymissolmuja. Harmaannutettuihin yksilösolmuihin liittyy lapsia ja/tai puolisoita, joiden avulla niiden haplolyypit on pakotettu tiettyyn asemaan, mutta joita ei ole piirretty kuvaan

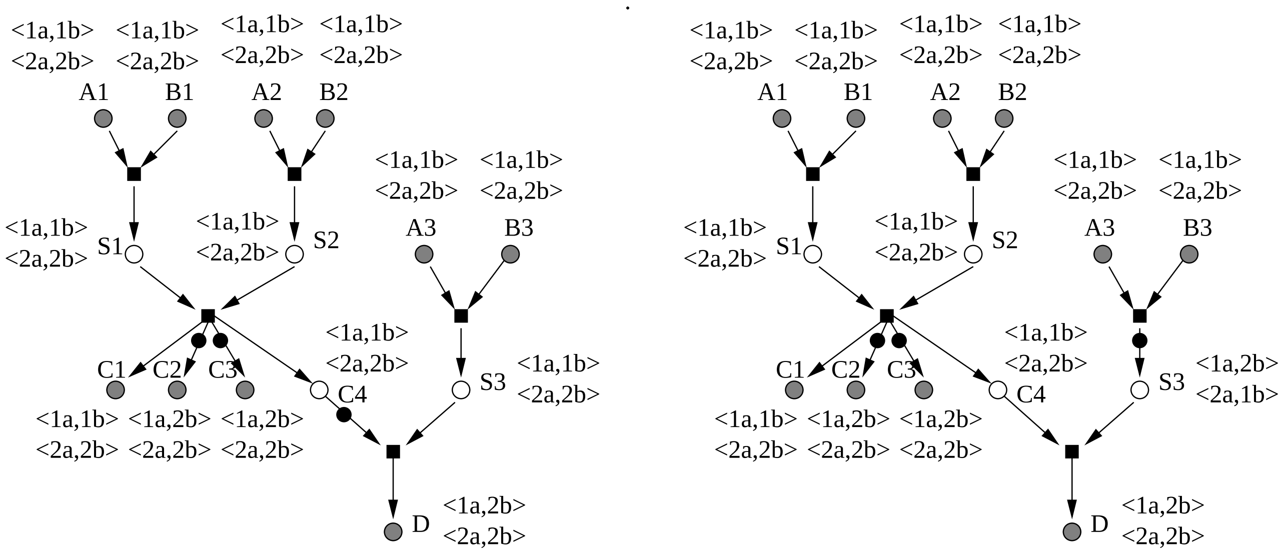
Laskemalla tarvittavien rekombinaatioiden määrä näissä kaikissa päädytään siihen, että pienin mahdollinen rekombinaatioiden määrä on kolme. Kuitenkin, vain silloin kun tasan yksi solmuista S_i on konfiguraatiossa H' ja kaksi muuta konfiguraatiossa H saadaan aikaan tilanne, jossa yksi näistä kolmesta rekombinantista on joku solmuista S_i. Muissa tapauksissa rekombinantit ovat kaikki solmujen S_i lapsissa. (Kuva 6.) Tämä on oleellista seuraavan vaiheen kannalta.

Seuraavaksi nimittäin tehdään sama pienen sukupuun konstruktio kaikille $a \in U$. Tämän jälkeen varsinainen puu saadaan samaistamalla solmut, jotka kuvaavat joukkoa S_i. Jokaisella solmulla S_i on edelleen vain yhdet vanhemmat, mutta lapsiryhmiä $(C_1 - C_4, D)$ on kutakin solmua kohden useita. Koko puu on selvästi konstruoitavissa polynomisessa ajassa.

Edellisessä yhdistämisessä ongelmaksi muodostuu se rajoite, että varsinaisessa sukupuussa kullakin yksilöllä tulee olla sukupuoli, ja että vain erisukupuoliset yksilöt voivat lisääntyä keskenään. Tämä ongelma voidaan kiertää duplikoimalla jokainen solmu S_i ja nimeämällä toinen mies- ja toinen naissolmuksi. Näin voidaan valita jokaisen lisääntymissolmun vanhemmat niin, että toinen on mies ja toinen nainen. Se, että kullakin solmun S_i instanssilla on aina samat haplotyypit voidaan varmistaa luomalla heille lapset, jotka pakottavat minimirekombinaatiopuun tähän.

3XCX3:n palautus. Jokaisessa alipuussa siis minimirekombinaatioiden määrä on 3, mutta on mahdollista sijoittaa tämä kolme niin, että eri alipuut voivat ”jakaa” tietyt rekombinaatiot. Jokainen S_i on nimittäin mukana kolmessa alipuussa, sillä jokaisessa S_i on kolme alkiota. Koko puun rekombinaatiot minimoi haplotyypitys, jossa jokaisessa alipuussa mahdollisimman suuri osa rekombinaatioista jaetaan jonkin toisen alipuun kanssa.

Lin ja Jiangin mukaan ”on helppo osoittaa”, että 3XCX3:n ratkaisu vastaa ylläkuvatun puun vähäisimpien rekombinaatioiden haplotyypitystä tietyllä rekombinaatioiden määrällä



Kuva 6: Solmuista S_1,S_2,S_3 muodostettu apupuu, samoin kun kuvassa 5. Vasemmalla: jos kiinnitetään kaikille solmuille S_i haplotyypitys H, vaatii puu vähintään kolme rekombinaatiota (merkitty mustilla palloilla nuolten päällä), joista kaksi on solmuissa C_2 ja C_3 ja kolmas joko solmussa C_4 tai solmussa D. Oikealla: jos taas vaihdetaankin S_3:n haplotyypitykseksi H', voidaan yksi rekombinaatioista "siirtää" S_3:lle. Helposti nähdään, että myös S_1:n tai S_2:n tai molempien muuttaminen samaan aikaan ei vähentäisi rekombinaatioita, ja että siirto voitaisiin tehdä myös S_1:lle tai S_2:lle vastaavasti (kuvien piirto jää tilanpuutteen vuoksi lukijalle).

[LJ03]. Intuitiivisesti on selvää, että jos ko. puun rekombinaatiot minimoivassa haplotyyppikonfiguraatiossa jokaisessa alipuussa tasan yksi saman alipuun solmuista S_i on haplotyypitystä H', 3XCX3:n ratkaisu ovat ne S_i, joiden haplotyypitys on H'. Jos ratkaisua ei ole, on pienin mahdollinen rekombinaatioiden määrä *suurempi* kuin tuo mainittu määrä.

Nyt siis jos on mahdollista vastata kysymykseen "onko mielivaltaiselle genotyypitetylle sukupuulle olemassa haplotyyppikonfiguraatiota, jossa tarvittavien rekombinaatioiden määrä on korkeintaan X", esimerkiksi etsimällä vähäisimpien rekombinaatioiden haplotyyppikonfiguraatio, samaa algoritmia voidaan käyttää vastaamaan 3XCX3-ongelmaan.

Päätelmä. Siis 3XCX3 on palautettavissa MRHC:hen, joten MRHC on NP-kova. □

4 Huomioita

MRHC:n NP-kovuustodistuksen valossa ei ole hämmästyttävää, että useimmat ongelman ratkaisemiseen käytetyt algoritmit käyttävät jotain enemmän tai vähemmän perusteltua heuristiikkaa ja/tai eivät pysty käsittelemään kovin suuria sukupuita [LJ03]. Mielenkiintoista on, että ilmeisesti myös MRHC sukupuille, joissa ei sallita edes suuntaamattomia syklejä, on NP-kova, ja että vastaavia todistuksia on löydetty myös muille samantapaisille sukupuuongelmille [LJ03] (Li ja Jiang siteeraavat toistaiseksi julkaisemattomia tuloksia).

Ongelman ratkaisemiseksi on kehitetty erilaisia heuristisia lähestymistapoja. Ensinnäkin vastaava 0-rekombinaatio-ongelma [1] on polynomiaalinen [LJ03]. Toiseksi, mikäli puun koko

[1]annettuna genotyypitetty sukupuu, on löydettävä haplotyyppikonfiguraatio, jossa ei ole lainkaan rekom-

tai markkerilokusten määrä on rajoitettu, erilaiset ahneetkin algoritmit suoriutuvat hyvin [LJ03, QB02].

NP-kovuustulos on mielenkiintoinen, sillä ongelma on tällä hetkellä laskennallisen genetiikan kentässä hyvin tärkeä. Analyysimenetelmien kehittyessä ja alan nykyisin herättämän suuren kiinnostuksen takia datamäärät kasvavat koko ajan, joten aikaisemmin riittävän hyvin toimineet algoritmit ovat nopeasti saavuttaneet suorituskykynsä rajat. Lisäksi osa heuristiikoista toimii hyvin huonosti nimenomaan bialleelisten markkerien tapauksessa [QB02], mikä on ongelma sen vuoksi, että SNP-kartat, jotka ovat tiheimpiä ja siksi mielenkiintoisimpia nykyisin käytössä olevista geneettisistä kartoista, ovat nimenomaan bialleelisia.

Käytössä olevat algoritmit kärsivät kahdentyyppisistä rajoitteista. Toisaalta on olemassa täsmälliseen ratkaisuun päätyviä algoritmeja, joiden mahdolliset käyttöalueet ovat rajoitettuja - yleensä siten, että niiden aikavaatimus kasvaa eksponentiaalisesti joko markkerien määrän tai sukupuun tai molempien kasvaessa [Oe03]. Toisaalta on olemassa stokastisia algoritmeja, jotka pystyvät käsittelemään suurempia datamääriä mutta eivät takaa globaalisti optimaalista ratkaisua; tällaisia ovat esimerkiksi Markovin ketju -malliin ja simuloituun jäähdyttämiseen perustuva SimWalk-ohjelmiston toteuttama malli [SL96] tai erilaiset sääntöpohjaiset heuristiset mallit [LJ03, QB02].

Toisaalta on hyvä kysymys, onko minimirekombinaatio-oletus järkevin mahdollinen tapa valita haplotyyppikonfiguraatiota. Mikäli ihmisen perimässä havaittujen haplotyyppiblokkien [GSN+02] taustalla todella on ilmiö, jossa tietyt DNA:n ominaisuudet *aiheuttavat* rekombinaatioiden huomattavasti suuremman todennäköisyyden tietyissä kohdissa, pitäisi tämä luonnollisesti ottaa huomioon mallissa.

Joillekin suurimpaan uskottavuuteen perustuville algoritmeille onkin mahdollista mahdollista määritellä eri todennäköisyys rekombinaatiolle eri markkereiden välillä. Uskottavimman haplotyyppikonfiguraation etsiminen myös tällaiselle mallille on mielivaltaisille todennäköisyyksille NP-kova; tässä työssä esitettu MRHC-ongelma on tämän yleisemmän mallin erikoistapaus, jossa rekombinaatiotodennäköisyys kaikkien markkereiden välillä on sama. Ei kuitenkaan ole varmaa, etteivätkö sellaiset erikoistapaukset, joissa rekombinaatiotodennäköisyys useiden markkereiden välillä on käytännössä nolla, voisi olla polynomiaalisesti ratkeavia.

Kiinnostavaa on myös se, että laboratoriomenetelmien kehittyessä koko haplotyypitysongelma saattaa pian olla käytännön kannalta historiaa; jos laboratoriodata olisi alunperinkin haplotyyppidataa genotyyppidatan suhteen olisi monien perinnöllisten tautien analysointi huomattavasti täsmällisempää. [DBG+01]

5 Yhteenveto

Tässä työssä esitettiin vähäisimpien rekombinaatioiden haplotyyppikonfiguraatio -ongelma genotyypitetyille sukupuille, ja todistus sille, että se on NP-kova. Lyhyesti pohdittiin tämän seurauksia ongelmille ja mainittiin heurististen ratkaisutapojen mahdollisuuksista ja ongelmista.

Viitteet

[DBG+01] J A Douglas, M Boehnke, E Gillanders, J M Trent, and Gruber S B. Experimentally-derived haplotypes substantially increase the efficiency of linkage

binaatioita, tai todettava, että tällaista konfiguraatiota ei ole

disequilibrium studies. *Nature Genetics*, 28(4):361–364, July 2001.

[GSN$^+$02] S B Gabriel, S F Schaffner, H Nguyen, J M Moore, J Roy, B Blumenstiel, J Higgins, M DeFelice, A Lochner, M Faggart, S N Liu-Cordero, C Rotimi, A Adeyemo, R Cooper, R Ward, E S Lander, M J Daly, and D Altshuler. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, June 2002.

[LJ03] Jing Li and Tao Jiang. Efficient rule-based haplotyping algorithms for pedigree data. In *Proc. of 7th Annual Conference on Research in Computational Molecular Biology (RECOMB03)*, pages 197–206. ACM, April 2003.

[Oe03] P Onkamo (ed.). Genetics for computer scientists. Kurssimateriaali samannimisellä kurssilla, Tietojenkäsittelytieteen laitos, Helsingin Yliopisto, 2003.

[QB02] D Qian and L Beckmann. Minimun-recombinant haplotyping in pedigrees. *Am J Hum Genet*, 70(6):1434–1445, 2002.

[SL96] E Sobel and K Lange. Descent graphs in predigree analysis: applications to haplotyping, location scores, and marker sharing statistics. *Am J Hum Genet*, 58:1323–1337, 1996.