TEKNILLINEN KORKEAKOULU
Teknillisen fysiikan ja matematiikan osasto

Leo Lahti

Vertaileva toiminnallinen genomianalyysi assosiatiivisella ryhmittelymenetelmällä

Diplomi-insinöörin tutkintoa varten tarkastettavaksi jätetty diplomityö

Työn valvoja dosentti Samuel Kaski
Työn ohjaaja diplomi-insinööri Janne Nikkilä

Helsinki, 4. joulukuuta 2003

TEKNILLINEN KORKEAKOULU DIPLOMITYÖN TIIVISTELMÄ

Tekijä:	Leo Lahti
Työn nimi:	Vertaileva toiminnallinen genomianalyysi assosiatiivisella ryhmittelymenetelmällä
English title:	Comparative functional genome analysis using associative clustering
Päivämäärä:	4. joulukuuta 2003 **Sivumäärä:** 63
Osasto:	Teknillisen fysiikan ja matematiikan osasto
Professuuri:	T-61, Informaatiotekniikka
Työn valvoja:	dos. Samuel Kaski
Työn ohjaaja:	dipl.ins. Janne Nikkilä

Helpommin tutkittavien malliorganismien kuten hiiren avulla tehtävä tutkimus voi auttaa ymmärtämään myös ihmisen geenien toimintaa. Lajeja vertailemalla saatava lisätieto on arvokasta, sillä geenien toiminnan selvittäminen on valtava urakka. Tämän työn tavoitteena on etsiä biologisen tutkimuksen kannalta kiinnostavia ihmisen ja hiiren vastingeenien ryhmiä.

Tutkimukseen käytettävä assosiatiivinen ryhmittely on uusi eksploratiivisen data-analyysin menetelmä. Eksploratiivisten menetelmien tavoitteena on nopean yleiskuvan muodostaminen tutkimusaineistosta ilman siihen liittyvää ennakkotietoa. Eksploratiivisten menetelmien avulla voidaan tuottaa hypoteeseja myöhemmän tutkimuksen tarpeisiin. Genominlaajuisen ilmenemisdatan avulla muodostettava vastingeenien assosiatiivinen ryhmittely maksimoi ihmisen ja hiiren geeniryhmien riippuvuudet Bayesilaisen riippuvuusmitan mielessä. Tällä tavalla löydetään vastingeeniryhmiä, joiden toiminnallinen yhteys on muuhun aineistoon verrattuna poikkeuksellinen. Työn toisena tavoitteena on uuden assosiatiivisen ryhmittelymenetelmän ominaisuuksien arviointi ja vertaaminen vaihtoehtoisiin menetelmiin.

Assosiatiivinen ryhmittely osoittautui genomisten datajoukkojen vertailevassa analyysissa käyttökelpoiseksi eksploratiiviseksi menetelmäksi. Se onnistuu vaihtoehtoisia menetelmiä paremmin yhdistämään kaksi tavoitetta, riippuvuuksien mallintamisen ja helppotulkintaisuuden. Tutkimuksessa löydettiin potentiaalisesti mielenkiintoisia vastingeenien ryhmiä ja muodostettiin uusia hypoteeseja geenien yhteyksistä. Mahdollisten virhelähteiden merkitys ei ole tutkimuksen onnistumisen kannalta ratkaiseva, sillä muodostettavat hypoteesit on joka tapauksessa vahvistettava biologisissa tutkimuksissa. Tulosten biologisen merkityksen selvittäminen vaatii lisätutkimuksia.

Avainsanat: assosiatiivinen ryhmittely, geenien ilmenemisdata, geenien ortologia, vertaileva toiminnallinen genomiikka

Author: Leo Lahti

Title of thesis: Comparative functional genome analysis using associative clustering

Finnish title: Vertaileva toiminnallinen genomianalyysi assosiatiivisella ryhmittelymenetelmällä

Date: 4. joulukuuta 2003 **Pages:** 63

Department: Department of Engineering Physics and Mathematics

Chair : T-61, Computer and Information Science

Supervisor: Doc. Samuel Kaski

Instructor: M.Sc. Janne Nikkilä

Better understanding of human gene function is often gained by research on model organisms such as mouse. Such additional information is valuable as understanding gene function and genetic networks in a genome-wide scale is a huge mission. The aim of this work is to find biologically interesting groups of orthologous mouse and human genes using two genome-wide expression data sets.

Associative clustering is a new tool for exploratory data analysis. Exploratory methods are general-purpose instruments that illustrate the essential features of a data set. They do not require prior information of research data and are often used to produce new hypotheses for the purposes of later study. Associative clustering method finds human and mouse gene clusters so as to maximize a Bayesian dependency measure of the two sets of clusters. This reveals orthologous gene groups that are functionally exceptional with respect to other data. In this work, we also evaluate features of the new method and compare its performance to alternative methods.

Associative clustering proves to be a useful exploratory method for comparative analysis of genome-wide expression data sets. Compared to alternatives, it is able to find a better compromise between dependency modeling and easily interpretable clusters. We could find find potentially interesting groups of orthologous genes and to form new hypotheses about gene function. Possible sources of error are not crucial for analysis as new hypotheses should be confirmed in biological studies. Further study is needed to find out the biological significance of the results.

Keywords: associative clustering, comparative functional genomics, gene expression data, gene orthology

Esipuhe

Tämä diplomityö on tehty neuroverkkojen tutkimusyksikössä Teknillisessä korkeakoulussa. Työn on rahoittanut Suomen Akatemia (projekti 52123).

Tahdon kiittää valvojaani dos. Samuel Kaskea ja ohjaajaani dipl.ins. Janne Nikkilää heidän ohjauksestaan ja kärsivällisistä vastauksista lukuisiin kysymyksiini. Heidän antamansa arvostelu ja palaute on suuresti tukenut työtäni. Kiitos kuuluu myös diplomityöni vastaavalle professorille Jaakko Hollménille.

Lisäksi haluan kiittää prof. Eero Castrénia ja fil. maist. Juha Knuuttilaa Helsingin yliopiston neurotieteen tutkimuskeskuksesta, sekä fil. maist. Petri Töröstä Kuopion yliopiston A. I. Virtanen-instituutista avusta työn biologisten näkökulmien tulkinnassa.

Tutkimusryhmämme jäsenten kanssa käydyt keskustelut ovat auttaneet minua hahmottamaan tutkimuskentän laajemmin, ja kannustavan ilmapiirin ansiosta työnteko on ollut antoisaa. Olen kiitollinen tutkimusyksikön henkilökunnalle mahdollisuudesta tämän mielenkiintoisen ja haastavan työn tekemiseen erinomaisissa puitteissa merinäköalalla.

Lopuksi tahdon kiittää vanhempiani, joiden rakkauden ansiosta minulla on mahdollisuus puuhailla tällaisia mielenkiintoisia juttuja tässä maailmassa.

Helsingissä 4. joulukuuta 2003

Leo Lahti

Sisältö

Luku 1

Johdanto

Useiden organismien kaikkien geenien eli *genomin* rakenteen kartoittamisen [2, 4, 12, 28, 42, 47, 65] jälkeen geenien tehtävän ja toiminnallisten yhteyksien selvittämiseksi on avautunut uudenlaisia mahdollisuuksia. Vertaileva toiminnallinen genomiikka tutkii lajien yhteyksiä geenien toiminnan tasolla genominlaajuisessa mittakaavassa [20, 41]. Tulevaisuuden haasteena on geenien biologisen merkityksen selvittäminen ja uuden tiedon hyödyntäminen lääketieteellisten ongelmien ratkaisussa.

Helpommin tutkittavien malliorganismien kuten hiiren avulla tehtävä tutkimus voi auttaa ymmärtämään myös ihmisen geenien toimintaa. Malliorganismien tutkimuksesta saatava lisätieto on arvokasta, sillä geenien toiminnan selvittäminen on valtava urakka, jossa tarvitaan monitieteistä yhteistyötä.

Uusi geenisirutekniikka [44, 54, 68] mahdollistaa jopa kymmenien tuhansien geenien ilmenemistasojen mittaamisen samanaikaisesti yhdessä näytteessä, kun aiemmin kyettiin tutkimaan vain muutamaa geeniä kerrallaan. Uudet menetelmät tuottavat suuria määriä dataa, johon tärkeä tieto geenien toiminnasta kätkeytyy [1, 21]. Laajojen datajoukkojen matemaattisesta mallintamisesta on tullut tärkeä geenitutkimuksen työväline (ks. [46, 69]).

Tutkimusmenetelmien valinnassa on huomioitava käytettävissä olevat voimavarat. Tavallisesti kannattaa aloittaa laskennallisesti kevyemmillä ja rakenteeltaan yksinkertaisilla eksploratiivisilla menetelmillä, jotka eivät vaadi ennakkotietoa tutkimusaineistosta. Eksploratiivisten menetelmien tavoitteena on tutkimusaineiston yleisten piirteiden hahmottaminen ja hypoteesien muodostus myöhemmän tutkimuksen tarpeisiin. Monimutkaisemmat menetelmät kuvaavat yleensä tutkittavaa ilmiötä paremmin, mutta ovat laskennallisesti raskaita ja voivat vaatia työlästä sovittamista tutkimuskohteeseen.

Erilaiset ryhmittelymenetelmät (ks. [49]) ja riippuvuuksien mallintaminen ovat yleisiä eksploratiivisen tutkimuksen välineitä. Ryhmittelymenetelmillä voidaan etsiä geeniryhmiä, joiden geeneillä on yhteisiä piirteitä. Yhteys voi liittyä esimerkiksi samankaltaiseen toimintaan tai toimintaan osana laajempaa geenejä yhdistävää verkostoa. Kunkin

geeniryhmän biologisesta merkityksestä voidaan muodostettujen ryhmien avulla tehdä epäsuoria päätelmiä, jos kyseiseen ryhmään kuuluu myös ennestään tunnettuja ja toiminnaltaan hyvin kartoitettuja geenejä. Ryhmittelymenetelmien mahdollisuudet geenien toiminnan tutkimuksessa huomattiin alunperin hiivalla tehdyissä kokeissa. Tiettyihin aineenvaihduntareitteihin liittyvät geenit päätyivät samoihin ryhmiin [15, 18], minkä seurauksena ehdotettiin, että ryhmittelyä voitaisiin käyttää geenien toiminnan ennustamiseen. Geenien toimintaa koskevia hypoteeseja löydetään ryhmittelyn avulla huomattavasti nopeammin kuin uusia kytköksiä perinteisillä laboratoriomenetelmillä. Riippuvuuksien mallintaminen auttaa ymmärtämään tutkimusaineistojen yhteyksiä. Riippuvuuksia voidaan mallintaa mm. korrelaatiomitoilla ja yhteisesiintymämalleilla (ks. esim. [62, 59]).

Genominlaajuisen ilmenemisdatan vertailevaan tutkimukseen soveltuvia menetelmiä on kehitelty toistaiseksi vähän. Tämän työn tavoitteena on etsiä biologisen tutkimuksen kannalta kiinnostavia ihmisen ja hiiren vastingeenien ryhmiä käyttämällä ryhmittelyyn genominlaajuista aineistoa vastingeenien ilmenemisestä. Vastingeeneille muodostetaan nk. assosiatiivinen ryhmittely, joka maksimoi ihmisen ja hiiren geeniryhmien riippuvuudet Bayesilaisen riippuvuusmitan mielessä. Tällä tavalla löydetään vastingeenejä, joiden toiminnallinen yhteys on muuhun aineistoon verrattuna poikkeuksellinen. Assosiatiivinen ryhmittely [59] on uusi eksploratiivisen data-analyysin menetelmä, joka soveltuu kahden aineiston riippuvuuksien mallintamiseen ilman ennakkotietämystä niiden rakenteesta. Tämän työn toisena tavoitteena on uuden menetelmän arviointi ja vertaaminen vaihtoehtoisiin menetelmiin.

Assosiatiivinen ryhmittely osoittautui vertailevassa toiminnallisessa genomianalyysissa käyttökelpoiseksi eksploratiiviseksi menetelmäksi. Se onnistuu yhdistämään kaksi hyvää ominaisuutta, helppotulkintaisten ryhmien muodostamisen ja riippuvuuksien mallintamisen, vaihtoehtoisia menetelmiä paremmin, ja sen muodostamat ryhmät yleistyvät kuvaamaan myös potentiaalisia uusia näytteitä. Tutkimuksessa löydettiin mielenkiintoisia vastingeenien ryhmiä, joille on olemassa biologinen tulkinta. Tämän lisäksi voitiin muodostaa uusia hypoteeseja geenien yhteyksistä sekä lajien välillä että niiden sisällä. Löydöksiä voidaan käyttää lähtökohtana biologisille jatkotutkimuksille.

Työn biologinen tausta, geenien ilmenemisen mittaamiseen käytettävät menetelmät ja tutkimuksessa tarpeelliset tietokannat esitellään luvussa 2. Luvussa 3 esitellään tutkimuksen tavoitteet ja tutkimusaineisto, sekä luodaan katsaus vertailevan toiminnallisen genomiikan tutkimusmenetelmiin. Luvussa 4 esitellään menetelmien ymmärtämisen kannalta oleelliset taustatiedot ja assosiatiiviselle ryhmittelylle vaihtoehtoiset menetelmät. Assosiatiivista ryhmittelyä käsittelevässä luvussa 5 esitellään tämän työn ensisijainen tutkimusmenetelmä. Kokeiden tulokset ja tulosten tulkinnat esitellään luvussa 6. Viimeisessä luvussa 7 työstä tehdään yhteenveto.

Luku 2

Biologisen taustan esittely

Tieto ihmisen ja muiden eliöiden rakenteesta ja elintoimintojen säätelystä sisältyy solujen perimäainekseen DNA:han ja sen sisältämiin geeneihin. Valtaosa perimäaineksesta sijaitsee solujen sisällä kromosomeissa, joissa jokaisella geenillä on hyvin määritelty sijainti. Genominkartoitusprojektien tehtävänä on DNA:n rakenteen selvittäminen kokonaisuudessaan. Haave on toteutunut jo lukuisten eliöiden kohdalla, mutta yksittäisten geenien tehtävä on edelleen usein tuntematon.

Genomien kartoittamisen jälkeisen ajan haasteena on geenien tehtävän selvittäminen ja lajien välisten yhteyksien tutkimus geenien toiminnan tasolla. Geenien toiminnan ymmärtämisestä on apua mm. lajien kehityshistorian tutkimuksessa, sairauksien diagnosoinnissa ja elimistön toiminnan ymmärtämisessä. Geenien toiminnasta ja yhteyksistä voidaan saada arvokkaita vihjeitä analysoimalla geenien ilmenemistä erilaisissa olosuhteissa ja hyödyntämällä aiempien tutkimusten tuloksena syntyneitä laajoja tietokantoja.

Yhteisestä kehityshistoriallisesta alkuperästä johtuen hyvinkin erilaisilla eliöillä on runsaasti samankaltaisen rakenteen omaavia geenejä, jotka liittyvät usein samoihin tehtäviin. Tämän ansiosta esimerkiksi ihmisen geenien toimintaan liittyvää tietoa voidaan saada tutkimalla geenien toimintaa malliorganismeissa kuten hiiressä tai hiivassa.

2.1 Geenit ja proteiinisynteesi

Geenit muodostuvat DNA:sta, joka on nukleotideista muodostuva pitkä kaksijuosteinen molekyyliketju. Nukleotideja on neljää erilaista tyyppiä, jotka ovat adeniini (A), tymiini (T), guaniini (G) ja sytosiini (C). Niiden järjestys määrää geenin sisältämän informaation. Nukleotidit sitoutuvat DNA:n kaksoisjuosteessa toisiinsa vastakkaisina pareina siten, että vain parit A-T ja G-C ovat mahdollisia. DNA muodostuu kahdesta tällä tavoin toisilleen komplementaarisesta vastinjuosteesta.

Geeneissä oleva tieto vaikuttaa solun toimintaan proteiinien valmistuksen eli proteiini-

synteesin kautta (ks. kuva 2.1). Proteiinit ovat avainasemassa solunsisäisissä biokemiallisissa prosesseissa, ja niiden määrää ja laatua säätelemällä DNA ohjaa koko solun toimintaa. Proteiinisynteesi alkaa DNA:n kopioimisella. DNA:n kaksoiskierre avautuu geenin alkukohdasta, ja toinen nukleotidijuosteista kopioidaan lähetti-RNA:n esiasteeksi. Yksijuosteinen lähetti-RNA muodostuu nukleotideista kuten DNA, mutta siinä tymiinin korvaa urasiili (U).

Aitotumallisilla eliöillä vain osa lähetti-RNA:n sisältämästä informaatiosta hyödynnetään proteiinin valmistuksessa. Tarpeettomat osat poistetaan nukleotidiketjusta RNA:n silmukoinnissa. Valmis lähetti-RNA siirtyy ulos solun tumasta ja kiinnittyy solulimassa sijaitsevaan ribosomiin, jossa proteiini valmistetaan lähetti-RNA:n sisältämän informaation perusteella. Ribosomi siirtyy pitkin lähetti-RNA-juostetta sitä mukaa, kun siirtäjä-RNA tuo paikalle lähetti-RNA:n määräämiä aminohappoja. Valmis aminohappoketju laskostuu kolmiulotteiseksi rakenteeksi muodostaen lopullisen proteiinin [13].

2.2 Geenien ilmenemisen mittaaminen

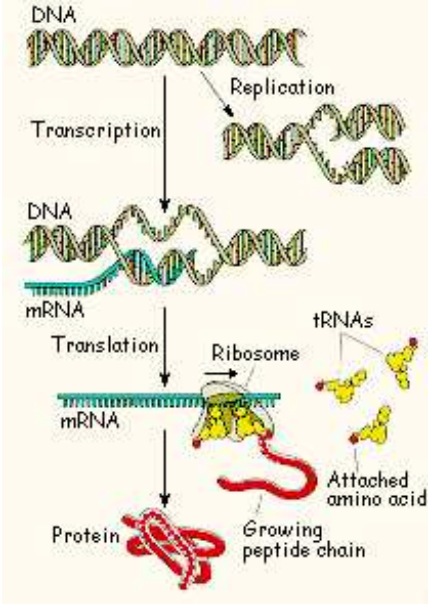
Geeniä vastaavan lähetti-RNA:n pitoisuus eli *ilmenemistaso* osoittaa geenin aktiivisuuden tutkittavassa näytteessä. Geenit toimivat eri tavoin elimistön eri osissa, ja tavallisesti vain pieni osa kaikista geeneistä ilmenee kerrallaan yksittäisessä näytteessä. Ilmenemistasojen voimakkuudet ja ajoitus ohjaavat solubiologisia prosesseja geenien toimiessa yhteistyössä monimutkaisina säätelyverkostoina.

Geenien toiminnan kartoittamiseksi niiden ilmenemistä mitataan lukuisissa olosuhteissa, kuten erilaisissa kudosympäristöissä tai aikasarjoina. Näin saadaan selville, minkälaisissa tilanteissa kukin geeni aktivoituu. Geenisiruanalyysin avulla tietoa saadaan sekä erilaisten geenien ilmenemisestä yksittäisessä näytteessä että yksittäisen geenin ilmenemisestä erilaisissa olosuhteissa. Ilmenemistasojen vertailu muodostaa pohjan laajoille analyyseille geenien toiminnasta. *Geenilastulla* voidaan mitata tuhansien geenien ilmenemistä samanaikaisesti annetussa näytteessä. Eri olosuhteissa mitatut yksittäisen geenin ilmenemistasot muodostavat *ilmenemisprofiilin* (kuva 2.2). Kukin ilmenemisprofiilin komponentti vastaa geenin ilmenemistasoa yhdessä mittausolosuhteessa, ja geeni voidaan esittää pisteenä eri mittausolosuhteita kuvaavassa data-avaruudessa.

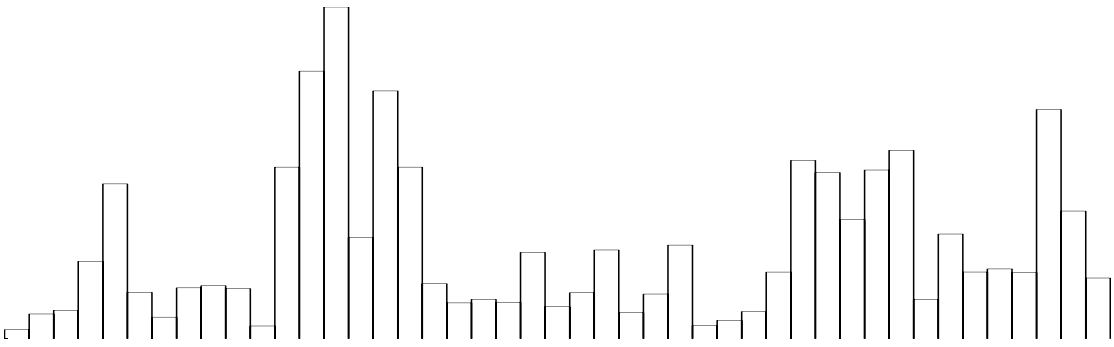
Geenisiruanalyysi on uusi ja käyttökelpoinen väline biologisten ongelmien ymmärtämisessä kokonaisvaltaisella tavalla. Geenisirut ovat parhaimmillaan seulontatutkimuksissa, esimerkiksi tuotettaessa uusia hypoteeseja jatkotutkimusta varten. Keskeiset tulokset on varmistettava toisilla menetelmillä ennen lopullisten päätelmien tekoa. Geenisiruanalyysin toteutuksessa on lukuisia vaiheita.

2.2.1 Geenilastu

Lähetti-RNA:n ilmenemistasojen mittaamiseksi näytteestä on kehitelty erilaisia geenisirumenetelmiä [44, 54, 68]. Affymetrixin kehittämä *geenilastu* [44] on yksi yleisimmin



Kuva 2.1: Proteiinisynteesin periaate. Lajin perimän sisältävä kaksoisjuosteinen DNA-molekyyli avautuu ja tarpeellinen osa kopioituu lähetti-RNA-molekyyliksi (mRNA). Tätä kutsutaan DNA:n transkriptioksi. Translaatiossa muodostetaan lähetti-RNA:n sisältämän tiedon perusteella aminohappoketju ribosomien ja siirtäjä-RNA:n (tRNA) avulla. Valmis aminohappoketju laskostuu kolmiulotteiseksi rakenteeksi muodostaen proteiinin. (Kuva on peräisin sivulta `http://faculty.virginia.edu/bio201/Home201.htm`)



Kuva 2.2: Ilmenemisprofiililla tarkoitetaan geenin ilmenemistasojen histogrammia. Kukin pylväs vastaa yhtä mittausolosuhdetta, joita voivat olla esimerkiksi erilaiset kudokset tai eri ajanhetkinä tehdyt mittaukset. Pylvään korkeus ilmoittaa geenin ilmenemistason kyseisessä mittauksessa.

käytetyistä ja luotettavimmista menetelmistä.

Koettimien valinta ja valmistus

Lähetti-RNA voi yksijuosteisena sitoutua komplementaariseen vastinjuosteeseensa. Geenisirumenetelmät hyödyntävät tätä ominaisuutta lähetti-RNA-molekyylien pitoisuuden mittauksessa. Geenilastulla kutakin mitattavaa lähetti-RNA-molekyyliä edustaa koetinjoukko, johon kuuluu tavallisesti 11-20 erilaista mittaukseen käytettävää nukleotidijuosteiden paria. Jokainen pari sisältää lyhyen mitattavalle geenille komplementaarisen, noin 25 nukleotidin mittaisen edustavan vastinjuosteen ja hieman virheellisen kontrollijuosteen. Nämä *oligonukleotidikoettimet* rakennetaan geenilastulle nukleotidi kerrallaan sekvenssitietokantojen, kuten GenBankin [9] sisältämien sekvenssitietojen perusteella. Kutakin tutkittavaa geeniä vastaavat oligonukleotidikoettimet valitaan tiettyjen edustavuuskriteerien ja muodostussääntöjen mukaisesti geenin lähetti-RNA:n eri osille. Yhdellä geenilastulla voi olla tuhansia koetinjoukkoja, joista kukin mittaa yhden geenin ilmenemistä tutkittavassa näytteessä.

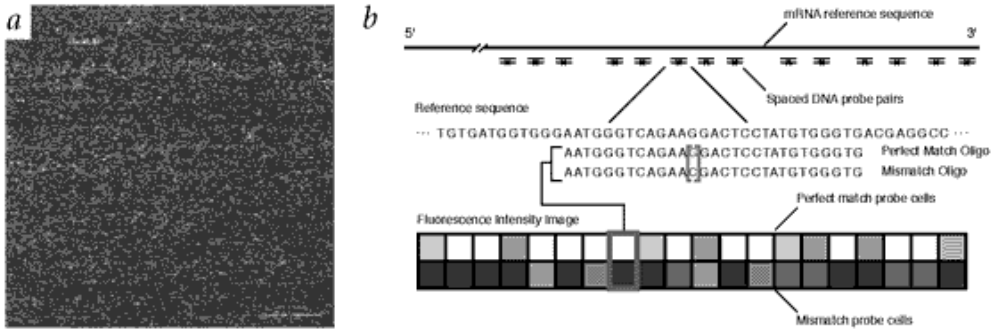
Tutkimusnäytteen käsittely ja hybridisointi

Tutkimusnäyte on käsiteltävä ennen geenilastumittausten tekemistä. Tutkittavasta näytteestä eristetyt lähetti-RNA-molekyylit muunnetaan ensin kaksijuosteisiksi hyödyntämällä molekyylien sitoutumiskykyä komplementaaristen nukleotidien kanssa. Kaksoisjuoste avataan värjäysmateriaalin läsnäollessa, jolloin värjäysmolekyylit sitoutuvat syntyviin RNA-juosteisiin. Värjäysmateriaalina voidaan käyttää esimerkiksi fluoresoivia molekyylejä. Syntyvät pitkät RNA-juosteet pilkotaan pienempiin osiin. Menettely vähentää RNA:n toisiorakenteen eli erilaisten laskosten ja sidosten vaikutusta tulevissa mittauksissa. Lopulta värjättyjen RNA-palojen seos voidaan asettaa geenilastulle, jolloin mitattavaa geeniä vastaavien RNA-molekyylien palaset sitoutuvat lastulla sijaitseviin komplementaarisiin vastinjuosteisiinsa. Tapahtumaa kutsutaan *hybridisaatioksi*. On odotettavissa, että mitattavaa geeniä vastaava RNA sitoutuu todellisiin vastinjuosteisiinsa paremmin kuin virheellisiin kontrollijuosteisiin.

Fluoresoivat molekyylit tuottavat valosäteilyä, jonka kokonaisintensiteetti riippuu fluoresoivien molekyylien määrästä. Mitä enemmän värjättyjä lähetti-RNA-molekyylejä sitoutuu koetinjoukolle, sitä suurempi on myös koetinjoukolta mitattava valosäteilyn intensiteetti. Intensiteettitasoja mittaamalla voidaan siten arvioida kullekin koettimelle sitoutuneiden RNA-juosteiden määrää (kuva 2.3).

Ilmenemistasojen mittaukset ja esikäsittely

Hybridisaation jälkeen geenilastusta otetaan värikuva skannerilla, joka on herkkä fluoresoiville väriaineille ja mittaa tarkasti kaikkien koettimien intensiteettiarvot. Alustava



Kuva 2.3: Geenien ilmenemisdatan tuottaminen geenilastulla: **(a)** Kutakin tutkittavaa geeniä vastaa geenilastulle sijoitettu koetinjoukko. **(b)** Tutkittavaa geeniä vastaavat oligonukleotidikoettimet valitaan lähetti-RNA:n eri osista. Tavallisesti kutakin geeniä vastaa geenilastulla 11-20 erilaista noin 25 nukleotidin mittaista koetintyyppiä ('perfect match'). Jokaisen koettimen parina on hieman virheellinen kontrollijuoste ('mismatch'). Kun tutkittavasta näytteestä eristetyt pilkotut ja värjätyt lähetti-RNA-juosteet asetetaan geenilastulle, ne sitoutuvat RNA-juosteitansa vastaaville komplementaarisille koettimille. Koettimille ja kontrollijuosteille sitoutuneiden lähetti-RNA-molekyylien pitoisuuksia verrataan, ja erilaisia koettimia vastaavista mittaustuloksista lasketaan keskiarvo. Geenin ilmenemistaso saadaan laskettua tämän avulla tiettyjen esikäsittelyjen jälkeen. Menettely vähentää huomattavasti taustakohinan vaikutusta mittaustuloksiin, mikä parantaa mittausten kvantitatiivista tarkkuutta ja toistettavuutta. Kuva on peräisin artikkelista [43].

kuva-analyysi tehdään tavallisesti Affymetrixin ohjelmistolla (Genechip 3.2), jonka yksityiskohdat ovat liikesalaisuuksia. Esikäsittelyllä vähennetään erilaisten biokemiallisten ilmiöiden aiheuttamaa epäoleellista taustakohinaa ja muokataan data helpommin käsiteltävään muotoon. Oikeaan vastinjuosteeseen sitoutuneiden molekyylien intensiteettiä merkitään symbolilla PM ('perfect match') ja virheellistä kontollijuostetta vastaavaa intensiteettiä symbolilla MM ('mismatch'). Geenikohtainen mittausarvo yksittäisen geenin lähetti-RNA:n ilmenemiselle saadaan laskemalla kyseistä geeniä vastaavan koetinjoukon mittaustulosten keskiarvo PM- ja MM-koettimissa. Mittaustuloksena käytetään näiden logaritmista suhdetta. Lopullinen geenin ilmenemisen voimakkuutta osoittava arvo eli *ilmenemistaso* ('average difference', AD) on verrannollinen tarkasteltavan geenin lähetti-RNA:n pitoisuuteen tutkittavassa näytteessä. Yksittäinen geenilastu voi sisältää koetinjoukkoja tuhansien geenien ilmenemisen samanaikaista mittaamista varten.

Geenilastun ominaisuuksista

Geenin eri osista valittavien edustavien vastinjuosteiden ja kontrollijuosteiden sekä mittausten keskiarvoistamisen ansiosta geenilastumenetelmällä on joitakin etuja muihin menetelmiin verrattuna. Geenilastut ovat vähemmän alttiita virheille ja niillä saadaan näkymään pienempiä muutoksia geenien ilmenemistasoissa. Geenilastuilla voi kuitenkin mitata vain yhden näytteen kerrallaan. Tämän seurauksena tutkittavan näytteen ja kont-

rollinäytteen vertaaminen on vaikeampaa kuin esimerkiksi cDNA-menetelmällä, jossa eri näytteisiin liittyvät lähetti-RNA-molekyylit hybridisoidaan samalle koettimelle. Eri näytteisiin liittyvät lähetti-RNA-molekyylit voidaan havaita näissä mittauksissa erilaisten värjäysmateriaalien ansiosta.

Geenilastujen valmistaminen yksittäistä tutkimusta varten on hankalaa ja kallista, ja tarjolla onkin pääasiassa joukko standardilastuja. Kaupallisesti valmistettavien geenilastujen monet yksityiskohdat ovat liikesalaisuuksia, ja tutkijat joutuvat luottamaan yrityksen julkaiseman informaation käyttökelpoisuuteen. Geenilastuilla tuotetun ilmenemisdatan suosio on kasvanut muihin geenisirumenetelmiin verrattuna, mutta toistaiseksi menetelmän korkeat kustannukset ovat rajoittaneet sen käyttöä.

2.2.2 Suhde muihin geenisirumenetelmiin

Muiden geenisirumenetelmien itsenäinen valmistus on geenilastuihin verrattuna helpompaa ja huokeampaa. Tällöin geenisiru voidaan myös paremmin räätälöidä tiettyyn tarkoitukseen. Muut menetelmät eivät kuitenkaan ole geenilastuun verrattuna yhtä herkkiä havaitsemaan pieniä ilmenemistasojen vaihteluita, ne ovat alttiimpia virheille ja niillä tuotettu data on kohinaisempaa.

Eri geenisirumenetelmillä tehtyjen mittausten vertailu tiedeyhteisössä on hankalaa, koska menettelytavat eivät ole vielä vakiintuneet yrityksistä [11] huolimatta.

2.2.3 Ilmenemisdatan laatu

Geenisiruanalyysissä on lukuisia vaiheita, joista kaikki aiheuttavat dataan kohinaa ja mahdollisia virheitä jo ennen varsinaisen analyysin aloittamista. Koetinjoukko voi olla huonosti valittu tai RNA-molekyylien kiinnittyminen geenilastulle epätasaista. Myös intensiteettitasojen mittaus, kuvankäsittely ja mittaustulosten esikäsittely voivat lisätä mittaustuloksiin virheitä ja kohinaa.

Usein geenien ilmenemistä analysoitaessa keskitytään niihin kudoksiin, jotka ilmenevät poikkeavasti. Rajaksi voidaan asettaa esimerkiksi muihin kudoksiin verrattuna kaksinkertainen ilmenemistaso. Sopivan rajan määrittely on kuitenkin hankalaa. Joidenkin geenien kohdalla pienikin ero voi olla merkitsevä, kun taas toisilla geeneillä voimakas ilmenemistasojen vaihtelu on tavallista, ja merkittävää olisi vasta moninkertainen ilmenemistasojen ero. Merkitsevyydestä ja geenin tavanomaisesta käyttäytymisestä saisi kuvan toistokokeilla, jolloin voitaisiin muodostaa kunnollinen virhemalli mitatulle datalle. Mitä useampi toisto on käytettävissä sitä paremmin kohinaiset mittaukset voidaan erotella muista näytteistä. Tutkimusnäytteiden saatavuus voi kuitenkin olla rajoitettu, ja luotettavien geenisirukokeiden tekeminen on kallista. Tavallisesti mittauksia tehdäänkin yksittäiseen näytteeseen liittyen vain muutamia, jolloin virherajat jäävät suuriksi. Geenilastua pidetään yhtenä luotettavimmista menetelmistä, ja lastuilla tuotettu data on osoittautunut käyttökelpoiseksi lukuisissa tutkimuksissa.

Ilmenemisprofiilien tulkinta ei ole suoraviivaista, sillä ensisijaisia vaikutuksia on usein hankalaa erottaa toissijaisista muutoksista ja taustakohinasta. Esikäsittelymenetelmien (ks. esim. [50]) avulla eri näytteisiin liittyvät mittaukset pyritään saamaan vertailukelpoisiksi, vähentämään kohinan vaikutusta ja karsimaan aineistosta tutkimuksen kannalta epäoleellista tietoa, joka voi hidastaa ja monimutkaistaa analyysia.

2.3 Geenien ortologia

Lajeilla on yhteinen kehityshistoriallinen alkuperä. Evoluution aikana lajit ovat eriytyneet, ja niiden perimän rakenne on muuttunut. Eri lajien perimässä on kuitenkin säilynyt yhteisiä piirteitä. Geenien nukleotidisekvenssien vahva samankaltaisuus voi olla merkki niiden yhteisestä kehityshistoriallisesta alkuperästä eli geenien *ortologiasta*. Eri lajien geenejä, joilla on yhteinen rakenne ja alkuperä, kutsutaan *vastingeeneiksi*. Niiden tehtävä on usein säilynyt samankaltaisena, mutta joissakin tapauksissa tehtävät ovat eriytyneet. Vastingeenejä voidaan etsiä biologisissa tutkimuksissa, mutta myös laskennallisesti tarkastelemalla kahden lajin geenien sekvenssihomologiaa eli geenien nukleotidisekvenssien samankaltaisuutta.

Lajien väliseen geenisekvenssien rinnastukseen on kehitetty erilaisia laskennallisia menetelmiä [45, 74]. Tässä työssä ihmisen ja hiiren vastingeenien selvittämiseen käytettiin HomoloGene-tietokannassa (ks. [72]) olevaa tietoa *putatiivisesti ortologisista* geenipareista. Rinnastusmenetelmät ovat epäsymmetrisiä; niiden tuottamien rinnastusten paremmuusjärjestys voi vaihdella riippuen siitä, kumman lajin geenisekvenssiä käytetään lähtökohtana etsittäessä sille parhaiten rinnastuvia sekvenssejä toisen lajin genomista. Putatiiviset ortologit ovat geenipareja, joiden rinnastus on kyseisille geeneille laskennallisesti paras molemmista lajeista käsin tarkasteltuna. Satunnaisia yhteyksiä karsivana lisäehtona rinnastettavilta sekvensseiltä vaaditaan vähintään 100 nukleotidin onnistunutta rinnastusta.

Varmaa tietoa ortologien kehityshistoriallisista yhteyksistä ei ole käytettävissä, joten virheellisten geenirinnastusten vaikutusta tutkimustuloksiin on hankalaa arvioida. Nykyisiä rinnastusmenetelmiä pidetään kuitenkin luotettavina.

2.4 Hyödynnettävät tietokannat

Geenitutkimus tuottaa valtavan määrän tietoa geenien rakenteesta, toiminnasta ja yhteyksistä muihin geeneihin. Jotta tutkimustietoa voitaisiin hyödyntää tehokkaasti, näiden tietojen järjestäminen on toteutettava niin, että tietojen nopea hakeminen on mahdollista. Tarkoituksiin on kehitetty erilaisia tietokantoja, joiden käyttö geenitutkimuksessa on rutiininomaista. Alla esitellään tässä tutkimuksessa käytetyt tietokannat.

GenBank. GenBank [8] on suuri DNA-sekvenssejä sisältävä tietokanta, jossa jokaiselle julkisesti saatavilla olevalle, yksittäisten laboratorioiden tai suurten sekvensointiprojektien kartoittamille DNA-sekvenssille ja niiden eri versioille on annettu oma tunnuksensa. GenBank-tunnusten avulla tutkimuksissa voidaan tarvittaessa viitata tiettyyn sekvenssiin yksikäsitteisesti. Geenit muodostavat vain osan perimäaineksesta, ja GenBank sisältääkin tietoja myös muista DNA-sekvensseistä. GenBank on arkistotietokanta, jossa yksittäisestä sekvenssistä voi olla runsaasti tietoa. Alkuperäinen ilmoittaja omistaa oikeudet luovuttamiinsa tietoihin, eikä niitä voida muuttaa.

RefSeq. Yhdysvaltalaisen National Institute for Biotechnology Information (NCBI)-organisaation 'Reference Sequence'-projekti (ks. [48]) on tuottanut tiivistä ja tutkittua tietoa tällä hetkellä tunnetuista geeneistä lukuisille organismeille. Tietoa voidaan hyödyntää esimerkiksi genomien lääketieteellisissä, toiminnallisissa ja vertailevissa tutkimuksissa. Kuhunkin projektin kirjaamaan geeniin liittyy geenin DNA-sekvenssiä hyvin edustava viitesekvenssi, joka johdetaan GenBank-arkiston sisältämistä sekvenssitiedoista.

LocusLink. LocusLink-tietokanta (ks. [48]) sisältää geenien sijaintitietoja, ja sen avulla voidaan selvittää annetun geenin täsmällinen paikka kromosomistossa. Kutakin geeniä vastaa tietokannassa yksikäsitteinen LocusID-tunnus. LocusLink sisältää myös lukuisan joukon muita geeneihin liittyviä tietoja, ja tarjoaa sijaintitietoja esimerkiksi ihmisen ja hiiren geenejä vastaaville viitesekvensseille.

HomoloGene. HomoloGene [72] on tietokanta, joka sisältää tietoja eri lajien geneettisistä yhteyksistä. Tietokannan sisältämät laskennalliset geenirinnastukset on tehty ihmiselle ja hiirelle käyttämällä 'discontiguous MegaBLAST'-algoritmia [74]. Menetelmä etsii kahdesta organismista ne geeniparit, joiden nukleotidisekvenssit ovat samankaltaisimpia. Sekvenssien samankaltaisuus määritellään monimutkaisella pisteytysmenetelmällä. Ihmisen ja hiiren geenien rinnastukset on tehty RefSeq-tietokannan sisältämien viitesekvenssien nojalla.

EASE. Geenien toiminnalliseen tutkimukseen soveltuvien data-analyysimenetelmien tavanomainen tulos on yksi tai usempia geenilistoja. Usein on tarpeen selvittää listan geenejä yhdistävät ominaisuudet ja muodostaa pikainen kokonaiskuva aineistosta. Tarkoitukseen kehitetty EASE-ohjelmisto [35] etsii annetun listan geeneihin liittyviä tilastollisesti merkityksellisiä teemoja, kuten osallisuutta samoihin aineenvaihduntareitteihin tai molekulaarisiin funktioihin. Tämän lisäksi ohjelmisto tarjoaa geenikohtaisen linkityksen muihin käyttökelpoisiin verkossa toimiviin työvälineisiin ja luo kuvailevia taulukoita. Ohjelmistosta on olemassa vapaasti käytettävä vuorovaikutteinen verkkosovellus osoitteessa `http://david.niaid.nih.gov/david/`.

Luku 3

Ihmisen ja hiiren vastingeenien tarkastelu

Ihminen ja hiiri eriytyivät yhteisestä kantamuodostaan noin 75 miljoonaa vuotta sitten. Noin 80 prosentilla hiiren geeneistä on sekvenssihomologian nojalla yksikäsitteinen vastingeeni ihmisen genomissa, ja ainoastaan yhdellä prosentilla hiiren geeneistä ei ole tällä hetkellä havaittavaa sekvenssihomologiaa ihmisen geenien kanssa [47]. Ihmisen ja hiiren proteiineja koodaavien geenien määräksi arvioidaan yleisesti noin 30 000, mutta arvioiden vaihteluväli on melko suuri. Geenien lisäksi perimässä on lukuisa joukko muita elimistön toiminnan säätelyn kannalta tarpeellisia osia.

Lajien geneettisten yhteyksien ansiosta ihmisen geenien toiminnasta voidaan saada tietoa tutkimalla geenien toimintaa malliorganismeissa kuten hiiressä. Hiiren käyttöä malliorganismina puoltavat vuosisadan ajan jatkuneen geneettisen tutkimuksen tuottama tieto, sekä mahdollisuudet hiiren perimän manipulointiin ja koeolosuhteiden kontrollointiin. Vastingeenien tutkimista voi käyttää lähtökohtana lajien vertailulle geneettisellä tasolla.

3.1 Vertaileva toiminnallinen genomianalyysi

Ihmisen elimistön toiminnasta on saatu lisätietoa jo pitkään tutkimalla fysiologisten systeemien toimintaa malliorganismeissa. Vertailun ulottaminen geenien toiminnan tasolle genominlaajuisessa mittakaavaassa on uusi tutkimusalue. Tämä *vertaileva toiminnallinen genomiikka* on erotettava vertailevasta genomiikasta, jossa vertailu kohdistuu geenien toiminnan sijasta genomien rakenteisiin (ks. esim. [47]).

Lajien sisäisillä tarkasteluilla löydettyjen geeniryhmien toiminnallinen vertailu lajien välillä on ensiaskel geenien toiminnan vertailevaan tutkimukseen. Toiminnallisesti samankaltaisten geenien ryhmiä on etsitty sekvenssihomologian perusteella mm. hiivalajeissa [26]. Sekvenssitiedon ohella vertailuissa on käytetty ortologisten geenien ilmenemisdataa [36]. Pelkän geenisirudatan perusteella muodostettuja lajien sisäisiä geeniryhmiä on ver-

tailtu ainakin hiiressä ja hiivassa [10]. Genomisessa mittakaavassa geenisiruilla tuotettua dataa on käytetty geenien luokitteluun perustuvissa tutkimuksissa myös yksittäisen lajin eri kantojen vertailuun bakteereilla ja hiivalla [6, 38, 55, 66] sekä taudinaiheuttajan isäntälajissa aiheuttamien infektioiden tutkimiseen [19]. Vastingeenien ilmenemisdatan tutkimuksessa on löydetty viitteitä siitä, että lajien säilymiselle tärkeiden vastingeenien ilmenemisessä tapahtuu evoluution aikana vähemmän eriytymistä kuin muiden vastingeenien ilmenemisessä [61].

Vertailevaa toiminnallista analyysia on tehty geenisekvenssien ja ilmenemisdatan lisäksi myös muilla genomisilla datajoukoilla. Esimerkkejä ovat aineenvaihduntaverkostojen analysointi singulaariarvohajotelmalla [24] ja proteiinien vuorovaikutusdatan tutkiminen valmiilla visualisointimenetelmillä [71]. Valmiit menetelmäpaketit ovat helppokäyttöisiä, mutta niiden varjopuolena on usein se, että menetelmän sovittaminen tutkimusaineistoon jää helposti puutteelliseksi, ja tulkinnan kannalta oleelliset tekniset yksityiskohdat saattavat jäädä loppukäyttäjälle epäselviksi.

Mainituissa tutkimuksissa on geenien ilmenemisdatan analysoimisen osalta turvauduttu tilastollisiin perusmenetelmiin kuten korrelaatiomittojen käyttöön tai tavanomaisilla ryhmittelymenetelmillä tuotettujen ryhmien vertailuun. Geenisiruilla tuotettujen laajojen datajoukkojen vertailevaan analyysiin on kehitelty toistaiseksi vähän menetelmiä, jotka yhdistävät aineistojen sisältämää tietoa jo analyysin laskennallisessa vaiheessa. Joitakin menetelmiä on kuitenkin olemassa. Yleistetyn singulaariarvohajotelman lähtökohtana ovat geenien ilmenemisdataa sisältävät matriisit $\hat{e}_1$ ja $\hat{e}_2$. Näissä on tietoa kahden eri lajin geenien ilmenemisestä yhteisissä koeasetelmissa. Matriiseille muodostettavat hajotelmat ovat muotoa $\hat{e}_1 = \hat{u}_1\hat{v}_1\hat{x}^{-1}$ ja $\hat{e}_2 = \hat{u}_2\hat{v}_2\hat{x}^{-1}$, missä $\hat{v}_1$ ja $\hat{v}_2$ ovat diagonaalisia. Kaksi datajoukkoa voidaan esittää tiivistetyssä esitysmuodossa, kun niille tehdään kytketty lineaarinen muunnos muunnosmatriisin $\hat{x}^{-1}$ esittämien kantavektoreiden määrittelemään aliavaruuteen. Singulaariarvohajotelma on perusteltu tapa kahdelle datajoukolle yhteisten piirteiden esittämiseksi, ja tuloksille on olemassa myös biologinen tulkinta. Yleistetty singulaariarvohajotelma paljastaa solubiologisten prosessien yhteyksiä lajien välillä [3], ja menetelmän avulla voidaan luokitella geenejä niihin liittyvien prosessien mukaisesti.

Biologisten datajoukkojen yhdistämiseksi ja tulkitsemiseksi on kehitelty myös rakenteeltaan monimutkaisempia malleja (ks. esim. [5, 33, 56, 67]). Rakenteellisemmat mallit kuvaavat usein tutkimuskohdetta paremmin kuin semiparametriset ryhmittelymallit, mutta niiden käyttäminen edellyttää tavallisesti huolellista sovittamista tutkimusaiheeseen ja runsasta etukäteistietoa aineistosta. Riippuvuuksia mallintavat eksploratiiviset ryhmittelymenetelmät tarjoavat näihin verrattuna keveän ja nopean vaihtoehdon. Ne eivät vaadi ennakkotietoa aineistosta, ja niitä voidaan käyttää aineiston tutkimisen ensimmäisenä vaiheena, jossa tavoitteena on aineiston huomionarvoisten piirteiden havaitseminen ja tarkempien tutkimussuuntien ehdottaminen.

3.2 Tutkimuksen tavoitteet

Tämän työn ensisijaisena tavoitteena on etsiä genominlaajuisesta ihmisen ja hiiren vastingeenien joukosta sellaisia vastingeenien ryhmiä, jotka voisivat liittyä biologisten jatkotutkimusten kannalta mielenkiintoisiin toiminnallisiin eroihin ja yhteyksiin lajien välillä. Tutkimus laajentaa artikkelin [62] tuloksia vastingeenien ja lajien välisten toiminnallisten yhteyksien tarkastelun suuntaan.

Yhteistä alkuperää olevilla lajin säilymisen kannalta tärkeillä vastingeeneillä on usein samankaltaisena säilynyt ilmenemiskäyttäytyminen, joka heijastelee lajeille yhteisiä fysiologisia toimintoja (ks. [61]). Geenien tutkiminen ryhminä on lajien vertailussa usein mielenkiintoisempaa kuin yksittäisten geenien vertailu (ks. [14, 61]). Lajien toiminnallisiin eroihin ja yhteyksiin liittyvien kiinnostavien geeniryhmien löytämiseksi laajasta aineistosta tarvitaan kuitenkin hyvin määriteltyjä tavoitteita. Tässä työssä etsitään sellaisia vastingeenien ryhmiä, joiden ilmenemiseen liittyvä yhteys on tilastollisesti poikkeuksellisen yleinen tai harvinainen.

Tutkimuksen toteuttamiseen soveltuva assosiatiivinen ryhmittelymenetelmä [59] on melko uusi, ja sen käytännön soveltamisesta on vasta vähän kokemuksia. Tämän työn toisena tavoitteena on uuden menetelmän toiminnan vertaaminen vaihtoehtoisiin menetelmiin, ja sen soveltamisessa huomionarvoisten ominaisuuksien kartoittaminen ja arviointi.

3.3 Tutkimusaineisto

Tässä työssä tarkastellaan julkisesti saatavilla olevaa geenilastuilla tuotettua geenien ilmenemisdataa [53, 62]. Datajoukko on edustava otos kahdesta tavanomaista nisäkkään fysiologista tilaa esittävästä *transkriptomista*, eli lähetti-RNA-molekyylien kokonaisuudesta. Mittauksia oli saatavilla 46:sta ihmisen ja 45:stä hiiren kudoksesta tai solulinjasta. Solulinja on soluviljelmä, joka suotuisissa olosuhteissa jatkaa jakaantumistaan määräämättömän ajan.

3.3.1 Esikäsittely

Esikäsittelyn tavoitteena on tutkimusaineiston muokkaaminen helpommin käsiteltävään muotoon. Tämän tutkimuksen aineistolle esikäsittelyt on tehty kuten artikkelissa [62]. Eri kudoksissa tehtyjen mittausten vertailukelpoisuuden varmistamiseksi ilmenemistasojen keskiarvo skaalattiin jokaisessa näytteessä arvoon 200. Suurimmat ja pienimmät arvot (2%) jätettiin huomiotta normalisointivakiota määritettäessä. Ilmenemistasot on laskettu Affymetrixin ohjelmistolla, ja ne ovat verrannollisia lähetti-RNA:n pitoisuuteen tutkittavassa näytteessä. Ilmenemistason AD=200 on arvioitu merkitsevän noin 3-5 lähetti-RNA-molekyylin pitoisuutta kussakin tutkimusnäytteen solussa. Useimmissa näytteissä geenilastuilla oli tehty 2-3 mittausta. Mittausvirheiden vaikutuksen vähentämiseksi kudoskohtaisista mittauksista laskettiin keskiarvo. Hyvin pienet arvot (AD$<$20)

merkitsevät joko erittäin matalaa ilmenemistasoa tai epäluotettavaa mittausta, ja ilmenemistasoksi asetettiin näissä tapauksissa AD=20.

Luettelo tutkituista kudoksista on annettu liitteessä A. Ensimmäisenä luetelluissa 21 kudoksessa geenien ilmenemistasojen mittaukset oli tehty molemmille lajeille. Muiden kudosten osalta mittaustuloksia oli saatavilla vain toisessa lajissa.

3.3.2 Näytteiden karsinta

Esikäsittelyn jälkeen eri näytteissä mitatut ilmenemistasot ovat vertailukelpoisia. Vain osa geeneistä ilmeni merkittävästi (AD>200 ainakin yhdessä kudoksessa). Ilmenemättömät geenit hidastavat ja monimutkaistavat laskentaa, ja niiden lisäarvo tälle tutkimukselle on olematon. Tämän vuoksi ne kannattaa poistaa aineistosta. Alkuperäisessä aineistossa oli mittauksia 6684 ihmisen ja hiiren putatiivisesti ortologisen geeniparin ilmenemisestä. Ilmenemättömien geenien karsinnan jälkeen käytettävissä oli 4499 geeniparia. Myös tarkasteltavan datajoukon dimensionaalisuuden pienentäminen nopeuttaisi laskentaa, mutta toisaalta se voisi kadottaa myös hyödyllistä informaatiota. Tässä työssä ulottuvuuksien karsintaan ei ollut laskentaresurssien riittävyyden kannalta tarvetta, joten se jätettiin tekemättä.

3.3.3 Normalisointi

Erilaisten mittausolosuhteiden johdosta eri näytteissä mitattujen ilmenemistasojen hajonnat eivät välttämättä ole vertailukelpoisia. Tietoa hajonnan merkityksestä eri näytteissä ei ollut käytettävissä, joten vertailun helpottamiseksi ilmenemistasojen varianssi normalisoitiin jokaisessa näytteessä arvoon 1.

3.3.4 Etäisyysmitat

Samankaltaisesti ilmenevillä geeneillä on usein samankaltaisia tehtäviä tai niihin liittyvät yhteiset säätelymekanismit (ks. [23, 63]). Toisaalta samankaltainen ilmeneminen voi olla merkkinä geenien tuntemattomista toiminnallisista yhteyksistä. Geenien ilmenemisprofiilit voidaan esittää pisteinä data-avaruudessa, jonka koordinaattiakselit vastaavat eri kudoksissa tehtyjä mittauksia. Ilmenemisprofiilin muodosta voidaan päätellä, missä olosuhteissa geeni erityisesti ilmenee. Samoissa olosuhteissa ilmenevillä geeneillä voi olla toiminnallisia yhteyksiä, vaikka niiden absoluuttisissa ilmenemistasoissa olisi huomattaviakin eroja. Toisaalta jo pelkkä ero absoluuttisissa ilmenemistasoissa saattaa olla kiintoisa vastingeenejä vertailtaessa. Samankaltaisesti ilmenevien geenien etäisyys data-avaruudessa on pienempi kuin eri tavoin ilmenevien geenien. Jotta datapisteiden etäisyyksiä ja sijainteja voidaan mielekkäästi vertailla, tarkasteltavaan data-avaruuteen on määriteltävä tietyt matemaattiset ehdot toteuttava etäisyysmitta (ks. [52]). Etäisyysmitta vaikuttaa aineiston tarkastelussa korostuviin piirteisiin.

Tässä työssä käytetään tavanomaista euklidista etäisyysmittaa, joka huomioi sekä ilmenemisprofiilien muodon että erot absoluuttisissa ilmenemistasoissa. Perinteinen euklidinen etäisyysmitta ei välttämättä painota erilaisia ilmenemisprofiilien piirteitä biologisten tulkintojen kannalta optimaalisella tavalla, mutta sen käyttö on yksinkertaista ja johtaa mielekkäiden geeniryhmien löytymiseen.

3.4 Datan yleispiirteiden alustava kartoitus

Aineiston ominaisuuksia tutkitaan usein alustavilla ryhmittelymenetelmillä (ks. [7, 22]) ennen monimutkaisempien mallien muodostamista. Kartoittamisen yhteydessä aineistosta voidaan löytää jatkoanalyysien kannalta huomionarvoisia piirteitä, minkä jälkeen erilaisten menetelmien soveltuvuutta kyseisen datajoukon tutkimiseen on helpompi arvioida.

Tavanomainen ensiaskel laajojen ja korkeaulotteisten datajoukkojen tutkimisessa on visuaalinen tarkastelu, joka helpottaa suuren tietoaineiston hahmottamista. Aineiston karkeiden yleispiirteiden hahmotteluun soveltuvat hyvin sen dimensionaalisuutta pienentävät menetelmät, jotka pyrkivät säilyttämään mahdollisimman hyvin tutkimusaineistoon sisältyvän oleellisen informaation. Yleisesti käytettyjä menetelmiä ovat mm. pääkomponenttianalyysi ja itseorganisoituva kartta (ks. esim. [34]).

3.4.1 Erikoistuneet geenit

Tässä työssä käytetty tutkimusaineisto on peräisin artikkelista [62], jonka tavoitteena oli geenitiedon louhintamenetelmien havainnollistaminen ja geenien toiminnan ja säätelyn eksploratiivinen tutkiminen mm. hierarkkisella ryhmittelyllä. Tutkimuksessa aineiston havaittiin sisältävän lukuisia vain yhdessä kudoksessa voimakkaasti ilmenevien geenien ryhmiä. Ihmiseltä löydettiin mm. yksinomaan kiveksissä, haimassa, maksassa, istukassa ja kateenkorvassa, sekä hiirellä lisäksi maitorauhasissa, kilpirauhasessa ja sylkirauhasissa voimakkaasti ilmenevien geenien joukkoja.

Myös tätä työtä varten muodostetuilla itseorganisoituvilla kartoilla havaittiin useita hiiren ja ihmisen geenien kasaumia alueilla, jotka vastasivat tiettyyn kudokseen erikoistunutta ilmenemistä. Aineistosta havaittiin myös sellaisia geenikasaumia, joihin ei liittynyt selvää kudoskohtaista erikoistumista.

3.4.2 Ortologien yhteydet

Artikkelissa [62] ihmisen ja hiiren ortologisten geenien toimintaa vertailtiin korrelaatiomitalla, joka huomioi ilmenemisprofiilin muodon, mutta ei eroja absoluuttisessa ilmenemistasossa. Useimpien ortologien ilmenemisprofiileilla oli selvä positiivinen korrelaatio (>0.5). Voimakkaimmin korreloivien ortologien ilmenemisprofiilien visualisointi havain-

nollisti hiiren ja ihmisen vastingeenien toiminnallisia samankaltaisuuksia. Lisäksi havaittiin negatiivisesti korreloivia ortologien ilmenemisprofiileja. Tulosten nojalla ehdotettiin, että vastingeenit ilmenevät usein ainakin osittain yhteisin tavoin, ja että eri tavoin ilmenevien vastingeenien tutkimus voisi tuoda lisätietoa lajien eriytymisessä tapahtuneista muutoksista. Voimakkaasti korreloivien vastingeenien toiminta oli usein keskittynyt yhteen kudokseen kuten maksaan, amygdalaan, kiveksiin, sydämeen, munuaisiin tai pikkuaivoihin.

Tutkimme hiiren ja ihmisen ortologien ilmenemisessä esiintyviä yhteyksiä alustavasti huomioimalla vain ne kudokset, joiden osalta mittaukset oli tehty molemmille lajeille. Tällöin sekä ihmisen että hiiren geenit voitiin esittää samassa 21-ulotteisessa dataavaruudessa. Keskimäärin ortologien etäisyys oli satunnaisesti valittuihin geenipareihin verrattuna merkitsevästi pienempi ainakin sisätulometriikassa (ks. [39]) lasketulla itseorganisoituvalla kartalla.

Luku 4

Käytettävät data-analyysimenetelmät

Jopa kymmenien tuhansien geenien samanaikainen analysointi kymmenissä tai sadoissa erilaisissa näytteissä tuottaa valtavan määrän tutkimustietoa. Aineiston analysointiin käytettävien menetelmien valinta riippuu paitsi tutkimuksen tavoitteista ja tutkimusaineiston ominaisuuksista, myös käytettävissä olevista voimavaroista ja vaaditusta tarkkuudesta. Aineiston tutkiminen helpottuu, jos käytettävissä on sen rakennetta kuvaava malli. Mallin sovittamisessa pyritään etsimään sellaiset parametrit, joilla mallin kuvaamat yleiset säännönmukaisuudet kuvaavat tutkittavaa aineistoa mahdollisimman hyvin. Hyvä malli yleistyy kuvaamaan myös uusia havaintoja.

Tämän työn tavoitteena on ihmisen ja hiiren vastingeeneihin liittyvien mielenkiintoisten toiminnallisten erojen ja samankaltaisuuksien etsintä eksploratiivisen data-analyysin keinoin. Tutkimusaineistona on genominlaajuinen joukko mittauksia ortologisten geeniparien ilmenemisestä. Ihmisen ja hiiren geenien ilmenemisprofiilit esitetään kahden monimuuttuja-avaruuden datapisteinä, joiden ryhmittelyyn on olemassa lukuisia menetelmiä. Ihmisen ja hiiren geeniryhmien yhteydet esitetään kontingenssitaululla, jonka avulla voidaan arvioida myös niiden tilastollisia riippuvuuksia.

Tutkimukseen ensisijaisesti käytetty assosiatiivinen ryhmittelymenetelmä esitellään luvussa 5. Sen tavoitteena on Bayesilaisen riippuvuusmitan mielessä optimaalisen ryhmittelyn muodostaminen kahdelle geenijoukolle. Tässä luvussa esitellään menetelmien ymmärtämisen kannalta oleellinen taustatieto ja työssä käytettävät data-analyysin perusmenetelmät. Lisäksi luodaan katsaus assosiatiiviselle ryhmittelylle vaihtoehtoisiin menetelmiin.

4.1 Ohjaamaton ryhmittely

Ryhmittelymenetelmille on ominaista annetun näytejoukon jakaminen ryhmiin, joissa kuhunkin ryhmään kuuluvilla näytteillä on jotakin yhteistä ja toisaalta piirteitä, jotka erottavat ne muiden ryhmien näytteistä. Ohjaamattomassa ryhmittelyssä ryhmien muodostamiseen käytetään ainoastaan ryhmiteltävään aineistoon itseensä sisältyvää tietoa. Ryhmät muodostuvat tällöin ohjaamattomasti ryhmittelymallin ja tutkimusaineiston ominaisuuksien perusteella. Ohjaamattomat ryhmittelymenetelmät ovat käyttökelpoisia etenkin silloin, kun tutkittavaan aineistoon liittyvää ennakkotietoa ei ole käytettävissä.

Ryhmittelyn avulla löydettäviä yhteyksiä voivat olla esimerkiksi geenien toiminnalliset luokat, osallisuus samoihin energia- tai aineenvaihduntareitteihin ja geenien koodaamien proteiinien perheet.

4.1.1 K-means

K-means-algoritmi (ks. esim. [39]) soveltuu aineiston ryhmittelyyn, kun ryhmien määrä on valittu ennakolta. Menetelmän tavoitteena on ryhmien sisäisen homogeenisuuden ja toisaalta ryhmien väliset erot maksimoivan ryhmittelyn muodostaminen. Tämä voidaan mieltää 'luonnolliseksi' tavaksi muodostaa K ryhmää tutkittavaan pistejoukkoon. Kustannusfunktiona käytetään yleensä ryhmien keskimääräistä kvantisaatiovirhettä. Ryhmän kvantisaatiovirhe Q on siihen kuuluvien datapisteiden $\{\mathbf{d}_r\}$ ja ryhmäkohtaisen mallivektorin $\mathbf{m}$ neliöllisten etäisyyksien summa

$$Q = \sum_r d(\mathbf{d}_r, \mathbf{m})^2,$$

missä d merkitsee käytettävää etäisyysmittaa.

Ryhmittelyn lähtökohtana ovat satunnaisesti valitut ja riittävän etäällä toisistaan sijaitsevat K mallivektoria. Mallivektorit määrittelevät näytteiden ryhmittelyn siten, että kukin datapiste kuuluu lähimmän mallivektorin osoittamaan ryhmään. Uudet mallivektorit lasketaan kullakin laskentakierroksella ryhmän pisteiden keskiarvona. Mallivektorien sijainnit voivat tällöin muuttua, mikä voi puolestaan aiheuttaa niiden määrittelemän ryhmittelyn muuttumiseen. Ryhmien muuttuminen johtaa seuraavalla laskentakierroksella uusien mallivektoreiden määrittämiseen. Ryhmittelyä jatketaan kunnes riittävä kustannusfunktion suppeneminen tai muu optimointikriteeri on saavutettu.

Satunnaisalustus vaikuttaa ryhmittelyn lopputulokseen. Kustannusfunktion paikallisten minimien vaikutuksen välttämiseksi ryhmittely tehdään muutamilla erilaisilla satunnaisalustuksilla, joista valitaan kustannusfunktion mielessä paras.

Voronoi-alueet

Mallivektoreiden avulla määritelty ryhmittely voidaan käsittää laajemmin koko avaruuden ositukseksi. Tarkasteltava data-avaruus X jaetaan osiin $\{V_i\}$ määrittelemällä mallivektoreiden joukko $\{\mathbf{m}_i\} \subset X$. Mielivaltainen data-avaruuden piste $\mathbf{x} \in X$ kuuluu valitun etäisyysmitan d mielessä lähimmän mallivektorin määräämään ryhmään: $\mathbf{x} \in V_j$, jos $d(\mathbf{x}, \mathbf{m}_j) < d(\mathbf{x}, \mathbf{m}_k)$ pätee kaikille $k \neq j$. Ositus on reunapisteitä lukuun ottamatta yksikäsitteinen. Jatkuvassa avaruudessa reunapisteiden todennäköisyys on nolla, ja ne voidaan jättää huomioimatta käytännön laskennassa.

Annetun mallivektorin $\mathbf{m}_j$ määrittelemää pistejoukkoa $V_j \subset X$ kutsutaan *Voronoi-alueeksi*. Voronoi-alueet ovat yhtenäisiä ja ainakin matalaulotteisissa avaruuksissa usein myös konvekseja.

Tutkittavien datapisteiden jakauma voidaan mallivektoreiden avulla esittää diskreetissä ja tiivistetyssä muodossa. Tämä on tavoitteena vektorikvantisaatiossa, johon K-means menetelmä on siten läheisessä yhteydessä.

4.2 Yhteisjakaumamallintaminen

Yksityiskohtaisimmin kahden aineiston X ja Y riippuvuuksia voitaisiin tutkia niiden yhteisesiintymien tiheyttä kuvaavan todennäköisyysjakauman $p(\mathbf{X}, \mathbf{Y})$ avulla. Tämä yhteisjakauma sisältäisi kaiken tiedon aineistojen riippuvuuksista. Tuntematon yhteisjakauma joudutaan käytännössä arvioimaan äärellisen havaintoaineiston perusteella. Täydellisessä yhteisjakaumamallintamisessa käytetään aineistojen riippuvuuden mallintamisen ohella huomattavasti voimavaroja aineistojen sisäisten jakaumien mallintamiseen, ja se on laskennallisesti hyvin raskasta. Tässä työssä olemme ensisijaisesti kiinnostuneita kahden aineiston riippuvuuksista. Aineistoista kiinnostavia ryhmiä erotteleva menetelmä voi kyetä jäljittämään riippuvuuksia jopa paremmin kuin yhteisjakauman täysi mallintaminen (ks. [51, 57, 58]).

Yhteisjakaumaa voitaisiin pyrkiä mallintamaan karkeasti esimerkiksi yhdistämällä tutkimusaineistot. Mallin opetukseen käytettävät dataparit $(\mathbf{x}, \mathbf{y}) \in X \times Y$ voitaisiin esittää yksittäisinä yhdistetyn avaruuden pisteinä $[\mathbf{x}, \mathbf{y}] \in X \cup Y$. Konkatenoitujen dataparien muodostama pistejoukko voitaisiin ryhmitellä tavanomaisella ryhmittelymenetelmällä. Ryhmiä tarkastelemalla havaittaisiin, miten kahden alkuperäisen data-avaruuden alueet ovat yhdistyneet ryhmittelyssä. Yhdistetyn avaruuden ositus saattaisi johtaa vaikeasti tulkittavien ryhmien muodostumiseen tapauksessa, jossa ositettava avaruus muodostuu kahdesta erillisestä komponentista, mutta ryhmittelymenetelmä etsii yhdistetyn avaruuden osituksen. Tässä työssä eksploratiiviseen riippuvuuksien mallinnustehtävään soveltuvatkin luultavasti paremmin menetelmät, jotka muodostavat omat osituksensa kahdelle erillisenä esitettävälle aineistolle.

4.3 Käytettävät riippuvuusmallintamisen menetelmät

Korrelaatioiden tai muiden etäisyysmittojen avulla olisi helppoa löytää geenipareja, joissa geenien ilmenemisessä on selviä eroja tai samankaltaisuuksia. Lisäehtoja asettamalla voidaan pienentää tarkemman tutkimuksen kohteeksi valittavien näytteiden määrää ja kohdentaa huomio sellaisiin tutkimusaineiston piirteisiin, joita ei onnistuta kuvaamaan pelkkien perinteisten tunnuslukujen avulla. Tässä työssä lisäehtona käytetään geeniryhmien riippuvuuksiin liittyvää vastingeenien tilastollisesti poikkeuksellista esiintymistiheyttä.

4.3.1 Kontingenssitaulut

Perinteisesti kontingenssitaulujen avulla on tutkittu diskreettien näytteiden yhteisesiintymiä ja riippuvuuksia, esimerkiksi eri kahvilaatujen ja maistajien arvostelmien suhdetta. Muodostettava kontingenssitaulu on kaksiulotteinen matriisi, jonka rivit vastaavat kahvilaatuja ja sarakkeet erilaisia luonnehdintoja kuten kitkerä, pehmeä tai imelä. Kontingenssitaulun ruudut ovat rivien ja sarakkeiden leikkauskohtia. Kukin ruutu kertoo miten monesti tietty kahvilaatu on arvioitu tietyn makuiseksi. Eri kahvilaaduille tehtyjen maistatusten ja erilaisten makuarvioiden kokonaismäärät muodostavat taulun reunajakaumat; pikakahvia on voinut maistaa kymmenen ihmistä, mutta pannukahvia vain viisi. Kontingenssitaulun tilastollisilla tarkasteluilla voidaan havaita satunnaisista vaihteluista poikkeavat reunajakaumien yhteydet, esimerkiksi se että pannukahvin makua pidetään yleisesti pehmeänä ja pikakahvia kitkeränä.

Riippumattomien reunajakaumien mallissa kontingenssitaulun ruutujen todennäköisyydet ovat kontingenssitaulun reunajakaumia vastaavien todennäköisyysjakaumien tuloja, jolloin esimerkiksi kahvin laadun ja maun välillä ei havaita riippuvuutta. *Riippuvien reunajakaumien mallissa* reunajakaumat oletetaan toisistaan tuntemattomalla tavalla riippuviksi, jolloin havaittu kontingenssitaulun jakauma heijastelee tuntematonta ruutujen yli määriteltyä multinomijakaumaa. Tällöin reunajakaumat eivät määrää kontingenssitaulun jakaumaa, mutta voivat jossain määrin vaikuttaa siihen. Yhteyksien päättely suoraan kontingenssitaulun jakaumasta voi olla hankalaa. Perinteisesti kontingenssitaulun riippuvuusmittana on käytetty χ^2-testiä, mutta reunajakaumien riippuvuuksia voidaan tutkia myös *Bayes-faktorin* avulla.

Kontingenssitaulun esittämää tietoa voidaan havainnollistaa erilaisilla visualisoinneilla (ks. kuva 6.1), ja kontingenssitauluja vertaamalla voidaan arvioida eri menetelmien onnistumista riippuvuuksien mallinnustehtävässä.

Hiiren ja ihmisen geeniryhmien riippuvuuksien tarkastelemiseksi muodostettavan kontingenssitaulun rivit vastaavat ihmisen ja sarakkeet hiiren geenien ryhmiä. Kummankin lajin geeniryhmiä kutsutaan tässä yhteydessä *reunaryhmiksi*, koska niiden voidaan ajatella muodostavan kontingenssitaulun reunat. Kontingenssitaulun ruutujen osoittamia ihmisen ja hiiren geeniryhmien pareja kutsutaan *yhteisryhmiksi*. Kontingenssitaulu esittää kunkin yhteisryhmän geeniparien lukumäärän.

4.3.2 Bayes-faktori riippuvuuden mittana

Bayesilainen teoria tarjoaa johdonmukaisen tavan käyttää todennäköisyyksiä ilmaisemaan epävarmuutta päättelyssä. Se poikkeaa tavanomaisesta tilastollisesta mallinnuksesta ja hypoteesin testauksesta ennakko-oletusten eksplisiittisen esittämisen suhteen. Perinteisessä tilastotieteessä ennakko-oletukset eivät ole suoraan mukana päättelyssä, vaan sisältyvät implisiittisesti luokkien todennäköisyysmalleihin, vertailtavaksi valittuihin hypoteeseihin ja hypoteesien hyväksymisehtoihin.

Bayesilaisen päättelyn peruselementtejä ovat tutkittavaa hypoteesia koskeva (*a priori*) ennakkotietämys, ja *posteriori*tietämys, jossa havainnosta saatu informaatio on yhdistetty ennakkotietämykseen. Päättely tapahtuu muodostamalla todennäköisyysjakaumat kiinnostuksen kohteena oleville muuttujille havaintojen ja ennakko-oletusten avulla.

Bayesilaista lähestymistapaa voidaan käyttää vastakkaisten hypoteesien $\bar{H}$ ja H todennäköisyyksien vertailuun hypoteesien todennäköisyyksistä tehtyjen ennakko-oletusten $p(\bar{H})$ ja $p(H)$ ja havaintoaineiston D valossa. Havaintojen D ehdollinen todennäköisyys hypoteesin H tapauksessa on $p(D|H)$. Merkitsemme vastaavasti myös muita ehdollisia todennäköisyyksiä. Havaintoaineiston todennäköisyyttä vastakkaisten hypoteesien valossa vertaava *Bayes-faktori* on muotoa

$$\frac{p(D|\bar{H})}{p(D|H)} = \frac{p(\bar{H}|D)}{p(H|D)} \cdot \frac{p(\bar{H})}{p(H)} \,, \tag{4.1}$$

ja seuraa suoraan todennäköisyyslaskennan perusaksioomista (ks. [27]).

Soveltaminen tutkimusaiheeseen

Bayes-faktoria voidaan käyttää mittana kontingenssitaulun reunajakaumien riippuvuudelle (ks. esim. [31]). Tällöin riippumattomien reunaryhmien malli H toimii nollahypoteesina, johon riippuvien reunaryhmien mallin $\bar{H}$ todennäköisyyttä verrataan. Havaintoaineistona on kontingenssitaulun multinomijakauma $\{n_{ij}\}$, jonka todennäköisyys on eri hypoteesien valossa erilainen. Reunajakaumien riippuvuuksista tehdyt ennakko-oletukset sisältyvät kontingenssitaulun prioreihin.

Kontingenssitaulun jakauma ja sen reunajakaumat ovat multinomiaalisia, ja kontingenssitaulun Bayesilaiseen tarkasteluun soveltuva Dirichlet'n priori (ks. [30]) on muotoa $\prod_{s=1}^{t} q_s^{\alpha_s - 1}$. Parametrit α_s ovat positiivisia prioriparametreja, ja multinomijakauman muuttujiin liittyvien todennäköisyyksien summa on $\sum_s q_s = 1$. Parametri t ilmaisee tarkastelun kohteena olevan multinomijakauman muuttujien lukumäärän. Kontingenssitaulun tarkastelussa esiintyviä multinomijakaumia ovat kontingenssitaulun jakauma ja sen reunajakaumat. Riippuvien reunaryhmien mallissa jokaiseen kontingenssitaulun ruutuun (i, j) liittyy oma prioriparametrinsa α_{ij}. Riippumattomien reunaryhmien mallissa priorit $\{\alpha_i\}$ ja $\{\alpha_j\}$ määritellään reunajakaumille, ja kontingenssitaulun ruutujen priorit

saadaan näiden ulkotulona: $\alpha_{ij} = \alpha_i \cdot \alpha_j$ kaikille (i, j). Sopivien prioriparametrien valinta riippuu ongelman luonteesta. Prioriparametrien valinnan jälkeen havaintoaineiston ehdolliset todennäköisyydet eri hypoteesien valossa $P(\{n_{ij}\}|H)$ ja $P(\{n_{ij}\}|\bar{H})$ voidaan laskea.

Bayes-faktorissa (4.1) esiintyvä hypoteesien prioritodennäköisyyksien suhde on tässä tutkimuksessa $p(\bar{H})/p(H) = 1$, koska kummankaan hypoteesin todennäköisyyden painottamiselle ei ole erityisiä perusteita. Kontingenssitaulun jakauman $\{n_{ij}\}$ nojalla laskettu Bayes-faktori yhtyy nyt hypoteesien posterioritodennäköisyyksien suhteeseen, joka on Dirichlet'n priorin tapauksessa muotoa (ks. [59])

$$\frac{P(\bar{H}|\{n_{ij}\})}{P(H|\{n_{ij}\})} \propto \frac{\prod_{ij} \Gamma(n_{ij} + \alpha_{ij})}{\prod_i \Gamma(n_{\cdot i} + \alpha_i) \prod_j \Gamma(n_{j\cdot} + \alpha_j)} \,, \tag{4.2}$$

missä $n_{\cdot i} = \sum_j n_{ij}$ ja $n_{j\cdot} = \sum_i n_{ij}$ merkitsevät kontingenssitaulun reunajakaumia. Kaavassa esiintyvä gammafunktio on muotoa

$$\Gamma(z) = \int_0^\infty t^{z-1} e^{-t}\, dt.$$

Gammafunktion ja kertoman välillä vallitsee mielenkiintoinen yhteys kaikille positiivisille kokonaisluvuille n; $\Gamma(n + 1) = n!$ (ks. [17]). Prioriparametrien asettaminen arvoon $\alpha_{ij} = \alpha_i = \alpha_j = 1$ kaikille (i, j) vastaa tilannetta, jossa ennakko-oletuksena on erilaisten havaintoaineistojen tasajakauma. Oletus on luonteva, kun ennakkotietoa jakaumasta ei ole käytettävissä. Tässä tapauksessa Bayes-faktori yhtyy hypergeometriseen todennäköisyyteen, jota on perinteisesti käytetty kontingenssitaulujen riippuvuusmittana [25, 73]: havaituilla reunajakaumilla $\{n_{i\cdot}\}$ ja $\{n_{\cdot j}\}$ todennäköisyys jakauman $\{n_{ij}\}$ esiintymiselle kontingenssitaulussa on riippumattomien reunaryhmien hypoteesin H tapauksessa

$$P_T \equiv P(\{n_{ij}\}|\{n_{i\cdot}\}, \{n_{\cdot j}\}, H) = \frac{\prod_i n_{i\cdot}! \prod_j n_{\cdot j}!}{N! \prod_{ij} n_{ij}!} \,, \tag{4.3}$$

missä $N = \sum_{ij} n_{ij}$ merkitsee näyteparien kokonaismäärää. Mitä suurempi tämä todennäköisyys on, sitä todennäköisemmin reunajakaumat ovat riippumattomia. Toisaalta reunajakaumien riippuvuuden todennäköisyys ja tasajakaumaa esittävää prioria vastaava Bayes-faktori ovat käänteisesti verrannollisia todennäköisyyteen (4.3) nähden. Tarkempia pohdintoja kontingenssitaulujen, multinomijakaumien ja Dirichlet'n priorien suhteesta löytyy esimerkiksi artikkeleista [30] ja [31].

Kontingenssitaulun reunajakaumien riippuvuutta voidaan mitata myös yhteisinformaatiolla. Äärellisen datan tapauksessa kontingenssitaulun heijastelema aito todennäköisyysjakauma on arvioitava kontingenssitaulun empiirisestä jakaumasta. Tästä johdettu yhteisinformaatio [16] on kuitenkin harhainen estimaatti. Bayes-faktorin käyttö äärellisten datajoukkojen yhteydessä on hyvin perusteltua. Se välttää empiirisen yhteisinformaation

näytteiden keruusta johtuvat epävarmuudet, mutta on asymptoottisesti verrannollinen yhteisinformaatioon suurilla näytemäärillä [58].

4.3.3 Riippuvuusmallintaminen K-means-menetelmällä

Täydellistä yhteisjakaumamallintamista keveämpi vaihtoehto kahden aineiston riippuvuuksien etsimiseksi olisi aineistojen riippumaton ryhmittely esimerkiksi tavanomaisella K-means-menetelmällä, ja esille tulevien riippuvuuksien tarkastelu kontingenssitaulun avulla. Ei ole kuitenkaan taattua, että aineistojen riippumaton ryhmittely paljastaisi tehokkaasti niiden riippuvuuksia.

Tässä tutkimuksessa valittiin sekä ihmisen että hiiren geeneille K-means-menetelmän kustannusfunktion mielessä paras ryhmittely kolmella eri satunnaisalustuksella tehdyn ryhmittelyn joukosta. Ryhmittelyn onnistumista mitattiin validointijoukolla. Mallin opettamiseen käytettiin siis vain osaa aineistosta, ja opetetun mallin onnistumista jäljelle jääneiden näytteiden ryhmittelyssä mitattiin kustannusfunktiolla. Näin sovittamisessa kyetään arvioimaan myös mallin yleistyskykyä. Tässä tapauksessa opetukseen ja mallin arviointiin käytettiin yhtä suuria datajoukkoja.

Mallin opetukseen käytettävää datajoukkoa sanotaan opetusjoukoksi, ja opetetun mallin arviointiin käytettävää joukkoa validointijoukoksi. *Ristiinvalidoinnissa* mallin opetus ja arviointi tehdään useilla erilaisilla tutkimusdatasta poimituilla opetus- ja validointijoukoilla, ja opetetuista malleista valitaan kustannusfunktion mielessä paras.

4.3.4 VQ-IB

Tehokkaaseen riippuvuuksien etsintään kontingenssitaulun avulla soveltuva menetelmä on 'Information bottleneck' (IB) [60]. Alkuperäinen IB-menetelmä ryhmittelee diskreettejä näytteitä yhdessä datajoukossa. Sen tavoitteena on sellaisten ryhmittelyn muodostaminen, joka maksimoi ryhmien ja koeasetelmien yhteyksiä esittävästä kontingenssitaulusta lasketun empiirisen yhteisinformaation. Tavanomainen esimerkki menetelmän soveltamisesta on tekstidokumenttien ryhmittely sanajakaumien perusteella.

Tämän tutkimuksen tavoitteena on kahden jatkuva-arvoisesta datasta muodostuvan aineiston ryhmittely, joka kuvaa mahdollisimman hyvin aineistojen riippuvuuksia ja yleistyy kuvaamaan myös uusia näytteitä. IB-menetelmä soveltuu tehtävään joidenkin muutosten jälkeen. Ryhmittelyn lähtökohtana on kahden tarkasteltavan data-avaruuden diskretointi atomaarisiksi Voronoi-alueiksi. Ryhmittelyn pohjana käytettävät atomaariset ryhmät muodostetaan vektorikvantisaatiosta tutulla mallivektorimenetelmällä (ks. kohta 4.1.1), minkä johdosta olemme antaneet menetelmän nimeksi VQ-IB.

Ensimmäisen datajoukon atomaarisia alueita käytetään IB-menetelmällä ryhmiteltävinä diskreetteinä näytteinä. Kutakin näytettä vastaa kontingenssitaulun rivi, jolta voidaan lukea tarkasteltavaan atomaariseen alueeseen kuuluvien datapisteiden määrän jakauma toisen data-avaruuden atomaarisiin alueisiin kuuluvien pisteiden pareina. Satun-

naisen alustuksen jälkeen ryhmittelyä optimoidaan näyte kerrallaan siten, että kontingenssitaulun rivien ja sarakkeiden empiirinen yhteisinformaatio maksimoituu. Näytteiden ryhmittelyn jälkeen datajoukkojen rooli vaihdetaan ja ryhmittely tehdään toisen data-avaruuden atomaarisille Voronoi-alueille. Prosessia toistetaan kustanusfunktiona käytettävän yhteisinformaation suppenemiseen tai ennalta asetetun askelmäärä saavuttamiseen asti.

Ryhmittely maksimoi atomaarisista Voronoi-alueista muodostettujen kahden aineiston ryhmien välisen empiirisen yhteisinformaation. Ryhmät eivät välttämättä ole yhtenäisiä, mutta yleistyvät koko data-avaruuteen.

Sovittaminen tutkimusaiheeseen

Sopiva atomaaristen Voronoi-alueiden määrä etsittiin validointijoukolla, ja alueet parametrisoitiin kummassakin avaruudessa K-means-menetelmällä käyttäen kolmea erilaista satunnaisalustusta. Ryhmittelyn optimointiin käytettiin sekventiaalista IB-menetelmää [60]. Käytettäväksi valittiin validointijoukon avulla IB:n kustannusfunktion mielessä paras kolmesta ryhmittelystä.

4.4 Optimointimenetelmät

Käytännössä mallin onnistumista tehtävässään mitataan kustannusfunktion avulla. Sopivat parametrit etsitään laskennallisesti optimoimalla kustannusfunktion arvoa. Tällöin mallin sovituksesta tulee laskennallisessa mielessä perinteinen optimointiongelma. Tässä työssä tarvittavia optimointimenetelmiä ovat konjugaattigradienttimenetelmä ja simuloitu jäähdytys.

4.4.1 Konjugaattigradienttimenetelmä

Konjugaattigradienttimenetelmä (ks. esim. [29]) on suunniteltu neliöllisten ongelmien optimointiin, mutta sitä voidaan käyttää sopivilla rajoituksilla myös muunlaisten ongelmien ratkaisemiseen. Käytännössä useimpien kustannusfunktioiden paikallinen käyttäytyminen on lähes neliöllistä ja konjugaattigradienttimenetelmä löytää riittävällä askelmäärällä paikallisen minimin mille tahansa differentioituvalle kustannusfunktiolle.

Uusi optimointisuunta valitaan jokaisella optimointiaskeleella funktion gradientin ja aiemman optimointisuunnan avulla. Minimoitava neliöllinen kustannusfunktio on muotoa

$$f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} + \mathbf{c}^T\mathbf{x} \quad ,$$

missä $\mathbf{x}$ ja $\mathbf{c}$ ovat D-ulotteisia reaaliarvoisia vektoreita ja A reaaliarvoinen $D \times D$-

neliömatriisi. Ensimmäinen optimointisuunta $\mathbf{d}_0$ valitaan aloituspisteessä $\mathbf{x}_0$ funktion gradientille vastakkaiseksi. Tämän jälkeen edetään seuraavien askelten mukaisesti.

1. Etsi kustannusfunktion minimi suunnassa $\mathbf{d}_i$, eli etsi λ_i siten, että $f(\mathbf{x}_i + \lambda_i \mathbf{d}_i)$ minimoituu. Olkoon $\mathbf{x}_{i+1} = \mathbf{x}_i + \lambda_i \mathbf{d}_i$.
2. Olkoon $\mathbf{d}_{i+1} = -\nabla f(\mathbf{x}_{i+1}) + \gamma_i \mathbf{d}_i$, missä
$$\gamma_i = \frac{\nabla f(\mathbf{x}_{i+1})^T \left(\nabla f(\mathbf{x}_{i+1}) - \nabla f(\mathbf{x}_i)\right)}{\|\nabla f(\mathbf{x}_i)\|^2}$$
3. Lopeta, jos $i + 1 = D$, muuten kasvata lukua i yhdellä ja palaa ensimmäiseen kohtaan.

Parametrin γ päivitykseen käytetään tässä niin kutsuttua Polak-Ribiéren menetelmää. Myös muita päivitysmenetelmiä on ehdotettu. Neliöllisten kustannusfunktioiden tapauksessa menetelmät ovat yleensä identtisiä, mutta ei-neliöllisten kustannusfunktioden tapauksessa Polak-Ribiéren menetelmää pidetään yleisesti laskennallisesti toimivimpana.

Neliöllisille ongelmille sovellettuna konjugaattigradienttimenetelmä suppenee D askeleella. Ei-neliöllisten ongelmien tapauksessa on käytettävä suurempaa askelmäärää. Laskennallisista epätarkkuuksista ja kustannusfunktion ei-neliöllisyydestä johtuen minimointisuuntien A-konjugaatti saattaa harhautua. Tästä syystä optimointi käynnistetään ei-neliöllisten ongelmien ratkaisemisessa ajoittain uudelleen. Yleisesti käytetty menetelmä on optimoinnin käynnistäminen saavutetussa optimipisteessä lasketun gradientin osoittamalle suunnalle vastakkaisesti D askelen välein; gradientin vastaluku osoittaa uuden suunnan. Laskentaa jatketaan, kunnes riittävä suppeneminen tai muu optimointikriteeri on saavutettu.

4.4.2 Simuloitu jäähdytys

Simuloitu jäähdytys (ks. esim. [34]) on satunnaisuuteen perustuva optimointimenetelmä, joka jäljittelee aineen molekyylirakenteen järjestäytymisen noudattamia säännönmukaisuuksia hitaassa lämpötilan laskussa. Optimaalista ratkaisua etsitään aiheuttamalla systeemiin pieniä satunnaismuutoksia. Satunnaisuuden hyödyntäminen uusien ratkaisuvaihtoehtojen etsimisessä on eduksi, koska se pienentää menetelmän häiriöalttiutta ja ehkäisee juuttumista paikallisiin minimeihin.

Menetelmän suppenemisesta globaaliin optimiin ei yleensä ole takeita. Etuna on, että se löytää yleensä melko hyvän ratkaisun, eikä kustannusfunktion sileydestä ja derivoituvuudesta tarvitse tehdä oletuksia. Kustannusfunktion sijasta simuloidussa jäähdytyksessä puhutaan usein systeemin 'energiasta', mikä korostaa analogiaa fysikaalisen jäähtymisilmiön kanssa.

Menetelmän soveltaminen etenee seuraavien askelten mukaisesti. Optimoitavan systeemin tilaan tuotetaan satunnainen muutos. Tämän jälkeen verrataan uuden ja vanhan

systeemin energioita. Muutos hyväksytään automaattisesti, mikäli systeemin uudella tilalla on matalampi energia. Myös muutos korkeampaan energiatilaan voidaan hyväksyä tietyllä todennäköisyydellä. Tämä vastaa fysikaalisen systeemin lämpöliikettä. Energiaa kasvattavan muutoksen hyväksymistodennäköisyys p_M optimointiaskeleella t saadaan kaavasta

$$p_M(t) = \exp\left(-\frac{(E(t+1) - E(t))}{T(t)}\right) , \tag{4.4}$$

missä $T(t)$ on askelmäärän mukana vähenevä systeemin 'lämpötilaa' osoittava funktio, ja $E(t)$ systeemin energia askelella t. Kokonaisenergiaa kasvattavat muutokset hyväksytään todennäköisemmin korkeammissa lämpötiloissa. Alkutilanteessa systeemin tilojen vaihtelu onkin lähes täysin satunnaista. Jäähdytyksen seurauksena systeemin energia vähitellen laskee, ja matalammissa lämpötiloissa energian kasvamiseen johtavat muutokset hyväksytään yhä epätodennäköisemmin. Vähitellen systeemi suppenee kohti lopullista ratkaisua. Optimointi voidaan lopettaa esimerkiksi siinä vaiheessa, kun merkittäviä muutoksia ei enää tapahdu.

Luku 5

Assosiatiivinen ryhmittely

Tämän työn tavoitteena on mielenkiintoisten hiiren ja ihmisen vastingeenien ryhmien etsintä genominlaajuisesta aineistosta. Tehtävään soveltuvia valmiita menetelmiä ei ole ollut tiettävästi saatavilla. Assosiatiivinen ryhmittely on lupaava ja teoreettisesti perusteltu eksploratiivinen menetelmä kahden geenien ilmenemistä kuvaavan aineiston riippuvuuksien mallintamiseen ja kiinnostavien geeniryhmien etsintään. Uuden menetelmän hyviä ominaisuuksia ovat riippuvuusmallintamisen ohella tulosten yleistyvyys ja helppotulkintaisuus.

Menetelmän toistaiseksi ainoat bioinformatiikkasovellukset ovat alustavia ja liittyvät vastingeenien ja geenien säätelytekijöiden tutkimiseen [59].

5.1 Periaate

Assosiatiivinen ryhmittely on kahden jatkuva-arvoisen monimuuttuja-avaruuden X ja Y ryhmittelymenetelmä, joka pyrkii mallintamaan datajoukkojen riippuvuuksia tunnetun paritiedon $\{(\mathbf{x}_k, \mathbf{y}_k)\}_k \subset \mathbb{R}^{d_x} \times \mathbb{R}^{d_y}$ avulla ilman ennakkotietämystä niiden rakenteesta. Menetelmä pyrkii löytämään kummastakin avaruudesta yhtenäisiä alueita, joilla sijaitsevien pisteiden parit erottuvat omaksi ryhmäkseen myös toisessa avaruudessa. Muodostettavat ryhmät yleistyvät koko data-avaruuteen. Yhtenäisyyden ansiosta ryhmillä on tiettyä sisäistä homogeenisuutta, mikä helpottaa niiden tulkintaa.

Kahden avaruuden ositusten väliset yhteydet esitetään kontingenssitaululla. Ositusten karteesiset tulot vastaavat kontingenssitaulun ruutuja, ja määräävät dataparien jakautumisen yhteisryhmiin. Menetelmä pyrkii löytämään sellaiset ositukset, että niitä vastaavan kontingenssitaulun reunajakaumien riippuvuus on mahdollisimman suuri. Riippuvuutta mitataan Bayes-faktorin (4.1) avulla. Perinteisesti Bayes-faktoria on käytetty mittana kontingenssitaulun reunojen riippuvuudelle (ks. esim. [31]). Assosiatiivisessa ryhmittelyssä ryhmät muodostetaan maksimoimalla kontingenssitaulusta laskettua Bayes-faktoria sen sijaan, että ainoastaan mitattaisiin valmiiden taulujen esittämiä riippuvuuksia.

Bayes-faktorin optimoinnin jälkeen kontingenssitaululla on yleensä joitakin satunnaisuudesta merkitsevästi poikkeavia ruutuja. Näissä näytteiden esiintymistiheyksiä ei voida selittää kontingenssitaulun reunajakaumien tulona. Poikkeuksellisten ruutujen osoittamissa yhteisryhmissä on tässä tutkimuksessa potentiaalisesti mielenkiintoisia geenipareja.

5.2 Toteutus

Kummankin avaruuden ryhmät parametrisoidaan Voronoi-alueina, ja sopiva ryhmittely etsitään optimoimalla mallivektoreiden sijaintia. Ryhmien parametrisointi mallivektoreiden avulla on laskennallisesti yksinkertainen menettelytapa.

5.2.1 Alustus

Assosiatiivisessa ryhmittelymenetelmässä ryhmien määrä on ennakolta kiinnitetty. Luonnollinen alustava ositus voidaan tuottaa luvussa 4 esitellyn K-means-menetelmän avulla. Valittu alustus voi vaikuttaa voimakkaasti ryhmittelyn lopputulokseen. K-means-menetelmän käyttö on perusteltua, koska se tuottaa mahdollisimman homogeenisia ryhmiä tukien pyrkimystä helposti tulkittavien ryhmien muodostamiseen. Ryhmittely tehtiin kolmella eri satunnaisalustuksella, joista valittiin paras validointijoukolla.

5.2.2 Optimointi

Menetelmän maksimoitavana kustannusfunktiona käytetään Bayes-faktoria (4.1), jonka avulla voidaan verrata kontingenssitaulun riippuvien ja riippumattomien reunajakaumien hypoteesien todennäköisyyksiä. Hypoteesien prioritodennäköisyyksien suhteen ollessa vakio kontingenssitaulun Bayes-faktori (4.1) yhtyy lausekkeeseen (4.2). Havaintoaineistojen tasajakaumaa vastaavilla prioriparametreilla $\alpha_{ij} = \alpha_i = \alpha_j = 1$ kaikille (i, j) tämä on muotoa

$$\frac{P(\bar{H}|n_{ij})}{P(H|n_{ij})} \propto \frac{\prod_{ij} \Gamma(n_{ij} + 1)}{\prod_i \Gamma(n_{\cdot i} + 1) \prod_j \Gamma(n_{j\cdot} + 1)} \,. \tag{5.1}$$

Lausekketta (5.1) käytetään tässä työssä assosiatiivisen ryhmittelyn kustannusfunktiona. Sen maksimointi on tehdyillä oletuksilla yhtäpitävää Bayes-faktorin (4.1) maksimoinnin kanssa.

Rajojen pehmennys

Optimointimenetelmien soveltamisessa tarvittava Bayes-faktorin gradientti on ns. 'kovien' ryhmien tapauksessa nolla tai sitä ei ole määritelty suhteessa Voronoi-alueiden

mallivektoreihin, riippuen niiden arvoista. Tässä tapauksessa kustannusfunktion (5.1) maksimoiva ryhmien mallivektoreiden optimointitehtävä on kombinatorinen. Kombinatoristen ongelmien ratkaiseminen on laskennallisesti hyvin raskasta, jopa mahdotonta. Laskennan helpottamiseksi ja äärellisten gradienttien tuottamiseksi turvaudutaan 'pehmennettyihin' rajoihin ja ositusten jatkuvuusoletukseen, jolloin mallin optimointiin voidaan käyttää perinteisiä optimointimenetelmiä. On huomattava, että tässä tapauksessa kunkin ryhmään kuuluvien näytteiden 'lukumäärä' saadaan tätä ryhmää vastaavan kuulumisfunktion integraalina koko avaruuden yli. Vastaava kontingenssitaulun jakauma $\{n_{ij}\}$ ei välttämättä muodostu kokonaisluvuista.

Ryhmittelyä sanotaan 'kovaksi', jos jokainen ryhmiteltävä piste voi kuulua vain yhteen ryhmään. 'Pehmeiden' ryhmien tapauksessa ryhmien rajat ovat sumeita, ja raja-alueiden pisteet voivat kuulua samanaikaisesti kahteen tai useampaan ryhmään, mutta eri voimakkuuksilla. Pisteen $x \in X$ *kuulumisaste* ryhmään i saadaan tässä tapauksessa kaavasta

$$g_i(\mathbf{x}) \equiv Z(\mathbf{x})^{-1} \exp\left(-\|\mathbf{x} - \mathbf{m}_i\|^2/\sigma^2\right).$$

Gaussiset kuulumisfunktiot $g_i(\cdot)$ kuvaavat reunoilta pehmennettyjä mallivektoreiden $\{\mathbf{m}_i\}$ avulla määriteltyjä Voronoi-alueita. Gaussisen jakauman hajontaparametri σ säätelee pehmennyksen voimakkuutta. Kuulumisasteiden summa on jokaisessa avaruuden pisteessä $\mathbf{x}$ täsmälleen 1. Tästä huolehtii normalisointivakio $Z(\cdot)$, joka määräytyy ehdosta $\sum_i g_i(\mathbf{x}) = 1$.

5.2.3 Laskenta

Pehmennetyillä rajoilla muokattua kustannusfunktiota (5.1) merkitään symbolilla BF'. Laskennallisista syistä optimoinnissa käytetään tämän logaritmia. Tällöin yksittäisten todennäköisyyksien tulot muuttuvat summiksi, joita on yleensä helpompi käsitellä. Tavallisesti valmiit optimointimenetelmät minimoivat kustannusfunktion arvoa. Bayes-faktorin maksimoinnin sijasta voidaan minimoida sen vastalukua. Tämän logaritmiksi saadaan nyt

$$\begin{aligned} -\log BF' = &\log \sum_{ij} \Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k) g_j^{(y)}(\mathbf{y}_k)\right) \\ &- \lambda^{(x)} \log \sum_i \Gamma\left(\sum_k g_i^{(x)}(\mathbf{x}_k)\right) \\ &- \lambda^{(y)} \log \sum_j \Gamma\left(\sum_k g_j^{(y)}(\mathbf{y}_k)\right), \end{aligned} \tag{5.2}$$

missä indeksit x ja y viittaavat kahteen tarkasteltavaan datajoukkoon. Tämä on ryhmittelyn laskennallisessa vaiheessa minimoitava kustannusfunktio, joka optimoidaan kon-

jugaattigradienttimenetelmällä suhteessa kahdessa tarkasteltavassa avaruudessa oleviin Voronoi-alueiden mallivektoreiden joukkoihin $\{\mathbf{m}_i^{(x)}\}$ ja $\{\mathbf{m}_j^{(y)}\}$.

Kaavassa (5.2) esiintyvien parametrien $\lambda^{(\cdot)}$ arvona on lähtökohtaisesti 1, kun se johdetaan kustannusfunktion (5.1) pehmennettyjen rajojen logaritmina. Näitä termejä painottamalla menetelmän toimintaa voidaan tarvittaessa *regularisoida*.

5.2.4 Regularisointi

Regularisoinnilla tarkoitetaan menetelmän muokkaamista optimoinnin helpottamiseksi, laskennallisten ongelmien kiertämiseksi ja erityisesti yleistyskyvyn parantamiseksi. Opetusjoukolla saatujen tulosten tulisi kuvata hyvin myös muita samaan ilmiöön liittyviä aineistoja. Jos näin ei ole, mallin sanotaan ylisovittuneen.

Joissakin tapauksissa ylisovittumista voidaan ehkäistä regularisoinnilla, joka estää mallin parametreja saamasta liian äärimmäisiä arvoja. Hyvin joustavalle mallille voidaan esimerkiksi antaa sopivia lisärajoituksia. Ennakkotietämyksen merkityksen korostaminen voidaan nähdä regularisointina, ja Bayesilainen lähestymistapa regularisoinnin toteuttamiseksi olisi sopivan priorin määritteleminen parametreille.

Tässä työssä käytettiin regularisointia, joka lähestyy kontingenssitaulun reunajakaumien entropiaa ja siten yleisesti suosii saman kokoisten ryhmien muodostamista. Kustannusfunktion (5.2) parametri $\lambda^{(\cdot)} > 1$ tuottaa tällaisen regularisoinnin vastaten IB-menetelmässä [64] reunaesitysten kompleksisuutta rajoittavaa termiä.

5.3 Mallin opettaminen ja arviointi

Assosiatiivisen ryhmittelymallin opetuksen aikana etsitään mallivektorit, joiden määräämä ryhmittely on Bayesilaisen riippuvuusmitan mielessä optimaalinen. Vaihtoehtoisten ryhmittelyjen onnistumista riippuvuuksien mallintamisessa voidaan mitata kustannusfunktion avulla.

5.3.1 Opetusparametrien valinta

Ennen varsinaisen ryhmittelymallin opettamista valittavia parametreja ovat kontingenssitaulun koko eli haluttu reunaryhmien määrä, ryhmien rajoille tehtävän pehmennyksen voimakkuutta säätelevä optimointiparametri σ ja reunajakaumia tasoittava regularisointiparametri λ.

Assosiatiivisen ryhmittelyn tulokset eivät ilmeisesti ole erityisen herkkiä ryhmien määrän valinnalle, vaikka se vaikuttaakin siihen, miten hienojakoisia riippuvuuksia menetelmä kykenee tuomaan esille. Hiiren ja ihmisen datajoukkojen koko ja dimensionaalisuus

oli samaa luokkaa, ja geenien ryhmittelyyn päätettiin käyttää samaa ryhmien määrää molemmissa aineistoissa. Kontingenssitaulun koko valittiin siten, että kuhunkin yhteisryhmään päätyi keskimäärin noin kymmenen geeniparia. Tällaiset ryhmät ovat sopivan kokoisia manuaalista tarkastelua ja tulkintaa ajatellen. Taulun kooksi saatiin 21×21 ruutua. Samaa ryhmien määrää käytetään myös vaihtoehtoisilla menetelmillä tehdyissä vertailukokeissa.

Sopiva optimointiparametrin σ arvo etsittiin validointijoukon avulla. Tutkittava datajoukko jaettiin kahteen osaan, joista ensimmäistä käytettiin assosiatiivisen ryhmittelymallin opettamiseen ja toista opetetun mallin arviointiin. Malli opetettiin 30 erilaisella parametrin σ arvolla, jotka valittiin alustavien kokeiden jälkeen logaritmisesti väliltä [0,005; 5]. Kustannusfunktion (5.1) mielessä parhaan ratkaisun tuottanut arvo valittiin käytettäväksi lopullisen mallin opettamisessa. Sopivaksi regularisointiparametrin λ arvoksi valittiin molemmissa avaruuksissa alustavien kokeiden nojalla 1,2.

Lopullisen ryhmittelymallin opettamiseen käytettiin koko tutkimusaineistoa.

5.3.2 Mallin arviointi

Rajojen pehmennystä kuvaavan ryhmien jatkuvuusapproksimaation mukaista kustannusfunktiota (5.2) käytetään ainoastaan optimoinnin laskennallisessa vaiheessa. Opetetun mallin ryhmien riippuvuuden arviointiin käytetään kaavan (5.1) mukaista kustannusfunktiota 'kovilla' ryhmillä.

5.3.3 Mielenkiintoisten yhteisryhmien tunnistaminen

Assosiatiivisen menetelmän tuottamista yhteisryhmistä kiinnostavia ovat ne, joissa näytteiden määrä poikkeaa merkitsevästi satunnaisuudesta. Assosiatiivisen ryhmittelyn tuottaman kontingenssitaulun jakaumaa verrattiin riippumattomien reunajakaumien nollahypoteesin mukaisiin satunnaisiin kontingenssitauluihin, joita tuotettiin vertailua varten tuhat kappaletta. Satunnaisuudesta merkitsevänä poikkeamana pidettiin tässä työssä p-arvoa $p < 0,001$.

5.4 Ominaisuudet ja tulkinta

Tilastollisesta satunnaisuudesta poikkeavat kontingenssitaulun ruudut osoittavat potentiaalisesti mielenkiintoisia vastingeenien ryhmiä. Ryhmittely on toteutettu geenien ilmenemisprofiileiden perusteella, ja ryhmien tulkinta aloitetaan tarkastelemalla ilmenemisprofiileja ja niiden yhteyksiä.

5.4.1 Tulkinnan tasot

Yhteisryhmän tulkinta on kaksitasoinen prosessi. Ensimmäisessä vaiheessa kiinnitetään huomiota yhteisryhmää vastaavien reunaryhmien koostumukseen. Toisessa vaiheessa tarkastellaan lähemmin yhteisryhmän geenejä.

Reunaryhmien tulkinta. Poikkeuksellisen pienet yhteisryhmät sisältävät tavallisesti korkeintaan muutaman geeniparin, joiden ilmeneminen ei välttämättä ole yhtenäistä. Tällaiset geeniparit ovat usein ns. 'outliereita' eli selvästi muusta datajoukosta erillisiä yksittäisiä pisteitä. Ne voivat olla satunnaisia mittausvirheistä ja biologisista vaihteluista johtuvia harhapisteitä, mutta voivat toisaalta liittyä harvinaisiin ja biologisen tutkimuksen kannalta mielenkiintoihin ilmiöihin.

Data-analyysin yleisenä ongelmana on 'outlier'-pisteiden huomiointi tutkimuksessa. Ne voivat hankaloittaa tutkimusta tarpeettomasti, mutta toisaalta niihin saattaa liittyä arvokasta uutta tietoa. Assosiatiivinen ryhmittely on perusteltu tapa poikkeuksellisten vastingeenien etsimiseksi. Jos poikkeuksellisen pienen yhteisryhmän geenipari ilmenee hyvin samankaltaisesti kuin reunaryhmien geenit, satunnaisten tekijöiden vaikutus geeniparin löytymiseen on saattanut olla pienempi. Tällä tavalla voitaisiin pyrkiä erottamaan aineistoon sisältyvät satunnaiset harhapisteet merkittävistä biologisista löydöksistä.

On toistaiseksi määrittelemättä, millä tavalla voidaan erottaa geeniryhmälle 'tyypilliset' ilmenemisprofiilit ja toisaalta 'outlier'-pisteet. Keskiarvojen ja hajontamittojen avulla tehdyt arviot ovat puutteellisia, koska nämä eivät kuvaa geenien jakaumaa ryhmän sisällä. Jakauman ja luottamusvälien määrittely vaatisi monimutkaisempia tarkasteluja.

Riippuvuusryhmittelyn paljastamien tilastollisesti poikkeuksellisen suurten yhteisryhmien tapauksessa keskittyminen yhteisryhmän geenien tulkintaan on luultavasti riittävää. Ei ole tiedossa, mitä tulkinnan kannalta tarpeellista lisätietoa reunaryhmien geenien tarkastelu voisi tuottaa tällaisissa tapauksissa.

Yhteisryhmien tulkinta. Yhteisryhmän tulkinnassa selvitetään sen geeniparien ominaisuudet. Näitä ovat voimakkaimmin ilmenevien kudosten tunnistaminen, kudoskohtaisen hajonnan arviointi ja ryhmän geenipareja yhdistävien ominaisuuksien tarkastelu ilmenemisprofiilien vertailun ja kirjallisuushakujen avulla. Lisäksi huomioidaan, onko kyse satunnaisuuteen verrattuna poikkeuksellisen yleisestä vai harvinaisesta yhteydestä.

5.4.2 Yhteisryhmien tulkinta

Alla esitellään lyhyesti menetelmän mahdollisuudet potentiaalisesti mielenkiintoisista lajien yhteyksistä kertovien vastingeeniryhmien etsimiseksi. Näistä esitetään tiivis yhteenveto taulukossa 5.1.

Vastingeenit	(0) harvinainen yhteys	(1) yleinen yhteys
(a) erilainen ilmeneminen	lajien erot, kohina	lajien erot
(b) yhtäläinen ilmeneminen	tapauskohtainen	lajien yhteydet

Taulukko 5.1: Assosiatiivisen ryhmittelyn mahdollisuudet mielenkiintoisten vastingeeniryhmien etsinnässä. Jotkin yhteisryhmät ovat satunnaisuuteen verrattuna poikkeuksellisen suuria tai pieniä ($p < 0,001$). Tämä viittaa yhteisryhmän vastingeenien välisiin poikkeuksellisen harvinaisiin tai yleisiin yhteyksiin, jotka voivat liittyä vastingeenien erilaiseen tai samankaltaiseen ilmenemiseen. Jos aineistossa on erityisen harvinaisia vastingeenien pareja, niiden ilmenemisessä on luultavimmin odottamaton ero (a0). Ero voi johtua tutkimuksen satunnaisista virhelähteistä, mutta toisaalta se voi liittyä kiintoisaan biologiseen ilmiöön. Eri tavoin ilmenevien odottamattoman yleiseen yhteyteen liittyvien vastingeenien ryhmät (a1) ovat potentiaalisesti hyvin mielenkiintoisia, mutta niiden löytyminen on melko epätodennäköistä. Samankaltaisesti ilmenevien vastingeenien ryhmienkään ei voida odottaa olevan erityisen harvinaisia (b0). Tällaisten tapausten tulkinta on hyvin tapauskohtaista. Samankaltaisesti ilmenevät vastingeenit ovat luultavasti muita yleisempiä, joten tällaisten yhteisryhmien (b1) löytyminen lienee todennäköisintä.

Poikkeuksellisen yleiset yhteydet. Samalla tavoin ilmenevien vastingeenien voitaisiin odottaa olevan tavallista yleisempiä, koska ainakin sukulaislajeille tärkeisiin ja yhteisinä säilyneisiin tehtäviin (mm. solusyklin ylläpito, suvunjatkaminen, lihakset) erikoistuneiden geenien rakenteen ja tehtävän on arveltu muuttuvan lajien eriytymisen aikana muuhun perimäainekseen verrattuna hitaammin (ks. [32, 40, 61]). Myös tutkimuksessa [62], josta tämän työn tutkimusaineisto on peräisin, ortologien enemmistöllä havaittiin ainakin osittain samankaltainen ilmenemisprofiili.

Poikkeuksellisen suurille samankaltaisesti ilmenevien vastingeenien yhteisryhmille voitaisiin ehdottaa tällaista lajin säilymisessä tärkeää tehtävää, ja löydös voisi korostaa lajien välisiä geenitason yhteyksiä.

Lajien eroista kertovat yhteisryhmät. Luultavasti harvinaisempaa, mutta lajien erojen tutkimisen kannalta mielenkiintoista voisi olla selvä ja odottamattoman yleinen ero vastingeenien ilmenemisessä. Tällaisten vastingeeniryhmien löytyminen voisi viitata säännönmukaisuuteen lajien eriytymisessä. Lajeista voisi esimerkiksi löytyä samaan tehtävään liittyviä, mutta eri kudoksissa toimivia ja sen vuoksi eri tavoin ilmenevien geenien ryhmiä. Tällaiset geenit voisivat liittyä esimerkiksi jonkin elimistössä vaikuttavan hormonin tuottamiseen. Esimerkiksi eläinkokeiden käyttökelpoisuuden edellytyksenä on vastingeenien yhtenäinen toiminta, joten tällainen löydös olisi potentiaalisesti hyvinkin kiintoisa.

Harvinaiset vastingeenien toiminnalliset yhteydet voisivat viitata lajien eroihin, olipa vastingeenien ilmeneminen samankaltaista tai erilaista. Eri tavoin ilmenevien vastingeenien voidaan odottaa ryhmittyvän poikkeuksellisen pieniin yhteisryhmiin, koska ne ovat luultavasti samankaltaisesti ilmeneviä vastingeenejä harvinaisempia. Vastingeenien ilmenemisessä havaittavat erot voivat liittyä esimerkiksi lajien eriytymisessä tapahtuneisiin muutoksiin proteiinisekvenssien tasolla (ks. [62]). Mielenkiintoisia ihmisen ja hiiren gee-

nien eroja voisi löytyä esimerkiksi kasvutekijöistä, maksasta ja aivoista. Ihmisen ja hiiren aivoissa on rakenteellisia eroja, ja hiiren maksa käsittelee huomattavasti ihmistä tehokkaammin erilaisia myrkkyjä. Kasvutekijät vaikuttavat geneettisesti lähisukuisten lajien kokoeroon.

Vastingeenien ilmenemisen erot liittyvät luultavammin kudoskohtaisiin painotuseroihin kuin täysin eri kudoksiin liittyviin ja jyrkästi eri tavoin ilmeneviin geeneihin.

Yhteisryhmien jakautuminen. Ihmisen ja hiiren geenien toiminnallisiin eroihin voisi viitata myös kontingenssitaulun avulla havaittu yhteisryhmien 'jakautuminen'. Tällaisessa tapauksessa löydettäisiin kaksi yhteisryhmää, joissa esimerkiksi hiiren geenit ilmenevät samankaltaisesti (kuuluvat samaan reunaryhmään), mutta ihmisen vastingeenit jakautuvat kahteen eri tavoin ilmenevään geeniryhmään. Kyse voisi olla geeneistä, joiden ilmenemisessä on toisessa lajissa tapahtunut muutoksia geenien tehtävien muuttumisen tai monipuolistumisen seurauksena.

Geenien yhteydet lajien sisällä. Geenien säätelyverkostoja tunnetaan toistaiseksi huonosti. Usein samojen säätelyreittien geenit tai vastingeenit ilmenevät melko samankaltaisesti (ks. [63]). Tapauksissa, joissa näin ei ole, tuntemattomia säätelyverkoston jäseniä voi olla vaikeaa havaita. Assosiatiivisessa ryhmittelyssä samaan yhteisryhmään kuuluvilla saman lajin geeneillä voi olla hyvinkin erilaisia ilmenemisprofiileita. Niillä on kuitenkin ominaisuuksia, jotka yhdistävät ne ryhmän muihin geeneihin. Assosiatiivisessa ryhmittelyssä käytetään vastingeenien paritietoa, joten aineiston ulkopuolinen lisätieto vaikuttaa kummankin lajin geenien ryhmittelyyn. Tämän ansiosta voidaan löytää sellaisia tutkimusaineistoon kätkeytyviä piirteitä, joita ei havaittaisi tavanomaisilla ryhmittelymenetelmillä.

Olisi perusteltua ehdottaa tuntemattomia yhteyksiä yhteisryhmän geeneille kummankin lajin *sisällä*, jolloin yhteys löydettäisiin tavallaan toisen lajin kautta 'kiertämällä'. Tällainen tilanne olisi mahdollinen esimerkiksi silloin, jos lajin geeneillä on yhteinen tehtävä tietyssä kudoksessa, mutta muissa kudoksissa niiden tehtävät ovat eriytyneet. Tällöin geenien ilmenemisprofiilit voivat olla hyvinkin erilaisia. Ainakin toisen lajin geenien ilmenemisprofiilien homogeenisuus tukee oletusta yhteisryhmän geenien todellisista biologisista yhteyksistä lajien sisällä. Homogeenisuus voi vaikuttaa oleellisesti geenien ryhmittymiseen, ja sen ansiosta assosiatiivinen ryhmittely voi sallia toisen lajin ilmenemisprofiileille suuremman hajonnan.

On tietysti mahdollista, että yhdessä lajissa kytkeytyneiden geenien ortologeilla ei ole toiminnallisia yhteyksiä toisessa lajissa, mikä pienentää tällaisen lähestymistavan mahdollisuuksia. Tiettävästi geenien ryhmittelyä lajin sisällä ei ole aiemmin tehty paritiedon avulla.

Luku 6

Tulokset

Assosiatiivista ryhmittelyä käytettiin ihmisen ja hiiren 4499 putatiivisesti ortologisen geeniparin ryhmittelyyn. Kaikki tutkitut geenit ilmenivät ainakin yhdessä kudoksessa (AD>200). Tutkimusaineisto muodostui ihmisen geenien 46-ulotteisista ja hiiren geenien 45-ulotteisista ilmenemisprofiileista. Työn tavoitteena oli (i) etsiä genominlaajuisesta aineistosta lajien eroihin ja yhteyksiin liittyviä mielenkiintoisia vastingeeniryhmiä ja (ii) verrata assosiatiivisen ryhmittelyn toimintaa vaihtoehtoisiin menetelmiin. Lisäksi (iii) kartoitettiin assosiatiivisen ryhmittelyn soveltamisessa ja tulkinnassa esille tulevia ominaisuuksia ja ongelmakohtia.

6.1 Geeniryhmien yhteydet kontingenssitaululla

Tulosten tulkinnan helpottamiseksi koko datalla opetettua mallia vastaavalle kontingenssitaululle muodostettiin visualisointi (kuva 6.1), joka esittää yhteisryhmien koon poikkeamat satunnaisuudesta. Muodostettu kontingenssitaulu auttaa tulkitsemaan kahden organismin geenien toiminnan välisiä riippuvuuksia. Kullakin ihmisen geeniryhmällä on selvä satunnaisuudesta poikkeava yhteys vain joihinkin hiiren geeniryhmiin.

Assosiatiivisen ryhmittelyn löytämät potentiaalisesti mielenkiintoiset vastingeenien ryhmät löytyvät niistä yhteisryhmistä, joissa on satunnaisuudesta merkitsevästi ($p < 0,001$) poikkeava määrä ortologisia geenipareja. Menetelmän soveltamisessa keskitytään tällaisten yhteisryhmien tulkintaan. Tilastollisesta satunnaisuudesta poikkeavia yhteisryhmiä oli yhteensä 63 kappaletta, mikä vastaa noin 14 prosentin osuutta kaikista yhteisryhmistä.

Kontingenssitaulun järjestäminen

Kontingenssitaulun kukin rivi liittyy yhteen hiiren ja sarake vastaavasti yhteen ihmisen geeniryhmään. Assosiatiivisen ryhmittelymallin opettamisen aikana muodostettu kon-

tingenssitaulun rivien tai sarakkeiden osoittama ryhmien järjestys ei kerro vierekkäisten ryhmien yhteyksistä ja on tässä mielessä mielivaltainen.

Tulkinnan helpottamiseksi kontingenssitaulu voidaan pyrkiä järjestämään niin, että vierekkäisillä ruuduilla on jotakin yhteistä. Esimerkiksi samankaltaisia ilmenemisprofiileja sisältävät ryhmät voidaan pyrkiä kuvaamaan mahdollisimman lähekkäin. Ryhmän geenejä edustavan malliprofiilin määritteleminen voi kuitenkin olla hankalaa. Tässä työssä kontingenssitaulun ruudut on järjestetty simuloidulla jäähdytyksellä p-arvojen mukaan niin, että samankaltaisia p-arvoja omaavat yhteisryhmät kuvautuvat kontingenssitaululla mahdollisimman läheisiin ruutuihin. Monissa tapauksissa tämä johtaa myös ilmenemisprofiilien mielessä samankaltaisten yhteisryhmien päätymiseen lähekkäin.

On huomattava, että läheinen sijainti kontingenssitaululla ei järjestämisen jälkeenkään aina kerro geenien toiminnallisesta samankaltaisuudesta.

Kontingenssitaulun reunajakaumat

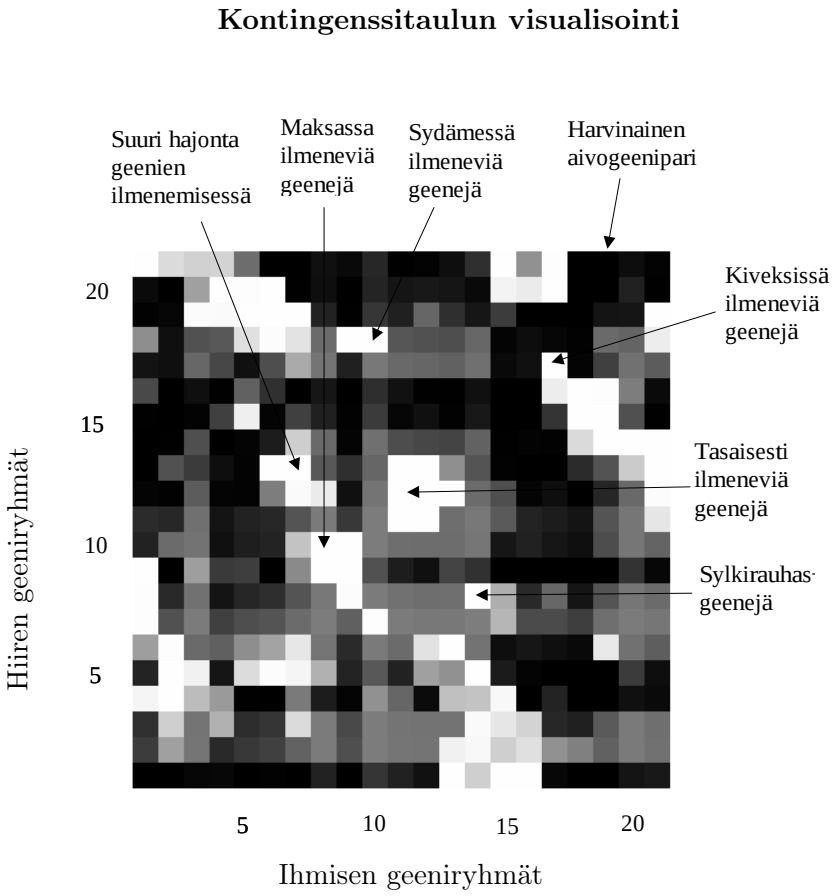
Ihmisen ja hiiren geeniryhmien koot eli kontingenssitaulun reunajakaumat on esitetty kuvissa 6.2 ja 6.3. Reunaryhmien välillä on tasajakaumaa suosivasta regularisoinnista huolimatta melko suuria kokoeroja. Pienimmässä ihmisen reunaryhmässä on vain 8 ja hiirellä vastaavasti 5 geeniä, kun suurimmat reunaryhmät sisältävät useita satoja näytteitä. Tutkimusaineisto sisältää piirteitä, jotka aiheuttavat reunaryhmien välille selviä kokovaihteluita.

6.2 Geeniryhmien tulkinta

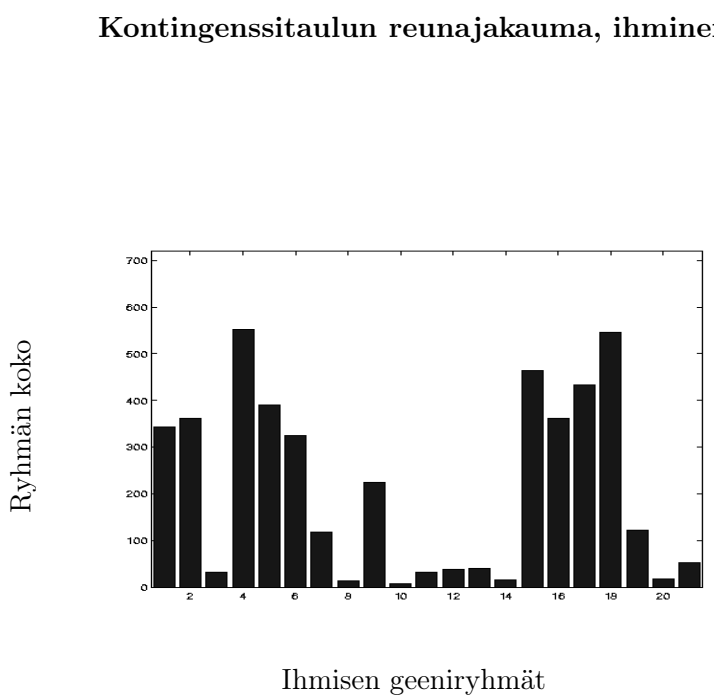
Geeniryhmän tulkinnan kannalta on oleellista tunnistaa kudokset, joissa geenien ilmeneminen on yhtenäistä ja toisaalta kudokset, joiden osalta menetelmä on sallinut ilmenemistasojen suuren hajonnan. Ryhmän geeneille voidaan ehdottaa niihin kudoksiin liittyviä toiminnallisia yhteyksiä, joissa geenien ilmeneminen on yhtenäistä ja muihin kudoksiin verrattuna voimakkaampaa.

Ilmenemistasojen tulkinnassa käytettiin alkuperäistä varianssinormalisoimatonta dataa, jossa ilmenemistasot ovat verrannollisia geeniä vastaavan lähetti-RNA:n pitoisuuteen tutkittavassa näytteessä. Eri kudoksiin liittyvät ilmenemistasojen hajonnat pyrittiin saamaan vertailukelpoisiksi tutkimusdatalle tehdyn varianssinormalisoinnin avulla. Ilmenemistasojen hajonnan vertailussa käytettiin varianssinormalisoidusta datasta laskettuja keskihajontoja.

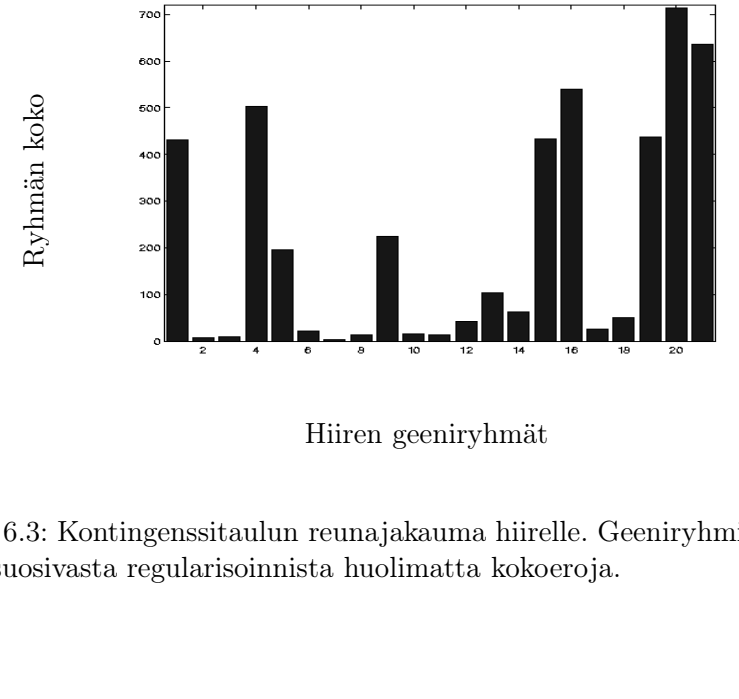
Lajien vertailussa keskityttiin yhteisiin kudoksiin. Ilmenemisprofiilien tulkinnan helpottamiseksi kudokset esitetään aina samassa järjestyksessä. Sopiva järjestys etsittiin simuloidulla jäähdytyksellä niin, että koko aineistossa samalla tavoin ilmenevät kudokset ovat mahdollisimman lähellä toisiaan. Lajeille yhteiset ja muut kudokset järjestettiin erikseen. Muiden kudosten osalta järjestystä ei käytetty tulosten tulkinnassa. Kudosten



Kuva 6.1: Ihmisen ja hiiren putatiivisesti ortologisilla geenipareilla opetetun assosiatiivisen ryhmittelymallin tuottaman kontingenssitaulun visualisointi. Mustat ruudut merkitsevät yhteisryhmiä, joissa on odottamattoman vähän geenipareja. Valkoisissa ruuduissa on satunnaisuuteen verrattuna poikkeuksellisen paljon geenipareja ($p < 0,001$). Harmaan sävyt näiden ääripäiden välillä kertovat, että yhteisryhmä ei ole tilastollisesti poikkeuksellinen. Taulun rivit ja sarakkeet on järjestetty simuloidulla jäähdytyksellä niin, että p-arvoltaan samankaltaiset ruudut ovat mahdollisimman lähekkäin. Kuvaan on merkitty tässä työssä lähemmin tarkasteltavat yhteisryhmät.



Kuva 6.2: Kontingenssitaulun reunajakauma ihmiselle. Geeniryhmien välillä on tasajakaumaa suosivasta regularisoinnista huolimatta kokoeroja.



Kuva 6.3: Kontingenssitaulun reunajakauma hiirelle. Geeniryhmien välillä on tasajakaumaa suosivasta regularisoinnista huolimatta kokoeroja.

järjestys on esitetty liitteessä A.

6.2.1 Geeniryhmien esittäminen

Biologisten vaihteluiden merkitys ilmenemistasojen ja niiden hajonnan arvioinnissa on tapauskohtaista, ja erilaiset geeniryhmän geenien ilmenemistä kuvaavat tunnusluvut on käsitettävä vain suuntaa antaviksi apuvälineiksi.

Tässä työssä geeniryhmän ilmenemisestä saadaan yksinkertainen ja suuntaa antava arvio käyttämällä malliprofiilina ryhmän ilmenemisprofiilien keskiarvoa ja vertailemalla ilmenemistasojen keskihajontaa eri kudoksissa. Ilmenemistasojen vaihteluita ryhmän sisällä voitaisiin arvioida lisäksi esimerkiksi selvittämällä ilmenemistasojen ääriarvot eri kudoksissa. Kudoskohtaiset ääriarvot kuvaavat kuitenkin ilmenemistasojen jakaumaa huonosti, ja niillä on puutteellinen yleistyskyky.

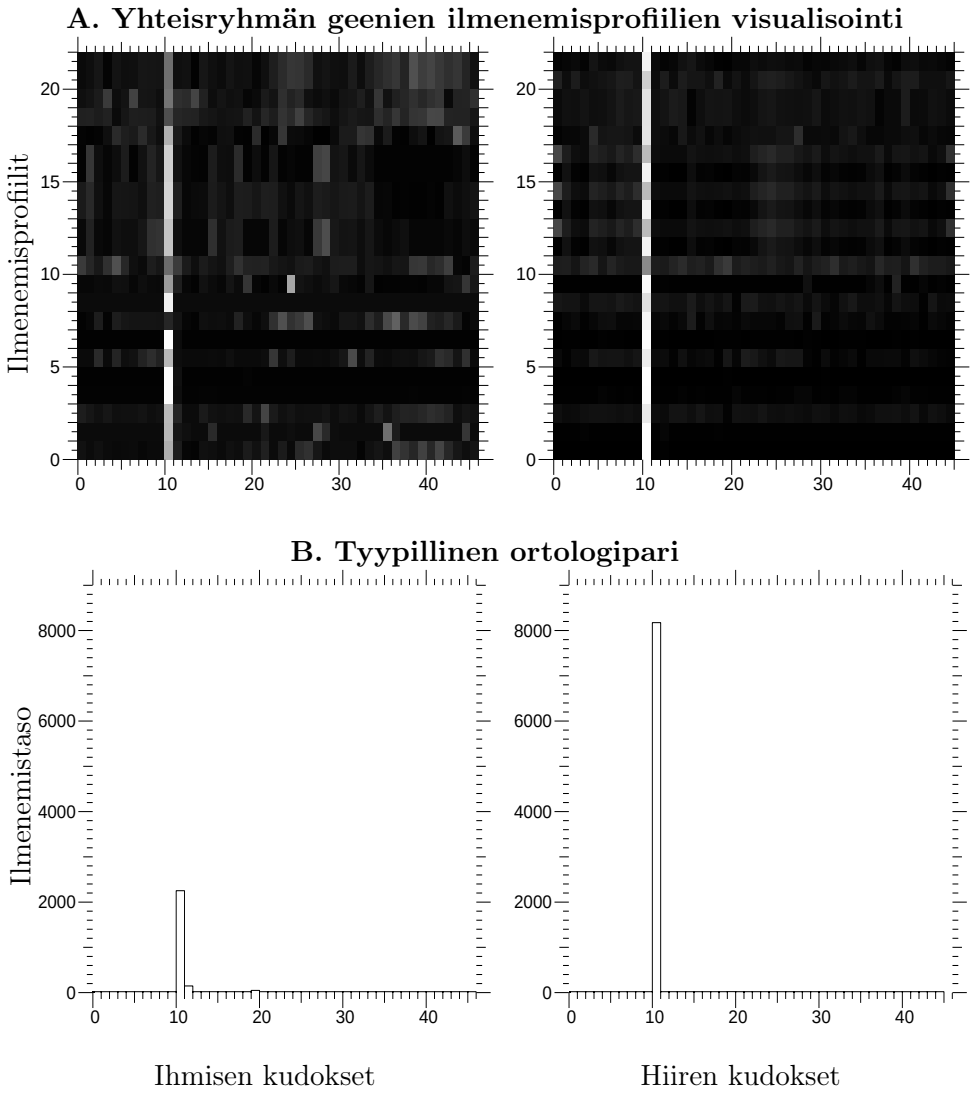
6.2.2 Yleinen yhteys ortologien ilmenemisessä

Poikkeuksellisen suuret yhteisryhmät viittavat sellaiseen vastingeenien ilmenemisessä esiintyvään yhteyteen, joka on jostakin syystä erityisen yleinen. Samankaltaisesti ilmenevien vastingeenien voidaan odottaa olevan aineistossa yleisempiä kuin eri tavalla ilmenevien vastingeenien. Assosiatiivisen ryhmittelyn avulla löydettiinkin useita esimerkkejä tällaisista vastingeenien yhteisryhmistä. Alla esitellään kolme tyypillistä, selkeimmin tulkittavaa ja potentiaalisesti mielenkiintoista tapausta.

Kiveksiin erikoistuneiden geenien ryhmä. Kuvassa 6.4 on erään poikkeuksellisen suuren yhteisryhmän geenien ilmenemisprofiilien visualisointi. Yhteisryhmän geenit ilmenevät voimakkaimmin kiveksissä, mutta ihmisen geenit ilmenevät hiiren vastingeeneihin verrattuna moninkertaisesti heikommin. Ilmenemistasoissa on selviä eroja myös lajien sisällä, ja havainnollisuuden vuoksi yhteisryhmän kaikkien geeniparien samanaikaisessa visualisoinnissa (kuva 6.4**A**) esitetään vain ilmenemisprofiilien muodot.

Kyseessä voi olla esimerkki molemmille lajeille yhteisiin, lajin säilymisen kannalta ratkaiseviin tehtäviin erikoistuneiden geenien ryhmästä. Yhteisryhmässä on myös sellaisia geenejä, joilla ei tiedetä olevan erityisiä kiveksiin liittyviä tehtäviä. Tätä voidaan ehdottaa, ja näillä geeneillä voi olla myös tuntematon toiminnallinen yhteys ryhmän muiden geenien kanssa.

Sydängeenien ryhmä. Kuvan 6.5 poikkeuksellisen suuren yhteisryhmän geenien ilmenemisprofiilien visualisointi paljastaa geenien voimakkaan ilmenemisen sydämessä. Ihmisen geenit ilmenevät voimakkaammin kuin hiiren geenit. Osa hiiren geeneistä ilmenee voimakkaasti myös muissa kudoksissa, joiden osalta käytettävissä ei ole tietoa ihmisen geenien ilmenemisestä. Yhteisryhmään on keskittynyt erityisesti lihasten kehitykseen liit-



Kuva 6.4: Kontingenssitaulun valkeasta ruudusta löydettiin poikkeuksellisen suuri määrä kiveksissä ilmeneviä vastingeenejä. **A.** Yhteisryhmän vastingeenien (21) ilmenemisprofiilien visualisointi. Vastingeenit ovat viereisissä kuvissa samalla rivillä, ja sarakkeet vastaavat kudoksia. Ihmisen geenit ilmenevät kiveksissä moninkertaisesti heikommin kuin hiiren geenit, ja geenien absoluuttiset ilmenemistasot vaihtelivat toisiinsa verrattuina moninkertaisesti. Visualisoinnissa esitetään ainoastaan ilmenemisprofiilien muodot, mikä tuo paremmin esille kivesten säännönmukaisesti voimakkaamman ilmenemisen muihin kudoksiin verrattuna. Pelkät ilmenemisprofiilien muodot saatiin esille normalisoimalla kunkin geenin ilmenemisprofiilin euklidiseksi pituudeksi 1. Harmaaskaala vaihtuu lineaarisesti mustasta valkoiseen välillä [0,1]. **B.** Tavanomainen esimerkki yhteisryhmän ilmenemisprofiileista on pari 4 (LocusID-tunnukset ihmiselle 7180 ja hiirelle 22024).

tyviä geenejä. Tämän yhteisryhmän geenien toiminta tunnetaan ilmeisesti melko hyvin, eikä ilmenemisprofiilien nojalla voida tehdä uusia ehdotuksia geenien toiminnasta.

Ylläpitogeenien yhteisryhmä. Assosiatiivinen ryhmittely voi löytää voimakkaasti tiettyihin kudoksiin erikoistuneiden geeniryhmien lisäksi myös muita poikkeuksellisen yleisellä tavalla ilmenevien geenien ryhmiä, jos niitä sisältyy tutkittavaan aineistoon.

Kontingenssitaulun keskellä kuvassa 6.1 havaittavan valkean alueen ruutuihin kuvautuneiden geenien ilmeneminen on samankaltaista, ja ne voidaan tulkita yhdeksi ryhmäksi. Yhteisryhmässä on pääasiassa ribosomaalisia proteiineja ja translaatiotekijöitä, sekä soluviestintään liittyvissä tehtävissä tarpeellisia proteiineja koodaavia geenejä. Nämä osallistuvat usein kaikille solutyypeille tärkeisiin geenien lukemiseen ja proteiinien valmistukseen liittyviin prosesseihin, ja niiden ilmeneminen eri kudoksissa on tasaista.

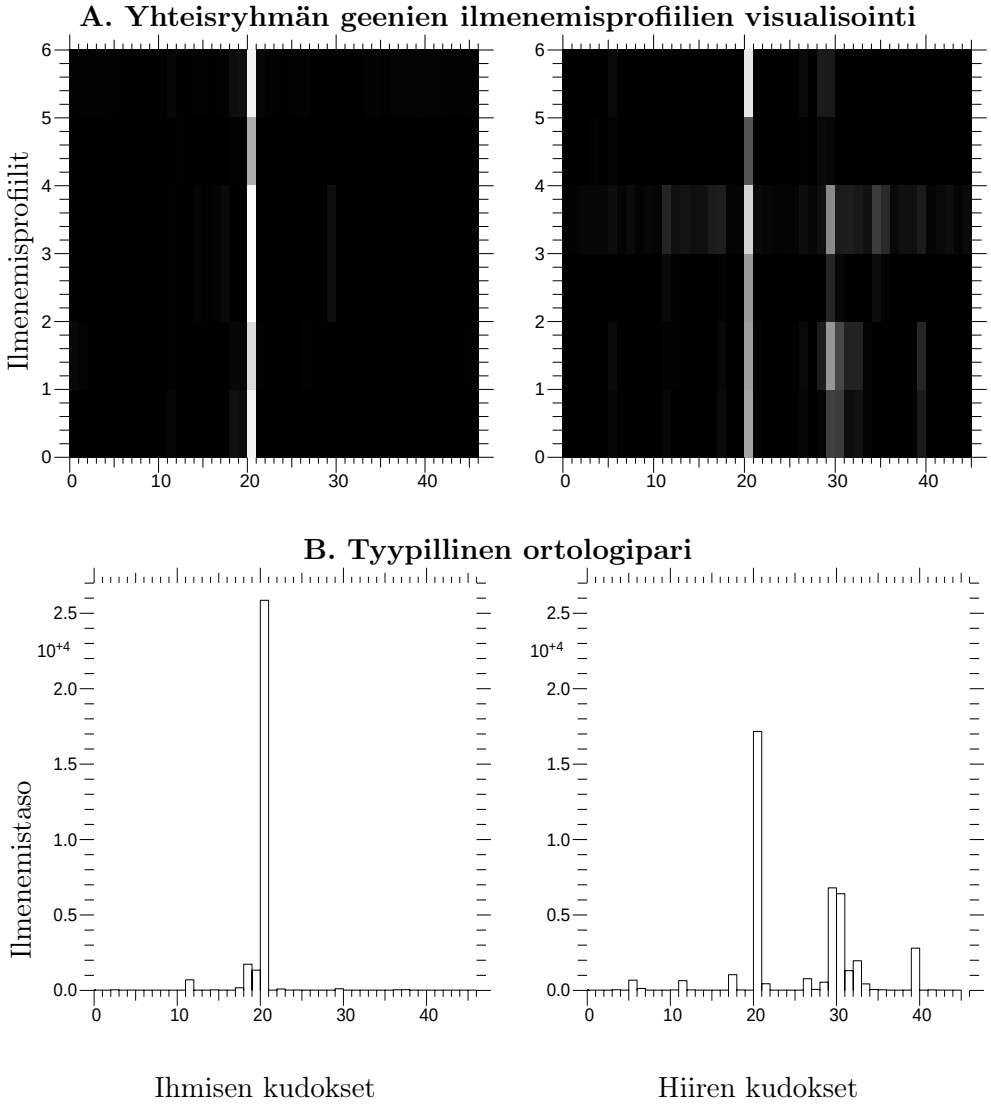
Tasaisesti kaikissa kudoksissa ilmeneviä geenejä voitaisiin mahdollisesti käyttää kontrolligeeneinä (ks. [70]) lajien vertailevissa tutkimuksissa. Niiden avulla voitaisiin mahdollisesti löytää kahden lajin geenien ilmenemistasoille sopiva normalisointitaso, joka tekisi eri lajeihin liittyvät mittaukset vertailukelpoisiksi. Tyypillisesti kontrolligeeneiksi pyritään valitsemaan sellaisia geenejä, jotka ilmenevät tasaisesti kaikissa annetuissa näytteissä. Kontrolligeenien on ilmettävä riittävän voimakkaasti, jotta niitä voitaisiin käyttää koetilanteissa varmasti ilmenevinä vertailukohteina muille geeneille. Löydetyn yhteisryhmän geeneillä on näitä ominaisuuksia, joskin ilmeneminen eri kudoksissa ei ole kaikissa tapauksissa aivan tasaista.

6.2.3 Harvinaiset yhteydet ortologien ilmenemisessä

Poikkeuksellisen pienet yhteisryhmät liittyvät vastingeenien harvinaisiin toiminnallisiin eroihin tai samankaltaisuuksiin. Erityisesti muuhun aineistoon verrattuna poikkeuksellisesti ilmenevät yksittäiset geenit eli 'outlierit' tulevat tällä tavoin esille. Tarkempia ennusteita tällaisten geeniparien luonteesta ei voida tehdä, ja löydökset on tulkittava tapauskohtaisesti. Osa 'outlier'-pisteistä voi johtua satunnaisista virheistä, mutta assosiatiivinen ryhmittely on perusteltu menetelmä kiinnostavien poikkeuksien etsimiseksi. Yhteisryhmän ja reunaryhmien yhtenäistä ilmenemistä voidaan käyttää lisätukena pyrittäessä erottamaan biologisesti kiinnostavat löydökset satunnaisista harhapisteistä.

Harvinainen yhteys aivogeenien välillä. Eräästä poikkeuksellisen pienestä yhteisryhmästä löytyi kiintoisa aivoissa toimivien geenien pari. Tämä oli yhteisryhmän ainoa geenipari. Tässä tapauksessa ihmisen geeni ilmenee erityisesti aivokudoksissa ja hiiren geeni tasaisen heikosti kaikissa kudoksissa (kuva 6.6**A**). Yhteisryhmän poikkeuksellisen pieni koko viittaa siihen, että tällä tavoin ilmenevät vastingeenit ovat erityisen harvinaisia.

Geenit ilmenevät samoissa kudoksissa kuin reunaryhmien geenit (kuva 6.6**B**). Tämä tukee sitä oletusta, että geenipari ei ole pelkkä mittausvirheistä tai ryhmien väljistä rajois-



Kuva 6.5: Poikkeuksellisen suuri sydämeen erikoistuneiden vastingeenien ryhmä. Visualisointi on selitetty kuvan 6.4 yhteydessä. **A.** Kaikki geenit ilmenevät voimakkaimmin sydämessä. Harmaaskaala sijoittuu lineaarisesti välille [0,27000]. **B.** Geenipari 0 on tyypillinen esimerkki yhteisryhmän ilmenemisprofiileista (LocusID-tunnukset ihmiselle 4151 ja hiirelle 17189).

ta johtuva satunnainen harhapiste. Geeniparin edustama harvinaisen yhteyden luonne voitaisiin tässä tapauksessa ennustaa myös reunaryhmien ilmenemisprofiilien perusteella.

Tulosten tulkinnassa on huomioitava, että aivokudoksissa hiiren ja ihmisen geenien ilmenemisessä voi olla eroja jo aivojen rakenteellisista eroista johtuen. Kirjallisuushakujen avulla voitiin tässä tapauksessa vahvistaa, että kyse on ihmisen ja hiiren alkionkehityksessä aktiivisesta aivogeenistä, joka aikuisilla yksilöillä ilmenee vain ihmisessä. Geeniparin ja sen kanssa tunnetussa toiminnallisessa yhteydessä olevien muiden geenien tarkempi tutkiminen voisi tuottaa lisätietoa lajien välisistä eroista.

Tämän tapaisten tutkimussuuntien ehdottaminen on eksploratiivisten menetelmien tavoitteena, joten löydös on rohkaiseva esimerkki assosiatiivisen ryhmittelymenetelmän mahdollisuuksista huomionarvoisten vastingeenien löytämiseksi. Löydöksen kiinnostavuutta lisää se, että ennakkoarvioiden mukaan lajien välisiä eroja voisi odottaa löytyvän aivoista.

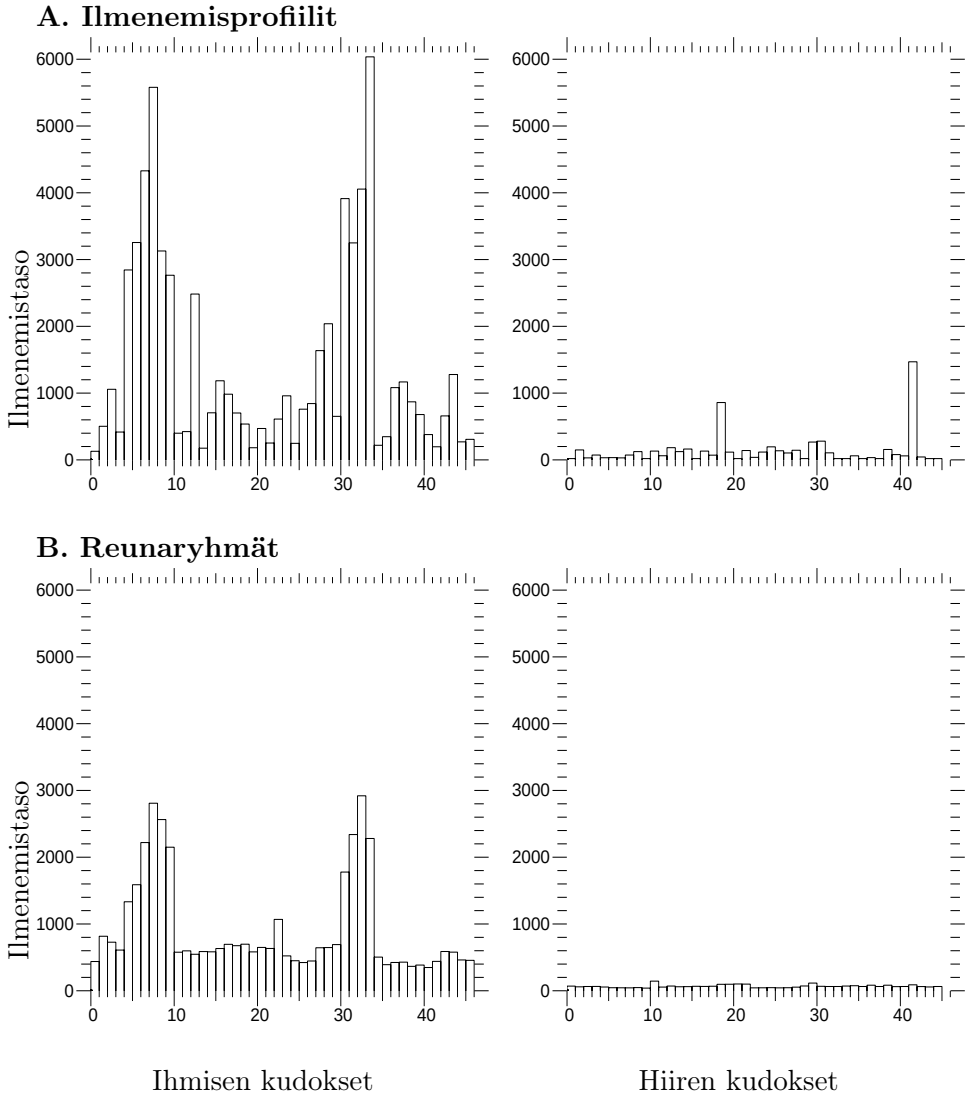
6.2.4 Geenien yhteydet lajin sisällä

Lajien toiminnallisen vertailun lisäksi yhteisryhmän geeneille voidaan ehdottaa yhteyksiä kummankin lajin sisällä.

Kiveksiin erikoistuneiden geenien yhteisryhmässä (kuva 6.4) oli geenejä, joiden osalta kiveksiin liittyvää tehtävää ei voitu varmentaa kirjallisuushakujen avulla. Tässä tapauksessa voitaisiin tutkia, onko yhteisryhmän geeneillä tuntemattomia kiveksiin liittyviä toiminnallisia yhteyksiä lajien sisällä. Usein selvästi tiettyyn kudokseen erikoistuneiden ja samalla tavoin ilmenevien vastingeenien toiminta tunnetaan hyvin, eikä uusia hypoteeseja yhteisryhmän geenien yhteyksistä voida muodostaa. Tällaiset tapaukset voidaan luultavasti löytää myös perinteisillä ryhmittelymenetelmillä. Sydämessä ilmenevä geeniryhmä (kuva 6.5) on tästä esimerkki. Assosiatiivinen ryhmittely kykenee kuitenkin löytämään myös vaikeammin havaittavia yhteyksiä.

Kaksi sylkirauhasgeeniä. Poikkeuksellisen suuressa kahden geeniparin yhteisryhmässä (kuva 6.7) on molekyylien pilkkomistehtäviin liittyviä geenejä. Molemmat ihmisen geenit ja toinen hiiren geeneistä ilmenevät voimakkaasti sylkirauhasissa. Toinen hiiren geeni ilmenee kilpirauhasessa.

Ensimmäinen ihmisen ja hiiren geenipari tuottaa hydrolyysissa eli eräässä molekyylien pilkkomistehtävässä tarvittavaa ainetta amylaasia. Geenien biologiset tehtävät eivät ilmeisesti ole homologisia geenisekvenssien samankaltaisuudesta huolimatta. Kumpikin geeni tuottaa hieman erilaista amylaasia, ja ihmisen amylaasigeeni ilmenee voimakkaammin haimassa. Toinen ihmisen geeni tuottaa eräiden molekyylien pilkkomisessa tarpeellista kallikreiinia, ja hiiren geeni erästä hermojen kasvutekijää, jolla on myös hydrolyysiin liittyvä tehtävä. Näiden geenien tuottamilla proteiineilla saattaa olla sama molekulaarinen funktio, mutta niiden biologinen merkitys on erilainen. Sekvenssihomologiaan perustuva vastingeenien määrittely on tässä mielessä biologisesti harhaanjohtava. Sen



Kuva 6.6: Esimerkki harvinaisesta yhteydestä ihmisen ja hiiren aivogeenien välillä. Vasemmanpuoleiset kuvat liittyvät ihmiseen ja oikeanpuoleiset hiireen. **A.** Geeniparin ilmenemisprofiilit (LocusID-tunnukset ihmiselle 1808 ja hiirelle 12934). **B.** Reunaryhmien keskiarvoprofiilit. Geeniparin samankaltainen ilmeneminen reunaryhmien kanssa tukee oletusta löydöksen biologisesta taustasta.

nojalla on kuitenkin mahdollista, että näillä ortologisilla geeneillä on yhteinen alkuperä, ja jopa osittain yhteisiä tehtäviä. Tässä tapauksessa yhteisryhmän geeneille onnistuttiin löytämään yhteys, joka liittyi molekyylien pilkkomisessa tarvittavien proteiinien koodaamiseen.

Samaan ryhmään päätyneiden ihmisen vastingeenien ilmenemisen homogeenisuus on saattanut vaikuttaa merkittävästi eri tavoin ilmenevien hiiren geenien ryhmittelyyn. Reunaryhmissä geenien keskimääräinen ilmeneminen on heikompaa kuin yhteisryhmässä ja ilmenemistasojen hajonta suurta verrattuna ilmenemisen absoluuttiseen tasoon. Onkin mahdollista, että satunnaisten tekijöiden vaikutus geenien päätymiselle samaan yhteisryhmään on ollut biologisia tekijöitä suurempi. Löydös saattaa kuitenkin olla esimerkki assosiatiivisen ryhmittelymenetelmän kyvystä hyvin eri tavalla ilmenevien, mutta osittain yhteisiin tehtäviin liittyvien geenien yhteyksien löytämiseksi lajin sisällä. Käyttämällä kahta hiiren geeniä tutkimuksen lähtökohtana voitaisiin ehkä saada lisätietoa esimerkiksi hermojen kasvutekijöiden ja molekyylien pilkkomiseen osallistuvien proteiinien yhteistyöstä.

Voimakkaasti maksassa ilmenevät geenit. Maksaan erikoistuneiden geenien yhteisryhmässä (kuva 6.8) oli erityisesti solun ulkoisiin toimintoihin liittyviä geenejä, jotka ilmenivät hyvin samankaltaisesti molemmissa lajeissa. Vaikka vastingeenien toiminta tunnetaan, kaikkien geenien yhteyksiä lajien sisällä ei saatu selville. Ryhmästä voisi löytyä ehdokkaita esimerkiksi jonkin maksan toimintaan liittyvän geenien säätelyverkoston puuttuviksi jäseniksi.

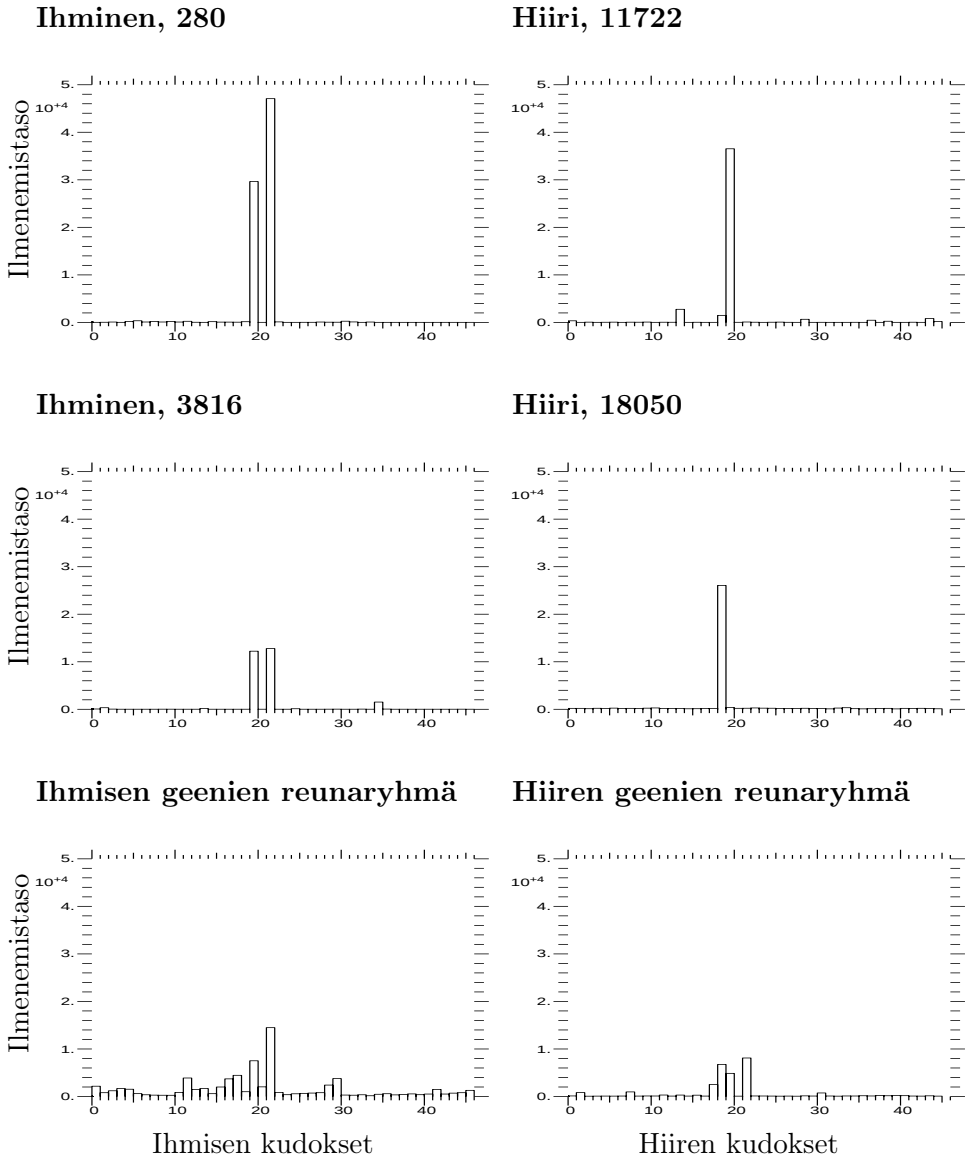
6.2.5 Muista mahdollisista sovelluksista

Tutkittavasta datasta ei löytynyt lupaavia esimerkkejä poikkeuksellisten yhteisryhmien jakautumisesta. Myöskään ortologien ilmenemiseen liittyviä yleisiä eroja tai harvinaisia samankaltaisuuksia ei löytynyt, mutta tällaisten tapausten löytyminen onkin vähemmän odotettavaa. Toisenlaisesta aineistosta saattaisi löytyä myös näitä piirteitä. On mahdollista, että niitä sisältyy myös tutkimusaineistoon, mutta assosiatiivisen menetelmän voimavarat kuluivat nyt voimakkaimmin esille tulevien geeniryhmien havaitsemiseen. Tiettyyn kudokseen erikoistuneet geenit, joita tässä työssä löydetään erityisen paljon, tunnetaan usein melko hyvin ja voitaisiin löytää myös tavanomaisilla menetelmillä. Siten niiden lisäarvo biologisen tutkimuksen kannalta saattaa olla muita tapauksia pienempi.

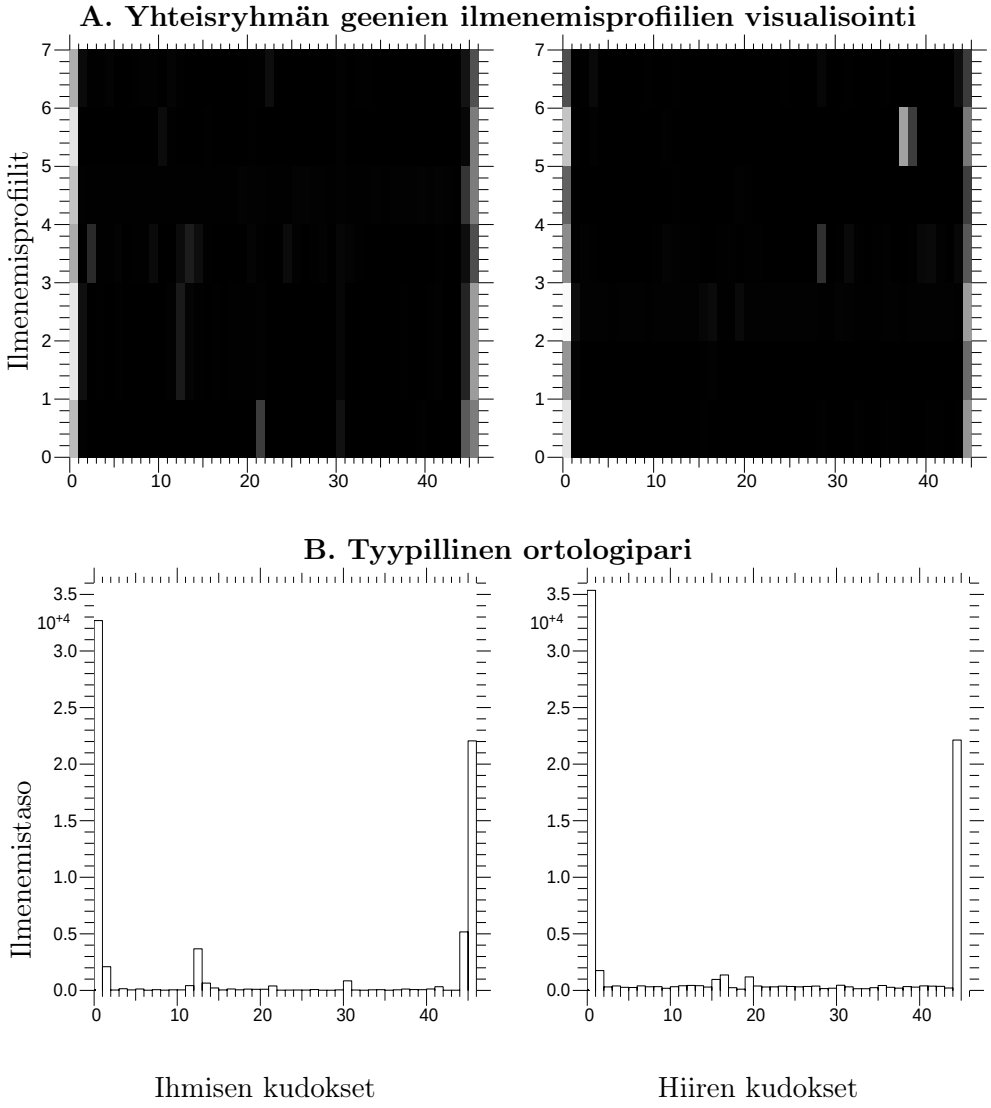
6.3 Vertailu vaihtoehtoisiin menetelmiin

Assosiatiivisen ryhmittelymenetelmän arviointia varten ihmisen ja hiiren geenit ryhmiteltiin kahdella vaihtoehtoisella menetelmällä, jotka on esitelty luvussa 4.

Assosiatiivisen menetelmän tuottamien ryhmien parametrisointi yksinkertaisina Voronoi-alueina tekee niistä VQ-IB-menetelmällä tuotettuihin ryhmiin verrattuna vähemmän



Kuva 6.7: Kaksi geeniparia sisältävä, mutta reunaryhmien kokoon nähden poikkeuksellisen suuri yhteisryhmä. Yhteisryhmän geenit osallistuvat mm. erilaisiin molekyylien pilkkomistehtäviin, ja ilmenevät toista hiiren geeniä lukuun ottamatta sylkirauhasissa. Ilmenemisprofiilin yllä mainitaan vastaavan geenin LocusID-tunnus. Alimmalla rivillä esitetään reunaryhmien keskiarvoprofiilit.



Kuva 6.8: Kontingenssitaulun valkeasta ruudusta löydettiin poikkeuksellisen suuri määrä maksassa ilmeneviä vastingeenejä. Visualisointi on selitetty kuvan 6.4 yhteydessä. **A.** Yhteisryhmän ilmenemisprofiilien visualisointi. Harmaaskaala sijoittuu lineaarisesti välille [0,36000]. **B.** Yhteisryhmälle tyypillinen profiilipari 2 (LocusID-tunnukset ihmiselle 5265, hiirelle 20703).

joustavia, mutta toisaalta yhtenäisiä. Riippuvuuksien mallintaminen tapahtuu ryhmien sisäisen homogeenisuuden kustannuksella, joten assosiatiivisen ryhmittelyn tuottamien ryhmien sisäisen hajonnan voidaan odottaa olevan suurempaa kuin K-means-menetelmällä tuotettujen ryhmien.

6.3.1 Riippuvuuksien mallintaminen

Kun kaksi aineistoa ryhmitellään toisistaan riippumattomasti tässä työssä käytetyllä K-means-menetelmällä, ja dataparien jakautumista yhteisryhmiin tarkastellaan kontingenssitaulun avulla, joissakin ruuduissa saattaa olla satunnaisuuteen verrattuna poikkeava määrä näytteitä. Tällainen ryhmittely ei kuitenkaan pyri kahden avaruuden välisten riippuvuuksien mallintamiseen, joka on assosiatiivisen ryhmittelyn tavoitteena. Diskretoidun aineiston ryhmittelyyn käytetty VQ-IB-menetelmä puolestaan mallintaa erityisen hyvin kahden datajoukon välisiä riippuvuuksia.

Sekä assosiatiivisen ryhmittelymenetelmän että VQ-IB:n tuottama parannus ryhmien riippuvuuden esittämisessä Bayes-faktorilla (5.1) mitattuna oli tilastollisesti merkitsevä K-means-menetelmään verrattuna ($p < 0,001$). VQ-IB onnistui odotetusti assosiatiivista ryhmittelyä paremmin riippuvuuksien mallintamisessa ($p < 0,02$). Vertailujen toteuttamiseen käytettiin kymmenkertaista ristiinvalidointia.

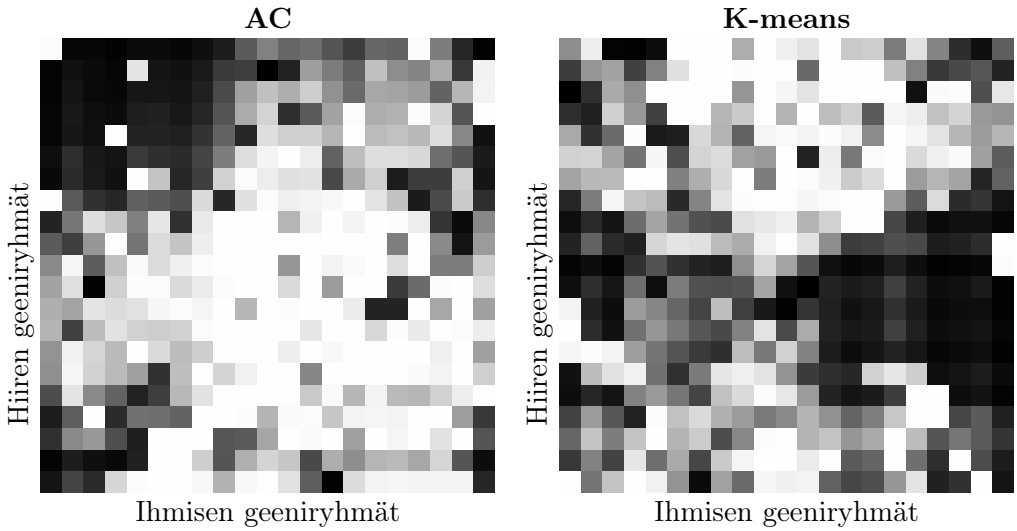
Assosiatiivisen ryhmittelyn paljastama riippuvuusrakenne voidaan havaita myös visuaalisesti. Assosiatiivisen ryhmittelyn tuottamassa kontingenssitaulussa huomattavasti useammat ruudut poikkeavat merkitsevästi riippumattomien reunajakaumien nollahypoteesista kuin K-means-menetelmällä tuotetussa (ks. kuva 6.9).

6.3.2 Reunaryhmien homogeenisuuden arviointi

K-means-menetelmä pyrkii muodostamaan mahdollisimman homogeenisia ryhmiä, mutta se ei pyri mallintamaan kahden avaruuden välisiä riippuvuuksia. VQ-IB mallintaa kahden aineiston riippuvuuksia, mutta ei huomioi muodostettavien ryhmien sisäistä homogeenisuutta. VQ-IB-menetelmällä tuotetut ryhmät eivät ole välttämättä yhtenäisiä. Tämän seurauksena tulosten yleistyskyky on huonompi ja tulkinta hankalampaa kuin assosiatiivisessa ryhmittelymenetelmässä.

Normalisoinnin ansiosta ilmenemisprofiilien hajonnat ovat vertailukelpoisia eri kudoksissa. Kudoskohtaisten varianssien summa kuvaa ilmenemistasojen kokonaisvaihtelun suuruutta tutkittavassa geeniryhmässä, ja tätä käytettiin reunaryhmien sisäistä homogeenisuutta arvioivana hajontamittana. Kolmella eri menetelmillä tuotettujen ryhmittelyjen sisäistä hajontaa kummankin aineiston ryhmissä verrattiin lisäksi satunnaisen ryhmittelyn tuottamien ryhmien sisäiseen hajontaan. Satunnainen ryhmittely antaa perustellun vertailukohdan arvioitaessa menetelmien kykyä satunnaisilmiöistä poikkeavien piirteiden löytämiseksi.

Tulokset (kuva 6.10) olivat ennakoitavissa. Assosiatiivisen ryhmittelymenetelmän ja K-



Kuva 6.9: Kontingessitaulun ruutujen poikkeaminen nollahypoteesin mukaisesta satunnaisuudesta, jossa reunajakaumat ovat riippumattomia. Musta merkitsee satunnaisuuteen yhtyvää geeniparien määrää. Mitä vaalempi ruutu on kyseessä, sitä voimakkaammin se poikkeaa satunnaisuudesta. Valkoisissa ruuduissa poikkeama on tilastollisesti merkitsevä ($p < 0,001$). Tässä p-arvot on saatu simuloimalla 1000 riippumattomien reunajakaumien mallia vastaavaa kontingenssitaulua, joihin menetelmien tuottamia kontingenssitauluja verrataan. Havainnollisen visualisoinnin tuottamiseksi kontingenssitaulun rivit ja sarakkeet on järjestetty simuloidulla jäähdytyksellä niin, että samankaltaisten p-arvojen ruudut ovat kontingenssitaululla mahdollisimman lähekkäin. Kuvassa esitettyjen kontingenssitaulun ruutujen sijainnit eivät ole vertailukelpoisia. Kuvasta nähdään kuitenkin, että assosiatiivisen ryhmittelyn (AC) tuottamassa kontingenssitaulussa on enemmän satunnaisuudesta poikkeavia yhteisryhmiä kuin K-means-menetelmällä tuotetussa. Harmaaskaalan merkitys poikkeaa kontingenssitaulun kaksisuuntaisesta visualisoinnista kuvassa 6.1

means-menetelmän tuottamien ryhmien sisäisessä hajonnassa ei ole tilastollisesti merkitseviä eroja. VQ-IB-menetelmällä tuotettujen ryhmien sisäinen hajonta oli näihin verrattuna kummassakin aineistossa suurempaa, eikä puolestaan selvästi poikennut satunnaisten ryhmien sisäisestä hajonnasta. Hajontamittojen virherajat ovat melko väljät, joten kovin vahvoja päätelmiä tulosten nojalla ei voida tehdä.

6.4 Assosiatiivisen ryhmittelymenetelmän arviointi

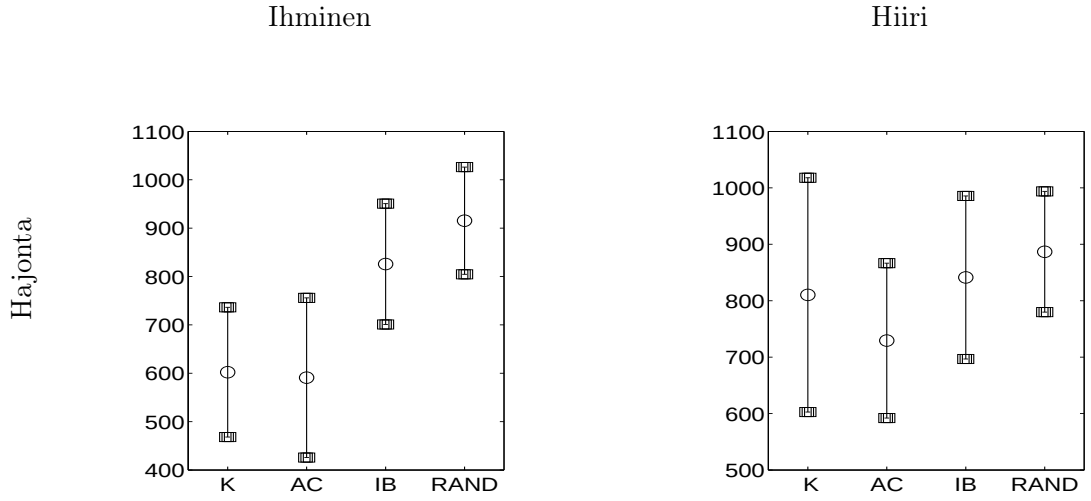
Ryhmien tulkinnan ongelmat. Assosiatiivinen ryhmittely ei kiinnitä huomiota ryhmien sisäiseen jakaumaan, eikä geeniryhmää voida aina esittää tyydyttävästi helppotulkintaisen malliprofiilin avulla. Ryhmä saattaa olla yksinkertaisesti liian suuri kuvatakseen aineiston oleellisia pienimuotoisempia vaihteluita, ja se voi muodostua useista selvästi erillisistä geeneistä tai geenikasaumista. Joissakin ryhmissä geeneillä ei ole lainkaan visuaalisesti havaittavaa yhtenäistä ilmenemistä (kuva 6.11). Tällaisissa tapauksissa pyrkimys ryhmälle 'tyypillisen' ilmenemisprofiilin esittämiseen voi tuottaa harhaanjohtavia tulkintoja ryhmän geenien ilmenemisestä ja yhteyksistä.

Voronoi-alueet voivat levittäytyä korkeaulotteisessa data-avaruudessa laajalle alueelle, jolloin ryhmän määrittelevä mallivektori saattaa olla hyvin erilainen ryhmän geenien ilmenemisprofiileihin verrattuna. Ryhmien parametrisointi Voronoi-alueina on yksinkertainen, mutta ei välttämättä optimaalisin tai intuitiivisesti paras tapa ryhmien määrittelemiseksi. Tarvittaessa laajennukset toisenlaisiin parametrisointeihin ja reunaryhmien rakenteisiin ovat ainakin periaatteessa suoraviivaisia.

Geenin ilmenemistaso on usein yhdistelmä monista samanaikaisista solunsisäisistä prosesseista, ja ilmenemistasojen ja niiden hajonnan merkitys on geenikohtaista. Vaikka assosiatiivinen ryhmittely ei ole erityisen herkkä esikäsittelymenetelmän valinnalle ja löytää implisiittisesti aineiston riippuvuuksien kuvaamiseen soveltuvan etäisyysmitan, mm. oppivien metriikoiden (ks.[37]) käyttö ryhmittelyn tukena voisi olla avuksi, jos tutkittavien geeniryhmien suuri hajonta vaikeuttaa tulosten tulkintaa.

Yhteisryhmien määrä. Poikkeuksellisen pienissä yhteisryhmissä geenien ja niiden yhteyksien yksityiskohtainen tarkastelu voidaan tehdä nopeasti. Suurten, jopa satoja geenipareja sisältävien yhteisryhmien manuaalinen tulkinta on työläämpää. Ryhmien määrän kasvattaminen voisi johtaa pienempien ja sisäisesti homogeenisempien geeniryhmien muodostumiseen. Tällaisten ryhmien tulkinta on helpompaa. Liian suuri ryhmien määrä voi aiheuttaa yhtenäisen geeniryhmän keinotekoisen pilkkoutumisen. Riittävän samankaltaisia näytteitä sisältävät ryhmät voidaan kuitenkin tulkita yhdeksi ryhmäksi, joten pienimuuotoisempien vaihteluiden kuvaaminen ei luultavasti tuhoaisi mahdollisuuksia suurten yhtenäisten geeniryhmien havaitsemiseen.

Tutkimusaineiston karsinta Tässä työssä vain osa tilastollisesti poikkeuksellisista yhteisryhmistä (yht. 63) valittiin lähempään tarkasteluun. Valinta tehtiin manuaalisesti



Kuva 6.10: Neljällä eri menetelmällä tuotettujen reunaryhmien homogeenisuutta arvioitiin laskemalla ryhmien tuotettujen ryhmien keskimääräinen sisäinen hajonta molemmissa data-avaruuksissa. 'K' merkitsee reunaryhmille riippumattomasti tuotettua K-means-ryhmittelyä, jonka voidaan odottaa tuottavan erityisen homogeenisia ryhmiä. AC merkitsee assosiatiivista ryhmittelyä ja IB puolestaan VQ-IB-menetelmää. Symbolilla 'RAND' merkitään satunnaista ryhmittelyä Tämä antaa hajonnalle eräänlaisen ylärajan. Ympyrät kuvaavat keskimääräisiä ryhmän sisäisiä vaihteluita ristiinvalidoinneissa (n=10) käytetylle arviointidatalle, ja neliöt osoittavat arvioidun 99 prosentin luottamusvälin. AC- ja K-means-ryhmät saattavat olla homogeenisempia kuin IB- ja RAND-ryhmät ($p<0{,}05$). Erot AC- ja K-means-menetelmien tai IB- ja RAND- menetelmien tuottamien ryhmien hajonnassa eivät ole tilastollisesti merkitseviä ($p>0{,}1$).

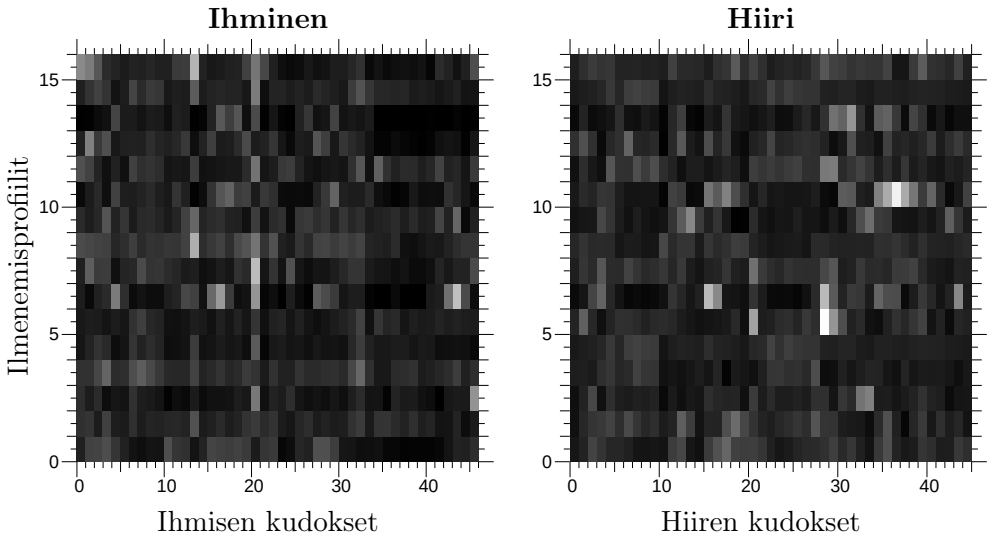
ilman yhtenäistä kriteeriä, koska geenien ilmenemiseen vaikuttavien lukuisten biologisten tekijöiden merkityksen arviointi onnistuu parhaiten manuaalisessa tarkastelussa. Eksploratiivisessa tutkimuksessa etsittävät hypoteesit eivät ole tarkasti rajattuja, eikä uuden menetelmän soveltamiselle ole olemassa vakiintuneita ja hyväksi havaittuja tapoja. Viisi poikkeuksellisen pientä yhteisryhmää (yht. 30) oli tyhjiä, ja muiden ryhmien karsinta tehtiin pääasiassa ilmenemisprofiilien suuren hajonnan perusteella. Jos satunnaisuudesta poikkeavien yhteisryhmien määrä on liian suuri manuaalisesti tarkasteltavaksi, voidaan keskittyä esimerkiksi mahdollisimman homogeenisiin yhteisryhmiin, joiden tulkinta on helpompaa. Toisaalta voitaisiin kiinnittää erityistä huomiota esimerkiksi tietyissä kudoksissa ilmeneviin geeniryhmiin.

Tutkimusaineistossa oli runsaasti tietoa geenien ilmenemisestä kudoksissa, joissa mittauksia oli tehty vain toisella lajilla. Lisätiedon toivottiin tukevan mielekkäiden ryhmien muodostumista, mutta toisaalta se saattaa johtaa tulkinnan kannalta ongelmallisempien korkeaulotteisten ryhmien muodostumiseen. Pelkkien yhteisten kudosten huomioinnissa menetelmä kykenisi ehkä keskittymään tarkemmin lajien yhteyksien kuvaamiseen. Joissakin tapauksissa tutkimusaineistossa oli yhteen geeniin liittyen useampia erilaisilla koetinjoukoilla tehtyjä mittauksia. Tällöin ryhmittelymallin opetukseen käytettävien geeniparien määrä kasvaa vastaavasti. Tällaisten näytteiden osuus oli vähäinen, mutta ne saattoivat vaikuttaa jonkin verran siihen, mitkä yhteisryhmät määriteltiin tilastollisesti poikkeaviksi. Vaikutus ryhmien koostumukseen lienee olematon. Tulkituissa tapauksissa huomattavin vääristymä liittyi kiveksiin erikoistuneeseen yhteisryhmään, jossa todellisia vastingeenejä oli 18 ja näitä vastaavia näyteaineiston geenipareja 22 kappaletta.

Menetelmän vahvuudet. Assosiatiivinen ryhmittely käyttää tietoa dataparien yhteisesiintymistä ja kykenee sen ansiosta huomioimaan geenien toiminnallisten yhteyksien kannalta oleellisten ja epäoleellisten vaihteluiden suunnat ryhmien rajojen määrittelyssä. Menetelmä käyttää siten implisiittisesti aineiston kuvaamiseen soveltuvaa etäisyysmittaa, mistä johtuen menetelmä ei ole erityisen herkkä tutkimusaineiston esikäsittelylle.

Yhteisryhmien muodostamisessa pyritään yhtenäisten ja helppotulkintaisten ryhmien muodostamiseen, mutta tavoitteesta voidaan joustaa riippuvuusmallintamisen hyväksi. Näin assosiatiivinen ryhmittely kykenee löytämään myös sellaisia geenien yhteyksiä, joita olisi vaikeaa havaita pelkän ilmenemisprofiilien vertailun avulla. Assosiatiivinen ryhmittely on perusteltu tapa muuhun aineistoon verrattuna poikkeuksellisella tavalla ilmenevien vastingeeniryhmien etsimiseksi.

Löydösten merkitys on arvioitava tapauskohtaisesti, sillä osa löydetyistä vastingeeneistä voi olla biologisen tutkimuksen kannalta hyödyttömiä satunnaisia harhapisteitä, jotka tulevat esille menetelmän tekniseen toteutukseen ja biologisiin vaihteluihin liittyvien virheiden johdosta.



Kuva 6.11: Esimerkki poikkeuksellisen suuresta yhteisryhmästä, jonka geenien ilmenemisessä ei ole visuaalisesti havaittavissa mainittavia yhteyksiä. Harmaaskaala sijoittuu lineaarisesti välille [0,12100].

Luku 7

Pohdinta

Verrattain kevytrakenteisena menetelmänä assosiatiivinen ryhmittely soveltuu eksploratiiviseen data-analyysiin, jossa etsitään aineiston ennalta tuntemattomia säännönmukaisuuksia ja suuntaviivoja jatkotutkimuksia varten. Se tarjoaa toimivan vaihtoehdon silloin, kun täydellinen todennäköisyysmallintaminen on liian raskasta tai vaatii liikaa ennakkotietämystä. Genomien toiminnallisessa vertailussa laajamittaisen eksploratiivisen tutkimuksen mahdollisuudet ovat selvät. Kyseessä on nuori tieteenala, ja lajien väliset yhteydet tunnetaan geenien toiminnan tasolla huonosti. Genominlaajuinen aineisto sisältää valtavan määrän tietoa, jonka tutkimuksessa on perusteltua keskittyä aluksi yleisten piirteiden havainnointiin ja tutkimuksen suuntaviivojen hahmotteluun. Löydösten merkitys jatkotutkimusten suuntamiselle voi olla suuri.

7.1 Virhelähteiden arviointi

Uuden menetelmän virhelähteet tunnetaan toistaiseksi huonosti. Työn tärkeimmät virhelähteet voivat liittyä menetelmän tekniseen toteutukseen ja tulosten tulkintaan. Virheitä voi aiheutua myös mm. tutkimusaineiston mittausvirheistä ja virheellisestä vastingeenien määrittelystä. Virhelähteet hankaloittavat tulosten tulkintaa, mutta niiden merkitys ei ole tämän tutkimuksen onnistumisen kannalta ratkaiseva. Muodostettavat hypoteesit on joka tapauksessa vahvistettava biologisissa tutkimuksissa.

7.2 Löydösten merkitys

Tietyllä tavalla ilmenevien vastingeenien määrä tutkimusaineistossa ei vielä kerro, onko kyseinen yhteys poikkeuksellinen. Tämän selvittämiseksi vastingeenejä on verrattava muihin näytteisiin. Assosiatiivisen menetelmän muodostamien yhteisryhmien geenipareilla on ortologian lisäksi geenien ilmenemiseen liittyvä tilastollisesti poikkeuksellinen yhteys.

Samalla tavoin ilmeneviä poikkeuksellisen yleisiä geenipareja löytyi odotetusti eniten. Tulos osoittaa, että assosiatiivinen ryhmittely kykenee löytämään säännönmukaisuuksia odotetulla tavalla ja tukee sitä oletusta, että vastingeenit usein ilmenevät samalla tavoin. Samankaltainen ilmeneminen on toisaalta tukena sille, että vastingeenien määrittäminen on tapahtunut oikein. Biologisissa tutkimuksissa voimakkaasti yhteen kudokseen erikoistuneet geenit on yleensä kartoitettu muita paremmin. Samalla tavoin ilmenevien vastingeenien arvo saattaakin biologisen tutkimuksen kannalta olla pienempi kuin muiden esille tulevien tapausten. Usein kyse on tiettyihin kudoksiin erikoistuneista geenipareista, jotka voitaisiin löytää helposti muillakin menetelmillä, esimerkiksi tarkastelemalla ilmenemisessä esiintyviä korrelaatioita ja karsimalla tasaisesti ilmenevät geenit pois tutkittavasta aineistosta. Tämä mahdollistaisi keskittymisen huonommin tunnettujen yhteyksien mallintamiseen. Tässä tutkimuksessa onnistuttiin myös selvästi yhdessä kudoksessa samankaltaisesti ilmenevien vastingeeniryhmien tapauksissa muodostamaan uusia hypoteeseja geenien yhteyksistä lajien sisällä. Eri tavoin ilmenevät vastingeenit voisivat olla biologisen tutkimuksen kannalta mielenkiintoisia, mutta niiden määrä on luultavasti pienempi. Tässä tutkimuksessa nämä saattoivat jäädä havaitsematta menetelmän voimavarojen kuluessa selvemmin esille tulevien yhteyksien mallintamiseen.

Assosiatiivinen ryhmittely osoittautui geenien ilmenemisdatan vertailevassa analyysissa käyttökelpoiseksi menetelmäksi, joka onnistuu yhdistämään vaihtoehtoisia menetelmiä paremmin kaksi hyvää ominaisuutta, riippuvuuksien mallintamisen ja helppotulkintaisten geeniryhmien muodostamisen. Luultavasti menetelmä ei ole erityisen herkkä aineiston esikäsittelyn ja malliparametrien valinnalle, koska se kykenee geenien paritiedon avulla etsimään implisiittisesti aineistojen riippuvuuksien soveltuvan etäisyysmitan. Löydösten nojalla voitiin muodostaa biologisia tulkintoja ja uusia hypoteeseja geenien yhteyksistä. Tulokset ovat suuntaa antavia, ja uudet hypoteesit on varmennettava biologisissa kokeissa ennen lopullisten päätelmien tekoa.

7.3 Lopuksi

Useampien sukulaislajien geneettinen vertailu voisi tuottaa selkeämmän kokonaiskuvan geenien yhteyksistä ja evoluution aikana tapahtuneiden muutosten luonteesta. Esimerkiksi geneettisesti hyvin lähisukuisten ihmisen ja simpanssin väliltä voisi löytyä mielenkiintoisia eroja. Muutaman geneettisesti kaukaisemman lajin, kuten seeprakalan ja banaanikärpäsen vertailu nisäkkäisiin voisi tuoda lisävaloa nisäkkäitä yhdistävien geneettisten piirteiden selvittämiseen. Geenisekvenssien voimakkaan eriytymisen johdosta vastingeenien määrittely on ongelmallista kehityshistoriallisesti kaukaisissa lajeissa. Myös tutkimusaineistojen yhteismitattomuus voi aiheuttaa vaikeuksia vertailtaessa lajeja, joilla ei ole juurikaan yhteisiä kudoksia. Usein voidaan kuitenkin löytää samoihin tehtäviin liittyviä geenejä, joiden ilmeneminen on samankaltaista. Vastingeenien yhteydet eivät aina ole suoraviivaisia, ja lajien vertailu geenien toiminnan tasolla on kiehtova tutkimusalue. Monilla ulkoisesti kovin erilaisilla lajeilla on hämmästyttävän paljon yhtäläisyyksiä perimän tasolla.

Liite A

Kudosten järjestys

Lista kudoksista, joiden osalta geenien ilmentymisdataa oli käytössä. 21 ensimmäistä kudosta ovat hiirelle ja ihmiselle yhteisiä. Kudokset on järjestetty simuloidulla jäähdytyksellä niin, että koko aineistossa samankaltaisesti ilmenevät kudokset ovat mahdollisimman lähellä toisiaan. Yhteiset (0-20) ja muut (21-45) kudokset on järjestetty erikseen. Muiden kudosten osalta järjestystä ei hyödynnetty tulosten tulkinnassa.

Kudoksen numero	Ihminen	Hiiri
0	Liver	liver
1	Kidney	kidney
2	Adrenal gland	adrenal gland
3	Placenta	placenta
4	DRG	drg
5	Spinal cord	spinal cord lower
6	Cerebellum	cerebellum
7	Cortex	cortex
8	Amygdala	amygdala
9	Caudate nucleus	striatum
10	Testis	testis
11	Trachea	trachea
12	Lung	lung
13	Spleen	spleen
14	Thymus	thymus
15	Ovary	ovary
16	Uterus	uterus
17	Prostate	prostate
18	Thyroid	thyroid
19	Salivary gland	salivary gland
20	Heart	heart
21	Pancreas	eye
22	Pituitary gland	olfactory bulb
23	Lymphoblastic molt-4	frontal cortex
24	Myelogenous k-562	hippocampus
25	THY-	hypothalamus
26	THY+	spinal cord upper
27	OVR278E	trigeminal
28	OVR278S	brown fat
29	Prostate Cancer	skeletal muscle
30	Corpus callosum	tongue
31	Thalamus	epidermis
32	Whole brain	bone
33	Fetal brain	bone marrow
34	Bukitts Raji	umbilical cord
35	Burkitts Daudi	bladder
36	K422	large intestine
37	WSU	small intestine
38	Ramos	stomach
39	GA10	snout epidermis
40	DOHH2	digits
41	HL60	mammary gland
42	A2058	lymph node
43	HUVEC	adipose tissue
44	Hep3b	gall bladder
45	Fetal liver	-

Kirjallisuutta

[1] The chipping forecast. *Nature Genetics Supplement*, 21:1–60, 1999.

[2] M. D. Adams, S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, et al. The genome sequence of *Drosophila melanogaster*. *Science*, 287:2185–2195, 2000.

[3] O. Alter, P. O. Brown, and D. Botstein. Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms. *PNAS*, pages 3351–3356, March 2003.

[4] S. Aparicio, J. Chapman, E. Stupka, N. Putnam, J. ming Chia, P. Dehal, A. Christoffels, S. Rash, S. Hoon, A. Smit, M. D. S. Gelpke, J. Roach, T. Oh, I. Y. Ho, M. Wong, C. Detter, F. Verhoef, P. Predki, A. Tay, S. Lucas, P. Richardson, S. F. Smith, M. S. Clark, Y. J. K. Edwards, N. Doggett, A. Zharkikh, S. V. Tavtigian, D. Pruss, M. Barnstead, C. Evans, H. Baden, J. Powell, G. Glusman, L. Rowen, L. Hood, Y. H. Tan, G. Elgar, T. H. anddagger Byrappa Venkatesh, D. Rokhsar, and S. Brenner. Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science*, 297:1301–1310, 2002.

[5] Y. Barash and N. Friedman. Context-specific bayesian clustering for gene expression data. *Journal of Computational Biology*, 9:169–191, 2002.

[6] M. Behr, M. Wilson, W. Gill, H. Salamon, G. Schoolnik, S. Rane, and P. Small. Comparative genomics of bcg vaccines by whole-genome DNA microarray. *Science*, 284:1520–1523, 1999.

[7] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–97, 1999.

[8] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler. Genbank. *Nucleic Acids Research*, 31(1):23–37, 2003.

[9] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler. Genbank. *Nucleic Acids Research*, 28:15–18, 2000.

[10] H. Bono and Y. Okazaki. Functional transcriptomes: comparative analysis of biological pathways and processes in eukaryotes to infer genetic networks among transcripts. *Current Opinion in Structural Biology*, 12:355–361, 2002.

[11] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. A. Ball, H. Causton, T. Gaasterland, P. Glenisson11, F. Holstege, I. Kim, V. Markowitz, J. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame) - toward standards for microarray data. *Nature Genetics*, 29(4):365–371, 2001.

[12] C. elegans Sequencing Consortium. Sequence and analysis of the genome of C. elegans. *Science*, 282:2012–2018, 1998.

[13] N. A. Campbell, J. B. Reece, and L. G. Mitchell. *Biology*. Benjamin/Cummings, 2001. Sixth edition.

[14] S. B. Carroll. Genetics and the making of homo sapiens. *Nature*, 422:849–857, 2003. Review.

[15] S. Chu, J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and L. Herskowitz. The transcriptional program of sporulation in budding yeast. *Science*, 282:699–705, 1998.

[16] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, New York, 1991.

[17] P. J. Davis. Leonhard euler's integral: A historical profile of the gamma function. in memoriam: Milton abramowitz. *American Mathematical Monthly*, 66:849–869, 1959.

[18] J. L. DeRisi, V. R. Iyer, and P. O. Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278:680–686, 1997.

[19] M. Diehn and D. Relman. Comparing functional genomic datasets: lessons from dna microarray analyses of host-pathogen interactions. *Current Opinion in Microbiology*, 4:95–101, 2001.

[20] R. Doolittle. Microbial genomes opened up. *Nature*, 392:339–342, 1997. News feature.

[21] R. D. (editor). Functional genomics. *Nature*, 405:819–846, 2000. Nature Insight.

[22] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences, USA*, 95:14863–14868, 1998.

[23] R. M. Ewing and J.-M. Claverie. EST databases as multi-conditional gene expression datasets. In *Proc. of Pacific Symposium on Biocomputing*, volume 5, pages 427–439, 2000.

[24] I. Famili and B. O. Palsson. Systemic metabolic reactions are obtained by singular value decomposition of genome-scale stoichiometric matrices. *Journal of Theoretical Biology*, 224:87–96, 2003.

[25] R. Fisher. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1934. Fifth edition.

[26] C. Gaillardin, G. Duchateau-Nguyen, F. Tekaia, B. Llorente, S. Casaregola, C. Toffano-Nioche, M. Aigle, F. Artiguenave, G. Blandin, M. Bolotin-Fukuhara, E. Bon, P. Brottier, J. de Montigny, B. Dujon, P. Durrens, A. Lepingle, A. Malpertuy, C. Neuveglise, O. Ozier-Kalogeropoulos, S. Potier, W. Saurin, M. Termier, M. Wesolowski-Louvel, P. Wincker, J. Souciet, and J. Weissenbach. Genomic exploration of the hemiascomycetous yeasts: 21. comparative functional classification of genes. *FEBS Letters*, 487:134–149, 2000.

[27] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, USA, 2003. Second edition.

[28] A. Goffeau et al. The yeast genome directory, 1997. Nature supplement.

[29] G. H. Golub and C. F. V. Loan. *Matrix Computations*. Johns Hopkins University Press, London, 1996. Third edition.

[30] I. Good and J. Crook. The Bayes/non-Bayes compromise and the multinomial distribution. *Journal of American Statistical Association*, 69:711–720, 1974.

[31] I. J. Good. On the application of symmetric Dirichlet distributions and their mixtures to contingency tables. *Annals of Statistics*, 4(6):1159–1189, 1976.

[32] R. Hardison, J. Oeltjen, and W. Miller. Long human-mouse sequence alignments reveal novel regulatory elements: a reason to sequence the mouse genome. *Genome Res.*, 7:959–966, 1997.

[33] A. J. Hartemink, D. K. Gifford, T. S. Jaakkola, and R. A. Young. Combining location and expression data for principled discovery of genetic regulatory network models. In *Proc. of Pacific Symposium on Biocomputing*, volume 7, pages 462–473, 2002.

[34] S. Haykin. *Neural Networks, a Comprehensive Foundation*. Prentice Hall, New Jersey, second edition, 1999.

[35] D. Hosack, G. D. Jr., B. Sherman, H. Lane, and R. Lempicki. Identifying biological themes within lists of genes with ease. *Genome Biology*, 4(R70), 2003.

[36] J. L. Jiménez, M. P. Mitchell, and J. G. Sgouros. Microarray analysis of orthologous genes: conservation of the translational machinery across species at the sequence and expression level. *Genome Biology*, 4(R4), 2002.

[37] S. Kaski and J. Sinkkonen. Metrics that learn relevance. In *Proceedings of IJCNN-2000, International Joint Conference on Neural Networks*, volume V, pages 547–552. IEEE Service Center, Piscataway, NJ, 2000.

[38] M. Kato-Maeda, J. Rhee, T. Gingeras, H. Salamon, J. Drenkow, N. Smittipat, and P. Small. Comparing genomes within the species *Mycobacterium Tuberculosis*. *Genome Research*, 11:547–554, 2001.

[39] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, 2001. Third edition.

[40] B. Koop. Human and rodent dna sequence comparisons: A mosaic model of genomic evolution. *Trends Genet.*, 11:367–371, 1995.

[41] E. Lander. The new genomics: Global views of biology. *Science*, 274:536–539, 1996. News feature.

[42] E. Lander et al. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[43] R. J. Lipschutz, S. P. A. Fodor, T. R. Gingeras, and D. J. Lockhart. High density synthetic oligonucleotide arrays. *Nature Genetics Supplement*, 21(1):20–24, 1999.

[44] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Horton, and E. L. Brown. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology*, 14:1675–1680, 1996.

[45] B. Ma, J. Tromp, and M. Li. Patternhunter: Faster and more sensitive homology search. *Bioinformatics*, 18(3):440–445, 2002.

[46] O. Monni, S. Hautaniemi, and O. Kallioniemi. Geenisirutekniikka ja siihen liittyvä bioinformatiikka. *Duodecim*, 118:1157–1166, 2002. (In Finnish).

[47] Mouse Genome Sequencing Consortium. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420:520–562, 2002.

[48] K. Pruitt, T. Tatusova, and D. Maglott. NCBI reference sequence project: update and current status. *Nucleic Acids Research*, 31:34–37, 2003.

[49] J. Quackenbush. Computational analysis of microarray data. *Nature Reviews Genetics*, 2:418–427, 2001.

[50] J. Quackenbush. Microarray data normalization and transformation. *Nature Genetics Supplement*, 32:496–501, 2002.

[51] Y. D. Rubinstein and T. Hastie. Discriminative vs. informative learning. In D. Heckerman, H. Mannila, D. Pregibon, and R. Uthurusamy, editors, *Proc. ACM KDD*, pages 49–53. AAAI Press, 1997.

[52] W. Rudin. *Principles of mathematical analysis*. Mc-Graw-Hill, Singapore, 1976. Third edition.

[53] R. Sandberg and R. Yasuda. From the cover: Regional and strain-specific gene expression mapping in the adult mouse brain. *Proceedings of the National Academy of Sciences of the USA*, 97:11038–11043, 2000.

[54] M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270:467–470, 1995.

[55] G. K. Schoolnik. Functional and comparative genomics of pathogenic bacteria. *Current Opinion in Microbiology*, 5:20–26, 2002. Review.

[56] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17(Suppl 1):243–252, 2003.

[57] J. Sinkkonen and S. Kaski. Clustering based on conditional distributions in an auxiliary space. *Neural Computation*, 14:217–239, 2002.

[58] J. Sinkkonen, S. Kaski, and J. Nikkilä. Discriminative clustering: Optimal contingency tables by learning metrics. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the ECML'02, 13th European Conference on Machine Learning*, pages 418–430, Berlin, 2002. Springer.

[59] J. Sinkkonen, J. Nikkilä, L. Lahti, and S. Kaski. Associative clustering by maximizing a Bayes factor. Technical Report A68, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland, 2003.

[60] N. Slonim. *The information bottleneck: theory and applications*. PhD thesis, Hebrew University, Jerusalem, 2002.

[61] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–255, 2003.

[62] A. I. Su, M. P. Cooke, K. A. Ching, Y. Hakak, J. R. Walker, T. Wiltshire, A. P. Orth, R. G. Vega, L. M. Sapinoso, A. Moqrich, A. Patapoutian, G. M. Hampton, P. G. Schultz, and J. B. Hogenesch. Large-scale analysis of the human and mouse transcriptomes. *PNAS*, 99:4465–4470, 2002.

[63] C. Tilstone. Vital statistics. *Nature*, 424:610–612, 2003. News feature.

[64] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *37th Annual Allerton Conference on Communication, Control, and Computing*, pages 368–377, Urbana, Illinois, 1999.

[65] J.-F. Tomb, O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D. Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, E. K. Hickey, D. E. Berg, J. D. Gocayne, T. R. Utterback, J. D. Peterson, J. M. Kelley, M. D. Cotton, J. M. Weidman, C. Fujii, C. Bowman, L. Watthey, E. Wallin, W. S. Hayes, M. Borodovsky, P. D. Karp, H. O. Smith, C. M. Fraser, and J. C. Venter. The complete genome sequence of the gastric pathogen helicobacter pylori. *Nature*, 388:539–547, 1997.

[66] J. Townsend, D. Cavalieri, and D. Hartl. Population genetic variation in genome-wide gene expression. *Molecular Biology and Evolution*, 20:955–963, 2003.

[67] O. G. Troyanskaya, K. Dolinski, A. B. Owen, R. B. Altamna, and D. Botstein. A Bayesian framework for combining heterogeneous data sources for gene function prediction (in saccharomyces cerevisiae). *Proceedings of National Academy of Sciences*, 100(14):8348–8353, 2003.

[68] V. E. Velculescu, L. Zhang, B. Vogelstein, and K. W. Kinzler. Serial analysis of gene expression. *Science*, 270:484–487, 1995.

[69] O. G. Vukmirovic and S. M. Tilghman. Exploring genome space. *Nature*, 405:820–822, 2000.

[70] J. A. Warrington, A. Nair, M. Mahadevappa, and M. Tsyganskaya. Comparison of human adult and fetal expression and identification of 535 housekeeping/maintenance genes. *Physiol. Genomics*, 2:143–147, 2000.

[71] M. Werner-Washburne, B. Wylie, K. Boyack, E. Fuge, J. Galbraith, J. Weber, and G. Davidson. Comparative analysis of multiple genome-scale data sets. *Genome Research*, 12:1564–1573, 2002.

[72] D. Wheeler, D. Church, S. Federhen, A. Lash, T. Madden, J.U.Pontius, G. Schuler, L. Schrimi, E. Sequeira, T. Tatusova, and L. Wagner. Database resources of the national center for biotechnology. *Nucleic Acids Research*, 31(1):28–33, 2003.

[73] F. Yates. Contingency tables involving small numbers and the χ^2 test. *Journal of the Royal Statistical Society Supplement*, 1:217–239, 1934.

[74] Z. Zhang, S. Schwartz, L. Wagner, and W. Miller. A greedy algorithm for aligning dna sequences. *Journal of Computational Biology*, 7:203–214, 2000.