Geenieksprессioiden klusterointi

Katja Saarela
Katja.Saarela@cs.helsinki.fi

Klusterointimenetelmät-seminaari
Helsingin yliopisto, tietojenkäsittelytieteen laitos
Raportti C-2002-54, s. 64-75, marraskuu 2002

Tiivistelmä

Tässä työssä tarkastellaan geenien ekspressiodatan klusterointia, jonka avulla pyritään selvittämään tuntemattomien geenien toimintaa. Geeni on perinnällistä ominaisuutta ohjaava DNA-jakso, joka sisältää tiedon tietyn proteiinin rakenteesta. Geenin ekspressoituminen tarkoittaa tuon tiedon käyttämistä proteiinin valmistuksessa. DNA-mikrosirutekniikan avulla voidaan tutkia samanaikaisesti kymmenien tuhansien geenien ekspressiotasot solu- tai kudosnäytteessä. Klusteroinnin tarkoituksena on löytää toiminnaltaan samankaltaiset geenit. Klusterointialgoritmina on perinteisesti käytetty hierarkkista klusterointia, mutta tässä työssä esitellään myös muita ekspressiodatan klusterointiin soveltuvia menetelmiä. Lopuksi tarkastellaan vielä lyhyesti vaihtoehtoisia ja täydentäviä menetelmiä ekspressiodatan klusteroinnille.

1 Geeni, ekspressio ja DNA–mikrosirutekniikka

Lähes kaikki ihmisen geenien DNA-sekvenssit on tunnistettu, ja ne on saatavilla tietokannoista. Useimpien geenien merkitys tunnetaan kuitenkin huonosti. Ekspressiodatan klusteroinnin avulla voidaan saadaan tietoa tuntemattomien geenien toiminnasta, geenien säätelystä ja säätelyryhmistä sekä ympäristövaikutuksista. Geenien toiminnan täsmällinen tunteminen olisi oleellista lääketieteellisten ongelmien ratkaisussa, kuten esimerkiksi lääkeainekehityksessä ja diagnostiikassa. Yksi mahdollisuus geenien toiminnan selvittämiseksi on tutkia, mitkä geenit toimivat yhdessä. Tämä voidaan muotoilla klusterointiongelmaksi seuraavasti: luokittele geenit toimintansa perusteella ryhmiksi siten, että ryhmään kuuluvilla geeneillä on keskenään enemmän toiminnallista samankaltaisuutta kuin siihen kuulumattomilla geeneillä.

Tämä luku toimii johdantona mikrosirutekniikan käyttöön geenieksprессioiden mittauksessa. Biologisen johdannon tarkoituksena on esitellä, millaista ja millaisin menetelmin hankittua dataa klusterointialgoritmeille annetaan syötteeksi.

1.1 Geenit ja ekspressio

Ihmisen perinnöllinen tieto on kromosomeissa. Kromosomit ovat DNA-kaksoiskierrettä. DNA-kaksoiskierre koostuu emäspareista siten, että adenosiinin (A) parina on tymiini (T) ja cytosiinin (C) parina on guaniini (G). Kromosomi sisältää sekä koodaavia osia, geenejä, että ei-koodaavia osia, eksoneita. Geeni on perinnällistä ominaisuutta ohjaava DNA-jakso, joka sisältää tiedon tietyn proteiinin rakenteesta. [Alb02]

Geenin tyypillisin tehtävä on proteiinin valmistuksessa tarvittavan tiedon säilytys. Tieto on talletettu DNA-kaksoiskierteeseen, josta se voidaan lukea kierre avaamalla. Kun kierre avataan ja tietoa aletaan käyttää hyödyksi, sanotaan geenin olevan ekspressoitunut. Kierteelle voidaan rakentaa komplementti, RNA, emästen liittymissääntöjen mukaan ja tälle vastaavasti komplementti. Jälkimmäinen komplementti on identtinen alkuperäisen geenisekvenssin kanssa ja tätä kutsutaan proteiiniksi. Proteiinit toimivat useimmiten entsyymeinä. Entsyymi katalysoi eli nopeuttaa jonkin tietyn reaktion tapahtumista. Oleellista on huomata, että kaikki solut sisältävät kaikki geenit, mutta ainoastaan tietyt geenit toimivat eli ekspressoituvat tietyissä soluissa [Oll02].

Elimistössä ravintoaineista tuotetaan elintoimintoihin tarvittavaa energiaa energia-metabolian avulla, jossa ravintoaineet pilkkoutuvat ja muuttuvat energiaa sisältäviksi molekyyleiksi. Tämä muodostaa reaktioista koostuvan metaboliaverkon, jossa kutakin reaktiota katalysoi jonkin geenin koodaama entsyymi. Näin siis geenien, proteiinien ja elimistön biokemiallisten reaktioiden välillä on yhteys: geenin DNA-sekvenssistä saadaan RNA, josta saadaan proteiini, joka toimii entsyyminä ja katalysoi tiettyä reaktiota. Tämä yhteys geenin ja proteiinin välillä johtaa seuraaviin päätelmiin:

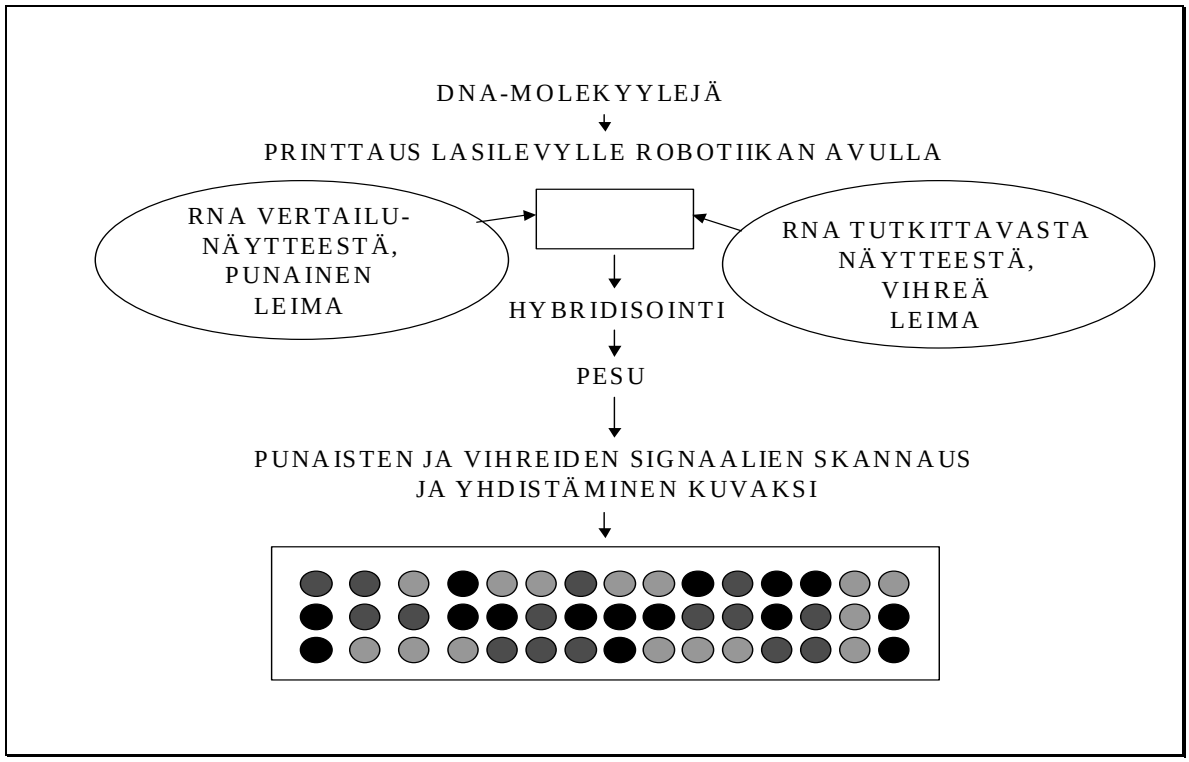
- jos halutaan selvittää geenien toiminnallisuutta, voidaan tarkastella niiden koodaamien entsyymien katalysoimia reaktioita ja
- vastaavasti geenien toimintaa tarkastelemalla voidaan saada tietoa organismissa tapahtuvista reaktioista.

Yksi tapa selvittää tuntemattomien geenien toimintaa on tutkia, mitkä geenit toimivat yhdessä eli mitkä geenit ovat ekspressioprofiililtaan samanlaisia. Mutta miten geenitason tietoa voidaan hankkia?

1.2 DNA-mikrosirutekniikka

Perinteisessä geenitutkimuksessa on tutkittu kerrallaan yhden geenin toimintaa ja merkitystä. DNA-mikrosirutekniikka on suhteellisen uusi menetelmä, jonka avulla voidaan tutkia samanaikaisesti kymmenien tuhansien geenien ekspressiotasot solu- tai kudosnäytteessä. DNA-mikrosirutekniikka on erittäin suosittua monilla biologian osa-alueilla solu- ja molekyylibiologiasta kliiniseen patologiaan ja immunologiaan. Lisäksi biotekninen ja lääketeollisuus hyödyntää DNA-mikrosiruja uusien lääkkeiden kehitystyössä ja testaamisessa. Seuraavassa esitellään DNA-mikrosirutekniikka pääpiirteissään. Menetelmän tunteminen on tärkeää, jotta saadaan käsitys klusterointialgoritmille annettavan datan tarkkuudesta.

Menetelmään tutustuminen kannattaa aloittaa tutkimalla kuvaa 1 sivulla 66. Ensimmäisessä vaiheessa valitaan tutkittavat geenit, jotka ovat siis DNA-molekyylejä. Samat geenit G_1-G_g otetaan useasta eri koetilanteesta K_1-K_k, jolloin muodostuu kuvan alalaidassa oleva taulukko. Itse asiassa kyseessä on $g \times k$ matriisi, jossa g on tarkasteltavien geenien ja k tarkasteltavien kokeiden määrä. Koska matriisi on kooltaan suuri, ei pelkkä silmämääräinen tarkastelu riitä ekspressioprofiililtaan samanlaisten geenien löytämiseksi vaan tarvitaan klusterointimenetelmiä geenien ryhmittelemiseksi. Näytteitä voidaan ottaa esimerkiksi erilaisista kudoksista, jonkin biokemiallisen prosessin eri vaiheista tai vaikkapa tietynlaisesta syöpäkudoksesta taudin eri vaiheissa. Valitut DNA-molekyylit, itse asiassa kunkin DNA-sekvenssin toinen puolisko, kiinnitetään lasilevylle robotiikan avulla.



Kuva 1. DNA-mikrosirutekniikka [Alb02].

Jos geeni on ollut tietyssä koetilanteessa aktiivinen eli ekspressoitunut, se on tuottanut toisesta DNA-ketjusta komplementin eli RNA:n. Nyt kuhunkin kohtaan lasilevyllä laitetaan DNA:n lisäksi kahta RNA-näytettä. Toinen on värjätty punaiseksi ja on peräisin vertailunäytteestä, esimerkiksi terveestä kudoksesta ja toinen on värjätty vihreäksi ja on peräisin tutkittavasta näytteestä, esimerkiksi syöpäkudoksesta.

Seuraava vaihe, hybridisointi, tarkoittaa RNA:n kiinnittymistä DNA:han. Tämä kiinnittyminen tapahtuu luvun 1.1 alussa mainitun koodauksen mukaisesti. Jos DNA on komplementaarinen siihen lisättyyn RNA:han nähden, RNA kiinnittyy kohdassa olevaan DNA:han. Tämä kiinnittyminen ei kuitenkaan ole täydellistä: joskus RNA kiinnittyy DNA:han, vaikka kyseessä ei olisikaan täydellinen komplementti ja toisinaan taas RNA ei kiinnity, vaikka kyseessä olisi täydellinen komplementti. RNA saattaa myös irrota kiinnittymisen jälkeenkin tai hybridisointia seuraavassa pesuvaiheessa. Tämä aiheuttaa luonnollisesti epätarkkuutta mittaustuloksiin ja epätarkat mittaustulokset puolestaan vaikuttavat klusterointitulokseen.

Jos täplässä on enemmän vihreäksi kuin punaiseksi leimattua RNA:ta, värjäytyy piste vihreäksi ja jos taas punaiseksi värjättyjä on enemmän kuin vihreitä, värjäytyy piste punaiseksi. Jos taas punaisia ja vihreitä on yhtä paljon, värjäytyy piste mustaksi. Tällöin siis vertailutilanteeseen nähden enemmän ekspressoituneita ja vähemmän ekspressoituneita on yhtä paljon, joten keskimäärin tilanne on sama kuin vertailutilanteessa.

Pesun jälkeen signaalit vielä skannataan ja yhdistetään kuvaksi. Mikrosirutekniikan käyttö perustuu kokonaisuudessaan viisivaiheiseen prosessiin [Tam02].

1. Mikrosirun ja koeasetelman suunnittelu. Tässä vaiheessa valitaan tutkittavien geenien joukko G sekä kokeiden joukko K. Alue on selkeästi biologien erikoisosaamista, joten siihen ei tässä työssä tarkemmin syvennytä. Mainittakoon, että matemaattisin menetelmin voidaan kuitenkin arvioida tarvittavien kokeiden määrää.
2. Kuvankäsittely. Tähän vaiheeseen kuuluu kuvassa 1 esitetty vaihe ”signaalien skannaus ja yhdistäminen kuvaksi”.

3. Koetulosten säilytys ja organisointi. Tämä vaihe vaatii huolellisesti suunniteltua tietokantaa, jotta tuhansien koetulosten tiedot saadaan pidettyä järjestyksessä.
4. Ekspressioprofiilien vertailu ja samanlaisten geenien luokittelu.
5. Klusterien biologisen merkityksen ymmärtäminen.

Prosessin kolmas vaihe alkaa siitä, kun mikrosiru, jolle tutkittavat näytteet ovat kiinnittyneet, luetaan fluorosenssimikroskoopilla. Tulokseksi saadaan kuvat sirun fluorosenssista kahdella eri aallonpituudella, jotka kuvaavat testi- ja vertailunäytteiden sekvenssien hyridisoitumista kussakin sivun testipisteessä. Kuvankeräysvaiheen jälkeen geenisirujen fluorosenssitasot mitataan kussakin sirun testipisteessä, vähennetään taustafluorosenssi ja muutetaan kuvan informaatio lukuarvoiksi. Analyysiohjelmat normalisoivat testi- ja vertailunäytteen väliset intensiteettierot ja muodostavat kullekin geenille punaisen ja vihreän fluoresenssin välisen suhteen. Tämä luku kertoo suoraan geenieksprression vilkkaudesta testinäytteessä suhteessa vertailunäytteeseen. [MHK02]

Yhdessä tutkimuksessa analysoidaan usein kymmeniä näytteitä, joiden tuloksia verrataan yhteen vertailunäytteeseen. Näin olen tilastollisen analyysin lähtökohtana on joukko geenieksprression vilkkautta kuvaavia suhdelukuja. Usein on tarpeellista normalisoida suhdeluvut esimerkiksi keskiarvon tai hajonnan suhteen. Mikrosirujen kuvankäsittelyyn, normalisointiin ja tilastolliseen käsittelyyn on kehitetty useita menetelmiä [esimerkiksi CFS97, HZB01]. Myös artikkelissa [MS02] on esitetty yksityiskohtaisesti erään datasetin esiprosessoinnin vaiheet, mutta tämä aihe joudutaan tässä työssä sivuuttamaan.

Tässä työssä tarkastellaan tarkemmin ainoastaan neljättä vaihetta, jossa tarkoituksena on luokitella samanlaisen ekspressioprofiilin omaavat geenit samaan ryhmään kuuluvaksi käyttämällä erilaisia klusterointialgoritmeja. Biologisena oletuksena klusterointimenetelmiä käytettäessä on, että ainakin lähes kaikki geenistä tuotettu RNA käytetään proteiinin valmistuksessa, jolloin pelkkää RNA:ta tarkastelemalla saadaan tietää, mitä proteiineja solussa on kyseisessä koetilanteessa. Lisäksi täytyy olettaa, että geenit, jotka tuottavat samanlaisen ekspressioprofiilin, saattavat osallistua samaan biokemialliseen reaktiopolkuun tai olla toiminnaltaan muutoin samankaltaisia.

2 Geenieksprressioiden klusterointi

Kun DNA-mikrosirutekniikalla on saatu tuotettua geenien ekspressioprofiileista kuva (prosessin vaiheet 1-3), voidaan siirtyä vaiheeseen 4 eli geenien vertailuun ja luokitteluun. Jos luokittelussa käytetään klusterointia, on seuraavaksi valittava käytettävä metriikka ja klusterointimenetelmä. Näihin perehdytään tässä luvussa.

2.1 Metriikan valinta

Kun geenit halutaan ryhmitellä niiden toiminnallisuuden perusteella klustereiksi, on oleellista ensin määritellä, milloin geeni 1 on toiminnaltaan samanlaisempi kuin geeni 2 verrattuna geeniin 3. Tätä määrittelyä kutsutaan etäisyysmetriikan valinnaksi. Metriikan valinta vaikuttaa oleellisesti klusterointitulokseen. Taustalla on oletuksemme siitä, millainen klusterointi on laadullisesti hyvä. Näin siis klusteroinnin lopputuloksen laatua ei voi arvioida irrallaan metriikasta. Yleisin käytössä oleva metriikka on euklidinen etäisyys, jossa kahden pisteen x ja y välinen etäisyys kolmidimensionaalisessa avaruudessa määritellään seuraavasti:

$$d_{12} = \sqrt{(x_1 - y_1)^2 (x_2 - y_2)^2 (x_3 - y_3)^2} \ . \qquad (1)$$

Tämä voidaan yleistää n-ulotteiseksi:

$$d_{12} = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \tag{2}$$

jossa x_i ja y_i ovat geenien X ja Y mitatut ekspressiot koetilanteessa i. Metriikoita, jotka eivät toteuta kolmioepäyhtälöä, kutsutaan semimetrisiksi etäisyysmitoiksi. Monia semimetrisiä etäisyysmittoja käytetään yleisesti ekspressiodatan analysoinnissa [Qua01] .

Euklidisen etäisyyden ohella toinen varsin usein käytetty metriikka perustuu korrelaatiokertoimen käyttöön. Kun koetilanteita on N kappaletta, geenejä G kappaletta ja G_i kuvaa geenin ekspressiota koetilanteessa i, voidaan geenien X ja Y välinen korrelaatiokerroin ilmaista muodossa

$$S(X,Y) = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{X_i - X_{offset}}{\Phi_X} \right) \left(\frac{Y_i - Y_{offset}}{\Phi_Y} \right) \tag{3}$$

jossa

$$\Phi_G = \sqrt{\sum_{i=1}^{N} \frac{(G_i - G_{offset})^2}{N}} \, . \tag{4}$$

Jos G_{offset} on ekspressioiden keskiarvo, on Φ_G sama kuin G:n keskihajonta ja S(X, Y) on tällöin Pearsonin korrelaatiokerroin. [Eis98]

Metriikan valinta tehdään luonnollisesti biologisten oletusten perusteella. Jos esimerkiksi tiettyjen geenien toiminnan tiedetään olevan hyvin lähellä toisiaan ja ne tiettyä metriikkaa käytettäessä luokitellaan samaan klusteriin, voidaan metriikkaa ainakin näiltä osin pitää hyvänä. Näin siis tutkittavien geenien joukkoon kannattaa ottaa muutama tunnettu geeni, jotka ovat toiminnaltaan hyvin samanlaisia tai mahdollisimman erilaisia ja kokeilla, miten ne klusteroituvat. Jos tunnettujen geenien osalta klusterointi vaikuttaa hyvältä, voidaan olettaa metriikan valinnan olevan hyvä myös tuntemattomien geenien osalta. Jos eri metriikoilla saadaan samansuuntaisia tuloksia, kannattanee metriikaksi valita käytännöllisistä syistä yksinkertaisin. Koska metriikan valinta vaikuttaa siis klusterointituloksiin, täytyy klusterointimenetelmien vertailu tehdä samaa metriikkaa käyttäen.

2.2 Hierarkkinen klusterointi

Eisen *et al.* [Eis98] esittivät ensimmäisenä hierarkkisen klusterointialgoritmin ekspressio-datalle. Se on vielä nykyisinkin eniten käytetty menetelmä ekspressiodatan klusteroinnissa. Menetelmän etuja ovat yksinkertaisuus ja helppo visualisoitavuus. Lisäksi se on toteutettavissa laskennallisesti tehokkaasti. Menetelmän heikkous on, että huonosti tehtyä klustereiden yhdistämistä ei voi enää jälkeenpäin muuttaa. [Eis98, Qua01]

Hierarkkinen klusterointi voidaan jakaa *bottom-up* ja *top-down* –strategioita käyttäviin menetelmiin [Sal02]. *Bottom-up* –klusteroinnin alkutilassa jokainen geeni muodostaa oman klusterinsa. Klusteroinnin eteneminen voidaan esittää algoritmina seuraavasti:

1. Alustus: {Geeni $_1$} ∈ Klusteri $_1$, … , {Geeni $_n$} ∈ Klusteri $_n$, etäisyysmitan d määrittely
2. Lasketaan kaikkien klusterien pareittaiset etäisyydet
3. Yhdistetään lähimmät klusterit
4. Jos d(Klusteri $_i$, Klusteri $_j$) = d(Klusteri $_i$, Klusteri $_k$), valitaan sovitun periaatteen mukaan, kumpi klustereista j ja k yhdistetään klusterin i kanssa
5. Toistetaan askelia 2-4 kunnes kaikki klusterit on liitetty samaan klusteriin

Algoritmin tuloksena syntyy yksi klusteripuu eli dendrogrammi, jossa geenien läheisyys puussa kuvaa geenien samankaltaisuutta. Erityisesti vielä samaan klusteriin kuuluvien geenien oletetaan olevan toiminnaltaan samankaltaisia tai osallistuvan samaan biokemialliseen reaktiopolkuun.

Top-down –strategiassa toimitaan päinvastaisesti: aluksi kaikki geenit kuuluvat samaan klusteriin ja joka askeleella jaetaan hajanaisin klusteri kahtia. Näin jatketaan, kunnes jokainen geeni muodostaa oman klusterinsa. Tuloksena saadaan klusteripuu, mutta nyt varsinaiseksi klusteroinniksi ei tarvitse valita koko puuta, vaan voidaan valita siitä vain osa tiettyyn tasoon asti tai jopa siten, että solmut ovat eri tasoilta. Siis vaikka puu pilkotaankin yhden geenin kokoisiin osiin, voi käyttäjä itse valita, milloin mikin klusteri näyttää hyvältä ja lopettaa klusterin pilkkomisen siihen. Tässä klusterin hyvyys viittaa jälleen klusteroinnin biologiseen mielekkyyteen.

Hierarkkiset klusterointimenetelmät voidaan jakaa kuuteen ryhmään sen perusteella, miten klusterien välinen etäisyys on määritelty [Qua01]. Seuraavassa esitellään lyhyesti nämä eri menetelmät.

- Pienimpään etäisyyteen perustuva klusterointi. Tässä menetelmässä klusterien välinen etäisyys määritellään siten, että se on kahden klusterin lähimpien jäsenten välinen etäisyys. Näin kaksi klusteria yhdistetään, jos kaksi niiden jäsenistä on riittävän lähellä toisiaan. Tämä johtaa puumaisiin muodostelmiin, kun yksittäiset jäsenet liitetään yksi kerrallaan klusteriin.
- Suurimpaan etäisyyteen perustuva klusterointi. Tässä klusterien yhteen liittämisen kriteerinä on klusterien kauimmaisten jäsenten välinen etäisyys. Tällä menetelmällä saadaan tiiviitä klustereita, jotka ovat kooltaan varsin samanlaisia.
- Keskimääräiseen etäisyyteen perustuva klusterointi. Tässä menetelmässä klusterien välinen etäisyys on määritelty siten, että se huomio kaikkien pisteiden väliset etäisyydet. Keskimääräinen etäisyys voidaan määritellä monin eri tavoin, esimerkiksi laskemalla keskiarvo jokaisen klusterin 1 ja klusterin 2 jäsenten välisestä etäisyydestä.
- Painotettu keskimääräinen etäisyys. Tässä huomioidaan edellisen menetelmän tapaan jokaisen pisteen etäisyys jokaisesta pisteestä, mutta nyt otetaan huomioon myös klusterin koko, jolla etäisyydet painotetaan. Tätä menetelmää on syytä käyttää, jos klusterien koon oletetaan vaihtelevan suuresti.
- Ryhmän sisäinen klusterointi. Tässä menetelmässä tarkastellaan yksittäisten pisteiden sijaan klusterien keskimääräistä etäisyyttä. Voidaan esimerkiksi laskea klustereille keskipisteet ja verrata niiden avulla klusterien välisiä etäisyyksiä.
- Wardin menetelmä. Tämä menetelmä perustuu keskimääräisen neliösumman laskentaan. Klusterit liitetään toisiinsa siten, että neliöllinen virhe on mahdollisimman pieni.

2.3 Muita klusterointimenetelmiä ekspressiodatalle

Seuraavassa esitellään lyhyesti joitain muita ekspressiodatan klusteroinnissa käytettyjä menetelmiä. Käsittely on joidenkin menetelmien osalta pintapuolinen, sillä menetelmät on esitelty toisaalla tässä julkaisussa (sekoitemallit [Aut02], itseorganisoituva kartta [Hel02]). Sen sijaan simuloituun jäähdytykseen ja pienimpiin virittäviin puihin perustuvat menetelmät käsitellään tässä yhteydessä hieman tarkemmin. Myös k:n keskipisteen menetelmää voidaan käyttää ekspressiodatan klusterointiin [Qua01], mutta tätä menetelmää ei tässä yhteydessä käsitellä. Tämän kappaleen tarkoituksena on lähinnä valottaa geeniekspressioihin liittyvän tietojenkäsittelytieteellisen tutkimuksen monipuolisuutta ja laajuutta.

2.3.1 Mallipohjainen lähestymistapa ja sekoitemallit

Todennäköisyyksiin perustuva mallipohjainen lähestymistapa on vaihtoehto heuristisille klusterointialgoritmeille. Mallipohjaisessa klusteroinnissa oletetaan, että data on generoitu äärellisestä joukosta todennäköisyysjakaumia. Hyvän klusterointialgoritmin ja oikean klusterien määrän sijaan on tarkoituksena etsiä tilanteeseen mahdollisimman hyvin sopiva malli. On huomattava, että tämän aihepiirin artikkeleissa käytetään klusterin synonyyminä sekoitekomponenttia (mixture component).

Yeung *et al.* [Yeu01] kehittivät äärellisiin malleihin perustuvan menetelmän, jonka avulla voidaan klusteroida ekspressiodataa arvioimalla ensin oikea klusterien määrä. Medvedovic *et al.* [MS02] kehittivät menetelmää siten, että klusterien määrän arvioinnin epävarmuus huomioidaan. He käyttivät ääretöntä bayesilaista sekoitemallia, jossa klusterien määrää ei tarvitse arvioida etukäteen. Oikea klusterien määrä saadaan sen sijaan käyttämällä mallina keskiarvoa kaikista malleista, joissa on käytetty kaikkia mahdollisia klusterien määriä.

Myös McLachlan *et al.* [MBP02] ovat tutkineet sekoitemallien käyttöä ekspressiodatan klusteroinnissa. He tutkivat kahta syöpäkudoksien ekspressioprofiileita sisältävää datasettiä. He löysivät kummallekin aineistolle klusteroinnin, joka ei ollut sama kuin kliinisesti tehty, mutta jolla sen sijaan oli yhteys kudosten biologiseen taustaan. Näin siis on mahdollista, että klusteroinnin tuloksena saadaan pätevä luokittelu, joka ei ole ihmissilmin havaittavissa.

2.3.2 Itseorganisoituva kartta

Itseorganisoituva kartta (SOM) on neuroverkkopohjainen menetelmä, jossa SOM järjestää geenit partitioiksi sen perusteella, miten samankaltaisia niiden ekspressiovektorit ovat käytetyn referenssivektorin kanssa. Käyttäjän on annettava partitioille geometrinen muoto, tyypillisesti kaksidimensionaalinen hila, joka vastaa klusterien määrän valintaa k-keskisessä klusteroinnissa. Aluksi joka partitiolle generoidaan satunnaisesti alustettu referenssivektori. Seuraavaksi valitaan satunnainen geeni ja sijoitetaan se käytetyn metriikan mukaiseen lähimpään partitioon. Sen jälkeen referenssivektoria muokataan siten, että se on enemmän valitun geenin ekspressiovektorin kaltainen. Vastaavasti muokataan myös hilaa lähellä olevia referenssivektoreita. Iteraatiota toistamalla saadaan geenit jaettua partitioihin. Tässä menetelmässä termin klusteri synonyyminä käytetään siis termiä partitio. [Qua01]

Vastaavasti kuten k:n keskipisteen menetelmässä on ongelmana klusterien määrän valinta, on itseorganisoituvia karttoja käytettäessä päätettävä geometria etukäteen. Tässä voidaan käyttää apuna esimerkiksi pääkomponenttianalyysia (PCA) [Qua01] tai kappaleessa 2.3.3 esitettävää simuloitua jäähdytystä. Sekä k:n keskipisteen menetelmän että

itseorganisoituvien karttojen tulokset riippuvat alkugeometrian valinnasta. Ne toimivat hyvin, jos klusterirajat ovat säännöllisiä, mutta vaikeuksiin joudutaan, jos klusterien rajat ovat monimutkaisempia tai jos esimerkiksi saadut klusterit eivät ole erillisä vaan klustereilla on yhteisiä alkioita [XOX02]. Xu *et al.* [XOX02] tarjoavat avuksi pienimpiin virittäviin puihin perustuvan menetelmän, johon perehdytään kappaleessa 2.3.4.

2.3.3 Simuloitu jäähdytys

Simuloidun jäähdytyksen idea on fysiikassa: tarkoituksena on matkia todellisen fysikaalisen systeemin hakeutumista energeettisesti edullisimpaan olomuotoonsa. Metriikkana voidaan euklidista etäisyyttä, joka on määritelty kaavassa 2. Annettuna klusterien määrä K, voidaan ekspressioprofiilien jakauma klusterien yli optimoida minimoimalla lauseke 5. Kustannusfunktioon voitaisiin myös lisätä termi, joka maksimoi klusterien välisen etäisyyden.

$$E(K) = \frac{1}{K}\left[\sum_{i \in Ck}\sum_{j \in Ck} d_{ij}\right] \tag{5}$$

jossa i ∈ Ck tarkoittaa vektoria i joka kuuluu klusteriin numero k. E:n minimoinnissa käytetään simuloitua jäähdytystä. Aluksi vektorit ovat klustereissa satunnaisesti jakautuneena. Jokaisella iteratiivisella askeleella satunnaisesti valittu vektori otetaan klusteristaan ja sijoitetaan toiseen satunnaisesti valittuun klusteriin. Lasketaan uusi arvo E^{uusi} ja verrataan sitä arvoon E^{vanha}. Jos E^{uusi} on suurempi kuin E^{vanha}, uusi tila hyväksytään ja sitä pidetään seuraavan iteraation alkuarvona. Jos sen sijaan E^{vanha} on suurempi kuin E^{uusi}, uusi tila hyväksytään todennäköisyydellä

$$P = \exp\left[-\frac{\left(E^{uusi} - E^{vanha}\right)}{T}\right] \tag{6}$$

jossa parametri T kuvaa lämpötilaa, kun parametrin E voidaan sanoa kuvaavan systeemin energiaa. Algoritmi takaa, että riittävän monen iteraation jälkeen systeemi noudattaa Bolzmannin jakaumaa annetussa lämpötilassa. Simuloitua jäähdytystä käyttämällä löydetään globaali optimi, kunhan jäähdytys on tarpeeksi hidas. Liian nopeassa jäähdytyksessä jäädään lokaaliin minimiin, joka on tyypillistä fysikaalisten kappaleiden käyttäytymisessä. Lukashin *et al.* [LF02] käyttivät jäähdytyksessä kaavaa $T_{n+1} = cT_n$, jossa n on askeleen numero ja 1-c on positiivinen ja lähellä nollaa. He osoittivat, että jäähdytys on riippumaton satunnaisluvun generoinnista kun $1\text{-}c \leq 10^{-6}$. Aineistona heillä oli simuloitua aikasarjadataa.

2.3.4 Pienimpiin virittäviin puihin perustuva menetelmä

Xu *et al.* [XOX02] esittävät menetelmän, jonka avulla monidimensionaalisen datan klusterointiongelma voidaan muuttaa pienimmän virittävän puun osituksen etsimiseksi. Tällöin tosin joudutaan luopumaan joistain datan sisäisistä riippuvuuksista. He osoittavat kuitenkin, että klusterointia koskevasta oleellisesta informaatiosta ei jouduta luopumaan, sillä jokainen klusteri vastaa yhtä alipuuta eikä klusteria kuvaava alipuu ole päällekkäinen minkään muun klusterin kanssa.

Määritellään pienin virittävä puu seuraavasti: olkoon D = $\{d_i\}$ ekspressiodatajoukko ja d_i = (e_i^1,…, e_i^t) geenin i ekspressiotasot aikavälillä [1,t]. Painotetun suuntaamattoman verkon G(D) = (V,E) solmujoukko on V = { d_i | $d_i \in$ D} ja kaarijoukko E = { (d_i, d_j) | d_i ,$d_j \in$ D ja i ≠ j}.Täten verkko E(D) on täydellinen. Jokaisella kaarella (u,v) ∈ E on paino w, joka kuvaa geenien välistä etäisyyttä. Metriikkana voi olla esimerkiksi Euklidinen etäisyys (kaava 2) tai korrelaatiokerroin (kaava 3). Yhtenäisen painotetun verkon G(D) virittävä puu on G(D):n sellainen yhtenäinen syklitön aliverkko, joka sisältää G(D):n jokaisen solmun. Minimaalinen virittävä puu on painoltaan pienin virittävä puu. Painotetun verkon pienin virittävä puu voidaan löytää esimerkiksi ahneella Kruskalin algoritmilla. [XOX02]

Xu *et al.* [XOX02] esittävät kolme pienimpään virittävään puuhun perustuvaa klusterointialgoritmia. Ensimmäinen perustuu pitkien kaarien karsintaan, toinen on luonteeltaan iteratiivisesti etenevä lokaalin minimin nopeasti löytävä algoritmi ja kolmas algoritmi puolestaan etsii globaalin optimin. Kaikki vaativat syötteenä alipuiden määrän, joten tarvittavien klusterien määrä on myös tätä menetelmää käytettäessä pystyttävä aluksi arvioimaan.

3 Vaihtoehtoja ja täydentäviä menetelmiä klusteroinnille

Klusterointimenetelmät ovat ohjaamatonta oppimista, mutta geeniekspressioita voidaan luokitella myös käyttämällä ohjattuja oppimismenetelmiä, kuten tukivektorikoneita (SVM) tai päätöspuita. Monesti pelkkä geenien luokittelu ei kuitenkaan riitä, vaan tarvitaan myös täydentäviä menetelmiä biologisesti oikeellisen kuvan saamiseksi. [Alt01]

Ekspressiodatan avulla voidaan esimerkiksi muodostaa koko genomin kattavia geenin säätelyverkkoja. Yhtenä vaihtoehtona on ollut mallintaa verkko kaksiarvoiseksi siten, että säätelysuhteen olemassaolo merkitään nollalla ja puuttuminen ykkösellä. Myös geenien yhteyksien lineaarista mallinnusta on tehty. Hieman täsmällisempiin malleihin on päästy käyttämällä Bayesiläistä mallinnusta. Ekspressiodataa voidaan käyttää paitsi säätelyverkon rakentamiseen, myös potentiaalisimman metaboliaverkon topologian valintaan. Tällöin selvä yhteys klusterointiin löytyy: jos tietyn polun reaktioita katalysoivien entsyymien geenit ovat ekspressoituneet, polku on ilmeisesti aktiivinen. [Alt01]

Hanish *et al.* [Han02] esittävät, miten metaboliaverkkoihin ja expressiodataan liittyvät etäisyysmitat voidaan yhdistää. Metaboliaverkko on suunnattu verkko, jossa kukin solmu vastaa reaktiota ja reaktioon liittyy sitä katalysoiva entsyymi. Kaaret yhdistävät reaktiot reaktiopoluiksi. Kahden geenin välinen etäisyys verkossa λ_{net} on määritelty niiden koodaamien entsyymien katalysoimien reaktioiden lyhimmäksi etäisyydeksi reaktiopoluilla. Ekspressiodataan perustuvana etäisyysmittana λ_{net} voidaan käyttää vaikkapa Euklidista etäisyyttä. He yhdistävät nämä etäisyysmitat kaavassa 7, jossa g_i ja g_j ovat objekteja o_i ja o_j vastaavat geenit ja v_i ja v_j niitä vastaavat verkon solmut.

$$\Delta(o_i, o_j) = 1 - 0.5 \cdot \lambda_{\exp}(g_i, g_j) + \lambda_{net}(v_i, v_j) \qquad (7)$$

Näin siis yhdistämällä klusteroinnin avulla saatua tietoa geeniryhmien aktiivisuuksista metaboliamittauksiin, voidaan saada täsmällisempi kuva metaboliaverkon topologiasta.

4 Yhteenveto

Geenien ekspressioiden klusteroinnissa on tarkoituksena selvittää geenien toimintaa tutkimalla, mitkä geenit toimivat yhdessä tai ovat keskenään vuorovaikutuksessa. Koska samankaltaiset ekspressioprofiilit omaavat geenit ekspressoituvat samankaltaisissa koetilanteissa, geenien koodaamat entsyymit saattavat sijaita lähekkäin metaboliaverkossa. Geenien toiminnan täsmällisestä tuntemisesta on hyötyä esimerkiksi lääkeaineiden kehittämisessä ja diagnostiikassa. On kuitenkin huomattava, että klusterointi antaa vasta alustavaa tietoa geenin toiminnasta, ja ennen päätelmien tekoa ja tulosten julkaisemista tulisi tulokset varmistaa toisilla menetelmillä tai testata esitettyä hypoteesia biologisessa kokeessa.

Klusteroinnin hyvyyttä arvioitaessa tulee muistaa, että ainoa todellinen kriteeri klusteroinnin hyvyydelle on sen biologinen merkitys. Jos geenit G_1, G_2, G_3 ja G_4 tulevat samaan klusteriin ja ne todellisuudessakin toimivat yhdessä, klusterointi oli näiden geenien osalta onnistunut. Ongelmana vain on, että klusteroinnin avulla pyritään juuri selvittämään, mitkä geenit toimivat yhdessä. Näin siis samaan klusteriin voidaan eri metriikoita ja algoritmeja käyttäen luokitella geenit G_1, G_2, G_3 ja G_4 tai G_1, G_2, G_4 ja G_5 tai vaikkapa geenit G_1, G_3, G_4 ja G_5. Ainoa keino selvittää, mikä klusterointi on oikeellisin, on selvittää kunkin geenin G_1 - G_5 toiminta ja tämän jälkeen valita klusteroinneista pätevin tai todeta, että kaikki edellä mainitut klusteroinnit olivat yhtä virheellisiä. Käytännössä kuitenkin kukin kolmesta klusterointituloksesta johtaa biologisiin jatkotutkimuksiin. Karsintaa voidaan tehdä myös sillä perusteella, että jos esimerkiksi varmasti tiedetään, että geenit G_i ja G_j eivät koskaan toimi yhdessä, voidaan nämä geenit sisältävä klusteri sulkea pois. Lisäongelmia tuo mittausdatan epätarkkuus, jonka vuoksi ei edes voida varmasti sanoa, että geeni olisi ekspressoitunut tietyssä koetilanteessa, vaikka kyseinen kohta olisikin värjäytynyt punaiseksi mikrosirudatassa.

Metriikan valinnassa kannattaa hyödyntää olemassa olevaa tietoa geenien toiminnasta. Klusterointialgoritmin valinnalle sen sijaan ei ole yhtä helppoa antaa biologisia perusteluja, koska klusterointitulosten absoluuttista hyvyyttä voidaan vertailla vasta sitten, kun geenien toiminta tunnetaan täydellisesti. Valintakriteereinä onkin näin ollen pääasiassa algoritmin yksinkertaisuus ja ymmärrettävyys sekä tulosten helppo visualisoitavuus. Algoritmien aika- ja tilavaatimus ei tässä sovelluksessa ole kovinkaan keskeistä, koska datan tuottaminen DNA-mikrosirutekniikalla on ainakin vielä nykymenetelmillä varsin hidasta ja kallista verrattuna tietokoneajoon. Näin siis esimerkiksi laskennallisesti raskaamman menetelmän käytölle ei ole estettä, jos sitä käyttäen vain saadaan relevantimpia tuloksia. Toki tällöin on kiinnitettävä huomiota datan esitykseen. Ekspressiomatriisi voidaan esimerkiksi ilmoittaa bittimatriisina, jolloin geenit voidaan kussakin koetilanteessa jaotella kahteen luokkaan ”ekspressoituu” ja ”ei ekspressoidu”.

Tämän sovellusalueen suurimpana haasteena on mielestäni sen monitieteisyys, sillä geeni-ekspressiodatan klusteroinnissa tarvitaan ainakin kolmen eri osa-alueen asian-tuntijoiden tietämystä. Biologit keskittyvät DNA-mikrosirutekniikan kehittämiseen ja automatisointiin, insinöörit kuvankäsittelypuoleen ja tietojenkäsittelytieteilijät puolestaan soveltavat erilaisia klusterointialgoritmeja, joiden hyvyyden määrittäminen jää kuitenkin viime kädessä biologeille. Tämän työn tarkoituksena oli keskittyä lähinnä tietojenkäsittelijöiden rooliin yhteistyössä, mutta selvittää myös perusteita muista osa-aluista, jotta keskustelu näiden alojen asiantuntijoiden kanssa olisi mahdollista.

Viitteet

[Alb02] B. Alberts, A. Johnson, J. Lewis, M. Raff, K., Roberts, P. Walter. *Molecular biology of the cell*. Garland Science, New York, 2002.

[Alt01] R. Altman. Whole-genome expression analysis: challenges beyond clustering. *Current opinion in Structural Biology* 11 (2001) 340-347.

[Aut02] I. Autio. Sekoitemallit ja EM-algoritmi. Teoksessa *Klusterointimenetelmät-seminaari*, raportti C-2002-54, Helsingin yliopisto, tietojenkäsittelytieteen laitos , marraskuu 2002, 127-138.

[CFS99] T. Chen, V. Filkov, S. Skiena. Identifying gene regulatory networks from experimental data. International conference on research in computational molecular biology (RECOMB), 1999.

[Eis98] M. Eisen, P. Spellman, P. Brown, D Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 95 (1998) 14863-14868.

[Han02] D. Hanisch, A. Zien, R. Zimmer, T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics* 18 (2002) 145-154.

[Hel02] P. Hellemaa. Klusterointi itseorganisoituvalla kartalla. Teoksessa *Klusterointimenetelmät-seminaari*, raportti C-2002-54, Helsingin yliopisto, tietojenkäsittelytieteen laitos , marraskuu 2002, 1-11.

[HZB01] K. Hess, W. Zhang, K. Baggerly, D. Stivers, K Coombes. Microarrays: handling the deluge of data and extracting reliable information. *Trends in Biotechnology,* 19 (2001), 463-368.

[Leh02] K. Lehmussaari. Hierarkkinen klusterointi. Teoksessa *Klusterointimenetelmät-seminaari*, raportti C-2002-54, Helsingin yliopisto, tietojenkäsittelytieteen laitos , marraskuu 2002, 76-85.

[LF01] A. Lukashin, R. Fuchs. Analysis of temporal gene expression profiles: clustered by simulated annealing and determinig the optimal number of clusters. *Bioinformatics* 17 (2001) 405-414.

[MBP02] G. J. McLachlan, R. W. Bean, D. Peel. A mixture model-based approach to the clustering of microarray expression data. *Bioinformatics* 18 (2002) 413-422.

[MS02] M. Medvedovic, S. Sivaganesan. Bayesian infinite mixture model based clustering of gene expression profiles. *Bioinformatics* 18 (2002) 1194-1206.

[MHK02] O. Monni, S. Hautaniemi, O. Kallioniemi. Geenisirutekniikka ja siihen liittyvä bioinformatiikka. *Duodecim*, 118 (2002), 1157-1166.

[Oll02] V. Ollikainen. *Simulation Techniques for Disease Gene Localization in Isolated Populations*. Väitöskirja. Raportti A2002-2, Helsingin yliopisto, tietojenkäsittelytieteen laitos, 2002.

[Qua01] J. Quackenbush. Computational analysis of microarray data. *Nature*, June 2001, 418-427.

[Sal02] M. Salmenkivi. Laskennallisen biologian kurssimateriaali. Helsingin yliopisto, tietojenkäsittelytieteen laitos, 2002.
http://www.cs.helsinki.fi/u/salmenki/labi-s01.fi.html

[Tam02] J. Tamames. Bioinformatics methods for the analysis of expression arrays: data clustering and information extraction. *Journal of biotechnology* 98 (2002) 269-283.

[XOX02] Y. Xu, V. Olman, D. Xu. Clustering gene expression data using a graph.theoretic approach: an application of minimum spanning trees. *Bioinformatics* 18 (2002) 536-545.

[Yeu01] K. Y Yeung, C. Fraley, A. Murua, A. E. Raftery, W. L. Ruzzo. Model based clustering and data transformations for gene expression data. *Bioinformatics* 17 (2001) 977-987.