Geeniekspressiohahmojen haku moniulotteisesta datasta

Aija Niissalo

Helsinki 1.2.2008

Tiedon louhinnan seminaari, Ohjaaja: Hannu Toivonen

HELSINGIN YLIOPISTO
Tietojenkäsittelytieteen laitos

Sisältö

1 Johdanto

Emäs- ja aminohapposekvenssien samankaltaisuutta voidaan hakea tehokkaasti eliöiden genomien DNA-molekyyleistä ja solujen proteiineista. Mm. BLAST-hakutuloksia [ASG90] hyödyntäen on voitu päätellä sekvenssien välisiä toiminnallisia ja evoluutioon liittyviä suhteita. Geenien ekspressio tarkoittaa geenien ilmentymistä eli solun tapahtumasarjaa, jossa DNA:n sisältämä emäs-koodi kopioidaan RNA:ksi ja edelleen tämän RNA-emäsjärjestyksen ohjaamaksi proteiiniksi. (Geeni)ekspressiota mitataan tehokkaasti mm. DNA-mikrosirujen (DNA-array, DNA-chip, microarray) laboratoriokokeilla. Mikrosirumittaus on ns. tehoseulontamenetelmä, joka tuottaa kerralla monta tuhatta ekspressioarvoa reaalilukuna. Hahmonhaku reaaliarvoisesta geenien ekspressiodatasta on luonteeltaan erilaista kuin merkkijonosekvenssien haku. Ekspressiodata on moniulotteista: geenejä on tyypillisesti tuhansia ja olosuhteita jopa satoja erilaisia samassa laboratoriomittaussarjassa. Kuitenkin helppoudessaan BLAST:in kaltaiselle joka-biologin hahmonhakuvälineelle olisi käyttöä valtavan ekspressiodatamäärän analysoinnissa.

Ekspressiodataa on varastoituna julkisissa tietokannoissa nyt jo yli sadalle eliölle, monille tuhansille mittaussarjoille ja sadoille tuhansille mikrosiruille. Määrät kasvavat jatkuvasti. Samanlaisten laboratoriomittausten ja mittaussarjojen etsiminen metatiedon avulla voi viedä kohtuuttomasti aikaa ja käyttäjän pitää tietää, millaista mittaussarjaa hän on hakemassa. Geenien ekspressiotietoja voidaan käyttää mm. solun toimintamekanismien mallintamisessa, sairauksien luokittelussa, sairausennusteissa sekä lääkkeiden vaikutusten arvioinnissa.

Ajankohtainen ja kiinnostava kysymys on, onko pelkästään ekspressiodatan lukuarvojen suuruuteen perustuva yleiskäyttöinen hakukone mahdollinen. Tietokannoissa olevien mittaussarjojen hakutuloksiin saataisiin silloin paremmuusjärjestys. Haku halutaan tehokkaaksi, jolloin hakuun ei annetta kaikkien geenien ekspressioarvoja vaan käyttäjän mielestä tärkeä tai muuten kiinnostava joukko. Olosuhteita ja muuta metatietoa ei hakutilanteessa tarvittaisi. Samansuuntainen arvojakauma, geenien ekspressiohahmo eli profiili, kysely- ja tietokantamittaussarjan välillä voisi paljastaa yllättävää samankaltaisuutta eri koeolosuhteiden synnyttämissä tiloissa. Hakutuloksen biologinen tulkinta ja tilastollinen analyysi voitaisiin tehdä vastauksena saatavien kokeiden tarkkoja tietoja tulkitsemalla.

Tässä kirjoituksessa esittelen geenien ekspressiohahmojen hakualgoritmit, jotka perustuvat reaaliarvoisten mittauspisteiden (1) Euklidiseen etäisyyteen, (2) Pearso-

nin korrelaatioon, (3) Spermanin rank-korrelaatioon tai (4) pienimmän neliösumman sovitukseen singulaariarvohajotelmaa käyttäen (Singular Value Decomposition, SVD). Hakualgoritmien kokeilutietokanta on ladattu opetusministeriön hallinnoiman Tieteen tietotekniikan keskuksen, CSC:n YeastBASE-hiivatietokannan käyttöliittymästä: `http://yeastbase.csc.fi/`. Kokeilutietokantaan otin YeastBASE:n cDNA-sirujen (kokeiluun sopivan) ekspressiodatan.

Luvussa 2 kerrotaan, mitä geenieksрessio on ja miten sitä mitataan. Lisäksi formalisoidaan hakutehtävä ja mainitaan, millaisia hakukoneita on nykyisin käytössä. Luvussa 3 esitellään hakualgoritmeissa käytetyt reaaliarvoisen datan etäisyysmitat sekä moniulotteisen ekspressiodatan analyysiongelmia. Itse hakualgoritmit kuvataan luvussa 4. Loppupäätelmät ja hakutulosten vertailu ovat viimeisessä luvussa 5. Liitteissä annetaan biologisten käsitteiden sanasto ja lyhyt DNA-mikrosirujen kuvaus.

2 Geenieksрessio ja hakutehtävä

Tässä luvussa kuvaan, mitä geenieksрessio on ja miten sitä mitataan sekä mitä laboratoriomittauksessa kullekin geenille saatu ekspressioarvo tarkoittaa. Lisäksi formalisoin hahmonhakutehtävän, joka kohdistuu yhden tai monen mittauksen tuottamaan reaaliarvoiseen datamatriisiin. Tavoitteena on löytää hahmoja, joiden perusteella kyselynä annetun ja tietokannassa olevien mittausten samankaltaisuutta voidaan arvioida. Jos samanlaisia hahmoja löytyy, voidaan tehdä hypoteeseja solujen tiloista. Nämä testataan uusilla laboratoriokokeilla.

2.1 Geenieksрessio

Geenieksрessio on (DNA-molekyylissä olevien) geenien koodaamien proteiinien tuottamista solussa. 1990-luvulta lähtien ekspression suuruutta on voitu mitata kvantitatiivisesti ns. tehoseulontamenetelmillä, kuten DNA-mikrosirujen laboratoriokokeilla [SSD95, LDB96, EiB99]. Mikrosirumittaukset perustuvat Edvin Southernin tekemään keksintöön, jolla solunäytteen sekvenssit löytyvät DNA-fragmentteja ja komplementaarista RNA:ta hybridisoimalla [Sou75]. cDNA-siru mittaa näytesolujen (esim. $10^8 - 10^9$ kpl) hetkellisen tilan poikkeamaa kontrollitilasta ja antaa tuloksena eliön geenien ilmentymisen suhteelliset reaalilukuarvot yhdessä mittauksessa. Geenejä on tuhansia, ja näiden samanaikaisen ilmentymisen mittaustulos voi kertoa jotain oleellista näytteen biologisesta tilasta, voidaan esimerkiksi ennustaa syöpäpo-

tilaan selviytymismahdollisuus. Koska näyte on peräisin suuresta solupopulaatiosta, yhdellä sirulla mitataan itse asiassa monen tilan yhteisvaikutusta. Tämä on tärkeä muistaa, kun hahmonhakua mallinnetaan ja tuloksia verrataan.

Näyteliuoksessa olevat ekspression aikaansaamat molekyylit on leimattu eri fluorisoivalla värillä kuin kontrolliliuoksen. Leima-aineiden lähettämän valon intensiteetti mitataan laserlaitteella. (Tarkempi sirukuvaus löytyy liitteestä 2, DNA-mikrosiru, jos tämä teknologia ei ole tuttu lukijalle. Ainakin liitteen viimeiseen kuvaan kannattaa tutustua.) Olkoon näytteen ekspression aikaansaama (laserlaitteella mitattu ja normalisoitu) valon intensiteetti geenille i ja sirulle j I_{red} ja vastaavan kontrollinäytteen intensiteetti I_{green}. Tällöin yksittäinen ekspressioarvo datassa on:

$$g_{ij} = log_2 \frac{I_{red}}{I_{green}}.$$

Geenien ekspressiolla tarkoitetaan jatkossa näitä arvoja. Arvot esitetään datamatriisina, jossa riveillä yksilöidään geenit ja sarakkeilla sirut.

Ekspressiodata-arvojen muutosta kutsutaan geenien ekspression profiiliksi (gene expression profile). Geenisuuntainen muutos, geeniprofiili (gene profile), kuvaa yhden geenin suhteellisen ekspressiotason vaihtelua monessa eri sirulla tehdyssä mittauksessa. Sirusuuntainen muutos, siruprofiili (array profile), kuvaa yhdellä sirulla mitattua kaikkien geenien suhteellisen ekspressiotason vaihtelua toisiinsa nähden. Näiden geenien ekspressiohahmojen etsintä formalisoidaan seuraassa aliluvussa moniulotteisen ekspressiodatan hakutehtävänä ekspressioarvojen muodostamassa vektoriavaruudessa, joka esitetään matriisiesityksenä. Näin hahmoja voidaan vertailla ja hyödyllisiä vastaavuuksia louhia.

Muutama mikrosirujen hahmonhakukone on toteutettu. Hughesin ja kumppaneiden tutkimus [HMJ00] on ensimmäinen suuri koosteisen ekspressiodatan tutkimus. Koska data oli tuotettu homogeenisissa koeolosuhteissa, kokoava hierarkkinen klusterointi [SSZ98, ESB98] Pearsonin korrelaatioetäisyyden kanssa ryhmitteli yhteen biologisesti samankaltaisia tiloja. Lamb ja kumppanit [LCP06] ovat kehittäneet lääkeaine-geeni-sairaus-yhteyden etsivän hakukoneen, Connectivity Map, ihmisen heterogeeniselle geenien ekspressiodatalle. Justin Lamb luokittelee [Lam07] esimerkiksi eurooppalaisen ArrayExpress ja vastaavan amerikkalaisen Gene Expression Omnibus (GEO) ekspressiotietokannat super-koosteiksi, joilla Connectivity Mapin kaltainen datan koostaminen on ratkaisematta. Helppo käyttäjäkyselyiden haku (ekspressioarvoja käyttäen) ei niissä ole mahdollista. The Connectivity Mapin hakualgoritmi noudattaa parametritonta, järjestykseen perustuva hahmonhakustrategiaa, joka pe-

rustuu muunneltuun Kolmogorov-Smirnov statistiikkaan. Hakualgoritmi on Subramanian ja kumppaneiden [STM05] formalisoima GSEA-algoritmi, Gene Set Enrichment Analysis. CellMontage-sovelluksella [FKT+07] voi hakea kyselyekspressioarvojen samanlaisuutta esimerkiksi osasta GEO-tietokannan dataa. Menetelmä käyttää Spearmanin rank-korrelaatiokerrointa samanlaisuuden mittana. Alusta (platform, esim. tietty mittausinstrumentti ja eliö) pitää kyselyssä valita. Haku tehdään sirukohtaisesti, koko mittaussarjan hahmoa ei haeta (kuten ei Connectivity Mapissäkään).

2.2 Hakutehtävän formalisointi

Ekspressiomittaussarja sisältää yhden tai monta (DNA-siru)mittausta. Mittaussarjan tulos on reaalilukuarvoinen datamatriisi, jossa geenit ovat rivejä ja sarakkeet mittauksia, yhden sirun tuottamia ekspressioarvoja. Tehtävä on tiedonhaku näistä vaihtelevaulotteisista datapistejoukoista (koe- eli mittaussarjoista), joita datamatriisit kuvaavat ja joissa yksi piste vastaa yhtä mittausta.

Kyselyn kohteena oleva tietokanta koostuu datapistejoukoista, jotka sijaitsevat eriulotteisissa reaalilukuavaruuksissa ja joiden koordinaatistojen vastaavuus ei ole triviaalia. Datapistejoukkojen vastaavuusongelman aiheuttaa mittaussarjojen erilainen koko; mittauksia voi olla yhdessä mittaussarjassa jopa yli sata tai vain yksi, jolloin datamatriisi kutistuu yhdeksi pystyvektoriksi. Geeniavaruuden ulottuvuus oletetaan samaksi kaikille mittauksille. Geeniavaruuden moniulotteisuuteen liittyy kuitenkin ongelmia.

Kyselynä annetaan uusi datapistejoukko ja vastauksena halutaan tietokannan datapistejoukot järjestettynä sen mukaan. kuinka samanlaisia ne ovat kyselypistejoukon kanssa. Jotta vastaaminen onnistuu, pitää ratkaista eriulotteisten ja toisiaan vastaamattomien avaruuksien kohdentaminen toistensa suhteen siten, että vertailukelpoisuus toteutuu. Myös yhden datapisteen vastaavuuden etsiminen on mielenkiintoinen tapaus. Seuraavaksi formalisoidaan edellä kuvattu tilanne. Samoja merkintöjä käytetään jatkossa.

Vektorit, matriisit ja joukot merkitään lihavoidulla fontilla; reaaliluvut, kuten lukumäärät, indeksit ja arvot merkitään kursiivilla. Reaalilukujen joukolle on oma merkintä: $\mathbb{R}$. Vektoreita merkitään myös alkioittain: $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)$.

Jokainen ekspressiomittaus antaa ekspressioarvon kullekin mitatulle geenille. Näin saadaan ekspressioarvoja, g_i, missä i kuuluu järjestettyyn joukkoon geenit, $\boldsymbol{G} =$

$\{1, 2, \cdots, m\}$. Voidaan olettaa että joukko $\boldsymbol{G}$ on aina sama ja sen alkioiden järjestys on määrätty. Käytännön nimenvastaavuusongelmat ja puuttuva data ratkaistaan erikseen. Kukin mittaus antaa datan $\boldsymbol{g} = (g_1, g_2, \cdots, g_m)$, joka on mittauksen ekspressioarvoista muodostettu vektori. Mittaussarja sisältää n kappaletta edellä määriteltyjä mittauksia $\boldsymbol{g}_j$, missä j kuuluu joukkoon mittaukset, $\boldsymbol{A} = \{1, 2, \cdots, n\}$. Mittaussarja voi sisältää mm. kontrollimittauksia, yksittäisiä eri olosuhteissa tehtyjä mittauksia tai/ja aikasarjamittauksia (esim. 'control/control', 'treatment/control', 'time1/control', 'time2/control'). Mittaussarja antaa n datavektoria. Ne yhdessä muodostavat mittaussarjan $\boldsymbol{D}$, joka on alussa mainittu datapistejoukko. Mittaukset vaihtelevat eri mittaussarjoissa sekä tulkinnaltaan että lukumäärältään.

Ekspressiodatan tietokanta $\boldsymbol{DB}$, johon haku kohdistuu, on kokoelma tällaisia $m \times n$ -datamatriiseja $\boldsymbol{D}$. Kukin tietokannan datamatriisi $\boldsymbol{D}_k$ kuuluu joukkoon $\mathbb{R}^{m \times n}$, jossa m on geenien lukumäärä ja n on sirujen lukumäärä siten, että m on kaikille sama, mutta n voi vaihdella. k on mittaussarjan indeksi tietokannassa. Datamatriisi $\boldsymbol{D}_k$ koostuu mittausarvoista g_{ij} siten, että i kuluu joukkoon $\boldsymbol{G}$ ja j kuluu joukkoon $\boldsymbol{A}$. Silloin tietokanta $\boldsymbol{DB}$ kuluu joukkoon $\mathbb{R}^{m \times \sum n}$ ja $\boldsymbol{D}_k$ kuluu joukkoon $\boldsymbol{DB} = \{\boldsymbol{D}_1, \boldsymbol{D}_2, \cdots, \boldsymbol{D}_K\}$, missä K on mittaussarjojen lukumäärä tietokannassa. Koko tietokannan sirujen määrää, $\sum n$, merkitään jatkossa p:llä.

Kyselyn lähtökohtana on tilanne, jossa on suoritettu uusi mittaussarja $\boldsymbol{Q}$ (erikoistapauksena yksittäinen mittaus), ja halutaan tietää, onko tietokannassa mittaussarjaa, jossa suhteellinen ekspressio on muuttunut samalla tavoin kuin uudessa mittaussarjassa. Lisäksi ollaan kiinnostuneita vain joistakin geeneistä, ei koko joukosta $\boldsymbol{G}$. Kyselyssä on annettava ne geenit joita kysely koskee (= jokin joukon $\boldsymbol{G}$ osajoukko) ja näille uudessa laboratoriokokeessa saadut mittausarvot. Geenejä valitaan M kpl. Mittausarvoja on kullakin geenillä N kappaletta, jos kokeessa tehtiin N mittausta. Kyselyssä on $M \times N$ mittausarvoa. Näin kysely $\boldsymbol{Q}$ kuluu joukkoon $\mathbb{R}^{M \times N}$, jossa M on kyselygeenien lukumäärä ja N on kyselysirujen lukumäärä ja $\boldsymbol{G}_M$, joka sisältyy joukkoon $\boldsymbol{G}$, on kyselygeenien joukko. Kysely koostuu mittausarvoista q_{ij} siten, että i kuluu joukkoon $\{1, 2, \cdots, M\}$ ja j kuluu joukkoon $\{1, 2, \cdots, N\}$.

Esimerkki: Alla olevassa taulukossa on kolmen geenin ekspressioarvot kyselyn datamatriisille $\boldsymbol{Q}$, jossa on neljä sirua, tietokantamatriisille $\boldsymbol{D1}$, jossa on myös neljä sirua ja tietokantamatriisille $\boldsymbol{D2}$, jossa on vain kaksi sirua. Ensimmäisellä kyselysirulla s1 on mitattu aineella A käsiteltyjen solujen ekspression suhdetta käsittelemättömien solujen ekspressioon. Toisella s2, kolmannella s3 ja neljännellä s4 sirulla vastaavasti aineilla B, C ja D käsiteltyjen solujen ekspression suhdetta kontrolliin.

	s1	s2	s3	s4	s5	s6	s7	s8	s9	s10
mittaussarja	Q	Q	Q	Q	D1	D1	D1	D1	D2	D2
geeni1	-1	-2	-1	-2	-4	-3	-4	-3	-4	-4
geeni2	2	6	5	1	3	2	4	1	5	2
geeni3	5	1	2	3	2	4	1	2	2	3

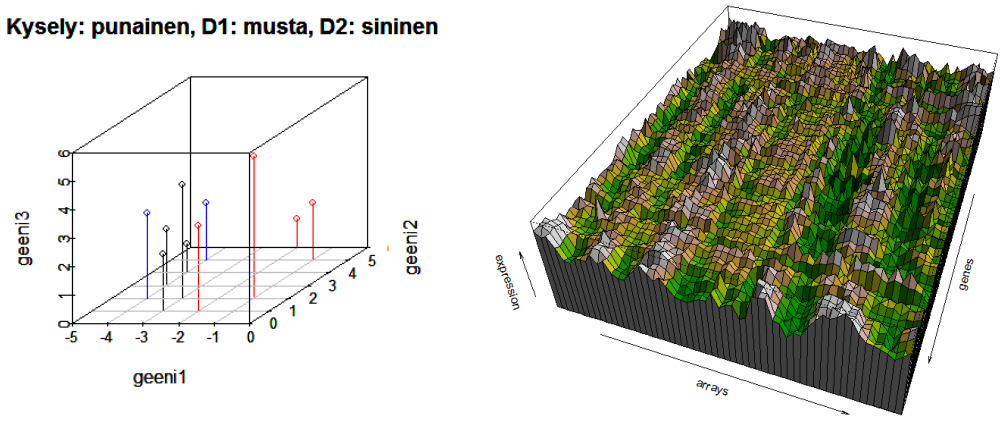
Kyselyn geeniprofiili geenille 1 on (-1, -2, -1, -2), jossa $q_{11} = -1$, $q_{12} = -2$, $q_{13} = -1$ ja $q_{14} = -2$. Siruprofiili sirulle s1 on (-1, 2, 5), jossa q_{11} = -1, $q_{21} = 2$, $q_{31} = 5$. Geeni 1 on alaspäin ekspressoitunut, sen ekspressio on vähäisempää näytteessä kuin kontrollinäytteessä. Muut geenit ovat ylöspäin ekspressoituneita.

Kuvassa[1] 1 on (a) geenien suhteen kolmiulotteisessa avaruudessa kuvattu esimerkin kyselymittaussarja $\boldsymbol{Q}$ (neljä mittausta, punainen) ja kaksi tietokannan mittaussarjaa ($\boldsymbol{D}_1$: neljä mittausta, musta; $\boldsymbol{D}_2$: kaksi mittausta, sininen) sekä (b) geenien ekspressioarvojen säännöllisiä hahmoja havainnollistava esimerkki. Jos kuvan (a) kolmen datapistejoukon kuvan kääntäisi niin, että pisteet esitettäisiin mittausavaruudessa (eri sirut avaruuden suuntina), datapisteitä olisi jokaisella mittaussarjalla kolme, yksi jokaista geeniä kohden. Avaruudet olisivat eriulotteiset (mutta tehollinen avaruus olisi kolmiulotteinen, koska geenejä on vain kolme: kolme vektoria, tässä datapistettä, voi virittää vain kolmiulotteisen avaruuden). Silloin eriulotteisten avaruuksien kohdentaminen toistensa suhteen ei ole triviaalia. SVD-matriisihajotelma (Singular Value Decomposition, SVD) tarjoaa linjaus- ja skaalausoperaatiot (ts. kääntö ja venytys) niin, että rivi- ja sarakeavaruuksien linjaus onnistuu hallitusti. Kun datamatriisilla operoidaan vektoriin, se voidaan SVD:n avulla esittää linjaus-skaalaus-linjaus-operaationa (tarkemmin seuraavassa luvussa). Seuraavassa kuvassa 2 ovat esimerkin kaikki siruprofiilit.

3 Etäisyysmitat ja moniulotteisuus

Alla esitellään etäisyysmitat, joita on käytetty luvun 4 hahmonhakualgoritmeissa. Lisäksi on muutama huomio ekspressidatan ominaisuuksista.

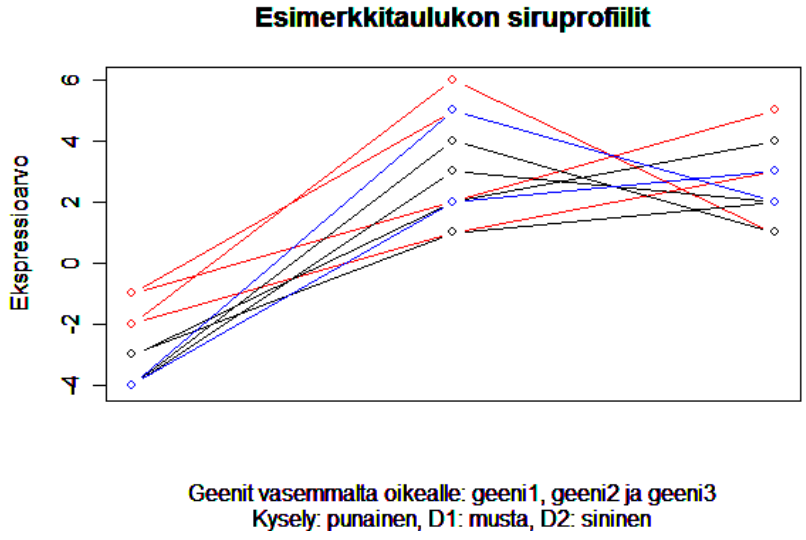
[1] Kuvat on tuotettu R-ohjelmistolla [IG96]



(a) Kysely ja kaksi tietokantadatapistejoukkoa

(b) Osa yhden mittaussarja datasta

Kuva 1: (a) Vastaus kysymykseen, kumpi datapistejoukko vastaa kyselyä, ei ole triviaalia näinkään yksinkertaisessa esimerkissä. (b) Spellmanin ja kumppaneiden [SSZ98] solukiertotutkimuksissa käytetty ekspressiodata osoitti solukiertoon liittyvää säännöllistä geenieksprressiovaihtelua. Kuvassa on 70 geenin ekspressioarvojen suuruudet yhtenä, geenien identiteetti toisena ja siruidentiteetti kolmantena suuntana.



Kuva 2: Siruprofiili kuvaa yhdellä sirulla olevien geenien ekspressioarvojen muutosta. Jokaisella sirulla on mitattu kolmen geenin ekspressiotaso. Värit vastaavat kutakin mittaussarjaa. Sinisellä sarjalla on vain kaksi sirua. Kahdella muulla on neljä.

3.1 Euklidinen etäisyys ja korrelaatiomitat

Euklidinen etäisyys on geometrinen etäisyysmitta ja Minkovskin etäisyyden erikoistapaus. Euklidisessa etäisyydessä yksittäisen koordinaatin suuri numeroarvo voi hallita etäisyyden numeroarvoa. Etäisyys voi kasvaa miten suureksi tahansa. Jos mittaukset ovat tarkkoja ja poikkeavia havaintoja (outlier) ei ole, Euklidinen etäisyys mittaa hyvin mittausten vastaavuutta. Sitä käytetään monissa geenieksprression analyyseissä mittausten epätarkkuudesta huolimatta.

Pearsonin korrelaatiokerroin on numeerinen mitta satunnaismuuttujien väliselle lineaariselle riippuvuudelle. Arvo on -1:n (vastakkaissuuntainen korrelaatio) ja 1:n (samansuuntainen korrelaatio) välillä, 0 merkitsee, että muuttujat eivät korreloi. Geometrisesti tulkittuna korrelaatio on verrannollinen vektoreiden välisen kulman kosiniin. Pearsonin korrelaatio on kohinaherkkä. Pearsonin korrelaatiota käytetään paljon geenieksprression klusteroinneissa.

Spearmanin ρ on ei-parametrinen järjestyskorrelaatio, joka on Pearsonin korrelaatio siten, että havaintoarvojen (esimerkiksi ekspressioarvojen) suuruuden järjestysnumeroita käytetään korrelaatiolaskennan arvoina. Spearmanin järjestyskorrelaatiokerroin, ρ, mittaa kahden muuttujan havaintoarvojen suuruusjärjestyksien yhteensopivuutta. Se sopii esimerkiksi tilanteeseen, jossa ekspressiomittaukset on tehty eri teknologialla.
Laskentakaavoissa vektoreina ovat $\boldsymbol{x} = (x_1, x_2, \cdots, x_m)$ ja $\boldsymbol{y} = (y_1, y_2, \cdots, y_m)$.

Kaava 1 *Vektoreiden Euklidinen etäisyys, tason pisteiden etäisyyden yleistys:*

$$d(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Kaava 2 *Vektoreiden Pearsonin korrelaatiokerroin, kun* s_x, s_y *ovat* $\boldsymbol{x}$*:n ja* $\boldsymbol{y}$*:n koordinaattien keskihajonnat (viiva merkitsee keskiarvoa):*

$$cor(\boldsymbol{x}, \boldsymbol{y}) = \frac{\sum_{i=1}^{m} (x_i - \bar{x})(y_i - \bar{y})}{m s_x s_y}$$

Kaava 3 *Vektoreiden Spearmanin* ρ*, kun RANK merkitsee vektorin koordinaattiarvon suuruuden järjestysnumeroa (1, 2 jne.):*

$$RE_i = RANK(x_i) - RANK(y_i)$$

$$\rho(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{6 \sum_{i=1}^{m} RE_i^2}{m(m^2 - 1)}$$

Esimerkkitaulukossa: Pienin Euclidinen etäisyys (tietokantasiruille) on kyselysirun $s1$ ja $s6$ välillä: $d(\boldsymbol{s1}, \boldsymbol{s6}) = \sqrt{(-1-(-3))^2 + (2-2)^2 + (5-4)^2} = \sqrt{5}$. Paras Pearsonin korrelaatio $s1$:lle on myös $s6$:lla: $cor(\boldsymbol{s1}, \boldsymbol{s6}) = \frac{\sum_{i=1}^{3}(s1_i-2)(s6_i-1)}{3*3*3.6} = 0.97$. Paras Spearmanin korrelaatio $s1$:lle tietokantariruista on kolmella siruilla $s6$, $s8$ ja $s10$, kaikilla korrelaatio on yksi ja ekspressiotason suuruusjärjestys on kaikille $(geeni1^{RANK}, geeni2^{RANK}, geeni3^{RANK}) = (1, 2, 3)$.

3.2 Matriisin singulaariarvohajotelma, SVD

SVD:tä on käytetty mm. tiedonhaussa (information retrieval) ja kuvahahmojen etsinnässä sekä erilaisissa signaalintunnistustehtävissä. Myös PageRank [BrP98] ja HITS-algoritmit [Kle99] hakevat ominaisvektoreita (ks. alla). Pääkomponenttianalyysillä (Principal Component Analysis, PCA) visualisoidaan myös ekspressiodataa. PCA on SVD keskitetylle datamatriisille. Alter ja kumppanit [ABB00] ovat käyttäneet menestyksekkäästi SVD:tä solujen eri tilojen identifioimiseen ja luokitelleet geenejä esimerkiksi sen mukaan, kuuluko niiden toiminta oleellisesti hiivan solukierron tiettyihin vaiheisiin. He ovat laajentaneet menetelmän käyttöä yleistetyllä SVD:llä niin, että eri lajienkin geeniekspressiota on voitu verrata.

SVD [PFT86, GoV83, Eld07] esittää matriisin $\boldsymbol{D}$, joka kuuluu joukkoon $\mathbb{R}^{m \times n}$, hajoitettuna kolmeen osaan: $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T$ siten, että matriisin $\boldsymbol{U}$ sarakkeet ovat matriisin $\boldsymbol{D}\boldsymbol{D}^T$ ominaisvektoreita ja matriisin $\boldsymbol{V}$ sarakkeet matriisin $\boldsymbol{D}^T\boldsymbol{D}$ ominaisvektoreita ja $\boldsymbol{\Sigma}$:n diagonaalin, $(\sigma_1, \sigma_2, \cdots, \sigma_n)$, alkiot ovat matriisin $\boldsymbol{D}\boldsymbol{D}^T$ kuten myös matriisin $\boldsymbol{D}^T\boldsymbol{D}$ ominaisarvojen neliöjuuria, eli singulaariarvoja. Singulaariarvot ovat diagonaalilla järjestyksessä suurimmasta pienimpään. Ominaisvektorit (geeniekspression yhteydessä: ominaisgeenit, $\boldsymbol{V}$, ja ominaissirut, $\boldsymbol{U}$) ja singulaariarvot vastaavat kolmikkona toisiaan niin, että matriisin ominaisgeenit kuvautuvat (projisoituvat) mittausmatriisin kuvauksessa skaalatuille ominaissiruille ja päinvastoin: $\boldsymbol{D}\boldsymbol{v}_i = \boldsymbol{\sigma}_i\boldsymbol{u}_i$ ja $\boldsymbol{D}^T\boldsymbol{u}_i = \boldsymbol{\sigma}_i\boldsymbol{v}_i$. Matriisi koostuu ominaisgeeneistään ja -siruistaan: $\boldsymbol{D} = \sum_{i=1}^{n} \sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$, ominaisgeenien ja -sirujen alkiot sekä singulaariarvot antavat kertoimet lineaarikombinaatioon, jolla matriisin voi muodostaa. Singulaariarvot kertovat summassa olevien osamatriisien normin ja ominaisvektorit rivi- ja sarakeavaruuden ortonormaalit kantavektorit, kukin summan jäsen on rank-1 matriisi.

Ominaisvektoreiden alkioiden lukuarvot kertovat, mitkä geenit ja sirut ovat voimakkaimmin vaikuttaneet ko. lineaarikombinaation syntymiseen. Singulaariarvojen suhteet kuvaavat eri ulottuvuussuuntien merkitystä. Singulaariarvot pienenevät nopeasti, jos matriisissa on paljon lineaarista riippuvuutta. Jos kahden mittaussarjan

(muutamat) ensimmäiset ominaissirut ($\sim$ pääkomponentit) ja singulaariarvot vastaavat hyvin toisiaan ja singulaariarvot pienenevät nopeasti, mittaussarjojen voisi olettaa kuvaavan (jossain määrin) solujen samojen tilojen joukkoa.

Kyselymittaus voidaan projisoida tietokannan mittaussarjaan ominaissirujen avulla niin, että kyselysirujen lyhimmät etäisyydet tietokannan mittaussarjaan voidaan määrätä (pienimmän neliösumman sovitus, Least Square, LS), jos ekspressioarvojen oletetaan olevan lineaarisesti riippuvaisia. Projektio tehdään pseudokäänteismatriisin $\boldsymbol{D}^{-1} = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}^{T}$, avulla, missä $\boldsymbol{\Sigma}^{+}$ on singulaaiarvojen käänteisluvuista muodostettu diagonaalimatriisi. Vektori $\hat{\boldsymbol{q}} = \boldsymbol{U}_k^T\boldsymbol{q}_j$ on projektio, joka linjaa kyselysirun tietokantamittaussarjan ominaissiruille. Operaatio $\boldsymbol{V}\boldsymbol{\Sigma}^{+}\hat{\boldsymbol{q}}$ skaalaa ja linjaa em. vektorin tietokantamittaussarjan ominaisgeeneille ja tuloksena saadaan vektori $\boldsymbol{fitX}$. Tämä projisoidaan edelleen tietokantamittaussarjan mittausavaruuteen. Yhden kyselysirun etäisyys mittaussarjasta on edellä kuvatun projektiosovituksen ja ko. sirun etäisyys eli jäännöksen normi, joka merkitään $||\boldsymbol{D}_k\boldsymbol{fitX} - \boldsymbol{q}_j||_2$. Koko kyselymittaussarjan etäisyys määritellään em. normien summaksi: $\sum_{j=1}^{n} ||\boldsymbol{D}_k\boldsymbol{fitX} - \boldsymbol{q}_j||_2$.

Kokeellisen datan laatu voi olla erittäin vaikea havaita etukäteen. Pienimmän neliösumman ongelmat voivat olla sekä ylimäärättyjä (overdetermined, datapisteitä on enemmän kuin parametrejä) että alimäärättyjä (underdetermined, datassa esiintyy monitulkintaisia parametrikombinaatioita). SVD:llä LS-sovitusta laskettaessa tätä ei tarvitse etukäteen tietää [PFT86]. Myös SVD ovat herkkä poikkeaville havannoille, siksi ne on poistettu tietokannasta.

3.3 Moniulotteisuus ja Simpsonin paradoksi

Euklidisen avaruuden tilavuus kasvaa eksponentiaalisesti, kun uusia ulottuvuuksia lisätään[2]. Moniulotteisen avaruuden tilavuus on keskittynyt avaruuden laidoille. Avaruus on enimmäkseen tyhjä, mikä merkitsee sitä että datan sisäinen rakenne on yleensä pienempiulotteinen. Siksi luokittelualgoritmeissa datavektorit voidaan usein projisoida aliavaruuteen ja luokitella siellä. Parhaan aliavaruuden löytäminen riippuu louhintatehtävästä. Hakuavaruuden kokoa rajoitetaan luvun 4 hahmonhakualgoritmeissa rajaamalla haku vain kyselyn sisältämiin geeneihin. Ekspressiodata on kuitenkin melko tasaisesti jakautunutta origon ympärille, joten vaikka geenejä otettaisiin hakuun tuhansia, hakutulokset eivät yleensä vääristy liikaa. Tämä voi olla

[2]Tämä on dynaamisen ohjelmoinnin keksijän Richard Bellmanin nimeämä 'moniulotteisuuden kirous'

hiivan datan ominaisuus, jonka olemassa olo mm. ihmisen datan kohdalla pitää testata erikseen.

Yksi mikrosirudatan analyysituloksia arvioitaessa huomioon otettava asia on riippuvuuksien ja riippumattomuuksien hienovaraisuuksien huomioiminen. Tämä liittyy 'Simpsonin paradoksina' tunnettuun tilastolliseen ilmiöön. Seuraava *esimerkki* [HMS01] tuntuu ristiriitaiselta. Lääke A parantaa vanhat, suhteessa 30 parantunutta 90 lääkkeen käyttäjästä (30/90), ja nuoret, suhteessa kymmenen parantunutta kymmenestä lääkkeen käyttäjästä (10/10), paremmin kuin lääke B, jolle vastaavat luvut ovat 2/10 vanhoille ja 48/90 nuorille. Kuitenkin, kun ryhmät yhdistetään ja katsotaan kuinka tilastollisesti on käynyt, lääke B (, jolle vastaavat luvut ovat 50/100) parantaa yhdistetyssä ryhmässä paremmin kuin lääke A (, jolle vastaavat luvut ovat 40/100). Ryhmien koon erilaisuus ja parantumismittausten suhteellisuus aikaansaa paradoksin. Koska solunäytteessä mitatut RNA-suhteet voivat olla peräisin (esimerkiksi) kahdesta eri solupopulaatiosta ja populaatioiden kokoero ei ole tiedossa, saattaa mittaustulos näyttää siltä, että joku geeni säätelee toista geeniä alentamalla ekspressiota, vaikka itse asiassa molemmissa solupopulaatioissa erikseen geeni säätelee toista geeniä ekspressiota lisäämällä[3]. Myös (yllä olevan esimerkin tapaan) solunäytteiden solupopulaatiot ovat eri kokoisia ja mittaukset ovat suhteellisia. Tämän tyyppiset rajoitukset ekspressiomittauksissa edellyttävät hahmonetsintätulosten varmentamista muulla tavoin ja biologisen asiantuntemuksen käyttöä tuloksien tulkinnassa.

4 Hakualgoritmit

Haku pyrkii löytämään jotakuinkin samansuuntaiset ja normiltaan samanlaiset vektorijoukot. Tätä varten määritellään yleiseen tapaukseen heuristinen laatuluku L, jolla samanlaisuutta arvioidaan. Hakuongelman ratkaisut käyn läpi yksinkertaisemmasta tilanteesta hankalampaan. Yksiulotteisessa tapauksessa käytän samanlaisuuden kriteerinä Euklidista etäisyyttä tai korrelaatiomittoja. Yleisen tapauksen koostan yksiulotteisista tapauksista em. laatuluvun avulla. SVD ratkaisee ulottuvuusongelmat geenien ja sirujen suuntaan samanaikaisesti. SVD-LS haussa lasken LS-sovituksen jäännöksen (residual) normin kullekin kyselysirulle erikseen (jokaista tietokantamittaussarjaa kohden) ja summaan sen jokaiselle tietokantamittaussarjalle

[3]Tommi Jaakkola kurssilla: Intensive course on modeling biological networks, Otaniemi elokuu 2007

erikseen. Pienin summa voittaa.

Haku tehdään vain hakugeenit, $\boldsymbol{G}_M$, sisältävään osatietokantaan, $\boldsymbol{DB}_M \in \mathbb{R}^{M \times p}$, missä p on kaikkien sirujen lukumäärä tietokannassa. Geenisuuntaista profiilia ei etsitä; se tulee implisiittisesti mukaan, jos siruprofiilit ovat riittävän samanlaisia. Kunkin mittaussarjan kyselysiruja lähinnä olevat mittaukset määräävät mittaussarjan vastaavuuden kyselyn kanssa. Kaukana olevat mittaukset eivät huononna tulosta, jos lähellä olevia löytyy.

Selityksissä käytän merkintöjä: (1) $SORT$ merkitsee järjestämistä, pienimmästä suurimpaan. (2) $SORT^{desc}$ merkitsee järjestämistä, suurimmasta pienimpään. (3) $RANK$ ja $RANK^{rev}$merkitsevät järjestysnumeron antamista arvon suuruuden perusteella - rev merkitsee, että suurin arvo saa pienimmän järjestynumeron. (4) $\bar{x}$ on lukujen x_i keskiarvo. (5) K on tietokannassa $\boldsymbol{DB}$ olevien mittaussarjojen $\boldsymbol{D}_k$ lukumäärä ja (6) p on kaikkien tietokantasirujen lukumäärä.

4.1 Yksiulotteinen tapaus

Kyselysyötteenä on yksi vektori, jos kyselyssä ei ole kuin yksi mittaus: $\boldsymbol{Q} = (q_{11}, \cdots, q_{M1})$. Tietokannan mittaussarja on (mallissa) samanlainen kuin kysely, jos jonkin mittaussarjasirun Euklidinen etäisyys kyselysirusta on lähellä nollaa geenivaruudessa mitattuna tai jos kyselyn siruprofiili korreloi hyvin mittaussarjan jonkun siruprofiilin kanssa. Ensimmäisenä on triviaali tapaus. Jos vain yhden geenin ekspressiota on mitattu, vektori on itse asiassa skalaari ja vertailussa käytetään vain arvojen erotusta. Seuraavaksi oletetaan, että geenejä on useampia.

Edellisen luvun etäisyyksiä käyttäen voi mittausten vastaavuuden määrätä kolmella eri kriteerillä. (1) Kuinka kaukana ekspressioarvot ovat toisistaan? (2) Kuinka hyvin mittaukset korreloivat, kun negatiivinen korrelaatio katsotaan huonommaksi kuin ei korrelaatiota ollenkaan? (3) Kuinka hyvin eri geenien ekspression keskinäinen vaihtelu mittaustilanteessa korreloi, kun vain ekspressiotason suuruuden järjestys geenien kesken merkitsee? Näistä saamme kolme yksinkertaista algoritmia, Algoritmi 1, Algoritmi 2 ja Algoritmi 3. Jokainen tietokannan mittaussarja, $k \in \{1, 2, \cdots, K\}$, saa järjestysnumeron perustuen siihen, kuinka etäällä lyhimmällä etäisyydellä kyselystä oleva siru on sijoittunut tai kuinka hyvin mittaussarjan parhaiten korreloiva siru on sijoittunut.

Kullekin algoritmille tehdään sama **alustus**: Erota tietokannasta $\boldsymbol{DB}$ geenejä $\{1, \cdots, M\}$ vastaavat rivit. Osatietokanta, johon kysely kohdistuu, on nyt $\boldsymbol{DB}_M^{M \times p}$,

missä p on tietokannan sirujen lukumäärä. Yksittäinen mittausarvo osatietokannassa on g_{ij}, missä $i \in \{1, 2, \cdots, M\} ja\ j \in \{1, 2, \cdots, p\}$.

Algoritmi 1 Ehaku$(\boldsymbol{Q} = \boldsymbol{q_1} = (q_{11}, \cdots, q_{M1}), \boldsymbol{DB}_M)$
Hakualgoritmi yksiulotteiseen hahmon hakuun Euklidisen etäisyyden avulla.

1. **Laske** $d_j \leftarrow d(\boldsymbol{q_1}, \boldsymbol{g_j}) = \sqrt{\sum_{i=1}^{M}(q_{i1} - g_{ij})^2}, \forall j \in \{1, 2, \cdots, p\}$
2. **Laske** $d_k \leftarrow min_{\boldsymbol{g}_j \in \boldsymbol{D}_k}(d_j), \forall k \in \{1, 2, \cdots, K\}$
3. **return** $SORT(d_k); k \in \{1, 2, \cdots, K\}$

Algoritmi 2 Phaku$(\boldsymbol{Q} = \boldsymbol{q_1} = (q_{11}, \cdots, q_{M1}), \boldsymbol{DB}_M)$
Hakualgoritmi yksiulotteiseen hahmon hakuun Pearsonin korrelaation avulla

1. **Laske** $cor_j \leftarrow cor(\boldsymbol{q_1}, \boldsymbol{g_j}) = \frac{\sum_{i=1}^{M}(q_{i1} - \bar{q})(g_{ij} - \bar{g_j})}{M s_q s_{g_j}}, \forall j \in \{1, 2, \cdots, p\}$
2. **Laske** $cor_k \leftarrow max_{\boldsymbol{g}_j \in \boldsymbol{D}_k}(cor_j), \forall k \in \{1, 2, \cdots, K\}$
3. **return** $SORT^{desc}(cor_k); k \in \{1, 2, \cdots, K\}$

Algoritmi 3 Shaku$(\boldsymbol{Q} = \boldsymbol{q}_1 = (q_{11}, \cdots, q_{M1}), \boldsymbol{DB}_M)$
Hakualgoritmi yksiulotteiseen hahmon hakuun Spearmanin korrelaation avulla

1. **Laske** $\rho_j \leftarrow \rho(\boldsymbol{q}_1, \boldsymbol{g}_j) = 1 - \frac{6\sum_{i=1}^{M} RE_i^2}{M(M^2-1)}, \forall j \in \{1, 2, \cdots, p\}$
2. **Laske** $\rho_k \leftarrow max_{\boldsymbol{g}_j \in \boldsymbol{D}_k}(\ \rho_j), \forall k \in \{1, 2, \cdots, K\}$
3. **return** $SORT^{desc}(\ \rho_k); k \in \{1, 2, \cdots, K\}$

4.2 Yleinen tapaus: heuristinen ratkaisu

Yleisessä hakutapauksessa mittaussarjojen vastaavuus määritellään heuristisen laatuluvun avulla. Parhaat mittaussarjojen vastaavuudet haetaan ensin joka kyselysirulle erikseen ja tietokannan sirut saavat laatuluvun kyselysirun ja tietokantasirun parittaisen etäisyysmitta-arvon suuruuden järjestysnumerona (pienin etäisyys voittaa). Tämän voin laskea yksiulotteisen tapauksen kolmea eri kriteeriä käyttäen. Jokaisen tietokantakokeen k lopulliseen laatulukuun L_k summataan *kutakin kyselysirua* parhaiten vastaavan sirun laatuluku tässä mittaussarjassa (eli N kpl). Laatulukujen summan arvo määrää mittaussarjan sijoituksen vastauksessa. Kyselysyötteenä

on matriisi, jos kyselyssä on monta mittausta: $\boldsymbol{Q}^{M\times N}$, joka koostuu vektoreista q_{ij} siten, että i kuuluu joukkoon $\{1,\cdots,M\}$ ja j kuuluu joukkoon $\{1,\cdots,N\}$, M on geenien lukumäärä kyselyssä ja N sirujen. Vastaavuudessa ei tietokantamittausten mittaussarjakohtainen lukumäärä vaikuta hakuun, vain kyselymittausta (kutakin mittaussarjaa kohden) lähinnä olevat mittaussarjan mittaukset määräävät samanlaisuuden. Jos kyselysirulle löytyy mittaussarjassa hyvin samanlainen siru, se riittää. Silloin mittaussarjassa on tehty ainakin yksi mittaus, jossa samanlaisuus toteutuu. Vaikka muita mittauksia on lisäksi tehty, siitä ei rankaista. Tietokannan yksittäisellä sirulla voi olla paras vastaavuus moneen kysely-siruun nähden. Näin käy välttämättä, jos tietokannan mittaussarjassa ei ole ollut kuin yksi siru. Algoritmissa 4 käytetään yhtä yksiulotteisen tapauksen hakutavoista **Ehaku**, **Phaku** tai **Shaku** samassa haussa. Tietokannan alustus osatietokannaksi tehdään kuten yksiulotteisessa tapauksessa.

Algoritmi 4 Lhaku$(\boldsymbol{Q}^{M\times N}, \boldsymbol{DB}_M)$
Hahmon haku laatuluvun avulla

1. *Alusta* $L_k \leftarrow 0, \forall k \in \{1,2,\cdots,K\}$
2. **for** i *in* $\{1,2,\cdots,N\}$

$$L_i^{1-ulotteinen} \leftarrow \begin{cases} RANK(\textbf{Ehaku}(\boldsymbol{q}_i, \boldsymbol{DB}_M))tai \\ RANK^{rev}(\textbf{Phaku}(\boldsymbol{q}_i, \boldsymbol{DB}_M))tai \\ RANK^{rev}(\textbf{Shaku}(\boldsymbol{q}_i, \boldsymbol{DB}_M)) \end{cases}$$

$$L_k \leftarrow L_k + L_{ki}^{1-ulotteinen}, \forall k \in \{1,2,\cdots,K\}$$

3. **return** $SORT(L_k)$ *;* $k \in \{1,2,\cdots,K\}$

4.3 Pienimmän neliösumman sovitus SVD:llä

Pienimmän neliösumman sovitus SVD:llä lasketaan luvussa 3 kuvatulla tavalla:

Jokaiselle kyselysirulle lasketaan pienimmän neliösumman jäännöksen (residual) normi osatietokantamittaussarjojen pseudokäänteismatriisin $\boldsymbol{D}^{-1} = \boldsymbol{V}\Sigma^{+}\boldsymbol{U}^{T}$ avulla (algoritmissa oleva tietokantaindeksi k identifioi tietokantamittaussarjan). Jos kyselysiru $\boldsymbol{q}$ olisi (singulaarisen) $\boldsymbol{D}$:n rangilla, projektiolla $\boldsymbol{DfitX} = \boldsymbol{q}$ on ääretön määrä ratkaisuja. Jos taas kyselysiru ei olisi (singulaarisen) $\boldsymbol{D}$:n rangilla, ratkaisuja ei olisi ollenkaan. Kummassakin tapauksessa lähin mahdollinen (pienimmän

neliösumman mielessä) $\boldsymbol{fitX}$ löytyy SVD:n avulla, mikä minimoi jäännöksen normin $||\boldsymbol{D}\boldsymbol{fitX} - \boldsymbol{q}||_2$ (tässä 2-normi) [PFT86, GoV83]. Algoritmikuvassa 5 on tämä esitetty pseudokoodina. Alustus sisältää taas hakugeenien osatietokantaan rajoittumisen, kuten muissakin algoritmeissa. Itse SVD-hajotelma lasketaan esimerkiksi R-ohjelmistolla. Rutiini SVD:n laskuun löytyy useista ohjelmistoista. SVD:n voi laskea $O(mn^2)$ liukulukuoperaatiolla.

Algoritmi 5 *Hahmon haku SVD-LS-menetelmällä*

1. **for** k *in* $\{1, \cdots, K\}$
   $[\boldsymbol{U}_k, \boldsymbol{\Sigma}_k, \boldsymbol{V}_k] \leftarrow SVD(\boldsymbol{D}_k)$
   Alusta $Residual_k \leftarrow 0$
   for j *in*$\{1, 2, \cdots, N\}$
   $\hat{\boldsymbol{q}} \leftarrow \boldsymbol{U}_k^T \boldsymbol{q}_j$ *; kyselysirun projektio ominaissirukannassa*
   $\boldsymbol{fitX} \leftarrow \boldsymbol{V}_k \Sigma^+ \hat{\boldsymbol{q}}$ *; sovitus*
   $Residual_k \leftarrow Residual_k + ||\boldsymbol{D}_k \boldsymbol{fitX} - \boldsymbol{q}_j||_2$ *; normin summaus*
2. **return** $SORT(Residual_k)$*;* $k \in \{1, 2, \cdots, K\}$

Esimerkkitaulukossa: Euklidisen etäisyyden avulla laskettaessa $\boldsymbol{D}_1$ on niukasti lähempänä kyselyn mittaussarjaa verrattuna $\boldsymbol{D}_2$:een. Siruja $s1$ ja $s4$ lähinnä ovat $\boldsymbol{D}_1$:n sirut $s6$ ja $s8$. Sekä siruja $s2$ ja $s3$ lähinnä on $\boldsymbol{D}_2$:n siru $s9$. Vaa'ankielenä on $\boldsymbol{D}_2$:n siru $s10$ - se ei ole kuin vasta kolmanneksi lähinnä siruun $s4$ nähden. Jos se olisi ollut toisena, tulos olisi ollut 'tasapeli', koska $\boldsymbol{D}_1$:n sirut olivat toisena molemmissa 'ei-ykkösenä'-tapauksissa. Pearsonin korrelation avulla laskettaessa $\boldsymbol{D}_1$ voittaa selvästi. Sen siru $s6$ korreloi parhaiten sirujen $s1$ ja $s4$ kanssa, vastaavasti $s7$ korreloi parhaiten sirujen $s2$ ja $s3$ kanssa. $\boldsymbol{D}_2$:n siru $s9$ on toisena sirujen $s1$ ja $s4$ suhteen, siru $s10$ on vasta kolmantena sirujen $s2$ ja $s3$ suhteen. Sija kaksi on sekin $\boldsymbol{D}_1$:llä. Spermanin korrelaatiolla laskettaessa tulos on tasapeli. Näin käy helposti, jos vain kolme geeniä on haussa. SVD-LS menetelmällä $\boldsymbol{D}_1$ voittaa selvästi. Sille kaikkien kyselysirujen jäännöksen normi on käytännössä nolla. $\boldsymbol{D}_2$:lle kyselysirun $s2$ jäännöksen normi on alle yksi, muille yli yksi. Tarkemmin: $||\boldsymbol{D}_2 x - s1||_2 = 3.40, ||\boldsymbol{D}_2 x - s2||_2 = 0.84, ||\boldsymbol{D}_2 x - s3||_2 = 1.97$ ja $||\boldsymbol{D}_2 x - s4||_2 = 1.07$.

5 Hakutulosten vertailu ja loppupäätelmät

Menetelmieni hakutuloksia olen verrannut (yksittäistulosten ja biologisen relevanssiarvion lisäksi) erillisellä ohjelmalla. Ohjelma vertaa hakutulosten mittaussarjojen

tulosparemmuusjärjestyksen alku- ja loppupääleikkausten sisältämien alkioiden lukumäärää. Näin voi arvioida, löytyykö hakutuloksille yhteistä konsensusta paremmuusjärjestykseen. Vertailu koskee hiivatietokannan, jossa on 31 mittaussarjaa, 479 sirua ja 5939 geeniä, hakutuloksia. Pearsonin ja Spearmanin korrelaatiot tuottavat hyvin samanlaisia tulosjärjestyksiä. Seuravaksi eniten samankaltaisuutta on SVD-LS- ja Pearsonin korrelaatio -menetelmillä. Euklidinen haku on tuottaa samanlaisia tulosjärjestyksiä Pearsonin korrelaation kanssa, mutta harvemmin kuin kaksi edellistä paria. Kaikkien eri menetelmien yhteiskonsensus löytyy myös usein. Pienet vaihtelut paremmuusjärjestyksessä tasottuvat leikkausvertailulla, Spearmanin ρ ja Kendallin τ ovat kohinaherkempiä konsensusarvioinnissa kuin em. leikkauksen käyttö.

Euklidisen etäisyyden ja laatuluvun käyttö näyttää löytävän mittaussarjojen samanlaisuutta, jos poikkeavia havaintoja ei tietokannassa ole ja haku sisältää riittävän paljon lukuarvoltaan erilaisesti ekspressoituneita geenejä. Euklidisessa haussa merkittävästi ekspressoituneet geenit yleensä määräävät samanlaisuuden eri mittausten välille.

Pearsonin korrelaatiolla voi hakea lineaarista riippuvuutta kyselyn ja tietokannan mittausten välille. Lukuarvojen suuruudella ei silloin ole väliä, kunhan arvovaihtelu on ollut samansuuntaista. Jos kyselyllä tai tietokannan mittaussarjalla on tasaista numeroarvojen vääristymää tai mittausten kontrollitaso on eri, Pearsonin korrelaatio voi löytää samantapaiset kokeet, joita Euklidisella etäisyydellä tehty haku ei löydä. Pearsonin korrelaatio on herkkä datassa esiintyvälle kohinalle ja lineaarisuuden löytyminen on osittain sattumanvaraista, jos datassa lähes kaikki arvot ovat lähellä nolla-tasoa olevia arvoja. Euklidista etäisyysmittaa käytettäessä pienillä lukuarvoilla ei tule suuria eroja, mutta Pearsonin korrelaatiota käytettäessä normiltaan pienten vektoreiden suunnat voivat vaihdella merkittävästi, vaikka normi itsessään olisi merkityksetön eroavaisuutta etsittäessä. Jos mittaukset olisivat tarkkoja, silloin suunnan vaihtelu on tietenkin merkitsevä eroja etsittäessä.

Spearmanin järjestyskorrelaation ongelma on esimerkiksi mittausten lukuarvojen tarkkuusraja ja geenijoukon koko. Läheisillä arvoilla olevien geenien järjestykseen voi tulla mittaustarkkuusrajan määräämää sattumanvaraisuutta. Liian pieni määrä geenejä ei erottele mittauksia tarpeeksi. Järjestykseen perustuvan korrelaation hyöty tulee sen parametrittomuudesta. Mm. eri alustojen (mittaustekniikoita on erilaisia ja niiden tuottamat ekspressioarvot eivät suoraan ole vertailukelpoisia) mittauksia voidaan silloin vertailla keskenään.

SVD:hen perustuva menetelmä ei etsi suoraan kahden mittauksen välistä lineaarista riippuvuutta, vaan koko mittaussarjaan liittyvää lineaarisuutta, jolloin koko mittaussarjan hahmo erottuu. SVD:n etsimän lineaarisuuden merkitys ei ole selvää solubiologian epälineaarisissa prosesseissa - mikä on menetelmän huono puoli. Tietokannan esianalyysit kuitenkin osoittivat, että mittausarvojen välillä on lähes lineaarista riippuvuutta suuressakin geenijoukkossa (mittausarvojen lineaarisuutta testattaessa). Press ja kumppaneiden mukaan [PFT86] SVD:n käyttö pienimmän neliösumman parametrien arvioinnissa ja itse asiassa koko LS-menetelmä voidaan tulkita tilastolliseksi maksimaalisen todennäköisyyden arvioksi (maximum likelihood estimation) siten, että LS-menetelmällä löydetyt parametrit maksimoivat mitatun datan todennäköisyyden. Luulen, että yksi syy SVD:n tehoon niin monella sovellusalalla perustuu osittain siihen, että SVD:llä LS-laskenta voidaan tehdä mahdollisimman häiriöttömästi. SVD tasoittaa vertailuongelmaa, joka syntyy paljon ja vähän siruja sisältävien mittaussarjojen välille. Laatuluvulla ja Euklidisella etäisyydellä tai korrelaatioilla vertailtaessa monta sirua sisältävästä mittaussarjasta löytyy tilastollisesti helpommin (ehkä liiankin helposti) samanlainen siru kyselysirun kanssa kuin SVD:llä vertailtaessa (tämä koskee esimerkiksi aikasarjadataa). Laatuluvun käyttöä voisi kokeilla niin, että lisäisi sille sopivan painotuksen, joka huonontaisi laatua, jos kyselyllä ja tietokantamittaussarjalla on paljon kokoeroa. Perustavanlainen kysymys on silloin, millaista vastaavuutta halutaan.

[PFT86] on hyvä lähde SVD:n soveltamisessa, jos SVD:n käyttö kiinnostaa lukijaa.

Hakukoneen kehittämisessä voisi seuraavaksi kokeilla ICA:n (Independent Component Analysis) ja NMF:n (Nonnegative Matrix Factorization) käyttöä. ICA:n etu SVD:hen nähden on, että sillä voi analysoida tilastollista riippumattomuutta. NMF löytää tunnetusti paikallisia hahmoja.

Lähteet

ABB00 Alter, O., Brown, P. O. ja Botstein, D., Singular value decomposition for genome-wide expression data processing and modelling. *Proc. Natl. Acad. Sci. USA*, 97,18(2000), sivut 10101–10106.

ASG90 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. ja Lipman, D. J., Basic local alignment search tool. *Journal of Molecular Biology*, 215, sivut 403–410.

BrP98 Brin, S. ja Page, L., The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30, sivut 107–117.

EiB99 Eisen, M. B. ja Brown, P. O., Dna arrays for analysis of gene expression. *Methods Enzymol*, 303, sivut 179–205.

Eld07 Eldén, L., *Matrix Methods in Data Mining and Pattern Recognition.* SIAM, 2007.

ESB98 Eisen, M. B., Spellman, P. T., Brown, P. O. ja Botstein, D., Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95, sivut 14863–14868.

FKT+07 Fujibuchi, W., Kiseleva, L., Taniguchi, T., Harada, H. ja Horton, P., Cellmontage: Similar expression profile search server. *Bioinformatics, on line version.* [`http://bioinformatics.oxfordjournals.org/cgi/content/short/23/22/3103`, on line version].

GoV83 Golub, G. ja Van Loan, C., *Matrix computations.* The Johns Hopkins University Press, Oxford, 1983.

HMJ00 Hughes, T., Marton, M., Jones, A., Roberts, C., Stoughton, R., Armour, C., Bennett, H., Coffey, E., Dai, H., He, Y., Kidd, M., King, A., Meyer, M., Slade, D., Lum, P., Stepaniants, S., Shoemaker, D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M. ja Friend, S., Functional discovery via a compendium of expression profiles. *Cell*, 102,1(2000), sivut 109–126.

HMS01 Hand, D., Mannila, H. ja Smyth, P., *Principles of Data Mining.* The MIT Press, 2001.

IG96 Ihaka, R. ja Gentleman, R., R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5,3(1996), sivut 299–314.

Kle99 Kleinberg, J., Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46, sivut 604–632.

Lam07 Lamb, J., The connectivity map: a new tool for biomedical research. *Genome Res.*, 7,1(2007), sivut 54 – 60.

LCP06 Lamb, J., Crawford, E., Peck, D., Modell, J. W., Blat, I. C., Wrobel, M. J., Lerner, Jim Brunet, J.-P., Subramanian, A., Ross, K. N., Reich, M., Hieronymus, H., Wei, G., Armstrong, S. A., Haggarty, S. J., Clemons, P. A., Wei, R., Carr, S. A., Lander, E. S. ja Golub, T. R., The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *Science*, 313,5795(2006), sivut 1929–1935.

LDB96 Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Norton, H. ja Brown, E. L., Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nat Biotech*, 14,12(1996), sivut 1675–1680.

PFT86 Press, W., Flannery, B., Teukolsky, S. ja Vetterling, W., *Numerical recipies, the art of scientific computing.* Cambridge University Press, Cambridge, 1986.

Sou75 Southern, E., Detection of specific sequences among dna fragments separated by gel electrophoresis. *J Mol Biol.*, 98, sivut 503–517.

SSD95 Schena, M., Shalon, D., Davis, R. W. ja Brown, P. O., Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270,5235(1995), sivut 467–470.

SSZ98 Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. ja Futcher, B., Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9,12(1998), sivut 3273–3297. [Myös `http://yeastbase.csc.fi/`, 3.5.2007].

STM05 Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. ja Mesirov, J. P., Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. USA*, 102,43(2005), sivut 15545–15550. [GSEA `http://www.broad.mit.edu/gsea/`, 27.7.2007].

TuL05 Tuimala, J. ja Laine, M., toimittajat, *DNA Microarray Data Analysis*. Picaset Oy, Helsinki, 2005. [Myös `http://www.csc.fi/english/research/sciences/bioscience/books`].

Liite 1. Sanasto

Sanasto sisältää tämän kirjoituksen biologista käsitesanastoa ja sellaisia asioita, joita saattaa tulla vastaan, kun kirjoituksen lähteisiin tutustuu. Biologiassa on paljon poikkeuksia ja muuttuvaa tietoa, joten sanojen selityksetkin muuttuvat ja riippuvat asiayhteydestä. Yleisimmät käsitteet on selitetty ensin.

- *Eliöt - eukaryootit ja prokaryootit*

  Taksonomisesti eliöt voidaan jakaa kolmeen ryhmään: arkkieliöihin (arkit, *Archaea*), bakteereihin (*Bacteria*) ja aitotumallisiin (*Eukarya*). Solutyypin mukaan jaoteltuna arkit ja bakteerit kuuluvat prokaryootteihin (esitumallisiin), joilla ei ole solukalvon erottamaa tumaa, ja eukaryootteihin (aitotumallisiin), joilla on sekä tuma että muita kalvopäällysteisiä DNA:ta sisältäviä rakenteita (esimerkiksi: eläimillä mitokondriot ja kasveilla viherhiukkaset). Taksonomisena ryhmänä voidaan käyttää myös kliinejä (clade). Kliinin jäsenet jakavat piirteitä, jotka ne ovat perineet yhteiseltä esi-isältä. *Neomura*-kliiniin on ehdotettu kuuluvaksi sekä arkit että eukaryootit. Samoin on ehdotettu, että *Neomura* olisi kehittynyt bakteereista, tarkemmin nykyäänkin maaperässä runsaana esiintyvistä aktinobakteereista (*Actinobacteria*). Hiiva ja ihminen kuuluvat eukaryootteihin. Leivontahiiva (*Saccharomyces cerevisiae*) on molekyyli- ja solubiologiassa laajalti käytetty eukaryoottinen mallieliö (model organism), jolle on olemassa lajikohtaisia tietokantoja, mm. Standfordin yliopiston yhteydessä oleva *Saccharomyces*Genome Database, SGD. Viruksia ei lueta eliöiden taksonomiaan, koska ne ovat soluttomia ja niillä ei ole aineenvaihduntaa. Virukset käyttävät eliöitä isäntäsoluinaan. Muita soluttomia elämänmuotoja ovat esimerkiksi prionit.

- *DNA*

  DNA on solun makromolekyylirakenteinen nukleiinihappo, johon on koodattu eliöiden solujen ja joidenkin virusten geneettinen materiaali. Koodin tietoyksikköinä ovat neljä emästä, joiden järjestys on kaiken elollisen toiminnan ja kehityksen takana. Eliön lisääntyessä geneettinen materiaali kopioituu ja välitetään jälkeläisille. Eukaryooteissa DNA on järjestäytynyt kromosomeiksi solun tumassa. Prokaryooteissa DNA on rengasmaisena molekyylinä solun sisällä ilman tumaa nukleoidiksi kutsutulla alueella. Eläin- ja kasvisolujen mitokondriot sekä kasvisolujen kloroplastit (viherhiukkaset) sisältävät oman rengasmaisen DNA:nsa kuten bakteereissa.

- *RNA*

  RNA on DNA:sta kopioituva ribonukleiinihappo, jota käytetään solun sisällä DNA-emästen järjestyksen tiedon siirtoon, kopiointiin ja jossain tapauksissa geneettisenä materiaalina. RNA:han kopioituu vain pieni osa DNA:ta ja sen rakennekin on hieman toisenlainen.

  RNA-tyyppejä on kolme lähetti-RNA (messenger RNA, mRNA), ribosomi-RNA (ribosomal RNA, rRNA) ja siirtäjä-RNA (transfer RNA, tRNA). Virusten geneettinen materiaali voi olla RNA:ta.

- *Geeni* (Gene)

  Geenillä on useita määritelmiä riippuen tarkkuudesta ja yhteydestä, jossa asiasta puhutaan. Yleensä geeni on sellainen osa DNA:ta, jota koodaamalla (transkriptiossa) saadaan muodostettua yksi solussa toimiva proteiini. Geenejä eliöillä on tuhansia. Esimerkiksi hiivalla n. 6000 ja ihmisellä yli 20 000 (ennen väitettiin, että 40 000), bakteereillakin useampi tuhat. Eukaryooteilla geenin koodaavien osien, eksonien, välissä on proteiinia koodaamattomia jaksoja, introneja. Intronien alussa ja lopussa on jaksoja, joiden avulla ne voidaan poistaa (silmukointi) ennen proteiinisynteesiä.

- *Genomi*

  Genomi on organismin koko perintöaines, joka on koodattu DNA:han.

- *Sekvenssi*

  Sekvenssi on DNA:ssa olevien emästen (A, C, G ja T), RNA:ssa olevien emästen, proteiinien aminohappojen (yl. 20 kirjainkoodia) tai muun biologisen koodausosasten merkkijonojen järjestys.

- *Nukleotidi*

  Nukleotidi on monomeeri, joista nukleiinihapot DNA ja RNA muodostuvat. Nukleotidi koostuu fosfaatti-, sokeri- ja emäsosasta. RNA:n sokeriosana on riboosi ja DNA:n deoksiriboosi. DNA:n emäsosia ovat guaniini, adeniini, sytosiini ja tymiini. RNA:n emäsosat ovat guaniini, adeniini, sytosiini ja urasiili. Ihmisen DNA sisältää kolme miljardia nukleotidia. Viruksellakin niitä voi olla tuhansia.

- *Aminohappo*

Aminohappo on orgaaninen happo- ja emäspään (korboksyyli- ja amino-ryhmä) sisältävä yhdiste, joita yhdistämällä saadaan muodostettua elämään tarvittavat proteiinit. Eliöt käyttävät kahtakymmentä (nykytiedon mukaan kaksi lisää - eräällä mikrobilla) erilaista aminohappoa tuhansien erilaisten proteiinin muodostamiseen. Kolmen emäksen järjestys DNA:ssa vastaa yhtä luonnon aminohappoa.

- *Proteiini* (Valkuaisaine)

  Proteiinit ovat polymeroituneita aminohappoja (polypeptidejä). Yksi proteiini voi sisältää satoja ja jopa tuhansia aminohappo-osia. PDB:ssä (Protein Data Bank) on lähes 40 000 proteiinin rakenne. Proteiineilla on ensinnäkin primäärinen rakenne, joka kertoo aminohappojärjestyksen, josta se muodostuu. Aminohapot liittyvät toisiinsa peptidisidoksin. Sekundaarisella rakenteella viitataan muun muassa proteiinin muotoon vaikuttaviin rikkisiltoihin ja vetysidoksiin. Tertiäärinen rakenne tarkoittaa proteiinin avaruudellista rakennetta kokonaisuudessaan. Kvaternaarinen rakenne tarkoittaa usean aminohappoketjun ryhmittymää. Esim. hemoglobiini on neljän yhteen liittyneen proteiinin muodostama tetrameeri. Myös useimmat entsyymit ovat proteiineja.

- *Transkriptio*

  Solun proteiinisynteesiprosessin eräs osa on transkriptio, jossa DNA:n koodin mukainen emäsjärjestys välittyy aminohappojärjestykseksi syntyvään proteiinimolekyyliin (RNA:n muodostuminen DNA:sta). Transkriptiotekijät (transcription factor) säätelevät monimutkaisella prosessilla miten ja milloin jotakin proteiinia solussa aletaan koodata. Transkriptiotekijät mm. paikallistavat DNA:n promoottorialueita (alue, jonka jälkeen jonkun proteiinin koodaus voi alkaa) sitoutumalla suoraan DNA:n ympärille sopivalla avaruudellisella tavalla - usein lisäksi muiden transkriptiotekijöiden avustamana.

- *Metabolia* (Aineenvaihdunta)

  Metabolia on biologinen prosessi, jossa ravintoaineita muokataan, jotta saataisiin energiaa ja rakennusaineita solun tai eliön käyttöön. Metabolisia prosesseja on anabolisia ja katabolisia. Anabolisissa prosesseissa rakennetaan monimutkaisempia molekyylejä yksinkertaisista rakennusaineista. Katabolisissa prosesseissa monimutkaisemmat yhdisteet hajotetaan pienemmiksi.

- *Metaboliitti* (Aineenvaihduntatuote)

Metaboliitti on aineenvaihdunnassa syntynyt yhdiste, josta usein syntyy edelleen muita yhdisteitä. Metaboliaverkoilla kuvataan solun aineenvaihdunnan tapahtumia suunnattuna verkkona, jossa kukin solmu vastaa reaktiota ja reaktioon liittyy sitä katalysoiva entsyymi. Kaaret yhdistävät reaktiot reaktiopoluiksi.

- *cDNA* (complementary DNA)

  cDNA on komplementaarinen DNA-juoste, joka on syntetisoitu käyttäen mallina transkriptiossa syntynyttä RNA:ta.

- *Geeniekspressio* (geenin ilmeneminen - gene expression)

  Geeniekspressio[4] on DNA:n koodaaman proteiinin tuottamista solussa. DNA:ssa olevat geenit eivät ilmene (samankaan eliön) joka solussa samalla tavalla. Ilmeneminen on monimutkaisesti ohjattua ja riippuu solun laadusta ja ympäristön olosuhteista. Ekspressiossa yhdistyy mRNA-molekyylien, DNA-sekvenssien ohjaama, transkriptio ja mRNA-molekyylien ohjaaman proteiinin tuottaminen translaatiossa. Väljemmin puhuttaessa pelkkää transkriptiota kutsutaan ekspressioksi. Eri geenien yhtäaikaista ekspressiota voidaan mitata DNA-siruilla.

- *DNA-siru* (mikrosiru - siru - DNA microarray - chip)

  Katso liite 2: DNA-mikrosiru.

- *BLAST* (Basic Linear Alignment Search Tool)

  BLAST on eräs käytetyimmistä sekvenssien hakumenetelmistä. Se on heuristinen menetelmä ja sen sisältämä ohjelmistovalikoima mahdollistaa tietokantahakujen tekemisen riippumatta siitä, ovatko hakusekvenssi ja tietokannan sisältämät sekvenssit proteiini- tai nukleotidisekvenssejä. Monipuolisten hakumahdollisuuksien lisäksi BLAST:n etuja ovat käytettyjen algoritmien nopeus, herkkyys sekä ohjelmiston vapaa saatavuus. Useimmille BLAST on tuttu ennen kaikkea NCBI:n ja EBI:n kaltaisten keskusten www-palvelimista. Näiden kaikille avointen palvelujen lisäksi monet laboratoriot ja palvelukeskukset ovat toteuttaneet omia BLAST-palveluja. Altschul ja kumppaneiden Basic Local Alignment Search Tool [ASG90] on tieteen eniten viitattu artikkeli.

[4]Viime vuosina on pystytty mittaamaan yhden solun ekspressiota. Animaatio yhden solun ekspressiosta löytyy esimerkiksi Yeshiva yliopiston Singer Laboratoriosta osoitteesta `http://singerlab.aecom.yu.edu/supplements/natrevmcb_v5p855/movies03.htm`.

- *SNP* (Single Nucleotide Polymorphism, snipit)

  SNP on mutaatio (emäsosan vaihtuminen) yhdessä nukleotidissa, joka periytyy. Sen täytyy esiintyä vähintään yhdessä prosentissa populaatiota, jotta se hyväksytään snipiksi. Snipit edustavat 90% ihmisten genomien välisistä eroista ja niitä ilmenee 100 - 300 emäksen välein sekä introneissa että eksoneissa. Koska tällaiset perimäerot selittävät suurimman osan yksilöiden välisestä eroavuudesta, luonnonvalinta kohdistuu erityisen voimakkaasti niihin. Snippejä käytetään paljon laboratoriokokeiden suunnittelussa, kun haetaan sopivaa kohtaa DNA:ssa, johon koe tulisi kohdistaa. Eri organismien snipeille on olemassa tietokantoja.

- *PCR* (Polymerace Chain Reaction)

  Menetelmä, jolla syntetisoidaan DNA:n jostain segmentistä monta kopiota. Siinä käytetään lämpökäsittelykiertoa, jotta polymeraasireaktio saadaan halutunlaiseksi. Menetelmää käytetään monien geneettisten tutkimusten osana.

- *Elektroforeesi* (Electrophoresis)

  Elektroforeesi on analysointimenetelmä, jolla eri painoiset DNA-sekvenssit voidaan erotella. Tulokset kuvataan ja näin syntyy kuvainformaatiota, jota on tarpeen analysoida ja säilyttää. Esimerkiksi PCR-menetelmällä (ks. edellä) tuotettuja DNA-sekvessejä voidaan analysoida elektroforeesin avulla.

- *Proteomiikka ja proteomi*

  Proteomiikka tutkii proteiinien rakennetta ja ominaisuuksia huipputehokkaita laitteita ja teknologioita hyödyntäen. Miikka-päätettä käytetään samaan tapaan monilla biotieteen aloilla. Proteomi on solun (tai koko eliön solujen) tuottamien proteiinien kokonaisuus samoin kuin genomi on geenien ja muun perimäaineksen kokonaisuus. Transkriptomi on puolestaan solun tai solujoukon mRNA-molekyylien kokonaisuus.

Liite 2. DNA-mikrosiru

DNA-mikrosirussa (tarkemmin cDNA, ks. sanasto) on lasi- tai nailonalusta, johon on kiinnitetty kokoelma cDNA-koettimia (cDNA probe). cDNA-koettimet sisältävät 50-500 emäsparia kukin. Koettimina voidaan käyttää myös olikonukleotideja. Olikonukleotideissa on vain 10-80 nukleotidia.

Mikrosirukokeessa hybridisoidaan sirun koettimia leimattujen näytemolekyylien kanssa. Näytemolekyylit ovat peräisin laboratoriokokeen solunäytteen (usein punainen leima) ja kontrollisolunäytteen (usein vihreä leima) ekspressiokokeesta. Koettimet vastaavat (lähes) kaikkia eliön geenejä ja leimatut molekyylit vastaavat (jossain määrin) koetilanteen transkriptiossa (ks. sanasto) tuotetun mRNA:n määrää. Koettimien ja leimattujen molekyylien nukleotidien emäsosat pyrkivät hybridisoitumaan aina pareittain A T:hen ja C G:hen (jos A, adeniini, on koettimessa niin T, tymiini, on leimatussa molekyylissä - eli esimerkiksi ATACG tarttuisi TATGC:hen).

Koska koettimet ovat spesifisiä, paikallaan siruhilassa ja paikka tiedetään, voidaan solunäytteessä tapahtunutta geeniekspression kvantitatiivista suuruutta arvioida leimattujen molekyylien hybridisaatiosuhteen määrää mittaamalla. Mittaustulokset annetaan usein käsitellyn näytteen cDNA-juosteen ja kontrolli-cDNA-juosteen hybridisaatiomäärien suhteen kaksikantaisena logaritmina. Hybridisaation määrä arvioidaan laser-laitteella mittaamalla leimausaineiden lähettämän valon intensiteetit.[5]. Sirun pinnalla on säännöllinen hila (grid) ja jokaisessa hilapisteessä (spot, näitä on jopa kymmeniätuhansia) on identifioitu koetinmolekyylijoukko siten, että yhtä laatua on aina vain yhdessä hilapisteessä. Hilapisteessä voi olla esimerkiksi 10^7 - 10^8 yhden geenin kloonia ja hilapisteitä voi yhdellä sirulla olla esimerkiksi kymmenentuhatta kappaletta. Hybridisointitapahtumassa näyteliuoksen leimatut molekyylit tarttuvat emäsjärjestystänsä vastaaviin koettimiin. Kun hybridisaatio on stabiloitunut kylliksi, ylimäärä näyteliuoksesta huuhdellaan pois, siru kuivataan ja hilaan tarttuneet leimatut molekyylit kuvataan laserlaitteella.

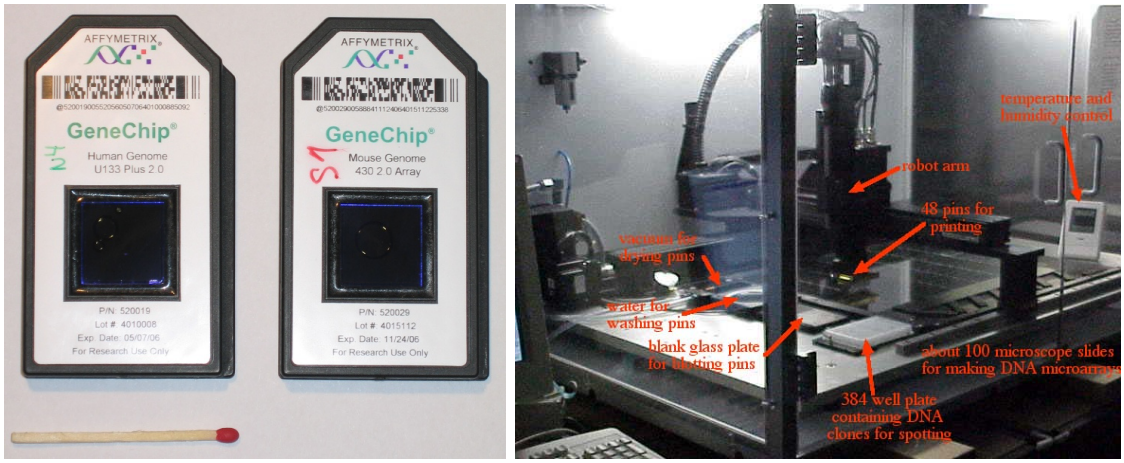
Kuvassa 3 a[6] on kaksi Affimetrix:in sirua ja kuvassa 3 b[7] on robotti, joka valmistaa laboratorioon omatekoisia siruja.

Kuvassa 4, on CSC:n julkaiseman DNA Microarray Data Analysis -kirjasta [TuL05]

[5]Samaan tarkoitukseen käytetään fluoresoivien väriaineiden lisäksi radioaktiivisuuteen perustuvaa leimaamista.

[6]Kuva 1 a on Wikipediasta ja julkaistu 'GNU Free Documentation License' -lisenssillä.

[7]Kuva 1 b on sivulta `http://www.bio.davidson.edu/Courses/genomics/arrays/arrayer.html`; sieltä löytyy myös liikkuvaa kuvaa mikrosirujen valmistuksesta.



(a) Affymetrix GeneChip -siruja (b) Sirujen printtausrobotti

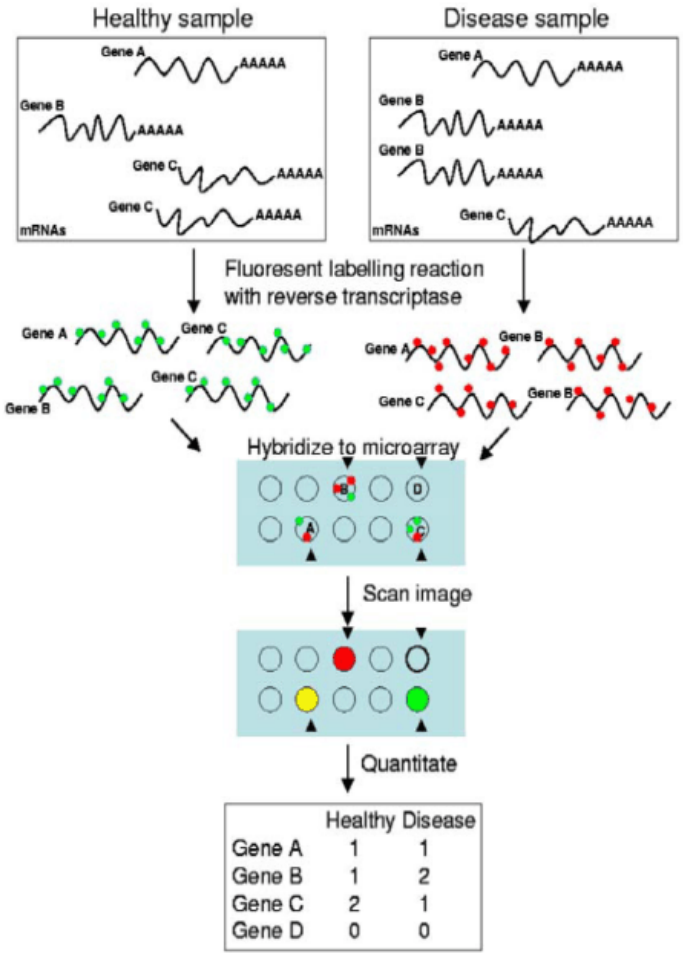
Kuva 3: a-kuvassa on ihmisen ja hiiren ekspressiomittauksiin Affymetrix Inc.:in valmistamat DNA-mikrosirut. Sirulle mahtuu koettimia niin paljon, että eliön lähes kaikkien geenien ilmeneminen voidaan mitata. Affymetrix:in GeneChip-sirujen tuottama ekspressioarvo kuvaa geenin suhteellista ekspressiota saman mittauksen muiden geenien ekspressioon. Siru tuotetaan *in situ* syntetisoinnilla, jolloin koettimet 'rakennetaan' fotolitografisella maskilla suoraan omille hilapaikoilleen. Näin tuotetuilla siruilla geenejä vastaavat paikat ovat samassa sirutyypissä identtiset. Koska sirut ovat tietyn standardin mukaisia, eri sirujen tuottamaa dataa on helpompi vertailla keskenään kuin omatekoisten sirujen. Laboratoriot valmistavat kuitenkin siruja itse, koska silloin koetinvalikoima voidaan suunnitella omaa tarvetta vastaavaksi (ja ne ovat myös halvempia). Koettimien näytevalikoima printataan sirun pinnalle robotiikkaa hyväksi käyttäen. b-kuvassa on mikrosiru robotti The Davidson Trust:in Department of Biology:n Davidson College:sta. Tällaisen sirun (kaksiväri-DNA-siru) tuottama ekspressioarvo on suhteellista geenin eri näytteiden välillä - koenäytteen (esimerkiksi: sairas kudos) ja kontrollinäytteen (esimerkiksi: verrokki kudos).

kuva, jossa on esimerkki mikrosirukokeen työnkulusta.

Sirusta saadaan kuva-analyysin ja normalisoinnin tuloksena monen tuhannen geenin ekspressioarvot yhtä aikaa. Mittaustulokset annetaan (kaksivärisiruilla) usein sirukokeen cDNA-juosteen ja kontrolli-cDNA-juosteen hybridisaatiomäärien suhteen logaritmina.

Logaritmimuunnos tehdään, koska suhteellisessa mittauksessa näytteen kaksinkertaisen ekspression suhde kontrolliin olisi arvoltaan 2 ja kaksi kertaa pienemmän olisi 0.5. Kun arvot muutetaan logaritmisiksi, suhteen arvot $[0, 1]$ päätyvät välille $[-\infty, 0]$ ja suhteen arvot $[1, \infty]$ päätyvät välille $[0, \infty]$. Näin yli- ja aliekspressoituminen (kontrolliin verrattuna) skaalautuu samalla tavalla. Kirjallisuudessa toivotaan usein, että suhteelliseen ekspressiodataan saataisiin yhteinen kontrollistandardi, jolloin arvojen keskinäinen vertailu olisi helpompaa.

Sirut valmistetaan eliökohtaisesti. Siruja voi hankkia teollisesti valmistettuna (esim. Affymetrix, ks. kuva 3) tai niitä voidaan valmistaa laboratoriossa korkean teknologian erityislaitteilla.



Kuva 4: Kuvassa näkyy tyypillisen kaksiväri-DNA-mikrosirun (koenäyte ja kontrollinäyte värjätään eri väreillä) työnkulku. Taulukon arvoista saadaan ekspressiomittauksen arvot (esimerkiksi) ottamalla kaksikantainen logaritmi sairaan ja terveen näytteen intensiteettiarvoista - geeni A: $0 = log_2(1/1)$, geeni B: $1 = log_2(2/1)$, geeni C: $-1 = log_2(1/2)$, geenillä D arvo puuttuu (se voidaan arvioida esimerkiksi KNN-menetelmällä, K-Nearest Neighbors, jos dataa on tarpeeksi). Silloin yli- ja aliekspressoituminen skaalautuu samalla tavalla.