Introduksjon til analyse av cDNA mikromatrisedata

Marit Holden og Anders Løland

Norsk Regnesentral, Gaustadalléen 23, Postboks 114 Blindern, 0314 Oslo
Korrespondanse: Marit Holden, Telefon: 22 85 25 00 Telefax: 22 69 76 60 E-post: marit.holden@nr.no

SAMMENDRAG

cDNA mikromatriser gjør det mulig å studere rollen til tusenvis av gener samtidig, med enorme mengder data som resultat. For å finne ut mer om en biologisk problemstilling er det viktig å velge riktig forsøksdesign. Ulike forsøksdesign gir ulik grad av usikkerhet og kan gi mulighet for svar på ulike spørsmål. Det vanligste forsøksdesignet er referansedesign, der alle celleprøvene sammenlignes mot en felles referanse som er representert på alle mikromatrisene. En cDNA mikromatrise er en glassplate med biologisk materiale som brukes til å undersøke genuttrykket til to ulike celleprøver. Dataene oppnås ved å skanne glassplaten. Deretter filtrerer man bort dårlige målinger og systematiske feil fjernes ved normalisering. Til slutt kommer den statistiske analysen. Målet med denne kan blant annet være å finne grupper av gener ved hjelp av klustring eller å finne ut hvilke gener som er forskjellig uttrykt i celleprøvene ved hjelp av multippel hypotesetesting. Dette kan for eksempel gi svar på hvilke gener eller grupper av gener som er aktive i kreftvev.

Holden M, Løland A. **An introduction to the analysis of cDNA microarray data**. *Nor J Epidemiol* 2003; **13** (2): 291-296.

ENGLISH SUMMARY

cDNA microarrays allow for studying thousands of genes simultaneously, resulting in enormous amounts of data. In order to investigate a biological question, the experimental design is crucial. A different experimental design will result in a different degree of uncertainty and may be suited to answer different questions. The most common design is the reference design, where all cell samples are compared to a common reference represented on all microarrays. A cDNA microarray or chip is a glass slide covered with biological material. This chip can be used to investigate the gene expression of two different cell samples. The data are obtained by scanning the glass slide. Bad measurements are filtered out and systematic variation is removed through normalisation. Finally, the data are analysed statistically. The objective of the analysis may be to find groups of genes using clustering or to find out which genes that are differentially expressed in the cell samples using multiple hypothesis testing. The analysis may for example result in a list of genes or groups of genes that are active in a cancer tissue.

INNLEDNING

Det genetiske materialet i menneskets celler er stort sett kartlagt. Den store utfordringen i årene framover blir å forstå rollen til hvert gen, dvs. hvilken funksjon det har, og hvordan de ulike genene virker sammen. Slik kunnskap kan bl.a. brukes til å utvikle nye og bedre medisiner, til å stille bedre og mer presise diagnoser og til å skreddersy behandling til det enkelte individet.

Det som måles ved hjelp av mikromatriseteknologien er genuttrykk, dvs. hvor aktive de ulike genene i en celle er. Ulike celletyper har ulikt genuttrykk. Det er disse forskjellene i genuttrykk man er mest interessert i å måle. Genuttrykk kan brukes til å skille sykt fra friskt vev, påvise effekt av behandling og gi mer detaljert informasjon om enkelt-gener eller samspillet mellom gener.

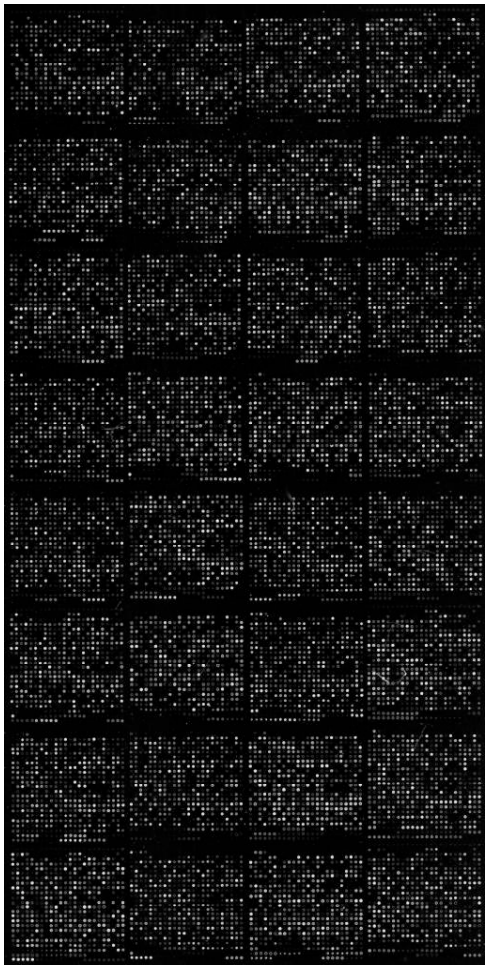
Nye teknikker, for eksempel cDNA mikromatriser, gjør det mulig å studere rollen til tusenvis av gener samtidig. Slike teknikker produserer enorme mengder data. For å få mest mulig ut av disse dataene er det nødvendig å bruke og utvikle avanserte statistiske metoder. Mikromatriseforsøk er svært ressurskrevende. Blant annet derfor er et godt gjennomtenkt forsøksdesign før mikromatriseforsøkene utføres svært viktig.

For å oppnå gode resultater er det nødvendig med nært samarbeid mellom forskere i statistikk og forskere i genetikk, biokjemi, medisin og/eller biologi. For at resultatet av den statistiske analysen skal bli best mulig, bør det legges stor vekt på at man sammen planlegger hvilke mikromatriseforsøk som skal utføres.

HVA ER EN cDNA MIKROMATRISE?

En cDNA mikromatrise er en glassplate. På denne glassplaten er det trykket flekker (spotter) med materiale fra ulike gener (cDNA kloner). Mikromatrisen brukes så til å undersøke genuttrykket til to ulike celleprøver, for eksempel en med tumorceller og en med normale celler. Dette gjøres ved at man først isolerer mRNAet i hver av de to celleprøvene. Mengden av mRNA-molekyler for et bestemt gen brukes som mål på hvor uttrykt dette genet er i den undersøkte celleprøven. For hver celleprøve lages en oppløsning som

inneholder mRNAet fra prøven. De to prøvene farges med to forskjellige fargestoffer, henholdsvis rødt og grønt (Cy5 og Cy3). For eksempel kan tumorprøven farges rød og normalprøven farges grønn. Deretter blandes de to oppløsningene og helles over glassplaten. mRNA fra et gitt gen vil binde seg (hybridisere) til flekker med materiale fra dette genet. En gitt flekk vil da inneholde mye rødt/grønt fargestoff dersom den røde/grønne prøven inneholdt mye mRNA fra det tilsvarende genet. Etter hybridisering vaskes materialet som ikke har bundet seg til noe vekk. Det blir så tatt to bilder av glassplaten med en laserskanner, ett som registrerer mengden av rødt fargestoff, og ett som registrerer mengden av grønt fargestoff. Bildene er vanligvis 16-bits TIFF gråtonebilder. Et eksempel på et slikt bilde av en mikromatrise er vist i Figur 1. Det er disse bildene som sendes videre til bildeanalyse og etterfølgende statistisk analyse. For å se på bildene visuelt slås ofte de to bildene sammen til et fargebilde der flekkene er røde om genet tilsvarende flekken var mye mer uttrykt i den røde prøven enn i den grønne, grønne om det er omvendt og gule om genet var omtrent like mye uttrykt i de to prøvene. I det følgende vil vi beskrive de ulike trinnene i analysen. En oversikt over disse er vist i Figur 2.



Figur 1. Eksempel på en cDNA mikromatrise. Kilde: Mikromatriseprosjektet, http://www.med.uio.no/dnr/microarray/.

BILDEANALYSE

Det første steget i analysen av dataene, dvs. de to TIFF gråtonebildene oppnådd for hver mikromatrise, er å finne intensiteten til hver flekk for rødt og grønt fargestoff. I tillegg må man finne bakgrunnsintensiteten til hver flekk. Denne bakgrunnsintensiteten skyldes blant annet uspesifikk binding, dvs. at mRNA binder seg til mikromatrisen andre steder enn til materialet som ble trykket for det tilsvarende genet. Bakgrunnen varierer over mikromatrisen, dvs. det er ulik bakgrunn i ulike områder på mikromatrisen. Signalet for den røde celleprøven beregnes derfor vanligvis som signalet i flekken minus signalet til den lokale bakgrunnen til flekken. Tilsvarende beregnes signalet for den grønne celleprøven. For et gitt gen antar man at det målte signalet er proporsjonalt med mengden mRNA i prøven man undersøker. For hver flekk beregnes derfor forholdet (ratioen R/G = rødt signal/grønt signal) mellom signalene til de to celleprøvene som undersøkes på mikromatrisen. Om forholdet er vesentlig større enn 1 betyr det at dette genet er mer uttrykt i den røde celleprøven enn i den grønne, om det er vesentlig mindre enn 1 er det omvendt, ellers er genet omtrent likt uttrykt i de to celleprøvene. Ofte ser man på $\log_2$-ratioen, $\log_2(R/G)$, istedenfor ratioen selv fordi verdien 1 da betyr at genuttrykket er dobbelt så høyt i rød prøve som i grønn, –1 at det er halvparten, 2 at det er firedobbelt, –2 at det er en fjerdedel osv. En annen grunn til å se på $\log_2$-ratioen, er at det kan lette modelleringen og kan rettferdiggjøre en implisitt eller direkte antagelse om normalfordeling.

Det finnes både kommersielle og allment tilgjengelige pakker for å trekke ut intensiteter fra cDNA mikromatrisedata. Vanligvis består bildeanalyseprosessen av følgende tre steg:

1. Finn hvor i bildet de ulike flekkene befinner seg. Den regelmessige strukturen til flekkene brukes for å finne plasseringen av disse.
2. Segmenter bildet inn i forgrunn (piksler som hører til selve flekken) og bakgrunn (område i nærheten av flekken) for hver av flekkene.
3. Finn intensitetsverdien for hver flekk og evt. også kvalitetsmål som kan brukes til å finne ut om vi bør inkludere verdien i videre analyse eller ikke. Dette kan være mål som intensitet i bakgrunn i forhold til i forgrunn eller flekkstørrelse eller form.

Etter bildeanalysen kommer filtrering og normalisering.

FILTRERING OG NORMALISERING

Endel flekker bør ikke inkluderes i den videre analysen fordi den beregnete intensiteten i flekken er for usikker. Dette kan for eksempel skyldes støv eller andre forstyrrelser på mikromatrisen. Slike flekker ekskluderes vanligvis manuelt ved visuell inspeksjon av bildet. En annen grunn til at enkelte flekker som oftest tas ut

av analysen er at intensiteten i flekken er for lik den lokale bakgrunnen. Da blir den beregnete intensiteten i flekken svært usikker. Noen velger da å ekskludere flekken om minst en av den røde og grønne målingen er usikker, andre tilordner en lav verdi til slike målinger. Det fins ulike **filtreringskriterier**, dvs. ulike måter å beregne når intensiteten i flekken er for lik bakgrunnen. Man kan f.eks. filtrere bort flekker der intensitet i flekk / intensitet i bakgrunn < 1,4. Grensen 1,4 kan eventuelt byttes ut med en annen grense eller man kan bruke et annet kvalitetsmål enn intensitet i flekk / intensitet i bakgrunn. Grensen for hva som filtreres bort kan være avhengig av kvaliteten på mikromatrisen. (1) foreslår en måte å finne et slikt adaptivt filtreringskriterium på. Dette kriteriet er basert på korrelasjonen mellom repeterte flekker. Repeterte flekker er samme gen trykket flere ganger på samme mikromatrise. Jo høyere korrelasjon mellom repeterte flekker, jo lavere kan grensen for hva som filtreres bort være. At enkelte flekker filtreres bort medfører manglende data. Hva gjør man så med disse manglende dataene i den videre analysen? Man kan

- imputere de manglende verdiene,
- inkludere bare gener uten manglende verdier i datasettet som analyseres eller
- bruke metoder som tillater manglende verdier i datasettet.

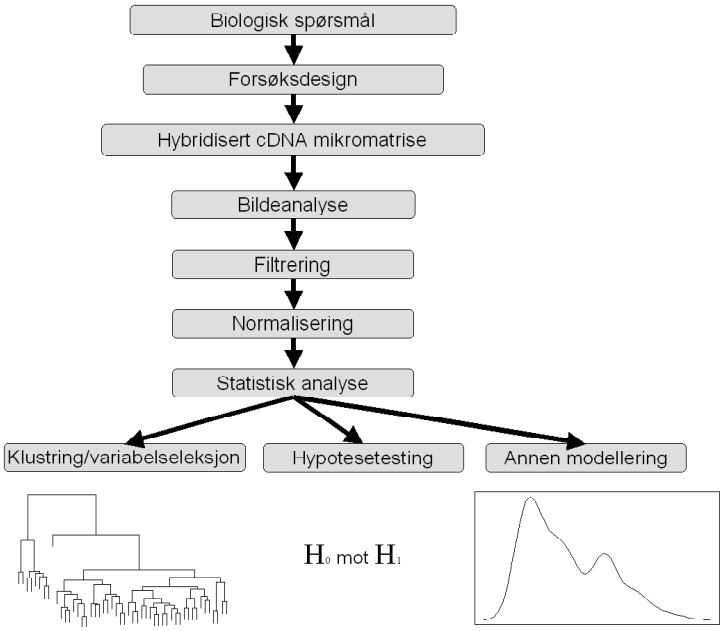
Normalisering av dataene består i å fjerne systematiske feil, og da først og fremst feil som skyldes at de to ulike fargestoffene virker noe ulikt inn på de målte signalene. Normalisering av $\log_2$-ratioene er viktig å gjøre enten før eller som en del av den statistiske analysen. Det mest vanlige er å gjøre det før resten av den statistiske analysen. Samme celleprøve vil få ulike målte signaler avhengig av om det farges med rødt eller grønt fargestoff. Det fins forskjellige, mer eller mindre avanserte, metoder for normalisering av dataene. Alle har som mål å fjerne best mulig effekten av at to ulike fargestoffer er brukt, uten å degradere dataene for mye. En vellykket normalisering vil medføre at likt uttrykte gener får en $\log_2$-ratio rundt 0. Se for eksempel (2) for mer om normaliseringsmetoder.

Det er disse filtrerte og normaliserte $\log_2$-ratiodataene det vanligvis tas utgangspunkt i for de ulike statistiske metodene som brukes for å analysere mikromatrisedataene.

FORSØKSDESIGN

Det første man må gjøre når man skal bruke mikromatrisedata til å finne ut mer om en biologisk problemstilling, er å velge forsøksdesign. Da må man bestemme hvilke celleprøver som skal undersøkes, hvilke celleprøver som skal på hvilke mikromatriser, hvor mange ganger biologisk materiale skal repeteres, hvor mange mikromatriser som skal produseres, hvilke gener skal være med på mikromatrisen, hvor disse genene skal plasseres i forhold til hverandre og hvor mange ganger hvert gen skal repeteres på mikromatrisen. Hva man velger her er avhengig av hvilket biologisk problem som skal studeres, dvs. hva vi ønsker å få svar på og hvilke statistiske metoder vi velger for å analysere dataene. Tid, penger og mangel på biologisk materiale kan være begrensende faktorer. De ulike formene for repetisjon er med ikke bare for å oppnå lavest mulig usikkerhet i de konklusjonene som skal trekkes, men også for at det i det hele tatt skal være mulig å finne ut det man ønsker finne et svar på.

Når det gjelder valg av hvilke celleprøver som skal på hvilke mikromatriser er det mest brukte forsøksdesignet **referansedesign** (se Figur 3 for en visuell



Figur 2. Oversikt over de ulike trinnene i analysen av mikromatrisedataene.

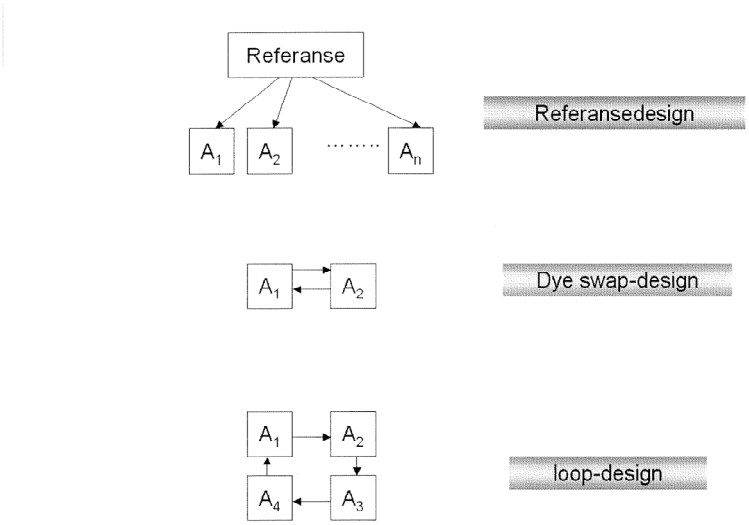
beskrivelse). Her sammenlignes alle celleprøvene som skal undersøkes mot en felles referanse. I alle mikromatriseforsøkene farger man da referansen med en farge, f.eks. rød, og celleprøvene som skal undersøkes med den andre fargen, f.eks. grønn. De ulike celleprøvene kan dermed sammenlignes via referansen. Ulempen med denne typen design er at halvparten av målingene gjøres på referansen som vi i utgangspunktet ikke er interessert i. Dermed får vi lite effektive eksperimenter. Fordelen er at det er lett å inkludere flere celleprøver enn planlagt på forhånd, så lenge man bare har nok av referansematerialet. Et alternativ til referansedesign er **dye-swap-design** der man sammenligner to prøver med hverandre direkte ved å ha dem på samme mikromatriser: en mikromatrise der celleprøve A_1 er rød og celleprøve A_2 er grønn; en annen mikromatrise der celleprøve A_1 er grønn og celleprøve A_2 er rød. Dette er et naturlig design dersom målet med hele eksperimentet er å sammenligne to ulike grupper. Det kan for eksempel være to ulike tumortyper, to ulike behandlinger, før og etter behandling osv. Skal man undersøke mer enn to ulike grupper, A_1, … , A_n, er det mulig å bruke **loop-design**. På den første mikromatrisen farges A_1 grønn og A_2 rød, på den andre A_2 grønn og A_3 rød, … , på den n-te mikromatrisen farges A_n grønn og A_1 rød. På denne måten er det mulig å sammenligne hvilken som helst av A_1, … , A_n mot hvilken som helst annen av A_1, … , A_n. Ulempen er at det kommer inn mer usikkerhet jo lenger fra hverandre i loopen to grupper som sammenlignes er. Dersom det mangler data på en eller flere av mikromatrisene i loopen begrenser det også hva det er mulig å undersøke. Disse problemene unngår man i stor grad ved referansedesign. Det generelle problemet unngår man imidlertid ikke: Det er svært mange gener (titusenvis), men ofte svært få repetisjoner (to-tre), og vi ønsker ofte å si noe om hvert gen. Dette må det tas hensyn til i den videre analysen. Se (3) for mer om forsøksdesign for cDNA mikromatrisedata.

STATISTISK ANALYSE

Dataene fra mikromatriseforsøk er svært støyfylte. Denne støyen er komplisert og kommer fra de mange forskjellige stegene i prosessen fra man har en celleprøve til man har data for genuttrykk. De statistiske metodene bør ta hensyn til denne støyen på en best mulig måte. Støyen kommer både fra selve produksjonen av den ferdig hybridisert mikromatrisen, fra skanningen der de to gråtonebildene oppnås og fra bildeanalysesteget. Målet med analysen kan være å, til tross for støyen, påvise hvilke grupper av gener som er like eller forskjellige.

For hver celleprøve man undersøker ved hjelp av en eller flere mikromatriser kan det lages en genekspresjonsvektor, dvs. en vektor med beregnet $\log_2$-ratio for hvert av de undersøkte genene. En genekspresjonsmatrise lages så fra disse genekspresjonsvektorene. I denne matrisen vil hver rad representere et gen og hver kolonne inneholde en genekspresjonsvektor for en prøve. Slike data er ofte utgangpunktet for ulike klustrings-, variabelseleksjons- og klassifikasjonsmetoder. Andre statistiske metoder tar utgangspunkt i de målte signalene/intensitetene, istedenfor i de beregnete $\log_2$-ratioene, og modellerer disse signalene direkte. Nedenfor skal vi se litt på ulike analysemetoder, både $\log_2$-ratio-basert og intensitetsbaserte.

Klustring er antagelig den mest brukte metoden for analyse av mikromatrisedata. Klustring består i å dele enheter inn i grupper basert på gitte egenskaper for hver enhet slik at gruppene er homogene og godt adskilte. Når det gjelder gener kan dette være å dele inn genene i grupper med gener som er mest mulig likt uttrykt innenfor hver av prøvene. For celleprøver kan det være å dele prøvene inn i grupper med mest mulig like genekspresjonsvektorer, for eksempel slik at normale celler havner i en gruppe, kreftceller i en annen. Andre eksempler er å dele celleprøver inn i grupper for ulike krefttyper, eller å dele kreftceller inn i grupper



Figur 3. Noen typer forsøksdesign for mikromatriseforsøk. Hver boks representerer en prøve som undersøkes på mikromatrisene. Hver pil beskriver en mikromatrise der prøven som står før pilen farges rød og prøven som står etter pilen farges grønn for denne mikromatrisen.

avhengig av om celleprøven ble tatt før eller etter behandling. Noen klustringsmetoder deler enten genene eller prøvene inn i grupper. Andre metoder deler inn i grupper med hensyn på både gener og prøver på en gang.

Den mest brukte klustringsmetoden er hierarkisk klustring (4). Metoden starter med alle egenskapsvektorer i sitt eget kluster. Deretter slås de to klustrene som ligner mest på hverandre sammen. Dette gjentas inntil alle egenskapsvektorene er i samme kluster. Det mest brukte målet på avstand mellom to klustre er gjennomsnittlig avstand mellom punkter i det første klusteret og punkter i det andre klusteret. Ved å se på klustringsresultatet visuelt velges de klustrene som fantes etter et visst antall iterasjoner av algoritmen som de oppnådde klustrene. Metoden kan brukes til å dele enten genene eller prøvene inn i grupper. Det fins mange andre klustringsmetoder blant annet "K-means"-klustring (4), "Self-organizing maps" (SOMs) (5), "Gene shaving" (6), "Plaid models" (7) og metoder basert på singulærverdidekomposisjon (8). Se (9) for en oversikt.

Målet med et mikromatriseforsøk er ofte å finne gener som er ulikt uttrykt i forskjellige grupper av celleprøver. Det vil si at vi ønsker å identifisere gener som er forskjellig uttrykt i de ulike gruppene, men mest mulig likt uttrykt innenfor hver gruppe. Dette kalles **variabelseleksjon**. Disse ulikt uttrykte genene kan vi så senere bruke for eksempel til å finne ut hvilken gruppe nye, ukjente celleprøver hører til. Dette gjøres ved hjelp av metoder for **diskriminantanalyse** og **klassifikasjon** som "Support vector machine" (10), nærmeste nabo-klassifikasjon (11), klassifikasjonstrær (12), voteringsklassifikasjon (13,14) og vektet genvotering (15).

Man kan også finne de ulikt uttrykte genene ved hjelp av hypotesetesting. For hvert gen vil da H_0 være at genet er likt uttrykt i gruppene som undersøkes, H_1 at det ikke er det. Her er det viktig å bruke metoder for **multippel hypotesetesting**. Dersom man ikke tar hensyn til at mange hypoteser testes på en gang, vil man identifisere mange flere gener som forskjellig uttrykt enn det det virkelig er. Metoder som blir brukt i forbindelse med multippel hypotesetesting er blant annet Bonferroni-korreksjon og Benjamini-Hochbergs prosedyre for å kontrollere andelen falske positive (16).

(17) bruker variansanalyse (**ANOVA**) for å finne gener som er ulikt uttrykt i forskjellige typer celleprøver. For å sette opp en ANOVA-modell for de målte røde og grønne intensitetene for hver flekk må man identifisere effekter eller kilder til variasjon som påvirker signalet. Dette kan være ulike individer, behandlinger, tumortyper, mikromatriser, fargestoff, gener osv., og samspill mellom disse effektene. Ofte vil man være interessert i å finne samspillet mellom gen og f.eks. behandling eller tumortype for å finne hvilke gener som er forskjellig uttrykt som resultat av ulik behandling eller tumortype.

Hvis man for eksempel har identifisert mikromatrise, fargestoff, celletype og gen som de viktigste effektene som påvirker signalet, kan man sette opp følgende modell for det målte signalet y_{advg}

$$\log_2 y_{advg} = \mu + A_a + D_d + V_v + G_g + AG_{ag} + DG_{dg} + VG_{vg} + \varepsilon_{advg}$$

der μ er gjennomsnittlig $\log_2$-intensitet; A_a er effekten av mikromatrise a; D_d er effekten av å bruke fargestoff d; G_g er effekten av gen g; V_v er effekten av celletype v; AG_{ag} er samspillet mellom mikromatrise a og gen g, dvs. effekten av en flekk; DG_{dg} er den genspesifikke fargestoffeffekten; VG_{vg} er den genspesifikke effekten av celletype, dvs. den effekten vi er interesserte i; ε_{advg} er modell- og målefeilen. Her har vi antatt at materiale fra hvert gen bare er trykket i én flekk på hver mikromatrise. Ved å sette opp en slik modell og finne estimater for de ulike effektene og samspillene mellom dem, kan man finne hvilke gener som er ulikt uttrykt for for eksempel celletype 1 og 2 ved å se på differensen $VG_{1g} - VG_{2g}$ for hvert gen g. Ved hjelp av bootstrappingsteknikker fant (17) de genene der disse differensene var signifikant forskjellige fra 0, dvs. gener som er forskjellig uttrykt for de to celletypene.

MOT EN MER HELHETLIG MODELL

Metodene beskrevet over består vanligvis av flere trinn. Først bildeanalyse, så normalisering og til slutt annen statistisk analyse for for eksempel identifisering av ulikt uttrykte gener. Estimatene man oppnår etter hvert trinn er usikre. Vanligvis overser man dette og bare plugger inn de oppnådde estimatene i neste trinn uten å ta hensyn til usikkerheten ved estimatene eller hvilken modell som lå til grunn for resultatene oppnådd fra forrige trinn. Ved å bruke **Bayesianske hierarkiske modeller** (18) har man muligheten til å unngå disse problemene ved å få med usikkerheten fra alle trinnene inn i samme modell. Man vil på denne måten kunne oppnå bedre estimater og, kanskje aller viktigst, oppnå realistiske usikkerhetsestimater for disse estimatene. Man kan også få et bedre bilde av hvilke bidrag de ulike trinnene gir. I (19) utvikler vi en modell for statistisk analyse av mikromatrisedata innenfor et Bayesiansk rammeverk der vi utnytter fordelene ved denne typen modeller.

LITTERATUR OG PROGRAMVARE

Det finnes endel bøker som omhandler analyse av cDNA mikromatrisedata. (20) beskriver alle trinnene i Figur 2 og gir også en oversikt over relevant programvare. Boken har en egen hjemmeside http://www.arraybook.org/ med linker til nyttig programvare. (21) gir en lignende oversikt over teknikker, programvare og databaser, mens (22) gir en oversikt over statistiske analyseteknikker.

Et problem en fort støter på ved analyse av mikromatrisedata er at datamengdene blir svært store. For å

avhjelpe analysen og administrasjonen av slike data, kan for eksempel BASE (23) brukes. BASE står for BioArray Software Environment (http://base.thep.lu.se/) og er gratis. Et annet nyttig, gratis dataverktøy er PubGene (24) (http://www.pubgene.org). PubGene er et bioinformatikksystem som gir oversikt over publiserte biologiske relasjoner mellom gener. Systemet utnytter informasjon fra den samlede vitenskapelige litteraturen til å gi en framstilling av hvordan gener fungerer sammen i store, komplekse nettverk. Etter at man har funnet en liste med viktige gener ved statistisk analyse av mikromatrisedata, kan dette verktøyet blant annet brukes til å undersøke nærmere de delene av nettverket der disse genene forekommer.

REFERANSER

1. Jenssen TK, Langaas M, Kuo WP, Smith-Sørensen B, Myklebost O, Hovig E. Analysis of repeatability in spotted cDNA microarrays. *Nucleic Acids Res* 2002; **30** (14): 3235-44.
2. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 2002; **30** (4): e15.
3. Churchill GA. Fundamentals of experimental design for cDNA microarrays. *Nat Genet* 2002; **32** (Suppl): 490-5.
4. Johnson RA, Wichern DW. Applied Multivariate Statistical Analysis. Prentice Hall International, USA, 1988.
5. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, Lander ES, Golub TR. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999; **96** (6): 2907-12.
6. Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol* 2000; **1** (2): research0003.1-0003.21.
7. Lazzeroni L, Owen A. Plaid models for gene expression data. *Statistica Sinica* 2002; **12** (1): 61-86.
8. Golub TR, Van Loan CF. Matrix Computations. John Hopkins University Press, Baltimore, 1983.
9. Aas K. Microarray Data Mining: A Survey. Notat Nr. SAMBA/02/01, Februar 2001, Norsk Regnesentral.
10. Brown MPS, Grundy WN, Lin D, Cristianni N, Sugnet CW, Furey TS, Ares M, Jr, Haussler D. Knowlegde-based analysis of microarray gene expression data using support vector machines. *Proc Natl Acad Sci USA* 2000; **97** (1): 262-7.
11. Duda RO, Hart PE. Pattern classification and Scene analysis. John Wiley & Sons, 1973.
12. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth, Belmont, CA, 1984.
13. Breiman L. Bagging predictors. *Machine Learning* 1996; **24**: 123-40.
14. Freud Y, Shapire RE. Experiments with a new boosting algorithm. In: Proceedings of the 13th International Conference on Machine Learning, 1996: 148-156.
15. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 1999; **286** (5439): 531-7.
16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *JRSS* 1995; 57: 289-300.
17. Kerr MK, Martin M, Churchill GA. Analysis of variance for gene expression microarray data. *J Comput Biol* 2000; **7**: 819-37.
18. Berger JO. Statistical Decision Theory and Bayesian Analysis, 2nd edn. Springer-Verlag, 1985.
19. Frigessi A, Glad I, Holden M, Lyng H, van de Wiel M. Towards a Bayesian analysis of cDNA microarray data. In preparation.
20. Parmigiani G, Garrett ES, Irizarry RA, Zeger S. The Analysis of Gene Expression Data. Springer-Verlag, 2003.
21. Baldi P, Wesley Hatfield G. DNA microarrays and gene expression: from experiments to data analysis and modeling. Cambridge University Press, 2002.
22. Speed T. Statistical analysis of gene expression microarray data. Chapman and Hall, 2003.
23. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg Å, Peterson C. BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data. *Genome Biol* 2002; **3** (8): software0003.1-0003.6.
24. Jenssen T-K, Lægreid A, Komorowski J, Hovig E. A literature network of human genes for high-throughput analysis of gene expression. *Nat Genet* 2001; **28** (1): 21-8.