5. Hansson L, Hedner T, Lund-Johansen P et al. Randomised trial of effects of calcium-antagonists compared with diuretics and blockers on cardiovascular morbidity and mortality in hypertension: the Nordic Diltiazem (NORDIL) study. Lancet 2000;356:359-65.
6. Dahlöf B, Devereux RB, Kjeldsen SE et al. Cardiovascular morbidity and mortality in the Losartan intervention for endpoint reduction in hypertension study (LIFE): a randomised trial against atenolol. Lancet 2002;359:995-1003.
7. Carlberg B, Samuelsson O, Lindholm LH. Atenolol in hypertension: is it a wise choice? Lancet. 2004;364:1684-9. Erratum in: Lancet. 2005;365: 656.
8. Christensen KL, Buus NH. Absence of the normal stress response during betablocker therapy: are the clinical trials interpreted correctly? J Hypertens 2005;23(suppl):2779.
9. Schrader J, Luders S, Kulschewski A et al. Morbidity and mortality after stroke, eprosartan compared with nitrendipine for secondary prevention: principal results of a prospective randomized controlled study (MOSES). Stroke 2005;36:1218-26.
10. Lou M, Blume A, Zhao Y et al. Sustained blockade of brain AT1 receptors before and after focal cerebral ischemia alleviates neurologic deficits and reduces neuronal injury, apoptosis and inflammatory responses in the rat. J Cereb Blood Flow Metab 2004,24:536-47.

Microarray-dataanalyse

Bioinformatiker Rehannah H.A. Borup, læge Claudio Csillag, overlæge Ole Haagen Nielsen & professor Finn Cilius Nielsen

H:S Rigshospitalet Klinisk Biokemisk afdeling, KB 3014, og Amtssygehuset i Herlev, Medicinsk Gastroenterologisk Afdeling C

Mange sygdomsbehandlinger har lidt under mangel på en præcis diagnostik og klassifikation, men flere resultater tyder på at DNA-*microarray*-analyser kan vende denne udvikling. *Microarray*-baseret klassifikation har f.eks. været anvendt ved mange tumorformer, og i næsten alle tilfælde er det vist, at den patologiske diagnose kan optimeres. *Microarrays* peger således frem mod en mere individuel patientdiagnostik. *Microarray*-baserede undersøgelser er forbundet med store datamængder, og dette har nødvendiggjort udviklingen af nye beregningsmetoder og matematiske værktøjer. Denne statusartikel giver en kort oversigt over de vigtigste beregningsprincipper.

Microarray-platforme

Der er overordnet to typer DNA-*microarrays*-*complementary DNA* (cDNA) *arrays*, også kaldet *spottet arrays* [1] og oligonukleotid-*arrays* [2]. cDNA-*arrays* er sammensat af polymerasekædereaktions-opformerede cDNA-sekvenser fra et cDNA-bibliotek, der kobles til et objektglas. Fordelene ved brug af cDNA-*arrays* er lave produktionsomkostninger og fleksibilitet i design. I modsætning til cDNA-*arrays* er højdensitetsoligonukleotid-*arrays* oftest præfabrikerede. Oligonukleotidprober kobles direkte til underlaget ved brug af *ink-jet*-teknologi eller laves de novo ved en fotolitografisk proces [2]. Fordelen ved oligonukleotid-*arrays* er, at det er let at rette de korte probesekvenser mod de mest specifikke dele af mRNA. Ydermere eliminerer in situ-syntese af probesekvenser håndtering af bakteriebiblioteker og opformering af sekvenser og dermed risikoen for krydskontaminering af prober. Den mest udbredte oligonukleotid-*array*-platform fremstilles af firmaet Affymetrix (Santa Clara, CA, USA). Probedensiteten forbedres løbende, og den nuværende generation af humane *microarrays* indeholder omkring 1.300.000 prober, som tilsammen detekterer ca. 48.000 forskellige mRNA. Næste generation af *arrays*, de humane exon-*arrays*, indeholder mere end dobbelt så mange prober og kan anvendes til en fuld transkript- og alternativ splicingsanalyse.

Microarrays anvendes til at måle mængden af mRNA i celler og væv med. Ekspressionsværdien udtrykkes enten som en relativ værdi, der angiver mRNA-forholdet mellem to prøver – en kontrol og en test, som hybridiseres samtidig til *array*'et (*spotted arrays*), eller ved en absolut værdi for mængden af mRNA i en specifik prøve (oligonukleotid *arrays*). Mærkning af prøver til *spotted arrays* sker ved inkorporering af fluoroforekoblede nukleotider i cDNA. Almindeligvis mærkes kontrolmateriale med grøn (cyanin-3) og testmateriale, f.eks. tumorvæv med rød (cyanin-5).

Mærkning af prøver til Affymetrix oligonukleotid-*arrays* starter med, at total RNA revers transkriberes til dobbeltstrenget cDNA, som derefter in vitro-transkriberes til cRNA under indkobling af biotinylerede nukleotider, der kan binde en fluoreofor. Efter hybridisering af prøven til det enkelte *array* aflæses den bundne mængde med en laserskanner. Rådata fra en *microarray*-analyse er derfor en datafil med over en million felter indeholdende intensiteter, der reflekterer genernes ekspressionsniveau.

Microarray-dataanalyse

Microarray-teknologien er almindeligvis meget reproducerbar. Selve proceduren er forbundet med en ganske lille usikkerhed – formentlig under 2%'s variation på de fleste kommercielle platforme. Som ved mange andre analysemetoder spiller den præanalytiske variation en væsentlig rolle, og i forskningssammenhænge er et godt eksperimentelt design en dyd. Det, der primært adskiller dataanalyse af *microarrays* fra andre teknologier, er mængden af data. Mange overrumples af de mange niveauer og metoder for præprocessering og

matematisk normalisering (*low level*-analyse) af data før den egentlige dataanalyse kan begynde, samt de endnu mere omfattende statistiske og datalogiske modeller, der anvendes til at omsætte de mange tusinde datapunkter til biologisk eller klinisk relevant information (*high level*-analyse) (**Figur 1**).

Præprocessering og signalekstraktion

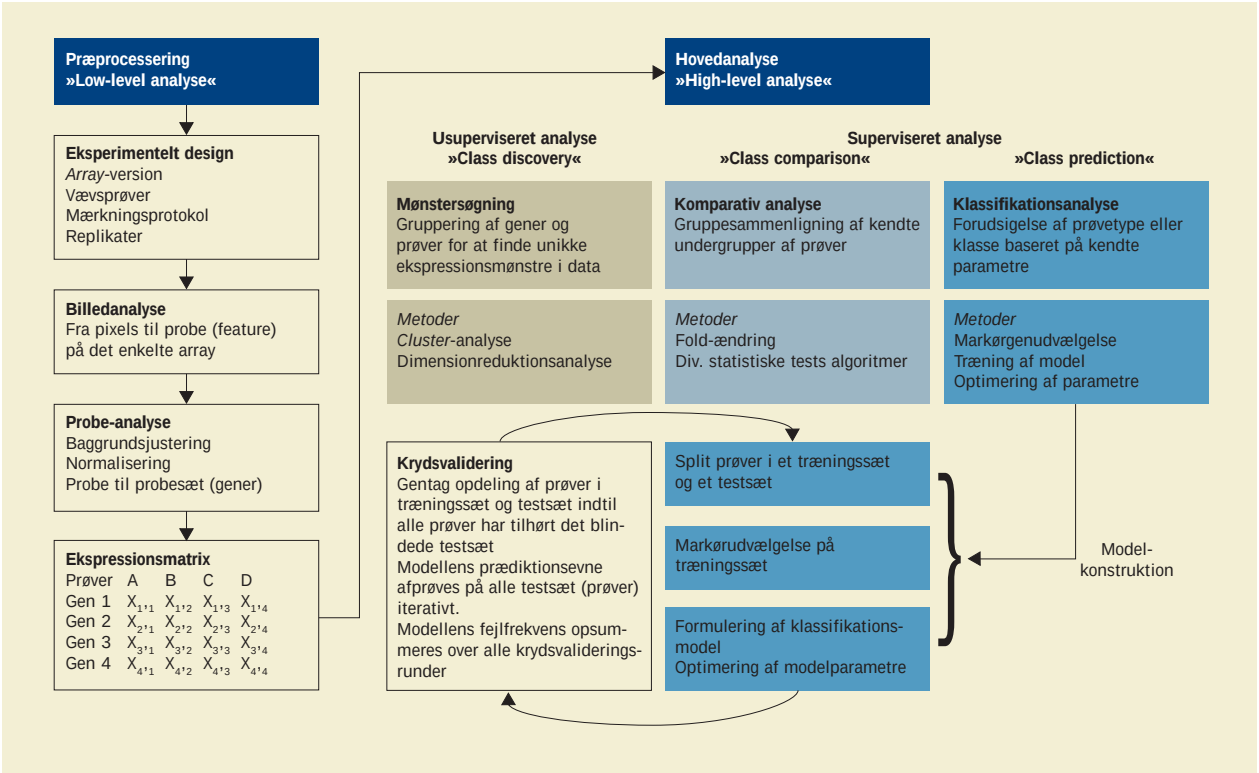
I den følgende beskrivelse fokuseres der på metoder anvendt på Affymetrix GeneChip data.

Første trin i *microarray*-dataanalysen er præprocessering af rådata, dvs. den datafil, der indeholder aflæste fluorescensintensiteter for de enkelte prober. Først udføres billed- og kvalitetskontrol ved visualisering af *array*'et, og for den enkelte probe bestemmes en gennemsnitlig intensitet i form af gennemsnittet af pixel i det 11-18 micron store probefelt (billedanalyse, Figur 1). Dernæst justeres for baggrundsintensitet for at give forhøjet *signal to noise*-ratio. Gennemsnitsintensiteterne gemmes i en såkaldt cel-fil, som kan importeres i en række af *microarray*-dataanalyseprogrammer, som efterfølgende kan anvendes til at normalisere datafilerne med. Under normaliseringen korrigeres intensiteterne for systematisk variation, som er introduceret under prøvemærkning, hybridisering og skanning. Prøverne bliver dermed sammenlignelige, og den tilbageværende variation repræsenterer ideelt set den biologiske forskel imellem prøverne.

Normaliseringen foregår på probeniveau, men for at kunne udføre genekspressionsanalyse, skal de multiple probedata for hvert probesæt summeres til at give en ekspressionsværdi, der repræsenterer det enkelte mRNA. Denne probesummering er særegen for Affymetrix, da hvert mRNA er repræsenteret af op til 22 prober (probeanalyse, Figur 1).

Hovedanalyse

Formålet med *microarray*-analyser er ofte identifikation af gener, der er forskelligt udtrykt mellem forskellige grupper af prøver (komparativ analyse). Et relateret formål er anvendelse



Figur 2. Skematisk oversigt over et fuldt *microarray*-dataanalyse-forløb. Præprocessering foretages på alle rådatafiler, der indeholder fluorescensintensiteter for den enkelte probecelle målt under skanning af *microarray*'et. Alle intensiteter omsættes til en talværdi, der reflekterer, hvor meget fluorescensmærket prøvemateriale, der er bundet til den enkelte probe. I denne proces korrigeres der for baggrundssignal, før intensiteten fra alle prober, der præsenterer et givet gen, sammenfattes til en genekpressionsværdi under probeanalysen. Præanalysen af *microarray*-filerne giver som resultat sammenlignelige (normaliserede) ekspressionsværdier for alle de mRNA, der undersøges på *microarray*'et. Ekspressionsværdier for alle *microarrays* i et projekt sammenfattes ofte i en ekspressionsmatrix, hvor gener er angivet i rækker og prøver i søjler. Hovedanalysen afhænger af den biologiske hypotese, der testes, og valget af metode reflekterer derfor analysens eksperimentelle design og formål. Mønstersøgning bruges ofte til at finde nye genekspressionsstrukturer med eller til at definere nye undergrupper af prøver med. Komparativ analyse er standardmetode for et kontrol- vs. et eksperimentforsøg, hvor man sammenligner to eller flere grupper af prøver mod hinanden, med henblik på at selektere de gener, der er differentielt reguleret mellem de forskellige grupper. Klassifikationsanalyse anvendes til prognose af f.eks. kræftsubtyper. Klassifikationsanalyse kan overordnet opdeles i tre trin: 1) selektion af markørgener, 2) specifikation af den klassifikationsmodel, som afgiver det prognostiske resultat – herunder optimering af parametre i modellen (matematisk algoritme) og 3) validering af *performance* på nye prøver. Optimering og fejlestimering foregår ved krydsvalidering, dvs. at prøver i analysen skiftevis holdes udenfor under markørselektion og modeloptimering og efterfølgende anvendes, når man tester modellens anvendelighed mht. at placere dem i den korrekte klasse.

af *microarrays* til indentificering af molekylære markører eller ekspressionssignaturer (mønstre) til klassifikation eller opdeling af vævsprøver i henhold til sygdomskategori (klassifikationsanalyse).

Man anvender to overordnede analysemetoder – usuperviseret analyse, hvor man ikke anvender information om prædefinerede klasser i dataanalysen, og superviseret analyse, hvor man inddrager kendte kliniske parametre, såsom behandlingsrespons og overlevelsestid i dataanalysen (**Figur 2**).

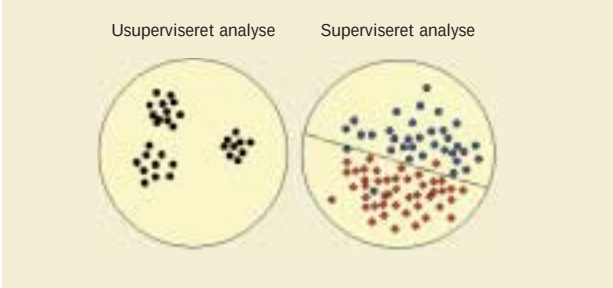
I den usuperviserede analyse bruges der ofte mønstersøgnings- og grupperingsalgoritmer såsom hierarkisk *cluster*-analyse og *self-organizing maps* (SOM). Hierarkisk *cluster*-analyse starter med, at to gener, der har et korreleret ekspressionsmønster, grupperes. Algoritmen køres iterativt, indtil alle gener er placeret i *clusters*. Ligheder mellem grupperede gener visualiseres med en træstruktur kaldet et dendrogram. Længden af grenene i dendrogrammet angiver ligheden mellem genernes ekspression, idet kortere grene angiver større similaritet. Til forskel fra den hierarkiske *cluster*-metode grupperer SOM-algoritmen gener i *clusters*, hvor antallet af *clusters*, der dannes, er angivet som inputparameter. Ved brug af *cluster*-analyse er det muligt i micro-*array*-data at visualisere og finde strukturer, der har relation til de biologiske tilstande som undersøges [3-5].

Den superviserede analyse har typisk to formål: at identificere gener, der er differentielt udtrykt mellem grupper af prøver, og at finde gener, vha. hvilke man kan forudsige prøvers klassetilhørsforhold.

En superviseret klassifikationsanalyse tager udgangspunkt i en prædefineret gruppeinddeling af prøver. Eksempelvis kan man indsamle prøver fra kræftpatienter, før man påbegynder behandling, og derefter opdele prøverne i henhold til, om patienterne efter behandling er gået i komplet remission eller ikke har responderet på kemoterapi. I dette tilfælde er målet med datanalysen at definere de gener, som man bedst kan bruge til at beskrive hhv. godt og dårligt behandlingsrespons med, og at anvende disse gener til at opstille en matematisk model, som kan anvendes, når man skal forudsige fremtidige patienters behandlingsrespons.

Eksempler på kræftstudier, hvor prognostiske metoder eller klassifikationsmetoder er anvendt, omfatter bl.a. studier af akut leukæmi [4], diffust storcellet B-celle-lymfom [6, 7] og blærekræft [8]. *Microarrays* er også anvendt til diagnosticering af ukendte primærtumorer baseret på ligheder med kendte kræftformer.

Der er tre stadier i en klassifikationsanalyse; det første stadie er genselektion, det andet stadie er specifikation af den matematiske algoritme (prædiktor) og dens indgående parametre på basis af de udvalgte gener, og det tredje stadie i analysen er validering af modellen på uafhængige datasæt. Typisk opdeles prøverne i to grupper, et træningssæt og et testsæt. Genselektion og parameterspecifikation foregår på træningssættet, og validering af modellen udføres på testsættet.



Figur 2. Usuperviseret og superviseret analyse.
Formålet med en usuperviseret analysemetode er at identificere unikke ekspressionsmønstre eller inddele ekspressionsprofiler (prøver) i grupper uden at inddrage viden omkring f.eks. sygdomsstadie eller gruppetilhørsforhold (eksempelvis tumortype). En superviseret analysemetode inddrager forudgående viden om de prøver, der analyseres (f.eks. rød eller blå label). Viden om prøvetype bruges til at selektere markører (gener), der har den største korrelation med de enkelte undertyper. Baseret på de udvalgte markører bygges en *classifier* (matematisk model), hvis parametre optimeres gennem krydsvalidering til at give den bedste opdeling af prøverne i rød og blå kategori. Nye prøver kan herefter klassificeres som blå eller rød type, afhængigt af på hvilken side af den stiplede linje de falder, når de køres igennem *classifier*-modellen.

Alfa og omega for en prædiktor er, at man vha. af den opnår at kunne give korrekte forudsigelser af nye kliniske prøver, hvis klassetilhørsforhold er ukendt. Disse forudsigelser udgør resultatet fra modellen.

En af de større faldgruber i forbindelse med træning af en klassifikationsmodel på *microarray*-data er *overfitting* eller *underfitting*. *Overfitting* kan forekomme, når antallet af parametre (gener) i modellen er meget større end antallet af prøver, og modellen bliver for kompleks. Ved *overfitting* opstilles der ud fra data i træningssættet en prædiktor, som er så »god«, at man vha. den ud over de generelle forskelle, f.eks. mellem to kræftsubtyper, også kan inddrage variationer i data, der er irrelevante med hensyn til klassifikation af subtyper. Dermed kan man vha. modellen bedre opdele prøverne i træningssættet i de »rigtige« grupper, men man kan ikke benytte den til generaliseret at virke på nye uafhængige data fra patienter med den samme sygdom. Ved *underfitting* opstilles en model, som enten er for simpel, når man skal indlære og beskrive de essentielle egenskaber i de data, der modelleres, eller hvor de indgående parametre ikke er tilstrækkelig optimerede [9].

En metode til at minimere *overfitting* på er at estimere og optimere en fejlfrekvens for alle de varianter af modellen, der opstilles under gentagne krydsvalideringer. I en krydsvalidering udelukkes prøverne i træningssættet på skift, og de bedste gruppediskriminerende gener udvælges i de tilbageværende prøver. Disse gener fødes ind i en række af klassifikationsmodeller, som hver især trænes, optimeres og testes på den eller de prøver, der er udelukket fra træningssættet. Denne proces gentages flere gange, og den gennemsnitlige succesrate for hver gennemkørsel anvendes til estimering af en overordnet fejlfrekvens for den enkelte klassifikationsmodel. Endelig bliver den optimale model, dvs. den som med et passende antal gener kan give en acceptabel prognostisk nøjagtighed

(f.eks. 90% rigtige klassifikationer), udvalgt. Modellens prognostiske fejlrate på trænings- og testsætdata sammenlignes med den observerede fejlrate, som modellen giver, når den køres på tilfældigt permuterede datasæt (tilfældig gruppering af prøver), for at afgøre hvorvidt en lige så god klassificering kan opnås tilfældigt. Til sidst bestemmes den sande prognostiske succesrate ved afprøvning af modellens klassfikationsevne på nye uafhængige data.

Som eksempel på ovennævnte metoder gives i det følgende en kort beskrivelse af et *microarray*-studie, hvori man klassificerede diffus storcellet B-celle-lymfom (DLBCL)-subtyper. I dette studie indsamlede og analyserede vi 52 lymfomprøver med henblik på en *cross-platform* (*spotted arrays* versus Affymetrix GeneChips)-validering af markører og DLBCL-undergrupper. Præprocessering blev foretaget med dChip (DNA-chip analyzer, http://www.dchip.org). Cel-filer blev importeret i dChip-programmet og normaliseret, og ekspressionsværdier for det enkelte gen blev beregnet. En undergruppe på 34 prøver, som konsistent kunne kategoriseres som hhv. *germinal center B-cell* (GCB) eller *activated B-cell* (ABC)-subtyper baseret på tidligere publicerede genmarkørlister fra andre *microarray*-analyser blev anvendt i en superviseret analyse [5, 6]. Vi udførte en gruppesammenligning med henblik på at definere et sæt gener til brug i en prædiktor, dvs. den matematiske algoritme, som skal trænes og optimeres til at klassificerer GCB- og ABC-type-lymfomer. Welch t-test-gruppesammenligning med testsandsynlighed under 0,05 førte til, at 4.586 gener blev udvalgt. De udvalgte gener blev rangordnet baseret på en *signal to noise*-score og testet for succes i GCB- og ABC-klassifikation med multiple krydsvalideringsrunder. Fire klassifikationsalgoritmer, *weighted voting* og K-*nearest neighbours* (K-NN) med tre forskellige K-værdier (K lig 3, 5 og 7) blev testet med et varierende antal markørgener som input (40 genlister, der indeholdt 1-4.000 gener) [4, 10]. I alle modeller sås ens succesrate med hensyn til klassifikation af GCB- og ABC-undergrupper baseret på krydsvalideringstest, og der var ikke mulighed for at definere en gruppe af markørgener med særlig god klassifikations-performance. En forbedret og mere specifik genliste blev herefter udvalgt ved at korrigere de t-test-udvalgte 4.586 markører for falsk positive med en såkaldt *multiple testing correction*-procedure kaldet Holmes step down maxT test. Denne komparative analyse resulterede i 78 højt signifikante gener, som gav god adskillelse af de to lymfomsubtyper. De 78 markørgener blev valideret på uafhængige prøver fra andre publicerede studier [6, 7] og gav i et af tilfældene en forbedret klassifikation af GCB- og ABC-grupper med signifikant forskel i femårsoverlevelsen på 63% for GCB og 43% for ABC i forhold til, hvad der tidligere var rapporteret om [7, 10].

Perspektiver

Det er veletableret, at man ved ekspressionsanalyse kan genfinde de overordnede histologiske grupperinger i kræftsygdomme. Samtidig har man i en mangfoldighed af studier påvist, at *microarrays* kan bidrage med en yderligere definition af nye grupper. Kræfttyper detekteret med *microarray*-analyse inkluderer ofte »små« molekylære ændringer, der har betydning for prognose og behandlingsrespons. Ved hjælp af *microarray*-teknologien kan molekylære markører udvælges, og de kan anvendes, når man skal definere grupper af tumorer, der udviser forskelle i behandlingsrespons og aggressivitet. En naturlig følge af udviklingen inden for området er muligheden for at definere globale tumormarkører for kendte kræftsygdomme. Disse globale tumormarkører er anvendelige på tværs af *spotted* og oligonukleotid–*microarray*-systemer og kan benyttes i den diagnostiske proces. Et skridt på vejen mod validering af generelle prognostiske markører er gennemførelse af større, prospektive, kliniske forsøg, som blandt andet indbefatter en standardprotokol for ensartet indsamling og håndtering af væv og grænseværdier for det minimale tumorcelleindhold i vævsprøven samt større fokus på de matematiske metoder, der anvendes til behandling af *microarray*-data. Valg af normaliseringsalgoritme og efterfølgende statistiske anlysemetoder er af stor betydning for kvaliteten og validiteten af *microarray*-analysen, og det skal understreges, at et *microarray*-datasæt kan og bør analyseres på forskellige måder med henblik på at få ekstraheret mest mulig biologisk information.

Korrespondanceansvarlig: *Rehannah H.A. Borup,* RH Microarray Center, Klinisk Biokemisk Afdeling, H:S Rigshospitalet, DK-2100 København.
Email: rborup@rh.dk/rehannah@email.dk

Antaget: 27. september 2005
Interessekonflikter: Ingen angivet

Litteratur

1. Duggan DJ, Bittner M, Chen Y et al. Expression profiling using cDNA microarrays. Nature Genet 1999;21:10-4.
2. Lipshutz RJ, Fodor SP, Gingeras TR et al. High density synthetic oligonucleotide arrays. Nature Genet 1999;21:20-4.
3. Eisen MB, Spellman PT, Brown PO et al. Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci USA 1998;95:14863-8.
4. Golub TR, Slonim D, Tamayo P et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.
5. Alizadeh AA, Eisen MB, Davis RE et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503-10.
6. Rosenwald A, Wright G, Wing C et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large B-cell lymphoma. N Engl J Med 2002;246:1937-47.
7. Shipp MA, Ross KN, Tamayo P et al. Diffuse B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8:68-74.
8. Dyrskjøt L, Thykjær T, Kruhøffer M et al. Identifying distinct classes of bladder carcinoma using microarrays. Nat Gen 2003;33:90-6.
9. Hastie T, Tibshirani R og Friedman J. The Elements of statistical learning. New York: Springer-Verlag, 2001.
10. Poulsen CB, Borup R, Nielsen FC et al. Microarray-based classification of diffuse large B-cell lymphoma. Eur J Haematol 2005;74:453-65.