Ny kunnskap fra måling av genuttrykk gjennom DNA-mikromatrise

DNA-mikromatrise (mikroarray) er en metode som gir oss muligheten til en helhetlig forståelse av biologiske prosesser i levende organismer ved parallell avlesning av genuttrykksdata for titusener av gener. Gener med samme funksjon synes å ha liknende uttrykksmønster, men å tolke uttrykksmønsteret av 30 000–50 000 gener krever helt nye metoder for dataanalyse. Datagruvedrift er den sentrale metoden for dataanalyse med mål å oppdage ny kunnskap. Slike metoder gjør det mulig å «lære» regler for sammenhengen mellom genuttrykk og genfunksjon. Reglene er bygd på kunnskap om sammenhenger mellom genuttrykk og genfunksjon, og kan benyttes til å klassifisere ukjente gener. Sentralt i kunnskapsoppdagelse og datagruvedrift står gjenbruk av eksisterende biologisk og medisinsk kunnskap. Det trengs effektive metoder for å trekke ut slik kunnskap fra eksisterende kilder, som for eksempel litteraturdatabaser (Medline, etc.). Vi omtaler her en metode for å oppdage ny kunnskap med utgangspunkt i mikromatrise genuttrykksmålinger. Metoden frembringer modeller som beskriver sammenheng mellom genuttrykk og genfunksjon. Modellene kan så brukes til å danne hypoteser om funksjonen for ukjente gener og finne nye funksjoner for kjente gener. Metoden har vært testet på offentlig tilgjengelige genuttrykksdata, men blir også brukt i vårt eget mikromatrisesystem. Prinsippene er av generell betydning og vil få anvendelse i vurderingen av andre komplekse datasett, for eksempel i klinisk praksis hvor store datamengder foreligger for hver enkelt pasient.

Jan Komorowski
Torgeir R. Hvidsten
Tor-Kristian Jenssen
Dyre Tjeldvoll
Institutt for datateknikk og informasjonsvitenskap
Norges teknisk-naturvitenskaplige universitet
7491 Trondheim

Eivind Hovig
Avdeling for tumorbiologi
Det Norske Radiumhospital
0310 Oslo

Astrid Lægreid
Arne K. Sandvik
Institutt for fysiologi og biomedisinsk teknikk
Norges teknisk-naturvitenskaplige universitet
7489 Trondheim

Komorowski J, Hvidsten TR, Jenssen T-K, Tjeldvoll D, Hovig E, Lægreid A, Sandvik AK.

New knowledge derived from measurement of gene expression with the DNA microarray method.

Tidsskr Nor Lægeforen 2001; 121: 1229–32.

Background. The cDNA microarray method offers the first possibility of obtaining a global understanding of biological processes in living organisms, by simultaneous read-outs of tens of thousands of mRNAs. Initial experiments suggest that genes with similar function have similar expression patterns.

Material and methods. Understanding this level of biological complexity will, however, require completely new approaches to data analysis. Computer science methods, such as data mining and knowledge discovery, can synthesize interpretable if-then rules that model the relation between gene expressions and functions and use the rules to classify unknown genes. The huge body of existing biological and medical knowledge makes it necessary to develop methods for extracting knowledge from such repositories.

Results. Models of relations between gene expressions and gene functions in a data set from a publicly available source are synthesized semi-automatically and applied to classify unknown genes. Encouraging results have been achieved. The method is applied in the analysis of data from our microarray system which has recently become operational.

Interpretation. The principles are of general importance and will be used to evaluate a wide range of complex data sets like decision support in clinical medicine, for situations in which physicians need to handle a large volume of data for each patient.

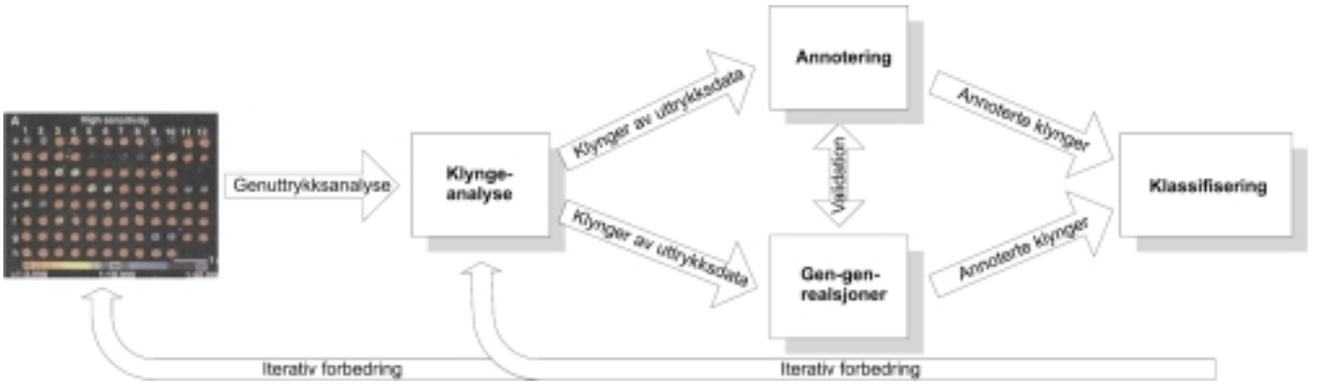
Liv hos mennesker og dyr beror på et komplekst samspill mellom mange forskjellige celler og celletyper som styres av en lang rekke ytre og indre faktorer. Hittil har molekylærbiologer og molekylærmedisinere undersøkt genetiske og biokjemiske mekanismer med metoder som gir svært lite informasjon i forhold til det store antall parametere som styrer levende organismer. Ved hjelp av mikromatriseteknologien (1) er det nå mulig å måle aktiviteten eller uttrykksnivået til flere tusen gener i én analyse. I tillegg pågår en rekke forskningsprogrammer som generer data i store mengder, som for eksempel Humant genom-prosjektet. Dette fører til at enorme mengder med opplysninger om levende systemer blir tilgjengelig. Tilgangen på data er så overveldende at man nå ser behovet for nye metoder for informasjonsbehandling. Mulighetene for å finne ny kunnskap fra disse data ser ut til å være svært store og vi nevner bare noen av dem her: Forbedret diagnose av sykdommer som for eksempel subklassifisering av kreft for bedre beregning av prognose og behandlingsvalg, medisiner skreddersydd til et individs genetiske profil, eller oppdagelse av parametere som vil komme til nytte innenfor epidemiologi via økt tilgang til molekylær, cellulær og annen biomedisinsk informasjon.

Oppdagelse av nye egenskaper ved gener

Å oppdage nye egenskaper ved gener, for eksempel funksjon, er en utfordrende oppgave. Mennesket har sannsynligvis 30 000–50 000 gener. Alle disse genene finnes i alle celler, men bare noen er uttrykt (tatt i bruk) i en gitt celle til et gitt tidspunkt. De uttrykte genene bestemmer cellens fenotype og funksjonelle tilstand. Derfor kan måling av genuttrykk brukes til å beskrive cellens aktiveringsgrad i øyeblikket, eksempelvis synteseaktivitet, differensieringsgrad og proliferasjonsstatus.

Mikromatriseanalyse gir oss muligheten til å se organiseringen av cellulært liv gjennom kvantitative målinger av genuttrykk (mRNA-nivå). Dette kan brukes til å forstå hvordan celler fungerer (2). Cellers reaksjoner på eksterne stimuli og overganger mellom tilstander kan observeres. I tillegg kan sykdommer studeres gjennom genuttrykk. Forskjellen i genuttrykk i friske og syke celler kan analyseres, og denne kunnskapen kan benyttes til sykdomsforståelse og diagnose.

I dag kjenner vi funksjonen til bare ca. 5 000 gener. Mikromatriseanalyse vil spille en viktig rolle i arbeidet for å bestemme biologisk funksjon for de resterende genene. En fundamental antakelse i denne forskningen er at gener med lik funksjon har liknende uttrykksprofil. Ikke-veiledede læringsmetoder, som for eksempel algoritmer (strategier) for klyngeanalyse, kan brukes til å gruppere gener med liknende uttrykksprofil.



Figur 1 *Et skjema for kunnskapsoppdagelse fra genuttrykk: Genene blir først gruppert basert på likhet i uttrykk ved hjelp av en algoritme for klyngeanalyse. Deretter blir genene i klyngene merket med typisk funksjon/prosess (annotert). Annoteringene gir et treningssett som kan brukes til å lære dataprogrammet en klassifikator for funksjonsforutsigelse av ukjente gener*

Klyngeanalysen gir en rekke muligheter til å visualisere, oppsummere og beskrive komplekse mikromatrisedata, og er således et nyttig verktøy for å redusere mengdene med informasjon som krever manuell tolking. Likevel er det ikke sannsynlig at klyngeanalyse kan gi alle svarene, rett og slett fordi det er opp til eksperter å komplettere analysen med bakgrunnskunnskap, dvs. finne sammenhenger mellom klynger av liknende uttrykksprofiler og biologisk funksjon. Et annet problem med klyngeanalyse er at det ikke finnes noen riktige svar. Gitt to metoder som produserer forskjellige klynger, er det vanskelig å sammenlikne og bestemme hvilken som er best. Få gode, objektive kriterier finnes, og selv med omfattende bakgrunnskunnskap er det ofte ikke mulig å finne en sammenheng mellom uttrykksprofiler gruppert ved klyngeanalyse og biologisk funksjon for genene i klyngen (3).

Vi mener at kunnskapsoppdagelse fra genuttrykk krever nye tilnærmingsmetoder hvis kompleksitetsbarrieren skal brytes. I vår metode prøver vi å komme bort fra de rent ikke-veiledede læringsmetodene som ser ut til å dominere forskningen i dag, og vise hvordan man kan bruke forskjellige typer bakgrunnskunnskap for å oppdage og validere ny viten. For å få til dette bruker vi offentlig tilgjengelige genomdatabaser, artikkeltitler og abstrakt fra Medline og Gene Ontology (4). Den sistnevnte er en systematisering av geners molekylære funksjon (f.eks.enzymaktivitet), betydning i biologisk prosess (f.eks. rolle i komplekse synteseveier) og subcellulære lokalisering (f.eks. mitokondrier). Resultatene blir bedømt av biomedisinske forskere.

Foreløpig har bare et ganske begrenset antall forskere sett på problemene omkring analyse av genuttrykksdata. Den vanligste metoden er sammenhopende hierarkisk klyngeanalyse (5) der genene/klyngene med mest likt uttrykk suksessivt slås sammen og danner et hierarki. Ett praktisk eksempel er bruk av genuttrykksdata i differensialdiagnostikk mellom akutt lymfatisk og akutt myelogen leukemi (6). Det har også vært arbeidet med metoder for læring fra genuttrykksdata med tanke på klassifisering av ukjente gener (7). Det er sannsynlig at flere firmaer involvert i genuttrykksanalyse har utviklet egne «pakker» med verktøy og metoder og av konkurransehensyn ikke gjør disse offentlig tilgjengelige. Vi presenterer her hovedstrukturen i kunnskapsoppdagelsessyklusen, med en forklaring på hvert steg i syklusen: klyngeanalyse, annotering, validering og klassifisering. Artikkelen avsluttes med konklusjoner og en diskusjon rundt kommende forskning.

En syklus for kunnskapsoppdagelse fra genuttrykk

DNA-mikromatriseteknologi gjør biomedisinske forskere i stand til å måle uttrykksnivået til tusener av gener i et eneste eksperiment, og i tidsstudier kan man følge endringer i genuttrykk for svært mange gener gjennom en biologisk respons. Tidlige eksperimenter (8) viste at gener med samme funksjon tenderer til å produsere liknende uttrykksprofil i mikromatrise hybridiseringseksperiment. Med ett unntak (7), bruker de fleste metodene ikke-veiledet læring for å oppdage de funksjonelt signifikante genene. Ikke-veiledet læring definerer først likhet mellom genuttrykk til to gener, og bruker deretter denne definisjonen til å finne grupper av gener med liknende uttrykksprofil. Hierarkisk klyngeanalyse (8) og selvorganiserende kart (9) er blant annet brukt. Den sistnevnte algoritmen er basert på ideen bak kunstige nevrale nett og tilhører den familien av algoritmer for klyngeanalyse som gjennom trinnvis tilnærming gradvis gir gode klynger. Disse ikke-veiledede metodene gjør ikke bruk av bakgrunnskunnskap. En veiledet læringsmetode, derimot, benytter gener med kjent funksjon eller prosess til å lære karakteristiske trekk som kjennetegner disse genene. Veiledet læring er derfor i prinsippet mer kraftfullt, siden den kan bruke kjente tilfeller til å lære en definisjon av klasser (et eksempel på en prosessklasse kan være kolesterolbiosyntese). Slike definisjoner kan så benyttes til å klassifisere ukjente gener.

Selv om ikke-veiledet læring er attraktiv som metode, vil den bare delvis være egnet til å lære funksjonelle klasser fra genuttrykk. Kompleksiteten i de underliggende prosessene som reflekteres gjennom endringer i genuttrykk, krever sofistikerte likhetsdefinisjoner og gjør det usannsynlig at mange nye oppdagelser kan gjøres. I tillegg vil biomedisinske eksperter før eller siden måtte kvalitetskontrollere klyngene og tilordne dem klasser. Dette er en formidabel oppgave som selv med full adgang til alle dagens genomdatabaser kan ta opptil flere timer for ett enkelt gen. Vi foreslår en noe annerledes tilnærming som er skjematisk illustrert i figur 1. Genuttrykksprofilene blir først gruppert ved hjelp av en algoritme for klyngeanalyse som egner seg for biologisk tolking. Så blir genene i klyngene merket med typisk funksjon/prosess (annotert) ved bruk av to kilder til kunnskap. Den første kilden til (formalisert) biologisk kunnskap er Gene Ontology (4). Den andre kilden til kunnskap er utviklet av oss ved å analysere Medline-siteringer og bygge gen-gen-relasjoner hvis genene er nevnt i samme artikkel. I tillegg kan disse gen-gen-relasjonene brukes til å validere allerede eksisterende annoteringer. Når vi har annotert genene, er det mulig å bedømme klyngenes kvalitet. Annoteringene gir oss også en samling gener med kjent funksjon (treningssett) som kan brukes til å lære en modell. Denne modellen er et sett av regler som representerer forholdet mellom uttrykksprofiler og biologisk funksjon. Modellens klassifiseringskvalitet kan evalueres med statistiske metoder og brukes til å klassifisere ukjente gener, altså fremsette hypoteser om deres biologiske funksjon.

Datagruvedriftsmetoder

Vi vil her illustrere våre metoder ved hjelp av data fra en publisert studie (10) av serumrespons hos fibroblaster i cellekultur. Ved hjelp av DNA-mikromatrise ble uttrykk av 8 613 gener målt ved 12 forskjellige tidspunkter i en periode på 24 timer etter serumstimulering. Man fant 517 gener som viste en vesentlig forandring i uttrykksnivå og derfor ble valgt ut for videre analyse. Iyer og medarbeidere (3) valgte en klyngeanalysemetode som krevde lik forandring av uttrykksnivå over hele perioden på 24 timer.

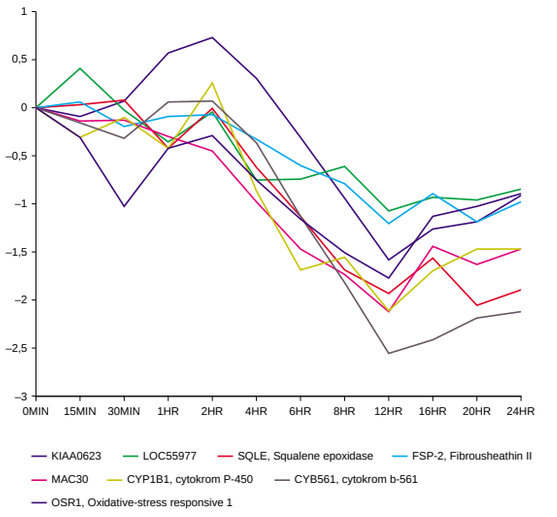
Klyngeanalyse

Å kreve at endringene i genuttrykkene skal være like over hele intervallet på 24 timer virker noe urealistisk med tanke på den komplekse biologiske prosessen som studeres. Vi valgte i stedet en klyngeanalyse som undersøker alle mulige subintervaller og som dermed grupperer gener ut fra liknende uttrykksprofil i et kortere tidsrom, samtidig som de samme genene kan ha ulike uttrykksforløp i et annet subintervall. Dermed får vi klynger som reflekterer karakteristiske forandringer i genuttrykksmønster over alle mulige tidsrom innenfor de 24 timene. Denne klyngeanalysen fanger opp det store spennet i genuttrykksmønster som reflekterer kompleksiteten av molekylære hendelser i fibroblastene, og forenkler dermed oppgaven med biologisk tolking (fig 2).

Validering av klynger

Etter å ha utført ikke-veiledet læring i form av klyngeanalyse, blir neste skritt å validere klyngene ved å annotere hvert gen til den prosessen genet er kjent for å delta i. Manuell annotering utført av eksperter er en svært tidkrevende oppgave, og ideelt sett burde denne annoteringen vært gjort automatisk fra lagre av biologisk kunnskap. Vi arbeider med å utvikle metoder for automatisert genannotering ved å bruke informasjon om annoterte gener fra modellorganismer og biologisk kunnskap fra litteraturdatabaser. Behovet for en strukturert metode for å relatere gener til hverandre på basis av funksjon og prosess har skapt genontologier. Ontologi (læren om «eksistens») i denne forstand er klassifisering av gener i forhold til biologisk funksjon. En genontologi er også en hierarkisk struktur, et tre, der foreldrenodene gir en mer generell beskrivelse av et gen enn barnenodene. Løvnodene er presise beskrivelser av hvert enkelt gen (fig 3). Som man kan se av figuren er genontologi (4) delt inn i tre kategorier på toppnivå: prosess, funksjon og subcellulær lokalisering. Dette er en formalisert kilde av biologisk kunnskap av høy kvalitet og vil bli viktig i studier av genuttrykk.

Kunnskapen om annotasjoner kan nå brukes til å validere klyngene fra klyngeanalysen beskrevet tidligere. Hvis gener i en klynge har liknende annotasjon (er liknende plassert i ontologien), tyder dette på at klyngen beskriver den funksjonen eller prosessen annotasjonen indikerer. I disse tilfellene styrkes hypotesen om at de ikke-ontologiklassifiserte genene (gener som vi ikke har noe kunnskap om og som derfor heller ikke har noen annotasjon) som finnes i denne klyngen også innehar denne, eller en nært beslektet, funksjon eller prosess. I realiteten gir annoteringene oss et verktøy for å teste kvaliteten, eller den prediktive styrken, til klyngene før vi gjør antakelser om de ukjente genene som finnes i klyngen. Vi vil vise dette ved å gå tilbake til eksemplet i figur 2. Denne klyngen inneholder åtte gener hvorav fire er navngitt. To av de navngitte genene, squalene epoxidase og cytokrom P-450, har annotasjon «kolesterolbiosyntese». Et annet, cytokrom b-561, har annotasjon «elektrontransport», noe som også kan indikere en mulig rolle i kolesterolsyntesen. Genet oxidative-stress responsive 1 koder for et protein indusert av redoksprosesser, dvs. en type prosess som også er involvert i kolesterolbiosyntese. EST, KIAA0623, FSP-2 og MAC30 er ikke navngitt og koder for proteiner med ukjente funksjoner. For biomedisinere er relasjonen mellom de navngitte annoterte genene slående. Klyngen med gener involvert i redoksprosesser, og mer spesifikt kolesterolsyntese, i dette eksemplet antyder at noen av genene med ukjent funksjon også kan være involvert i liknende prosesser (11).

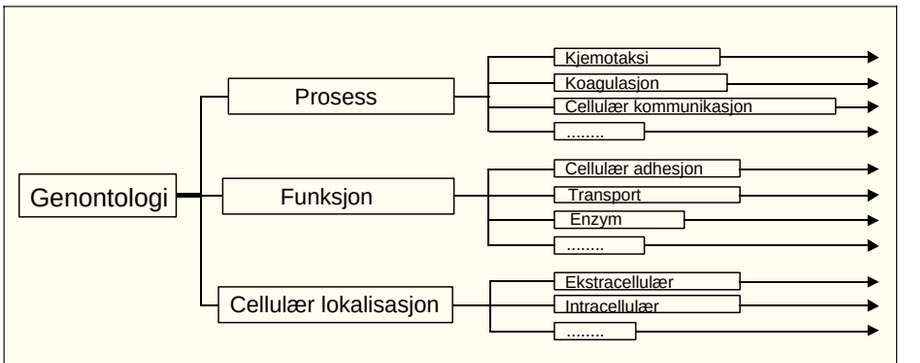


Figur 2 *Et eksempel på en klynge av gener med liknende uttrykksprofil (synkende genuttrykk i subintervallet 2–12 timer). Åtte gener, hvorav fire navngitte, tilhører denne klyngen og deres genuttrykksnivå er plottet som funksjon av tid. Alle de fire navngitte genene, squalene epoxidase, cytokrom P-450, cytokrom b-561 og oxidative-stress responsive 1, har annotasjon som indikerer en rolle i kolesterolsyntese. Genene EST, KIAA0623, FSP-2 og MAC30 koder for proteiner med ukjente funksjoner*

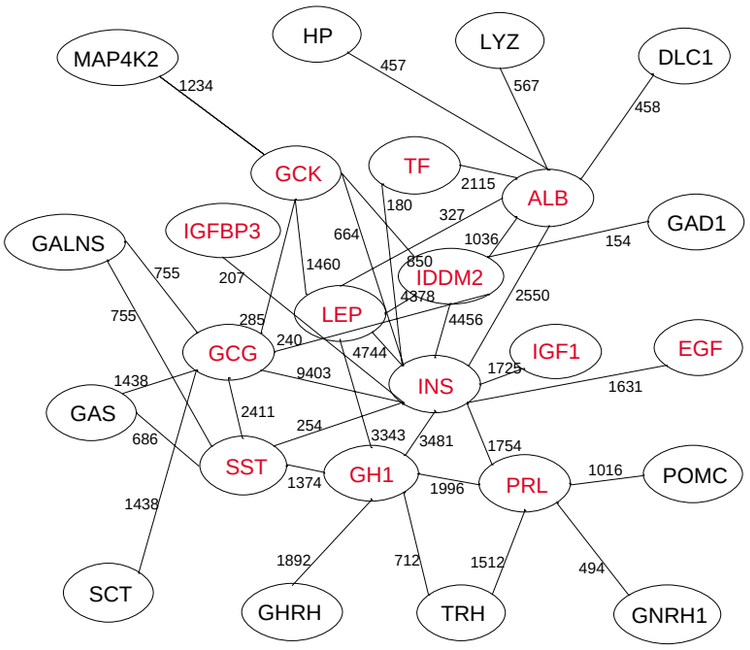
Validering av annotasjoner – PubGene

Mye av den biologiske kunnskapen som trengs for å annotere gener er ikke organisert i en ontologi, men finnes som tekstdokumenter. Vi presenterer her en metode for å finne gen-gen-relasjoner fra slike dokumenter. Disse relasjonene kan brukes til å validere allerede eksisterende annotasjoner og klynger, men også til å finne nye annotasjoner. Vi har laget en database av gen-gen-relasjoner ved å analysere siteringer i Medline. To gener er relaterte hvis de er nevnt i samme artikkel. Gen-gen-relasjonene i databasen utgjør derfor et litteraturbasert nettverk av gener. En webside er konstruert og tilbyr en presentasjon av gen-gen-nettverket der brukeren kan studere relasjoner og navigere rundt blant naboene til hvert gen (12).

Databasen som utgjør vår egen liste av humane gener er satt sammen ved å samle og



Figur 3 *Genontologi (bare toppnivåene er vist). Genontologier gir en hierarkisk organisering av kunnskap om gener*



Figur 4 *Gen (nabo)-nettverk fra Medline for insulin (INS)-genet. Naboene til genet opptrer som noder med kanter som er vektet med antall gjensidige siteringer. Sentrale naboer er genene LEP (leptin), IGF1 (insulinliknende vekstfaktor 1), PRL (prolaktin), GCG (glukagon), SST (somatostatin), GH1 (veksthormon 1), EGF (epidermal vekstfaktor), IDDM2 (insulinavhengig diabetes mellitus 2), GCK (glukokinase), ALB (albumin), og TF (transferrin)*

strukturere data fra offentlig tilgjengelige databaser og omfatter en liste med 13 712 gener. Alle publikasjoner i Medline (vel ti millioner) ble gjennomsøkt etter siteringer for disse genene. I om lag 15 % av artiklene ble ett eller flere gener funnet omtalt. I det grafiske vinduet på vår hjemmeside er naboene til hvert gen presentert som et nettverk sentrert rundt dette genet, som vist for insulingenet i figur 4. Naboene til genet opptrer som noder med kanter som er vektet med antall gjensidige siteringer. Brukeren kan navigere rundt i nettverket ved å klikke på en nabo i nettverket og dermed fokusere på dette genet. Dette arbeidet viser at automatisert litteraturanalyse kan utnyttes til å trekke ut biologisk meningsfull informasjon. I tillegg viser vi at metoden kan brukes i stor skala.

Læring av genklassifikatorer

En bedre metode enn den rene klyngeanalysen er å lære definisjoner av klasser av gener med samme biologiske funksjon. Annotasjonene deler genene i slike klasser, for eksempel «kolesterolbiosyntese». Vi får da samlinger av genuttrykksprofiler som er antatt å være representative for biologiske prosesser og som danner utgangspunkt (treningssett) for veiledet læring. En egnet metode er å anvende Pawlaks «rough set»-rammeverk (14, 15) for regellæring og klassifisering slik det er brukt i ROSETTA-systemet (16, 17). Denne metoden lærer definisjoner i form av HVIS-SÅ-regler fra treningssettet. Disse reglene beskriver en prosess ved hjelp av tidsforløpet til genene annotert til denne prosessen. Et gen, kjent eller ukjent, klassifiseres til den prosessen som har en definisjon som ligger nærmest dette genets tidsforløp. Klassifiseringskvaliteten kan på denne måten valideres med statistiske metoder på testsett av kjente gener. Denne valideringen angir utsagnskraften i senere forutsigelser av funksjon for ukjente gener. En detaljert beskrivelse av disse fremgangsmåtene vil føre for langt, men detaljer kan finnes i Hvidsten og medarbeidere (13).

Konklusjon

Et sett med verktøy for kunnskapsoppdagelse fra genuttrykksanalyse, tilknytting av gener til biologisk funksjon og visualisering er konstruert (11, 13, 18). Disse verktøyene er nå i bruk i vårt prosjekt for utvikling av genomklassifikatorer fra mikromatrisedata. Biomedisinske forskere finner verktøyene interessante fordi bakgrunnskunnskapen ofte er vel ivaretatt. Vi har styrket vår hypotese om at kunnskapsbaserte verktøy sannsynligvis vil ha et fortrinn innen kunnskapsoppdagelse i komplekse biomedisinske sammenhenger. Selv om våre metoder er anvendt på mikromatrisedata, er disse prinsippene av generell betydning og vil få anvendelse i vurderingen av alle slags komplekse datasett, for eksempel som beslutningsstøtte i kliniske sammenhenger hvor store mengder informasjon foreligger for hver enkelt pasient.

Litteratur

1. Schena M, Shalon D, Davis R, Brown PO. Quantitative monitoring of gene expression patterns with a cDNA microarray. Science 1995; 270: 467–70.
2. Debouck C, Goodfellow PN. DNA microarrays in drug discovery and development. Nat Genet 1999; 21 (suppl 1): 48–50.
3. Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF et al. The transcriptional program in the response of human fibroblasts to serum. Science 1999; 283: 83–7.
4. Gene Ontology. http://genome-www.stanford.edu/GO/ (5.10.2000).
5. Schalkoff R. Pattern recognition – statistical, structural and neural approaches. New York: John Wiley, 1992.
6. 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: 531–7.
7. Brown MPS, Grundy WN, Cristianini N, Sugnet CW, Furey TS, Ares M et al. Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci 2000; 97: 262–7.
8. Eisen M, Spellman P, Brown P, Botstein D. Cluster analysis and display of genome-wide expression pattern. Proc Natl Acad Sci USA 1998; 95: 1464–80.
9. Kohonen T. The self-organizing map. Proceedings of the IEEE 1990; 78: 1464–80.
10. The transcriptional program in the response of human fibroblasts to serum on the WEB: http://genome-www.stanford.edu/serum/ (28.2.2000).
11. Hvidsten TR, Jenssen T-K, Komorowski J, Lægreid A, Sandvik AK, Tjeldvoll D. Template-based gene expression analysis. I: Miyano S, Shamir R, Takagi T, red. Currents in computational molecular biology – RECOMB 2000. Tokyo: Universal Academy Press, 2000: 10–1.
12. The PubGene home page: www.idi.ntnu.no/grupper/KS-grp/microarray/pubgen/genes.cgi (16.10.2000).
13. Hvidsten TR, Komorowski J, Sandvik AK, Lægreid A. Predicting gene function from gene expressions and ontologies. Pacific Symposium on Biocomputing 2001; 6: 299–310. http://psb.stanford.edu/psb01/.
14. Pawlak Z. Rough sets. International Journal of Computer and Information Sciences 1982; 11: 341–56.
15. Skowron A, Komorowski J, Pawlak Z, Polkowski L. A rough set perspective on data and knowledge. I: Klösgen W, Zytkow J, red. Handbook of data mining and knowledge discovery. Oxford: Oxford University Press, 2001: 20–41.
16. Komorowski J, Skowron A, Øhrn A. The Rosetta system. I: Klösgen W, Zytkow J, red. Handbook of data mining and knowledge discovery. Oxford: Oxford University Press, 2001: 1–9.
17. Komorowski J, Øhrn A. Modelling prognostic power of cardiac tests using rough sets. Artificial Intelligence in Medicine 1999; 15: 167–91.
18. Jenssen T-K, Lægreid A, Komorowski J, Hovig E. PubGene: Discovering and visualising gene-gene relations. I: Miyano S, Shamir R, Takagi T, red. Currents in Computational Molecular Biology – RECOMB 2000. Tokyo: Universal Academy Press, 2000: 48–49.

○