INFORMATICA MEDICA SLOVENICA

SDMI

Revija Slovenskega društva za medicinsko informatiko
Informatica Medica Slovenica
LETNIK 11, ŠTEVILKA 1
ISSN 1318-2129
ISSN 1318-2145 on line edition
http://lsd.uni-mb.si/ims

O REVIJI

Informatica Medica Slovenica je interdisciplinarna strokovna revija, ki objavlja prispevke s področja medicinske informatike, informatike v zdravstvu in zdravstveni negi, ter bioinformatike. Revija objavlja strokovne prispevke, znanstvene razprave, poročila o aplikacijah ter uvajanju informatike na področjih medicine in zdravstva, pregledne članke in poročila. Še posebej so dobrodošli prispevki, ki obravnavajo nove in aktualne teme iz naštetih področij.

Informatica Medica Slovenica je strokovna revija Slovenskega društva za medicinsko informatiko. Revija je dostopna na naslovu http://lsd.uni-mb.si/ims. Avtorji člankov naj svoje prispevke v elektronski obliki pošiljajo glavnemu uredniku po elektronski pošti na naslov janez.stare@mf.uni-lj.si. Revijo prejemajo vsi člani društva. Informacije o članstvu v društvu oziroma o naročanju na revijo so dostopne na tajništvu društva (Drago Rudel, drago.rudel@mf.uni-lj.si).

VSEBINA

Uvodnik ■

Eksperimentalna biomedicina se je v zadnjem desetletju korenito spremenila. Nepisana pravila kot je en gen – en doktorat ne držijo več. Raziskave se danes lahko osredotočajo na biološke funkcije v kateri sodelujejo večje skupine genov, proteinov in metabolitov. Z novimi, nedavno razvitimi tehnologijami lahko raziskovalci istočasno opazujejo izražanje tisočih genov, interakcije med tisočimi proteinov, in koncentracije metabolitnih produktov celotnega opazovanega biološkega sistema. Če je pri klasični biomedicini za analizo eksperimentalnih podatkov zadostoval svinčnik in papir, danes slednje ni možno brez intenzivne uporabe računalnikov. Eksperimentalna biomedicina je dandanes pravzaprav tesno povezana z informatiko in sodobnimi metodami za shranjevanje in obdelavo podatkov. Od zasnove eksperimentov do končne obdelave danes raziskovalci na področju biomedicine sodelujejo s statistiki, informatiki, računalniškimi inženirji in podatkovnimi analitiki. Disciplina, ki vse to povezuje, se imenuje bioinformatika. Korenine bioinformatike lahko poiščemo seveda veliko prej, a je področje doživelo velik razmah prav v zadnjih desetih letih. Tako tudi v Sloveniji, kjer so raziskovalci na področju bioinformatike o svojih izsledkih že pred vrsto leti poročali na raznih srečanjih in v revijah. A se je prvi dogodek, ki je bil namenjen izključno tej disciplini, zgodil prav lani, ob koncu leta 2005. Na prvem srečanju bioinformatikov, ki je potekal na Kemijskem inštitutu pod pokroviteljstvom Slovenskega društva za medicinsko informatiko in Slovenskega biokemijskega društva se je zbralo 112 raziskovalcev in strokovnjakov iz 34 raziskovalnih, akademskih in privatnih inštitucij iz vse Slovenije. Ob obilici referatov, predstavljenih na srečanju, se je ponudila ideja po ureditvi posebne številke revije Informatica Medica Slovenica na temo bioinformatike. Tako je tudi nastala pričujoča številka, katere članki obdelujejo različne teme s področja. Bralec bo prav gotovo opazil, da se večina prispevkov – z nekaj pomembnimi izjemami – ukvarja z analizo podatkov o genskih ekspresijah. Številko zato pričenjamo s splošnejšimi članki, ki opisujejo tehnologijo mikromrež in pripadajoče statistične pristope. V nadaljevanju se vrstijo članki, ki opisujejo izbrane metode za analizo ekspresijskih podatkov in uporabo teh v medicinski diagnostiki in prognostiki, proučevanju varnosti hrane in funkcijski genomiki. Sledi članek o encimski kinetiki in molekularnem modeliranju, pričujočo izdajo revije pa zaključuje splošnejši prispevek o fenomiki in sistemski biologiji.

Blaž Zupan, Gregor Anderluh, Janez Stare

■ **Infor Med Slov**: 2006; 11(1): 1

Strokovni članek ■

Tehnologija DNA mikromrež in njena uporaba v medicini

Peter Juvan, Damjana Rozman

Izvleček. Tehnologija DNA mikromrež omogoča celostne študije genoma (DNA) in transkriptoma (RNA). Na ravni transkriptoma sledimo izražanju genov (ekspresijske mikromreže), na ravni genoma pa iščemo področja v genomu, kjer je prišlo do spremembe v številu kopij DNA zaporedij (primerjalna genomska hibridizacija), ugotavljamo razlike v DNA zaporedjih (čipi za ponovno sekvencioniranje, SNP čipi), ali pa iščemo regulatorna DNA zaporedja, na katera se vežejo izbrani regulatorni proteini (tehnologija čip-čip). Priprava in uporaba mikromrež vključuje veliko bioinformacijskega dela, in sicer pri izboru nukleotidnih zaporedij, ki bodo predstavljala posamezne gene, pri načrtovanju poskusov, ter pri zajemanju, urejanju, shranjevanju in analizi pridobljenih podatkov.

DNA microarray technology and its applications in medicine

Abstract. DNA microarray technology enables large-scale genome (DNA) and transcriptome (RNA) studies. At transcriptome level we track expression levels of individual genes (expression microarrays), and at genome level we search for DNA copy number aberrations (comparative genomic hybridization), detect differences in DNA sequences (re-sequencing and SNP chips) and search for specific transcription factor binding sites (chip-on-chip technology). Microarray design and the application of that technology involves a lot of bioinformatc work, e.g. selection of nucleotide sequences representing individual genes, planning experiments, capturing, managing and storing data and their analysis.

■ **Infor Med Slov**: 2006; 11(1): 2-15

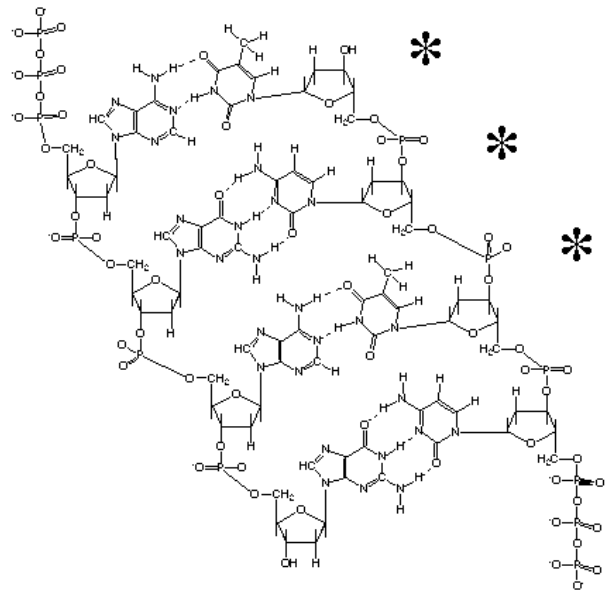
Instituciji avtorjev: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani (PJ), Center za funkcijsko genomiko in bio-čipe, Medicinska fakulteta, Univerza v Ljubljani (DR).

Kontaktna oseba: Peter Juvan, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Tržaška 25, SI-1001 Ljubljana. email: peter.juvan@fri.uni-lj.si.

Uvod

Organizem je kompleksen sistem, kjer tisoči genov in njihovih produktov (RNA in proteinov) usklajeno delujejo in ustvarjajo čudežnost življenja. Posamezen gen si lahko predstavljamo kot besedo v slovarju genoma, pristope, ki omogočajo celostno razumevanje izražanja genoma, pa kot orodja, ki nam pomagajo razumeti, kako se besede genoma povezujejo v smiselno besedilo.

Do nedavnega je v genomskih raziskavah prevladovalo pravilo "en gen – en poskus". Ta pristop podrobnega raziskovanja posameznih genov je od odkritja DNA leta 1953 pa do danes zaznamoval življenjske poti mnogih raziskovalcev, kamenčki posameznih odkritij pa so se sestavljali v mozaik skozi leta in desetletja. Vendar življenje celice ni sestavljeno iz izoliranih enot. Celica je splet signalnih, presnovnih, transportnih in drugih procesov, ki so tesno povezani med seboj in so v nenehnem stiku s svojim okoljem. Vsaka sprememba, ki jo povzročimo v celici, tudi če je ciljana na spremembo ene same podenote, na primer spremembe izražanja enega samega gena, se odraža v desetinah ali stotinah sprememb vzdolž različnih poti. Da bi čim bolje spoznali zakonitosti življenja, je smiselno uporabljati celostne (globalne) pristope, ki nam omogočajo sledenje mnogoterim lastnostim v enem samem poskusu. V zadnjih nekaj letih se je temu najbolj približala tehnologija DNA mikromrež, ki je eno od najučinkovitejših orodij za globalne študije izražanja genoma. Sprva se je pojem "DNA čip" uporabljal le takrat, ko je šlo za litografsko pripravljene čipe, s pojmom "DNA mikromreža" pa smo opisovali vse preostale tehnologije priprave. Zaradi napredka tehnike, ki z natančnimi roboti omogoča nanašanje molekul DNA z visoko gostoto, je ločnica med "čipom" in "mikromrežo" pogosto nejasna. Tako včasih govorimo o čipu tudi takrat, kadar so izbrane molekule DNA, ne glede na izvor in način priprave, nanešene na trdno podlago v visoki gostoti, na primer več tisoč skupin molekul na ploščici velikosti objektnega mikroskopskega stekelca, medtem ko mikromreža vsebuje manjše število skupin molekul DNA. Uporabljajo se še pojmi genomski čip, mreža genov in bio-čip. Slednji izraz je najbolj splošen, saj predpona "bio" opisuje, da se na biološkem čipu lahko nahajajo biomolekule katerekoli vrste – DNA, RNA, proteini in druge.



Slika 1 Podroben prikaz hibridizacije – komplementarnega parjenja dveh enojnih verig nukleinskih kislin (* označuje fluorescentno označeno verigo).

Vrste DNA čipov in mikromrež

DNA čipi so mikroskopske skupine tisočih DNA molekul z znanimi nukleotidnimi zaporedji, ki jih pritrdimo na podlago. Vsak gen je na mikromreži ali čipu zastopan z vsaj eno skupino identičnih DNA molekul. Organizirana razporeditev skupin DNA molekul predstavlja matrico, na kateri po hibridizaciji določimo razliko v izražanju genov med poskusnim vzorcem in kontrolo. Postopek hibridizacije temelji na komplementarnem parjenju baz A-T in G-C po modelu Watsona in Cricka (slika 1). Mikromrežo z organizirano razporeditvijo tarčnih DNA molekul izpostavimo fluorescentno ali radioaktivno označeni preizkusni snovi (imenovani proba ali sonda), ki jo pripravimo iz preiskovanih celic ali tkiv.

Hibridizacijski signal na določenem mestu matrice nam izraža identiteto nukleotidnega zaporedja, velikost signala pa je merilo za količino izraženega genskega produkta. Prvi DNA čipi so se uporabljali za spremljanje celostnega (globalnega) izražanja genov, vedno več pa je tudi uporabe za določanje sprememb na ravni genoma. Dandanes se DNA čipi najpogosteje uporabljajo za:

1. ugotavljanje količine izraženih genov (transkriptom ali ekspresijsko profiliranje) ter

2. določanje nukleotidnih zaporedij in sprememb na ravni genoma – sekvencioniranje, iskanje enojnih nukleotidnih polimorfizmov (SNP) in mutacij, primerjalno genomsko hibridizacijo (CGH) ter iskanje regulatornih DNA zaporedij (metoda čip-čip – kromatinska imunoprecipitacija z analizo čipov).

Obstaja več vrst DNA mikromrež, od raziskovalnega vprašanja oziroma namena uporabe pa je odvisno, katera je najprimernejša. Mikromreže s kratkimi oligonukleotidi, ki so sintetizirani na matrici in situ, so sprva imenovali DNA čipi, saj je tehnologija priprave podobna pripravi računalniških čipov. Vodilno vlogo pri pripravi teh čipov ima podjetje Affymetrix (http://www.affymetrix.com), ki je metodologijo tudi razvilo in patentiralo. Pri Affymetrixovi tehnologiji je vsak gen zastopan s preko 10 kratkimi nukleotidi, specifičnost hibridizacije posameznega nukleotida pa se določa na podlagi odsotnosti hibridizacije z nukleotidom, ki se od tarčnega razlikuje za eno samo bazo.

Obstajajo tudi mikromreže z dolgimi oligonukleotidnimi sondami, mikromreže s komplementarnimi DNA (cDNA) sondami in mikromreže z več 100 kb dolgim odseki genomske DNA. Razen pri tehnologiji Affymetrix so sonde sintetizirane po klasičnih molekularno-bioloških postopkih in so kasneje z nanašalnim robotom (angl. spotterjem) nanešene na podlago.Tehnologija dolgih oligonukleotidov (50 – 80 bp) uporablja eno do tri oligonukleotidne probe za posamezni gen. Skupina Patricka Browna s sodelavci z Univerze Stanford je razvila tako imenovane klasične DNA mikromreže, kjer so na stekleno ploščico nanešene 300 do 500 bp komplementarne DNA (cDNA), od katerih vsaka predstavlja en gen. Le-te pridobimo iz celičnih informacijskih RNA (mRNA) s pomočjo gensko-specifičnih začetnih oligonukleotidov v reakciji obratnega prepisovanja in verižnega pomnoževanja s polimerazo (RT-PCR). Tudi zelo dolge odseke genomske DNA pridobimo s klasičnimi molekularno-biološkimi postopki in jih naknadno nanesemo na podlago.

Kljub temu, da so bili Affymetrixovi DNA čipi za ekspresijsko profiliranje na tržišču prvi, pa zaradi visoke cene in zaprtosti sistema (podatki o nukleotidnem zaporedju na čipu uporabljenih oligonukleotidov niso dostopni, uporabnik ne more spreminjati genov na čipu) za mnoge raziskovalne laboratorije niso dostopni. V prihodnosti bo sicer cena komercialno dostopnih čipov in mikromrež padala in postajala dostopnejša, vendar pa je uporaba klasičnih mikromrež na principu cDNA še vedno smiselna za usmerjene študije omejenega števila genov (čipi nizke gostote), kot je npr. določanje bolezenskih markerjev izbranega, že dobro definiranega obolenja.

DNA čipi za ekspresijsko profiliranje

Za ekspresijsko profiliranje se uporabljajo tako oligonukleotidne kot cDNA mikromreže. Obstajata dve glavni smernici priprave ekspresijskih DNA mikromrež. V prvem primeru čip vsebuje gene, ki jim je skupna fizična lokacija (npr. človeški kromosom 22) ali kar vse gene določenega organizma. Na tržišču so že dostopni DNA čipi celotnega genoma kvasovke *S.cerevisiae*, mnogih bakterij, pa tudi celotnega človeškega, mišjega in podganjega genoma. Glede na sedanjo tehnologijo je na stekleno ploščico moč nanesti do 50.000 genov. Celotni človeški genom se pri čipih podjetij Affymetrix in Agilent (http://www.agilent.com) nahaja na eni stekleni ploščici. Pri drugem pristopu čip vsebuje izbrane

gene, ki so med seboj smiselno (tematsko) povezani. Načrtovanje takih čipov izhaja iz a priori biokemičnega znanja o funkciji genov in iz poznavanja fiziologije ter patofiziologije proučevanih obolenj. Tematski čipi raziskovalcem ponujajo neomejene možnosti tvorjenja in uresničevanja idej na temeljnem, aplikativnem in klinično-diagnostičnem področju. Na tržišču je vedno večje število usmerjenih čipov, ki se ukvarjajo z izražanjem genov med onkogenezo (onko-čipi) in tekom drugih pogostih bolezni. Mnogi se razvijajo v smislu diagnostičnih orodij, ki bi lahko našla pot tudi v klinično prakso.

Priprava čipov nizke gostote

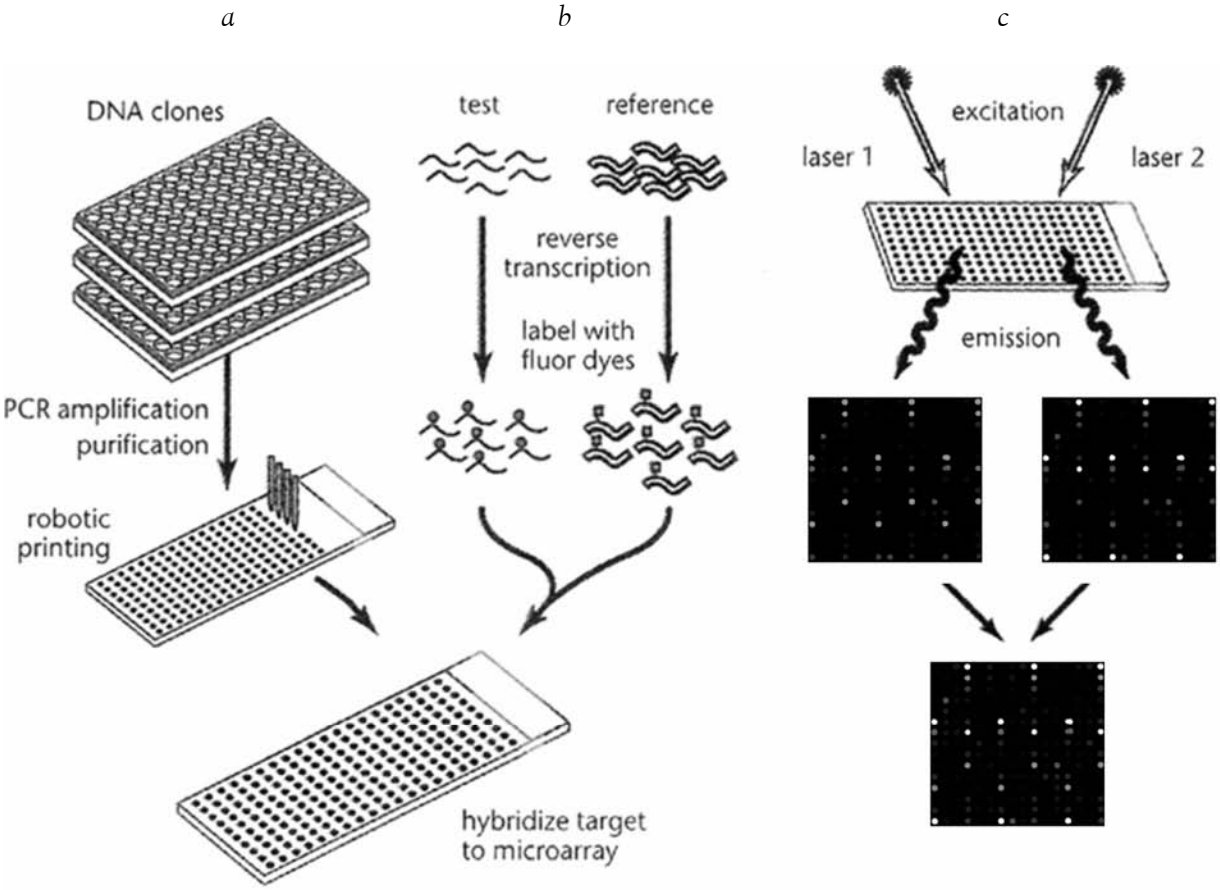
Ne glede na to, ali se odločimo za pripravo čipa, ki bo vseboval dolge oligonukleotide, ali za čip na podlagi cDNA, začetna faza načrtovanja zahteva veliko bioinformacijskega dela. Izbor in število genov sta odvisna od raziskovalnega vprašanja, ki ga želimo s tehnologijo DNA mikromrež reševati. Dostop do nukleotidnih zaporedij vseh genov, ki so bila določena v akademskih inštitucijah, je omogočen preko mnogih strežnikov, kot je npr. strežnik Nacionalnega centra za biotehnološke informacije v ZDA (NCBI, http://www.ncbi.nih.gov). Preprosteje je, če imamo dostop do katere od velikih urejenih cDNA knjižnic, ki vsebuje v klasične vektorje klonirane cDNA večjega števila (oziroma vseh) genov posameznih genomov ali tkiv.

Tudi pri izboru in pripravi nukleotidnih zaporedij, ki bodo predstavljala posamezen gen, imajo ključno vlogo orodja bioinformatike. Vsako nukleotidno zaporedje mora specifično predstavljati en sam gen in mora imeti čim manjšo homologijo z vsemi ostalimi nukleotidnimi zaporedji na čipu. Od pravilnega načrtovanja čipa in poznavanja lastnosti nukleotidnih zaporedij je odvisna končna interpretacija rezultatov. V primeru čipa na osnovi dolgih oligonukleotidov izberemo za vsak gen 2 – 3 od 50 do 80 baz dolga zaporedja, ki specifično predstavljajo en sam gen. V primeru cDNA čipov izberemo za vsak gen dva kratka oligonukleotida, s pomočjo katerih v verižni reakciji s polimerazo (PCR) pomnožimo odsek cDNA iz obratno prepisane RNA, ki je izolirana iz celice ali tkiva, kjer se izbrani gen izraža (slika 2a). Kot matrica za pomnoževanje posameznih genov so zelo primerne tudi urejene cDNA knjižnice, kjer je možno vse gene pomnožiti z dvema univerzalnima oligonukleotidoma, katerih prijemališči sta v vektorju, v katerem so cDNA klonirane, vendar je pri tem pristopu velika možnost križnih kontaminacij klonov.

Pred nanosom na trdno podlago je pri pomnoženih cDNA potrebno preveriti integriteto in koncentracijo nukleotidnih zaporedij. Za preverjanje integritete je primerna agarozna elektroforeza ali mikrokapilarna elektroforeza z elektroforeznim čipom (tehnologija Agilent Technologies). Preverjanje koncentracije in čistosti DNA poteka spektrofotometrično z merjenjem UV-absorpcije pri 260nm in 280 nm, pri čemer je za veliko število vzorcev neobhodno potreben spektrofotometer na mikrotiterske ploščice.

Po kontroli koncentracije in kakovosti DNA vzorce razredčimo z enim volumnom dvakrat koncentrirane nanašalne raztopine in jih v vnaprej določenem zaporedju razporedimo na 384-mestni mikrotiterski ploščici. Računalniški program nanašalnega robota mora prenesti motiv DNA vzorcev na mikrotiterski ploščici v nov, vnaprej izbrani motiv DNA vzorcev, kot se bodo pojavili na čipu (slika 2a). Zaradi statistične obdelave je priporočljivo, da je vsak DNA vzorec na čipu predstavljen vsaj v duplikatu. Nanašalni robot ima običajno večje število igel (tudi do 48), ki jih po želji odvzemamo ali dodajamo in s tem prilagajamo hitrost nanašanja in geometrijo čipa. Skupne lastnosti vseh nanašalnih robotov na tržišču so velika mehanska natančnost (natančnost nanašanja na 3 – 5 mikrometrov) in pripadajoča programska oprema za njihovo krmiljenje, ki praviloma omogoča, da so posamezne točke matrice medsebojno oddaljene največ 10 mikrometrov. Na tržišču je dostopnih več vrst nanašalnih robotov. Roboti lahko vsebujejo kapilare, ki posrkajo nekaj femtolitrov raztopine DNA in jo nato odložijo na trdno podlago čipa.

Prednost teh nanašalcev je majhna poraba vzorca DNA, slabost pa občutljivost kapilar, ki se zaradi majhnega premera pogosto mašijo, in možnost kontaminacije vzorcev, če sistem spiranja ni popoln. Druga vrsta nanašalnih robotov ima namesto kapilar igle, pri čemer za nanašanje raztopine DNA na trdno podlago čipa izkorišča pojav površinske napetosti kapljevin. Njihova slaba lastnost je počasnost, saj je obisk mikrotiterske ploščice z vzorci potreben pred vsakim nanosom posebej. Obstaja tudi kombinacija med kapilaro in iglo, tako imenovana razcepljena igla (angl. split pin), pri kateri količina raztopine DNA, ki se nabere v razpoki, zadostuje za večje število (tudi več kot 100) nanosov. Prednost tega sistema je velika hitrost nanašanja brez težav s spiranjem, ki so prisotne pri kapilarnem sistemu.



Slika 2 Prikaz postopka priprave in analize DNA čipa po Patrick-Brownovi tehnologiji. Faze: *a* - priprava DNA čipa vključuje sintezo tarčnih DNA molekul in njihovo pritrditev na podlago z nanašalnim robotom; *b* - postopek hibridizacije s testnim in referenčnim fluorescentno označenim vzorcem nukleinskih kislin; *c* - analiza signala z optičnim laserskim čitalcem.

Podlago čipa navadno predstavlja kemično obdelano steklo (nanos oligonukleotidov in cDNA zahteva drugačne kemično obdelane steklene ploščice), lahko pa je to tudi najlonska ali nitrocelulozna membrana. V primeru obdelanega stekla pride do kemične reakcije med nukleotidi DNA in aktivnimi skupinami na premazu stekla. DNA še dodatno pritrdimo z zamreženjem pod vplivom ultravijolične svetlobe. Sledi spiranje v več raztopinah detergenta in soli padajoče ionske jakosti ter hitro odstranjevanje vodnih kapljic z vakuumsko centrifugo. Slednje je ena izmed kritičnih stopenj postopka za pripravo kvalitetnih čipov, saj v primeru počasnega sušenja vodne

kapljice pustijo obris, ki je viden tudi po hibridizaciji in otežkoča kvantifikacijo signalov na področju motnje.

Hibridizacija in odčitavanje

Po zgoraj opisanem postopku obdelani čip je pripravljen za hibridizacijo z izbranim fluorescentno označenim preiskovanim vzorcem. Preiskovani vzorec predstavlja iz celic ali tkiv izolirana RNA, ki jo s kompleti za označevanje prevedemo v fluorescentno označeno tkivno ali celično cDNA. Fluorescentno označeno cDNA skupaj s hibridizacijsko raztopino enakomerno porazdelimo po čipu (slika 2b). Hibridizacijo lahko izvajamo ročno, v vodni kopeli, ali na avtomatski hibridizacijski postaji, kjer je možno hibridizirati do nekaj deset čipov hkrati. S kontrolne plošče lahko uravnavamo temperaturni in časovni interval za vsak čip posebej, tako med hibridizacijo kot tudi med kasnejšim spiranjem. V primerjavi z ročno hibridizacijo in spiranjem daje aparat za avtomatsko hibridizacijo bolj ponovljive rezultate hibridizacije, hkrati pa se izognemo tudi mehanskim poškodbam čipa in motnjam, ki lahko nastanejo zaradi učinka vodnih kapljic med postopkom ročnega spiranja.

Hibridizacijski signal odčitamo z večlaserskim optičnim čitalcem (angl. scanner), ki omogoča sledenje fluoroforom različnih valovnih dolžin (slika 2c). Navadno imamo na istem čipu le dva različno obarvana fluorescentna signala. Ker pa je na tržišču dostopnih vse več različnih fluoroforov, se priporoča čitalec z vsaj dvema laserjema, ki skupaj s filtri omogoča zaznavanje 4 – 6 različnih fluoroforov. Čitalec mora imeti dovolj visoko ločljivost, da zazna točke, ki so medsebojno oddaljene le nekaj mikrometrov. Rezultat odčitavanja je računalniška slika, kjer je jakost hibridizacijskega signala ponazorjena z intenziteto slikovnih pik. Na tržišču je cel spekter dvo- in tri-laserskih čitalcev, ki imajo pri enakem številu laserjev medsebojno primerljive cene.

Litografsko pripravljeni čipi visoke gostote

Zaradi tehnične zahtevnosti in težavne standardizacije v raziskovalnih laboratorijih običajno ni mogoče pripravljati litografskih čipov, so pa ti čipi dostopni komercialno. To tehnologijo je vpeljala in najdlje razvila ameriška firma Affymetrix, po kateri čipi tudi nosijo ime. Vsak gen je na Affymetriksovem čipu predstavljen z najmanj desetimi oligonukleotidi, dolgimi do 20 bp. Ker so tako kratka nukleotidna zaporedja podvržena nespecifični hibridizaciji, še posebej pri sorodnih genih, ima vsak oligonukleotid na čipu tudi kontrolni oligonukleotid, ki se od pravega razlikuje za en sam nukleotid. Po hibridizaciji in odčitavanju signala se pri določanju izražanja posameznih genov upošteva le tiste oligonukleotide, kjer je hibridizacijski signal močan pri pravi probi in odsoten pri kontrolni probi. Zaradi popolnoma drugačne tehnologije priprave čipa je tudi postopek hibridizacije s tehničnega stališča drugačen, kot je opisano v prejšnjem razdelku. Prav tako je potreben čitalec večje občutljivosti, saj so in situ sintetizirani oligonukleotidi medsebojno oddaljeni manj kot naknadno na trdno podlago nanešene molekule DNA. Potrebna je tudi posebna programska oprema (GCOS - GeneChip Operating Software), ki s pomočjo ustrezne podatkovne zbirke vsakega od kratkih oligonukleotidov poveže z odgovarjajočim genom. Kljub "zaprtosti" sistema, ki zahteva popolnoma nov sklop opreme (hibridizacijska komora, laserski čitalec in programska oprema), zaradi velike standardizacije uporabniku zagotavlja visoko raven ponovljivosti meritev in možnost študije izražanja genov na ravni celotnih genomov (človeški genom je trenutno z več kot 47.000 transkripti dostopen na enem Affymetriksovem čipu). Visoke cene teh čipov (med 500 in 1000 EUR/čip) zaenkrat onemogočajo rutinsko uporabo malim uporabnikom in dajejo prednost dostop velikim industrijskim in kliničnim partnerjem. S stališča manjših uporabnikov so komercialni čipi visoke gostote pomembni predvsem kot prva stopnja raziskav, ki se bodo kasneje osredotočile na omejeno število genov z uporabo mnogo cenejših DNA čipov nizke gostote. Litografski čipi visoke

gostote so perspektivni tudi za uporabo v farmacevtski industriji, predvsem pri testiranju zdravilnih učinkovin na izražanje celotnega človeškega genoma.

DNA čipi za določanje nukleotidnih zaporedij in sprememb na ravni genoma

Za drugo vrsto DNA mikromrež je značilno, da ne spremljamo ravni izražanja genov (raven mRNA), temveč določamo nukleotidno zaporedje na izbranih delih genoma (raven DNA). Za namen sekvencioniranja se razvijajo "prekrivajoče" mikromreže (angl. tiling microarrays), kjer so kromosomi po celotni dolžini pokriti s prekrivajočimi se 20 bp dolgimi oligonukleotidi. S tovrstnimi čipi je možno določiti genotip posameznika, tako nukleotidno zaporedje normalnih alelov kot tudi okvarjenih bolezenskih alelov. Mikromreže za sekvencioniranje celotnih genomov nekateri preprostejših organizmov (npr. različni soji kvasovk *S.cerevisiae*) vsebujejo prekrivajoča se zaporedja celotnega genoma kvasovke.[1] Pri pripravi takih mrež je zelo pomembna bioinformatika, saj je potrebno zadostiti predpostavki, da lahko vsako mesto v genomu teoretično vsebuje katerokoli od štirih baz: A,C,G ali T.

Podoben princip (vsako mesto v genomu lahko vsebuje katerokoli od štirih baz) velja tudi pri iskanju enojnih nukleotidnih polimorfizmov (SNP) in/ali mutacij. Razlika z mikromrežami za sekvencioniranje je v obsegu DNA zaporedja na čipu. Mikromreže za sekvencioniranje morajo vsebovati celotno področje genoma, katerega nukleotidno zaporedje želimo preveriti, SNP čipi pa so lahko omejeni npr. na posamezni polimorfni gen, ki je odgovoren za določeno obolenje. Posamezniki se namreč v svojem genskem zapisu med seboj razlikujemo tudi po polimorfnih nukleotidnih zaporedjih, med katere spadajo tudi SNP. Analiza DNA iz različnih osebkov bo na mikromreži pokazala prisotnost ali odsotnost signala, ki potrjuje prisotnost oziroma odsotnost določenega SNP, z boleznijo povezanega. Tako lahko s čipom, ki vsebuje vse doslej poznane mutacije gena za cistično fibrozo (eno izmed najpogostejših monogenskih obolenj), v enem samem poskusu ugotovimo, katere mutacije tega gena so prisotne pri preiskovanem osebku. SNP čipi že nadomeščajo dolgotrajnejše metode, kjer se je vsaka mutacija preiskovala ločeno.

Primerjalna genomska hibridizacija (CGH) na čipu je visoko ločljivostna metoda za preiskovanje kvantitativnih sprememb na ravni genoma.[2] Kvantitativne spremembe vključujejo delecije (primanjkljaj oziroma izrez določenega odseka genoma), insercije (vključitev nukleotidnih zaporedij v genom), amplifikacije (pomnožitve določenega odseka DNA) in kromosomske prerazporeditve, kjer se deli dednine med posameznimi kromosomi nepravilno izmenjajo. Klasične citogenetske metode, s katerimi so preiskovali te spremembe (kariotipizacija, proganje kromosomov, fluorescentna in situ hibridizacija – FISH, idr.) imajo pomanjkljivo ločljivost in zahtevajo deleče se celice. Na CGH čipu so nanešeni klonirani poljubno dolgi (tudi do nekaj 100 kb) odseki DNA točno določene lokacije na kromosomu. To omogoča zelo natančno določevanje sprememb in njihovo preslikavo na določeno nukleotidno zaporedje. Resolucija CGH čipov je odvisna od razdalje med posameznimi področji DNA na kromosomu. Nekatera podjetja že razvijajo prekrivajoče se CGH čipe, ki lahko odčitajo spremembe na 1 bp natančno. Metoda se v največji meri uporablja v onkologiji, in sicer v diagnostične namene (prepoznavanje vrste raka) in s tem povezano prognozo, uporabna pa je tudi za določevanje kromosomskih aberacij različnih genetskih obolenj in za prenatalno diagnostiko različnih kromosomskih abnormalnosti, npr. Downovega sindroma (trisomija kromosoma 21). Analiza DNA različnih osebkov na mikromreži pokaže prisotnost ali odsotnost signala, kar potrjuje prisotnost oziroma odsotnost določene kromosomske spremembe, povezane z boleznijo.

Mnoga obolenja nastanejo tudi zaradi napačnega uravnavanja izražanja genov, ki nastane zaradi napačne vezave regulatornih proteinov

(transkripcijskih faktorjev) na regulatorne odseke genov (promotorje). Obstaja mnogo klasičnih metod za iskanje interakcij med regulatornimi odseki DNA in proteini, v zadnjih letih pa je v porastu metoda čip-čip. Gre za oligonukleotidne čipe, ki lahko vsebujejo zaporedja celotnega genoma (*S.cerevisiae*), pri kompleksnejših organizmih pa vsebujejo le posamezne dele genoma (posamezni kromosom ali neprevedene – regulatorne regije kromosoma). Metoda se zaenkrat uporablja bolj v raziskovalne namene, saj ima tehnične omejitve pri pripravi vzorca za procesiranje na čipu. Za analizo moramo proteine, ki so v celici vezani na regulatorne odseke DNA, trajno pritrditi na DNA, nato pa to DNA izločiti iz zmesi in jo analizirati na čipu. Prvega dela postopka ne moremo izvajati na celem organizmu, ampak le na izoliranih celicah.[3]

Vloga informatike pri tehnologiji DNA mikromrež

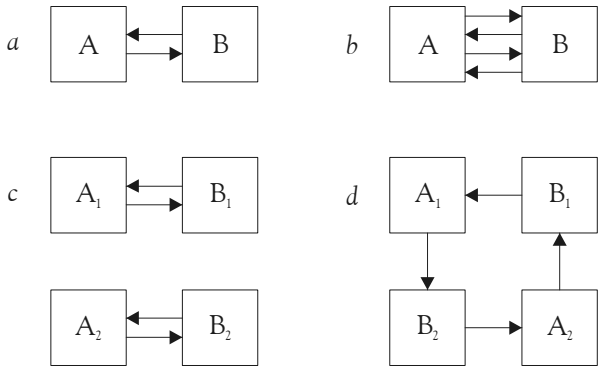
Tehnologija DNA mikromrež je prinesla nove izzive tudi na področju računalniških znanosti in statistike. Prednost tehnologije DNA mikromrež ni v samem načinu merjenja izraženosti genov, pač pa v zmožnosti merjenja več tisoč genov naenkrat. Preskok v obsegu meritev iz posameznih genov na več tisoč genov je zahteval razvoj novih standardov in metod za upravljanje, analizo in vizualizacijo podatkov ter njihovo algoritmično implementacijo na računalnikih, zaradi česar je področje analize podatkov DNA mikromrež v zadnjih letih postalo eno izmed najbolj obetavnih in donosnih področij za raziskovalce s področja računalniških znanosti in statistike.

Razvoj tehnologija DNA mikromrež je v veliki meri prehitel razvoj ustreznih analitičnih metod. V praksi pogosto prihaja do primera, da se podatki analizirajo naknadno, torej v času, ko so meritve že opravljene oziroma poskusa ni več možno spreminjati. Pri analizi se pogosto pokažejo pomanjkljivosti, ki so plod neoptimalne zasnove poskusa. Zato je sodelovanje biologov in analitikov pomembno že v času načrtovanja poskusa. Za uspešno izvedbo biološkega poskusa je pomembno, da analitik dobro razume njegov namen, da sodeluje tako pri njegovi zasnovi kot tudi pri analizi podatkov in ovrednotenju rezultatov, in da se biolog zaveda nevarnosti in omejitev povezanih s tehnologijo DNA mikromrež.

Področje bioinformatike, ki se najširšem smislu nanaša na zajemanje, urejanje, shranjevanje in analizo raznovrstnih bioloških podatkov, zahteva od raziskovalcev hkratno poznavanje metod bioloških, računalniških in statističnih znanosti ter spretnost pri uporabi računalnika kot eksperimentalnega orodja. Čeprav stranskega pomena v primerjavi z analizo, urejanje, poimenovanje in shranjevanje podatkov trenutno zavzema večino časa bioinformatikov, saj je sistematičnost na tem področju predpogoj za uspešno aplikacijo analitičnih metod. Mnogo objav s področja bioinformatike predstavlja nove oziroma izboljšane analitične metode in algoritme, zlasti s področja strojnega učenja, prilagojene za analizo podatkov DNA mikromrež. Pogosto se zastavlja vprašanje, katera izmed teh metoda je trenutno "najboljša". Zavedati se moramo, da izbira analitične metode ni tako pomembna kot dobra zasnova biološkega poskusa, saj nobena metoda za analizo ne more odtehtati pomanjkljivosti pri njegovi zasnovi.

Načrtovanje poskusa

Zaradi visokih stroškov, ki so povezani z uporabo tehnologije DNA mikromrež, se načrtovanje poskusa v nasprotju s tipičnim biomedicinskim poskusom pogosto prične s stališča denarnih sredstev. Osnovno vodilo na ta način zasnovanega poskusa ni testiranje določene hipoteze, pač pa tvorjenje hipotez oziroma upanje, da pridemo do smernic za nadaljnje raziskave. Tak pristop je zgrešen že v osnovi. Poskus mora biti osnovan na biološkem vprašanju, na katerega želimo odgovoriti, vprašanje pa mora biti osredotočeno in zastavljeno dovolj ozko, da lahko nanj odgovorimo v okviru denarnih sredstev, ki so nam na voljo.



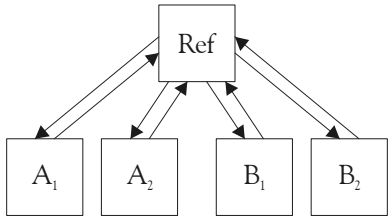
Slika 3 Različni načrti poskusa z dvobarvnimi DNA mikromrežami, ki vključujejo dva tretmaja (A in B). Načrta *a* in *b* predvidevata dve oziroma štiri tehnične ponovitve z zamenjavo barv (angl. dye swap); *c* in *d* sta osnovana na dveh neodvisnih bioloških ponovitvah tretmajev (ponazorjeno z indeksom pri oznaki tretmaja); *c* predvideva biološko ponovitev načrta *a*; *d* prikazuje enostaven krožni načrt.

Pri načrtovanju poskusa moramo upoštevati različne komponente variance, ki je prisotna na posameznih korakih poskusa. V grobem jo lahko razdelimo na biološko varianco, ki je prisotna zaradi genetskih razlik med organizmi in vpliva okolja, tehnično varianco, do katere pride med postopkom izolacije, označevanja in hibridizacije, ter meritveno napako laserskega čitalca. Posamezne komponente variance zajamemo z ustreznimi ponovitvami, in sicer s ponovljenimi biološkimi vzorci, tehničnimi ponovitvami poskusov, in reproduciranimi probami znotraj posameznih čipov. Statistične teste lahko osnujemo na podlagi katerekoli izmed zajetih komponent variance, razlika je le v interpretaciji njihovih rezultatov. Če nas zanima učinek določenega tretmaja, moramo teste osnovati na biološki varianci. Statistični testi na podlagi tehnične variance nam pokažejo razlike znotraj posameznih skupin. Prispevek posamezne komponente variance lahko ocenimo s korelacijo; pri ponovitvah znotraj posameznega čipa je ponavadi višja od 95%, pri tehničnih ponovitvah pade tipično na 60 do 80%, pri bioloških ponovitvah pa je lahko samo še 30%.[4] Ponovljivost poskusa bo veliko višja, če se izognemo biološkim ponovitvam in opravimo samo tehnične, vendar so rezultati takega poskusa varljivi – statistično značilne razlike med skupinami lahko odražajo zgolj naključne razlike med posameznimi osebki.

Ustreznost poskusa lahko na preprost način ocenimo s številom prostostnih stopenj (angl. degrees of freedom), ki ustreza številu neodvisnih enot poskusa, od katerega odštejemo število različnih tretmajev. Neodvisne enote poskusa ustrezajo osebkom, ki bi teoretično lahko bili podvrženi poljubnemu tretmaju, in katerih vzorci so bili tretirani neodvisno drug od drugega v vseh fazah poskusa. Če pridemo do števila 5 ali več, smo na dobri poti. V določenih okoliščinah lahko število osebkov presega maksimalno število hibridizacij, ki jih lahko opravimo, ali pa količina RNA posameznega osebka ne zadošča za predvideno število hibridizacij. V takih primerih lahko osebke združimo v skupine (angl. pools), pri čemer skupine predstavljajo nove neodvisne enote poskusa. Z zmanjšanjem števila neodvisnih enot poskusa se zmanjša komponenta biološke variance, ne pa tudi tehnična komponenta, kar poveča verjetnost napačnih zaključkov statističnih testov. V splošnem je bolje narediti večje število manjših skupin oziroma več bioloških in manj tehničnih ponovitev.

Pomemben korak pri načrtovanju poskusa je določiti število tehničnih ponovitev, kar je pri dvobarvnih čipih tesno povezano z odločitvijo, katere pare vzorcev bomo hibridizirali na istem čipu. Do učinkovite zasnove poskusa lahko pridemo z upoštevanjem nekaj enostavnih pravil.[5] Načrt predstavimo z usmerjenim grafom, kjer vozlišča predstavljajo biološke vzorce (neodvisne enote poskusa), usmerjene povezave pa hibridizacije, pri katerih vzorec na začetku povezave obarvamo z rdečim (Cy5) in vzorec na koncu povezave z zelenim (Cy3) barvilom. Slika 3 prikazuje različne zasnove poskusa za neposredno primerjavo vzorcev A in B. Načrta *a* in *b* predvidevata samo tehnične ponovitve (na podlagi zamenjave barv), *c* in *d* pa tudi biološke. Učinkovitost posamezne primerjave je odvisna od dolžine in števila poti med primerjanima vzorcema.[6] Najbolj učinkovito je skupaj hibridizirati tiste vzorce, katerih primerjava je najbolj zanimiva. Primerjati je možno tudi vzorce,

ki niso hibridizirani skupaj, pod pogojem da med njimi obstaja pot v grafu. Načrti *b*, *c* in *d* na sliki 3 predvidevajo enako število hibridizacij (štiri), med seboj pa se razlikujejo v učinkovitosti posameznih primerjav. Krožna zasnova (načrt *d*) je učinkovita predvsem pri manjšem številu vzorcev, vendar se njena učinkovitost močno zmanjša v primeru neuspeha katerekoli od hibridizacij. Robustnost krožne zasnove lahko povečamo s prepletanjem. Potencialno pristranost primerjav minimiziramo z uravnoteženim načrtom, kjer iz vsakega vzorca naredimo sodo število tehničnih ponovitev, od katerih polovico označimo z enim, polovico pa z drugim barvilom.



Slika 4 Načrt poskusa z dvobarvnimi DNA mikromrežami, kjer vzorce (A_1 ... B_2) primerjamo preko skupne reference (Ref).

Vzorce lahko med seboj primerjamo tudi posredno preko skupne reference (slika 4). Večina današnjih poskusov je zasnovanih na ta način, saj ima kljub očitni neučinkovitosti (kar polovica meritev se nanaša na skupno referenco) mnogo prednosti, kot so npr. enaka učinkovitost vseh primerjav, možnost razširitve z novimi vzorci in manjša možnosti napake pri laboratorijskem delu. Pri uporabi skupne reference je pomembna izbira reference, ki bo "prižgala" vse probe na naši mikromreži, in je hkrati homogena, stabilna in prisotna v zadostni količini. Odločimo se lahko med nakupom univerzalne reference in lastno referenco, ki jo ustvarimo z združitvijo RNA iz vseh vzorcev, ki jih bomo analizirali. Prednost slednje je predvsem v količini RNA, ki je podobna kot pri naših vzorcih.

Dober načrt mora upoštevati dejstvo, da je za veljavnost statističnih testov potrebno pri vseh korakih elemente poskusa izbirati naključno. Najbolj pomembna je naključna izbira osebkov in tretmajev. Če na to izbiro ne moremo vplivati, moramo zagotoviti dovolj velik vzorec, da je raznolikost populacije dobro predstavljena. Naključna mora biti tudi izbira vzorcev za hibridizacije, ki jih bomo opravili v istem dnevu, in izbira barvila pri tehničnih ponovitvah poskusa. Čipi so ponavadi tiskani v serijah, ki se med seboj lahko močno razlikujejo v kakovosti. Pogosto se razlike v kakovosti pojavijo tudi glede na položaj čipa pri tiskanju in celo glede na pozicijo probe na čipu.[7] Zato je pomembno, da tudi čipe izbiramo naključno, in da so probe na čipu razporejene v naključnem vrstnem redu (idealno bi bilo, če bi bila razporeditev prob na vsakem čipu drugačna, vendar bi to močno otežilo njihovo izdelavo in analizo).

Shranjevanje podatkov, standardi in ontologije

Podatki, ki jih pridobimo z uporabo DNA mikromrež, so v veliki meri odvisni od pogojev, pri katerih je bil izveden poskus. Za njihovo pravilno interpretacijo moramo poznati podrobnosti izdelave čipa (pozicije in zaporedja transkriptov), pripravo vzorcev in z njimi povezane tretmaje, korake pri izvedbi poskusa, nastavitve laserskega čitalca in postopek normalizacije in transformacije podatkov. Velika količina podatkov in njihova raznolikost, ki je povezana z uporabo tehnologije DNA mikromrež, je zahtevala razvoj standardov za njihovo upravljanje, shranjevanje in izmenjavo. Pobuda je na tem področju prišla od mednarodne organizacije biologov, računalnikarjev in analitikov, imenovane Microarray Gene Expression Data (MGED) Society (http://www.mged.org). Organizacija združuje šest projektov s področja standardizacije, od katerih so najpomembnejši trije:

- Minimum Information About a Microarray Experiment (MIAME) je pobuda za standardizacijo opisa poskusa z uporabo tehnologije DNA mikromrež;[8]
- MicroArray and Gene Expression (MAGE) združuje podatkovni model (MAGE-OM) in

jezik (MAGE-ML) za predstavitev podatkov DNA mikromrež;[9]

- MGED Ontology (MO) je ontologija za označevaje poskusov z DNA mikromrežami, ki predpisuje termine z vidika načrtovanja poskusa, zgradbe čipa, priprave vzorcev, hibridizacijskih protokolov in analize podatkov.[10]

Standardi so pomembni za dvig kakovosti podatkov, izmenjavo in pravilno interpretacijo podatkov, primerjavo podatkov različnih raziskovalnih skupin in različnih bioloških sistemov ter možnost reprodukcije poskusov. Standard MAGE omogoča izmenjavo podatkov med različnimi sistemi in njihovo medsebojno primerjavo. Tehtnost primerjave je odvisna od poznavanja parametrov poskusa, zaradi česar je pomembna standardizacija minimalnega nabora parametrov za opis poskusa oziroma standard MIAME. Računalniška avtomatizacija izmenjave podatkov in njihova medsebojna primerjava je možna le ob uporabi standardnega nabora terminov oziroma ob uporabi ontologij, pri katerih poleg terminov definiramo tudi logične opise uporabljenih terminov in njihove medsebojne relacije. Namen ontologija MO ni združiti termine s tako širokega področja, kot ga pokriva tehnologija DNA mikromrež, pač pa določiti ogrodje za uporabo terminov iz obstoječih ontologij. Ontologija MO vključuje mnogo terminov, ki se ne nanašajo izključno na tehnologijo DNA mikromrež (npr. termini za opis vzorcev, načrta poskusa ipd.), pač pa jih je možno aplicirati tudi na druge tehnologije funkcijske genomike (masno spektrometrijo, in situ hibridizacijo), zaradi česar je v prihodnosti predvidena njena razširitev z ontologijo FuGO (Functional Genomics Investigation Ontology), ki bo omogočala skladno označevanje poskusov s področja funkcijske genomike neodvisno od uporabljene tehnologije.

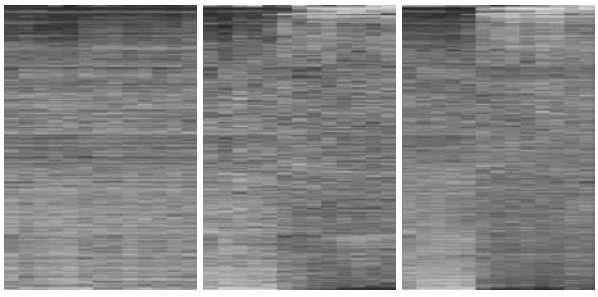
Standardi ne predpisujejo strukture podatkovnih zbirk, zaradi česar so različne raziskovalne skupine glede na lastne potrebe in omejitve razvile zbirke, ki se med seboj močno razlikujejo tako po strukturi kot tudi po funkcionalnosti. Gardiner-Garden in soavtorji[11] so pripravili njihov pregled in primerjavo, Anderle in soavtorji[12] pa so opisali potek tipičnega poskusa z DNA mrežami s stališča upravljanja s podatki in njihovega shranjevanja. Upoštevanje priporočil MGED in javna dostopnost podatkov je danes pri večini vplivnejših revij pogoj za objavo prispevka s področja DNA mikromrež. Javna podatkovna skladišča, med katerimi so največja evropski ArrayExpress Evropskega instituta za bioinformatiko (EBI),[13] ameriški GEO Nacionalnega centra za biotehnološke informacije (NCBI)[14] in japonski CIBEX Nacionalnega instituta za genetiko (NIG),[15] odpirajo možnost za nove ideje in primerjave, ki sicer ne bi bile možne znotraj posameznih institucij.

Analiza in odkrivanje zakonitosti v podatkih DNA mikromrež

Na prvi pogled se zdi, da lahko v namen analize podatkov DNA mikromrež uporabimo standardne statistične prijeme za določitev relacij med spremenljivkami, kjer posamezne probe (geni) na čipu predstavljajo neodvisne spremenljivke. Kljub temu je dandanes večina analiz opravljenih s pomočjo metod strojnega učenja. Glavni razlog je v naravi podatkov. Standardne statistične metode so prilagojene za podatke, kjer je število meritev vsaj za red velikosti večje od števila spremenljivk. Pri podatkih DNA mikromrež je slika ravno obrnjena: zaradi visoke cene meritev ali pomanjkanja primernih bioloških vzorcev imamo običajno opravka z relativno majhnim številom meritev (tipično med 10 in 100) in velikim številom spremenljivk (tipično reda velikosti od 1.000 do 10.000). Področje strojnega učenja je uspešno prodrlo na področje analize podatkov DNA mikromrež zaradi svojih bogatih izkušenj z analizo slabo determiniranih sistemov visokih dimenzij (npr. s področja avtomatskega razpoznavanja človeških obrazov). V resnici tudi na področju statistike poteka mnogo raziskav s področja analize takih sistemov, le da do nedavnega žal niso prodrle na področje analize podatkov DNA mikromrež. Podobno kot na drugih področjih znanstvenega raziskovanja je tudi

pri analizi podatkov DNA mikromrež za verodostojnost zaključkov namreč potrebna ocena njihove statistične značilnosti.

Kot primer praktične uporabe tehnologije DNA mikromrež naj povzamemo študijo s področja funkcijske genomike, katere glavni cilj je določiti biološke funkcije genom, skupinam genov in predvsem interakcijam med geni. Zmožnost hkratnega merjenja izraženosti več tisoč genov, ki jo omogoča tehnologija DNA mikromrež, ne ponuja samo mehanizma za tvorjenje hipotez oziroma obetavnih smernic za nadaljnje raziskave, pač pa ob ustrezni aplikaciji tudi orodje za potrjevanje njihove veljavnosti. Ekspresijsko profiliranje organizma D.discoideum[16] je pokazalo, da je rekonstrukcija regulatorne poti, ki uravnava prehod med rastjo in razvojem tega organizma, možna zgolj s pomočjo tehnologije DNA mikromrež in brez uporabe biološkega predznanja oziroma pristranosti. Slika 5 prikazuje ekspresijske profile dveh enojnih in ustrezne dvojne mutacije tega organizma, ki služijo kot generičen fenotip za določitev zaporedja vplivov mutiranih genov. Na podlagi podobnosti med fenotipi lahko s pomočjo klasične analize epistaze[17] in ocene statistične značilnosti[18] tudi na ravni transkriptoma potrdimo hipotezo, da gen *yakA* vpliva na gen *pufA* in ne obratno.



yakA$^-$ *pufA*$^-$ *yakA*$^-$*pufA*$^-$

Slika 5 Ekspresijski profili enojnih mutantov *yakA*$^-$ in *pufA*$^-$ ter dvojnega mutanta *yakA*$^-$*pufA*$^-$ organizma *D.discoideum*.

Uporaba in obeti tehnologije DNA mikromrež na področju medicine in farmakogenomike

Tehnologija DNA čipov in mikromrež ima veliko uporabnost pri odkrivanju genov, vključenih v bolezenske fenotipe, diagnosticiranju obolenj in odkrivanju novih učinkovin na področju farmakogenomike in toksikologije. Mnoga genetska obolenja so poligenska, kar pomeni, da ne nastanejo zaradi napake v enem samem genu, temveč je v razvoj bolezni vključenih več okvarjenih genov. Pri mnogih od teh obolenj je do sedaj poznan le en glavni "gen krivec", tehnologija DNA mikromrež pa omogočajo odkritje preostalih udeleženih genov.

Na področju medicine ponuja tehnologija DNA mikromrež možnost za določitev osnovnih vzrokov že znanih in za odkritje vzrokov še neznanih obolenj, nove strategije za iskanje t.i. bolezenskih markerjev in razvoj novih diagnostičnih orodij, s tem pa tudi izboljšano možnost za ustrezno preventivo in izbiro zdravljenja. Celostne študije izražanja genov z ekspresijskim profiliranjem močno spreminjajo naše vedenje o obolenjih in njihovih kompleksnostih. Vsak sprememba v celici povzroči niz sprememb, katerim je bilo pred tehnologijo mikromrež praktično nemogoče slediti.

V onkologiji mikromreže obetajo možnost natančnejše diagnoze in prognoze rakastih obolenj. Golub in soavtorji[19] so prvi pokazali, da lahko zgolj na podlagi profilov izražanja genov ločimo med različnimi vrstami raka in tako avtomatsko, brez biološkega predznanja, identificiramo nove vrste raka, hkrati pa tudi napovemo uspešnost kemoterapije. Mejnik za prenos tehnologije DNA mikromrež v klinično prakso predstavlja tri leta kasneje objavljena strategija za terapijo raka dojke, oblikovana po meri pacienta.[20] Na tržišču se že pojavljajo čipi nizke gostote s področja onkologije, s katerimi bi bilo moč ugotavljati vrste tumorjev. Številne raziskave v smeri oblikovanja diagnostičnih čipov za sledenje pogostih obolenj pri človeku so v polnem razmahu v različnih laboratorijih po svetu. Problem hitrega prenosa

diagnostičnih čipov v širšo klinično prakso poleg cene predstavljata še standardizacija postopkov in aparatur za odčitavanje signala.

DNA mikromreže so nepogrešljive tudi v farmakogenomiki, vedi, ki proučuje, kako dedni zapis (genetski faktorji) posameznika vpliva na odziv organizma na zdravila. Je nekakšen hibrid med farmakologijo in funkcijsko genomiko. Farmakogenomika vključuje razvijanje novih zdravil, ki bodo ciljala le eno tarčo in tako zmanjšala verjetnost nezaželenih stranskih učinkov. Mikromreže omogočajo prikaz celostnega učinka učinkovine na genom in s tem določitev zaželenih (zdravilni učinek) in nezaželenih tarč (stranski učinek).

Zelo hitro si v klinično prakso utirajo pot čipi za določanje zaporedja in sprememb v genomu, predvsem SNP čipi. Pogosto se srečujemo z vprašanjem, zakaj določeno zdravilo pri vseh pacientih ne deluje enako in zakaj so nekatera zdravila lahko za določeno skupino ljudi celo toksična. Vzrok je v enojnih nukleotidnih polimorfizmih (SNP), miniaturnih spremembah med posameznimi genomi, ki naredijo vsak osebek genetsko edinstven. Eden izmed ciljev farmakogenomike je poiskati povezave med terapevtskimi odzivi na učinkovino in genetskim profilom posameznika. S pomočjo tehnologije mikromrež lahko določimo, ali bo osebek npr. hitro ali počasi presnavljal določena zdravilo. To obeta, da bo nekoč moč doseči osebno terapijo, kjer bodo vrste in doze zdravil prilagojene genetskemu zapisu posameznika. Okolje, prehrana, starost, način življenja, splošno zdravstveno stanje ipd. sicer vplivajo na odgovor posameznika na zdravila, vendar je ključ do osebne terapije poznavanje razlik v zapisu genov, ki so odgovorni za presnovo zdravil.

SNP so tako postali eden izmed najpomembnejših tarč medicinskih in farmakogenomskih raziskav, saj lahko razkrijejo poti do novih tarč zdravilnih učinkovin. SNP in ostale diagnostične čipe pospešeno razvija firma Roche (http://www.roche.com/home.html). Zelo uporabni so SNP čipi, ki vsebujejo različice genov naddružine citokromov P450, vključenih v presnovo zdravil. Z njimi lahko napovemo, ali bo posameznik hitro ali počasi presnavljal izbrano zdravilno učinkovino. Podatki o mutacijah v genih, ki so odgovorni za presnovo zdravil, so neprecenljivega pomena za kliniko, ki se dnevno srečuje s problemom različne učinkovitosti zdravil pri različnih posameznikih. Navzkrižno učinkovanje zdravil je za posameznika škodljivo in ima lahko usodne posledice.

V klinično prakso si utirajo pot tudi čipi za primerjalno genomsko hibridizacijo (CGH). Njihova uporabnost je usmerjena v detekcijo genomskih anomalij pri genetskih boleznih (kot so Downov sindrom, sindrom Prader-Willi in Angelman sindrom) in raku. Zaradi visoke robustnosti, občutljivosti in hitrosti metode lahko kmalu pričakujemo razvoj komercialno dostopnih diagnostičnih orodij in njihovo rutinsko uporabo v diagnostične namene.

Stanje tehnologije DNA čipov v Sloveniji

Glede na naraščajoč interes in potrebe je bil junija 2001 ustanovljen Slovenski konzorcij za bio-čipe, h kateremu so pristopile slovenske akademske ustanove, klinične ustanove in farmacevtska industrija. Oprema za pripravo in analizo bio-čipov nizke gostote, kot tudi oprema za hibridizacijo čipov visoke gostote (tehnologija Affymetrix), ki je v lasti Slovenskega konzorcija za bio-čipe, je sedaj dostopna na Centru za funkcijsko genomiko in bio-čipe (CFGBC) na Medicinski fakulteti Univerze v Ljubljani (http://cfgbc.mf.uni-lj.si). Center je bil ustanovljen junija 2005 z namenom združiti infrastrukturo kot tudi interdisciplinarni kader (biološka in matematična znanja) za uporabo in kvaliteten razvoj tehnologije bio-čipov v Sloveniji.

V CFGBC smo slovenski raziskovalci Medicinske fakultete Univerze v Ljubljani, skupaj z raziskovalci Lek, d.d. in tujimi partnerji v okviru EU projekta STEROLTALK pripravili tematski cDNA čip Steroltalk, ki je usmerjen v študije homeostaze holesterola in presnove zdravil. Čip je

pripravljen v dveh različicah, in sicer za študije izražanja genov pri laboratorijski miški in pri človeku. Vsaka različica vsebuje 300 prob dolžine 300 do 500 bp, natisnjenih na stekleno ploščico. V prihodnosti naj bi se razvijal tudi v smeri diagnostike bolezni srca in ožilja.

Literatura

1. Gresham D, Ruderfer DM, Pratt SC, et al.: Genome-wide detection of polymorphisms at nucleotide resolution with a single DNA microarray. Science 2006; 311(5769): 1932-1936.
2. Oostlander AE, Meijer GA, Ylstra B: Microarray-based comparative genomic hybridization and its applications in human genetics. Clin Genet 2004; 66(6): 488-495.
3. Buck MJ, Lieb JD: ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics 2004; 83(3): 349-360.
4. Churchill GA: Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002; 32 Suppl: 490-495.
5. Kerr MK, Churchill GA: Experimental design for gene expression microarrays. Biostatistics 2001; 2(2): 183-201.
6. Yang YH, Speed T: Design issues for cDNA microarray experiments. Nat Rev Genet 2002; 3(8): 579-588.
7. Lee ML, Kuo FC, Whitmore GA, et al.: Importance of replication in microarray gene expression studies: statistical methods and evidence from repetitive cDNA hybridizations. Proc Natl Acad Sci U S A 2000; 97(18): 9834-9839.
8. Brazma A, Hingamp P, Quackenbush J, et al.: Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. Nat Genet 2001; 29(4): 365-371.
9. Spellman PT, Miller M, Stewart J, et al.: Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biol 2002; 3(9).
10. Whetzel PL, Parkinson H, Causton HC, et al.: The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics 2006; 22(7): 866-873.
11. Gardiner-Garden M, Littlejohn TG: A comparison of microarray databases. Brief Bioinform 2001; 2(2): 143-158.
12. Anderle P, Duval M, Draghici S, et al.: Gene expression databases and data mining. Biotechniques 2003; Suppl: 36-44.
13. Parkinson H, Sarkans U, Shojatalab M, et al.: ArrayExpress--a public repository for microarray gene expression data at the EBI. Nucleic Acids Res 2005; 33(Database issue): D553-555.
14. Barrett T, Suzek TO, Troup DB, et al.: NCBI GEO: mining millions of expression profiles--database and tools. Nucleic Acids Res 2005; 33(Database issue): D562-566.
15. Ikeo K, Ishi-i J, Tamura T, et al.: CIBEX: center for information biology gene expression database. C R Biol 2003; 326(10-11): 1079-1082.
16. Van Driessche N, Demšar J, Booth EO, et al.: Epistasis analysis with global transcriptional phenotypes. Nat Genet 2005; 37(5): 471-477.
17. Avery L, Wasserman S: Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet 1992; 8(9): 312-316.
18. Juvan P: Artificial intelligence methods for discovery of relationships in genetic data. PhD thesis. Faculty of Computer and Information Science, University of Ljubljana. Ljubljana, Slovenia, 2005.
19. Golub TR, Slonim DK, Tamayo P, et al.: Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999; 286(5439): 531-537.
20. van 't Veer LJ, Dai H, van de Vijver MJ, et al.: Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002; 415(6871): 530-536.

Technical Paper ■

Some observations on experimental design of microarray experiments

Lara Lusa

Abstract. Gene-expression microarrays measure simultaneously the expression of thousands of genes and are nowadays widely used in genomic research. The aim of this paper is to give a brief overview of the objectives of microarray gene-expression experiments and to describe some statistical issues related to study design and data preprocessing. Quality control, normalization, replication, validation and use of pooling of independent samples will be discussed.

■ **Infor Med Slov**: 2006; 11(1): 16-24

Author's institution: Istituto Nazionale per lo Sudio e la Cura dei Tumori, Milano and Fondazione Istituto FIRC di Oncologia Molecolare (IFOM).

Contact person: Lara Lusa, Department of Experimental Oncology, Istituto Nazionale per lo Studio e la Cura dei Tumori, Milano and Molecular Genetics of Cancer Group, Fondazione Istituto FIRC di Oncologia Molecolare (IFOM), Via Adamello 16, 20139 Milano, Italy. email: lara.lusa@ifom-ieo-campus.it.

Introduction

Gene-expression microarrays measure simultaneously the expression of thousands of genes and are nowadays widely used in biomedical research to pursue many different objectives.

Microarrays have been used since the end of the 90's.[1,2] Since then, it has become clear that adequate statistical methods play a crucial role in maximizing the potentials of this rapidly evolving field.

In early microarray studies statistics was often misused or not used at all, and this sometimes resulted in scientific contributions that presented results that were unreliable and not reproducible.[3-5]

At the same time, thanks to the extensive use of microarrays in biomedical research, novel statistical methods were developed while many old statistical methods and principles gained new popularity.

The aim of this paper is to give a brief overview of the objectives of microarray gene-expression experiments and to describe some statistical issues related to study design and data preprocessing. Quality control, normalization, replication, validation and use of pooling of independent samples will be disussed. Specific methods for data analysis have been discussed elsewhere[6,7] and will not be covered here.

Objectives and characteristics of gene expression microarray experiments

Most of the objectives of gene-expression microarray experiments can be categorized in three broad classes

- **class discovery objectives**: when the aim is to discover previously unknown subgroups of genes or subjects that are homogeneous in their gene expression (for example, Perou *et al.*[8] proposed a molecular classification of breast cancer identifying different subtypes with distinct gene expression profiles);

- **class comparison objectives**: when the aim is to compare two or more classes (phenotypes or experimental conditions) in terms of gene expression, identifying genes that are differentially expressed between them (for example, Hedenfalk *et al.*[9] compared expression profiles of sporadic breast cancers and of breast cancer tumors with mutations in BRCA1 and BRCA2 genes, and identified the genes that were differentially expressed between the three groups); class comparison can be seen as a special case of all those problems in which it is of interest to evaluate the association of gene expression with other variables such as, for example, expression levels of a biological marker, size of the tumor, survival time;

- **class prediction objectives**: when the aim is to develop classifiers based on expression profiles that predict an outcome (for example, van't Veer *et al.*[10] developed a predictor based on the expression of 70 genes to predict the relapse of breast cancer within 5 years after surgical treatment).

It is not uncommon for microarray experiments to pursue more than one of these aims at the same time.

From a statistical point of view, the most important peculiarity of microarray experiments is the large number of variables (genes) being measured for each subject. On the other hand, in most experiments the sample size (the number of subjects) is still very small. This is known as the "large p, small n" problem,[11] where p is the number of measured variables and n the sample size; due to this characteristic, the straightforward application of standard statistical procedures for data analysis of microarray experiments can be problematic.

The most dangerous consequence of the "large *p*, small *n*" problem in class comparison experiments is the so called *multiple testing problem.* Differentially expressed genes between the classes are generally identified performing hypothesis testing gene by gene. In order to control for false discoveries, several multiple testing procedures, which control in different ways for false discoveries, are available and should be used.[12-14,6,7]

Another consequence of the "large *p*, small *n*" problem is the possibility of easily overfitting data when constructing class predictors using gene expression data. Therefore, when independent data are not available for external validation, the performance of the predictor has to be properly evaluated using cross-validation or bootstrap techniques.[4,6,7]

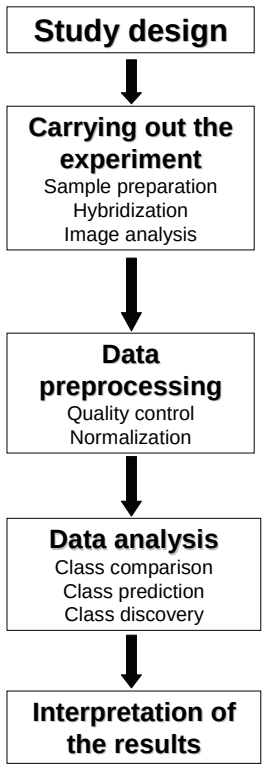


Figure 1 Schematic of the steps of a microarray experiment.

The use of appropriate statistical methods is important in all steps of a microarray experiment, starting from the experimental design. Figure 1 gives a schematic of the steps of a microarray experiment, which include: study design, carrying out of the experiment, quality control, normalization, data analysis and interpretation of the results. It can be noted that after the experiment has been carried out, an additional step is generally required before data analysis.

Preprocessing of the data

Quality control and normalization steps are sometimes referred to as *preprocessing of the data* and they differ substantially between one-channel (Affymetrix Gene Chips ®) and two-channel arrays (two-color cDNA arrays or long oligonucleotide arrays). The reason is related to the many differences between these arrays: the way in which gene expression is measured, the sources of systematic biases, the way image analysis is performed, with different algorithms and different outputs. [15]

Even within the same type of arrays usually there is no general agreement on which method should be used for the preprocessing of the data: many alternatives exist and some of these problems are topics of active statistical research[16-19]. Although the Affymetrix proprietary software includes a program for the preprocessing of GeneChips,[20] many *ad hoc* methods were independently developed for this aim, which generally perform better than the original method. [18]

With the quality control step, genes or samples that were not measured *reliably* are removed from the analysis. Since microarray data are normally very noisy, usually this step reduces greatly the number of genes that are eventually used in the analysis. A lot of useful information on the quality of the measurements can be derived from visual inspection of the images and from image analysis outputs. Generally, small spots, spots with relative large background, with weak or saturated signals are considered unreliable. However, determining, for example, how small a spot should be to be unreliable or, more in general, what a *reliable*

measurement is, is to a great extent arbitrary and not many commonly accepted rules or methods exist.

Within gene-expression microarray experimental procedures, there are many factors that are unrelated to the biological characteristics of the samples but that can influence the outcome of the experiment; the normalization step is aimed at removing these experimental artifacts which can produce systematic biases.

Such experimental artifacts can be due, among other technical reasons, to

- imbalances between RNA amounts,
- RNA amplification,
- RNA degradation,
- retro-transcription efficiency,
- efficiency in dye incorporation or dye fading,
- order in which arrays were hybridized,
- batch to which the slides belong,
- operator that performed the hybridizations or the RNA extraction,
- temperature, humidity or ozone level at the time of the hybridization,
- efficiency of the washing procedure.

Common choices for the normalization of the data include median centering the arrays and methods that adjust the gene expression depending on their intensity or location on the array;[21,17] sometimes just a subset of the genes, which expression is supposed not to change across arrays (*housekeeping genes*), is used for the normalization. However this approach can be difficult to apply since there is no general agreement on how to identify these genes.

Ideally, the normalization process should be incorporated in data analysis rather than separated from it and treated as a preprocessing step.[22] Some attempts in this sense have been made in the context of microarrays,[22-25] mostly using ANOVA models and considering the *nuisance* effects together with the effects of interest. This kind of approach has in principle many advantages, mostly because it does not suppose that no additional error is introduced by the normalization. However, modeling appropriately the nuisance effects can be difficult and computationally challenging and therefore this approach is seldom used in practice.

In general, there is no best method for normalizing data. The choice of a specific normalization method should depend on the characteristics of the data at hand and it should be made after a careful inspection of the data. When data characteristics allow it, the simplest methods should be used, so as to avoid making too many assumptions and limiting the overfitting of the data.

Experimental Design

Proper experimental design plays an essential role for correctly addressing the questions of interest and a clear definition of the objectives of the study is crucial for the correct planning of a microarray experiment.

While experimental design was largely neglected in the early stage of microarray use, its need is now becoming increasingly acknowledged, with a greater emphasis on the importance of replication.[26-28]

Replication

The use of replicates from independent subjects (*biological replicates*) cannot be avoided when the aim of the experiment is finding results that can be extended from the samples being analyzed to the populations to which they belong, *i.e.* drawing proper statistical inference. Multiple

measurements of the same subject (*technical replicates*) do suffice only in quality control studies, where just the evaluation of the reproducibility of the measurements and of the error associated with the array process is of interest.[27,28]

The distinction between biological and technical replicates has been a source of substantial confusion in early microarray experiments, where often only technical replicates were used.[27] Misuses are still frequent in experiments with cell lines or inbred animals, where biological variability is supposed to be negligible or very small.

Methods for the calculation of the number of independent biological replicates needed in a microarray experiment depend on the aim of the experiment, on the method used for data analysis and, in two-channel arrays, on the way samples are allocated to the arrays.

The main feature of two-channel arrays is that two samples are hybridized on the same array. Many alternative designs exist,[29] but the simplest way to allocate samples on the arrays is to use a *reference design*, in which an aliquot of a reference sample is labeled with the same label and hybridized on each array. This design has many advantages over its alternatives:[29] it allows the direct comparisons of any subset of arrays in class comparison problems, also across experiments if the same reference was used. Moreover, it is robust to the presence of bad quality arrays and data can be straightforwardly used also in class discovery and class prediction problems. Last but not least, it is easy to perform in the lab.

A disadvantage of the reference design is that half of the hybridizations are used for the reference sample, for which usually there is no biological interest. Other designs which allocate samples more efficiently or that have some *optimality* properties have been proposed, examples being the balanced block design,[28] the loop design[23] and the interwoven loop.[30] These methods for sample allocation are less flexible, less suited for class discovery problems, they require more complex methods for the analysis of data and are more sensitive to the presence of bad quality arrays.

For class comparison studies, methods based on power analysis and depending on the method in which samples are allocated on the arrays have been proposed by Dobbin and Simon,[29,31] which used classical statistical sample size reasoning, taking into account the multiple testing problem, and reviewed[31] previously proposed methods for sample size calculation for microarray experiments. These methods usually require some knowledge on gene variability in the population of interest and on the variability of the experimental error, both of which might be available from previous experiments or can be estimated by pilot studies.

Sample size estimation methods have been developed also for class prediction problems and are more complex, having to take into account the variability deriving from both the predictor construction and the choice of the genes to be included in the predictor (*feature selection*).[32-34]

Randomization and confounding

As mentioned in the section on the preprocessing of data, in microarray gene-expression experiments many factors that are unrelated to the characteristics of the samples can influence the outcome of an experiment. Normalization is generally used to correct for these effects. There are however some situations in which, due to bad design of the experiment, normalization cannot be effectively used for its purpose.

As an example, suppose that in an experiment all the samples of normal tissue are hybridized in one day, while all the tumor samples are processed on the following day, in which the level of the humidity unexpectedly and dramatically rises. When looking for the genes that are differentially expressed between normal and tumor tissue, it will be impossible to identify the genes that have a different expression in the two types of tissues from those which change was influenced by the humidity level.

Given the strong influence that experimental factors can have on the hybridization results, the usual principles of statistical study design should be applied also to microarray experiments, the most basic of which is the randomization of the samples to the levels of the known confounding factors.

Even though some methods for correcting for batch effects exist,[35] in most cases some care in the experimental design can avoid their use and make data analysis simpler and not dependent on additional assumptions.

Planning the validation of the results

In class comparison experiments a very common practice consists in validating the results obtained from the analysis of microarray data with a different technology, usually with real time – quantitative polymerase chain reaction (RT-QPCR).

Most of the times the validation is performed on the same samples that were used in the microarray experiment. In this case what is being validated is merely the validity of the microarray measurements and therefore this approach cannot be seen as a way to improve the confidence on the generalizability of the results. Moreover, the genes that are not identified as being differentially expressed from the microarray experiment are hardly ever validated, so this kind of validation can identify false positive but not false negative results.

When an independent set of samples is used to validate the findings of a microarray experiment, comparisons with the original findings are generally made comparing P-values rather than the sizes of the effects.

In class prediction problems the validations of the results is usually made evaluating the predictive accuracy of the model using cross-validation or bootstrap.[4,6,7]

When an independent set of samples is used to validate a predictive model developed from gene-expression microarray data, it is important that the predictive model is completely specified before applying it to the new data set.[36] This avoids running the risk of overfitting the model on the new data and obtaining biased estimates of the predictive accuracy. This problem is particularly interesting when RT-PCR or custom arrays are used to measure the genes that were included in the predictive model developed from original microarray data. In this case, since the measurements are made using different methods of measurements it seems legitimate to re-estimate the model on the independent data set. However, this cannot be claimed to be a completely independent validation and the model developed on the independent data set can still be prone to overfitting.

Pooling of samples

Whether to pool samples and hybridize them instead of individual samples on the arrays is another option in the design of gene-expression microarray experiments.

Pooling the RNA of independent samples is a necessary choice in microarray experiments where the amount of available RNA from each sample is not sufficient for obtaining a good quality array[37] and RNA amplification is not considered.

Even when the RNA quantity is not a concern, investigators often consider pooling of samples as a choice when designing their class comparison studies.[38,39] Pooling is seen as an effective way to cut the costs of the expensive microarray experiments, while reducing the biological variability, at the cost of loosing the individual information.

Many investigators have been misusing pooling, typically obtaining one pool per condition and then hybridizing one or multiple aliquots of the pools on the arrays. As noted above, independent biological replicates are needed in order to make inference on the populations to which samples

belong. Therefore, multiple independent pools for each class, each composed by different units, must be used in order to be able to extend the experimental findings from the sample to the classes to which the pools belong.[27]

Recently some papers addressed the issue of sample size requirements for microarray class comparison experiments with pooled samples and compared pooled and individual samples designs. It was shown that increasing the number of independent subjects included in the study, comparable precision or power to a non-pooled design can be obtained by a pooled design with fewer arrays.[40,41]

However, there is some experimental evidence that the major assumption underlying pooling, namely that the gene expression of the pool equals the average expression of the individual samples in that pool, may not hold. Shih *et al.*[41] showed that gene-expression of the pool can significantly differ from the average expression of the individual samples, especially for high signals and more markedly for Affymetrix data. Moreover, experimental data showed that the expected reduction of overall within-class variability in pooled samples can be observed for only a part of the genes[42-44](from 70 per cent to 40 per cent).

Kendziorski *et al.*[42] after a comprehensive experimental comparison of pooled and non-pooled experimental results, recommend that "pooling be done when fewer than 3 arrays are used in each condition". However, more than 2 independent replicates per condition should be used in each microarray experiment in order to apply statistical methods for the analysis of data, therefore the utility of pooling seems limited.

Especially when the biological variability of the samples is expected to be small compared to technical variability, pooling is not likely to be beneficial and a large number of individual samples is required in order to be able to reduce the number of arrays without loosing power.

Conclusions

The focus of this paper was mainly on experimental design, which constitutes a fundamental but often neglected step in each microarray experiment. This aspect, together with thoughtful validation of the results is crucial for transferring the discoveries of this powerful tool into clinical applications.

Acknowledgements

This work was partially supported by an Italy-U.S.A. Fellowship of the Istituto Superiore di Sanità on Oncological Pharmacogenomics-Seroproteomics. I would like to thank James F. Reid for helpful discussion.

References

1. Schena M, Shalon D, Davis RW, et al.: Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 1995; 270 (5235): 467-470.
2. Lockhart DJ, Dong H, Byrne MC, et al.: Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol. 1996;14(13):1675-80.
3. Lee M-LT, Kuo FC, Whitmore GA, et al.: Importance of replication in microarray gene expression studies: Statistical methods and evidence from repetitive cDNA hybridizations. Proceedings of the National Academy of Sciences U S A 2000; 97(18):9834-9839.
4. Simon R, Radmacher MD, Dobbin K, et al.: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. Journal of the National Cancer Institute 2003; 95(1): 14-18.
5. Ioannidis JPA: Microarrays and molecular research: noise discovery? Lancet 2005; 365(9458): 454-455.
6. Simon RM, Korn EL, McShane LM, et al.: Design and analysis of DNA microarray investigations. New York (NY) 2004; Springer-Verlag.
7. Speed T, editor: Statistical analysis of gene expression microarray data. Boca Raton (FL) 2003; Chapman & Hall/CRC.
8. Perou CM, Sorlie T, Eisen MB, et al.: Molecular portraits of human breast tumours. Nature 2000; 406: 747-52.

9. Hedenfalk I, Duggan D, Chen Y, et al.: Gene-expression profiles in hereditary breast cancer. N Engl J Med. 2001; 344(8): 539-48.
10. van't Veer LJ, Dai H, van de Vijver MJ, et al.: Gene expression profi ling predicts clinical outcome of breast cancer. Nature 2002; 415: 530-6.
11. West M: Bayesian factor regression models in the "large p, small n" paradigm. Bernardo JM, Bayarri M, Berger J, Dawid A, Heckerman D, Smith A,West M (Eds.), Bayesian Statistics 7, Oxford University Press, 2003, pp. 723-732.
12. Benjamini Y, Hochberg Y: Controlling the false discovery rate: A practical and powerful approach to multiple testing. J. R. Statist. Soc. Ser. B 1995; 57: 289-300.
13. Korn EL, Troendle JF, McShane LM, et al.: Controlling the number of false discoveries: application to high-dimensional genomic data. Journal of Statistical Planning and Inference 2004;124(2):379-398.
14. Storey JD, Tibshirani R: Statistical significance for genomewide studies. Proc Natl Acad Sci U S A 2003;100(16):9440-5.
15. Holloway AJ, van Laar RK, Tothill RW, et al.: Options available--from start to finish--for obtaining data from DNA microarrays II. Nat Genet. 2002;32 Suppl:481-9.
16. Li C, Wong W: Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. Proceedings of the National Academy of Science U S A 2000; 98: 31-36.
17. Irizarry RA, Hobbs B, Collin F, et al.: Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics 2003; 4:249-264.
18. Cope LM, Irizarry RA, Jaffee HA, et al.: A benchmark for Affymetrix GeneChip expression measures. Bioinformatics 2004;20(3):323-31.
19. Wu Z, Irizarry R, Gentlemen R, et al.: A model-based background adjustment for oligonucleotide expression arrays. Journal of the American Statistical Association 2004; 99(468): 909-917.
20. Affymetrix: Statistical algorithms reference guide:Technical report, 2001; Affymetrix.
21. Yang YH, Dudoit S, Luu P, et al.: Normalization of cDNA microarray data: a rubust composite method addressing single and multiple slide systematic variation. Nucleic Acid Research 2002; 30:4:e15.
22. Wu Z, Irizarry RA: A Statistical Framework for the Analysis of Microarray Probe-Level Data. Johns Hopkins University, Dept. of Biostatistics Working Papers. Working Paper 73 2005; http://www.bepress.com/jhubiostat/paper73.
23. Kerr M, Afshari C, Bennett L, et al.: Statistical analysis of a gene expression microarray experiment with replication. Statistica Sinica 2002; (12): 203-217.
24. Kerr M, Martin M, Churchill GA: Analysis of variance for gene expression microarray data. Journal of Computational Biology 2000; 7: 819-837.
25. Wolfinger R, Gibson G, Wolfinger E, et al.: Assessing gene significance from cDNA microarray expression data via mixed models. Journal of Computational Biology 2001; 8(6): 625-637.
26. Yang YH, Speed T: Design issues for cDNA microarray experiments. Nat Rev Genet. 2002; 3(8): 579-88.
27. Churchill GA: Fundamentals of experimental design for cDNA microarrays. Nat Genet 2002; 32 Suppl: 490-5.
28. Simon R, Radmacher MD, Dobbin K: Design of studies using DNA microarrays. Genet Epidemiol. 2002; 23: 21-36.
29. Dobbin K, Simon R: Comparison of microarray designs for class comparison and class discovery. Bioinformatics. 2002; 18(11): 1438-45.
30. Wit E, McClure JD: Statistics for Microarrays; Design, Analysis and Inference, Chichester 2004; John Wiley & Sons.
31. Dobbin K, Simon R: Sample size determination in microarray experiments for class comparison and prognostic classification. Biostatistics 2005; 6: 27-38.
32. Dobbin K, Simon R: Sample size planning for developing classifiers using high dimensional DNA microarray data. Biostatistics 2006; In Press. doi:10.1093/biostatistics/kxj036
33. Fu WJ, Dougherty ER, Mallick B, et al.: How many samples are needed to build a classifier: a general sequential approach. Bioinformatics 2005; 21: 63-70.
34. Mukherjee S, Tamayo P, Rogers S, et al.: Estimating dataset size requirements for classifying DNA microarray data. Journal of Computational Biology 2003; 10: 119-142.
35. Johnson WE, Rabinovic A, Li C: Adjusting batch effects in microarray expression data using Empirical Bayes methods. Biostatistics 2006; In Press. doi:10.1093/biostatistics/kxj036

36. Simon R: Development and Validation of Therapeutically Relevant Multi-Gene Biomarker Classifiers. J Natl Cancer Inst 2005; 97: 866-7.
37. Jin W, Riley RM, Wolfinger RD, et al.: The contributions of sex, genotype and age to transcriptional variance in Drosophila melanogaster. Nature Genetics 2001; 29: 389-395.
38. Agrawal D, Chen T, Irby R, et al.: Osteopontin identified as lead marker of colon cancer progression, using pooled sample expression profiling. Journal of the National Cancer Institute 2002; 94: 513-21.
39. Enard W, Khaitovich P, Klose J, et al.: Intra- and interspecific variation in primate gene expression patterns. Science 2002; 296: 340-343.
40. Kendziorski CM, Zhang Y, Lan H, et al.: The efficiency of pooling mRNA in microarray experiments. Biostatistics 2004; 4: 465-477.
41. Shih JH, Michalowska AM, Dobbin K, et al.: Effects of pooling mRNA in microarray class comparisons. Bioinformatics 2004; 20: 3318-3325.
42. Kendziorski C, Irizarry RA, Chen K-S, et al.: On the utility of pooling biological samples in microarray experiments. Proceedings of the National Academy of Sciences U S A 2005; 102: 4252-4257.
43. Han ES, Wu Y, McCarter R, et al.: Reproducibility, sources of variability, pooling, and sample size: important considerations for the design of high-density oligonucleotide array experiments. J Gerontol A Biol Sci Med Sci 2004; 59: B306-315.
44. Lusa L, Cappelletti V, Gariboldi M, et al.: Caution regarding the utility of pooling samples in microarray experiments with cell lines. International Journal of the Biological Markers 2006; In Press.

Izvirni znanstveni članek ■

Diagnostika raka z DNA mikromrežami – preprosti in razumljivi vizualni modeli

Minca Mramor, Gregor Leban, Janez Demšar, Blaž Zupan

Izvleček. V zadnjem času je tehnologija DNA mikromrež omogočila globalni vpogled v spremembe izraženosti genov v rakastem tkivu in postala praktično nepogrešljiva v raziskavah raka. V članku podajamo kratek pregled metod analize podatkov pridobljenih z mikromrežami. Uveljavljene metode za gradnjo diagnostičnih in drugih napovednih modelov iz genskih podatkov večinoma temeljijo na sorazmerno zapletenih in težko razumljivih računskih modelih. Kot pokažemo v članku je le-te moč nadomestiti s preprostimi, a učinkovitimi vizualizacijskimi tehnikami. V prispevku predstavljamo metodo VizRank, ki v množici možnih vizualizacij poišče take, ki omogočajo jasno ločitev diagnostičnih razredov z uporabo le nekaj spremenljivk na vseh preiskovanih bazah podatkov.

DNA Microarray Cancer Diagnostics – Simple and Effective Visual Models

Abstract. Today's DNA microarray technology enabled researchers to obtain a global view of human cancer gene expression and is becoming indispensable in cancer research. In the paper, we first present a short overview of the methods used in the analysis of microarray data. The majority of recently applied diagnostic prediction methods are based on complex computational methods and the models they produce are therefore hard to understand and interpret by biologists. Alternatively, we show that a relatively straightforward approach that searches through the space of possible data projections can find simple graphs that are easy to interpret, show good class separation, and include only a small number of genes.

■ **Infor Med Slov**: 2006; 11(1): 25-33

Instituciji avtorjev: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani (MM, GL, JD, BZ), Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, USA (BZ).

Kontaktna oseba: Minca Mramor, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Tržaška 25, 1000 Ljubljana. email: minca.mramor@fri.uni-lj.si.

Uvod

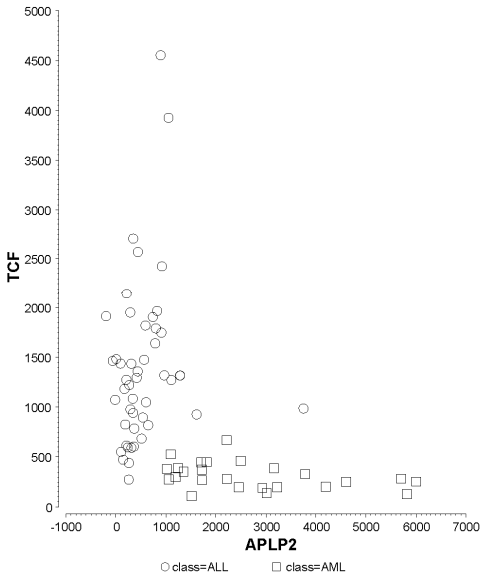
Rakasta obolenja so posledica progresivnih genetskih in epigenetskih sprememb, ki vodijo pretvorbo normalnih celic v njihove maligne derivate. Genetske spremembe so predvsem mutacije v onkogenih in tumor supresorskih genih, medtem ko epigenetski mehanizmi uravnavajo prepisovanje genov preko različnih mehanizmov, kot so npr. modulacija strukture kromatina, metilacija DNA in inaktivacija X kromosoma. Zaradi izredne raznolikosti različnih tipov raka in kompleksnosti same bolezni je natančna diagnostika raka velik izziv.[1,2] Pri nekaterih tipih raka se izziv začne že pri postavljanju začetne diagnoze (npr. levkemija, [3] glioblastomi),[4] pri mnogih drugih pa je težko napovedati odgovor na zdravljenje, ponovitev bolezni po končanem zdravljenju, razsoj metastaz... Trenutna diagnostična orodja, kot so klinična slika, TNM klasifikacija, slikovna diagnostika, histološki pregled tkiva in izbrani tumorski markerji, pogosto ne morejo zadovoljivo odgovoriti na pomembna vprašanja v diagnostičnem procesu, saj je njihova prognostična vrednost močno omejena.

V zadnjih nekaj letih se zato pospešeno razvijajo metode, ki omogočajo ugotavljanje sprememb v rakastih celicah na DNA, RNA in proteinskem nivoju. Trenutno najbolj razvita in uporabljana je tehnologija DNA mikromrež, s katero lahko simultano merimo količino več tisoč različnih mRNA molekul v biološkem vzorcu in iz tega sklepamo o izražanju pripadajočih genov. Številne raziskave so pokazale superiorne diagnostične zmožnosti DNA mikromrež za klasifikacijo rakastih obolenj v primerjavi s standardnimi morfološkimi kriteriji.[2-4,5] Cilji uporabe mikromrež v raziskavah raka so vpogled v proces karcenogeneze, identifikacija biomarkerjev za različne tipe raka, natančnejša klasifikacija ter izboljšanje in individualizacija zdravljenja z razvojem novih, usmerjenih terapevtikov. [5]

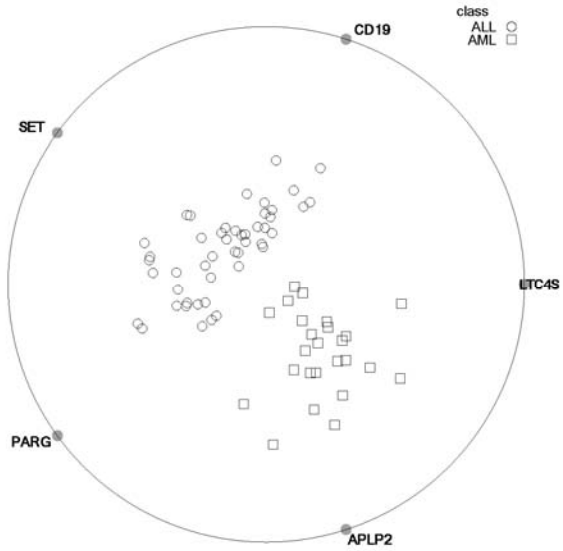
Največji problem podatkov, ki jih pridobimo z merjenji izražanja genov, je njihova visoka dimenzionalnost, saj navadno vključujejo več tisoč spremenljivk (genov) in majhno število vzorcev (bolnikov). Poleg tega so meritve izredno občutljive na zunanje dejavnike, zato je v podatkih pogosto prisotno veliko šuma, posledica pa je tudi slaba ponovljivost rezultatov v različnih eksperimentih. Stopnje v analizi podatkov pridobljenih z mikromrežami so navadno predobdelava podatkov, izbor spremenljivk in gradnja klasifikacijskih modelov iz podatkov z nenadzorovanim in nadzorovanim učenjem.[6,7] V ozadju tovrstnih analiz sta dva glavna cilja. Prvi je izbor majhnega števila genov, ki najbolje ločijo med napovednimi razredi in bi bili morda lahko primerni za nove klinične tumorske markerje posameznih tipov raka. Drugi cilj pa je izgradnja klasifikacijskih modelov za natančnejšo diagnostiko raka glede na izraženosti genov, ki istočasno omogoča ugotavljanje zanimivih interakcij med geni in odkrivanje novih spoznanj o nastanku in razvoju raka.

V naslednjem poglavju bomo podali kratek pregled najbolj uporabljanih metod na področju izbora spremenljivk in gradnje klasifikacijskih modelov. V drugem delu prispevka nato predstavljamo preprosto metodo za analizo in vizualizacijo podatkov pridobljenih z mikromrežami. Le-ta temelji na preiskovalnem algoritmu VizRank[8], ki preišče prostor možnih vizualizacij in za vsako oceni, kako dobro loči posamezne napovedne razrede, npr. vrsto raka. Za prikaz podatkov smo uporabili dve osnovni dvodimenzionalni vizualizacijski metodi - razsevni diagram za prikaz izraženosti dveh genov in radviz diagram[9] s tremi in več geni. Primera takih vizualizacij na podatkih o izražanju genov pri dveh vrstah levkemije[3] sta prikazana na sliki 1 (razsevni diagram) in sliki 2 (radviz diagram).

Na obeh primerih vizualizacij (slika 1 in 2) so napovedni razredi dobro ločeni. Poleg tega, da lahko iz grafov razberemo vlogo posameznih genov in njihov interakcijski učinek pri ločevanju diagnostičnih razredov, nam VizRank s tem, ko poišče vizualizacije z dobro ločljivostjo napovednih razredov, implicitno tudi omogoča izbor najpomembnejših genov za napoved vrste raka in identifikacijo potencialnih izstopajočih primerov.



Slika 1 Najbolje ocenjeni razsevni diagram za razločevanje med akutno limfoblastno levkemijo (ALL) in akutno mieloidno levkemijo (AML) na podlagi izraženosti genov APLP2 in TCF.



Slika 2 Najbolje ocenjeni radviz diagram za razločevanje med akutno limfoblastno levkemijo (ALL) in akutno mieloidno levkemijo (AML) na podlagi izraženosti petih genov, CD19, SET, PARG, APLP2 in LTC4S.

Zanimivo je, da so najboljše vizualizacije preiskovanih baz podatkov iz študije, ki jo predstavljamo v članku, praviloma vsebovale tudi nekatere znane, biološko relevantne gene.

Obe vizualizaciji lahko uporabljamo tudi za napovedovanje razredov novih primerov. Za ročno rabo je primernejši razsevni diagram. Če je pri nekem novem vzorcu vrednost gena ALPL2 enaka 500 in gena TCF 1500, sodi le-ta globoko v področje, ki ga zasedajo vzorci limfoblastne levkemije. Princip uporabe radviza je podoben, vendar pri njem zaradi bolj zapletene projekcije pri napovedovanju navadno potrebujemo računalnik, ki izračuna položaj novega primera v diagramu in določi njegov razred na podlagi bližnjih primerov.

Predstavljeni vizualizaciji sta torej uporabni pri reševanju obeh v uvodu predstavljenih problemov, izboru manjše množice genov uporabnih v diagnostiki in gradnji diagnostičnih modelov.

Analiza podatkov o izraženosti genov – pregled metod

Najpomembnejše stopnje v analizi diagnostičnih in prognostičnih podatkov o rakastih oboljenjih pridobljenih z mikromrežami so predobdelava podatkov (angl. *preprocessing*), izbor napovednih spremenljivk (angl. *feature selection*) ter gradnja klasifikacijskih modelov z nenadzorovanim učenjem oz. razvrščanje v skupine za odkrivanje novih razredov (angl. *class discovery*) in gradnja klasifikacijskih modelov z nadzorovanim učenjem oz. napovedovanje razredov (angl. *class prediction*). Predobdelava podatkov vključuje analizo slik mikromrež, normalizacijo podatkov, ki naredi podatke med različnimi eksperimenti in platformami primerljive, uporabo postopkov za obravnavo manjkajočih vrednosti in ponovljenih meritev izraženosti istega gena ter numerično transformacijo podatkov.[6,7,10] Natančnejša obravnava metod predprocesiranja presega okvire tega članka je pa, na primer, lepo podana v preglednem članku Pham in soavtorjev.[10] V

nadaljevanju bomo podali pregled najbolj uporabljanih metod na ostalih stopnjah analize.

Izbor spremenljivk

Podatki pridobljeni z mikromrežami so zaradi velikega števila spremenljivk podvrženi tako imenovanemu prekletstvu dimenzionalosti. Za uspešnost klasifikacijskih algoritmov namreč velja splošno pravilo, naj bi bilo število vzorcev (veliko) večje od števila atributov.[7] Pri podatkih o izraženosti genov je stanje prav nasprotno, zato se pred gradnjo klasifikacijskih modelov pogosto uporabljajo različne tehnike zmanjševanja dimenzionalnosti s stališča števila spremenljivk. Ločimo jih na metode konstrukcije novih spremenljivk iz množice obstoječih(angl. *feature extraction*) in metode izbora podmnožice spremenljivk (angl. *feature selection*).

Značilnost metod konstrukcije novih spremenljivk je, da iz obstoječih izraženosti genov tvorijo nove spremenljivke. Najbolj uporabljani pristopi na tem področju so analiza glavnih komponent (angl. *principal component analysis*, *PCA*), večrazsežnostno lestvičenje (angl. *multidimensional scaling*, *MDS*) in samoorganizirajoče karte (angl. *self-organizing maps, SOM*). Glavni praktični problem uporabe teh metod pri analizi podatkov je, da so nove spremenljivke metageni, ki so sestavljeni iz mnogih genov in zato nimajo znanih bioloških in strukturnih lastnosti. Poleg tega klasifikator, ki uporablja metagene, potrebuje podatke o izraženosti vseh genov, iz katerih so sestavljeni, zato metageni niso uporabni za diagnostične teste ali razvoj tumorskih markerjev.[7]

Za izbor spremenljivk obstajata dva v osnovi različna pristopa: univariatni, ki ga uporabljajo precejalne metode (angl. *filtering methods*) in multivariatni, na katerem temeljijo metode na principu ovojnice (angl. *wrapper methods*).[7,11] Precejalne metode, ki med geni na podlagi univariante ocene napovedne moči posameznih genov preprosto izberejo podmnožico najbolje ocenjenih genov, so znane tudi kot metode "en gen naenkrat", saj navadno ne upoštevajo interakcij med geni in ocenjujejo sposobnost razlikovanja med razredi za vsak posamezni gen. Te metode zato imenujemo kratkovidne. Med seboj se razlikujejo predvsem glede izbrane metrike za ocenjevanje genov, npr. t-test, razmerje med signalom in šumom (angl. *signal to noise*), Wilcoxonov test, ANOVA... Problem teh metod je, da lahko spregledajo gene, ki so sami slabo informativni, v povezavi z drugimi geni pa bi bili lahko za napovedi razredov zelo uporabni.[11]

Med precejalnimi metodami izstopa ReliefF, ki vsak gen ocenjuje v lokalnih kontekstih, ki jih določajo vrednosti ostalih genov.[12,13] ReliefF zato ni kratkoviden, žal pa na podatkih z velikim številom spremenljivk ne more dobro določiti kontekstov, zato za analizo genskih mikromrež ni najbolj primeren.

Drug pristop temelji na iskanju podmnožice genov, ki maksimizira klasifikacijsko točnost izbranega algoritma strojnega učenja. Pristop z ovojnico torej zgradi klasifikator na podlagi določene podmnožice genov in glede na oceno uspešnosti klasifikatorja oceni podmnožico genov. [7,11] Prednost teh metod je, da na ta način ocenjujejo kvaliteto skupine genov ter, če uporabimo primerne algoritme za gradnjo napovednih modelov, v oceni upoštevajo tudi vpliv možnih genskih interakcij. Preveriti vse podmnožice tisočih genov v naših podatkih pa je praktično nemogoče, zato so bile razvite različne hevristike za pregledovanje prostora genskih podmnožic. Kljub temu je glavna težava pristopov z ovojnico velika računska kompleksnost.[7]

Nenadzorovano učenje in odkrivanje razredov

Pri nenadzorovanem učenju ali razvrščanju v skupine (angl. *clustering*) ne upoštevamo informacije o klasifikacijskem razredu. Metode razvrščanja iščejo naravne skupine v multidimenzionalnih podatkih na podlagi izbrane metrike podobnosti med primeri ali geni.[12] Metode nenadzorovanega učenja so izredno popularne v analizi mikromrež, saj zanje ne potrebujemo nikakršnih hipotez in nobenih predpostavk o podatkih, vedno pa podatke razvrstijo v skupine,

ne glede na velikost vzorca in kvaliteto podatkov. Prav to pa je tudi glavni problem teh metod, saj dobljene skupine pogosto nimajo nikakršnega biološkega ozadja.[6,7] Največkrat uporabljena metoda nenadzorovanega učenja za podatke pridobljene z mikromrežami je hierarhično razvrščanje, katerega rezultat lahko grafično predstavimo v obliki dendrograma.[6,7,14] Nehierarhične oziroma delitvene metode med drugim vključujejo k-povprečno razvrščanje (angl. k-*means clustering*), mešano modeliranje (angl. *mixture modelling*) in mnoge druge.[7,15]

Klasifikacijski modeli in napovedovanje razredov

Klasifikacijske oz. napovedne modele gradimo s t.im. nadzorovanim učenjem, naloga tako dobljenih modelov pa je novemu primeru, opisanemu z množico spremenljivk, določiti, kateremu izmed možnih razredov pripada.[12] Napovedne modele praviloma gradimo na učni množici, ocenjujemo pa na testni množici primerov. Obstajajo številni algoritmi za napovedno modeliranje, vsi pa so do neke mere podvrženi prevelikemu prileganju podatkom (angl. *overfitting*). To se povečuje z večanjem kompleksnosti klasifikacijskega modela in navadno vodi k zmanjšanju napovedne točnosti na novih (ali testnih) podatkih.[6,7]

Od najbolj znanih metod za gradnjo napovednih modelov iz podatkov naštejmo tu le najbolj uporabljane na področju bioinformatike. Te so k-najbližjih sosedov, umetne nevronske mreže, Fisherjeva linearna diskriminantna analiza, naivni Bayesov klasifikator, metode podpornih vektorjev (SVM) in klasifikacijska drevesa. Natančen opis teh metod z obravnavo njihovih prednosti in pomanjkljivosti se nahaja v preglednem članku Asyali in sodelavci.[7] Nobena od klasifikacijskih metod pa ni splošno sprejeta kot najboljša ali optimalna. Glede na velikosti vzorcev, ki so navadno na voljo pri raziskavah z mikromrežami, pa imajo preprostejši modeli, poleg tega, da so lažje razumljivi, pogosto tudi boljše napovedne lastnosti od kompleksnejših modelov.[6]

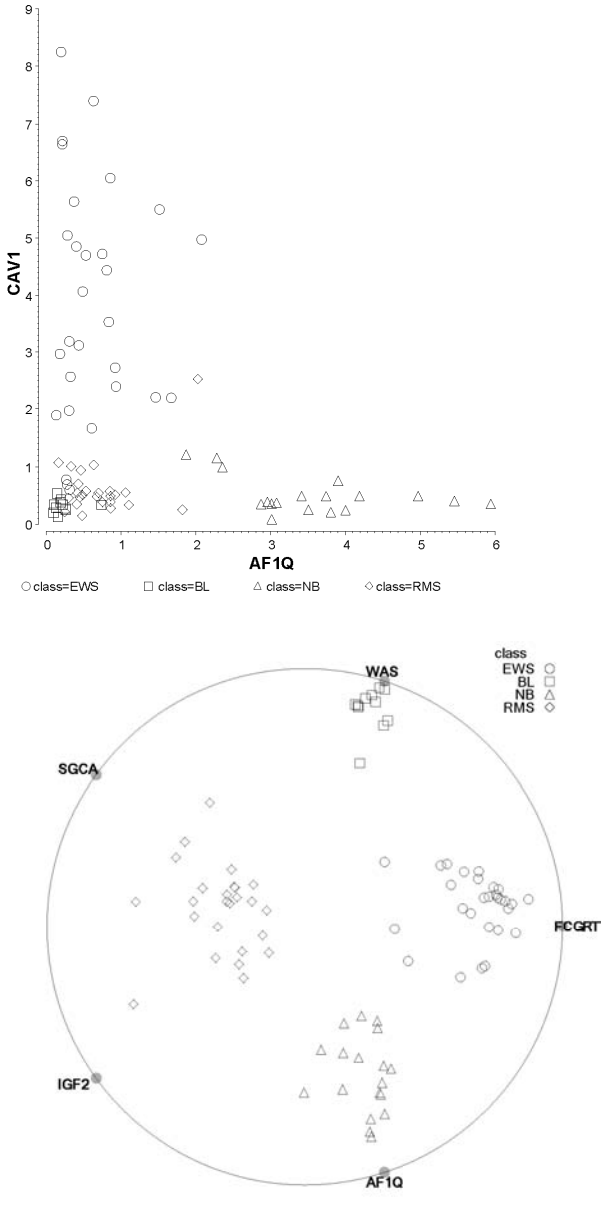
Metoda VizRank

Za vizualizacijo večdimenzionalnih podatkov o izraženosti genov smo izbrali dve dvodimezionalni geometrijski vizualizacijski metodi, razsevni diagram (slika 1) in radviz diagram (slika 2). Z razsevnim diagramom lahko prikažemo podatke glede na vrednosti dveh spremenljivk, radviz pa omogoča istočasni prikaz večjega števila spremenljivk. Na slikah 1 in 2 prikazujemo vizualizaciji, ki jasno ločita dve vrsti levkemije glede na izraženost majhnega števila genov. Nabori podatkov o izražanju genov pri bolnikih z rakom vsebujejo tisoče genov, zato ni trivialno poiskati dobro projekcijo z majhno podmnožico genov. Zaradi velikega števila genov ročno iskanje dobrih kombinacij genov in njihovih vizualizacij ne pride v poštev. Tako je na primer že za 100 genov možnih 4,450 različnih razsevnih diagramov, za 10,000 pa je teh 49,995,000. Pri iskanju dobrih n-teric genov za prikaz v projekciji radviz je problem še večji.

Za potrebe iskanja dobrih projekcij smo zato razvili računsko podprt postopek VizRank.[8] VizRank temelji na računsko določeni kvaliteti izbrane vizualizacije, ki priredi boljšo oceno vizualizacijam, ki bolje ločujejo med posameznimi napovednimi razredi, ter na hevrističnem preiskovanju prostora možnih vizualizacij. Oceno vizualizacije smo tako definirali s pomočjo metode *k*-najbližjih sosedov, ki poišče *k* sosednjih primerov glede na lego izbranega testnega primera v dvodimenzionalni projekciji (v poskusih v tem članku smo uporabili $k=5$). Ocena vizualizacije je delež testnih primerov, pri katerih je večina od najbližjih k sosedov v istem razredu kot testni primer. Takšna mera dobro razlikuje med projekcijami, v katerih so posamezni razredi dobro ločeni, in projekcijami, kjer se le-ti prekrivajo.[8]

Ker ovrednotenje vseh možnih projekcij tisočih spremenljivk ni mogoče, VizRank uporablja učinkovito hevristiko za preiskovanje prostora možnih projekcij. Spremenljivke najprej oceni z ReliefF-om,[13,14] nato pa projekcije uredi glede na vsoto ocen spremenljivk, ki nastopajo v njih. Če

projekcije ocenjujemo v takšnem vrstnem redu, je, kot kažejo poskusi, dovolj preiskati že zelo majhen del (navadno okoli 2 %) vseh možnih projekcij, da najdemo najboljše.



Slika 3 Najbolje ocenjeni razsevni (zgoraj) in radviz (spodaj) diagram za nabor podatkov o tumorjih v otroštvu (SRBCT) (EWS – Ewingov sarkom, BL – Burkitov limfom, NB – nevroblastom, RMS – rabdomiosarkom).

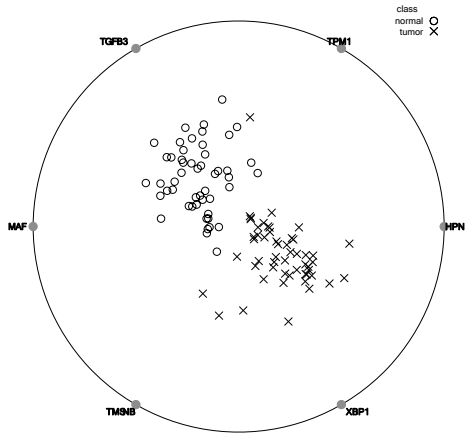
Poskusi in rezultati

Za eksperimentalni del raziskave smo uporabili osem naborov podatkov, ki so javno dostopni na spletu na strani http://www.broad.mit.edu/cancer, razen nabora podatkov SRBCT, ki je dostopen na strani http://research.nhgri.nih.gov/microarray/Supplement/. Nabori vsebujejo podatke o izraženosti 2308 do 12625 genov pri 40 do 230 bolnikih z rakom. Primeri so razvrščeni v dve do pet diagnostičnih skupin (različnih podvrst določenega raka). Glavne značilnosti vsakega nabora so povzete v Tabeli 1. Zadnji stolpec Tabele 1 prikazuje oceno najboljše projekcije radviz, to je povprečno verjetnost pravilne klasifikacije testnega primera.

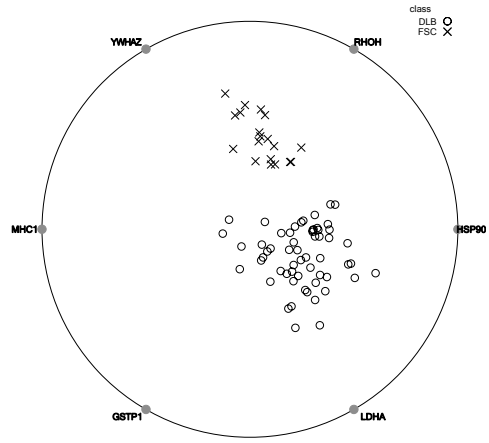
Tabela 1 Nabori podatkov.

Nabor podatkov	Št. vzorcev	Št. genov	Št. razredov	Radviz ocena
Levkemia	72	7074	2	100%
MLL	72	12533	3	100%
SRBCT	83	2308	4	100%
Prostata	102	12533	2	98%
DLBCL	77	7070	2	100%
Glio	50	12625	4	95%
Možgani	40	7129	5	93%
Pljuča	203	12600	5	97%

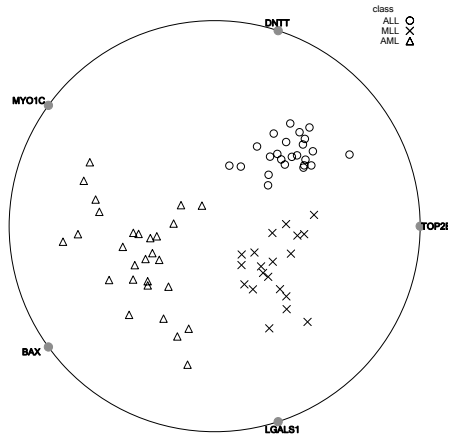
Za vse nabore podatkov smo z algoritmom VizRank poiskali vizualizacije, ki čim bolje ločijo med različnimi diagnostičnimi razredi. Izkazalo se je, da so na vseh najboljših vizualizacijah vrste raka jasno ločene (slike 1-4). V tem prispevku bolj podrobno opisujemo najboljše vizualizacije za en dvorazredni (Levkemija, sliki 1 in 2) in en večrazredni (SRBCT, slika 3) klasifikacijski problem. Slika 4 prikazuje najboljše radviz projekcije za preostale nabore podatkov, ocene pa so podane v Tabeli 1.



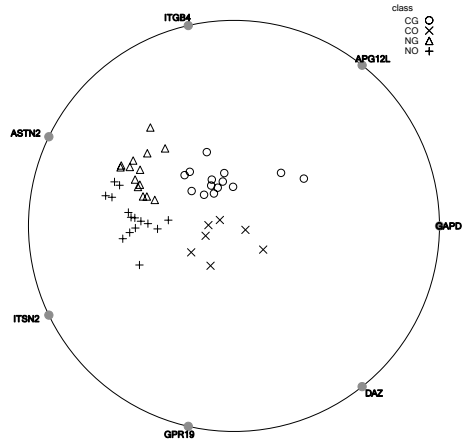
a) Prostata



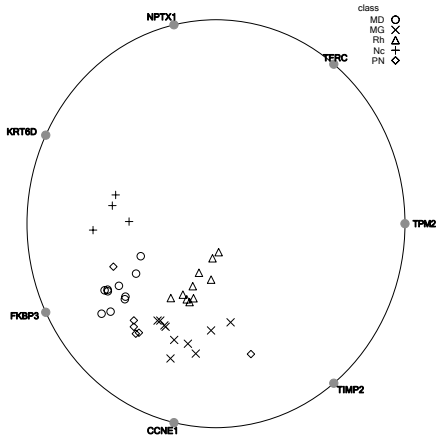
d) DLBCL



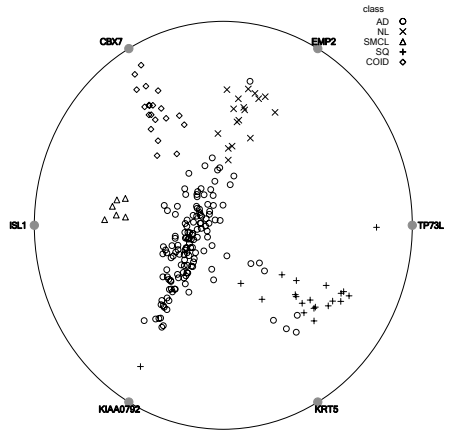
b) MLL



e) glioblastomi (glio)



c) možganski tumorji



f) pljučni tumorji

Slika 4 Najbolje ocenjeni radviz diagrami za razločevanje normalnega in tumorskega tkiva prostate (a), treh vrst levkemije (b), petih razredov možganskih tumorjev (c), folikularnih in difuznih velikoceličnih B-limfomov (d), štirih tipov glioblastomov (e) in petih razredov pljučnih tumorjev (f).

Najboljši razsevni diagram za nabor podatkov Levkemija (slika 1) prikazuje gena APLP2 in TCF. Razreda AML in ALL sta jasno ločena z le nekaj izstopajočimi primeri, zato ima diagram oceno 98 %. Slika 2 prikazuje radviz vizualizacijo s povsem jasno ločitvijo obeh levkemij in oceno 100 %. Za tako jasno ločitev je bilo potrebno uporabiti podatke o izraženosti petih genov (APLP2, SET, CD19, LTC4S in PARG) prikazanih na diagramu.

Večrazredni klasifikacijski problem predstavljen na sliki 3 je ločitev štirih vrst otroških tumorjev, ki imajo zelo podobne histološke značilnosti – sestavljeni so iz majhnih okroglih modrih nediferenciranih celic (*angl.* small round blue cell tumors oz. SRBCT). Ti štirje tumorji (Ewingov sarkom, nevroblastom, Burkitov limfom in rabdomiosarkom) predstavljajo težak diagnostični problem v pediatrični onkologiji, čimprejšnja pravilna diagnoza pa je nujna za ustrezno zdravljenje. Opazimo lahko, da na sliki 3 razsevni diagram, kjer lahko prikažemo le izraženost dveh genov, ne loči jasno med štirimi diagnostičnimi razredi, medtem, ko so razredi na vizualizaciji radviz s petimi geni povsem jasno ločeni.

Pomembno vprašanje, ki ga odpirata slika 4 in še bolj zadnji stolpec tabele 1, je, ali kažejo slike resnične vzorce ali pa so dobre ločitve le rezultat pretiranega prilagajanja podatkom. Med milijardami možnih vizualizacij bi celo pri popolnoma naključnih podatkih brez dvoma naleteli tudi na takšne z odlično ločenimi razredi. Za odgovor na to vprašanje v strojnem učenju navadno uporabljamo prečno preverjanje, ki nameni del podatkov gradnji modela (v našem primeru, vizualizacije), del pa testiranju klasifikacijske točnosti ali druge mere kvalitete. Rezultati takšnega poskusa kažejo, da so rezultati VizRanka primerljivi z najnaprednejšimi metodami strojnega učenja, torej predstavljene vizualizacije ne kažejo le naključno odkritih vzorcev.

Podrobnosti poskusa in njegovih rezultatov bomo zaradi obsežnosti opisali v drugem članku. V tem pa bomo na pomisleke o smiselnosti najdenih vizualizacij odgovorili s pomočjo ekspertnega predznanja. Pri analizi najboljše ocenjenih vizualizacij smo ugotovili, da v večini od njih nastopajo posamezni geni, ki so jasno povezani z določenimi vrstami raka. Tako na primer gen SGCA (sarkoglikan alfa) na vizualizaciji radviz za nabor podatkov SRBCT ločuje rabdomiosarkome od ostalih tumorjev. Rabdomiosarkom je mehko tkivni tumor, ki se razvije iz mišičnega tkiva in predstavlja nekaj manj kot 5 % rakov v otroštvu.[16] Produkt gena SGCA je protein sarkoglikan alfa, ki sodeluje v razvoju mišičnega tkiva in pri krčenju mišice. Gen je najmočneje izražen v skeletnih mišicah, v manjši meri pa tudi v srčni mišici in pljučih. [17] Ker se gen ne izraža v kostnem, limfnem in živčnem tkivu, od koder izvirajo preostali tumorji v tem naboru podatkov, je vloga SGCA v radvizu biološko smiselna.

Zaključek

V članku smo podali kratek pregled metod, ki se uporabljajo na različnih stopnjah analize podatkov pridobljenih z mikromrežami in so namenjenih diagnostiki rakastih obolenj. Pokazali smo, da je uveljavljene metode za gradnjo napovednih modelov iz genskih podatkov, ki večinoma temeljijo na zapletenih in nepredstavljivih računskih modelih, mogoče zamenjati s preprostimi, a učinkovitimi vizualizacijskimi tehnikami in postopki za preiskovanje prostora različnih vizualizacij. Za vse preiskovane nabore podatkov smo našli dvodimenzionalne vizualizacije z razsevnim ali radviz diagramom, ki jasno ločijo napovedne razrede. Predstavljene vizualizacije podatkov o izraženosti genov tudi dokazujejo, da so tumorski diagnostični razredi jasno ločljivi že z informacijo o izraženosti le nekaj najpomembnejših genov. Izbrani geni so pogosto biološko povezani z vrsto raka, ki ga ločujejo. Predlagane vizualizacije lahko služijo kot enostavni in razumljivi diagnostični modeli, obenem pa omogočajo odkrivanje podobnosti in razlik med različnimi vrstami raka in prepoznavo potencialnih tumorskih markerjev.

Literatura

1. Ponder BA: Cancer genetics. Nature 2001; 411(6835): 336-41.
2. Khan J, Wei JS, Ringnér M, et al.: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nature medicine 2001; 6(1): 673-679.
3. Golub TR, Slonim DK, Tamayo P, et al.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. Science 1999; 286(5439): 531-537.
4. Nutt CL, Mani DR, Betensky RA, et al.: Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification. Cancer Res 2003; 63(7): 1602-1607.
5. Ramaswamy S, Golub TR: DNA microarrays in clinical oncology. J Clin Oncol 2002; 20(7): 1932-41.
6. Allison DB, Cui X, Page GP, et al.: Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet 2006; 7(1): 55-65.
7. Asyali MH, Colak D, Demirkaya O, et al.: Gene expression profile classification: a review. Current Bioinformatics 2006; 1(1): 55-73.
8. Leban G, Bratko I, Petrovic U, et al.: VizRank: finding informative data projections in functional genomics by machine learning. Bioinformatics 2005; 21(3): 413-414.
9. Hoffman PE, Grinstein GG, Marx K, et al.: DNA Visual and Analytic Data Mining. IEEE Visualization 1997, 1997; 1: 437-441.
10. Pham TD, Wells C, Crane DI: Analysis of Microarray Gene Expression Data. Current Bioinformatics 2006; 1(1): 37-53.
11. Wang Y, Tetko IV, Hall MA, et al.: Gene selection from microarray data for cancer classification-a machine learning approach. Comp Biol Chem 2005; 29(1): 37-46.
12. Kononenko I: Strojno učenje. Ljubljana 2005: Založba FE in FRI.
13. Kononenko I, Simec E: Induction of decision trees using RELIEFF. In Mathematical and statistical methods in artificial intelligence. New York 1995: Springer Verlag.
14. Eisen MB, Spellman PT, Brown PO, et al.: Cluster analysis and display of genome-wide expression patterns. PNAS 1998; 95(25): 14863-14868.
15. Datta S, Datta S: Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics 2003; 19(4): 459-66.
16. Pizzo PA, Poplack DG: Principles and Practice of Paediatric Oncology (4th edition). Philadelphia 2001: Lippincott Williams and Wilkins.
17. UniProt – the universal protein knowledgebase Web Page. http://www.ebi.uniprot.org/entry/Q16586.

Izvirni znanstveni članek ■

Pristop k podatkovni analizi genskih mikromrež na področju varnosti hrane

Katarina Cankar, Jeroen van Dijk, Kristina Gruden, Andrej Blejec, Jim McNicol, Esther Kok

Izvleček. Tehnika mikromrež omogoča vpogled v izražanje nekaj tisoč genov naenkrat. Obdelava rezultatov velikih nizov podatkov, pridobljenih z mikromrežami, je velik izziv. Prispevek predstavlja splošno problematiko analize tovrstnih podatkov in naš pristop k temu na primeru uporabe mikromrež za proučevanje varnosti prehrane.

An approach to the analysis of DNA microarray data and its use in food safety

Abstract. DNA microarray technique enables the study of expression of a few thousands genes concurrently. Large datasets obtained by such experiments present an analytical challenge. The paper reviews some problems in analyzing such data and presents our approach to microarray data analysis as an example of use of microarrays in food safety.

■ **Infor Med Slov**: 2006; 11(1): 34-39

Institucije avtorjev: Nacionalni inštitut za biologijo, Ljubljana, Slovenija (KC, KG, AB), RIKILT Institute of Food Safety, Wageningen, Nizozemska (JvD, EK), Scotish Crop Research Institute, Dundee, Velika Britanija (JM).

Kontaktna oseba: Katarina Cankar, Nacionalni inštitut za biologijo, Oddelek za rastlinsko fiziologijo in biotehnologijo, Večna pot 111, 1000 Ljubljana. email: katja.cankar@nib.si.

Uvod

Tehnike molekularne biologije nam omogočajo vedno večji vpogled v genski kod različnih organizmov. Ker pa se organizmi v različnih pogojih različno odzivajo, nas zanima tudi, kateri geni so vključeni v določenih pogojih in kateri geni sodelujejo pri določenih procesih.

Starejše tehnike za preučevanje izražanja genov so dovoljevale spremljanje enega ali nekaj izbranih genov v različnih pogojih. Genske mikromreže pa nam omogočajo vpogled v izražanje več tisoč genov v enem vzorcu naenkrat. Čeprav s to tehniko v kratkem času pridobimo veliko podatkov, njihova analiza predstavlja zahteven izziv, saj delamo z zelo velikim številom spremenljivk naenkrat.[1]

Na področju proučevanja fiziologije rastlin je bila ta tehnika uporabljena že za študije cirkadianih ritmov, obrambe rastlin ob okužbi, odziva na stres, razvojnih faz rastlin ter asimilacije nitratov.[2] Namen našega pristopa pa je ugotoviti, ali lahko tehniko mikromrež uspešno uporabimo za ocenjevanje varnosti nove hrane (npr. gensko spremenjenih rastlin). Novo hrano se dandanes testira z vrsto tarčno usmerjenih testov: [3,4] opravi se preko dvesto testiranj, pri katerih se izvede analizo sestavin, hranilne vrednosti ter teste za toksičnost in alergenost. Vsi testi so opravljeni v primerjavi z hrano s podobnimi lastnostmi, ki je že na trgu.

Kljub velikemu številu testiranj pa obstaja možnost, da pride pri pripravi novega živila do nepričakovanih oziroma neželenih učinkov, ki jih s testi ne bi mogli odkriti. Za odkrivanje takšnih sprememb so zelo primerne netarčne metode, ki omogočajo širši pogled v spremembe v rastlini. V Evropski uniji zato poteka preizkušanje metod transkriptomike, proteomike in metabolomike z namenom zaznavanja sprememb v novih rastlinah.

Uporaba genskih mikromrež nam bo omogočila boljši vpogled v fiziologijo poljščin in veliko obeta tudi kot tehnika za ocenjevanje varnosti novih rastlin. Zanima nas, ali se različne prakse v kmetijstvu dejansko odražajo v izražanju genov, hkrati pa rezultati predstavljajo zbirko podatkov o variabilnosti v izražanju genov, s katero bomo kasneje lahko primerjali gensko spremenjene rastline.

Pri kompleksnih poskusih, kjer uporabljamo mikromreže, smo soočeni z ogromno količino podatkov in njihova obdelava ter izločanje pomembnih podatkov, ki odgovorijo na zastavljeno raziskovalno vprašanje, je velik izziv. Obdelava podatkov genskih mikromrež obsega več zaporednih stopenj.[2,5,6] Prvi pomemben korak je izbira genske mikromreže, primerne za naš poskus, ter dober načrt poskusa, ki bo odgovoril na zastavljeno raziskovalno vprašanje. Izvedbi poskusa sledi statistična analiza podatkov, ki obsega več korakov: analizo slike mikromreže, transformacijo in normalizacija podatkov, iskanje diferencialno izraženih genov, iskanje zakonitosti in razlago biološke vloge diferencialno izraženih genov.

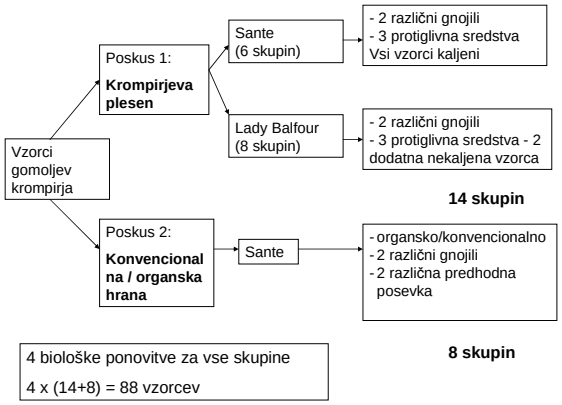
Prvi cilj analize podatkov je izločiti mikromreže, pri katerih hibridizacija ni bila uspešna, nato pa iz analize izločiti tudi posamezne točke mikromrež nezadostne kakovosti. Po izločanju nekakovostnih podatkov moramo podatke transformirati ter normalizirati, da lahko primerjamo podatke iz različnih mikromrež. Šele nato lahko vidimo, kateri vzorci imajo podoben profil izražanja genov, ter izražanje katerih genov se značilno razlikuje med posameznimi tretmaji.

Pristopov k analizi mikromrež je več in znanstveniki si niso enotni o optimalnem načinu analize podatkov. Statistična orodja, primerna za analizo podatkov mikromrež, ter ustrezna programska oprema so še vedno v razvoju. V nadaljevanju prispevka predstavljamo naš pristop k tej problematiki.

Načrt poskusa

Vzorce krompirja smo pridobili iz poskusa, izvedenega na Univerzi v Newcastlu v Veliki

Britaniji. Izvedena sta bila dva poljska poskusa (slika 1). V prvem poskusu (krompirjeva plesen) gre za organsko gojen krompir, ki je bil izpostavljen okužbi z glivo *Phytophthora infestans,* ki povzroča krompirjevo plesen. Uporabljeni sta bili dve različni sorti krompirja, gojeni z različnimi gnojili ter različnimi protiglivnimi sredstvi. V drugem poskusu pa nas je zanimala razlika med organskim in konvencionalnim pridelovanjem hrane. V tem poskusu sta bili uporabljeni dve vrsti gnojil, poleg tega pa so bili na polju predhodno posejani različni pridelki.

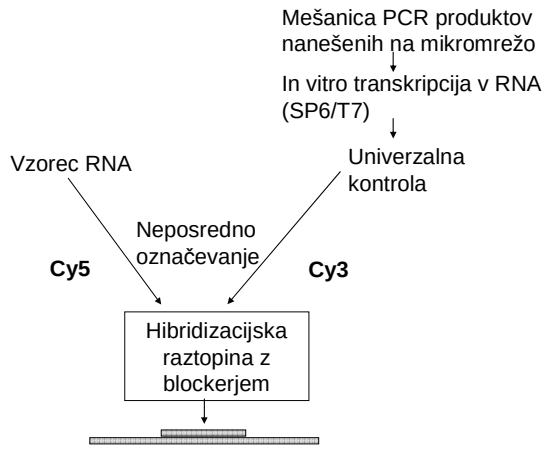


Slika 1 Shema poskusa.

Izvedba poskusa

Vzorce smo neposredno označili s fluorescentnim barvilom Cy5. Uporabili smo univerzalno kontrolo, ki je bila mešanica vseh PCR produktov, ki so bili natisnjeni na mikromreži. Kontrola je bila označena s fluorescentnim barvilom Cy3. Uporaba enake kontrole na vseh mikromrežah nam je omogočila neposredno primerjavo med vsemi vzorci (slika 2).

Zaradi velike variabilnosti rezultatov, pridobljenih z mikromrežami, in zaradi velikega števila manjkajočih vrednosti je pri poskusih z mikromrežami pomembna uporaba ponovitev. V našem poskusu so bili poskusi na polju izvedeni v štirih ponovitvah, tako da smo imeli štiri biološke ponovitve. Poleg tega smo hibridizacijo mikromrež izvedli v laboratoriju dvakrat. Skupno smo hibridizirali 176 mikromrež.



Slika 2 Priprava univerzalne kontrole in potek hibridizacije mikromrež.

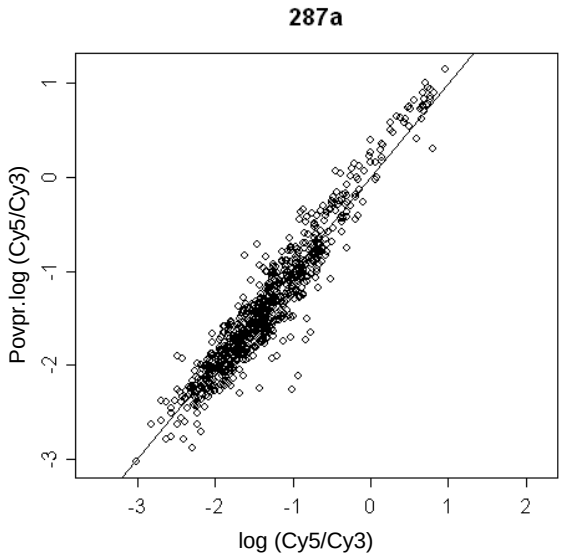
Analiza rezultatov

Analiza slike

Prvi korak analize rezultatov je analiza slike mikromreže po hibridizaciji. Slikanje mikromrež mora potekati z optimalnimi nastavitvami, da dosežemo optimalen signal posameznih točk mikromreže.[6] Sledi postavitev mreže, s katero določimo mesta posameznih točk mikromreže. Z računalniško analizo slike nato pridobimo podatke o intenziteti fluorescence obeh barvil (Cy5 in Cy3) na posameznih točkah mikromrež. Izmerimo tudi ozadje fluorescence, pri čemer smo se odločili za lokalno merjenje ozadja za vsako posamezno točko. Iz pridobljenih podatkov smo lahko izračunali razmerje med signalom in šumom.

Sledila je statistična analiza pridobljenih podatkov. Pred iskanjem diferencialno izraženih genov smo pripravili splošen pregled podatkov za oceno kvalitete naših poskusov. Analizirali smo intenziteto signala barvil Cy3 in Cy5, ozadje obeh

barvil ter frekvenco pozitivnih točk za posamezno hibridizacijo. Razpršenost podatkov za vsako posamezno mikromrežo ter odstopanje podatkov od povprečja celotnega poskusa smo preverili z uporabo razsevnih grafikonov (slika 3).



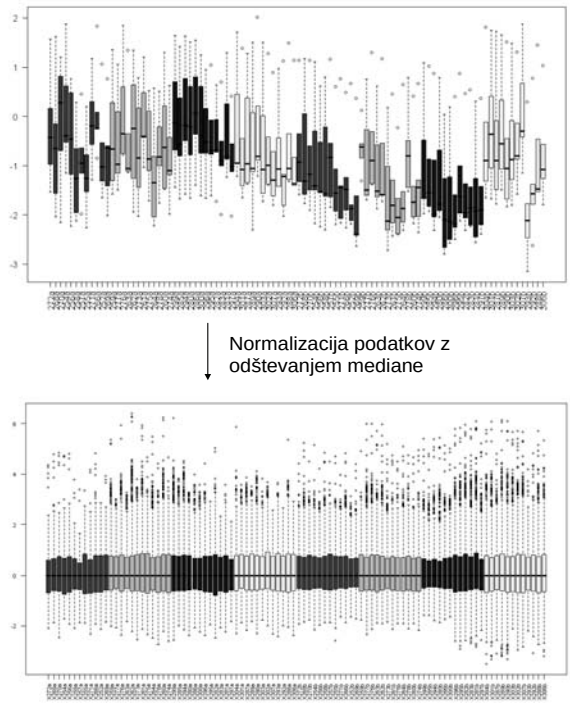
Slika 3 Razsevni grafikon za eno od analiziranih mikromrež.

Filtriranje podatkov in normalizacija

Pred nadaljnjo obdelavo podatkov smo želeli iz analize izločiti točke mikromrež, pri katerih signal ni bil zadosten. Odločili smo se za izločitev genov, pri katerih je bilo razmerje med signalom in šumom manjše od tri. Filter smo pripravili za obe barvili – Cy3 in Cy5. Povprečna intenziteta fluorescence se med posameznimi mikromrežami po hibridizaciji razlikuje, zato smo podatke normalizirali z odštevanjem mediane vrednosti posamezne mikromreže (slika 4).

Iskanje podobnosti med vzorci in identifikacija diferencialno izraženih genov

Za analizo podatkov, pridobljenih z mikromrežami, je možnih več postopkov, ki temeljijo na statističnih analizah. Podatke, pridobljene z mikromrežami, lahko uporabimo za iskanje podobnosti med vzorci, pri čemer upoštevamo vrednosti izražanja za veliko število genov. Po drugi strani pa želimo analizirati posamezne gene ter poiskati gene, ki so se različno odzvali med različnimi skupinami.

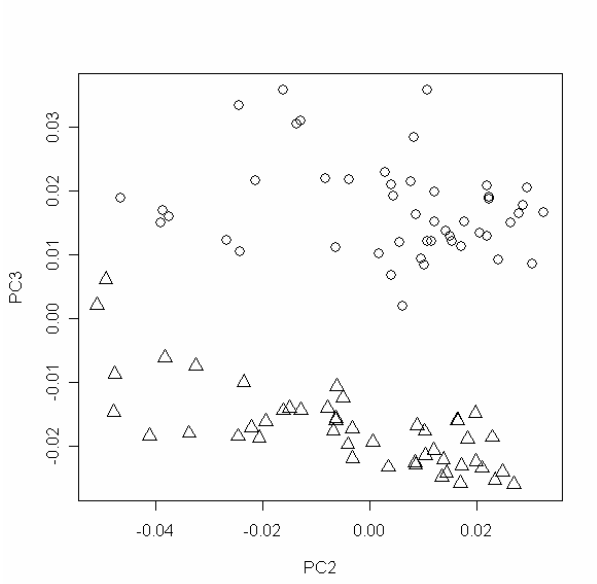


Slika 4 Normalizacija podatkov z odštevanjem mediane. Vsak izmed zabojev z ročaji predstavlja eno izmed mikromrež poskusa "krompirjeva plesen". Od vseh podatkov na tej mikromreži smo odšteli mediansko vrednost te mikromreže. S tem omogočimo primerljivost med podatki za različne vzorce.

Analiza glavnih komponent

Normalizirane podatke smo najprej analizirali z analizo glavnih komponent (principal component analysis, PCA).[7] Po tej metodi skupin ne določimo vnaprej, ampak (poenostavljeno povedano) iščemo komponente, s katerimi lahko razložimo kar največ variabilnosti v našem naboru podatkov, ter iščemo povezave med spremenljivkami (slika 5).

Primerjalno je bila izvedena tudi analiza glavnih koordinat (principal coordinate analysis), s katero smo dobili podobne rezultate.



Slika 5 Popolna ločitev sort krompirja za poskus "krompirjeva plesen" po metodi glavnih komponent. Sorta Lady Balfour je označena s krogi, sorta Sante pa s trikotniki.

ANOVA

Z analizo variance smo nato iskali statistično značilne razlike v izražanju posameznih genov za načrtno variirane spremenljivke (sorto krompirja, gnojilo, protiglivno sredstvo idr.). Zaradi kompleksne zasnove poskusa smo uporabili model analize variance, ki v prvem nivoju upošteva vpliv bioloških in tehničnih ponovitev, v drugem nivoju pa smo iskali značilne razlike med skupinami glede na variirane spremenljivke ter iskali morebitne interakcije med spremenljivkami. Z analizo variance smo tako pridobili podatke o genih, ki so se najbolj značilno razlikovali med posameznimi tretmaji. Za posamezne gene smo nato lahko natančno proučili profile izražanja pri različnih vzorcih.

Obravnavanje manjkajočih vrednosti

Pri analizi rezultatov mikromrež poseben problem predstavlja obravnavanje manjkajočih vrednosti. Pred analizo želimo odstraniti čimveč nezanesljivih rezultatov, ki bi lahko v končne rezultate vnašali pristranost. Manjkajoče vrednosti predstavljajo problem za statistične analize, saj na primer metoda glavnih komponent (v osnovni obliki) ne deluje, če so med podatki manjkajoče vrednosti, analiza variance pa predpostavlja uravnoteženost nabora podatkov. Manjkajoče vrednosti lahko nadomestimo z enotno vrednostjo (npr. srednjo vrednostjo izraženosti posamezne mikromreže), lahko jih med analizo zanemarimo ali pa uporabimo eno od metod za nadomeščanje manjkajočih vrednosti.

Iskanje biološkega pomena rezultatov

Po končani statistični analizi sledi iskanje biološkega pomena dobljenih rezultatov. Zanima nas, kakšno funkcijo imajo diferencialno izraženi geni, ter ali sodelujejo pri istih bioloških procesih. Primer programa, ki omogoča preslikavo ekspresijskih podatkov na metabolne poti, je program MapMan[8], ki je prilagojen tudi za nekatere rastlinske vrste. Rezultate lahko primerjamo tudi z zbirkami podatkov mikromrež za isti oziroma soroden organizem, pri katerem so bili izvedeni podobni poskusi, pri čemer nas zanima, ali pride do podobnega odziva genov.

Zaključki

S poskusom smo dobili širši vpogled v izražanje genov v različnih kmetijskih pogojih vzgoje. Pri tem je bil zelo pomemben dober eksperimentalni načrt, ki omogoča kasnejšo zanesljivo obdelavo podatkov. Uporaba univerzalne kontrole nam je omogočila primerjavo med velikim številom vzorcev ter olajšala analizo razlik med posameznimi spremenljivkami v vzorcu. Prav tako smo zanesljivost podatkov povečali z velikim številom ponovitev za posamezen vzorec.

Kljub temu je analiza tako velikega niza podatkov zelo zahtevna. Uporabljene metode za analizo mikromrež se med seboj dopolnjujejo. Z metodo glavnih komponent smo lahko videli, kateri vzorci tvorijo skupine, z analizo variance pa smo analizirali posamezne gene ter ugotovili, kateri geni se najbolj značilno razlikujejo med

proučevanimi skupinami vzorcev. S kombinacijo metod smo tako našli diferencialno izražene gene, ki jih nameravamo v prihodnosti še potrditi z metodo PCR v realnem času.

Podatki, pridobljeni s hibridizacijo mikromrež, predstavljajo trd oreh za statistične analize zaradi velike variabilnosti, ki nastane zaradi tehničnih razlik med posameznimi hibridizacijami. Zanesljivost rezultatov je omejena tudi z majhnim številom ponovitev v primerjavi z velikim številom analiz, ki jih lahko izvedemo v enem poskusu. Velik problem pa predstavlja tudi veliko število manjkajočih vrednosti. V prihodnosti bo zato potrebna uvedba metod, ki boljše delujejo na takšnih nizih podatkov. Pri analizi podatkov mikromrež je pomembno povezovanje znanja iz bioloških ved z znanji iz statistike in računalništva.

Literatura

1. Tilstone C: DNA microarrays: vital statistics. *Nature* 2003;424 (6949) :610-12.
2. Aharoni A, Vorst O: DNA microarrays for plant functional genomics. Plant Mol.Biol. 2002;48 (1-2): 99-118.
3. Kok EJ, Kuiper HA: Comparative safety assesment for biotech crops. *Trends biotechnol.* 2003; 21(10): 439-444.
4. Kuiper HA, Kok EJ, Engel KH: Exploitation of molecular profiling techniques for GM food safety assessment. *Curr Opin Biotechnol.* 2003; 14(2): 238-43.
5. Knudsen S: Guide to analysis of DNA microarray data New York 2004: John Wiley & sons, Inc.
6. Leung YF, Cavalieri D: Fundamentals of cDNA microarray data analysis. Trends Genet. 2003; 19(11):649-59.
7. Yeung KY, Ruzzo WL: Principal component analysis for clustering gene expression data. *Bioinformatics* 2001; 17(9): 763-74.
8. Thimm O, Blasing O, Gibon Y, et al.: MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 2004; 37: 914-939.

Research Paper ■

Application of closed itemset mining for class labeled data in functional genomics

Petra Kralj, Ana Rotter, Nataša Toplak, Kristina Gruden, Nada Lavrač, Gemma C. Garriga

Abstract. This paper applies a recently introduced methodology of closed itemset mining for class labeled data to potato microarray data. The study shows the discovered rules that best distinguish between virus resistant and virus sensitive transgenic potato lines. The discovered rules are interpretable and meaningful to domain experts.

■ **Infor Med Slov**: 2006; 11(1): 40-45

Authors' institutions: Institut Jožef Stefan, Ljubljana, Slovenia (PK, NL), National Institute of Biology, Ljubljana, Slovenia (AR, NT, KG), Nova Gorica Polytechnic, Nova Gorica, Slovenia (NL), Universitat Politecnica de Catalunya, Barcelona, Spain (GCG).

Contact person: Petra Kralj, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. email: petra.kralj@gmail.com.

Introduction

Microarray technology offers researchers the ability to simultaneously examine expression levels of hundreds or thousands of genes in a single experiment. Knowledge about gene regulation and expression can be gained by dividing samples into control samples, in our case mock infected plants, and treatment samples, in our case virus infected plants. Studying the differences between gene expression of the two groups (control and treatment) can provide useful insights into complex patterns of host relationships between plants and pathogens. [1]

Since the dimensions of microarrays are typically very large, statistical and data mining methods have to be used in order to draw significant conclusions from the data. Careful data preprocessing has to be done before using statistics or data mining. Data preprocessing includes, but is not limited to, filtering of data, leaving out low intensity signals or high background (noisy) signals. At later preprocessing stages, irrelevancy filtering[2] can also be used.

The task of data mining on microarray data differs from traditional data mining tasks because microarray domains are characterized by very large numbers of attributes (genes) relative to the number of examples (observations, samples). Standard classification rule learning algorithms do not perform well on microarray data because of this dimensionality problem.

This work applies a recently developed approach named RelSets[3] to microarray data. The approach[3] uses closed itemset mining to detect relevant rules of the form

IF Conditions THEN Class

from class labeled data. First, all frequent closed itemsets are found on data instances labeled as positive. In the second phase, itemsets that would form irrelevant rules are removed. Only relevant rules are returned.

In our study, we aimed to find differences between two classes of resistance in four transgenic potato lines. For this purpose, 48 potato samples were used leading to 24 microarrays. We applied the algorithm RelSets[3] on these microarray data. The resulting rules are meaningful to biology experts.

The paper is organized as follows. First, the algorithm for mining closed itemsets from class labeled data is reviewed. Next, the biological experiment is outlined with the description of the data preparation steps. The data mining task is then outlined, followed by data mining results, their interpretation and conclusions.

Methodological background

The closed set for class labeled data technique[3] used in our experiment is based on the theory of closed itemset mining[4], upgraded by the recently developed theory of relevancy. [2]

Closed itemsets

Searching for descriptions from data has been addressed in descriptive data mining, in particular association rule learning. [5] An innovative insight was provided by closure systems, [4] aimed at compacting the whole dataset into a reduced system of relevant sets of items that formally conveys the same information as the complete dataset.

Let E denote a set of training examples, described by a fixed set of features $F = \{f1, \ldots, fn\}$. Features are logical variables representing attribute-value pairs (called *items* in association rule learning). Each example e is represented as a tuple of features f from F with an associated class label.

From the point of view of data mining algorithms, closed itemsets are maximal sets among any other itemsets occurring in the same examples. Formally, let $supp(X)$ denote the number of examples where the itemset X is contained. Then, set $X \subseteq F$ is said to be a *closed itemset* when there is no other

set $Y \subseteq F$ such that $X \subset Y$ and $supp(X) = supp(Y)$. [4]

Feature and rule relevancy

The rule induction problem can be viewed as a process of searching the space of concept descriptions. In our case, the space of descriptions to be searched is the space of itemsets (conjunctions of features) that form rule conditions. Some descriptions in this hypothesis space may turn out to be more relevant than others for characterizing and/or discriminating the target class. [6]

A rule is said to *cover* an example if the condition part of the rule is satisfied for that example. A rule *correctly covers* an example if the rule covers the example and the predicted class of the rule matches the class label of the example. The rule *incorrectly covers* an example if the rule covers an example and the class of the rule differs from the class label of the example.

Quality of rule *R* is measured by *rule coverage*, determined by two quantities: the number of correctly covered examples *TP(R)* (*True Positives*) and the number of incorrectly covered examples *FP(R)* (*False Positives*). Good rules correctly cover many examples (many true positives) and incorrectly cover as few examples as possible (few false positives).

The notion of relative *relevancy of features*[2] can be generalized to apply to rules. [3]

Feature *f1* is *relatively irrelevant* with respect to feature *f2* if $TP(f1) \subseteq TP(f2)$ and if $FP(f2) \subseteq FP(f1)$. A feature is *relatively irrelevant* if there exists another feature in the dataset compared to which it is irrelevant.

The definition of feature relevancy can be generalized to rule relevancy as follows. Rule *R1* is *relatively irrelevant* with respect to rule *R2* if $TP(R1) \subseteq TP(R2)$ and if $FP(R2) \subseteq FP(R1)$. A rule is *relatively irrelevant* if there exists another rule in the ruleset compared to which it is irrelevant. [3]

The RelSets algorithm

A closed itemset mining algorithm and a rule relevancy filter are used in the approach applied in this paper. In this section we briefly recall the algorithm for closed itemset mining for labeled data, called RelSets. [3]

The input to RelSets is the dataset and one parameter: the minimum true positive count *(minTP)*. This is a constraint that implies that only rules that cover at least *minTP* positive examples should be constructed.

The dataset is first divided into two parts depending on the class label of the examples: the positive examples *P* and the negative examples *N*.

Closed itemsets in the positive examples are mined with a minimum support constraint *(minTP)*. These closed sets can be directly interpreted as rules:

IF Closed set THEN Positive

These rules have high true positives count since they were built with a *minTP* constraint. The theory[3] proves that these are all the most specific rules that have the potential to be relevant. Nothing is yet known about the coverage of false positives.

In the second phase RelSets confronts the rules found in the first phase with the negative data. It removes relatively irrelevant rules on the negative data. A maximum false positives count constraint can be imposed.

The RelSets algorithm returns all the relatively relevant rules which fulfill the minimum true positive count constraint. The algorithm is complete in the sense that it finds all the most specific rules satisfying the constraints. This is very appropriate for microarray data since not many

examples are available and the complete search of the space is very adequate.

Biological experiment

The goal of our experiment is to investigate differences between virus sensitive and resistant transgenic potato lines. Since potato cultivation is economically important worldwide, infection pathway research is motivated not only by scientists but also by the industry.

Four transgenic potato lines (two of them resistant and two of them sensitive to a viral infection) were tested. Plants from each transgenic line were divided into four groups: one half was infected with potato virus PVY^{NTN} and the other half was mock inoculated. One PVY^{NTN} inoculated group and one mock inoculated group of each transgenic line were harvested 8 hours and the rest 12 hours after the infection. Every experiment was repeated 3 times, thus yielding 48 samples. Each microarray was hybridized with a virus inoculated sample and mock inoculated sample from the same transgenic line, yielding to 24 microarray experiments (Figure 1). Depending on the individual microarray design, red or green labeling for different samples was used.

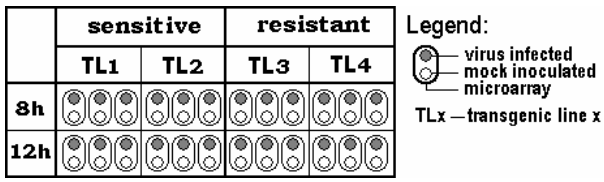


Figure 1 Schematic representation of the experiment. Each potato sample is represented by a circle: empty circles represent mock infected samples and full circles represent virus infected samples. An ellipse around two dots represents one microarray.

The dimensions of the initial data matrix after image scanning were 31200 x 24. As each gene is represented twice (as duplicate) on a microarray, the actual data dimensions are 15600 x 24. Data were filtered using the image analysis software ArrayPro Analyzer®. Spots that were unevenly distributed, had stained background (low signal-to-noise ratio) and low intensity signal on both channels (red and green) were left out of further analysis. After this first filtering, an average of 20,000 expression values per array remained for further analysis.

Two expression values for the same gene in a microarray were averaged. Second data filtering had two conditions: if 10 out of 24 microarray experiments for a given gene resulted in expression values in the interval (-0.3 , +0.3) or were missing values, the gene was discarded from further analysis. Both conditions for filtering were chosen arbitrarily to yield a suitable number of potentially regulated genes for further analysis. The dimensionality of data matrix was thus reduced to 6377 x 24.

Data mining task and results

The data mining task was to find differences in gene expression levels characteristic for virus sensitive potato transgenic lines, discriminating them from virus resistant potato transgenic lines and vice versa. For this purpose we used the RelSets[1] algorithm.

Our dataset contains 12 examples. Each example is a pair of microarrays (8 and 12 hours after infection) from the same transgenic line. All the data was discretized by using expert background knowledge. Features of the form $|$*gene expression value*$| > \pm 0.3$ were generated and enumerated.

Three groups of features were generated:

- the gene expression levels 8 hours after infection (feature numbers $\leq$ 12493)
- the gene expression levels 12 hours after infection (feature numbers between 12494 and 24965)
- the difference between gene expression levels 12 and 8 hours after infection (feature numbers $\geq$ 24966)

We ran our algorithm twice: once the sensitive examples were considered positive and once the resistant ones were considered positive. In both cases the constraint of minimal true positive count was set to 4, and in the first phase the algorithm returned 22 rules. The second part of the algorithm, which involves rule relevancy filtering, filtered the rules to just one relevant rule with true positive rate 100% and false positive rate of 0%. The results gained are shown below, where features are represented by numbers.

IF (13031 13066 19130 23462 24794 25509 29938 33795 33829 35003 35190 36266) THEN *sensitive* (TP=6)(FP=0)

Twelve features determined the potato sensitivity class for the samples used.

IF (16441 20474 20671 24030 25141 29777 30111 32459 33225 33248 33870 34108 34114 34388 37252 37484) THEN *resistant* (TP=6)(FP=0)

Sixteen features determined the potato *resistance* class for the samples used.

Biological interpretation

Based on the samples tested it seems that the response to the infection after 8 hours is not strong enough to distinguish between resistant transgenic lines and sensitive ones. None of the gene expression changes after 8 hours appeared significant to the data mining algorithm. However, gene expression levels after 12 hours and the comparison of gene expression difference (12-8) seem to determine the resistance to the infection with potato virus PVY^{NTN} for the transgenic lines tested.

According to the mechanism of virus plant interaction, genes that proved to be important for rule building, appearing in the output rules, have been classified into the following categories:

- *rec*: genes, whose products are responsible for sensing the infections by viral pathogen (receptors) and whose products are part of the cell membrane
- *sig*: genes, whose products are responsible for intracellular signaling transduction
- *TF*: genes, whose products are influencing transcription in cell nucleus
- *def*: genes, whose products are effectors for defense
- *hk*: housekeeping genes, whose expression was historically accepted as constant regardless of the physiological state of the plant
- *uf*: unknown function.

The distribution of the genes, appearing in class discriminating rules, determining whether a transgenic line is sensitive or resistant, can be viewed from Table 1. It can be argued that the downregulation of the first part of virus-plant interaction (reception and signaling) is an indicator of plant's sensitivity for infection, whereas upregulation of the second part (transcription and defense) of the interaction determines the resistance in plants tested.

Table 1 Functional distribution of genes, important to determine the sensitivity or resistance of a potato.

	sensitive	resistant
rec	1	0
sig	5	3
TF	2	4
def	7	6
hk	1	4
uf	1	3

The pattern can be visualized in Figure 2 and Figure 3. Housekeeping genes for which expression levels were argued to remain unchanged regardless of the treatment given to the plant seem to be important for determining the resistance of samples tested (Table 1).

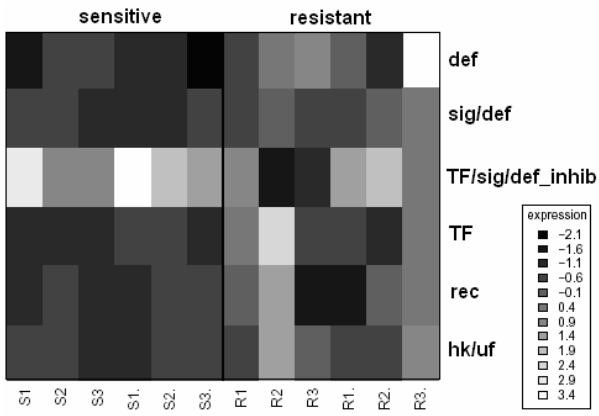


Figure 2 Heatmap for sensitive transgenic lines. Genes that have been found to be important for determining the sensitivity in samples tested were grouped into parts of plant-host interaction pathway: rec, sig, TF, def, hk, uf and their combinations if found. The heatmap shows that most of the genes of sensitive transgenic lines (marked with S, left side of the heatmap) were downregulated. Products of upregulated genes are inhibitors, important for signaling and defense pathways.

Conclusions

Using data mining on microarray data is a challenge because of the unusual dimensionality of the data specific for these domains. The recently developed method for closed itemset mining for labeled data shows no difficulties when applied to this kind of data. Furthermore, the results are in a form of comprehensible rules that are easy to be interpreted by domain experts. The approach was proven to perform well on microarray data, where the goal was to find differences between virus resistant and virus sensitive potato transgenic lines. The results proved to be meaningful to domain experts.

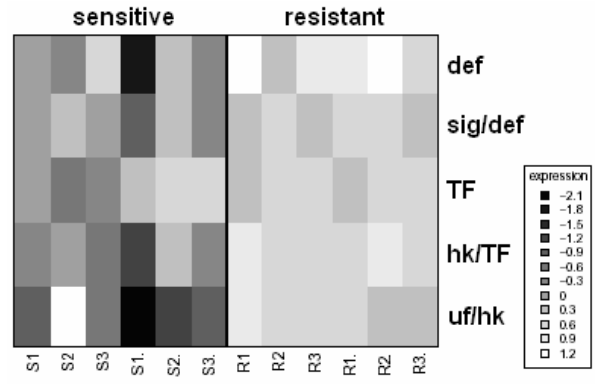


Figure 3 Heatmap for resistant transgenic lines. The heatmap shows that genes that have been found to be important of determining resistance in sample tested were upregulated (right side of the heatmap). Genes were grouped into parts of plant-host interaction pathway: rec, sig, TF, def, hk, uf and their combinations if found.

Literature

1. Taiz L , Zeiger E: Plant physiology, second edition 1998;(372:374). Sinauer Associates.
2. Lavrač N, Gamberger D: Relevancy in constraint-based subgroup discovery. In Boulicaut JF, De Raedt L, Mannila H (eds.) Constraint-Based Mining and Inductive Databases. Lecture Notes in Computer Science 2004: Springer, 243-266.
3. Garriga GC, Kralj P, Lavrač N: Subgroup discovery by closed itemset mining from labeled data, Jožef Stefan Institute Technical Report, No. 9351, December 2005.
4. Carpineto C, Romano G: Concept Data Analysis: Theory and Applications. 2004: Wiley.
5. Agrawal R, Srikant R: Fast algorithms for mining association rules. In Proceedings of the 20th International Conference on Very Large Databases 1994: 207-216.
6. Gamberger D, Lavrač N, Železny F, et al.: Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. Journal of Biomedical Informatics 2004; (37): 269-284.

Research Paper ■

Subgroup discovery: An experiment in functional genomics

Nada Lavrač, Dragan Gamberger

Abstract. Functional genomics is a typical scientific discovery domain characterized by a very large number of attributes (genes) relative to the number of examples (observations). This work presents an approach to subgroup discovery in supervised inductive learning of short rules that are appropriate for human interpretation. The approach is based on the subgroup discovery rule learning framework, enhanced by methods of restricting the hypothesis search space by exploiting the relevancy of features that enter the rule construction process as well as their combinations that form the rules. A multi-class functional genomics problem of classifying fourteen cancer types based on more than 16000 gene expression values is used to illustrate the methodology.

■ **Infor Med Slov**: 2006; 11(1): 46-51

Authors' institutions: Jožef Stefan Institute, Ljubljana, Slovenia (NL), Rudjer Bošković Institute, Zagreb, Croatia (DG).

Contact person: Nada Lavrač, Department of Knowledge Technologies, Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia. email: nada.lavrac@ijs.si.

Introduction

Construction of understandable and explainable models is important for scientific discovery as well as for the generation of actionable knowledge. It is possible to extract the most informative features or attributes from complex classifiers (the attributes with this property are called disease markers) but logical connections among these features or attributes are missing. This disables the construction and expert interpretation of models describing the target class. In contrast, short rules, despite being potentially less accurate than the complex classifiers, are much more appropriate for scientific discovery tasks in which the interpretability of induced models is of ultimate importance.

Functional genomics is a typical scientific discovery domain characterized by a very large number of attributes (genes) relative to the number of examples (observations). The danger of data overfitting is crucial in such domains. This work presents an approach to subgroup discovery, complemented by an approach which can help in avoiding data overfitting in supervised inductive learning of short rules that are appropriate for human interpretation. The approach is based on the subgroup discovery rule learning framework, enhanced by methods of restricting the hypothesis search space by exploiting the relevancy of features that enter the rule construction process as well as their combinations that form the rules.

This paper presents an approach, based on the subgroup discovery rule learning framework, enhanced by a method for filtering of irrelevant features. The results of its application on a multi-class functional genomics problem, aimed at classifying fourteen cancer types based on more than 16000 gene expression values, illustrate the use of the proposed methodology.

Subgroup Discovery

Subgroup discovery is a form of supervised inductive learning of subgroup descriptions for the target class in a two class domain. The descriptions have the form of rules built as logical conjunctions of features. Features are logical conditions that have values true or false, depending on the values of attributes which describe the examples in the problem domain. Subgroup discovery rule learning is therefore a form of two-class propositional inductive rule learning. Multi-class problems can be solved as a series of two-class learning problems, so that each class is once selected as the target class while examples of all other classes are treated as non-target class examples.

Formally, the task of subgroup discovery is defined as follows: given a population of individuals and a specific property of the individuals that we are interested in, find population subgroups that are 'most interesting', e.g., are as large as possible and have the most unusual distributional characteristics with respect to the property of interest.

Standard classification rule learning algorithms can be adapted to perform subgroup discovery. For instance, subgroup discovery algorithms, CN2-SD[1] and Apriori-SD[2] are adaptations of classification rule learners: CN2-SD is an adaptation of CN2[3] and Apriori-SD is an adaptation of APRIORI-C[4] and APRIORI[5]. These algorithms take as input the training examples described by discrete attribute values.

Method

In this work, subgroup discovery is performed by the SD algorithm[6,7], implemented in the on-line Data Mining Server (DMS), publicly available at http://dms.irb.hr, a relatively simple iterative beam search rule learning algorithm.

The input to SD consists of a set of examples E (E=P U N, P is the set of target class examples, N the set of non-target class examples) and a set of features F constructed for the given example set. For discrete (categorical) attributes, features have the form Attribute = value, while for continuous (numerical) attributes they have the form Attribute > value or Attribute < value. The output of the SD algorithm is a set of rules with optimal covering properties on the given example set. As in classification rule learning, an induced rule (subgroup description) has the form of a (backwards) implication:

$$Class \leftarrow Cond.$$

In terms of rule learning, the property of interest for subgroup discovery is the target class (Class) that appears in the rule consequent, and the rule antecedent (Cond) is a conjunction of features (attribute-value pairs) selected from the features describing the training instances.

A rule with ideal covering properties is true for all target class examples and not true for all non-target class examples. Target class examples covered by a rule are also called true positives, TP, while non-target class examples covered by the rule are called false positives, FP. All remaining non-target class examples not covered by the rule are called true negatives, TN. An ideal rules has TP=P and TN=N. In the proposed subgroup discovery approach, the following rule quality measure q is used in heuristic search of rules:

$$q = |TP| / (|FP| + g)$$

where g is a user defined generalization parameter. High quality rules will cover many target class examples and a low number of non-target examples. The number of tolerated negative examples, relative to the number of covered target class cases, is determined by parameter g.

The flexibility of subgroup discovery is due to its search of rules that satisfy groups of examples of the target class, not necessary excluding all of the non-target examples. Sizes of subgroups are not defined in advance but the algorithm tends to make them as large as possible. Due to this flexibility the algorithm is able to incorporate different rule relevancy methods with the goal to prevent the construction of target class subgroup descriptions which do not have sufficient supportive evidence for being significantly different from non-target samples. An equally important part of the methodology for avoiding overfitting is that each feature that enters the subgroup discovery algorithm should itself be a relevant target class descriptor.

Relevancy of features

The relevancy of features is determined by a combination of methods for restricting the hypothesis search space and for eliminating features with low covering properties. The later methods based on absolute and relative relevancy are universally applicable to any domain and their use is suggested in all feature based inductive learning tasks. The restrictions of the hypothesis search space are related to the form of rules and to the properties of the domain. In this section we present an effective approach that can strongly reduce the number of features and its application is suggested for descriptive induction tasks in gene expression domains.

The features are restricted to simple forms only, because their complex forms may enable that, despite testing feature covering properties, features with insufficient supportive evidence may enter the rule construction process. For example, for discrete attributes the simple features have the form $Ai=a$. No complex logical forms like ($Ai=a$ & $Aj=b$) or ($Ai=a$ V $Aj=b$) are acceptable. The first form is not needed as all potential conjunctions are tested by the beam search procedure of the subgroup discovery algorithm. The second form is dangerous because, for example, the feature $Ai=a$ may be relevant while the feature $Aj=b$ may be irrelevant. Their combination $Ai=a$ V $Aj=b$ may be even more relevant than $fi=a$ itself, which may cause that

condition *fj=b* may be included into the finally constructed rules while its inclusion is not justified by its covering properties on the training set. Notice that if both conditions *fi=a* and *fj=b* are relevant, it does not mean that by restricting the form of used features some important logical combinations of features will be ignored. In the subgroup discovery approach both features can build separate subgroup descriptions and - if they are relevant - they both have a chance to appear in the final set of induced rules.

Results

The gene expression domain, described by Ramaswamy et al.[8] and Gamberger et al.,[9] and used in our experiments, is a domain with 14 different cancer classes and 144 training examples in total. Eleven classes have 8 examples each, two classes have 16 examples and only one has 24 examples. The examples are described by 16063 attributes presenting gene expression values. The domain can be downloaded from http://www-genome.wi.mit.edu/cgi-bin/cancer/datasets.cgi . There is also an independent test set with 54 examples.

Gene expression scanners measure signal intensity as continuous values which form an appropriate input for data analysis. The problem is that for continuous valued attributes there can be potentially many boundary values separating the classes, resulting in many different features for a single attribute. Another possibility is to use presence call (signal specificity) values computed from measured signal intensity values by the Affymetrix GENECHIP software. In all experiments we used only the presence call values. The presence call has discrete values A (absent), P (present), and M (marginal). The M value can be interpreted as a *do not know* state, so for every attribute there are only two distinct features *Attribute = A* and *Attribute = P* generated for each attribute. The reason is that features presented by conditions like *Ai* is true (*Ai* is present) or *Aj* is false (*Aj* is absent) are very natural for human interpretation. A more important reason for using GENECHIP presence call values (instead of continuous signal intensity values) is that the approach can help in avoiding overfitting, as the feature space is very strongly restricted: instead of many features per attribute we have only two. Also, as the measured gene expression values are not completely reliable (which is reflected by the fact that for the same sample measured values may change from one measurement to another), some robustness of constructed rules is welcome, which is achieved by treating the marginal presence call attribute value M as a *do not know* state. The value can neither be used to support the relevancy of a feature or a rule, nor it can be used for prediction purposes. In this way it additionally restricts the hypothesis search space.

The experiments were performed separately for each cancer class so that a two-class learning problem was formulated where the selected cancer class was the target class and the examples of all other classes formed non-target class examples. In this way the domain was transformed into 14 inductive learning problems, each with the total of 144 training examples and with between 8 and 24 target class examples. For each of these tasks a complete procedure consisting of feature construction, elimination of irrelevant features, and induction of subgroup descriptions in the form of rules was repeated. Finally, using the SD subgroup discovery algorithm, for each class a single rule with maximal q value has been selected, for q being the heuristic of the SD algorithm, and g being equal 5 in all experiments presented in this work. The rules for all 14 tasks consisted of 2-4 features. The induced rules were tested on the independent example set. The procedure was repeated for all 14 tasks with the same default parameter values and tested on an independent test set. The results are presented in Table 1.

The table presents measured covering properties both on the training set and on the test set. Although the obtained covering values on the training sets are very good, the measured prediction quality on the test sets is for many

classes very low, significantly lower than those reported by Ramaswamy et al.[8] For 7 out of 14 classes the measured precision on the test sets is 0%. But from the table an interesting and important relationship between prediction results on the test set and the number of target class examples in the training set can be noticed. There are very large differences among the results on the test sets for various classes (diseases) and the precision higher than 50% has been obtained for only 5 out of 14 classes. There are only three classes (lymphoma, leukemia, and CNS) with more than 8 training cases and all of them are among those with high precision on the test set, while for only two out of eleven classes with 8 training cases (colorectal and mesothelioma) high precision was achieved.

Table 1 Covering properties on the training and on the independent test set for rules induced for 14 classes. Sensitivity is $|TP|/|P|$, specificity is $|TN|/|N|$, while precision is defined as $|TP|/(|TP|+|FP|)$.

Cancer	Training set			Test set		
	Sens.	Spec.	Prec.	Sens.	Spec.	Prec.
breast	5/8	136/136	100%	0/4	49/50	0%
prostate	7/8	136/136	100%	0/6	45/48	0%
lung	7/8	136/136	100%	1/4	47/50	25%
colorectal	7/8	136/136	100%	4/4	49/50	80%
lymphoma	16/16	128/128	100%	5/6	48/48	100%
bladder	7/8	136/136	100%	0/3	49/51	0%
melanoma	5/8	136/136	100%	0/2	50/52	0%
uterus_adeno	7/8	136/136	100%	1/2	49/52	25%
leukemia	23/24	120/120	100%	4/6	47/48	80%
renal	7/8	136/136	100%	0/3	48/51	0%
pancreas	7/8	136/136	100%	0/3	45/51	0%
ovary	7/8	136/136	100%	0/4	47/50	0%
mesothelioma	7/8	136/136	100%	3/3	51/51	100%
CNS	16/16	128/128	100%	3/4	50/50	100%

The classification properties of rules induced for classes with 16 and 24 target class examples (lymphoma, leukemia and CNS, presented below) are comparable to those reported by Ramaswamy et al.,[8] while the results on eight small example sets with 8 target examples were poor.

The following rule was found for the lymphoma class:

Lymphoma ← CD20_receptor EXPRESSED AND phosphatidylinositol_3_kinase_regulatory_alpha_subunit NOT_EXPRESSED.

For the leukemia class, we have the following rule:

Leukemia ← KIAA0128_gene EXPRESSED AND prostaglandin_d2_synthase_gene NOT EXPRESSED.

The best-scoring rule for the lymphoma class contains a feature corresponding to a gene routinely used as a marker in diagnosis of lymphomas (CD20), while the other part of the conjunction (the PI3K gene) seems to be a plausible biological co-factor. The best-scoring rule for the leukemia class contains a gene whose relation to the disease is directly explicable (Septin 6).

Lastly, we address the rule found for the CNS class:

CNS ← fetus_brain_mRNA_for_membrane_glycoprotein_M6 EXPRESSED AND CRMP1_collapsin_response_mediator_protein_1 EXPRESSED.

Conclusion

For larger training sets the subgroup discovery methodology enabled effective construction of relevant knowledge. The result, illustrated in Figure 1, demonstrates that mean values of rule sensitivity and precision are significantly higher for three tasks with 16 and 24 target class examples than for eleven tasks with only 8 target class examples. The mean values for the specificity are also higher but they were over 95% already for small target class sets.

The induced rules for lymphoma, leukemia and CNS were evaluated by a domain expert and most of features used in them were recognized as known disease markers for the target class cancers.[9] Expert evaluation proved the relevancy of induced rules. Both good prediction results on an independent test set as well as expert interpretation of induced rules show the

effectiveness of described methods for avoiding overfitting in scientific discovery tasks. Mostly bad results for tasks with only eight target class examples demonstrate that the methods can not be successful in all situations, especially those with a very small number of examples.

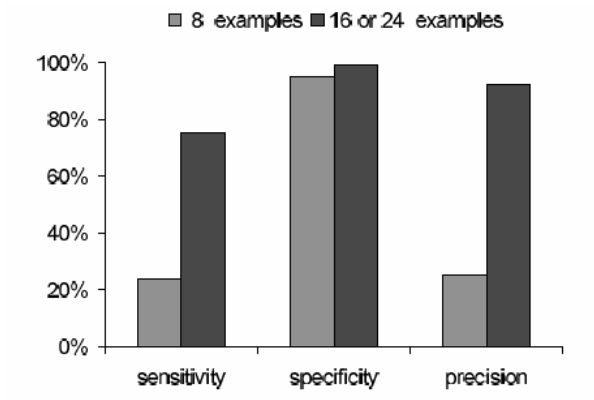


Figure 1 Mean values of sensitivity, specificity, and precision measured on the independent test set versus the number of target class cases in the training set.

In spite of the number of findings in agreement with the bio-medical state-of-the-art, discovery of known factors in the considered malignancies was not the ultimate goal of this study. The main goal of the methodology is the discovery of unknown and never thought-off relationships, in a form instantly understandable to an expert. The presented experiments have succeeded in discovering human understandable rules, some of which have uncovered interesting regularities in the data.

Literature

1. Lavrač N, Kavšek B, Flach P, et al.: Subgroup discovery with CN2-SD. Journal of Machine Learning Research 2004; 5: 153-188.
2. Kavšek B, Lavrač N: APRIORI-SD: Adapting association rule learning to subgroup discovery. In: Proceedings of the 5th International Symposium on Intelligent Data Analysis 2003; 230-241, Springer.
3. Clark P, Niblett T: The CN2 induction algorithm. Machine Learning 1989; 3(4):261-283.
4. Jovanovski V, Lavrač N: Classification rule learning with APRIORI-C. In Progress in Artificial Intelligence: Proceedings of the 10th Portuguese Conference on Artificial Intelligence 2001; 44-51, Springer.
5. Agrawal R, Srikant R: Fast algorithms for mining association rules. In: Proceedings of the 20th International Conference on Very Large Databases 1994; 207-216.
6. Gramberger D, Lavrač N: Expert-guided subgroup discovery: methodology and application. Journal of Artificial Intelligence Research 2002; 17:501-527.
7. Gamberger D, Krstačić A, Krstačić G, et al.: Data analysis based on subgroup discovery: experiments in brain ischaemia domain. In Proceedings of the 10th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology 2005; 52-56, University of Aberdeen.
8. Ramaswamy S, et al.: Multiclass cancer diagnosis using tumor gene expression signitures. In Proc. Natl. Acad. Sci USA 2001; 98(26): 15149-15154.
9. Gamberger D, Lavrač N, Železny F, et al.: Induction of comprehensible models for gene expression datasets by subgroup discovery methodology. Journal of Bioinformatics 2004; 37: 269-284.

Izvirni znanstveni članek ■

Odkrivanje pravil uravnavanja izražanja genov z razvrščanjem na podlagi pravil

Tomaž Curk, Blaž Zupan, Uroš Petrovič, Gad Shaulsky

Izvleček. Zapis in struktura regulatornih regij genov pogojujeta program izražanja genov v odzivu celice na notranje in zunanje dražljaje. Ključen predpogoj za uspešno eksperimentalno potrditev in razumevanje programov uravnavanja izražanja genov je računalniško odkrivanje relacij oziroma pravil, ki opisujejo povezavo med izmerjenim izražanjem in strukturo regulatorne regije gena. Težava pri tovrstnih analizah je njihova izjemna kombinatorična kompleksnost; možnih pravil, ki jih je potrebno pri analizi preveriti, je zelo veliko. Navadno imamo namreč na razpolago mnogo potencialnih vezavnih mest transkripcijskih faktorjev, s katerimi je možno opisati strukturo regulatornih regij. V članku opisujemo metodo, ki z uporabo hevrističnih pristopov gradi omenjena pravila in predlagamo različne načine predstavitve rezultatov tovrstne analize.

Rule-based clustering for discovery of patterns in gene expression regulation

Abstract. The genetic response programs of cells to their internal state and outside environment are predominately determined by the sequence and structure of gene regulatory regions. Computational discovery of relations between gene expression, sequence and structure of regulatory regions is a prerequisite for experimental validation and a successful understanding of such programs. Given a large base of regulatory elements (*i.e.* putative or known transcription factor binding sites) which can be used to infer rules, the main obstacle posed is the high combinatorial explosion of possible rules which need to be tested. The rule-based clustering method we developed, combined with an effective presentation of discovered rules, can successfully handle this combinatorial problem.

■ **Infor Med Slov**: 2006; 11(1): 52-59

Institucije avtorjev: Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Slovenija (TC, BZ), Baylor College of Medicine, Houston, USA (BZ, GS), Institut Jožef Stefan, Ljubljana, Slovenija (UP).

Kontaktna oseba: Tomaž Curk, Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Tržaška 25, SI-1001, Ljubljana. email: tomaz.curk@fri.uni-lj.si.

Uvod

Temeljni korak na poti določitve in razumevanja mehanizmov ter programov uravnavanja genov je analiza regulatornih regij genov.[1] Regulatorne regije so deli zapisa DNA, na katere se vežejo posebni proteini, imenovani transkripcijski faktorji, ki vzbujajo ali zavirajo transkripcijo gena v bližini regulatorne regije. Vezava transkripcijskih faktorjev je le eden izmed načinov uravnavanja izražanja genov in posledično tvorbe proteinov. Na izražanje namreč vplivajo tudi uravnavanje na nivoju strukture in oblike kromatina, epigenetski učinki, post-transkripcijsko uravnavanje, translacijsko in post-translacijsko uravnavanje ter drugi nivoji uravnavanja.[1] Ker je podatkov o slednjih zelo malo, se večina študij osredotoča na postopke iskanja povezav med vsebino regulatornega zaporedja DNA in izmerjenim izražanjem gena. Tehnologija DNA mikromrež omogoča vzporedno merjenje izražanja genov celotnega genoma, vendar so tako dobljene meritve dokazano nezanesljive.[2] Zato je nujno potrebna previdnost pri interpretaciji dobljenih rezultatov. Pomembnejše rezultate lahko na primer še dodatno preverimo z uporabo zanesljivejših metod (kot je na primer metoda PCR), ki pa običajno ne dovoljujejo hkratnega merjenja velikega števila genov.[3]

Prvi korak analize odnosov med genskim zaporedjem in njegovim izrazom je določitev regulatorne regije in vezavnih mest. Regulatorna regija se navadno nahaja v neposredni bližini kodirajočega dela gena ali pa se celo z njim prepleta. Področji se ločita v pogostosti posameznih nukleotidov ter pogostosti zaporednih trojk nukleotidov (kodonov), ki v kodirajočem delu gena določajo zaporedje aminokislin končnega proteina. Te in podobne lastnosti regulatornih in kodirajočih regij s pridom uporabljajo algoritmi za napovedovanje regulatornih regij.[4] Ker se regulatorna področja navadno ne prepisujejo v mRNA, je drugi, posredni in eksperimentalni način določanja regulatornih regij odkrivanje delov zaporedja DNA, ki se sploh kdaj prepišejo v mRNA (ang. *EST – expressed sequence tags*). Področja v bližini genov, ki se nikoli ne prepišejo, so kandidati za regulatorne regije.

Naslednji korak analize je določitev potencialnih vezavnih mest transkripcijskih faktorjev v odkritih regulatornih regijah. Vezavno mesto je navadno krajše zaporedje (4 do 20 nukleotidov),[1] ki je v manjših variacijah posameznih nukleotidov ohranjeno v regulatornih regijah reguliranih genov. Za računalniško obdelavo je najbolj primerna predstavitev vezavnega mesta v obliki matrike pogostosti nukleotidov na posamezni poziciji zaporedja, kar lahko predstavimo tudi grafično, v obliki tako imenovanih "logo-v" (tabela 1 in slika 1). Tovrstna, eksperimentalno potrjena zaporedja, najdemo tudi v javnih bazah podatkov, katere primer je baza TRANSFAC.[5] V primeru, da analiziramo gene s (še) neznanimi regulatorji oziroma neznanimi vezavnimi mesti, je možno uporabiti programska orodja, ki z lokalno ali globalno poravnavo zaporedij poiščejo pogosta, krajša podzaporedja,[6] ki jih v nadaljnji analizi lahko obravnavamo kot potencialna vezavna mesta. Podroben opis in primerjavo tovrstnih orodij podaja pregledni članek Tompa in sodelavcev.[7]

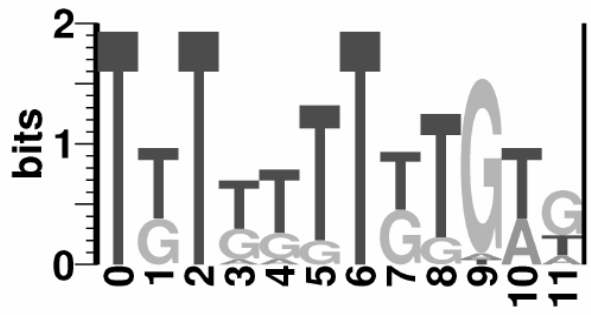
Tabela 1 Matrična predstavitev vezavnega mesta.

	0	1	2	3	4	5	6	7	8	9	10	11
A	0	0	0	0.2	0.1	0	0	0	0	0.1	0.6	0.1
C	0	0	0	0	0	0	0	0	0	0	0	0
G	0	0.5	0	0.4	0.4	0.1	0	0.8	0.4	0.8	0	0.6
T	1.0	0.5	1.0	0.4	0.5	0.9	1.0	0.2	0.6	0.1	0.4	0.3

Pojasnilo: Prikazane so frekvence štirih nukleotidov A, C, G in T na posamezni poziciji (od 0 do 11) potencialnega vezavnega mesta.

Večina postopkov namenjenih preučevanju povezave med prisotnostjo potencialnih vezavnih mest transkripcijskih faktorjev v regulatornih regijah in izmerjenim izražanjem genov temelji na začetnem razvrščanju genov v skupine (ang. *clustering*) na podlagi izražanja, funkcije ali drugega kriterija. Razvrščanju sledi iskanje vezavnih mest značilnih za posamezno skupino genov.[8,9] Uspeh

teh pristopov je močno odvisen od števila skupin, kar je navadno parameter metode razvrščanja, ki ga mora uporabnik podati vnaprej. Rahlo spremenjeni začetni pogoji ali drugače izbran prag pri razvrščanju v skupine lahko privede do popolnoma različnih skupin in posledično do popolnoma drugačnega nabora na ta način odkritih značilnih vezavnih mest v posamezni skupini. Druga, večja pomanjkljivost teh pristopov je, da se osredotočajo samo na izključujoče se podskupine genov, čeprav je znano, da se isti gen lahko odziva na več načinov oziroma opravlja več funkcij. Primer takšnega gena podajamo v razdelku z eksperimentalnimi rezultati.



Slika 1 Enostaven in razumljiv grafični prikaz vezavnega mesta določenega v tabeli 1. Prikazana je ohranjenost nukleotidov na posamezni poziciji vezavnega mesta.

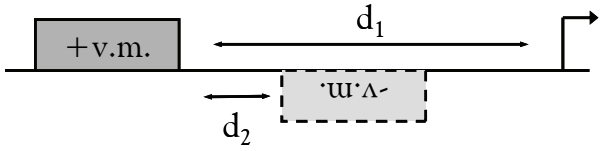
Komplementaren pristop zgoraj opisanemu prične s podatki o vezavnih mestih, ter nato išče takšne opise regulatornih regij, ki so skupne samo skupinam podobno izraženih genov.[10] Primer tovrstnega pristopa je delo Chianga in sodelavcev,[11] kjer za vsako vezavno mesto izračunajo povprečno izražanje skupine vseh genov, ki to mesto vsebujejo. Povprečno izražanje posamezne skupine nato primerjajo z izražanjem naključne, enako velike skupine genov in izračunajo statistično značilnost prisotnosti vezavnega mesta in koherence izražanja skupine genov. Takšno neodvisno obravnavanje posameznih vezavnih mest zanemarja dejansko kombinatorično naravo uravnavanja izražanja genov, ki jo določa vezava skupine transkripcijskih faktorjev na različna vezavna gena. Bolj napredne metode zato preverjajo koherenco izražanja skupin genov, katerih regulatorne regije vsebujejo kombinacijo več potencialnih vezavnih mest.[10,12] Ozko grlo tovrstnih pristopov je izčrpno, kombinatorično iskanje, ki ga te metode navadno uporabljajo in zato tipično preiskujejo samo kombinacije dveh ali največ treh vezavnih mest. Enostaven izračun pove, da se že pri preverjanju tisoč vezavnih mest število možnih dvojk in trojk hitro povzpne v več sto milijonov. Preiskovanje postane še toliko bolj zahtevno oziroma skoraj neizvedljivo, če želimo v analizo vključiti tudi razdalje med vezavnimi mesti, razdalje med vezavnimi mesti in določenimi pomembnimi mejniki v strukturi genov (na primer mesto začetka transkripcije ali translacije, itd.), njihovo orientacijo, število pojavitev posameznega vezavnega mesta, itd. Zato se večina postopkov omejuje na kombinatorično iskanje opisov z največ tremi ali izjemoma štirimi elementi.[12]

Da bi presegli zgoraj omenjene kombinatorične omejitve smo razvili hevristično metodo preiskovanja prostora opisov oziroma pravil, ki opisujejo kompleksne strukture regulatornih regij. Naša metoda razvrščanja na podlagi pravil uporablja informacijo o podobnosti izražanja genov in tako usmerjeno preiskuje le najbolj perspektivne (in koherentne) podskupine genov, ki imajo tudi podobno regulatorno strukturo.

Opisni jezik in iskanje pravil

Cilj našega postopka je poiskati pravila, ki opišejo skupno regulatorno strukturo genov, katerih izražanje je med seboj čimbolj podobno. V našem postopku podobnost med genskimi izrazi določamo na podlagi Pearsonove korelacije.

Za opis strukturnih lastnosti regulatornih regij smo uporabili bogat opisni jezik, s katerim skušamo zajeti razdaljo vezavnih mest od položaja začetka transkripcije ali translacije, medsebojno razdaljo različnih vezavnih mest ter njihovo relativno in absolutno orientacijo glede na neki referenčni položaj (slika 2).



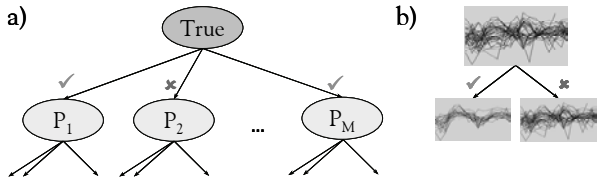
Slika 2 Elementi uporabljenega jezika hipotez, ki omogoča opis medsebojne razdalje vezavnih mest (v.m., razdalja d_2), razdalje od ATG (to je, mesta začetka translacije, razdalja d_1) ter orientacijo vezavnih mest v smeri branja DNA (+) ali v nasprotni smeri (-).

Pri tako bogatem opisnem jeziku je možnih pravil izredno veliko. Posledica je izjemno velik preiskovalni prostor in velika nevarnost, da se metode, ki v tem prostoru iščejo napovedna pravila, preveč prilagodijo podatkom (ang. *overfitting*). Da bi ta problem omilili, smo se pri razvoju metode hevrističnega iskanja zgledovali po metodi gradnje dreves za razvrščanje kot so jo predlagali Blockeel in sodelavci,[13] in tako razvili bolj splošno metodo iskanja pravil, ki poskuša preiskati le najbolj obetavne dele prostora. Vsak nadaljnji korak iskanja novih podskupin je v našem postopku ocenjen in izbran na podlagi korelacije izražanja genov v trenutno odkritih skupinah. Za nadaljnjo izostritev izberemo pravila, ki opisujejo le najbolj obetavne podskupine.

Postopek zahteva neko "ciljno" množico genov za katere želimo poiskati pravila in jih tako razvrstiti v skupine. Ta množica je lahko celoten genom, lahko pa je rezultat predhodne analize podatkov meritev DNA mikromrež, kjer na primer izberemo samo gene s statistično značilno spremembo izražanja v nekem ali več preučevanih pogojih (mutacija gena, zunanji, kemijski, mehanični, temperaturni ali kakšni drugi vplivi). Postopek prične z množico vseh genov, kar predstavimo s pravilom oziroma njegovim pogojem za proženje *TRUE*. To je tudi edino pravilo v začetni množici odkritih pravil. Sklepni del pravila je povprečno izražanje genov, ki jih opisuje pogojni del. Postopek iskanja poskuša izostriti trenutno odkrita pravila z dodajanjem novih pogojev. Na primer, pogojni del pravila, ki ga zapišemo z "M_1" in ki zahteva prisotnost vezavnega mesta M_1, lahko izostrimo z dodatnim pogojem, da je to vezavno mesto orientirano v pozitivno smer, kar zapišemo z "M_1+." Prvotni pogoj "M_1" lahko izostrimo tudi z zahtevo, da se vezavno mesto M_1 pojavlja na razdalji od -100 do -80 nukleotidov relativno glede na začetek translacije (ATG), kar zapišemo kot "M_1 @-100..-80(ref:ATG)." Za referenco lahko uporabimo neko drugo vezavno mesto, na primer M_2, kar zapišemo z "M_1 @ -100..-80(ref:M_2)." Prvotno pravilo "M_1" lahko izostrimo tudi z dodatno zahtevo o prisotnosti vezavnega mesta M_2, kar zapišemo z "M_1 in M_2."

Vsako izostreno pravilo pokrije manj genov od prvotnega pravila, vendar pa pri postopku zahtevamo, da pravila pokrijejo vsaj N ciljnih genov. N je parameter algoritma, ki ga določi uporabnik. Hkrati pa se mora podobnost med geni, ki ustrezajo izostrenemu pogoju, značilno povečati glede na medsebojno podobnost izražanja genov, ki ustrezajo prvotnemu pravilu. Če so ti pogoji izpolnjeni, nov pogoj oz. pravilo dodamo v množico pravil za nadaljnjo izostritev. V nasprotnem primeru pa ga le ocenimo in obdržimo, če se uvrsti med K najboljših pravil (uporabnik določi vrednost parametra K). Da bi še dodatno omejili preiskani prostor je velikost množice pravil za nadaljnjo izostritev omejena na največ L najboljših pravil (L je parameter algoritma, ki ga določi uporabnik). Kvaliteto pravila merimo s povprečno razdaljo izražanja genov v skupini, ki jo pravilo opisuje. Značilnost povečanja podobnosti genov med prvotnim in izostrenim pravilom merimo z uporabo F-testa, kjer v osnovi testiramo zmanjšanje variance v razdaljah med geni znotraj prvotne in izostrene skupine. Razdalja v izražanju genov je v našem postopku definirana s formulo: 1.0 – Pearsonova korelacija. Pravilo sprejmemo, če pokrije vsaj N ciljnih genov in opisuje skupino genov, katerih povprečna medsebojna razdalja je manjša od parametra D, ki ga določi uporabnik. Za osnovni korak preiskovalnega algoritma glej sliko 3, za preiskovalni algoritem pa sliko 4.

Velja poudariti, da lahko sprejeta pravila pokrijejo tudi gene izven ciljne množice. Metodo lahko zato uporabimo za iskanje genov, ki sicer niso bili vključeni v ciljno skupino, čeprav bi jih na podlagi izražanja in strukture regulatorne regije morali obravnavati skupaj z ostalimi geni v ciljni skupini.



Slika 3 a) Osnovni korak preiskovanja je izostritev posameznega pravila v trenutni množici odkritih pravil. b) Opazovana podobnost genov v skupini, ki jo opisuje izostreno pravilo, se mora značilno povečati v primerjavi s skupino prvotnega pravila. Dve različni izostritvi istega prvotnega pravila lahko privedeta do dveh različno homogenih podskupin (kljukica predstavlja značilno, križec pa neznačilno povečanje koherence izražanja genov dveh izostrenih pravil).

množica L najboljših pravil za izostritev B = {True}
množica K najboljših odkritih pravil R = {}
WHILE B ≠ {}
  iz B vzemi najboljše pravilo Pb
  FOR EACH k IN 1..število vseh vezavnim mest
    Pravilo Pb izostri z vezavnim mestom Mk in tako tvori pravilo Pn.
    IF pravilo Pn sprejemljivo AND značilno povečanje v podobnosti med pokritimi geni v Pb in Pn THEN v B dodaj novo pravilo Pn in v B ohrani le L najboljših pravil.
  Pravilo Pb dodaj v R, če je med K najboljšimi.
vrni množico K najboljših odkritih pravil R.

Slika 4 Preiskovalni algoritem. Izostritev pravila Pb z vezavnim mestom Mk se dejansko opravi na več načinov, z dodajanjem pogojev o prisotnosti, orientaciji in razdalji vezavnega mesta Mk glede na že prisotne člene v pravilu.

Predlagana metoda se razlikuje od klasičnih prekrivnih algoritmov za iskanje pravil, kot je na primer algoritem CN2,[14] saj omogoča odkrivanje prekrivajočih skupin genov. Osnovna verzija algoritma CN2 namreč iterativno odstranjuje primere (v našem primeru gene), ki jih v dani iteraciji opiše najboljše odkrito pravilo, ter nato ponovi iskanje na tako zmanjšani množici. To ponavlja dokler ne pokrije vseh primerov.

Predlagana metoda pa odkriva nova pravila vse dokler je pravila možno izostriti in se pri tem ne ozira na dejansko pokritost skupine genov s pravili.

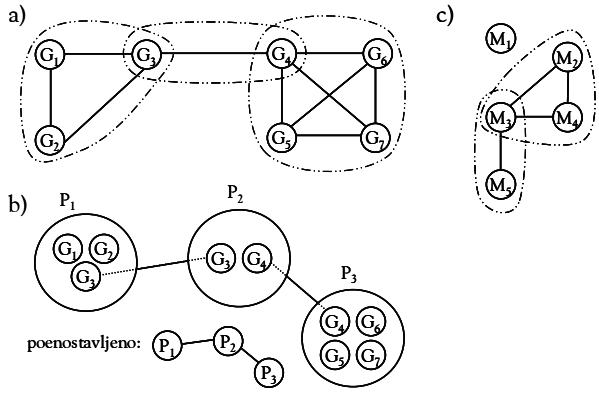
Izčrpno preiskovanje prostora relativno enostavnih (kratkih) pravil hitro preraste v neobvladljiv problem zaradi prej omenjene kombinatorične eksplozije. Izrazita prednost tu opisanega hevrističnega pristopa je zmožnost učinkovitega opisovanja uravnavanja izražanja genov, kjer lahko za osnovo obravnavamo več tisoč vezavnih mest in iz njih tvorimo kompleksnejše opise. V razdelku z eksperimentalnimi rezultati navajamo število hevristično preiskanih pravil in to primerjamo s številom pravil, ki bi jih sicer morali preveriti z izčrpnim preiskovanjem.

Prikaz odkritih pravil

Zaradi bogatega opisnega jezika in predvsem zaradi zmožnosti odkrivanja prekrivajočih skupin je število odkritih pravil oziroma skupin lahko zelo veliko. Da bi uporabnik lažje analiziral odkrita pravila, jih je smiselno prikazati z uporabo grafov, ki omogočajo boljši vpogled v skupne značilnosti in strukturo odkritih pravil in podskupin.

Kot osnovni prikaz odkritih pravil smo uporabili graf, kjer med seboj povežemo vse gene, ki jih opisuje posamezno pravilo (slika 5a). Pri velikem številu genov in odkritih pravil lahko tovrstni izris hitro postane zasičen in nepregleden. Naslednji višji nivo abstrakcije, s katerim lahko prikažemo iste rezultate vendar na dosti manj zasičen način, je graf pravil (slika 5b). Tu vsako vozlišče predstavlja eno pravilo. Dve pravili povežemo, če opisujeta vsaj en ali poljubno izbrano število skupnih genov. Z višanjem praga dobimo vse manj povezan graf, ohranjene povezave pa lahko kažejo na veliko strukturno in izrazno podobnost vpletenih genov. Zadnji nivo abstrakcije (slika 5c) opisuje podobnosti med pravili na podlagi skupnih členov, ki nastopajo v pravilih. Vozlišča so v tem primeru členi pravil. Povežemo jih, če nastopajo v (vsaj enem) istem pravilu. Podobno lahko tudi tu spreminjamo prag zahtevane prisotnosti člena v

različnem številu pravil in tako opazujemo kateri členi so centralni, torej nastopajo v mnogo pravilih in so zato morda vezavna mesta nekih splošnih regulatorjev. Opazujemo lahko tudi kateri členi postanejo hitro nepovezani in lahko predstavljajo zato vezavna mesta, ki določajo specifično izražanje podskupine genov.



Slika 5 a) Mreža genov. b) Prikaz povezanosti pravil oziroma skupin genov, ki jih pravila določajo. c) Prikaz podobnosti pravil, kjer povežemo pravila s skupnimi členi. Pravilo P_1 ("M_1")testira prisotnost vezavnega mesta M_1, pravilo P_2 ("M_2 in M_3 in M_4") prisotnost vezavnih mest M_2, M_3 in M_4, pravilo P_3 ("M_3 in M_5") pa prisotnost vezavni mest M_3 in M_5 v regulatornih regijah genov.

Eksperimentalni rezultati

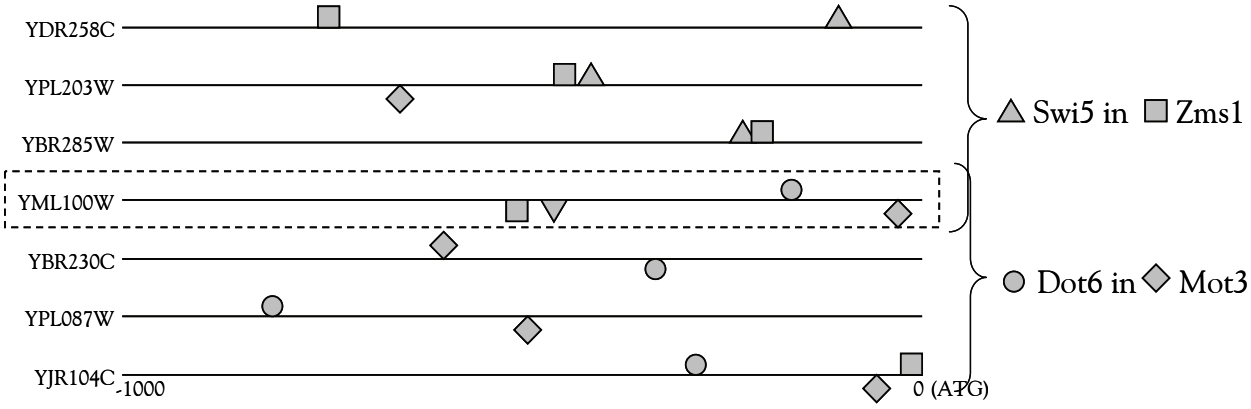
Z metodo razvrščanja na podlagi pravil smo analizirali regulatorne regije kvasnih genov, katerim so Gasch in sodelavci[15] izmerili in določili značilno spremembo izražanja v različnih stresnih pogojih. Za ciljno množico smo izbrali 281 genov s povečanim izražanjem v stresnih pogojih. Vzeli smo regulatorne regije dolžine 1000 nukleotidov od mesta začetka translacije (ATG, na poziciji 0). Za podatke o vezavnih mestih smo uporabili bazo znanih in eksperimentalno potrjenih vezavnih mest,[16] ter jim dodali vezavna mesta, ki smo jih odkrili s programom za lokalno poravnavo zaporedij MEME[6]. Pri gradnji pravil smo tako upoštevali približno 3100 vezavnih mest. Iskali smo skupine z najmanj štirimi ciljnimi geni (N=4) in povprečno Pearsonovo korelacijo genov v skupini nad 0.45 (D=1.0 – 0.45=0.55). Velikost množice pravil za nadaljnjo izostritev je bila omejena na L=1000 pravil. Glede na dano dolžino regulatorne regije, so relativne razdalje med elementi v razponu od –1000 do 1000 nukleotidov. Razdalje med elementi smo zaokroževali na 40 nukleotidov natančno, kar pomeni, da smo obravnavali 50 (=2000/40) možnih različnih razdalj med elementi. Večina odkritih pravil je sestavljenih iz dveh členov. Najdaljše odkrito pravilo vsebuje štiri člene in opisuje razdalje med štirimi vezavnimi mesti ter njihovo orientacijo. Število vseh možnih pravil, ki opisujejo samo prisotnost in orientacijo štirih vezavnih mest je ogromno (vezavnih mest je 3100, ki so lahko s pozitivno, negativno ali brez določene orientacije, faktor 3):

$$\binom{3100 \cdot 3}{4} \approx 3.11 \cdot 10^{14}$$

Če bi želeli izčrpno preiskati tudi pravila, ki opisujejo razdalje med vezavnimi mesti, bi se zgornje število možnih pravil povečalo za faktor $50^4 = 6.25 \cdot 10^6$. Naša metoda hevrističnega iskanja je na tej domeni pregledala $3.3 \cdot 10^9$ pravil oziroma manj kot 0.0011% prostora vseh možnih pravil s štirimi členi, kar se je na delovni postaji s procesno enoto Pentium 4, 3.4 GHz izvajalo približno eno uro in dvajset minut.

S predstavljeno metodo smo uspešno odkrili in opisali regulatorne regije skupine genov, za katere je bilo predhodno že pokazano, da so povezane s transkripcijskimi faktorji Msn2p, Msn4p in Yap1p. Prav tako smo odkrili druga, domnevna regulatorna vezavna mesta, ki nastopajo v kombinaciji z že znanimi mesti, kar nakazuje možnost novih mehanizmov uravnavanja. S pregledom grafa pravil (ni prikazan) smo hitro odkrili dve prekrivajoči skupini, prikazani na sliki 6. Pravili opisujeta regulatorne regije dveh različnih skupin genov s skupnim genom YML100W. Pregled znanih določitev funkcij (ang. *annotation*) v genski ontologiji[17] (ang. *Gene Ontology*) nam pove, da je značilni biološki proces

zgornjih treh genov na sliki 6 *celični metabolizem proteinov* (ang. *cellular protein metabolism*), značilni biološki proces spodnjih treh genov pa *odziv na stres* (ang. *response to stress*). Pregled določitve funkcije gena YML100W (imenovanega tudi TSL1) nam pokaže, da sta genu eksperimentalno določeni obe funkciji. Ta primer posredno potrjuje zmožnost metode za odkrivanje funkcijsko smiselno se prekrivajočih skupin genov.



Slika 6 Primer dveh enostavnih odkritih pravil, ki zahtevata le prisotnost posameznih vezavnih mest, in prikaz regulatorne regije genov, ki jih pravili opisujeta. Posamezni simbol ponazarja vezavno mesto. Obe pravili ("Swi5 in Zms1" in "Dot6 in Mot3") opisujeta regulatorno regijo gena YML100W. Zaporedja so poravnana glede na začetek translacije (ATG) na skrajni desni. Prikazana je regulatorna regija dolžine tisoč nukleotidov.

Zaključek

Dobljeni eksperimentalni rezultati kažejo, da je možno dokaj učinkovito in relativno hitro poiskati kompleksne opise regulatornih regij skupin med seboj podobno izraženih genov. Različni prikazi dobljenih rezultatov še dodatno pripomorejo k boljšemu razumevanju in biološki interpretaciji. Glavno uporabnost predstavljene metode vidimo v luči iskanja dodatnih dokazov, da so geni v neki teoretično ali pa eksperimentalno določeni skupini dejansko tudi regulatorno medsebojno povezani oziroma imajo skupne regulatorje. Metoda lahko na ta način postavi izbrane hipoteze, ki jih je moč preveriti z naknadnimi eksperimenti. Implementacija tu predstavljene metode v obliki spletne aplikacije, ki bo omogočila biologom enostavno uporabo opisanega orodja, je v delu.

Literatura

1. Wasserman WW, Sandelin A: Applied bioinformatics for the identification of regulatory elements. Nat Reviews Genet 2004; 5: 276-87.
2. Kothapalli R, Yoder SJ, Mane S, et al.: Microarray results: how accurate are they? BMC Bioinformatics 2002; 3:22.
3. Chuanqui RF, Bonner RF, Best CJM, et al.: Post-analysis follow-up and validation of microarray experiments. Nature Genetics Supplement 2002; 32:509-514.
4. Bajic VB, Tan SL, Suzuki Y, et al.: Promoter prediction analysis on the whole human genome, Nature Biotechnology 2004; 22:1467-1473.
5. Wingender E, Dietze P, Karas H, et al.: TRANSFAC: a database on transcription factors and their DNA binding sites. Nucleic Acids Res 1996; 24(1): 238-41.
6. Bailey TL, Elkan C: Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol 1994; 2: 28-36.
7. Tompa M, Li N, Bailey TL, et al.: Assessing computational tools for the discovery of

transcription factor binding sites, Nature Biotechnology 2005; 23:137-144.
8. Pennacchio LA, Rubin EM: Genomic strategies to identify mammalian regulatory sequences. Nat Reviews Genet 2001; 2: 100-9.
9. Conlon EM, Liu XS, Lieb JD, et al.: Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci USA 2003; 100(6): 3339-44.
10. Pilpel Y, Sudarsanam P, Church GM: Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet 2001; 29(2): 153-9.
11. Chiang DY, Brown PO, Eisen MB: Visualizing associations between genome sequences and gene expression data using genome-mean expression profiles. Bioinformatics 2001; 17(supp. 1):S49-S55.
12. Beer MA, Tavazoie S: Predicting gene expression from sequence, Cell 2004; 117:185-198.
13. Blockeel H, De Raedt L, Ramon J: Top-down induction of clustering trees. Machine Learning, Proceedings of the 15th International Conference 1998; Morgan Kaufmann.
14. Clark P, Niblett T: The CN2 induction algorithm. Machine Learning 1989; 3(4): 261-283.
15. Gasch AP, Spellman PT, Kao CM, et al.: Genomic expression programs in the response of yeast cells to environmental changes. Mol Biol Cell 2000; 11(12): 4241-57.
16. Lee TI, Rinaldi NJ, Robert F, et al.: Transcriptional regulatory networks in Saccharomyces cerevisiae. Science 2002; 298:799-804.
17. The Gene Ontology Consortium: Gene Ontology: tool for unification of biology. Nature Genetics 2000; 25:25-29.

Izvirni znanstveni članek ■

Encimska kinetika in molekularno modeliranje se dopolnjujeta pri proučevanju reakcijskih mehanizmov

Jure Stojan

Izvleček. Kinetika holinesteraz je zelo zapletena in kaže več odstopanj od klasične Michaelis-Mentenove hiperbole. Zato smo želeli uskladiti rezultate kinetičnih meritev z znanimi kristalografskimi podatki in tako predlagati reakcijsko shemo hidrolitične pretvorbe acetiltioholina - umetnega kromogenega substrata. S pomočjo molekularnega modeliranja smo prikazali pomembne vmesne produkte v reakcijskem mehanizmu in s steričnim oviranjem deacetilacije razložili inhibicijo holinesteraze s prebitkom substrata.

Enzyme kinetics and molecular modeling complement each other at study of reaction mechanisms

Abstract. The kinetics of cholinesterases is complex and show several deviations from classical Micaelis-Menten hyperbola. Therefore, we tried to correlate the results of kinetic measurements with available crystallographic data and suggest an appropriate reaction scheme for the hydrolysis of acetylthiocholine - an artificial chromogenic substrate. By means of molecular modeling we anticipated important reaction intermediates and suggested steric hindrance of deacetylation as a basis for the inhibition by the excess of substrate.

■ **Infor Med Slov**: 2006; 11(1): 60-65

Institucija avtorja: Medicinska fakulteta, Univerza v Ljubljani, Ljubljana, Slovenija.

Kontaktna oseba: Jure Stojan, Medicinska fakulteta, Univerza v Ljubljani, Vrazov trg 2, 1000 Ljubljana, Slovenija. email: stojan@mf.uni-lj.si.

Uvod

Encimska kinetika poskuša s proučevanjem hitrosti katalitične pretvorbe substrata ugotoviti mehanizem encimskega delovanja to je, razjasniti dogodke v poteku reakcije na molekulski ravni. Matematične osnove zapletenih algoritmov za analizo kinetičnih podatkov v encimatiki so že dolgo znane. Vendar pa je njihov prenos v praktično uporabo, zaradi izredne računske zahtevnosti, postal mogoč šele v zadnjem desetletju. Ne le regresijske metode za direktno prilagajanje parametrov eksplicitnih enačb za hitrost ali časovni potek encimskih reakcij, ampak tudi reševanje zapletenih sistemov diferencialnih enačb in prilagajanje njihovih parametrov je vse bolj dosegljivo.[1] Toda narava kinetičnih podatkov v encimatiki je taka, da njihova interpretacija ni enolična, kar pomeni, da pravega izmed konkurenčnih reakcijskih mehanizmov ni mogoče izbrati z gotovostjo.[2] Pomembne odločitve v poteku kinetične analize se tako največkrat sprejemajo na osnovi nekinetičnih podatkov, ki vsakokrat nakažejo smer nadaljevanja. Tu pridejo v poštev različne kromatografske tehnike, s katerimi je mogoče identificirati vmesne reakcijske produkte in metodologija rekombinantne DNA, ki z usmerjeno mutagenezo razkrije specifično vlogo ključnih aminokislinskih ostankov. Najbolj cenjena pa je strukturna informacija, ki prihaja iz kristalografije z X-žarki oz. NMR analize bioloških molekul v raztopini. Uporaba naštetih tehnik je žal zelo draga, zato pa je s povečanjem moči in predvsem nizko ceno računalniške opreme postala dostopnejšo še ena od metod encimskega proučevanja, t.j. molekularno modeliranje. Slednja temelji na izkušnjah obeh metod strukturne biologije in jih v obliki matematičnih algoritmov uporablja za napovedovanje 3D strukture novih ali dinamike delovanja že znanih bioloških molekul. Pri tem lahko izredno olajša pomebne odločitve v poteku kinetične analize in dodatno razkrije nove možne razlage.[3]

Predstavitev problema

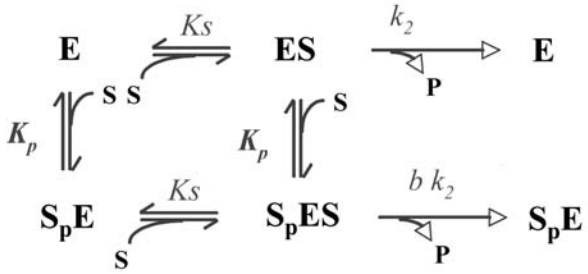
Acetilholinesteraza je encim, ki v večstopenjski reakciji hidrolitično razcepi nevrotransmiter acetilholin in s tem prekine živčni impulz v holinergični sinapsi ali signal za kontrakcijo v žično-mišičnem stiku.[4] Encim, ki je eden najuspešnejših, pretvori tudi do 10000 substratnih molekul v sekundi. Molekula acetilholinesteraze je sestavljena iz več kot 530 aminokislinskih ostankov in sodi v skupino α-β hidrolaz z globoko ugreznjenim aktivnim centrom. Pri katalizi kemično sedelujejo stranske verige Ser-His-Glu triade, in sicer kot sistem za prenos nabojev ter večje število hidrofobnih ostankov v steni lijaka aktivnega mesta (glej sliko 5). Kemični mehanizem hidrolitične pretvorbe substrata poteka po principih kislinsko-bazne in kovalentne katalize, seveda po predhodni prilagoditvi substrata na dnu aktivnega mesta. Kinetika te reakcije je zelo nenavadna in po dobrih 50 letih intenzivnega proučevanja še vedno v mnogih pogledih nerazjasnjena. Več kot 70 znanih kristalnih struktur petih različnih holinesteraz je sliko precej izostrilo, vendar predvsem v interakcijah z inhibitorji in manj pri reakciji s substratom.[5]

Kinetika holinesteraz

Odvisnost hitrosti od koncentracije substrata pri holinesterazah ni hiperbolična. Predvsem pri visokih koncentracijah substrata so encimi bolj ali manj popolnoma inhibirani, kar so že zelo zgodaj razlagali kot medsebojno tekmovanje molekul substrata. Predvideli so različne funkcionalne predele aktivnega centra, kamor se hkrati lahko vežeta dve substratni molekuli.

Slika 1 prikazuje Kinetični model po Webb-u[6] iz leta 1963. E je prosti encim, S substrat, P produkti, ES je kompleks encim-substrat, ko je substrat vezan v aktivnem centru, SpE je kompleks encim-substrat, ko je substrat vezan na perifernem mestu in SpES kompleks z zasedenima obema vezalnima mestoma; Ks je disociacijska konstanta za vezavo substrata v aktivni center, Kp za vezavo na

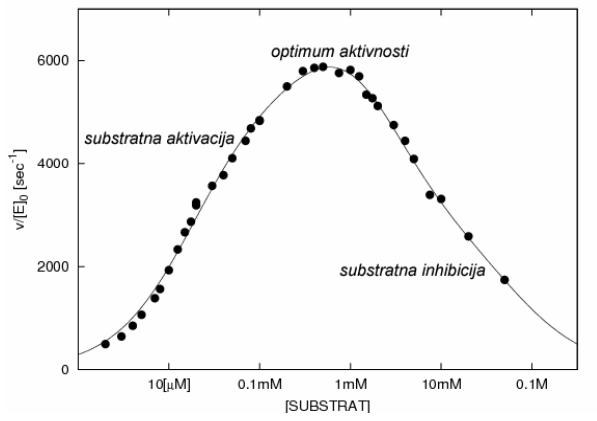
periferno mesto, k2 je hitrostna konstanta prvega reda za pretvorbo substrata v produkte in b propocionalnostni faktor.



Slika 1 Kinetični model po Webb-u.[6]

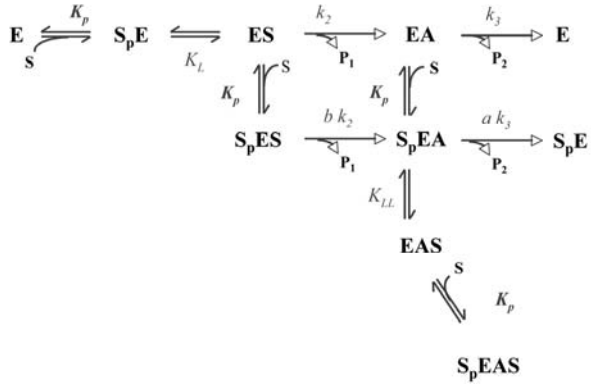
Ko smo pred leti ugotovili, da je večina holinesteraz, naravnih in specifično mutiranih, pri srednjih koncentracijah substrata precej aktivnejša kot predvideva Michaelis-Mentenova kinetika, je bilo potrebno revidirati kinetične modele in s tem tudi razumevanje molekulskega mehanizma reakcije.[3]

Tipična odvisnost acetilholinesterazne aktivnosti ($v_0/[E_T]$) od koncentracije substrata ([S]) je prikazana na sliki 2.



Slika 2 Aktivnost acetilholinesteraze iz Električne jegulje v odvisnosti od koncentracije substrata acetiltioholina.

Ena od razširitev reakcijske sheme holineteraz je kinetični model, ki predvideva sočasno vezavo dveh substratnih molekul na prosti in acetilirani encim ter aktivacijo in inhibicijo v odvisnosti od koncetracije substrata, kot so ga predlagali Stojan in sod.[7] (slika 3). Pomen simbolov je analogen kot pri Webb-ovem modelu; kompleksi, ki vsebujejo 'A' predstavljajo kovalentne vmesne reakcijske produkte z acetiliranim Ser200; k_2 je hitrostna konstanta prvega reda za acetilacijo in k_3 hitrostna konstanta pseudo prvega reda za deacetilacijo.



Slika 3 Kinetični model, kot so ga predlagali Stojan in sod.[7]

Za pomoč pri konstrukciji tega in podobnih modelov si pomagamo z molekularnim modeliranjem dogodkov med interakcijo holinesteraze s substratom.

Metodologija

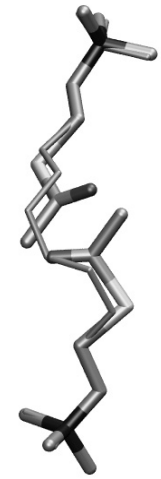
Za izdelavo molekularnih modelov nekaterih pomembnih reakcijskih intermediatov v poteku hidrolize acetiltioholina, t.j. analoga naravnega substrata, ki se uporablja za rutinsko zasledovanje holinesteraznih aktivnosti,[8] smo postavili nekaj izhodiščnih predpostavk: a/ ker se struktura encima v prisotnosti množice različnih inhibitorjev v razrešenih 3D strukturah praktično ne spreminja, smo v celotnem poteku grajenja molekularnih modelov fiksirali Cα okostje polipeptidne verige; b/ za kandidatno področje, kjer se verjetno nahajajo substratne molekule, smo izbrali prostor v aktivnem centru, kjer se nahajajo vezani inhibitorji; c/ prilagajanje substratnih molekul v kandidatno področje mora biti tako, da zahteva čim manj spreminjanja položaja stranskih

verig; d/ vmesni produkti, ki so kinetično predvidljivi morajo biti sterično možni; e/ manj znanstven, a zelo pomemben kriterij je estetska dopadljivost končnega izdelka.

Obstoj nekaterih intermediatov v poteku reakcije je očiten in ga je mogoče predvideti na osnovi sorodnih kristalografskih podatkov: n. pr. acetiliran serin v aktivnem mestu je analogen fosforiliranemu ali karbamoiliranemu serinu (Ser200) v kristalnih strukturah kovalentnih kompleksov z organofosfati (PDB koda 1SOM) in karbamati (PDB koda 1OCE). Po drugi strani pa je glavni teoretični izziv, ki ga nakazujejo kinetični poskusi, popolnoma inhibiran encim v prisotnosti zelo visokih koncetracij substrata. V našem kinetičnem modelu lahko tako situacijo razložimo s popolnoma zasedenim, kar pomeni popolnoma blokiranim aktivnim centrom. V njem se nahajata dve substratni molekuli, ki reakcijo ustavita na stopnji acetiliranega encima.

Za manipulacijo makro- in substratnih molekul smo uporabljali: WHAT IF, program za molekularno modeliranje in načrtovanje zdravil,[9] Gaussian 03, paket programov za ab initio računanje strukure malih molekul, CHARMM (Chemistry at HARvard Molecular Mechanics) program za makromolekularne simulacije, oprimizacije in dinamiko,[10] Molden, program za vizualizacijo molekulske in elektronske strukture[11] ter Swiss-PdbViewer za prikazovanje in strukturno poravnavanje makro- in substratnih molekul.[12]

Tipičen postopek modeliranja kompleska med acetiliranim encimom in dvema substratnima molekulama je potekal takole: substratno molekulo acetiltioholona smo zgradili s programom Molden in jo z Gaussian 03 ab initio optimizirali v vakuumu (pri optimizaciji je bil uporabljen 6-31g* bazni set). Geometrijo in naboje vseh 26 atomov smo uporabili pri izdelavi seta CHARMM-ovih parametrov. Podobna obravnava ocetne kisline in acetilatnega ostanka ni bila potrebna, saj je set parametrov že v CHARMM-ovi bazi. Kot izhodiščno strukturo za modeliranje položaja acetiltioholinskih molekul smo uporabili kompleks vretenčarske acetilholinesteraze iz ribe *Torpedo californica* (pacifiški električni skat) z dvopolnim inhibitorjem dekametonijem, PDB koda 1ACL. Struktura dekametonija namreč ustreza velikosti dveh acetilholinskih molekul, če bi le-ti bili med seboj povezani z metilnima skupinama acetatov (slika 4).

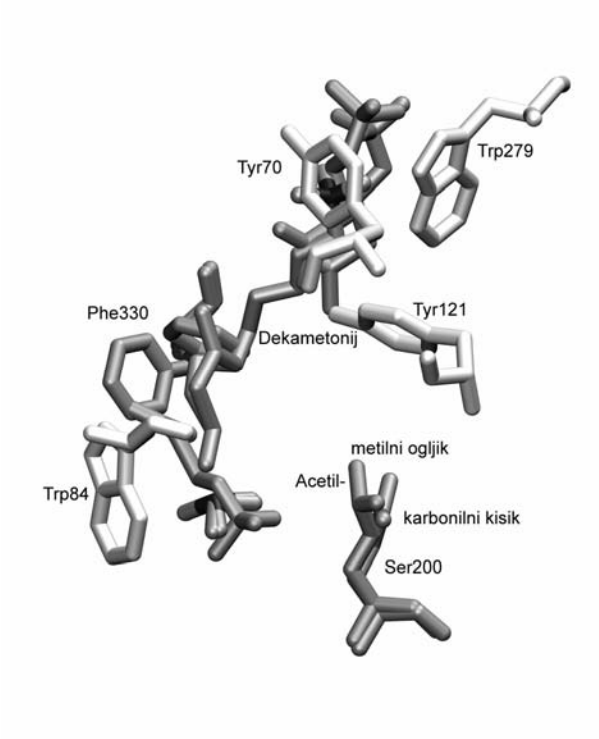


Slika 4 Začetna poravnava dveh molekul substrata na konformacijo dekametonija (oranžno) iz strukture kompleksa z acetilholinesterazo (PDB koda 1ACL).

Za prilagajanje 4 težkih atomov tetrametilamino skupine v acetiltioholinu na obe ustrezni skupini dekametonija smo uporabili 'suppos' ukaz v programu Whatif. Izhodiščne koordinate acetilne skupine smo po analognem postopku pridobili iz kristalne strukture istega encima v kompleksu s somanom, PDB koda 1SOM. V nadaljevanju smo s programom CHARMM strukturo najprej relaksirali in nato v več zaporednih dinamičnih simulacijah v skupni dolžini 500 pikosekund z vmesnimi korekcijami dokončali. Kot zadnjo stopnjo smo opravili še 50 korakov QMMM minimizacije popolnoma sproščene strukture z 2.5 A debelim plaščem vode. Pri tej končni relaksaciji smo kvantno mehansko računali stranske verige obeh triptofanov (notranjega Trp84 in zunanjegaTrp279), stransko verigo aktivnega serina (S200) z acetilno skupino in obe substratni molekuli, vsega skupaj 101 atom.

Rezultat

Rezultat modeliranja je prikazan na sliki 5.



Slika 5 Model dveh acetilholinskih molekul v aktivnem mestu acetilirane vretenčarske acetilholinesteraze.

Aminokislinski ostanki, s katerimi se povežeta na vhodu (zgoraj) in na dnu aktivnega centra (levo spodaj) obe acetilholinski molekuli so prikazani svetlejše. Kation-π interakcije med obema triptofanoma in pozitivno nabitima tetrametilamonijevima glavama substratov so najpomembnejše vezi, podobno kot pri vezavi dekametonija. Zaradi prisotnosti kovalentno vezanega acetata na serinu (desno spodaj), se tista acetilholinska molekula, ki je ujeta v notranjosti aktivnega centra (levo spodaj) ne more orientirati v položaj, ki bi omogočal nadaljevanje katalitične pretvorbe. Acetilirani serin, obe substratni molekuli in Phe330 so prikazani dvojno. V razrešeni kristalni strukturi acetilholinesteraze po vsrkavanju acetilthioholina v 500 mM koncentraciji (PDB koda 2C4H) je Phe330 orientiran vodoravno.[13] Edina bistvena razlika med modeliranim in kristalografsko ugotovljenim kompleksom je prav orientacija Phe330 in posledično položaj acetatne skupine notranje molekule substrata, ki se postavi na njegovo drugo stran. Opozorimo naj še na razdaljo med karbonilno skupino zunanje substratne molekule (v bližini Trp279), ki tako v modelu kakor tudi v razrešeni kristalni strukturi, po razdalji sodeč, tvori vodikovo vez s hidroksilno skupino Tyr121.

Razprava

Molekularno modeliranje in vizualizacija dogodkov med potekom encimske reakcije sta tako teoretično kot praktično v veliko pomoč pri razjasnitvi molekulskega mehanizma proučevane reakcije. To še posebej velja v primeru, ko molekularno modeliranje kombiniramo z analizo kinetičnih podatkov in drugimi nekinetičnimi študijami. Uporabljamo ga lahko na več nivojih: vizualizacija potencialnih reakcijskih kompleksov nam odpira nove ideje in možne interpretacije, dinamične simulacije in kvantno-mehanski izračuni pa učinkovito preverjajo hipoteze oziroma privedejo do novih, zanesljivih informacij.

Primer vizualizacije intermediata holinesterazne reakcije, ko je encim popolnoma blokiran, ne kaže le ujemanja z interpretacijo natančne kinetične analize, ampak tudi to, da je mogoče z relativno preprostim in poceni pristopom, v razmeroma kratkem času (2-3 tedne s 4 AMD64 procesorji), s precejšno gotovostjo napovedati reakcijske dogodke. Seveda se velikokrat pojavljajo kritike, da so taki modeli le slabe ideje, ki nimajo dosti zveze z resničnostjo in da šele razrešitev kristalne strukture postavi stvari na pravo mesto. Do neke mere je to lahko res, a kadar je molekulske modele mogoče razviti po zgoraj naštetih kriterijih, je rezultat mnogo bliže resnici kot špekulaciji. Slednje se je pokazalo tudi v našem primeru, ko dve substratni molekuli in molekula enega od produktov popolnoma blokirajo aktivni center acetilholinesteraze. Kaže, da je substratno inhibicijo vretenčarske acetilholinesteraze resnično mogoče pripisati veliki gneči substratnih

molekul v aktivnem centru. Molekulski model, ki je bil izdelan v skladu z interpretacijo kinetičnih podatkov podpira to idejo, saj sterične lastnosti substratov in produktov dopuščajo takšo razlago. Dokončno veljavo modelu pa daje izredno ujemanje z nedavno razrešeno kristalno strukturo encima v prisotnosti 500 mM koncentracije acetiltioholina.[13]

Literatura

1. Stojan J: Analysis of progress curves in an acetylcholinesterase reaction: a numerical integration treatment. J Chem Inf Comput Sci 1997; 37: 1025-7.
2. Stojan J: Rational Polynomial Equation Helps to Select Among Homeomorphic Kinetic Models for Cholinesterase Reaction Mechanism. Chem Biol Interact 2005; 157-158, 173-179.
3. Stojan J, Marcel V, Estrada-Mondaca S, et al.: A putative kinetic model for substrate metabolisation by Drosophila acetylcholinesterase. Febs Lett 1998; 440: 85-8.
4. Massoulie J, Pezzementi L, Bon S, et al.: Molecular and cellular biology of cholinesterases, Prog. Neurobiol. 1993; 41: 31-91.
5. Sussman JL, Harel M, Frolow F, et al.: Atomic structure of acetylcholinesterase from Torpedo californica: a prototypic acetylcholine-binding protein, Science 1991; 253: 872-879.
6. Webb JL: Enzyme and metabolic inhibitors, general principles of inhibition. New York, Academic Press 1963; 118.
7. Stojan J, Brochier L, Alies C, et al.: Inhibition of Drosophila melanogaster acetylcholinesterase by high concentations of substrate. Eur J Biochem 2004; 271: 1364-71.
8. Ellman GL, Courtney KD, Andres V, et al. A new and rapid colorimetric determination of acetylcholinesterase activity. Biochem. Pharmacol. 1961; 7: 88-95.
9. Vriend G: WHAT IF: A molecular modeling and drug design program. J. Mol. Graph. 1990; 8: 52-56.
10. Brooks BR, Bruccoleri RE, Olafson BD, et al.: CHARMM: A Program for Macromolecular Energy, Minimization, and Dynamics Calculations. J. Comp. Chem. 1983; 4: 187.
11. Schaftenaar G, Noordik JH: Molden: a pre- and post-processing program for molecular and electronic structures, J. Comput.-Aided Mol. Design 2000; 14: 123-134.
12. Guex N, Peitsch MC: SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 1997; 18: 2714-2723.
13. Colletier JP, Fournier D, Greenblatt HM, et al.: Structural insights into substrate traffic and inhibition in acetylcholinesterase. The EMBO Journal advance online publication 2006; doi: 10.1038/sj.emboj.7601175.

Pregledni znanstveni članek ■

K orodjem bioinformatike za fenomiko in sistemsko biologijo

Uroš Petrovič, Mojca Mattiazzi, Tomaž Curk, Blaž Zupan, Igor Križaj

Izvleček. Sistemska biologija je veda, katere cilj je razumevanje bioloških procesov na sistemski ravni, z upoštevanjem kompleksnih interakcij med geni, proteini in drugimi elementi celice. Področje post-genomske biologije, ki se ukvarja z vplivom celotnega genoma na lastnosti celice, kar je nujen korak na poti k sistemski biologiji, se imenuje fenomika. Zaradi tehničnih omejitev pri genetski manipulaciji človeških celic je zaenkrat pri raziskavah na področju fenomike nujna uporaba modelnih organizmov. Ena glavnih omejitev za razmah sistemske biologije je pomanjkanje ustreznih orodij bioinformatike, katerih razvoj zato poteka vzporedno z razvojem novih eksperimentalnih pristopov pri modelnih organizmih, s končnim ciljem aplikacije na biologijo človeka.

Towards the bioinformatics tools for phenomics and systems biology

Abstract. The aim of systems biology is systems-level understanding of biological processes that takes into account complex interactions of genes, proteins and other cell elements. The area of post-genomic biology that deals with the effect of the whole genome on the cell characteristics – a necessary step towards systems biology – is called phenomics. Because of technical limitations in genetic engineering of human cells, the use of model organisms is currently inevitable in phenomics studies. One of the main limitations for the advancement of systems biology is the lack of appropriate bioinformatics tools. The development of these therefore takes place in parallel with the development of new experimental approaches in model organisms, with the ultimate goal to apply them also to human biology.

■ **Infor Med Slov**: 2006; 11(1): 66-71

Institucije avtorjev: Institut Jožef Stefan, Ljubljana, Slovenija (UP, MM, IK), Fakulteta za računalništvo in informatiko, Univerza v Ljubljani, Slovenija (TC, BZ), Baylor College of Medicine, Houston, USA (BZ).

Kontaktna oseba: Uroš Petrovič, Odsek za biokemijo in molekularno biologijo, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana. email: uros.petrovic@ijs.si.

Uvod

Eno ključnih vprašanj biologije je, kako genotip vpliva na fenotip. Klasična genetika je pri iskanju odgovorov na to vprašanje omejena na preučevanje relativno majhnega števila genov, medtem ko je fenotip vedno rezultat delovanja vseh genov, torej genoma. Za razumevanje vpliva celotnega in ne le delnega genotipa na fenotip je potrebno analizirati vplive čim večjega števila genotipov posameznega genoma na čim večje število merljivih in praviloma kvantitativnih lastnosti fenotipa.

Fenomika in modelni organizmi

Za dosego tega cilja je nujna uporaba modelnih organizmov, ki omogočajo uporabo orodij za natančno poseganje v genom. Na njih poteka razvoj pristopov, ki bodo v prihodnosti uporabni tudi za študij sistemske biologije v humani medicini; celovit pristop za študij fenomike pri modelnih organizmih je zatorej načrtovan tako, da bo lahko v čim večji meri prenesen na preučevanje človeka. Najdlje je razvoj tehnik in orodij, potrebnih za globalno analizo fenotipa, fenomiko, prišel pri preučevanju kvasovke *Saccharomyces cerevisiae*. Od določitve zaporedja celotnega genoma kvasovke leta 1996[1] so si sledili:

- izdelava prve DNA mikromreže, ki je vsebovala celoten genom[2] in omogoča globalno transkriptomsko analizo,
- proteinske mikromreže s celotnim proteomom,[3] ki omogočajo določitev fizičnih interakcij endogenih ali eksogenih proteinov, lipidnih molekul in drugih nizkomolekularnih molekul z vsemi kvasnimi proteini,
- priprava zbirke sevov s sistematično izbitimi vsemi neesencielnimi geni,[4] kar je omogočilo analizo fenotipa vseh mutant s posameznimi delecijami,
- zbirka sevov z vsemi geni s fuzijo z genom za zeleni fluorescirajoči protein,[5] ki omogoča lokalizacijo vseh proteinov v celici,
- zbirka sevov z možnostjo uravnavanega izražanja vseh kvasnih genov,[6] kar je omogočilo analizo fenotipa vseh mutant s prekomernim izražanjem posameznih genov, in
- zbirka sevov z vsemi geni s fuzijo z označevalcem "TAP",[7] ki je omogočila identifikacijo večine proteinskih kompleksov v celici kvasovke.

Razvoj teh tehnik je napravil iz kvasovke trenutno najbolj primeren modelni organizem za sistemsko biologijo, saj poleg analize fenotipa pri vseh možnih "preprostih" genotipih (hipomorfne mutacije posameznih genov oziroma prekomerno izražanje posameznih genov) lahko dajo podatke tudi o drugi pomembni komponenti sistema, to je o medsebojni povezanosti njegovih gradnikov (proteinske in genetske interakcije, koordinirano izražanje genov). Poleg kvasovke se v zadnjem času kot modelni organizmi uveljavljajo tudi bolj kompleksni organizmi, na primer nematod *Caenorhabditis elegans*.[8]

Uporaba sistemske biologije v medicini bo predvidoma omogočala hiter razvoj novih, natančno usmerjenih zdravil in razvoj posamezniku prilagojene medicine. Za dosego tega cilja bo potrebno razumeti procese v človeških celicah in organizmu na podobni ravni, kot danes razumemo celice kvasovk. Kot primer vzemimo cistično fibrozo, ki je ena najbolj znanih tako imenovanih monogenskih bolezni. Nastanek cistične fibroze povzroči mutacija v enem genu, imenovanem *CFTR*.[9] Vendar pa samo iz mesta mutacije v *CFTR* ne moremo sklepati na točen potek bolezni, v kolikor ne upoštevamo tudi aktivnosti drugih genov/proteinov. Na primer, za nastanek bolezenskih simptomov na osnovi mutacije v *CFTR* je nujno potrebna aktivnost proteinov Hsp70. Tako lahko znižana aktivnost Hsp70, bodisi zaradi mutacije bodisi zaradi inhibitorjev, blaži simptome cistične fibroze. Znano je, da lahko kot inhibitorji Hsp70 delujejo

butanojska in druge kratkoverižne maščobne kisline, ki so prisotne v fizioloških razmerah v celicah, kar pomeni, da na simptome cistične fibroze posredno vpliva tudi metabolizem lipidov. Takšnih primerov je v fiziologiji človeških celic še veliko in predstavljajo, zaradi medsebojne povezanosti celičnih procesov, pravilo in ne izjeme.

Bioinformatični pristopi in orodja

Za razumevanje medsebojne povezanosti genov oziroma proteinov v celici ali organizmu je potrebno odgovoriti na nekaj osnovnih vprašanj. Za vse gene v genomu moramo najprej poznati njihovo funkcijo ("Kaj gen/protein počne?"), pri tem pa se moramo zavedati, da je večina genov/proteinov udeleženih v več kot samo enem procesu ("Kaj vse gen/protein počne?"). Naslednja stopnja razumevanja celice kot sistema zahteva poznavanje mehanizma delovanja ("Kako to počne?") ter končno poznavanje vseh genetskih in proteinskih interakcij, ki nastopajo v celičnih procesih ("S kom geni/proteini sodelujejo in kako?"). Odgovore na ta v bistvu zelo preprosta vprašanja lahko na ravni celotnega genoma/proteoma da le več različnih eksperimentalnih tehnik, katerih združena interpretacija šele lahko predstavi celotno sliko. Orodja bioinformatike za dosego tovrstne interpretacije trenutno še niso razvita v zadostni meri.

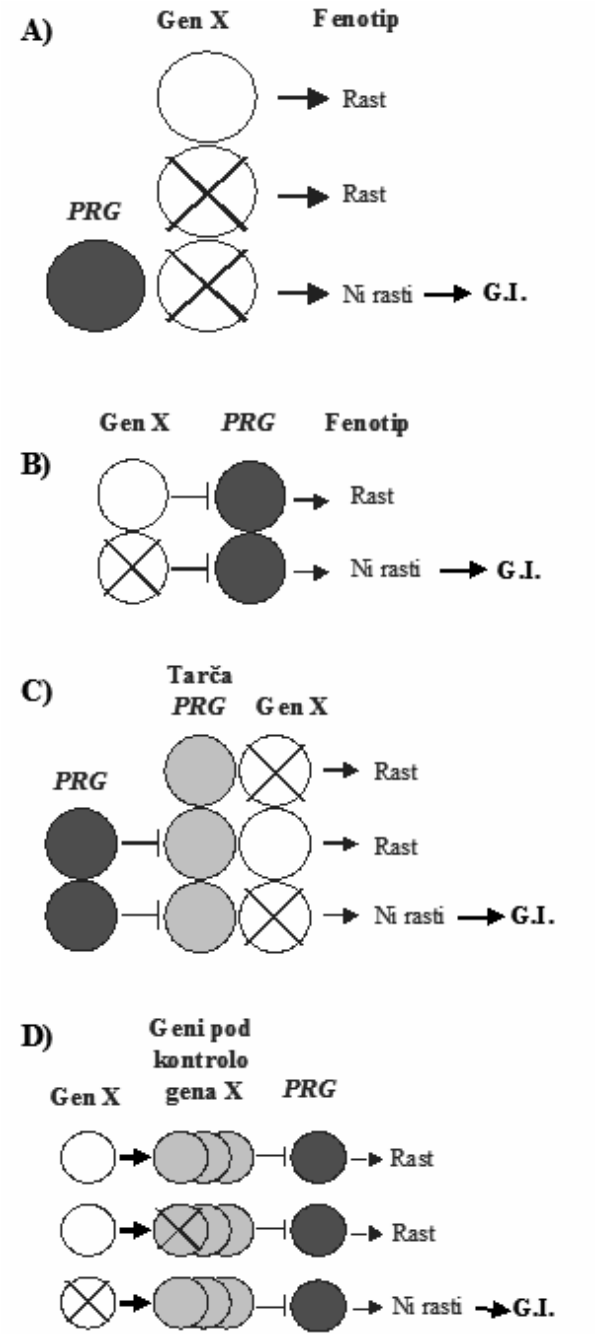
Orodja bioinformatike so nepogrešljiva pri prevajanju genoma z metodami funkcijske genomike na raven sistemske biologije in fenomike. Pri analizi fenotipov dvojnih mutant se na primer izkaže, da je število možnih genetskih mrež že pri manj kot desetih genih tako veliko, da odkrivanje mrež iz podatkov zahteva računalniško obdelavo in razvoj formaliziranih postopkov.[10] Pri fenomiki pa imamo opraviti na primer z analizo fenotipov vseh enojnih in dvojnih mutant, kar v primeru preprostega organizma kot je kvasovka pomeni približno 6.000 enojnih in 6.000 x 5.999 / 2 = 17.997.000 dvojnih mutant. Zato so potrebna nova orodja, ki so sposobna analize takšne količine podatkov in ki so hkrati sposobna integracije dodatnih podatkov o mutantah za reševanje konfliktov, do katerih neizogibno pride zaradi kopičenja napak pri eksperimentalnih pristopih na ravni celotnega genoma oziroma proteoma. Dodatni podatki so lahko raznovrstni, pomembno je le, da so relativno zanesljivi (na primer iz objavljenih ciljanih študij).

Za ponazoritev lastnosti orodij bioinformatike, ki so potrebna za analizo na ravni fenomike in sistemske biologije, vzemimo primer nadzorovanega oziroma prekomernega izražanja preiskovanega gena v zbirki sevov s sistematično izbitimi vsemi neesencialnimi geni. Nedavno je bila razvita tehnika, ki omogoča sistematično uvajanje dvojnih mutacij v genom kvasovke, z namenom določitve genetskih interakcij.[11] Genetska interakcija je definirana kot interakcija med dvema genoma, kjer ima mutacija v obeh genih za posledico fenotipsko lastnost, ki se ne pojavlja pri nobeni od posameznih enojnih mutacij. V opisanem primeru dobimo dvojne mutante, kjer je posledica ene mutacije prekomerno izražanje preiskovanega gena, posledica druge mutacije pa odsotnost izražanja drugega gena. Takšne dvojne mutante so uporaben model za določitev molekulske osnove delovanja preiskovanega gena v celici kvasovke kot modelu.[12]

Orodja bioinformatike za tovrstno analizo delovanja prekomerno izražanih genov še niso razvita, pričujoči primer pa nakazuje eno od smeri razvoja algoritmov, ki bodo predvidoma uporabljeni v teh orodjih. Z analizo ene same fenotipske lastnosti, zmanjšane hitrosti rasti, lahko določimo nabor genov, ki so v genetski interakciji s preiskovanim genom (slika 1A). Možnih razlag, zakaj obstaja genetska interakcija med preiskovanim genom in nekim drugim genom v genomu kvasovke, je več (slike 1B-D). Gen, ki je v genetski interakciji s preiskovanim genom, lahko inhibira delovanje preiskovanega gena, ki potencialno negativno vpliva na rast celice[12] (slika 1B); gen je lahko funkcijski homolog gena, ki je

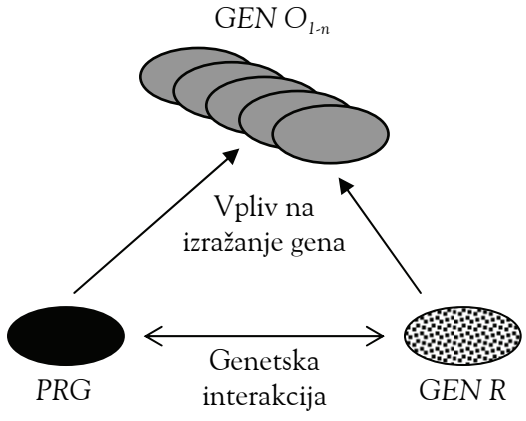
tarča neposrednega inhibitornega delovanja preiskovanega gena (slika 1C); gen je lahko aktivator skupine genov, ki inhibira delovanje preiskovanega gena, ki sicer negativno vpliva na rast celice (slika 1D). Že če upoštevamo samo eno dodatno fenotipsko lastnost, na primer povečanje hitrosti rasti kot posledico genetske interakcije, se število možnih mehanizmov delovanja preiskovanega gena bistveno poveča. Da bi ugotovili, kateri mehanizem je v danem primeru ustrezen, je potrebno poznati še druge lastnosti fenotipa kot samo hitrost rasti. Kadar v literaturi ni na voljo dovolj podatkov o določenem genu, je najhitrejša pot za določitev kompleksnega fenotipa, povezanega s preučevanim genom, določitev vpliva mutacije tega gena na transkriptom, saj ta način zahteva v teoriji le en eksperiment z uporabo DNA mikromreže s celotnim genomom. Na profil izražanja genoma mutiranega seva lahko gledamo kot na poseben fenotip, iz katerega je moč razbrati dinamično komponento genoma povezano s preučevanim genom. Podobno lahko postopamo tudi tedaj, ko želimo analizirati fenotipski odziv na izražanje eksogenega gena.

Kot primer določitve molekulske osnove delovanja eksogenega gena v kvasovki smo uporabili amoditoksin, večfunkcijsko fosfolipazo A_2, ki ima na sesalcih raznovrstne patofiziološke učinke, med njimi tudi nevrotoksičnega. Najprej smo določili vpliv nadzorovanega izražanja amoditoksina v celici kvasovke na izražanje celotnega transkriptoma, s čimer smo dobili kompleksno sliko fenotipa. Spremembe v izražanju smo zasledili predvsem pri genih, ki so neposredno udeleženi v odzivu celice na izražanje amoditoksina. Nato smo gen za amoditoksin s križanjem prenesli in nato nadzorovano izrazili v zbirki sevov z izbitimi vsemi posamičnimi neesencielnimi geni ter analizirali rast tako dobljenih dvojnih mutant.



Slika 1 Genetske interakcije. A: Osnovna definicija genetske interakcije med prekomerno izraženim preiskovanim genom in kvasnim genom (Gen X). B-D: Nekatere možne relacije med geni, ki privedejo do genetske interakcije (za podrobnosti glej tekst). Okrajšave: *PRG* – preiskovani gen; G.I. – genetska interakcija.

Počasnejša rast dvojne mutante kot posameznih enojnih mutant (genetska interakcija) kaže na funkcijsko povezavo med preiskovanim eksogenim genom in genom, ki je mutiran.[11-13] S to metodo smo identificirali predvsem regulatorne gene, ki posredno ali neposredno uravnavajo izražanje genov, identificiranih z zgoraj opisano analizo vpliva izražanja amoditoksina na transkriptom (slika 2). Naši eksperimenti so torej pokazali, da dobimo z določitvijo vseh genetskih interakcij preiskovanega gena ter vseh genov, ki se jim raven izražanja značilno spremeni kot posledica prekomernega izražanja preiskovanega gena, komplementarne rezultate. Za analizo tovrstnih rezultatov so potrebna nova orodja bioinformatike, ki kombinirajo analizo transkriptomskih podatkov z analizo mutant ter vključujejo tudi vse ostale razpoložljive podatke, kar pa se lahko hitro sprevrže v zelo kompleksen kombinatorični problem. Zato je nujno potrebno razviti in implementirati različne hevristične in optimizacijske pristope k reševanju tovrstnih kompleksnih problemov. Drugi pristop k obvladovanju kompleksnosti je večnivojsko in postopno odkrivanje in opisovanje rezultatov, od splošnega, kjer pokažemo le najbolj očitne lastnosti, do zelo podrobnega nivoja.[14]



Slika 2 Shematski prikaz tipičnega rezultata, dobljenega s kombinacijo identifikacije genetskih interakcij ter vpliva preiskovanega gena na transkriptom. Okrajšave: *PRG* – preiskovani gen; *GEN R* – regulatorni gen; *GEN* O_{1-n} – geni, ki so pod kontrolo regulatornega gena *GEN R*.

Nazorni prikaz dobljenih rezultatov, denimo v obliki genetskih mrež ali podobnih oblik hkratnega prikazovanja večje množice odkritih relacij med geni oziroma opazovanimi funkcionalni deli, je lahko odločilnega pomena za uspešno interpretacijo rezultatov in načrtovanje dodatnih analiz, za kar je spet potrebno razviti dodatna orodja. Pomembno orodje za vizualizacijo kompleksnih mrež je programski paket Pajek[15] (slika 3).

Slika 3 Mreža interakcij med amoditoksinom (v sredini grafa) in kvasnimi geni ter proteini, prikazana s programom Pajek. Različne oblike prikazujejo anotacije genov/proteinov: karo – gen v genetski interakciji z amoditoksinom; trikotnik – gen, na katerega izražanje vpliva amoditoksin; krog – transkripcijski faktor; kvadrat – ostali geni/proteini. Puščice prikazujejo vpliv transkripcijskih faktorjev na izražanje genov, polne črte genetske interakcije in prekinjene črte vpliv amoditoksina na izražanje genov. Številke označujejo kvantitativen vpliv amoditoksina na izražanje genov, beseda "lethal" pa podtip genetske interakcije, kjer je inhibicija rasti dvojne mutante popolna.

Pri razvoju orodij bioinformatike moramo še posebej stremeti k njihovi interaktivnosti, ki omogoča uporabniku, da enostavno in hitro preveri različne hipoteze. Čeprav so mnogokrat osnova numerični, kvantitativni podatki, je potrebno razviti in implementirati metode, ki generirajo simbolne in zato potencialno bolj

razumljive modele. Primer takih kvalitativnih modelov so tudi genske mreže.

Zaključek

V današnji poplavi javno dostopnih genetskih podatkov na spletu morajo orodja bionformatike ob navajanju rezultatov ponuditi povezave do relevantnih podatkov in objavljenih izsledkov, ki jih lahko raziskovalec uporabi za dodatno podkrepitev ali ovržbo dobljenih rezultatov in tako bistveno pospeši proces odkrivanja novih zakonitosti.

Zadnje in morda najbolj pomembno področje, na katero lahko posežejo orodja bioinformatike, pa je razvoj pristopov, ki pomagajo genetiku pri načrtovanju in izvajanju novih eksperimentov. Čeprav je razvoj tovrstnih tehnik za podporo raziskavam v sistemski biologiji še v povojih, nekaj nedavnih objav na tem področju priča, da je z uporabo pristopov umetne inteligence moč vsaj do določene mere avtomatizirati načrtovanje bioloških eksperimentov za potrebe odkrivanje novih bioloških znanj.[16,17]

Orodja bioinformatike na področju fenomike bodo omogočila razmeroma hitro in natančno določitev molekulske funkcije človeških genov na genomski ravni, kakor hitro bo razvoj eksperimentalnih tehnik dovolj napredoval, da bo omogočal analize, primerljive z modelnimi organizmi. Zato predstavljajo takšna orodja pomemben korak k sistemski biologiji, tako pri modelnih organizmih kot pri človeku.

Literatura

1. Goffeau A, Barrell BG, Bussey H, et al.: Life with 6000 genes. Science 1996; 274(5287): 546, 563-567.
2. DeRisi JL, Iyer VR, Brown PO: Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997; 278(5338): 680-686.
3. Zhu H, Bilgin M, Bangham R, et al.: Global analysis of protein activities using proteome chips. Science 2001; 293(5537): 2101-2105.
4. Winzeler EA, Shoemaker DD, Astromoff A, et al.: Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. Science 1999; 285(5429): 901-906.
5. Huh WK, Falvo JV, Gerke LC, et al.: Global analysis of protein localization in budding yeast. Nature 2003; 425(6959): 686-691.
6. Sopko R, Huang D, Preston N, et al.: Mapping pathways and phenotypes by systematic gene overexpression. Mol Cell 2006; 21(3):319-330.
7. Gavin AC, Aloy P, Grandi P, et al.: Proteome survey reveals modularity of the yeast cell machinery. Nature 2006; v tisku.
8. Kamath RS, Fraser AG, Dong Y, et al.: Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. Nature 2003; 421(6920):231-237.
9. Kerem B, Rommens JM, Buchanan JA, et al.: Identification of the cystic fibrosis gene: genetic analysis. Science 1989; 245(4922):1073-1080.
10. Zupan B, Demšar J, Bratko I, et al.: GenePath: a system for automated construction of genetic networks from mutant data. Bioinformatics 2003; 19(3): 383-389.
11. Tong AH, Lesage G, Bader GD, et al.: Global mapping of the yeast genetic interaction network. Science 2004; 303(5659):808-813.
12. Sopko R, Huang D, Preston N, et al.: Mapping pathways and phenotypes by systematic gene overexpression. Mol Cell 2006; 21(3):319-330.
13. Ye P, Peyser BD, Pan X, et al.: Gene function prediction from congruent synthetic lethal interactions in yeast. Mol Systems Biol 2005; msb4100034-E1.
14. Bornholdt S: Systems biology. Less is more in modeling large genetic networks. Science 2005; 310(5747): 449-451.
15. Batagelj V, Mrvar A: Pajek: program for large network analysis. http://vlado.fmf.uni-lj.si/pub/networks/pajek, 2006.
16. King RD, Whelan KE, Jones FM, et al.: Functional genomic hypothesis generation and experimentation by a robot scientist. Nature 2004; 427(6971): 247-252.
17. Zupan B, Bratko I, Demšar J, et al.: GenePath: a system for inference of genetic networks and proposal of genetic experiments. Artif Intell Med 2003; 29(1-2): 107-130.

Bilten SDMI ■

Zaključki kongresa "Zdravje na informacijski poti" (MI 2006), Zreče, 9.-11.4.2006

O kongresu

Slovensko društvo za medicinsko informatiko je med 9. in 11. aprilom 2006 v Zrečah organiziralo tradicionalni kongres medicinske informatike MI 2006 z naslovom "Zdravje na informacijski poti". Dogodek je privabil 177 udeležencev iz Slovenije in tujine, kar je največja udeležba na kongresih in strokovnih srečanjih društva doslej. Kongres je bil namenjen predstavitvi aktualnega dogajanja na področju zdravstvene informatike s poudarkom na uresničevanju nacionalne strategije eZdravje.

V okviru kongresa sta bili izvedeni delavnici na temo prenove zakona o zbirkah podatkov s področja zdravstva in na temo uporabe informacijskih tehnologij pri informiranju ter izobraževanju bolnikov in zdravstvenih delavcev.

Obravnavana so bila strateška področja zdravstvene informatike - standardizacija, baze podatkov, portali, poročanje, kartica zdravstvenega zavarovanja, komunikacije, elektronski zdravstveni zapisi, informacijska podpora čakalnim vrstam in upravljanju kakovosti v zdravstvu.

Predstavljene so bile nekatere izkušnje iz tujine (Danska, Velika Britanija, Srbija).

Pod pokroviteljstvom Ministrstva za zdravje je bila organizirana delavnica na temo uresničevanja strategije eZdravje, v okviru katere so bile obravnavane načrtovane zagonske naloge - konsolidacija infrastrukture, vzpostavitev zdravstvenega portala, informacijska podpora čakalnim dobam. Obravnavani so bili tudi organizacijski in finančni aspekti uresničevanja.

Na kongresu so bili predstavljeni aktualni projekti, obravnavana je bila informatika v zdravstveni negi. Krajša delavnica je bila namenjena obravnavanju področja varnosti informacij v zdravstvu. Predstavljenih je bilo tudi nekaj rešitev iz najsodobnejših strokovnih področij (bioinformatika, inteligentni sistemi).

Slika 1 Sekcija Informatika v zdravstveni negi uspeva mnogo bolje, kot je to napovedoval urednik te revije.

Zaključki kongresa

Uresničevanje strategije eZdravje

Slovensko društvo za medicinsko informatiko podpira pripravljeno strategijo eZdravje Ministrstva za zdravje RS in obenem opozarja na naslednje pomembne dejavnike pri uresničevanju strategije:

- Strategija eZdravje opredeljuje ambiciozne cilje za razvoj informacijskih rešitev v zdravstvu. Za izvedbo nadvse aktualnih, a zahtevnih projektov je potrebno zagotoviti finančne in kadrovske vire ter ustrezno koordinacijo.

- Predloge o širjenju uporabe omrežja državne uprave (omrežje HKOM) v zdravstvo je potrebno podrobno pretehtati in ugotoviti možnosti alternativnih rešitev in uporabe javnega internetnega omrežja.

- Standardizacija se kaže kot nujna infrastrukturna aktivnost na več področjih:

    - definiranje enotnih pomenov podatkov (podatkovni slovarji),

    - definiranje enotnih naborov, formatov podatkov za vodenje zbirk in za izmenjevanje podatkov med subjekti v zdravstvu,

    - varovanje informacij.

  Zato je potrebno čim preje ustanoviti nacionalne organe za koordinacijo aktivnosti na področju standardizacije in pristopiti k razvoju standardov za prednostne projekte.

- Zagotoviti je potrebno podporo in širitev področij delovanja uspešnim (tekočim) projektom ter prenos teh modelov delovanja v nove projekte (npr. uporaba izkušenj kakovosti v zdravstvu pri podpori projekta čakalnih dob).

- Ministrstvo za zdravje naj se pridruži pobudi MVŠZT, ki bo v koordinaciji Direktorata za informacijsko družbo že v letu 2006 pričela uresničevati nacionalni projekt "Varen dom" kot modelni dom tudi za ureditev domačega okolja bolnika za izvajanje dolgotrajne nege in zdravljenja na domu.

Slika 2 Občni zbor je minil mirno, kot že dolgo nobeden.

Ključni dejavniki za nadaljnjo informatizacijo zdravstva

Na konferenci so bili izpostavljeni naslednji ključni dejavniki za nadaljnjo informatizacijo v slovenskem zdravstvu:

- Spodbujati uvajanje konkretnih informacijskih rešitev, ki pripomorejo k izboljšanju kakovosti zdravstvenih storitev in k racionalizaciji poslovanja na način finančnih spodbud s strani ministrstva in ZZZS.

- Zagotavljanje ustrezne varnosti medicinskih podatkov je vse bolj izražena zahteva uporabnikov teh storitev. Nalog se je potrebno lotiti postopno (najprej zagotoviti najnujnejše ukrepe) in racionalno (ukrepe, ki veljajo za večino izvajalcev izvesti sočasno in enotno za podobne zdravstvene organizacije npr. zdravstvene domove, bolnišnice itd.). Zagotoviti je potrebno tudi ustrezno neodvisno preverjanje na nacionalni ravni. Na delavnici s

področja varnosti informacijskih sistemov je bil podan in podprt predlog za organizacijsko rešitev prenosa varnostno informacijskih standardov v prakso; konsolidacijo varnostnih standardov za podobne zdravstvene organizacije (npr. zdravstveni domovi, bolnišnice …); ter poenostavljeno preverjanje ustreznosti varnostnim standardom, kar bi pomenilo lažjo, hitrejšo in cenejšo uvedbo varnosti v zdravstvene informacijske sisteme.

- Priložnosti uporabe informacijskih tehnologij je poleg v institucionalnem zdravstvu potrebno izkoristiti tudi pri zagotavljanju zdravstvene oskrbe na domu.

- Udeleženci delavnice na temo uporabe informacijskih tehnologij pri informiranju ter izobraževanju bolnikov in zdravstvenih delavcev so opozorili da trenutna prizadevanja za vključitev informacijsko komunikacijske tehnologije v proces izobraževanja temeljijo na delu posameznikov entuziastov. Veliko več pozornosti je potrebno nameniti izobraževanju zdravstvenih delavcev s področja informatike v teku rednega izobraževanja na srednjih šolah in fakultetah, kakor tudi ob delu.

Slika 3 Predsednik programskega odbora je bil še enkrat več zadovoljen.

Prenova Zakona o zbirkah podatkov s področja zdravstvenega varstva

Udeleženci delavnice o prenovi Zakona o zbirkah podatkov s področja zdravstvenega varstva so oblikovali naslednje zaključke:

- Čim prej je potrebno imenovati delovno skupino, ki bo pripravila strokovne podlage za predlog novega zakona. Če bo predlog pripravljen za vladno proceduro do konca oktobra 2006, bo mogoče zakon sprejeti aprila 2007.

- Skupina mora biti interdisciplinarna, v njej morajo že od začetka sodelovati tudi pravniki.

- Potreben je naslednji pristop k pripravi zakona:

    - definirati informacije, ki jih deležniki potrebujejo,
    - definirati indikatorje, ki iz tega izhajajo,
    - opredeliti namen in cilje zbiranja podatkov,
    - opredeliti nabor podatkov,
    - opredeliti definicije podatkov,
    - opredeliti način zbiranja podatkov,
    - opredeliti ustrezno zakonsko ureditev.

- Z zbirkami podatkov je potrebno zagotoviti tako spremljanje zdravstvenega sistema kot zdravja prebivalstva.

- Potrebno je upoštevati evropske direktive o varovanju podatkov in veljavni Zakon o varstvu osebnih podatkov.

- Opredeliti je potrebno vloge letnih programov raziskovanj (po vzoru ureditve na področju državne statistike) kot možne ureditve

anketnega zbiranja podatkov, povezovanja podatkovnih zbirk, razvojnih projektov, …

- Glede priprave zakona obstajajo naslednja odprta vprašanja:
    - priprava novele zakona ali priprava novega zakona,
    - osebni identifikatorji v zdravstvu (poleg KZZ, tudi uporaba EMŠO),
    - vključitev podatkov o zdravstvenem zavarovanju.

Pri pripravi zakona je potrebno razmišljati tudi o natančni opredelitvi zahtev za elektronsko podpisovanje in elektronski arhiv podatkov v zdravstvu, s čimer bi zagotovili pravno enakovrednost elektronskega in papirnega arhiva.

Pripravi zakona mora slediti priprava podzakonskih aktov, ki bodo opredelili tehnološke izvedbene pogoje (varnost osebnih podatkov, izmenjave podatkov itd).

Slika 4 Ti, a naj tole pritisnem?

Problemi, ki zahtevajo nujno reševanje

Udeleženci kongresa so opozorili tudi na naslednje konkretne probleme, ki se pojavljajo pri izvajanju projektov in rednem delu v zdravstvu:

- Nujno je potrebno poenotiti pristope pri oblikovanju kazalnikov za spremljanje kakovosti dela v bolnišnicah. Trenutno na teh področjih med seboj neusklajeno delujeta zdravniška zbornica in ministrstvo.
- Nujno je potrebno poenotiti obstoječe baze podatkov o zdravilih in zagotoviti kakovostno in enotno zbirko teh podatkov na nacionalnem nivoju, saj je takšna zbirka nujna podlaga za bolnišnične projekte, elektronski recept in druge projekte.
- Nujno je potrebno zagotoviti pravne podlage in tehnične rešitve za uporabo baze podatkov o izvajalcih pri vseh subjektih zdravstva, saj je ta zbirka podatkov ključnega pomena za komuniciranje med subjekti v zdravstvu in podlaga za številne projekte.

Slika 5 Profesor Pajntar? Tole ni porodnišnica, veste! Saj se mi je zdelo, toliko žensk pa nobena noseča!

Pripravili: Ivan Eržen (predsednik SDMI), Tomaž Marčun, Polonca Truden Dobrin, Vesna Prijatelj, Brane Leskošek, Marija Trenz (fotografije)

■ **Infor Med Slov**: 2006; 11(1): 72-75