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

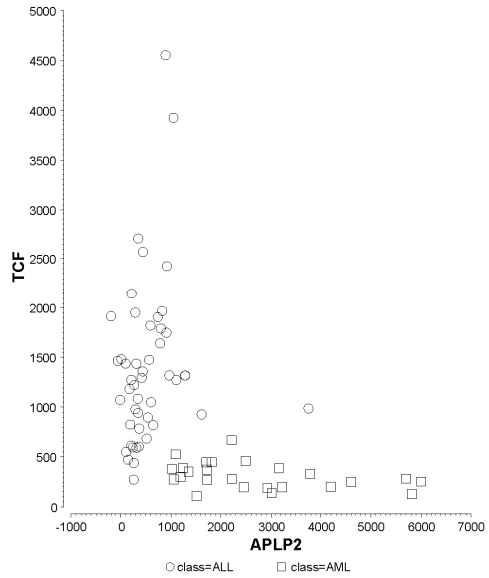
Rakasta obolenja so posledica progresivnih genetskih in epigenetskih sprememb, ki vodijo pretvorbo normalnih celic v njihove maligne derivate. Genetske spremembe so predvsem mutacije v onkogenih in tumor supresorskih genih, medtem ko epigenetski mehanizmi uravnavajo prepisovanje genov preko različnih 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]

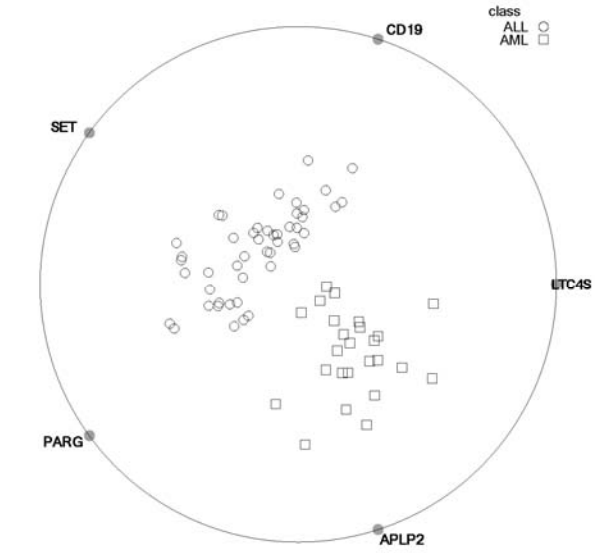
Največji problem podatkov, ki jih pridobimo z merjenji izražanja genov, je njihova visoka dimenzionalnost, saj navadno vključujejo več tisoč spremenljivk (genov) in majhno število vzorcev (bolnikov). Poleg tega so meritve izredno občutljive na zunanje dejavnike, zato je v podatkih pogosto prisotno veliko 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]

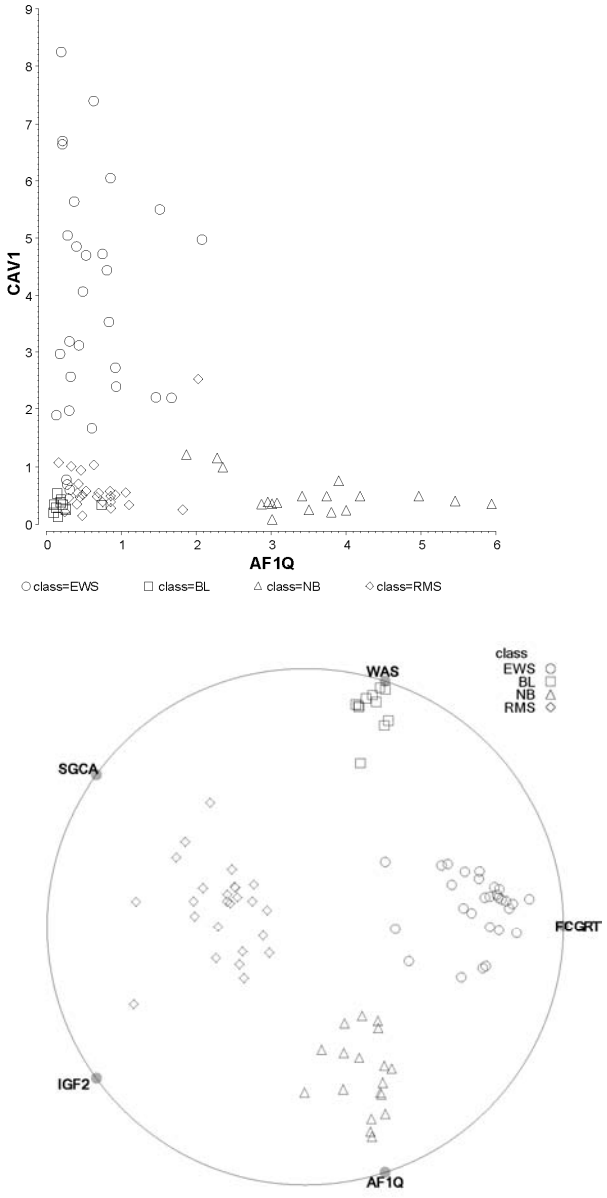
Metoda VizRank

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).

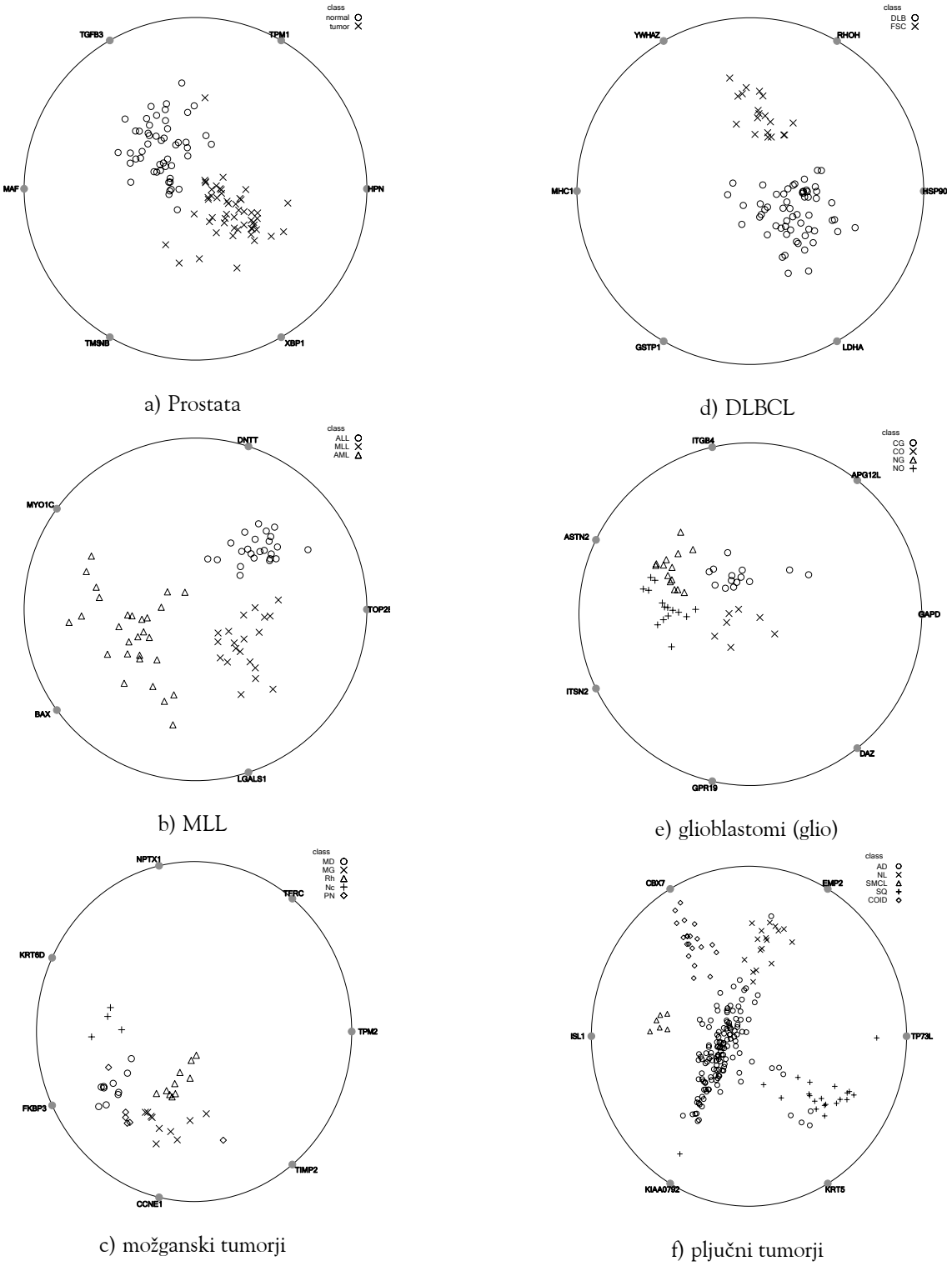
Poskusi in rezultati

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.



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.