Izvirni znanstveni članek ■

Uporaba skupin genov pri analizi podatkov o izraženosti genov pri raku

Minca Mramor, Marko Toplak, Tomaž Curk, Blaž Zupan

Izvleček. Uporaba skupin genov je močno izboljšala ujemanje med rezultati analize podatkov o izraženosti genov različnih raziskovalnih skupin in izboljšala napovedne točnosti modelov. V prispevku podamo kratek pregled metod za analizo podatkov o izraženosti genov na nivoju skupin genov in opišemo najpomembnejše baze znanj s podatki o izraženosti genov in o skupinah genov. Predstavimo nadgradnjo metode GSEA, ki omogoča izračun obogatenosti skupin genov v posameznem vzorcu, in je primerna za razvrščanje vzorcev na podlagi izraženosti skupin genov. Napovedna točnost metode podpornih vektorjev na tako pretvorjenih podatkih se ne spremeni, rezultate pa je moč lažje interpretirati zaradi uporabljenega predznanja o skupinah genov.

Utility of gene-sets in the analysis of cancer gene expression data

Abstract. The overlap between the results of different research groups studying the same cancer types is significantly improved if, instead of looking at individual genes, sets of genes with the same biological or molecular function are considered. In the paper, we present a short overview of the gene set analysis methods. We describe an extension to the GSEA method that is able to score gene sets in individual samples. We show that the classification performance of support vector machines is similar on the transformed and original data, but the models are – due to the use of domain knowledge – easier to interpret.

■ **Infor Med Slov**: 2008; 13(2): 1-10

Institucija avtorjev: Univerza v Ljubljani.

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

Uvod

Rak je bolezen genomskih sprememb: spremembe v zaporedju DNA, kromosomske preureditve in modifikacije, metilacije DNA, podvajanja in delecije genov skupaj vodijo v nastanek in napredovanje rakastih bolezni. Posledično je v rakasti celici motena regulacija transkripcije genov, kar vodi v spremenjeno izraženost mnogih genov. Dolgo je bilo za raziskovalce moteče dejstvo, da je pri raziskavah o izraženosti genov istega tipa raka ujemanje med najbolj spremenjeno izraženimi geni navadno izredno majhno. V zadnjem času pa je več študij pokazalo, da je ujemanje mnogo večje, če se iz nivoja genov dvignemo na nivo izbranih skupin genov. Kot skupine so navadno obravnavani geni, ki skupaj sodelujejo v bioloških poteh ali imajo v genski ontologiji pripisano isto molekularno funkcijo ali biološki proces (angl. *annotation*). Najnovejši rezultati tako kažejo, da moramo namesto k spremembam v posameznih genih pogled preusmeriti na funkcijske poti, v katerih ti geni nastopajo.[1-3]

Podatke o izraženosti genov lahko analiziramo na nivoju skupin genov na dva glavna načina:[4,5] (1) analiza posameznih genov (angl. *individual gene analysis*), kjer na podlagi seznama diferencialno izraženih genov določimo, katere skupine genov so zastopane bolj pogosto kot bi pričakovali po naključju in (2) analiza skupin genov (angl. *gene set analysis*), kjer gene najprej rangiramo glede na korelacijo med izraženostjo in fenotipom, ki ga opazujemo, nato pa vnaprej določene skupine genov ocenimo glede na izračunano korelacijo.

Pri analizi posameznih genov za vsak gen določimo ali je značilno diferencialno izražen med skupinami vzorcev, ki jih primerjamo. Rezultat take analize je množica genov, ki so izraženi nad določenim vnaprej postavljenim pragom. Številne metode in orodja nam nato omogočijo, da take množice primerjamo z biološko določenimi skupinam in na podlagi podatkov ugotovimo, katere skupine genov so zastopane bolj pogosto kot bi pričakovali glede na naključno porazdelitev.[4,5] Glavni problemi analize posameznih genov so velik vpliv izbranega praga, pogosto spregledane skupine genov pod postavljeno mejo, ki so v preiskovanih tkivih sicer različno izraženi, in napačna predpostavka o neodvisnosti izražanja genov.[5] Pregled metod za analizo posameznih genov je lepo podan v Khatri in Draghici.[6]

V zadnjem času je bila razvita vrsta alternativnih metod za analizo skupin genov, med katerimi je najbolj znana in uporabljana metoda analize obogatenosti skupin genov (angl. *gene set enrichment analysis*, GSEA).[7] Prednost teh metod je, da ne uporabljajo praga temveč uporabijo metriko, ki dobro oceni skupine genov, ki imajo zmerne, vendar skladne sprembe v izraženosti.[5] Izrazito učinkovitost pristopa so prvič prikazali Mootha in sod.[8] v raziskavi o izraženosti genov v mišicah bolnikov s sladkorno boleznijo tipa 2. Pokazali so, da z metodami analize posameznih genov niti en gen ni bil značilno diferencialno izražen med tkivi bolnikov s sladkorno boleznijo in posameznikov z glukozno intoleranco, medtem ko je metoda GSEA odkrila skupino genov vključenih v oksidativno fosforilacijo, katere raven izražanja je bila značilno znižana pri bolnikih s sladkorno boleznijo.

V članku bomo na kratko predstavili metode in orodja za analizo skupin genov in njihove nadgradnje ter opisali najpomembnejše javno dostopne baze znanj s podatki o izraženosti genov in o skupinah genov. Predstavili bomo novo metodo za izračun obogatenosti skupin genov za posamezne vzorce in jo primerjali z metodo ASSESS (angl. *Analysis of Sample Set Enrichment Scores*),[9] ki omogoča izračun obogatenosti skupin za posamezne vzorce v naborih podatkov z dvema razredoma. Pokazali bomo, da so napovedne točnosti metode podpornih vektorjev v prostoru skupin genov primerljive s tistimi v prostoru genov na naborih podatkov, ki vključujejo dva diagnostična razreda. Poleg tega so dobljeni modeli biološko lažje razložljivi.

Pregled metod za analizo genskih skupin

Glede na ničelno hipotezo, ki jo metode analize genskih skupin testirajo, sta jih Goeman in Buhlmann[10] razdelila na tekmovalne (angl. *competitive*) in samostojne (angl. *self-contained*). Posebno mesto zasedata metodi GSEA[7] in njena različica Gene Set Analysis (GSA),[11] ki ju glede na testirani hipotezi uvrščamo med mešane metode.

Tekmovalne metode

Metode iz te skupine tesirajo hipotezo, da je povezanost skupine genov z izbranim fenotipom enaka kot povezanost komplementa izbrane skupine. Tekmovalne metode tako ugotavljajo relativno obogatenost diferencialno izraženih genov v skupini genov v primerjavi s skupino vseh ostalih genov in iščejo skupine genov s koordiniranimi spremembami v izraženosti.[5,10] Primeri metod iz te skupine so PAGE,[12] ErmineJ[13] in ASSESS.[9]

Samostojne metode

Samostojne metode upoštevajo le gene v določeni skupini in testirajo ničelno hipotezo, da noben gen v skupini ni povezan s fenotipom. V tem primeru lahko že en sam diferencialno izražen gen iz skupine genov vpliva na značilno obogatenost te skupine. Zaradi te lastnosti lahko samostojne metode odkrijejo mnogo več obogatenih skupin genov kot tekmovalne.[5,10] Primeri samostojnih metod so PLAGE,[14] Goemanov globalni test[14] in SAM-GS.[15]

Mešane metode – metodi GSEA in GSA

Najbolj znana metoda za analizo podatkov o izraženosti genov s pomočjo skupin genov je metoda GSEA.[7] GSEA testira hipotezo, da nobena od izbranih vnaprej določenih skupin genov ni povezana s fenotipom. Metoda GSEA najprej z izbrano univariantno statistiko (npr. t-test) uredi gene glede na korelacijo med dvema biološkima stanjema (npr. dve vrsti raka), nato pa uporabi uteženo Kolmogorov – Smirnov statistiko za oceno obogatenosti vsake posamezne skupine genov z diferencialno izraženimi geni.[7] GSEA spada med mešane metode, ker je tekmovalna glede na posamezne skupine genov in samostojna glede na celoten nabor podatkov.[5]

Metoda GSA je nadgradnja metode GSEA, ki namesto prirejene Kolmogorov-Smirnov statistike uporablja "maxmean" statistiko, ki ima večjo statistično moč. Maxmean statistika je povprečje absolutno večjega pozitivnega ali negativnega dela ocene genov v skupini genov. Velika prednost metode GSA je možnost ocenjevanja skupin genov na podatkih z več kot dvema razredoma in na podatkih s kvantitativnim izidom (npr. podatki o preživetju).[11]

Metode za izračun obogatenosti skupin genov za posamezne vzorce

Ena od glavnih pomanjkljivosti metode GSEA je, da obogatenost skupin genov v določenem biološkem stanju izračuna za celotno bazo podatkov naenkrat, ne pa za posamezne primere oz. vzorce. [9] Tako je metoda primerna za analizo eksperimentov in postavljanje hipotez o sodelovanju posameznih genskih poti pri opazovanem procesu. Ker obravnavajo celotni nabor podatkov, pa ni uporabna pri kliničnem odločanju, kjer nas na primer zanima karakterizacija posameznega vzorca. V pričakovanju kliničnih aplikacij in uporabe podatkov o izražanju genov pri klinični diagnostiki so metode, ki bi znale oceniti obogatenost skupin genov za posamezne vzorce že v razvoju. V članku predstavimo nedavno razvito metodo ASSESS, ter jo primerjamo z metodo, ki smo jo razvili sami.

Viri podatkov in baze znanj

Vse zgoraj opisane metode za analizo podatkov o izraženosti genov s pomočjo skupin genov potrebujejo vhodne podatke o izraženosti genov in vnaprej določene skupine genov. V tem poglavju bomo zato na kratko opisali najpomembnejše baze

znanj s podatki o genskih izrazih in s podatki o genskih skupinah.

Baze znanj s podatki o genskih izrazih

Na svetovnem spletu obstajajo številne javno dostopne (in zasebne) baze znanj s podatki o genskih izrazih. Grobo oceno o razsežnosti nam lahko prikaže poizvedba o "gene expression databases" v iskalniku Google, ki vrne približno 206.000 zadetkov. Ena od boljših strani (http://ihome.cuhk.edu.hk/~b400559/arraysoft_public.html), žal nazadnje posodobljena leta 2004, našteje 23 javno dostopnih baz podatkov o genskih podatkih, med katerimi velja poleg Gene Expression Omnibus (GEO) in ArrayExpress izpostaviti Stanford Microarray Database (genome-www5.stanford.edu, SMD).

Zagotovo najpomembnejši javno dostopni bazi znanj o podatkih pridobljenih z mikromrežami sta GEO (www.ncbi.nlm.nih.gov/geo), ki je vzpostavljena pod okriljem NCBI (angl. *National Center for Biotechnology Information*) in ArrayExpress (www.ebi.ac.uk/arrayexpress), ki deluje pod okriljem Evropskega inštituta za bioinformatiko. Ta status sta dosegli tudi z odločitvijo založniških skupin kot sta Nature in PLoS, da je potrebno pred objavo članka, ki vsebuje rezultate o podatkih pridobljenih z metodo DNA mikromrež, omogočiti javen dostop do podatkov na straneh GEO ali ArrayExpress. Za raziskave na področju raka sta pomembni tudi javno dostopni bazi Oncomine (http://www.oncomine.org) in baza inštituta Broad (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi), ki deluje pod okriljem Massachusetts Institute of Technology in Harvardske Univerze.

Kot zanimivost lahko omenimo zaključke predstavljene v članku Piwowar in sod.,[16] kjer avtorji ugotavljajo, da javni dostop do podatkov o genskih izrazih ne koristi le celotni znanstveni skupnosti, temveč tudi avtorjem članka, saj so članki z javno dostopnimi podatki statistično značilno bolj opazni in citirani.

Baze znanj s podatki o skupinah genov

Pri analizi podatkov o izraženosti genov s pomočjo skupin genov je sestava in izbor skupin prav tako pomembna kot izbor metode za analizo. Skupine genov so pripravljene z uporabo raznolikih virov biološkega znanja. To so, na primer, podatki o pripisanih funkcijah genom v genski ontologiji, podatki o funkcijskih in metabolnih poteh iz javnih baz kot so KEGG, GenMAPP in Biocarta, podatki o koekspresiji genov v podatkih pridobljenih z mikromrežami, in podobni.

Pri analizi podatkov s pomočjo skupin genov se moramo zavedati, da je natančnost rezultatov odvisna od kakovosti pripravljenih skupin genov. Glavne pasti pri uporabi genske ontologije, ki veljajo tudi za skupine genov sestavljene iz drugih baz podatkov, so predstavljene v članku Yon Rhee in sod.[17] Najpomembnejše so nepopolno biološko znanje, nenatančne ali nepravilne elektronske anotacije in urejanje baz s časovnim zamikom.

Največja baza znanj s podatki o skupinah genov je MSigDB[7] (http://www.broad.mit.edu/gsea/msigdb/index.jsp), pripravljena za uporabo v programu GSEA. Poleg možnosti prenosa skupin na osebni računalnik med drugim omogoča iskanje in pregledovanje skupin genov, računanje prekrivanja med skupinami in pregled pripisov, ki opisujejo skupino. MsigDB trenutno vsebuje podatke o 5452 skupinah genov, razdeljenih na pet glavnih zbirk, označenih s C1 do C5, glede na uporabljeno biološko znanje. Glavne lastnosti zbirk so predstavljene v Tabeli 1.

Druga pomembnejša baza znanj s podatki o skupinah genov pripada metodi GSA (http://www-stat.stanford.edu/~tibs/GSA/).[11] Skupine genov so sestavljene glede na lokacijo na kromosomu, celični proces in izraženost genov v določenih vrstah raka. Uporabljeno je biološko znanje zbrano v SMD (*Stanford Microarray Database*). Glavna pomanjkljivost obeh omenjenih baz znanj s podatki o skupinah genov je, da sta primerni predvsem za analizo podatkov o izraženosti genov pri človeku. Baza znanj MsigDB vključuje nekatere

skupine, ki so primerne tudi za analizo podatkov o šimpanzih, miših, podganah, prašičih, opicah in navadni cebrici (angl. *zebra fish*).

Vse ostale metode uporabljajo skupine genov, ki so zgrajene na podlagi genske ontologije ali na podlagi bioloških poti iz baz znanj KEGG in Biocarta.

Tabela 1 Zbirke skupin genov iz baze znanj MsigDB in v članku uporabljene podzbirke.

Zbirka	Pod-zbirka	Opis skupin genov	Št. skupin
C1		sestavljene glede na lokacijo na kromosomu	386
C2			1892
	CP	standardne biološke poti iz 12 javno dostopnih baz znanj o funkcijskih poteh	639
	CGP	kemijske in genetske perturbacije	1186
C3		geni, ki imajo enake cis-regulatorne motive	837
C4		skupine izračunane z metodami za odkrivanje znanja iz podatkov	883
C5		sestavljene na podlagi genske ontologije (GO)	1454
	CC	GO celična komponenta	233
	MF	GO molekularna funkcija	396
	BP	GO biološki proces	825

Metoda za izračun obogatenosti skupin za posamezne vzorce

Opis metode

Razvili smo metodo, ki na vsakem posameznem primeru oz. vzorcu omogoča izračun obogatenosti vnaprej določenih skupin genov.

Metoda na vsakem vzorcu:

1. za vsak gen izračuna razmerje dvojiškega logaritma med izraženostjo gena v danem vzorcu in povprečno izraženostjo v vseh ostalih vzorcih ne glede na diagnostični razred,
2. gene rangira glede na to razmerje,
3. na tako rangiranih genih uporabi metodo GSEA za ocenjevanje obogatenosti skupin genov.

Predlagana metoda tako omogoča transformacijo podatkov, kjer novi nabori podatkov vključujejo iste vzorce kot originalni nabori, kot spremenljivke pa namesto genov nastopajo skupine genov. Numerično vrednost posamezne skupine genov pri določenem vzorcu predstavlja normalizirana ocena obogatenosti.

Namen predlagane metode je klinična prognostika in diagnostika, kjer moramo posamezen primer oz. vzorec uvrstiti v določeno skupino oz. razred. Razvščanje v skupine (angl. *classification*) je sicer na področju analize podatkov o genskih izrazih na področju raka dobro raziskano, a študije pri tem kot napovedne spremenljivke uporabljajo posamezne gene in ne skupin genov. Uporaba skupin genov bi, tudi zaradi tipične šumnosti podatkov o izražanju genov, lahko vodila k bolj natančnim napovedim, predvsem pa bi lahko olajšala oz. omogočila vsebinsko razumevanje napovedi.

Za gradnjo napovednih modelov smo izbrali metodo podpornih vektorjev (SVM), ki na podatkih o izraženosti genov navadno dosega boljše napovedne točnosti od ostalih metod strojnega učenja.[18] Za analizo uspešnosti razvrščanja na podlagi skupin genov smo primerjali napovedno točnost (CA) in površino pod krivuljo ROC (angl. *receiver operating curve*, mera AUC) zgrajenih napovednih modelov. Uspešnost napovednih modelov smo ocenili z metodo desetkratnega prečnega preverjanja.

Uporabljeni nabori podatkov

Za eksperimentalni del študije smo uporabili sedem naborov podatkov o izraženosti genov pri različnih vrstah raka (tabela 2). Vsi nabori so javno dostopni na strani inštituta Broad (http://www.broad.mit.edu/cgi-bin/cancer/datasets.cgi), razen nabor podatkov Garber in sod.,

ki je dostopen na strani SMD (http://genome-www.stanford.edu/lung_cancer/ adeno/). Nabori vsebujejo podatke o izraženosti od 7070 do 12625 genov pri 40 do 230 bolnikih z rakom. Vzorci so razvrščeni v dve do pet diagnostičnih skupin (različnih podvrst določenega raka).

Tabela 2 Nabori podatkov.

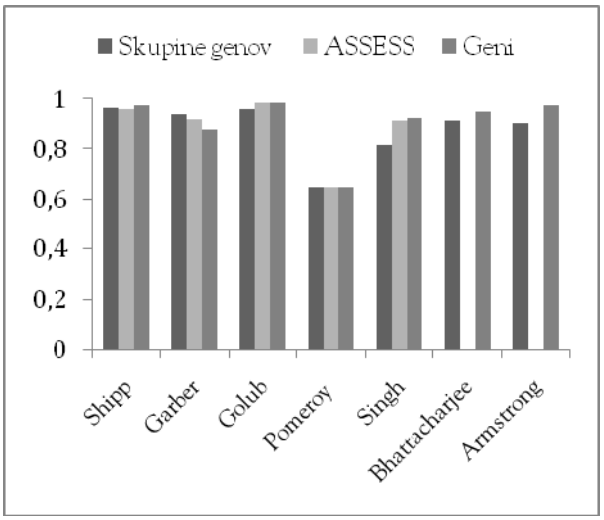
Nabor podatkov	Št. vzorcev	Št. genov	Št. razredov	Vrsta raka
Garber	50	12625	2	pljučni
Golub	72	7074	2	levkemija
Pomeroy	40	7129	2	možgani
Singh	102	12533	2	prostata
Shipp	77	7070	2	DLBCL
Armstrong	72	12533	3	MLL
Bhattacharjee	203	12600	5	Pljučni

Uporabljene skupine genov

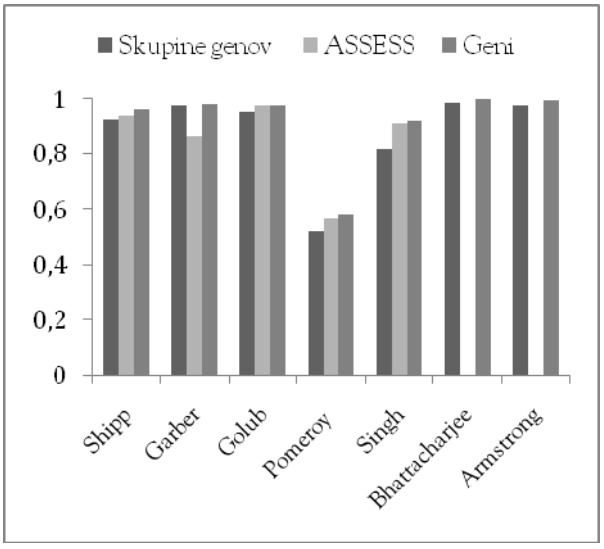
V raziskavi smo uporabili skupine genov iz baze znanj MsigDB. Uporabili smo tisti del zbirke C2, ki vsebuje standardne poti (C2, CP) in zbirko zgrajeno na podlagi genske ontologije (C5), dela, ki združujeta gene v skupine glede na enako molekularno funcijo (C5, MF) in biološki proces (C5, BP). Uporabili smo le skupine genov z manj kot 100 geni.

Rezultati naše metode in primerjava z metodo ASSESS

Primernost naše metode za transformacijo podatkov o genski ekspresiji smo ocenili s primerjavo napovednih točnosti modelov zgrajenih na transformiranih podatkih (podatkih, ki kot spremenljivke vsebujejo obogatenost skupin genov) z napovedno točnostjo modelov zgrajenih na originalnih podatkih (podatki, ki kot spremenljivke vsebujejo izražanje genov). Napovedne modele smo gradili z metodo SVM, napovedno točnost pa smo ocenili z merama CA in AUC, dobljenih z metodo desetkratnega prečnega preverjanja. Primerjava napovednih točnosti je prikazana na grafu 1, primerjava mer AUC pa na grafu 2.



Graf 1 Primerjava napovednih točnosti metode podpornih vektorjev, dobljenih z desetkratni prečnim preverjanjem na podatkih s skupinami genov (naša metoda in metoda ASSESS) in na originalnih podatkih o izraženosti genov.



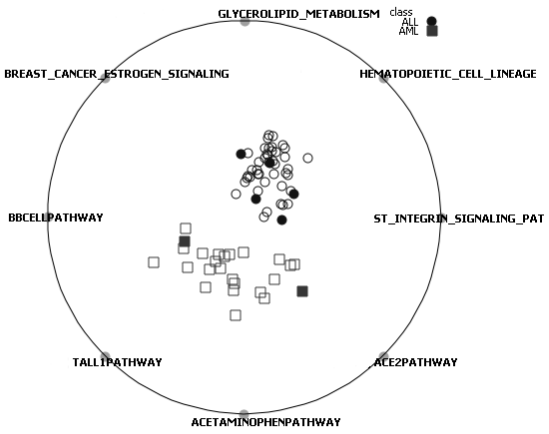
Graf 2 Primerjava mer AUC metode podpornih vektorjev, dobljenih z desetkratni prečnim preverjanjem na podatkih s skupinami genov (naša metoda in metoda ASSESS) in na originalnih podatkih o izraženosti genov.

V primerjavo smo vključili tudi metodo ASSESS,[9] ki je prav tako kot predstavljena metoda nadgradnja metode GSEA.[7] Omogoča izračun obogatenosti skupin genov za vsak posamezen

primer v naboru podatkov. Glavna razlika med našo metodo in metodo ASSESS je, da ASSESS pri ocenjevanju in rangiranju genov uporabi informacijo o razredu.

Na grafu 1 so za primerjavo prikazane klasifikacijske točnosti metode SVM dobljene z desetkratnim prečnim preverjanjem, kjer je obogatenost skupin genov izračunana z metodo ASSESS. Ker metoda ASSESS omogoča izračun obogatenosti le za dvorazredne nabore podatkov (vsi nabori razen Armstrong in Bhattacharjee), so točnosti prikazane le za te nabore. Graf 2 prikazuje primerjavo med metodami na podlagi mere AUC.

Transformacija spremenljivk v skupine genov z našo metodo omogoča dodatni vpogled v preiskovane podatke na nivoju skupin. Na sliki 1 je primer projekcije radviz, dobljene z metodo VizRank[19,20] na transformiranem naboru podatkov Goluba in sod.[21] o dveh vrstah levkemije. Projekcija prikazuje en korak (od desetih) prečnega preverjanja, kjer so učni primeri prikazani s praznimi znaki, testni pa s polnimi.



Slika 1 Projekcija Radviz enega koraka prečnega preverjanja s transformiranimi podatki Goluba in sod. o dveh vrstah levkemije. Učni podatki so prikazani s praznimi, testni pa s polnimi znaki.

Opazimo lahko, da so skupine genov katerih proteinski produkti so povezani s hematopoetsko celično linijo, z metabolizmom glicerolipidov in z estrogenim signaliziranjem pri raku dojke bolj obogatene pri primerih akutne limfocitne levkemije (ALL, krogci). Primeri iz razreda akutne mieloidne levkemije (AML, kvadratki) pa imajo večjo izraženost skupin genov vključenih v TALL1 signalno pot, v pot razgradnje actetaminofena in v regulatorno pot encima, ki pretvarja angiotenzin (angl. *angiotensin-coverting encime 2*, ACE2).

Diskusija

Uporaba skupin genov pri analizi podatkov o izraženosti genov omogoča identifikacijo bioloških procesov, ki so povezani s preiskovanim bolezenskim stanjem (npr., vrsta raka). Takšna analiza lahko odkrije lastnosti, ki pri analizi na nivoju posameznih genov ostanejo skrite, in vodi do razjasnitev sprememb v izraženosti genov pri rakastih sprembah iz drugega zornega kota. Glavne prednosti metod za analizo skupin genov, kot je npr., GSEA, pred metodami analize posameznih genov so:

- računanje obogatenosti skupin genov brez vnaprej določenega praga,
- možnost odkrivanja zmernih, a skladnih, sprememb v skupinah genov in bioloških poteh, ki jih metode analize posameznih genov spregledajo,
- večje ujemanje med nabori podatkov in med podatki pridobljenimi na različnih platformah o istih bioloških vprašanjih.[4,5]

V članku smo predstavili novo metodo, ki na enostaven način izračuna obogatenost skupin genov pri posameznem primeru in tako omogoča klinično prognostiko ali diagnostiko na nivoju skupin genov. Metodo smo primerjali s sorodno metodo ASSESS,[9] napovedno točnost obeh pa primerjali z napovednimi točnostmi, ki jih dobimo z razvrščanjem vzorcev na podlagi informacije o izraženosti genov, torej brez uporabe genskih

skupin (graf 1 in graf 2). Zaključimo lahko, da sta napovedna točnost in mera AUC naše metode morda malo slabši od metode ASSESS, vendar so razlike majhne. Prednost naše metode je njena preprostost, predvsem pa možnost uporabe na naborih podatkov, ki vsebujejo več kot dva razreda.

Napovedne točnosti modelov zgrajenih z izraženostjo posameznih genov so na večini preiskovanih dvorazrednih naborov podatkov primerljive s točnostmi modelov zgrajenih z izraženostjo skupin genov. Pri obeh naborih podatkov, ki vsebujeta več kot dva diagnostična razreda, pa so točnosti modelov zgrajenih z izraženostjo genov nekaj boljše.

V literaturi smo zasledili še dve metodi za izračun obogatenosti skupin pri posameznih vzorcih.[14,22] Obe uporabita metodo glavnih komponent na vnaprej določenih skupinah genov, ocenita korelacijo med razredom in prvo glavno komponento in s permutacijskimi testi določita skupine genov, ki so povezane z izidom. Metoda Chen in sod.[22] nadgradi metodo Tomfohr in sod.[14] z uporabo nadzorovane metode glavnih komponent, ki iz skupine genov izbere najpomembnejše gene z uporabo informacije o razredu in le na podlagi teh genov izračuna glavne komponente.

Glavna prednost napovednih modelov zgrajenih z izraženostjo skupin genov je večja informativnost in lažja biološka razložljivost dobljenih modelov. Kot primer si oglejmo vizualizacijo na sliki 1, ki prikazuje najboljšo projekcijo podatkov z dvodimenzionalno metodo radviz dobljeno z algoritmom VizRank[20] v enem izmed korakov desetkratnega prečnega preverjanja s transformiranimi podatki Goluba in sod.[21] VizRank oceni kvaliteto projekcije na podlagi ločenosti vzorcev iz različnih napovednih razredov. Originalni nabor podatkov vsebuje podatke o izraženosti 7074 genov pri 72 bolnikih z akutno limfocitno (ALL) ali mieloidno levkemijo (AML).

Pri tvorjenju krvnih celic izraz mieloiden opisuje bele krvne celice (levkocite), ki niso limfociti in nastajajo iz mieloidnih matičnih celic, ki so prisotne v kostnem mozgu. Medtem ko sklop uničujočih genetskih sprememb v limfoctih pripelje do nastanka ALL, AML vznikne iz ostalih belih krvniček (npr. monocitov ali granulocitov), ki so bile podvržene rakastim genetskim spremembam. Skupine genov Hemapoetic cell lineage, BBcellpathway in TALL1pathway na sliki 1 imajo vse pomembno vlogo pri tvorjenju in diferenciaciji krvnih celic in lahko vplivajo na nastanek levkemije. Poglejmo si primer motenega uravnavanja TALL1 signalne poti.

TALL1 signalna pot združuje genske produkte, ki sodelujejo pri prenosu signala preko BCMA (angl. *B-cell maturation factor*) in TACI receptorjev za tumorske nekrotske faktorje. Preko te signalne poti se uravnava izražanje genov, ki vplivajo na diferenciacijo limfocitov ter na vnetni in stresni odgovor.[23] Spremenjena aktivnost TALL1 signalne poti je dokazana pri bolnikih z avtoimunskimi boleznimi,[24] prav tako pa tudi pri limfocitnih rakastih obolenjih.[25] Opazimo lahko, da imajo na sliki 1 primeri iz razreda ALL (krogci) manjšo izraženost skupine genov vključenih v TALL1 signalno pot v primerjavi s primeri AML (korgci so bolj oddaljeni od sidrišča za TALL1 skupino genov kot kvadratki).

Sklep

Analiza podatkov o izraženosti genov se je do nedavnega osredotočala na opazovanje izraženosti posameznih genov. V zadnjem času pa se je pokazalo, da je predvsem pri raziskavah raka izrednega pomena vključevanje dodatnega znanja v analizo. To omogočajo metode uporabe skupin genov, ki pri analizi upoštevanjo znanje o biološki ali molekularni funkciji genov. Pri taki analizi imajo prednost metode tipa GSEA, saj ne uporabljajo vnaprej določenega praga za ločevanje bolj in manj izraženih genov in uporabljajo metriko, ki dobro oceni skupine genov, ki imajo lahko tudi zmerne, vendar skladne spremembe v izraženosti.

V članku smo preučili, kakšna je napovedna točnost metod, ki uporabljajo znanje o genskih skupinah in metod, ki tega znanja ne uporabljajo in napovedi tvorijo neposredno iz podatkov z genskimi izrazi. Ugotovili smo, da je napovedna točnost obeh pristopov primerljiva. Velika prednost metod, ki uporabljajo genske skupine, pa je zmanjšanje razsežnosti podatkov in gradnja napovednih modelov, ki nudijo dodaten, biološko lažje razložljiv vpogled v preiskovane podatke.

Na majhnem številu naborov podatkov z več diagnostičnimi razredi so modeli, zgrajeni na originalnih podatkih, dosegli boljše napovedne točnosti. V nadaljnjem raziskovalnem delu se bomo usmerili v izboljšanje predlagane metode, da bo uspešna tudi na podatkih z več razredi.

Literatura

1. Jones S, Zhang X, Parsons DW, et al.: Core Signaling Pathways in Human Pancreatic Cancers Revealed by Global Genomic Analyses. *Science* 2008; 321(5897): 1801-1806.
2. Parsons DW, Jones S, Zhang X, et al.: An Integrated Genomic Analysis of Human Glioblastoma Multiforme. *Science* 2008; 321(5897): 1807-1812.
3. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 2008; 455(7216): 1061-1068.
4. Manoli T, Gretz N, Grone HJ, et al.: Group testing for pathway analysis improves comparability of different microarray datasets. *Bioinformatics* 2006; 22(20): 2500-2506.
5. Nam D, Kim SY: Gene-set approach for expression pattern analysis. *Brief Bioinform* 2008; 9(3): 189-197.
6. Khatri P, Draghici S: Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics* 2005; 21(18): 3587-3595.
7. Subramanian A, Tamayo P, Mootha VK, et al.: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005; 102(43): 15545-15550.
8. Mootha VK, Lindgren CM, Eriksson KF, et al.: PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* 2003; 34(3): 267-273.
9. Edelman E, Porrello A, Guinney J, et al.: Analysis of sample set enrichment scores: assaying the enrichment of sets of genes for individual samples in genome-wide expression profiles. *Bioinformatics* 2006; 22(14): e108-116.
10. Goeman JJ, Buhlmann P: Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics* 2007; 23(8): 980-987.
11. Efron B, Tibshirani R: On testing the significance of sets of genes. *Ann Appl Stat* 2007; 1(1): 107-129.
12. Kim SY, Volsky DJ: PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics* 2005; 6: 144.
13. Lee HK, Braynen W, Keshav K, et al.: ErmineJ: tool for functional analysis of gene expression data sets. *BMC Bioinformatics* 2005; 6: 269.
14. Tomfohr J, Lu J, Kepler TB: Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics* 2005; 6: 225.
15. Dinu I, Potter JD, Mueller T, et al.: Improving gene set analysis of microarray data by SAM-GS. *BMC Bioinformatics* 2007; 8: 242.
16. Piwowar HA, Day RS, Fridsma DB: Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS ONE* 2007; 2(3): e308.
17. Yon Rhee S, Wood V, Dolinski K, et al.: Use and misuse of the gene ontology annotations. *Nat Rev Genet* 2008; 9(7): 509-515.
18. Statnikov A, Aliferis CF, Tsamardinos I, et al.: A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics* 2005; 21(5): 631-643.
19. Leban G, Zupan B, Vidmar G, et al.: VizRank: Data Visualization Guided by Machine Learning. *Data Mining and Knowledge Discovery* 2006; 13(2): 119-136.
20. 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.
21. 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.
22. Chen X, Wang L, Smith JD, et al.: Supervised principal component analysis for gene set enrichment of microarray data with continuous or

survival outcomes. *Bioinformatics* 2008; 24(21): 2474-2481.
23. Shu HB, Johnson H: B cell maturation protein is a receptor for the tumor necrosis factor family member TALL-1. *Proc Natl Acad Sci U S A* 2000; 97(16): 9156-9161.
24. Gross JA, Johnston J, Mudri S, et al.: TACI and BCMA are receptors for a TNF homologue implicated in B-cell autoimmune disease. *Nature* 2000; 404(6781): 995-999.
25. Laabi Y, Gras MP, Carbonnel F, et al.: A new gene, BCM, on chromosome 16 is fused to the interleukin 2 gene by a t(4;16)(q26;p13) translocation in a malignant T cell lymphoma. *EMBO J* 1992; 11(11): 3897-3904.