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.

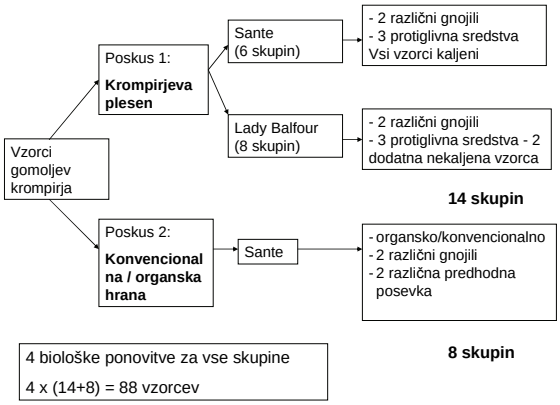
Prvi cilj analize podatkov je izločiti mikromreže, pri katerih hibridizacija ni bila uspešna, nato pa iz analize izločiti tudi posamezne točke mikromrež nezadostne kakovosti. Po izločanju nekakovostnih podatkov moramo podatke transformirati ter normalizirati, da lahko primerjamo podatke iz različnih 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.

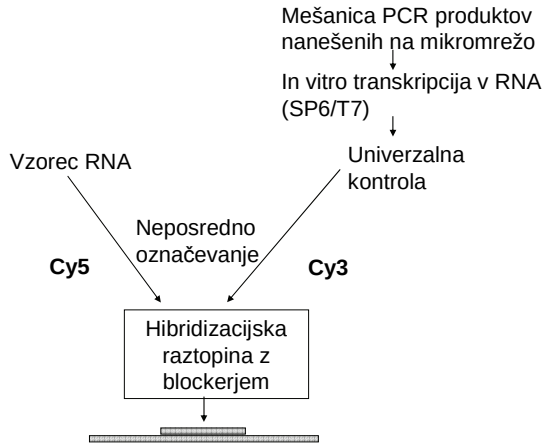


Slika 1 Shema poskusa.

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.

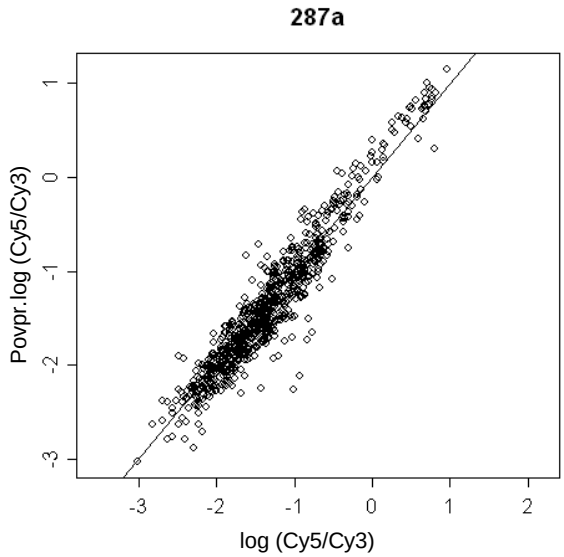
Analiza rezultatov

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).



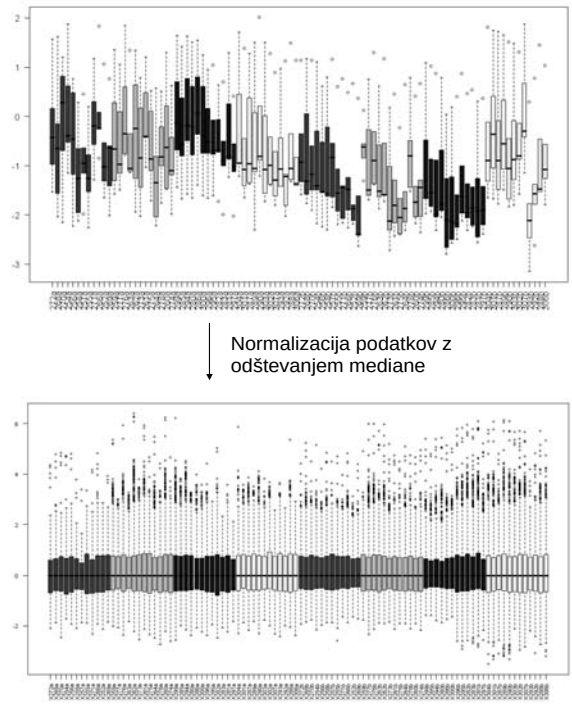
Slika 3 Razsevni grafikon za eno od analiziranih mikromrež.

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.

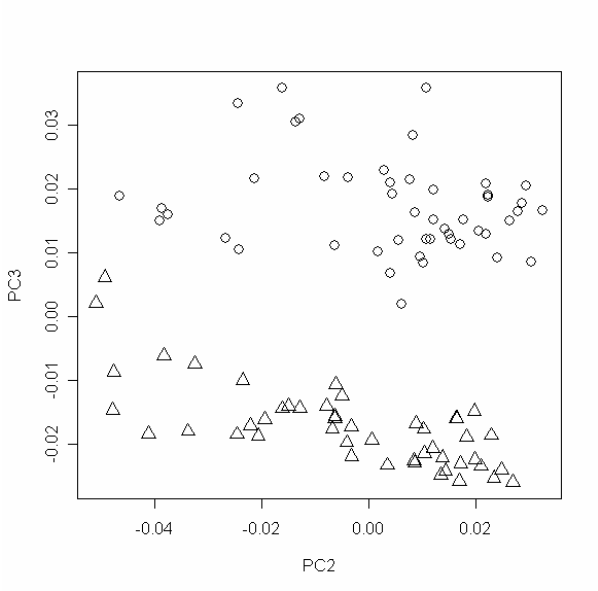


Slika 4 Normalizacija podatkov z odštevanjem mediane. Vsak izmed zabojev z ročaji predstavlja eno izmed mikromrež poskusa "krompirjeva plesen". Od vseh podatkov na tej mikromreži smo odšteli mediansko vrednost te mikromreže. S tem omogočimo primerljivost med podatki za različne vzorce.

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.