Podiely génových expresií v analýze microarray dát

Peter Bubelíny

MFF UK, KPMS, Sokolovská 83, CZ-18675 Praha 8
bubeliny@karlin.mff.cuni.cz

Abstrakt

Jeden z cieľov microarray experimentov je nájsť odlišne expresované gény získané z dvoch alebo viacerých štádií chorôb. Problém je, že génové expresie sú vysoko korelované. V tejto práci sú uvažované podiely génových expresií vytvorených z neusporiadaných alebo z usporiadaných párov génov. Pre HYPERDIP a TEL dáta (odlišné štádiá detskej leukémie) sa zdá, že tieto podiely pre rôzne páry sú približne nezávislé. Pre každú situáciu sú odhadnuté p-hodnoty pre testovanie, ktoré gény alebo ich podiely sú odlišne expresované. Tieto odhadnuté p-hodnoty sú zobrazené v histogramoch a tvary týchto histogramov sú porovnané. Porovnávacia štúdia ukázala, že tvary histogramov pre p-hodnoty spočítané z génových expresií a z ich podielov sa výrazne líšia.

Úvod

Microarray experiment, ktorý produkuje génové expresie na m rôznych génoch, môže byť reprezentovaný náhodným vektorom $\mathbf{X} = (x_1, ..., x_m)^T$ so vzájomne závislými zložkami. Uvažujme, že máme n výberov (slajdov) z $\mathbf{X}$. Potom môžeme microarray dáta pre m génov (niekoľko tisíc) z n slajdov (mnohonásobne menej - niekoľko desiatok) reprezentovať mxn dimenzionálnou maticou $X = \{\mathbf{X_1}|...|\mathbf{X_n}\} = \{x_{i,j}\}_{i,j=1}^{m,n}$, kde $x_{i,j}$ je génová expresia pre i-ty gén z j-tého slajdu. Uvažujme, že máme n_1 slajdov od ľudí trpiacich chorobou číslo jedna a n_2 slajdov od ľudí trpiacich druhou chorobou. Potom môžeme tieto dáta génových expresií reprezentovať dvomi maticami X a Y. Pre každé $i = 1, ..., n$ označme G_i^1 (pre prvú chorobu) a G_i^2 (pre druhú chorobu) rozdelenie génových expresií pre i-ty gén. Zaujímame sa o testovanie, ktoré gény sú odlišne expresované. To znamená, že chceme testovať hypotézy $H_i^0 : G_i^1 = G_i^2$ oproti alternatívam $A_i^0 : G_i^1 \neq G_i^2$ pre každé $i = 1, ..., n$ súčasne.
Génové expresie sú vysoko korelované medzi génmi. *Klebanov and Yakovlev* (2007) študovali rôzne microarray dáta a objavili, že priemer korelačných koeficientov medzi génmi bol v rozpätí od 0.84 do 0.97. Ďalej usporiadali gény podľa veľkosti rozptylu ich expresií (v rastúcom poradí) a definovali δ-postupnosť ako $\delta_{i,j} = X_{(2i),j} - X_{(2i-1),j}$, kde $X_{(l),j}$ je j-ta génová log-expresia l-tého génu vo vyššie zadefinovanom poradí. Oni objavili, že túto δ-postupnosť tvoria skoro nezávislé náhodné veličiny.
Pretože by sme chceli pracovať s expresiami (nie z log-expresiami), budeme uvažovať podiely génových expresií namiesto rozdielov génových log-expresií.

Dáta

V tejto práci budeme používať HYPERDIP a TEL dáta (dve štádiá detskej leukémie). Tieto dáta sa skladajú z $\log_2$-expresií pre 7084 génov a pre každý gén máme 88 (pre HYPERDIP) a 79 (pre TEL) slajdov. Pretože chceme pracovať s expresiami, transformujeme $\log_2$-expresie na expresie ako: $expresia = 2^{log_2 - expresia}$. Tieto génové expresie môžeme reprezentovať dvomi maticami $X = \{x_{i,j}\}$ a $Y = \{y_{k,l}\}$ s dimenziami 7084x88 and 7084x79. Preto pre každý gén $i = 1, ..., 7084$ máme génové expresie reprezentované náhodnými výbermi $\mathbf{x_i} = x_{i,1}, ..., x_{i,88}$ pre HYPERDIP (rozdelenie $x_{i,j}$ označme G_i^H) a $\mathbf{y_i} = y_{i,1}, ..., y_{i,79}$ pre TEL dáta (rozdelenie $y_{i,j}$ označme G_i^T).

Cieľ

Cieľom tejto práce je odhadnúť p-hodnoty pre testovanie, ktoré gény alebo podiely génov sú odlišne expresované s využitím HYPERDIP a TEL dát. Tieto odhadnuté p-hodnoty zobrazíme v histogramoch a tvary týchto histogramov porovnáme.

Test

Nech μ a ν sú dve pravdepodobnostné miery definované na Euklidovskom priestore R^d. Pre testovanie hypotézy $H : \mu = \nu$ použijeme permutačný test uvažovaný v *Szabo et al.* (2002). Označme $L(x, y)$ striktne negatívne definitnú funkciu, tj. $\sum_{i,j=1}^{s} L(x_i, x_j)h_i h_j \leq 0$ pre každé $x_1, ..., x_s$, $h_1, ..., h_s$ a $\sum_{i=1}^{s} h_i = 0$ s rovnosťou práve keď, všetky h_i sú nulové. Definujme
$N(\mu, \nu) = 2\int_{R^d}\int_{R^d} L(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) - \int_{R^d}\int_{R^d} L(x, y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) - \int_{R^d}\int_{R^d} L(x, y)\mathrm{d}\nu(x)\mathrm{d}\nu(y)$. Potom $\sqrt{N(\mu, \nu)}$ je metrika na priestore všetkých pravdepodobnostných mier na R^d.
Predpokladajme, že $\mathbf{x} = (x_1, ..., x_{n_1})'$ a $\mathbf{y} = (y_1, ..., y_{n_2})'$ sú dva nezávislé výbery skladajúce sa z n_1 a n_2 pozorovaní z μ and ν. Potom empirické vyjadrenie $N(\mu, \nu)$ je
$\hat{N}(\mathbf{x}, \mathbf{y}) = \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} L(x_i, y_j) - \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} L(x_i, x_j) - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} L(y_i, y_j)$. V tejto práci budeme používať Euklidovskú vzdialenosť ako striktne negatívne definitnú funkciu L. To znamená, že $L(x, y) = \sqrt{\sum_{i=1}^{d}(x_i - y_i)^2}$.
Nech $\mathbf{z} = (\mathbf{z_1}, \mathbf{z_2})' = (z_1, ..., z_{n_1+n_2})' = (x_1, ..., x_{n_1}, y_1, ..., y_{n_2})' = (\mathbf{x}, \mathbf{y})'$ je $n_1 + n_2$ génových expresií (podielov) pre $d \geq 1$ génov z dvoch výberov. Najprv spočítame $\hat{N}(\mathbf{x}, \mathbf{y})$, potom 10000-krát spermutujeme zložky $\mathbf{z}$ a spočítame $\hat{N}(\mathbf{z_1^{(i)}}, \mathbf{z_2^{(i)}})$, kde $\mathbf{z}^{(i)} = (\mathbf{z_1^{(i)}}, \mathbf{z_2^{(i)}})'$ je i-ta permutácia zložiek $\mathbf{z}$. Potom odhadneme p-hodnotu hypotézy H : *skupina génov (podielov) nie je odlišne expresovaná* ako

$$\hat{p} = \frac{1}{10000}\sum_{i=1}^{10000} I_{[\hat{N}(\mathbf{x},\mathbf{y}) \leq \hat{N}(\mathbf{z_1^{(i)}}, \mathbf{z_2^{(i)}})]}, \qquad (1)$$

kde $I_{[.]}$ je indikátorová funkcia . Odhadnuté p-hodnoty budeme označovať ako $\hat{p}$-hodnoty.

Usporiadanie génov

V tejto práci budeme uvažovať podiely génových expresií. Preto potrebujeme z génov vytvoriť dvojice, z ktorých budeme tieto podiely počítať. Ale ako máme vytvoriť tieto dvojice? Prvá možnosť je, že tieto dvojice vytvoríme bez usporiadania dát a dvojicu vytvoríme z $(2i-1)$-tého a $2i$-tého génu z usporiadania génov v databáze. Druhou možnosťou je usporiadať gény podľa odhadu veľkosti rozptylu ich génových expresií (v rastúcom poradí) a potom spojiť $(2i-1)$-ty a $2i$-ty gén. Tretia možnosť je inšpirovaná *Klebanov et al.* (2006). V tomto článku je objavená A-závislosť, ktorá sa objavuje v microarray dátach. Nechajme x a y byť génové expresie pre gény g_x a g_y. Povieme, že dvojica (g_x, g_y) je A-závislá, ak x a y splňujú podmienku $y = xz$, kde z je kladná náhodná veličina nezávislá na x. Použitím $\log_2$ transformácie na A-závislé náhodné veličiny dostávame, že Y=X+Z, kde $Y = \log_2 y$, $X = \log_2 x$ a $Z = \log_2 z$. S využitím nezávislosti x a z (teda aj X a Z) dostávame, že $\mathrm{Var}\, Y > \mathrm{Var}\, X$. Tento druh závislosti nie je symetrický. Génové páry A-závislých génov vytvárajú dlhé reťazce obsahujúce dokonca tisíce génov. Preto vzhľadom k A-závislosti môžeme usporiadať gény podľa veľkosti odhadnutých rozptylov ich $\log_2$-expresií.
Druhý problém je, ktoré dáta máme použiť pri usporiadaní (poradie pre HYPERDIP a TEL je iné)? Takže použijeme samostatne HYPERDIP dáta, TEL dáta a tiež ich spojíme (ako keby boli z rovnakého rozdelenia) a odhadneme spoločný rozptyl expresií ($\log_2$-expresií).

Expresie

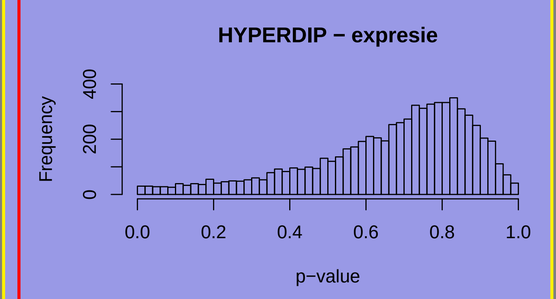
V tomto stĺpci budeme uvažovať génové expresie. Takže naše dáta môžeme reprezentovať dvomi maticami $X = \{x_{i,j}\}_{i,j=1}^{7084,88}$ a $Y = \{y_{i,j}\}_{i,j=1}^{7084,79}$, kde $x_{i,j}$ je i-ta génová expresia z j-tého slajdu HYPERDIP dát a $y_{i,j}$ je i-ta génová expresia z j-tého slajdu TEL dát.

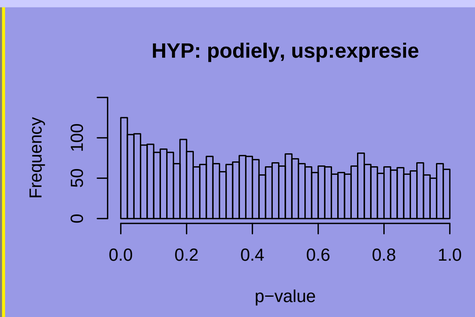
Podiely

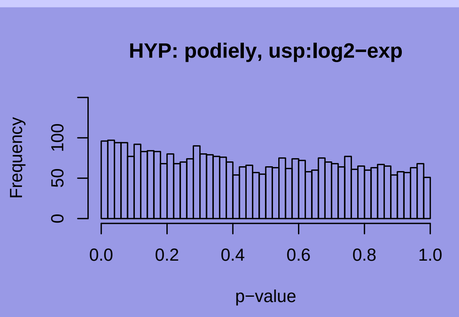
V tomto stĺpci budeme uvažovať podiely génových expresií. Takže tieto dáta môžeme reprezentovať dvomi maticami $X^* = \{x_{i,j}^*\}_{i,j=1}^{7084,88}$ a $Y^* = \{y_{i,j}^*\}_{i,j=1}^{7084,79}$, kde $x*_{i,j} = \frac{x_{(2i),j}}{x_{(2i-1),j}}$ a $x_{(i),j}$ je i-ta génová expresia z j-tého slajdu (ne)usporiadaných HYPERDIP dát a $y*_{i,j} = \frac{y_{(2i),j}}{y_{(2i-1),j}}$ a $y_{(i),j}$ je i-ta génová expresia z j-tého slajdu (ne)usporiadaných TEL dát.

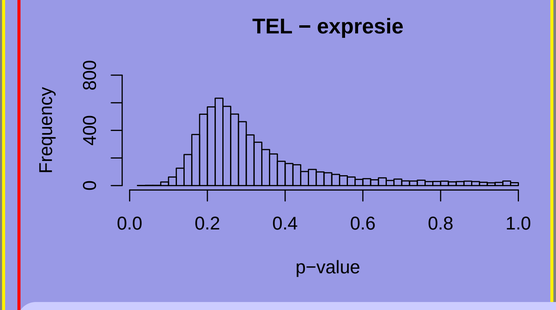
Všetky hypotézy platia

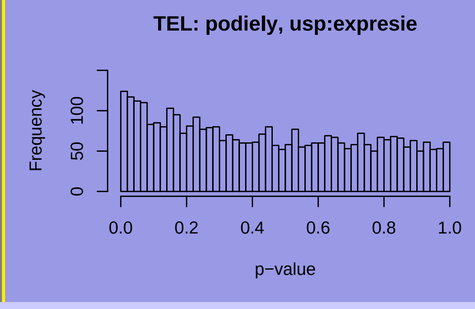
Najprv uvažujme HYPERDIP a TEL dáta samostatne. HYPERDIP dáta rozdelíme na dve polovice a tak vytvoríme dva výbery pre každý gén (podiel) skladajúce sa zo 44 slajdov (prvých 44 slajdov použijeme pre prvý výber a zvyšných 44 slajdov pre druhý výber). Tieto výbery sú z toho istého rozdelenia G_i^H, $i = 1, ..., 7084$ (respektíve G_i^{*H}, $i = 1, ..., 3542$ pre podiely génových expresií). Takže sa zaujímame o testovanie platných hypotéz $H_i^H : G_i^H = G_i^H$ ($H_i^{*H} : G_i^{*H} = G_i^{*H}$) súčasne pre všetky i. To isté urobíme aj pre TEL dáta. Pre každý gén (podiel) vytvoríme dva výbery skladajúce sa z 39 a 40 slajdov (prvých 39 slajdov použije pre prvý výber a zvyšných 40 slajdov pre druhý výber) a súčasne testujeme, ktorý takto vytvorený výber je odlišne expresovaný. Histogramy $\hat{p}$-hodnôt pre tieto situácie (histogramy pre neusporiadané podiely sú podobné histogramom pre usporiadané podiely - preto sú vynechané) sú na obrázkoch:

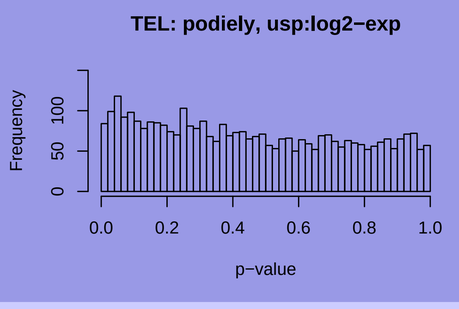








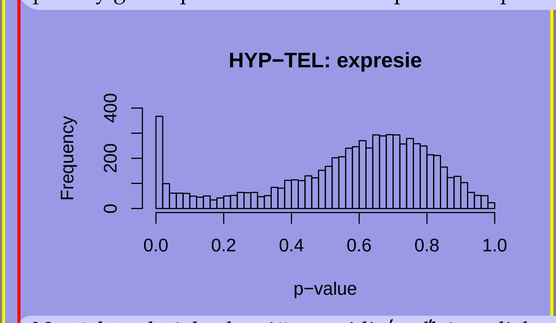


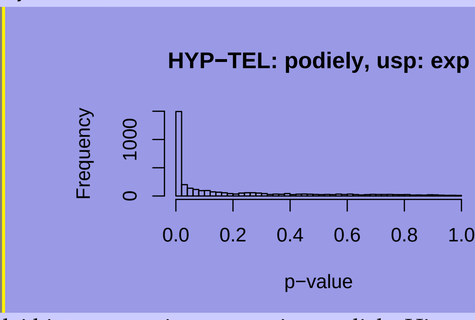


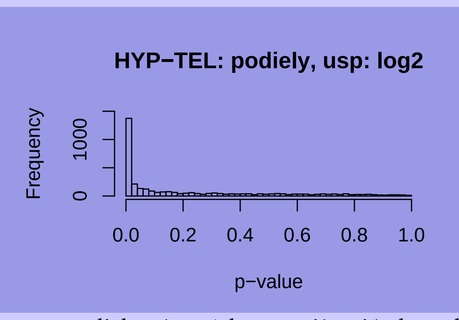
Keby génové expresie (podiely) boli nezávislé, p-hodnoty pre testovanie platných hypotéz by mali rovnomerné rozdelenie. Ale z obrázkov môžme vidieť, že histogramy pre génové expresie nevypadajú ako histogramy pre rovnomerne rozdelené náhodné veličiny. Tvar týchto histogramov môže byť pripísaný práve silnej závislosti medzi génmi. Naopak o histogramoch pre podiely génových expresií nemôžme povedať, že sa nepodobajú histogramom pre rovnomerne rozdelené náhodné veličiny. To potvrdzuje nízku závislosť medzi podielmi génových expresií.

Niektoré hypotézy neplatia

Doposiaľ sme uvažovali, že všetky hypotézy boli pravdivé. Teraz sa budeme zaujímať o to, ako sa situácia zmení, keď nie všetky hypotézy budú platné. Už nebudeme uvažovať HYPERDIP a TEL dáta samostatne ale spolu. Takže sa budeme zaujímať o testovanie, ktoré gény (podiely) sú odlišne expresované medzi HYPERDIP a TEL dátami. To znamená, že chceme testovať hypotézy $H_i : G_i^H = G_i^T$ pre každé $i = 1, ..., 7084$ ($H_i^* : G_{(i)}^{*H} = G_{(i)}^{*T}$, $i = 1, ..., 3542$) súčasne.
Histogramy $\hat{p}$-hodnôt pre génové expresie a pre usporiadané podiely vzhľadom k združenému rozptylu expresií a $\log_2$-expresií (tvary histogramov pre podiely génov pre ďalšie situácie sú podobné - preto sú vynechané) sú na obrázkoch:







Na týchto obrázkoch môžeme vidieť veľké rozdiely medzi histogramami pre expresie a podiely. Histogramy pre podiely génových expresií majú obrovský prvý stĺpec, histogramy pre génové expresie majú zase výrazný kopec okolo hodnoty 0.7. Nízke hodnoty $\hat{p}$-hodnôt pre podiely génových expresií naznačujú, že tam je veľa odlišne expresovaných podielov (oveľa viac ako odlišne expresovaných génov).

Páry

Niekto môže namietať, že je veľký rozdiel v tom, čo sme testovali. Pre génové expresie sme testovali 7084 hypotéz (jednu hypotézu pre jeden gén), ale pre podiely génových expresií sme mali iba 3542 hypotéz (jednu hypotézu pre dva gény). Pretože náš test je skonštruovaný aj pre testovanie náhodných vektorov, vytvoríme 3542 párov génov a budeme testovať, či združené rozdelenia pre tieto páry génov sú rovnaké pre HYPERDIP a TEL dáta. To znamená, že chceme testovať hypotézy $H_i^2 : (G_{(2i-1)}^H, G_{(2i)}^H) \stackrel{D}{=} (G_{(2i-1)}^T, G_{(2i)}^T)$ súčasne pre všetky $i = 1, ..., 3542$, kde $G_{(j)}^H$ je rozdelenie génovej expresie pre j-ty gén HYPERDIP dát a $G_{(j)}^T$ je rozdelenie génovej expresie pre j-ty gén TEL dát. Tak isto ako pre podiely génových expresií môžeme uvažovať neusporiadané, ale aj usporiadané páry. Tvary histogramov $\hat{p}$-hodnôt pre hypotézy H_i^2 pre páry génov sú podobné tvaru histogramu pre génové expresie (preto nie sú zobrazené).
V tabuľke sú zaznamenané počty hypotéz, ktoré by sme zamietli podľa Bonferroniho nerovnosti (kritická hodnota je $\frac{0.05}{3542}$) pre neusporiadané a pre usporiadané podiely a páry génových expresií.

	neuspor	HYP-exp	TEL-exp	združ-exp	HYP-$\log_2$	TEL-$\log_2$	združ-$\log_2$
podiely	618	659	750	671	610	651	601
páry	92	74	86	78	77	92	63

Z tejto tabuľky je vidieť, že pre podiely génových expresií zamietame približne 7-9 krát viac hypotéz ako pre páry.

Záver

V tejto práci sme študovali histogramy odhadnutých p-hodnôt pre testovanie, ktoré gény alebo podiely génov sú odlišne expresované pre HYPERDIP a TEL dáta. Uvažovali sme prípady, kde všetky hypotézy platili (použili sme HYPERDIP a TEL dáta samostatne), ale aj prípady, keď niektoré z hypotéz nemuseli platiť (použili sme tieto dáta súčasne). Ukázali sme, že tvary histogramov $\hat{p}$-hodnôt spočítaných pre génové expresie a pre podiely génových expresií sú značne odlišné. $\hat{p}$-hodnoty pre podiely génov mali tendenciu byť oveľa nižšie ako pre génové expresie. Preto testy využívajúce podiely génových expresií by mohli byť silnejšie ako testy s génovými expresiami. Naviac génové expresie sú vysoko závislé medzi génmi narozdiel od podielov. Preto by sme mali zamerať našu pozornosť na podiely génových expresií a preskúmať túto neznámu časť štatistiky.

Poďakovanie

Rád by som poďakoval Prof. Levovi Klebanovovi, DrSc. za cenné komentáre, poznámky a celkovú pomoc. Táto práca bola podporená grantom 201/05/H007.

Referencie

Szabo, A., Boucher, K., Caroll, W., Klebanov, L., Tsodikov, A. and Yakovlev, A. (2002). Variable Selection and pattern recognition with gene expression data generated by the microarray technology *Mathematical Biosciences*, 176, 77-98.

Klebanov, L. and Yakovlev, A. (2007). Diverse correlation structures in gene expression data and their utility in improving statistical inference, *The Annals of Applied Statistics*,Vol.1 No.2,538-559.

Klebanov, L., Jordan, C. and Yakovlev, A. (2006). A new type of stochastic dependence revealed in gene expression data, *Statistical Applications in Genetics and Molecular Biology*, Vol. 5 issue 1, Article 7.