Wybrane zagadnienia związane z testowaniem zbioru hipotez o hierarchicznej strukturze zależności.

Przemysław Biecek

IMPAN Oddział Wrocławski

Plan referatu

1. Zagadnienie testowania zbioru hipotez,
2. Zagadnienie testowania zbioru hipotez z zadaną relacją hierarchiczną,
3. Procedury testowania kontrolujące współczynniki FWER, PFER i FDR,
4. Przykład zastosowania uzyskanych wyników w zagadnieniu identyfikacji funkcji biologicznych,
5. Inne procedury testowania dla hierarchicznych struktur zależności.

Testowanie zbioru hipotez

Rozważmy zbiór eksperymentów $\mathcal{E} = \{\mathcal{E}^{(i)} : i \in I = \{1, ..., m\}\}$.
Każdemu eksperymentowi $\mathcal{E}^{(i)}$ odpowiada rozkład $P^{(i)}_{\theta_0^{(i)}}$ z rodziny rozkładów indeksowanej zbiorem $\Theta^{(i)}$.
Rozważmy zbiór hipotez $\mathcal{H} = \{H_0^{(i)} : i \in I\}$. Każda hipoteza zerowa jest związana z wyborem podzbioru $\Theta_0^{(i)} \subset \Theta^{(i)}$.

Hipoteza zerowa

Hipotezą zerową $H_0^{(i)}$ nazywamy przypuszczenie, że $\theta_0^{(i)} \in \Theta_0^{(i)}$.

Hipoteza alternatywna

Hipotezą alternatywną $H_A^{(i)}$ nazywamy przypuszczenie, że $\theta_0^{(i)} \notin \Theta_0^{(i)}$.

Testowanie zbioru hipotez

Rozważmy zbiór eksperymentów $\mathcal{E} = \{\mathcal{E}^{(i)} : i \in I = \{1, ..., m\}\}$.
Każdemu eksperymentowi $\mathcal{E}^{(i)}$ odpowiada rozkład $P^{(i)}_{\theta_0^{(i)}}$ z rodziny rozkładów indeksowanej zbiorem $\Theta^{(i)}$.
Rozważmy zbiór hipotez $\mathcal{H} = \{H_0^{(i)} : i \in I\}$. Każda hipoteza zerowa jest związana z wyborem podzbioru $\Theta_0^{(i)} \subset \Theta^{(i)}$.

Hipoteza zerowa

Hipotezą zerową $H_0^{(i)}$ nazywamy przypuszczenie, że $\theta_0^{(i)} \in \Theta_0^{(i)}$.

Hipoteza alternatywna

Hipotezą alternatywną $H_A^{(i)}$ nazywamy przypuszczenie, że $\theta_0^{(i)} \notin \Theta_0^{(i)}$.

Testowanie zbioru hipotez

	#przyjętych hipotez zerowych	#odrzuconych hipotez zerowych	
#prawdziwych hipotez zerowych	$U = \sum_i (1 - H_i)(1 - \psi_i)$	$V = \sum_i (1 - H_i)\psi_i$	m_0
#fałszywych hipotez zerowych	$T = \sum_i H_i(1 - \psi_i)$	$S = \sum_i H_i\psi_i$	m_1
suma	m - R	R	m

Symbolem $\psi_i \in \{0, 1\}$ oznaczamy wynik testowania dla hipotezy $H_0^{(i)}$. Symbolem $H_i \in \{0, 1\}$ oznaczamy stan hipotezy $H_0^{(i)}$.

Najpopularniejsze miary błędu

Per-family error rate (PFER)

$$PFER = E(V).$$

Family-wise error rate (FWER)

$$FWER = Pr(V > 0).$$

False discovery rate (FDR)

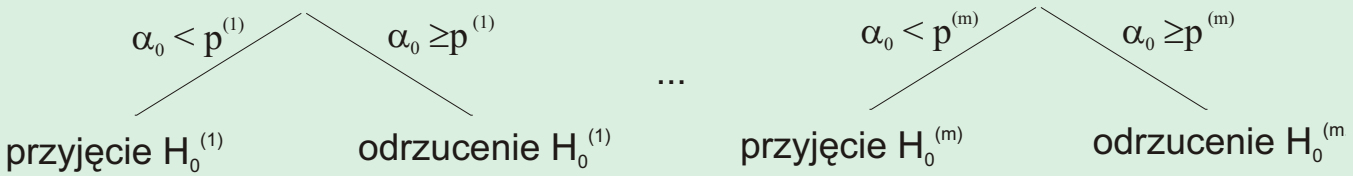
$$FDR = E(Q),$$

gdzie

$$Q = \begin{cases} V/R & R > 0, \\ 0 & R = 0. \end{cases}$$

Procedura jednokrokowa

Single step



Procedura Bonferroniego (1936)

Przyjmując dla każdego testu poziom istotności

$$\alpha_0 = \alpha/m, \tag{1}$$

kontrolujemy współczynnik PFER na poziomie α.

Korekta Sidaka

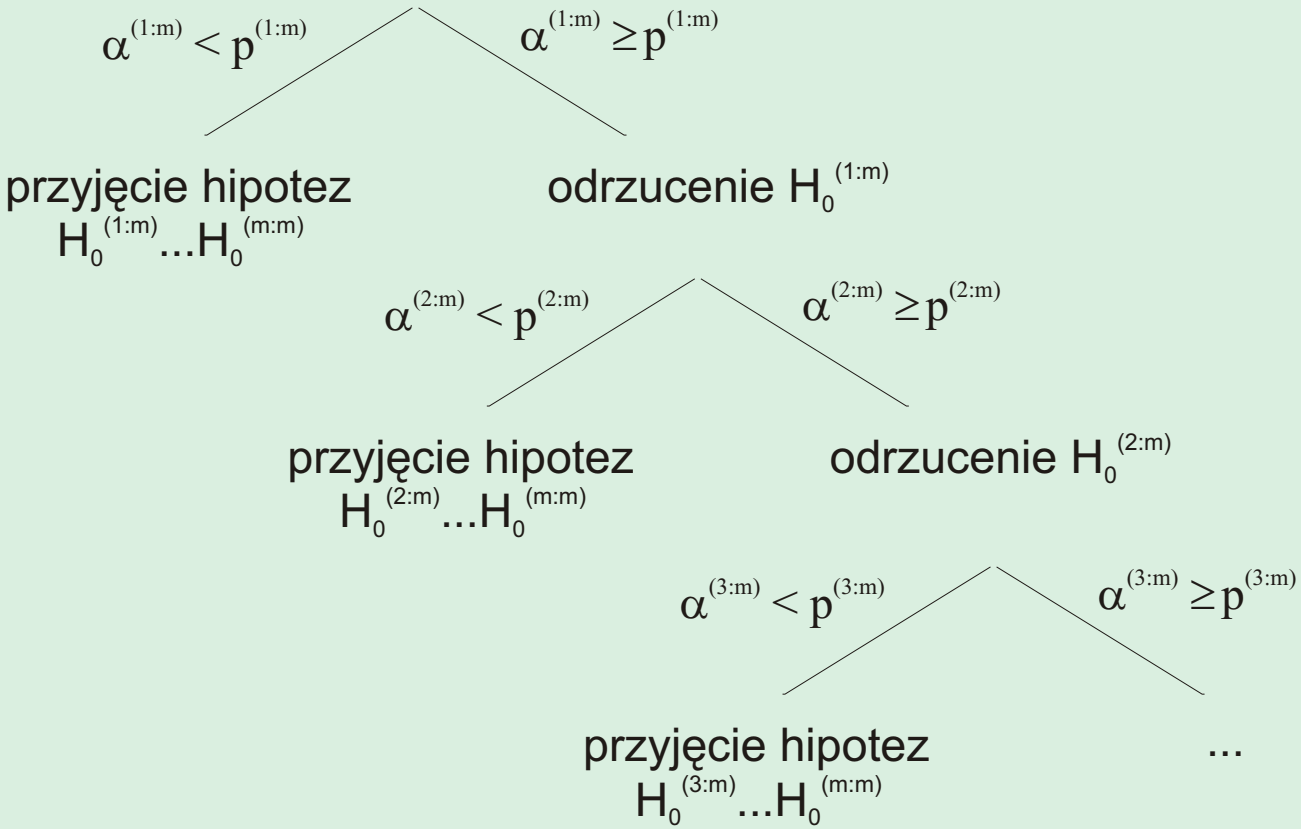
Jeżeli statystyki testowe są niezależne, to przyjmując poziomy istotności

$$\alpha_0 = 1 - (1-\alpha)^{1/m}, \tag{2}$$

kontrolujemy współczynnik błędu FWER na poziomie α.

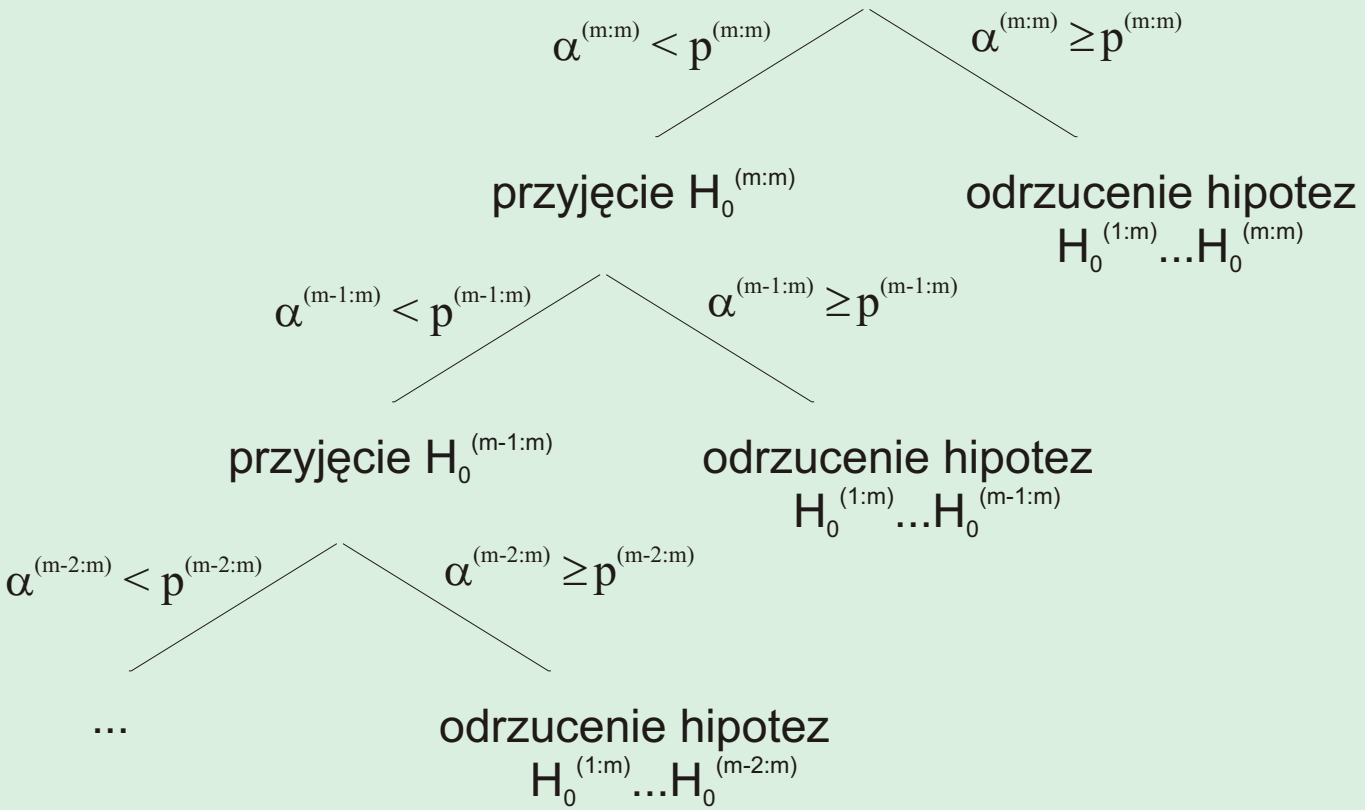
Procedura wielokrokowa

Step down



Procedura wielokrokowa

Step up



Procedura Holma (1979)

Przyjmując w procedurze step-down poziomy istotności

$$\alpha^{(i:m)} = \alpha/(m - i + 1), \tag{3}$$

kontrolujemy współczynnik FWER na poziomie α.

Procedura Hochberga (1988)

Przyjmując w procedurze step-up poziomy istotności

$$\alpha^{(i:m)} = \alpha/(m - i + 1), \tag{4}$$

kontrolujemy współczynnik FWER na poziomie α.

Procedura Benjaminiego Hochberga (1995)

Przyjmując w procedurze step-up poziomy istotności

$$\alpha^{(i:m)} = \frac{i}{m}\alpha, \tag{5}$$

kontrolujemy współczynnik FDR na poziomie α.

Literatura

- C.E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilitŕ. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze (1936) 8:3-62.
- S. Holm. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics (1979) 6: 65-70.
- Y. Hochberg. A sharper Bonferroni procedure for multiple tests of significance. Biometrika (1988), 75: 800-803.
- Y. Benjamini, Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Statist. Soc. B (1995) 57, No. 1, pp. 289-300.
- Y. Benjamini, D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. Annals of Statistics (2001) 29, 1165-1188.
- Y. Benjamini, A. M. Krieger, D. Yekutieli. Adaptive linear step-up procedures that control the false discovery rate (2005).

Literatura


S. Dudoit, Y.H. Yang, M.J. Callow, T.P. Speed. Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments. Statistica Sinica (2002), 12 111-139.


J. Storey. A direct approach to false discovery rates. J. R. Statist. Soc. B (2002) 64, Part 3, pp 479-498.


A. Owen. Variance of the number of false discoveries. J. R. Statist. Soc. B (2005) 67, 411-26.


P. Biecek. A modified Bonferroni correction for strongly correlated test statistics. Proceedings of the XI National Conference Application of Mathematics to Biology and Medicine (2005) ISBN: 83-903893-3-9.


J. Cheverud. A simple correction for multiple comparsions in interval mapping genome scans. Heredity (2001) 87 Issue 1 Page 52.

Testowanie zbioru hipotez z zadana relacją hierarchiczną $\mathcal{R}$

Zastosowanie do analiz z użyciem Gene Ontology

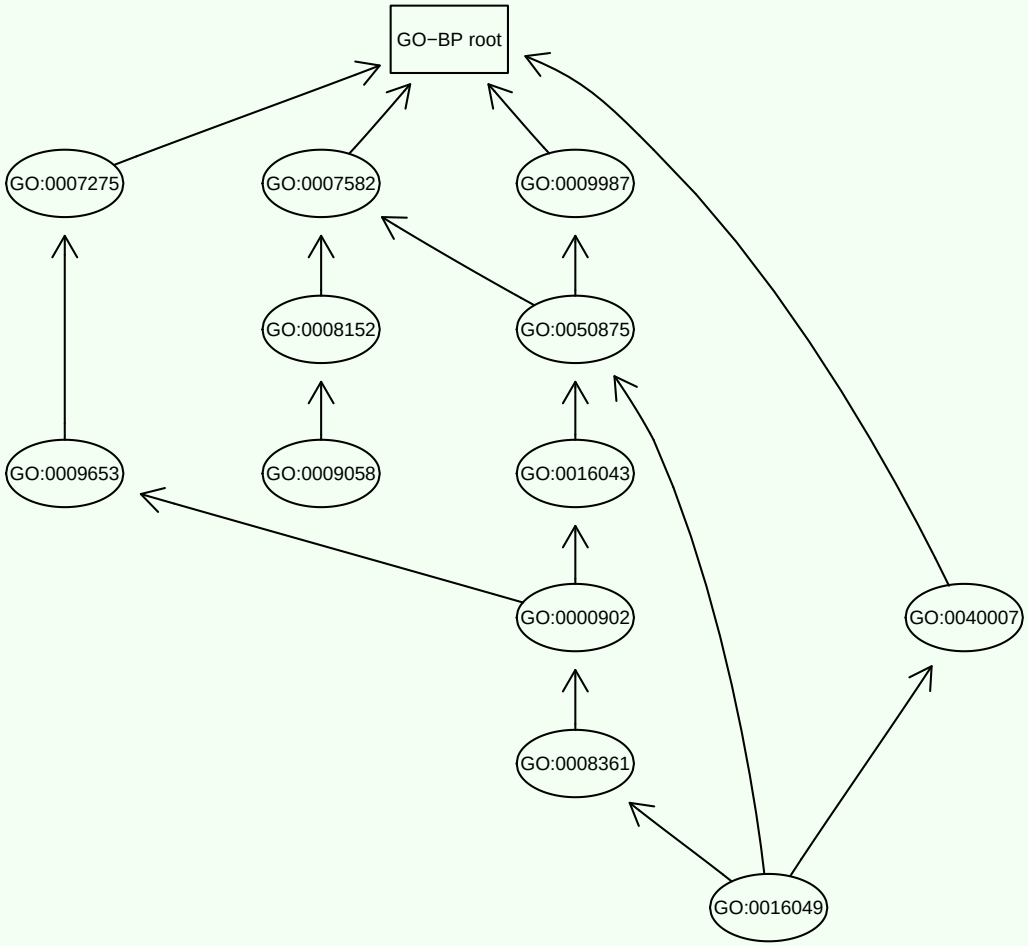
Rozważmy zbiór współdziałających genów i funkcję biologiczną f_i. Symbolem $\rho^{(i)}$ oznaczmy częstość występowania funkcji f_i w rozważanym zbiorze genów, symbolem $\rho_0^{(i)}$ w zbiorze wszystkich genów.

Hipotezy w rozważanym zagadnieniu biologicznym

$H_0^{(i)}$: $\rho^{(i)} = \rho_0^{(i)}$,

$H_A^{(i)}$: $\rho^{(i)} > \rho_0^{(i)}$.

Zastosowanie do analiz z użyciem Gene Ontology



Testowanie zbioru hipotez z relacją $\mathcal{R}$

Wiele osób pracowało nad kontrolą współczynnika FWER dla hipotez z zadaną relacją hierarchiczną (tz. *closure-testing* w badaniach *dose-response*). Pionierskie prace publikowali np. Gabriel (1969), R. Marcus i E. Peritz (1976), U. Naik (1977). Tematyka ta jest wciąż aktualna o czym może świadczyć praca H. Finnera (2002).

W wymienionych pracach określa się wspólną przestrzeń parametrów Θ dla wszystkich hipotez. Hipoteza $H_0^{(i)}$ odpowiada $\Theta_0^{(i)} \subset \Theta$ a relacja pomiędzy hipotezami wynika z postaci zbiorów $\Theta_0^{(i)}$.

Literatura

- R. Marcus, E. Peritz, K.R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. Biometrika (1976) 63(3):655-660.
- A.C. Tamhane, C. W. Dunnett, Y. Hochberg. Multiple test procedures for dose finding. Biometrics (1996), 52, 21-37.
- A.C. Tamhane, B.R. Logan. Multiple comparison procedures in dose response studies. Working Paper (2004) No. 04-017.
- H. Finner, K. Strassburger. The partitioning principle: A powerful tool in multiple decision theory. The Annals of Statistics (2002). Vol. 30, No. 4, 1194–1213.
- J. Goeman. Global testing as an alternative to single gene testing in genomic microarray data with some open multiple testing problems (2004). CIRM Seminar Luminy.
- P. Biecek. Multiple testing procedures for hierarchically related hypotheses. Zgłoszone do Biometrical Journal.

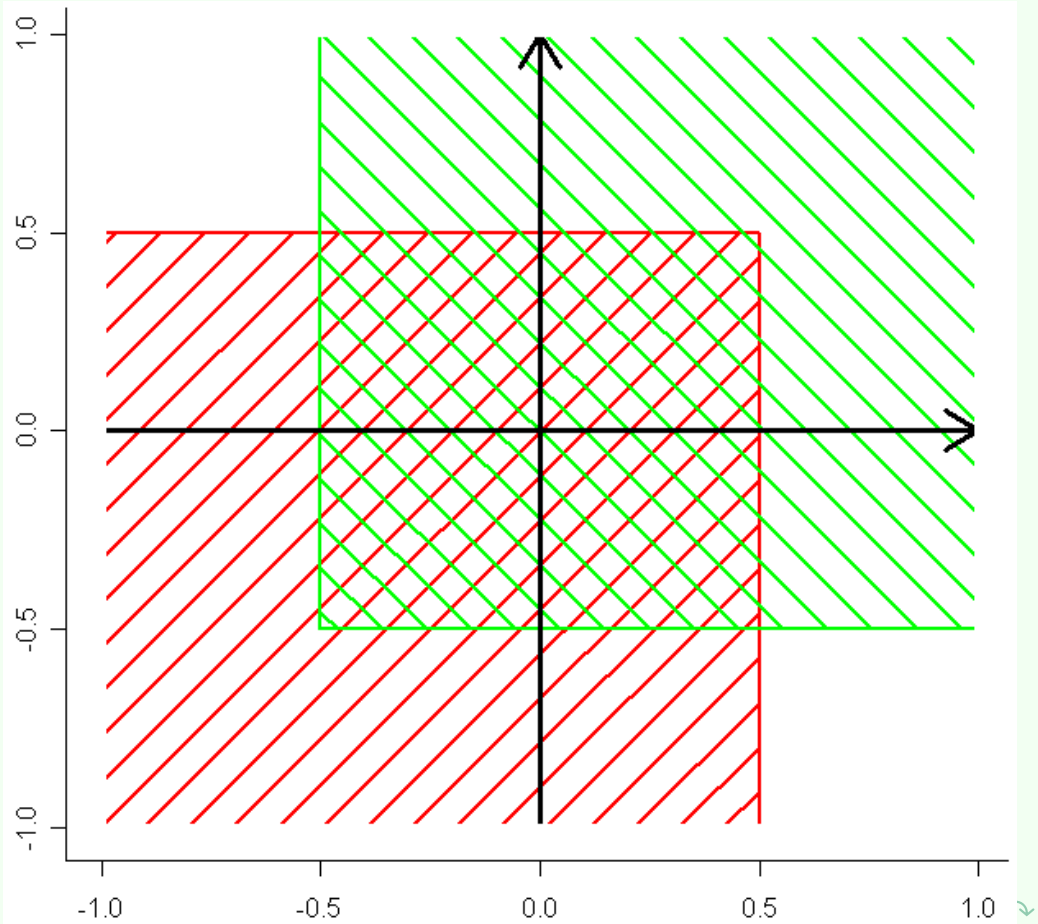
Testowanie zbioru hipotez z relacją $\mathcal{R}$

Obserwujemy $x_i \sim \mathcal{N}(\mu_i, 1)$, $i \in \{1, 2\}$, $(\mu_1, \mu_2) \in \Theta = R^2$.

Rozważmy hipotezy:

$$H_0^{(1)}: \quad -0,5 < \min(\mu_1; \mu_2);$$
$$H_0^{(2)}: \quad -0,5 < \mu_1; \mu_2 < 0,5;$$
$$H_0^{(3)}: \quad \max(\mu_1; \mu_2) < 0,5.$$

Tego podejścia nie można zastosować w analizach z wykorzystaniem Gene Ontology!!!



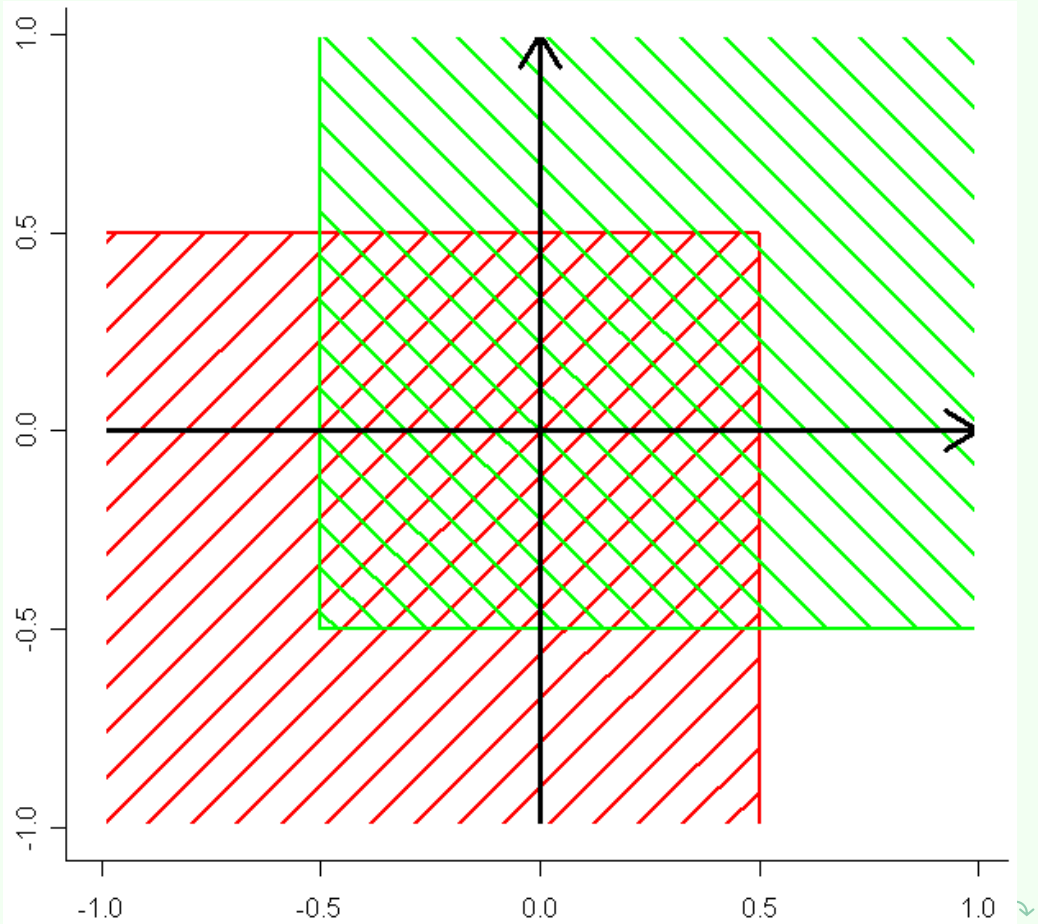
Testowanie zbioru hipotez z relacją $\mathcal{R}$

Obserwujemy $x_i \sim \mathcal{N}(\mu_i, 1)$, $i \in \{1, 2\}$, $(\mu_1, \mu_2) \in \Theta = R^2$.

Rozważmy hipotezy:

$$H_0^{(1)}: \quad -0,5 < \min(\mu_1; \mu_2);$$
$$H_0^{(2)}: \quad -0,5 < \mu_1; \mu_2 < 0,5;$$
$$H_0^{(3)}: \quad \max(\mu_1; \mu_2) < 0,5.$$

Tego podejścia nie można zastosować w analizach z wykorzystaniem Gene Ontology!!!



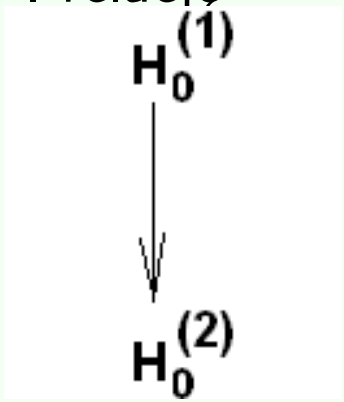
Testowanie zbioru hipotez z relacją $\mathcal{R}$

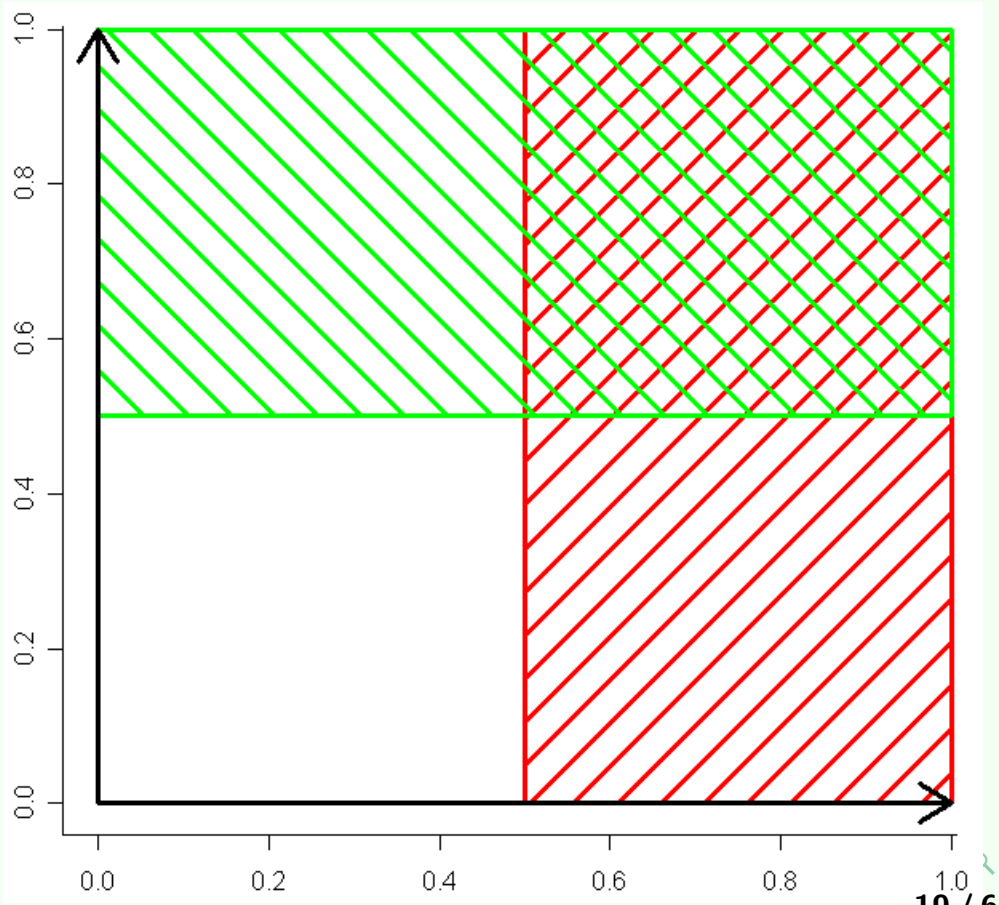
Obserwujemy $x_i \sim \mathcal{B}(\rho^{(i)})$, $i \in \{1, 2\}$, $(\rho^{(1)}, \rho^{(2)}) \in \Theta = [0, 1]^2$.

Rozważmy hipotezy:

$$H_0^{(1)} : \rho^{(1)} \leqslant 0,5;$$
$$H_0^{(2)} : \rho^{(2)} \leqslant 0,5.$$

I relacje





Testowanie zbioru hipotez z relacją $\mathcal{R}$

Rozważmy zbiór eksperymentów $\mathcal{E} = \{\mathcal{E}^{(i)} : i \in I = \{1, ..., m\}\}$. Każdemu eksperymentowi $\mathcal{E}^{(i)}$ odpowiada rozkład $P^{(i)}_{\theta_0^{(i)}}$ z rodziny rozkładów indeksowanej zbiorem $\Theta^{(i)}$.

Rozważmy zbiór hipotez $\mathcal{H} = \{H_0^{(i)} : i \in I\}$. Każda hipoteza zerowa jest związana z wyborem podzbioru $\Theta_0^{(i)} \subset \Theta^{(i)}$.

Symbol $\psi_i \in \{0, 1\}$ oznacza wynik testowania dla hipotezy $H_0^{(i)}$.
Symbol $H_i \in \{0, 1\}$ oznacza stan hipotezy $H_0^{(i)}$.

Testowanie zbioru hipotez z relacją $\mathcal{R}$

Relacja hierarchii

Symbolem $\mathcal{R}$ oznaczmy przechodnią, asymetryczną i antyzwrotną relację na zbiorze hipotez $\mathcal{H}$. Relacja $\mathcal{R}$ spełnia warunki

1. $R(i,j) = R(j,k) = 1 \Rightarrow R(i,k) = 1$;
2. $\forall_{i,j,k} R(i,i) = 0$;
3. $R(i,j) + R(j,i) \leqslant 1$.

Zgodność

Wektor $\psi = (\psi_1, ..., \psi_m)$ jest zgodny z relacją $\mathcal{R}$ jeżeli

$$(R(i,j) = 1) \Rightarrow (\psi_i \geqslant \psi_j).$$

Testowanie zbioru hipotez z relacją $\mathcal{R}$

Relacja hierarchii

Symbolem $\mathcal{R}$ oznaczmy przechodnią, asymetryczną i antyzwrotną relację na zbiorze hipotez $\mathcal{H}$. Relacja $\mathcal{R}$ spełnia warunki

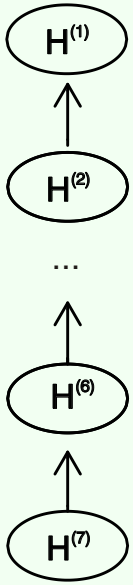
1. $R(i,j) = R(j,k) = 1 \Rightarrow R(i,k) = 1$;
2. $\forall_{i,j,k} R(i,i) = 0$;
3. $R(i,j) + R(j,i) \leqslant 1$.

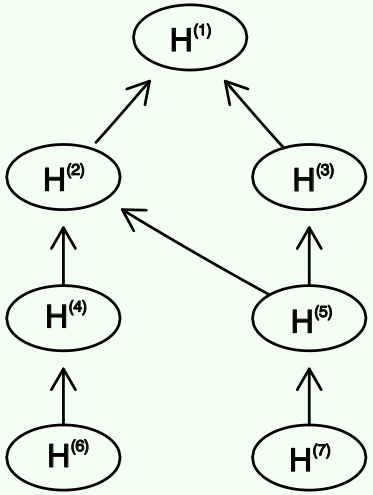
Zgodność

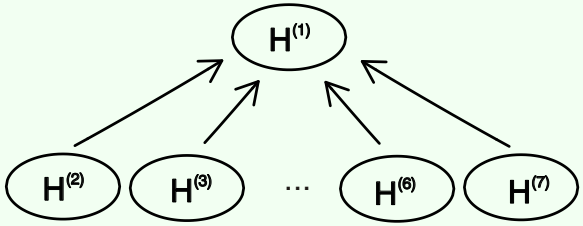
Wektor $\psi = (\psi_1, ..., \psi_m)$ jest zgodny z relacją $\mathcal{R}$ jeżeli

$$(R(i,j) = 1) \Rightarrow (\psi_i \geqslant \psi_j).$$

Przykłady relacji hierarchii







Kontrola współczynnika FWER

Domknięcie wyników testowania

Symbolem $\hat{\psi}$ oznaczamy domknięcie wyników testowania ψ, określone następująco

$$\hat{\psi}_i = \max(\psi_i, \{\psi_j : R(i,j) = 1\}).$$

Twierdzenie 1

Domknięcie wyników testowania nie zmienia współczynnika FWER.

Uwaga !!!

Domknięcie wyników testowania zwiększa współczynniki FDR i PFER.

Kontrola współczynnika FWER

Domknięcie wyników testowania

Symbolem $\hat{\psi}$ oznaczamy domknięcie wyników testowania ψ, określone następująco

$$\hat{\psi}_i = \max(\psi_i, \{\psi_j : R(i,j) = 1\}).$$

Twierdzenie 1

Domknięcie wyników testowania nie zmienia współczynnika FWER.

Uwaga !!!

Domknięcie wyników testowania zwiększa współczynniki FDR i PFER.

Przykład: Domknięcie procedury BH nie kontroluje FDR

Rozważmy zbiór 100 hipotez z relacją liniową, $H^{(1)} = 1$ oraz $H^{(i)} = 0$ dla $i \geqslant 2$. W procedurze Benjaminiego-Hochberga $p^{(2:100)} = \min\{p^{(i)} : i \geqslant 2\}$ jest porównywana z $\alpha^{(2:100)} = \frac{2}{100}\alpha$.

$$Pr\left(p^{(2:100)} \leqslant \frac{2}{100}\alpha\right) = 1 - \left(1 - \frac{2}{100}\alpha\right)^{99}.$$

Jeżeli $p^{(2:100)} \leqslant \frac{2}{100}\alpha$, to odrzucana jest jedna z 99 prawdziwych hipotez, a domknięcie odrzuci wszystkie do niej nadrzędne. Współczynnik FDR można więc oszacować z dołu

$$FDR \geqslant \left(1 - (1 - \frac{2}{100}\alpha)^{99}\right) \frac{1}{99} \sum_{i=1}^{99} \frac{i}{i+1} \approx \frac{2 * 99}{100}\alpha \frac{99 - ln(99)}{99} > \alpha.$$

Przykład: Domknięcie procedury BH nie kontroluje FDR

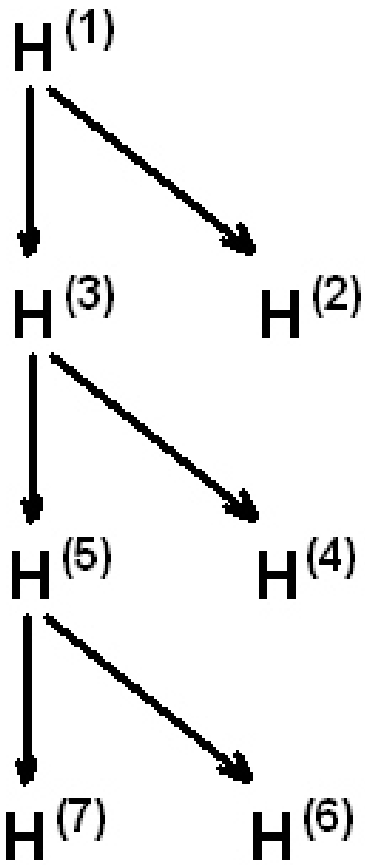
Rozważmy zbiór 100 hipotez z relacją liniową, $H^{(1)} = 1$ oraz $H^{(i)} = 0$ dla $i \geqslant 2$. W procedurze Benjaminiego-Hochberga $p^{(2:100)} = \min\{p^{(i)} : i \geqslant 2\}$ jest porównywana z $\alpha^{(2:100)} = \frac{2}{100}\alpha$.

$$Pr\left(p^{(2:100)} \leqslant \frac{2}{100}\alpha\right) = 1 - \left(1 - \frac{2}{100}\alpha\right)^{99}.$$

Jeżeli $p^{(2:100)} \leqslant \frac{2}{100}\alpha$, to odrzucana jest jedna z 99 prawdziwych hipotez, a domknięcie odrzuci wszystkie do niej nadrzędne. Współczynnik FDR można więc oszacować z dołu

$$FDR \geqslant \left(1 - (1 - \frac{2}{100}\alpha)^{99}\right) \frac{1}{99} \sum_{i=1}^{99} \frac{i}{i+1} \approx \frac{2 * 99}{100}\alpha \frac{99 - ln(99)}{99} > \alpha.$$

Procedura wstępująca (follow up)



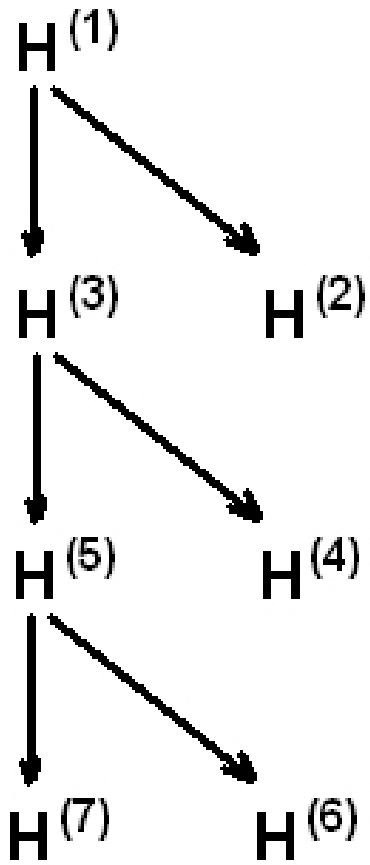
Krok 1: Testujemy hipotezę zerową $H_0^{(7)}$
$p^{(7)} > \alpha^{(7)}$, więc przyjmujemy hipotezę $H_0^{(7)}$

Krok 2: Testujemy hipotezę zerową $H_0^{(6)}$
$p^{(6)} < \alpha^{(6)}$, więc odrzucamy hipotezy
$H_0^{(1)}$, $H_0^{(3)}$, $H_0^{(5)}$, $H_0^{(6)}$

Krok 3: Testujemy hipotezę zerową $H_0^{(4)}$
$p^{(4)} > \alpha^{(4)}$, więc przyjmujemy hipotezę $H_0^{(4)}$

Krok 4: Testujemy hipotezę zerową $H_0^{(2)}$
$p^{(2)} > \alpha^{(2)}$, więc przyjmujemy hipotezę $H_0^{(2)}$

Procedura zstępująca (follow down)



Krok 1: Testujemy hipotezę zerową $H_0^{(1)}$
$p^{(1)}<\alpha^{(1)}$, więc odrzucamy hipotezę $H_0^{(1)}$

Krok 2: Testujemy hipotezę zerową $H_0^{(3)}$
$p^{(3)}>\alpha^{(3)}$, więc przyjmujemy hipotezy
$H_0^{(3)}$, $H_0^{(4)}$, $H_0^{(5)}$, $H_0^{(6)}$, $H_0^{(7)}$

Krok 3: Testujemy hipotezę zerową $H_0^{(2)}$
$p^{(2)}<\alpha^{(2)}$, więc odrzucamy hipotezę $H_0^{(2)}$

Kontrola współczynnika FDR

Twierdzenie 2

Symbolem $\mathcal{C}_i$ oznaczamy rodzinę zbiorów hipotez niebędących ze sobą w relacji

$$\mathcal{C}_i = \{C : i \in C \wedge \forall_{j,k \in C} R(j,k) = 0\},$$

a symbolem $\rho(B)$ oznaczamy liczbę hipotez w relacji do B

$$\rho(B) = 1 + \#\{j : \exists_{i \in B} R(j,i) = 1\}.$$

Procedura zstępująca z parametrami

$$\alpha^{(i)} = \min_{B \in \mathcal{C}_i} \{\alpha_{lin}^{(\rho(B),m)} / \#B\}, \tag{6}$$

kontroluje współczynnik FDR na poziomie α.

Kontrola współczynnika FDR

Twierdzenie 2 cd.

Parametry $\alpha_{lin}^{(i,m)}$ są wyznaczane następująco

$$\begin{aligned}
\alpha_{lin}^{(1,m)} &= \alpha, \\
\alpha_{lin}^{(i,m)} &= \min\left(0.5, \alpha\left[\sum_{k=i}^{m-1}\frac{k-i+1}{k}(1-\alpha_{lin}^{(k+1,m)})\prod_{l=i+1}^{k}\alpha_{lin}^{(l,m)} + \right.\right. \\
&\qquad \left.\left.\frac{m-i+1}{m}\prod_{l=i+1}^{m}\alpha_{lin}^{(l,m)}\right]^{-1}\right), \text{ dla } 1 < i < m, \\
\alpha_{lin}^{(m,m)} &= \min(0.5, m\alpha).
\end{aligned} \tag{7}$$

Kontrola współczynnika PFER

Twierdzenie 3

Symbolem $\phi(i)$ oznaczamy maksymalną moc zbioru hipotez który zawiera $H_0^{(i)}$ i z których żadne dwie nie są ze sobą w relacji

$$\phi(i) = \max_{A \in \mathcal{C}_i} \#A,$$

gdzie

$$\mathcal{C}_i = \{C : i \in C \wedge \forall_{j,k \in C} R(j,k) = 0\}.$$

Procedura zstępująca z parametrami

$$\alpha^{(i)} = \alpha / \left[\phi(i)(1+\alpha)\right] \tag{8}$$

kontroluje współczynnik PFER na poziomie α.

Kontrola współczynnika PFER

Twierdzenie 4

Symbolem $\rho(i)$ oznaczamy

$$\rho(i) = 1 + \#\{j : R(j,i) = 1\}.$$

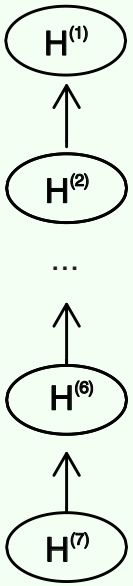
Procedura wstępująca z parametrami

$$\alpha^{(i)} = \alpha/[m * \rho(i)] \tag{9}$$

kontroluje współczynnik PFER na poziomie α.

Przykład relacji $\mathcal{R}_1$

$\mathcal{R}_1(i,j)$	j=1	j=2	j=3	j=4	j=5	j=6	j=7
i=1	0	1	1	1	1	1	1
i=2	0	0	1	1	1	1	1
i=3	0	0	0	1	1	1	1
i=4	0	0	0	0	1	1	1
i=5	0	0	0	0	0	1	1
i=6	0	0	0	0	0	0	1
i=7	0	0	0	0	0	0	0

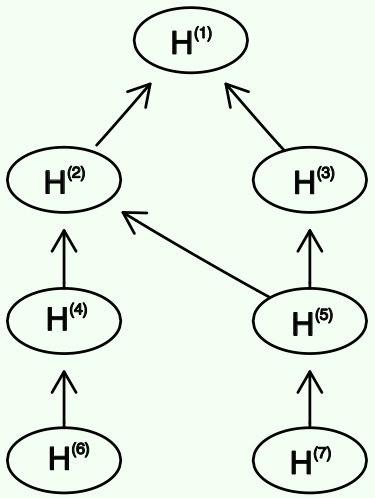


Parametry dla relacji $\mathcal{R}_1$

	Relacja $\mathcal{R}_1$				
	$\rho(i)$	$\phi(i)$	$\alpha_{FU}^{PFER}(i)$	$\alpha_{FD}^{PFER}(i)$	$\alpha_{FD}^{FDR}(i)$
i=1	1	1	0.00714	0.04761	0.05000
i=2	2	1	0.00357	0.04761	0.09523
i=3	3	1	0.00238	0.04761	0.13636
i=4	4	1	0.00178	0.04761	0.17391
i=5	5	1	0.00142	0.04761	0.20833
i=6	6	1	0.00119	0.04761	0.24000
i=7	7	1	0.00102	0.04761	0.35000

Przykład relacji $\mathcal{R}_2$

$\mathcal{R}_2(i,j)$	j=1	j=2	j=3	j=4	j=5	j=6	j=7
i=1	0	1	1	1	1	1	1
i=2	0	0	0	1	0	1	0
i=3	0	0	0	0	1	0	1
i=4	0	0	0	0	0	1	0
i=5	0	0	0	0	0	0	1
i=6	0	0	0	0	0	0	0
i=7	0	0	0	0	0	0	0



Parametry dla relacji $\mathcal{R}_2$

	Relacja $\mathcal{R}_2$				
	$\rho(i)$	$\phi(i)$	$\alpha_{FU}^{PFER}(i)$	$\alpha_{FD}^{PFER}(i)$	$\alpha_{FD}^{FDR}(i)$
i=1	1	1	0.00714	0.04761	0.05000
i=2	2	2	0.00357	0.02380	0.04762
i=3	2	2	0.00357	0.02380	0.04762
i=4	3	2	0.00238	0.02380	0.06818
i=5	4	2	0.00178	0.02380	0.08696
i=6	4	2	0.00178	0.02380	0.10417
i=7	5	2	0.00142	0.02380	0.10417

Scenariusz symulacji

Obserwacje pochodzą z rozkładu $\mathcal{N}(\mu, I_{7\times 7})$, $\mu = (\mu_1, ..., \mu_7)$.
Stawiamy hipotezy zerowe postaci

$$H_0^{(i)} : \mu_i = 0.$$

Rozważmy sześć następujących scenariuszy symulacyjnych

$$\begin{aligned}
\mu^{(1)} &= (0, 0, 0, 0, 0, 0, 0),\\
\mu^{(2)} &= (2, 2, 2, 2, 0, 0, 0),\\
\mu^{(3)} &= (3, 2.75, 2.5, 2.25, 0, 0, 0),\\
\mu^{(4)} &= (2, 2, 2, 2, 2, 2, 2),\\
\mu^{(5)} &= (3, 2.75, 2.5, 2.25, 2, 1.75, 1.5),\\
\mu^{(6)} &= (3, 0, 0, 0, 0, 0, 0).
\end{aligned}$$

Wyniki dla domknięć popularnych procedur testowania

		domknięcie single-step Bonferroni	domknięcie step-up Benjamini-Hochberg
Relacja	Wektor μ	PFER	FDR
$\mathcal{R}_1$	$\mu^{(1)}$	0.201	0.050
$\mathcal{R}_1$	$\mu^{(2)}$	0.073	0.046
$\mathcal{R}_1$	$\mu^{(6)}$	0.151	0.061
$\mathcal{R}_2$	$\mu^{(1)}$	0.151	0.050
$\mathcal{R}_2$	$\mu^{(2)}$	0.045	0.035
$\mathcal{R}_2$	$\mu^{(6)}$	0.102	0.056
$\mathcal{R}_3$	$\mu^{(1)}$	0.093	0.050
$\mathcal{R}_3$	$\mu^{(2)}$	0.294	0.028
$\mathcal{R}_3$	$\mu^{(6)}$	0.044	0.042

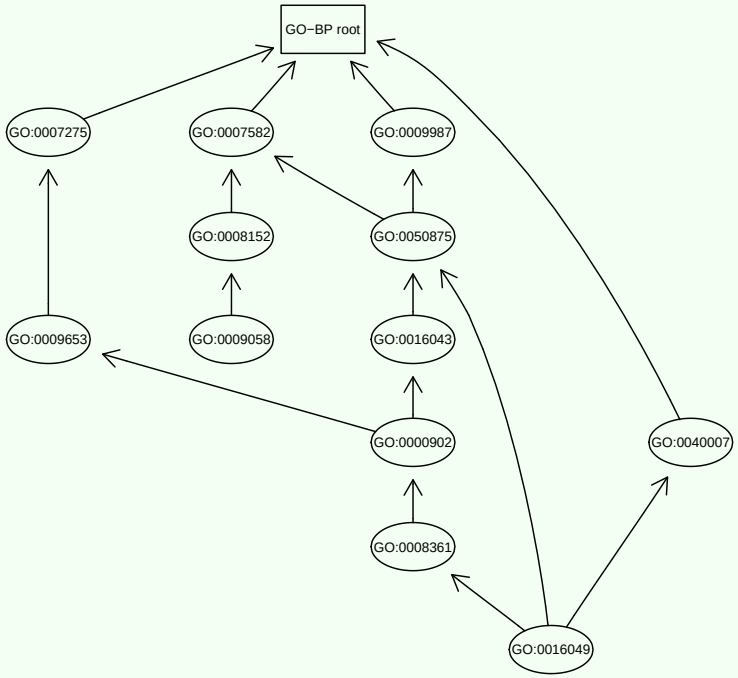
Wyniki dla proponowanych procedur testowania

Relacja	Wektor μ	follow-up PFER	follow-down PFER	follow-down FDR	follow-up E(S)	follow-down E(S)	follow-down E(S)
$\mathcal{R}_1$	$\mu^{(1)}$	0.050	0.050	0.050	—	—	—
$\mathcal{R}_1$	$\mu^{(2)}$	0.007	0.008	0.017	1.712	1.433	1.852
$\mathcal{R}_1$	$\mu^{(3)}$	0.007	0.022	0.035	2.503	2.761	3.236
$\mathcal{R}_1$	$\mu^{(4)}$	—	—	—	3.384	1.634	2.670
$\mathcal{R}_1$	$\mu^{(5)}$	—	—	—	3.425	3.258	4.842
$\mathcal{R}_1$	$\mu^{(6)}$	0.031	0.045	0.045	0.712	0.909	0.912
$\mathcal{R}_2$	$\mu^{(1)}$	0.050	0.050	0.050	—	—	—
$\mathcal{R}_2$	$\mu^{(2)}$	0.006	0.008	0.010	1.552	1.432	1.722
$\mathcal{R}_2$	$\mu^{(3)}$	0.006	0.022	0.023	2.314	2.679	3.034
$\mathcal{R}_2$	$\mu^{(4)}$	—	—	—	2.975	1.638	2.265
$\mathcal{R}_2$	$\mu^{(5)}$	—	—	—	3.092	3.184	4.186
$\mathcal{R}_2$	$\mu^{(6)}$	0.028	0.044	0.043	0.712	0.909	0.912
$\mathcal{R}_3$	$\mu^{(1)}$	0.050	0.050	0.050	—	—	—
$\mathcal{R}_3$	$\mu^{(2)}$	0.011	0.015	0.006	1.448	1.272	1.303
$\mathcal{R}_3$	$\mu^{(3)}$	0.011	0.022	0.007	2.225	2.365	2.394
$\mathcal{R}_3$	$\mu^{(4)}$	—	—	—	2.345	1.915	1.967
$\mathcal{R}_3$	$\mu^{(5)}$	—	—	—	2.784	3.069	3.116
$\mathcal{R}_3$	$\mu^{(6)}$	0.021	0.043	0.022	0.714	0.909	0.912

Zastosowanie do analiz z użyciem Gene Ontology

Gene Ontology składa się z trzech ontologii, każdej reprezentowanej przez graf skierowany acykliczny (directed acyclic graph DAG)

- MF - funkcje molekularne,
- BP - procesy biologiczne,
- CC - komponenty komórkowe.



Zastosowanie do analiz z użyciem Gene Ontology

Rozważmy zbiór współdziałających genów i funkcję biologiczną f_i. Symbolem $\rho^{(i)}$ oznaczmy częstość występowania funkcji f_i w rozważanym zbiorze genów, a symbolem $\rho_0^{(i)}$ w zbiorze wszystkich genów.

Hipotezy w rozważanym zagadnieniu biologicznym

$H_0^{(i)}$: $\rho^{(i)} = \rho_0^{(i)}$
Rozważana grupa genów **nie uczestniczy** w procesie (nie pełni funkcji) f_i.

$H_A^{(i)}$: $\rho^{(i)} > \rho_0^{(i)}$
Rozważana grupa genów **uczestniczy** w procesie (pełni funkcję) f_i.

Gene Ontology: Warsztat statystyczny

	badany zbiór	pozostałe geny	$\sum$
# posiadających funkcję f_i	k	C-k	C
# nie posiadających funkcji f_i	n-k	G-C-(n-k)	G-C
$\sum$	n	G-n	G

Dla każdego genu i każdej funkcji f_i wyznaczamy wartości p korzystając z jednostronnego dokładnego testu Fishera

$$p = Pr(X \geqslant k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-1}}{\binom{G}{n}}, \qquad (10)$$

- X – # genów posiadających funkcje f_i w badanym zbiorze,
- G – # wszystkich genów,
- C – # wszystkich genów posiadających funkcje f_i,
- n – # genów w badanym zbiorze,
- k - # genów posiadających funkcje f_i.

Wyniki dla danych rzeczywistych

Procedura testowania	#odrzuc. H_0	ocena wsp. błędu	#odrzuc. fałszywych H_0
Kontrola FDR (Tw. 2)	3892	FDR $= 0.01725$	3823
Kontrola PFER (Tw. 3)	1706	PFER $= 0.0209$	1683
Kontrola PFER (Tw. 4)	1191	PFER $= 0.0164$	1173
Domknięcie procedury Hochberga	2300	FWER $= 0.0382$	2255

Dla 1099 białek rozważano zbiór 271 hipotez, odpowiadających przypuszczeniu, że białko g_i uczestniczy w procesie biologicznym f_j.

Oceny V, S i R, oraz oceny współczynników PFER, FWER i FDR wyznaczono metodą one leave out cross validation.

Optymalny ranking p-wartości

Wspólna praca z:

- Adam Zagdanski (Politechnika Wrocławska, Uniwersytet w Toronto),
- Rafal Kustra (Uniwersytet w Toronto).

Scenariusz symulacji

- Rozważmy zbiór m hipotez postaci

$$\begin{aligned} H_0^{(i)} &: \mu_i = 0, \\ H_A^{(i)} &: \mu_i = \mu_A. \end{aligned} \tag{11}$$

- Dla każdej hipotezy zerowej $H_0^{(i)}$ określmy

$$\pi_1^{(i)} = \begin{cases} 0.1, \text{ dla } 1 \leqslant i \leqslant \frac{m}{3}, \\ 0.3, \text{ dla } \frac{m}{3} < i \leqslant \frac{2}{3}m, \\ 0.5, \text{ dla } \frac{2}{3}m < i \leqslant m. \end{cases} \tag{12}$$

- Stan każdej hipotezy zerowej losujemy z rozkładu dwumianowego: $H^{(i)} \sim \mathcal{B}(1, \pi_1^{(i)})$.
- Obserwacje $\{x(i)\}_{i=1,\ldots,m}$ losujemy z rozkładu $\mathcal{N}(\mu_i, 1)$, gdzie $\mu_i = H^{(i)}$.
- P-wartości wyznaczamy ze wzoru: $p^{(i)}(x(i)) = 1 - \Phi(x(i))$.

Współczynnik błędu FDR

Naturalnym estymatorem współczynnika FDR jest:

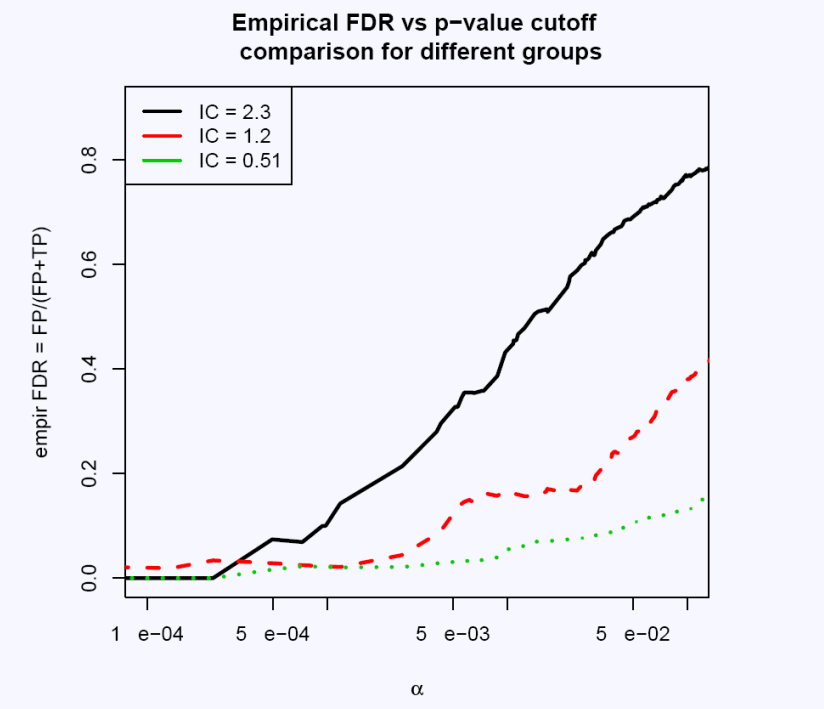
$$\widehat{FDR}(\alpha) = \frac{FP(\alpha)}{FP(\alpha) + TP(\alpha)}, \tag{13}$$

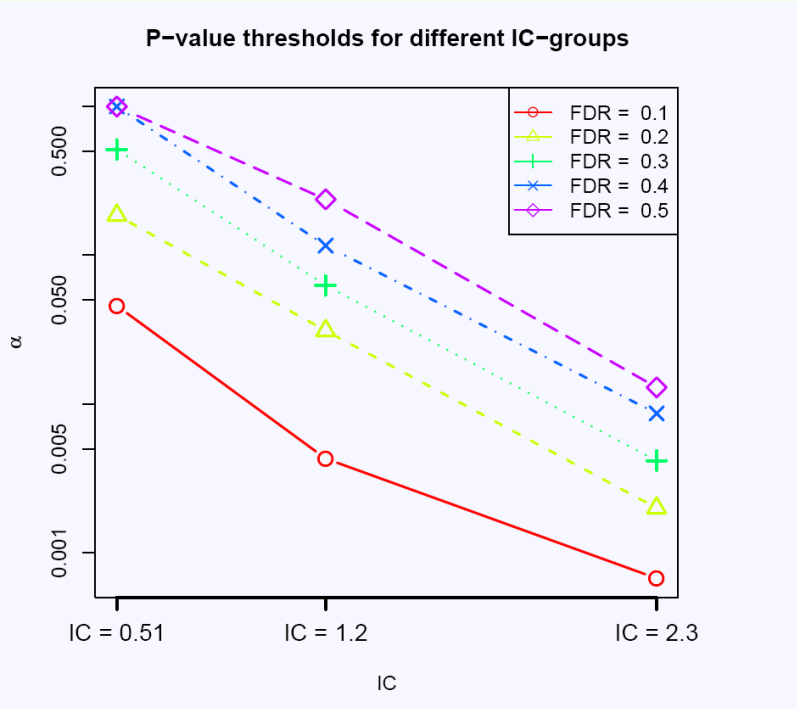
gdzie $FP(\alpha)$ i $TP(\alpha)$ oznaczają liczbę „false positives" i „true positives"

$$FP(\alpha) = \sum_{i=1}^{m} (1 - H^{(i)}) 1_{p^{(i)} < \alpha}, \tag{14}$$

$$TP(\alpha) = \sum_{i=1}^{m} H^{(i)} 1_{p^{(i)} < \alpha}. \tag{15}$$

Wyniki symulacyjne: FDR dla różnych grup hipotez





Propozycja zmodyfikowania poziomu istotności

$$\log(\alpha^{(\pi_1)}_{adjusted}) = \log(\alpha^{(\pi_1)}_{original}) + g\log(\pi_1) + h$$

gdzie g, h to parametry.

Literatura


Ch.R. Genovese, K. Roeder, and L. Wasserman.
False discovery control with p-value weighting.
Biometrika, 93(3):509–524, 2006.


L. Sun, R.V. Craiu, A.D. Paterson, and S.B. Bull.
Stratified false discovery control for large-scale hypothesis testing ...
Genetic Epidemiology, 30:519–530, 2006.


P. Biecek, A. Zagdański and R. Kustra.
Knowledge-based approach to handling multiple testing problem in functional genomics studies. (w przygotowaniu).


G. Xiao and W. Pan.
Gene function prediction by a combined analysis of gene expression data ...
Journal of Bioinformatics and Computational Biology, 3(6):1371–1389, 2005.


C. Stark, B.J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers.
BioGRID: a general repository for interaction datasets.
Nucleic Acids Research, 34:535–539, 2006.


M. Ashburner, C.A. Ball, J.A. Blake, et. al.
Gene Ontology: tool for the unification of biology. the gene ontology consortium.
Nature Genetics, 25(1):25–29, 2000.

Kryteria oceny: ROC, AUC & pAUC

- Krzywa Receiver Operating Characteristic (ROC)

$$(FPR(t), TPR(t)) : t \in [0, 1],$$

$$FPR(t) = FP/N = FP/(TN + FP),$$

$$TPR(t) = TP/P = TP/(TP + FN).$$

- Obszar pod krzywą (Area Under Curve, w skrócie AUC)

$$AUC = \int_0^1 ROC(u)du.$$

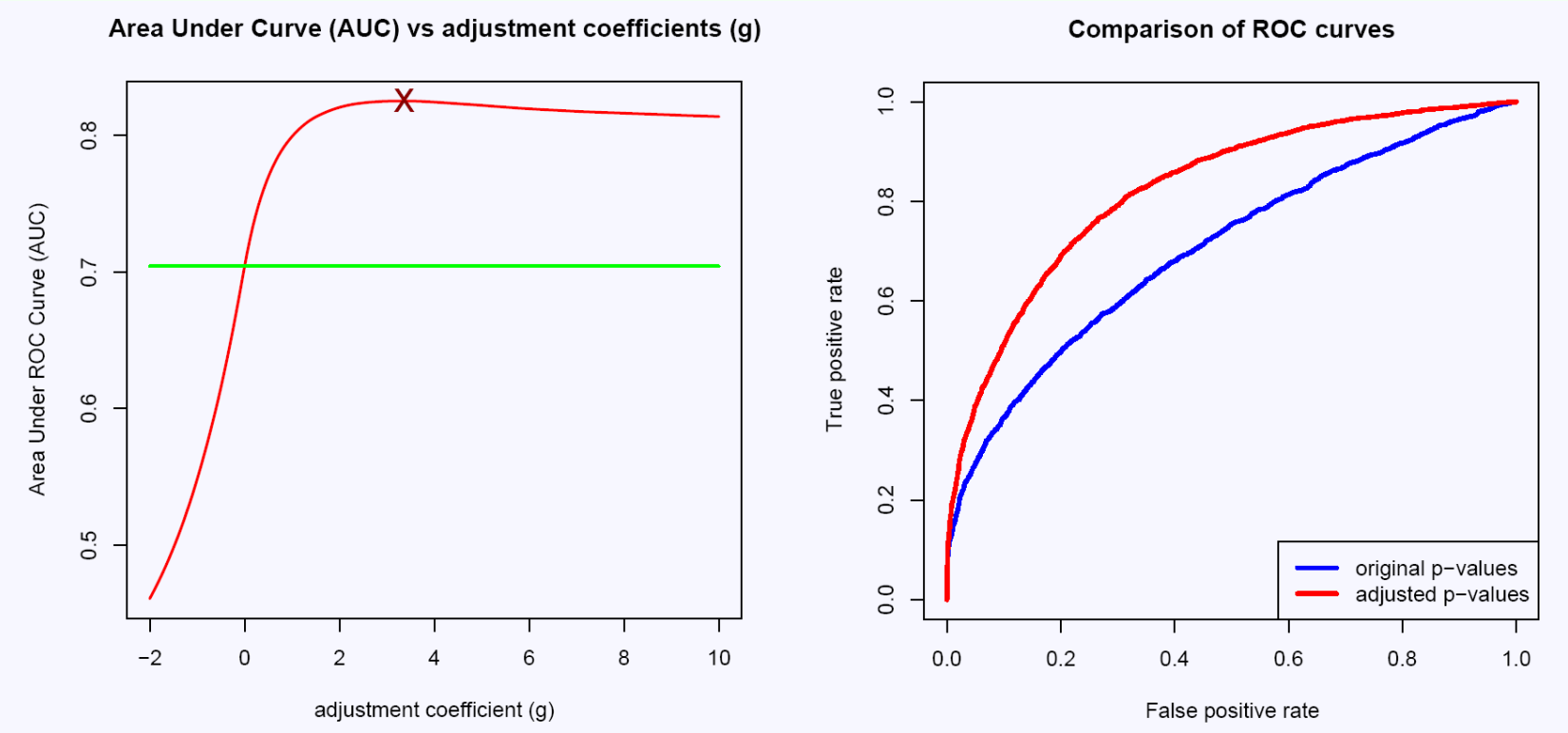
- Częściowy obszar pod krzywą (Partial Area Under Curve, w skrócie pAUC)

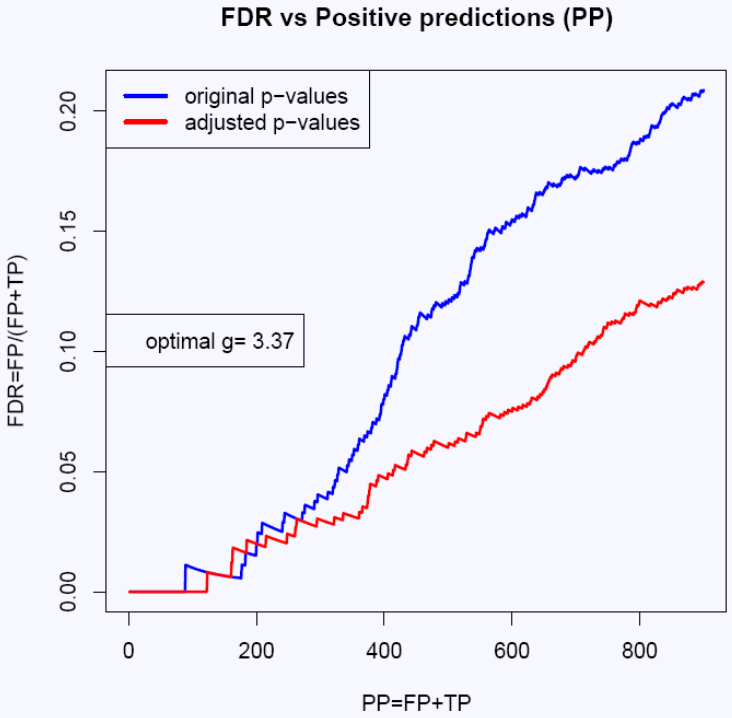
$$pAUC(t) = \int_0^t ROC(u)du.$$

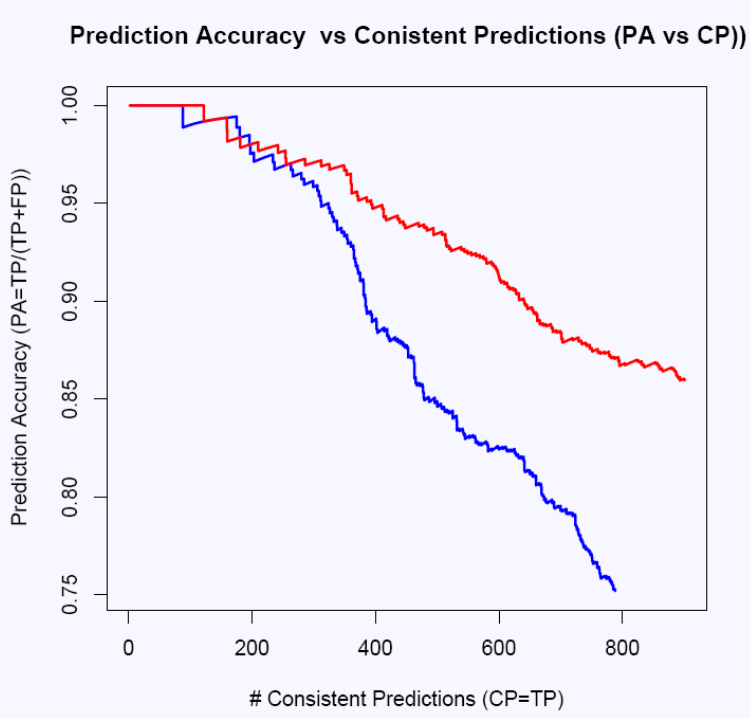
Modyfikacja p-wartości

$$p_{adjusted}^{(\pi_1)} = p_{original}^{(\pi_1)} \exp(g\pi_1)$$

gdzie g – parameter do oceny.




FDR vs Positive predictions (PP)
original p−values
adjusted p−values
optimal g= 3.37
FDR=FP/(FP+TP)
0.00
0.05
0.10
0.15
0.20
0
200
400
600
800
PP=FP+TP


Prediction Accuracy vs Conistent Predictions (PA vs CP))
Prediction Accuracy (PA=TP/(TP+FP))
0.75
0.80
0.85
0.90
0.95
1.00
0
200
400
600
800
Consistent Predictions (CP=TP)

Rozważania teoretyczne (kontynuacja)

- Przypuśćmy, że α jest dopasowywane dla ustalonego poziomu FDR równego γ, gdzie $\gamma \in (0, 1)$
- Ponieważ FDR maleje z malejącym α (monotonicznie), to dla ustalonego poziomu FDR (oznaczonego γ), istnieje jednoznacznie wyznaczone α (które spełnia **??**).
- Równanie (**??**) możemy przedstawić w postaci:

$$\frac{\pi_1}{1-\pi_1} \cdot \frac{\beta(\alpha)}{\alpha} = \frac{1-\gamma}{\gamma}.$$

- Więc, dla rosnącego π_1 „poziom odcięcia" dla p-wartości α musi rosnąć.

Zastosowanie do analiz z użyciem Gene Ontology

Rozważmy grupę współdziałających genów i funkcję biologiczną f_i. Symbolem $\rho^{(i)}$ oznaczmy częstość występowania funkcji f_i w rozważanej grupie genów, a symbolem $\rho_0^{(i)}$ w zbiorze wszystkich genów.

Hipotezy w zagadnieniu identyfikacji funkcji

$H_0^{(i)}$: $\rho^{(i)} = \rho_0^{(i)}$
Rozważana grupa genów **nie uczestniczy** w procesie (nie pełni funkcji) f_i.

$H_A^{(i)}$: $\rho^{(i)} > \rho_0^{(i)}$
Rozważana grupa genów **uczestniczy** w procesie (pełni funkcję) f_i.

GSFEA: Warsztat statystyczny

	badany zbiór	pozostałe geny	$\sum$
# posiadające funkcje f_i	k	C-k	C
# nie posiadające funkcji f_i	n-k	G-C-(n-k)	G-C
$\sum$	n	G-n	G

Dla każdego genu i każdej funkcji f_i wyznaczamy wartości p korzystając z jednostronnego dokładnego testu Fishera

$$p = Pr(X \geqslant k) = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-1}}{\binom{G}{n}}, \qquad (16)$$

- X – # genów posiadających funkcje f_i w badanym zbiorze,
- G – # wszystkich genów,
- C – # wszystkich genów posiadających funkcje f_i,
- n – # genów w badanym zbiorze,
- k - # genów posiadających funkcje f_i.

Schemat algorytmu modyfikacji p-wartości

Dla każdego kroku schematu 5 fold cross validation wykonać:

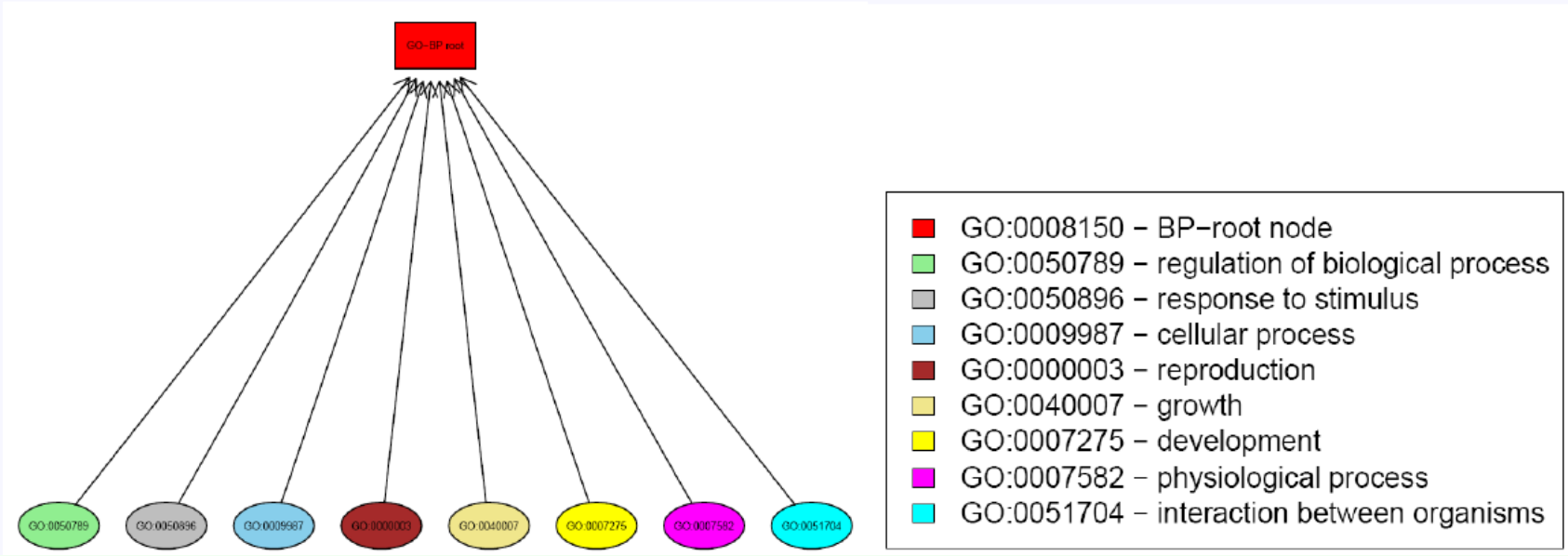
- Na zbiorze uczącym ocenić π_1 dla wszystkich rozważanych funkcji biologicznych f_i.
- Ocenić optymalny współczynnik modyfikacji (czyli $\hat{g}$) używając schematu cross validation (zagnieżdżonego). Optymalne $\hat{g}$ wybrać maksymalizując kryteria AUC lub pAUC.
- Wykonać transformacje p-wartości używając oceny współczynnika modyfikacji $\hat{g}$:

$$p_{adjusted}^{(\widehat{\pi_1})} = p_{original}^{(\widehat{\pi_1})} exp(\hat{g}\widehat{log(\pi_1)}).$$

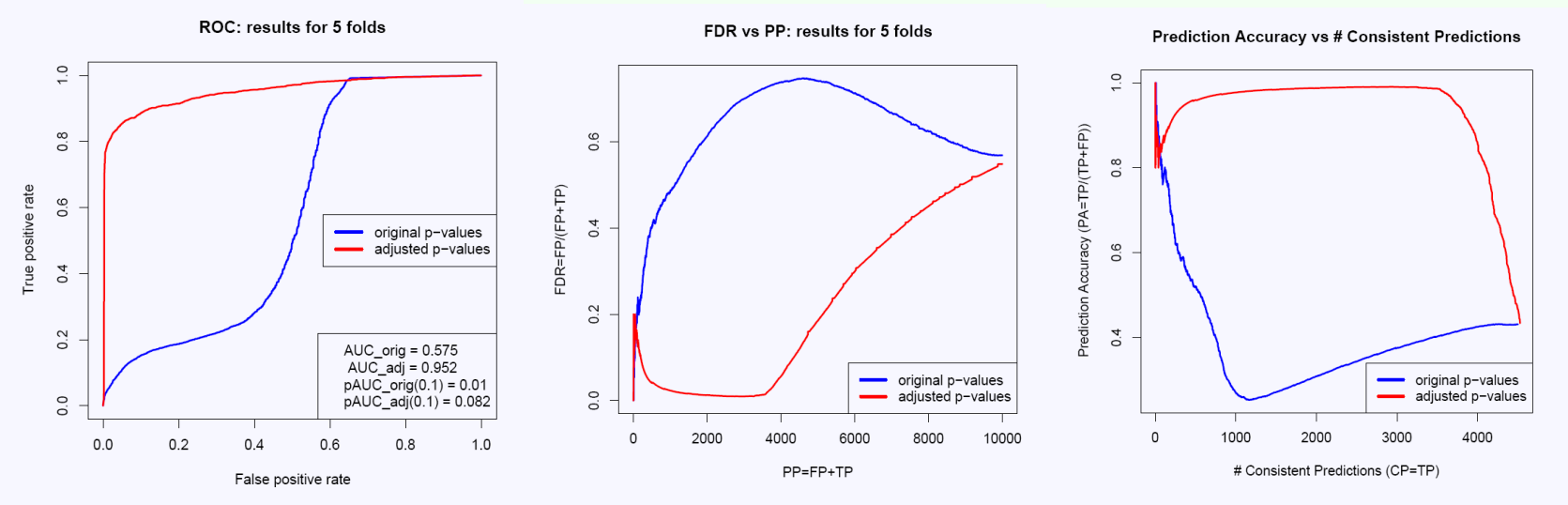
Wykorzystany zbiór danych

- Korzystamy ze zbioru danych o znanych interakcjach białko-białko (PPI) dla drożdży (inne analizy z wykorzystaniem tych danych patrz np. Xiao and Pan (2005)).
  Protein–Protein Interactions (PPI)
  BioGRID (June, 2006, v.2.0.20)
  http://www.thebiogrid.org/1;
- Korzystamy ze zbioru adnotacji funkcji biologicznych do genów drożdzy Gene-Ontology Biological Process (GO-BP) annotations (http://www.geneontology.org/1) Bioconductor pakiety: YEAST i GO (Marzec 2006, v.1.12.0);
- Uzasadnienie: około 70 { 80% białek wchodzących ze sobą w interakcje pelni tę samą funkcję biologiczną.

Przykład: pierwszy poziom Gene Ontology



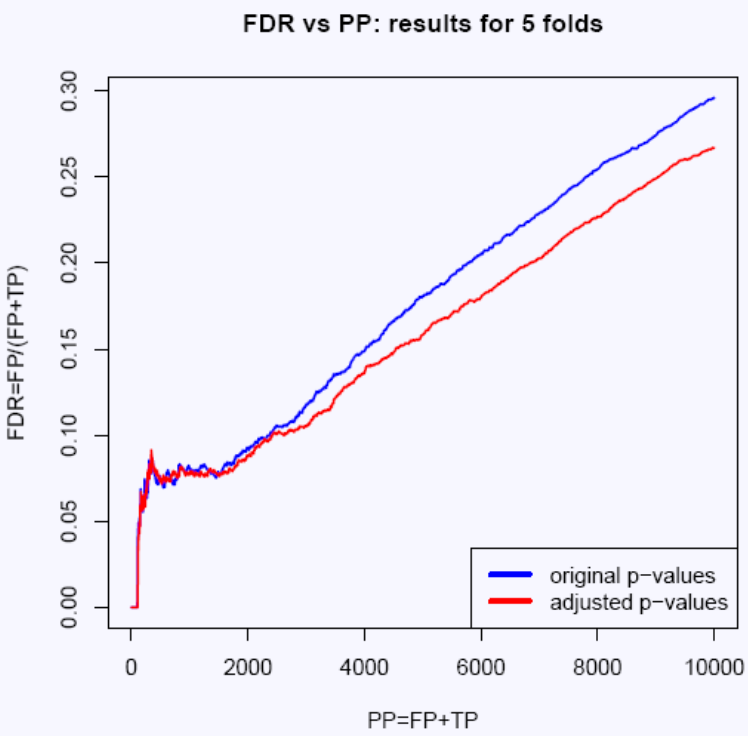
	GO:0050789	GO:0050896	GO:0009987	GO:0000003	GO:0040007	GO:0007275	GO:0007582	GO:0051704
$\hat{\pi}_1$	0.169	0.133	0.988	0.075	0.033	0.111	0.981	0.034
$\hat{IC}$	1.776	2.017	0.012	2.594	3.396	2.195	0.019	3.378

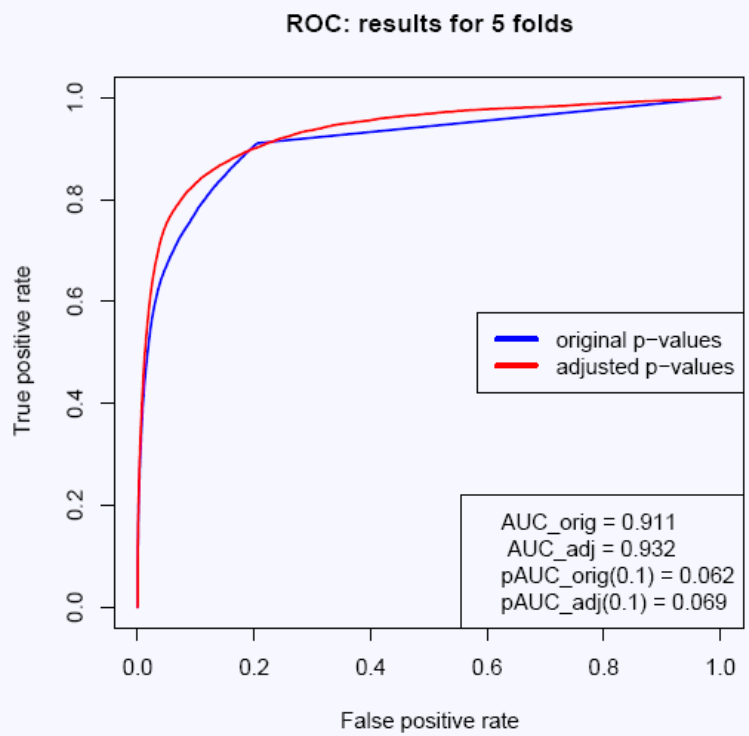


	fold1	fold2	fold3	fold4	fold5	all
bez mod.	0.010	0.011	0.008	0.011	0.010	0.010
z mod.	0.081	0.083	0.085	0.083	0.080	0.082
wzg. popr.	710.0%	654.5%	962.5%	654.5%	700.0%	720.0%

Częściowe pole pod krzywą (pAUC(0.1)). Wyniki dla 5-fold cross-validation.

Przykład: wszystkie informatywne funkcje i geny





	fold1	fold2	fold3	fold4	fold5	all
bez mod.	0.059	0.067	0.063	0.061	0.062	0.062
z mod.	0.067	0.073	0.069	0.068	0.070	0.069
wzg. popr.	13.56%	8.96%	9.52%	11.48%	12.90%	11.29%

Wybór liczby składników w modelu

Rozważmy model eksperymentu losowego $\mathcal{E}$

$$Y = X\beta + \sigma\varepsilon, \tag{17}$$

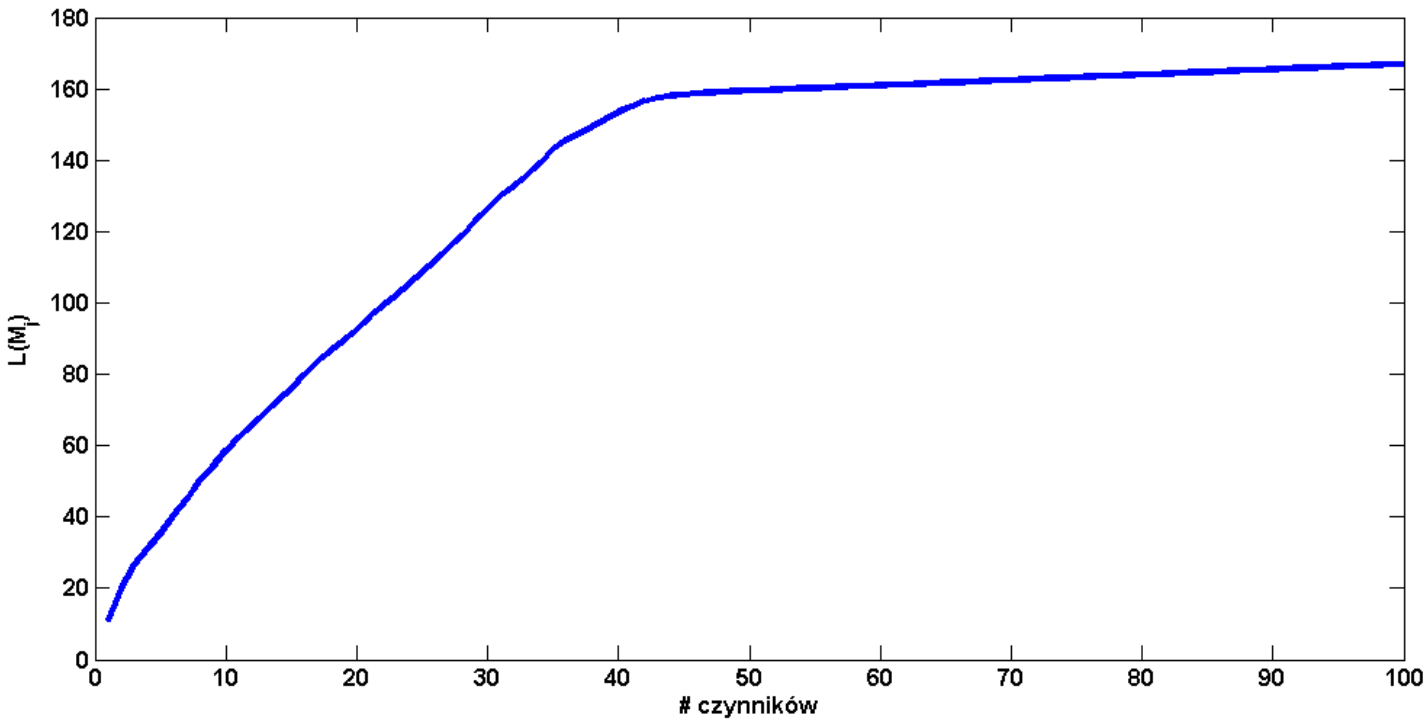
gdzie Y jest $n \times 1$ wektorem obserwacji, X jest macierzą pełnego rzędu o wymiarze $n \times p$, której $j$ta kolumna to n pomiarów $j$tego czynnika, β jest $p \times 1$ wektorem parametrów, σ jest pewną nieznaną stałą a ε jest n elementowym wektorem i.i.d. zmiennych o rozkładzie $\mathcal{N}(0, 1)$.
Przez M_0 oznaczmy indeksy niezerowych składowych wektora β.

Zagadnienie

Identyfikacja składowych wektora β, które $\beta_i \neq 0$ (lub równoważnie, identyfikacja $\beta_i = 0$).

Strategia wyboru modelu „forward selection"

1. W kroku pierwszym rozważamy pusty model $M^{best(0)} = \{\}$. Wartość funkcji wiarygodności oznaczmy $L(M^{best(0)}|Y)$.
2. W drugim kroku rozważamy p modeli postaci $M_i = \{i\}$. Określmy $L(M^{best(1)}|Y) = \max_{i \in \{1 \ldots p\}} L(M_i|Y)$, a model, dla którego $L(M_i|Y) = L(M^{best(1)}|Y)$ oznaczmy przez $M^{best(1)}$.
3. Wykonujemy test ilorazu funcji wiarygodności, okreslający czy $M^{best(1)}$ jest istotnie lepszy niż $M^{best(0)}$ na poziomie istotności $\alpha^{(1)}$.
4. Jeżeli w wyniku testu odrzucimy model $M^{best(1)}$ to kończymy procedurę przeszukiwania. W przeciwnym przypadku za najlepszy dotychczas model uznajemy $M^{best(1)}$ i przechodzimy do następnego kroku.
5. Najlepszy model w kroku i oznaczmy $M^{best(i)}$. Rozważmy $p - i + 1$ modeli $M_i = M^{best(i)} \cup \{i\}$. Oznaczmy $L(M^{best(i+1)}|Y) = \max_i L(M_i|Y)$, a model dla którego $L(M_i|Y) = L(M^{best(i+1)}|Y)$ oznaczmy przez $M^{best(i+1)}$.
6. Wykonujemy test ilorazu wiarygodności, określający czy $M^{best(i+1)}$



Przykładowe wartości $L(M^{best(i+1)}|Y)$. W symulowanym przypadku $n = 200$, $p = 100$, $\sigma^2 = 1$, zbiór M_0 składa się z 45 indeksów, a współczynniki $\beta_i = 0.5$ dla $i \in M_0$ i $\beta_j = 0$ dla $j \in M_0^C$.

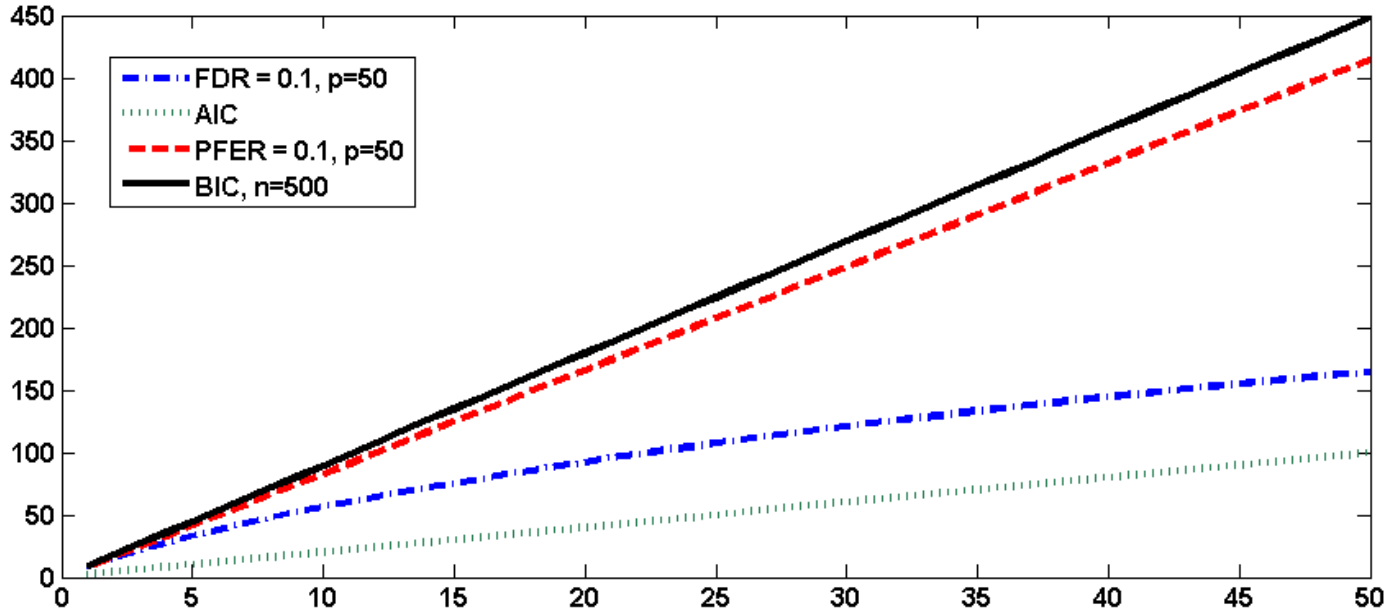
Procedura „forward selection" wybiera model maksymalizujący wartość $\mathcal{S}(M_i)$

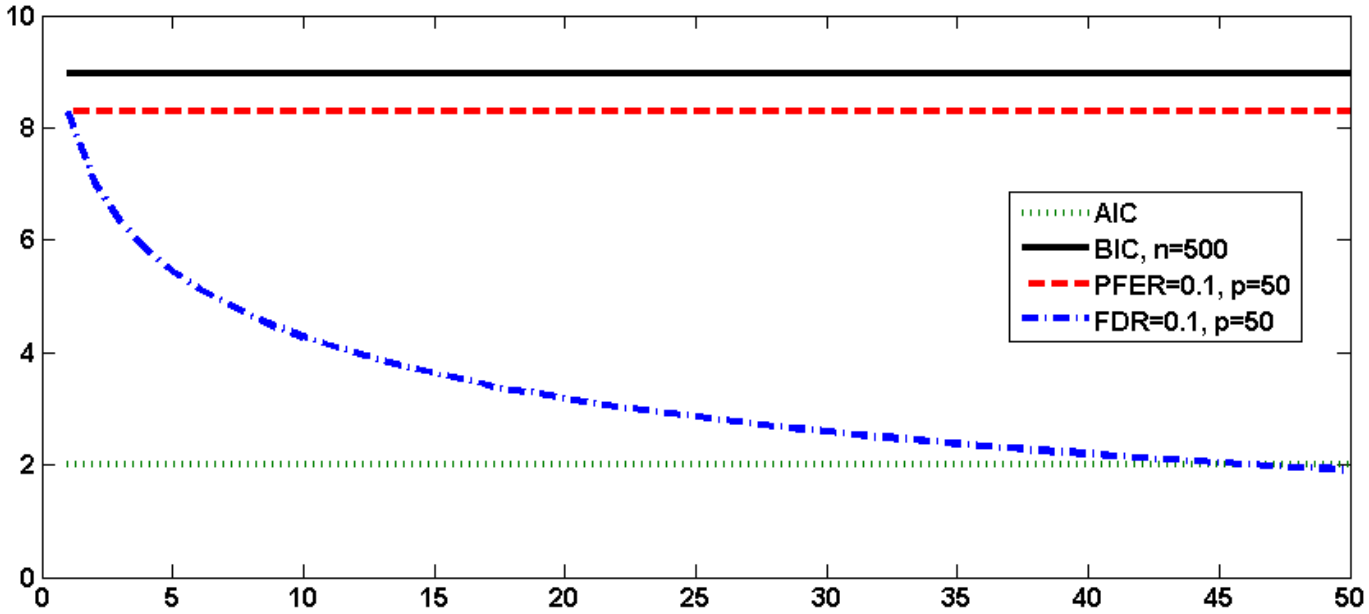
$$\mathcal{S}(M_i) = -2\log L(M_i|Y) - \sum_{j=1}^{|M_i|} c^{(j)} = -2\log L(M_i|Y) - \sum_{j=1}^{|M_i|} \left(\chi_1^2\right)^{-1}(1-\alpha^{(j)})$$

gdzie $\sum_{1=1}^{|M_i|} c^{(i)}$ to kara za rozmiar modelu. Jeżeli $\alpha^{(j)}$ są równe, to powyższe krytrium sprowadza się do kryterium GIC (Generalized Infortmation Criteria), które wybiera model maksymalizujący

$$\mathcal{G}(M_i) = -2\log L(M_i|Y) + \lambda|M_i|$$

gdzie λ to pewna nieujemna stała. Dla $\lambda = 2$ kryterium GIC sprowadza się do AIC, dla $\lambda = \log_2(n)$ sprowadza się do kryterium BIC.


FDR = 0.1, p=50
AIC
PFER = 0.1, p=50
BIC, n=500
450
400
350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
35
40
45
50


10
8
6
4
2
0
0
5
10
15
20
25
30
35
40
45
50
AIC
BIC, n=500
PFER=0.1, p=50
FDR=0.1, p=50

ρ	$\lvert M_0\rvert$	β	n	p	BIC	AIC	$FDR = 0.1$
0	1	0.5	100	50	0.01	3.50	**0.00**
0	1	0.4	200	50	**0.00**	2.21	**0.00**
0	1	0.5	200	50	**0.01**	2.59	**0.01**
0	10	0.5	100	50	-4.77	3.58	**-0.49**
0	10	0.5	200	50	**-0.03**	2.48	0.21
0	10	0.4	200	50	-1.12	2.52	**0.26**
0	30	0.4	200	50	-26.98	1.76	**1.07**
0	30	0.3	500	50	-0.59	1.10	**0.60**
0.9	1	0.4	500	50	**0.00**	1.55	0.01
0.9	10	0.4	500	50	-2.18	1.75	**0.29**
0.9	10	0.5	200	50	-3.20	1.90	**0.11**
0.9	10	0.6	200	50	-2.22	3.12	**1.09**

Średnie różnice pomiędzy liczbą niezerowych β_i a oceną tej liczby. Rozważano 50 markerów. ρ to korelacja pomiędzy statystykami testowymi ($\sigma = 1$). Wyniki uśrednione ze 100 powtórzeń. W każdym wierszu wytłuszczono najlepsze wyniki.

Podsumowanie

- Sformułowano nowe zagadnienie testowania zbioru hipotez i wskazano procedury kontrolujące popularne współczynniki błędu.
- Wykazano brak kontroli współczynników FDR i PFER po domknięciu popularnych procedur testowania.
- Przedstawiono procedurę wykorzystującą informację o różnicach w częstościach wystąpień fałszywych hipotez.,
- Przedstawiono wyniki zastosowań tych procedur w analizach z użyciem Gene Ontology.
- Przedstawiono zastosowanie otrzymanych wyników do kryterium wyboru modelu.