R Shiny를 이용한 다중검정 인터페이스 개발

전남대학교 수학/통계학과 석사과정 박주성

참여인력

발표자 : 박주성

저자 :

박주성(전남대학교 수학/통계학과)

김신준(한국농촌경제연구원, KREI)

오영재(대전 통계청)

교신저자 : 정재식(전남대학교 통계학과 교수)

Abstract

- Bioinfomatics에서 두 집단의 대사체 간의 차이를 찾기 위해서 다중검정법을 사용해야 하며, 이것에 대응되는 방법 중 하나가 FDR(False Discovery Rate) Control Method를 사용하는 것이다.
- 대사체 데이터 연구하는 분야에서 FDR Control Method를 사용하여 유의미한 차이를 얻기 위해선 통계 분야 종사자에게 도움을 구하거나 통계를 직접 공부해야 할 필요가 있다. 이는 어떤 측면에서 비용이 드는 것으로 생각할 수 있다.
- FDR Control Method의 1D Method와 2D Method 대표적인 예시 한 가지씩을 오미자 data(중국산, 한국산의 혼합)를 다중검정 인터페이스에 적용한 결과를 확인한다.

연구 배경

Large-Scale Hypothesis table

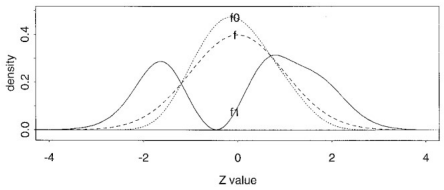
Unknown			Group		Observed				Unknown
Proportion		H_i	Actual	Normal	Abnormal	t_i	z_i	p_i	Decision	Density
N	π_1 Pr{non-null}	H_1	F	$\mu = \mu_{ni}$ $\sigma = \sigma_{ni}$	$\mu = \mu_{di}$ $\sigma = \sigma_{di}$	t_1	z_1	p_1	F	$f_1(z)$
		H_2	F			t_2	z_2	p_2	F	
		H_3	F			t_3	z_3	p_3	T	
		H_4	F			t_4	z_4	p_4	F	
		⋮	⋮			⋮	⋮	⋮	⋮	
		$H_{N*\pi_1}$	F			$t_{N*\pi_1}$	$z_{N*\pi_1}$	$p_{N*\pi_1}$	F	
	π_0 Pr{null}	$H_{N*\pi_1+1}$	T		$\mu = \mu_{ni}$ $\sigma = \sigma_{ni}$	$t_{N*\pi_1+1}$	$z_{N*\pi_1+1}$	$p_{N*\pi_1+1}$	T	$f_0(z)$
		$H_{N*\pi_1+2}$	T			$t_{N*\pi_1+2}$	$z_{N*\pi_1+2}$	$p_{N*\pi_1+2}$	T	
		⋮	⋮			⋮	⋮	⋮	⋮	
		H_{N-2}	T			t_{N-2}	z_{N-2}	p_{N-2}	T	
		H_{N-1}	T			t_{N-1}	z_{N-1}	p_{N-1}	F	
		H_N	T			t_N	z_N	p_N	T	

Mixture dist. & Zero-Assumption

- Bradley Efron이 제안한 가정으로 데이터 전체의 분포가 Null과 Non-Null의 분포의 합으로 이루어져 있다면, 전체 분포에서 0(혹은 평균) 근처에서는 Null에 대응되는 분포의 확률만 존재한다는 가정

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z) : Mixture\ dist.$$

$$f(z) = \pi_0 f_0(z)\ \ for\ z \in Z = [-c, c] : Zero - Assumption$$



- Efron etal. 2001 p.1155
- Zero-Assumption의 예시

FDR Control Method

1D fdr : Bradley Efron etal.

- **Empirical bayes analysis**를 local fdr를 추정 : $fdr1d(z) = \pi_0 f_0(z)/f(z)$

Zero-Assumption 등을 동반한 methods를 사용, $\hat{f}, \hat{f_0}, \hat{\pi}$ 를 이용하여 $fdr1d(z)$를 추정

2D fdr : Alexander Ploner etal.

- **2차원**의 local fdr를 추정 (fdr1d의 취약점 보완) : $fdr2d(z_1, z_2) = \pi_0 f_0(z_1, z_2)/f(z_1, z_2)$

$\hat{f}, \hat{f_0}, \hat{\pi}$ 를 이용하여 $fdr2d(z_1, z_2)$를 추정 (통계량으로 t값과 분산을 사용)

인터페이스 사용

- Data : schisandra data (오미자)
- 27개의 중국산과 30개의 한국산 데이터의 혼합으로 3200개 가량의 유전자 intensity들이 대응됨

인터페이스의 사용 단계

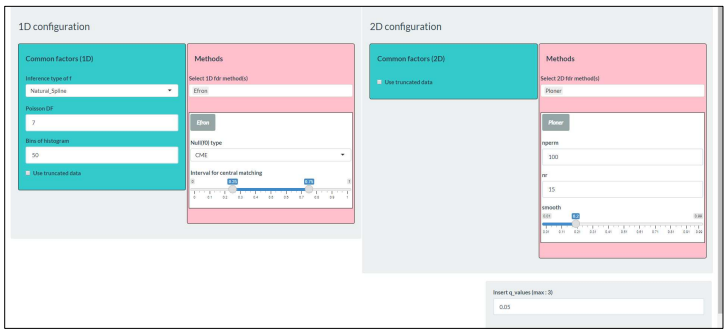
① Data Load : 데이터를 csv 파일 형태로 불러들임

② Method selection : 분석하고자 하는 방법론 및 파라미터 선택

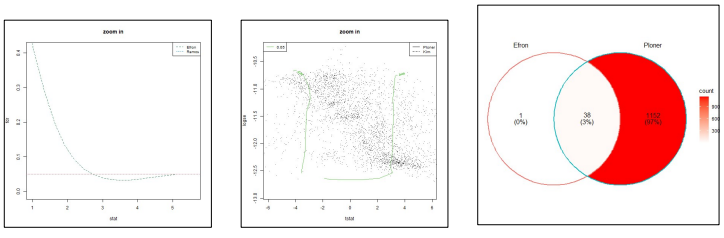
③ Result : 시각화 정보 및 유의한 특성(유전자) 색출

① Data Load

② Method selection



③ Result



1D: Efron　　2D : Ploner　　Venn Diagram

Reference

[1] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological), 57(1):289–300, 1995.

[2] Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. Journal of the American statistical association, 96(456):1151–1160, 2001.

[3] Alexander Ploner, Stefano Calza, Arief Gusnanto, Yudi Pawitan. *Bioinformatics*, Volume 22, Issue 5, 1 March 2006, Pages 556–565

This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research, NO.5120200913674) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF)