

Protocolo de análisis de datos genéticos en Shiny

Antonio Rodríguez Gómez

Supervisado por Mercè Farré

Universidad Autònoma de Barcelona

Septiembre 13, 2018

Caso de estudio

- Definir el problema
- Se diseña un experimento con el objetivo de. . .

Estructura del proyecto

Problema → Protocolo → Aplicación → Conclusiones

Datos de expresión genética

- Presentar la obtención de los datos
- Los datos son. . .
- La medida es..

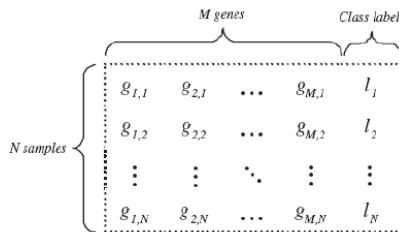


Figure 1: Matriz de los datos

Protocolo de análisis

Objetivos del análisis

- 1 Detectar genes que se expresan. . .
- 2 Buscar diferencias entre tratamientos..
- 3 Patrones. . .

Protocolo de análisis

Hipotesis planteadas

- 1 En algun gen (y cuál o cuales) hay diferencias significativas entre los niveles de expresión entre tatamientos?
- 2 Entre cuáles tratamientos (parejas) hay diferencias significativas?

En algun gen (y cuál o cuales) hay diferencias significativas entre los niveles de expresión entre tatamientos?

- ANOVA para comparar medias entre grupos/tratamientos.
- Cada gen es una variable. (Expresión del gen)
- Diseño balanceado y normalidad en los datos.
- *Ejemplo: En una muestra de 26 individuos, hacemos un experimento con 4 tratamientos. Si aplicamos ANOVA para el gen TFF3:*

Gen	$F_{k-1, N-k}$	P-valor
TFF3	5.29	0.005

- Suponemos que tenemos 20 genes, aplicamos el ANOVA para cada gen y listo?

En algun gen (y cuál o cuales) hay diferencias significativas entre los niveles de expresión entre tatamientos?

- Cuando realizamos un test podemos cometer el error de rechazar la hipótesis nula cuando realmente es cierta. (Error de tipo I)

Cálculo de probabilidades

- Si H_0 es cierta: $P(FP) = \alpha$ y $P(VP) = 1 - \alpha$
- $P(\text{Almenos un } FP \text{ en } m \text{ tests}) = 1 - (1 - \alpha)^m$
- Para $m = 1$, $P(\text{Almenos...}) = 1 - (1 - 0.05)^1 = 0.05$
- Para $m = 50$, $P(\text{Almenos...}) = 1 - (1 - 0.05)^{50} = 0.92$

En algun gen (y cuál o cuales) hay diferencias significativas entre los niveles de expresión entre tatamientos?

- Corregir los p-valores para controlar el error de tipo I

False Discovery Rate (FDR)

- Proporción esperada de falsos positivos entre todos los tests considerados como significativos.
- El método para controlar el FDR: Benjamini&Hochberg.

Nom del gen	Funció del gen	Estadístic F	P-valor	P-valor (Benjamini&Hochberg)
TFF3	Barrier Function	5.2927	0.0086	0.0603*
OCN	Barrier Function	1.2861	0.3094	0.4572
ZO1	Barrier Function	0.9128	0.4544	0.5979
MUC2	Barrier Function	7.2461	0.0022	0.0272***
MUC13	Barrier Function	2.1769	0.1261	0.2425
SI	Enzymed/Hormone	3.1867	0.0488	0.1525
DAO1	Enzymed/Hormone	1.2813	0.3109	0.4572
HNMT	Enzymed/Hormone	0.5736	0.6396	0.6953
ANPEP	Enzymed/Hormone	1.9658	0.1553	0.2773
GCG	Enzymed/Hormone	1.1972	0.3390	0.4709

- Falta...

Entre cuáles tratamientos (parejas) hay diferencias significativas?

- De los genes que han salido significativos en el ANOVA, hacemos comparaciones 2 a 2 entre las medias de los grupos.
- Utilizamos el método de Tukey:

Análisis descriptivo

- 1 Heatmap
- 2 Representación ACP
- 3 LinePlot

Aplicación en Shiny

La idea principal era crear una herramienta de análisis para los grupos de investigación.

- Desarrollada en R con el paquete Shiny.
- Interactiva, sencilla y accesible.
- Repositorio público.

<https://github.com/djangosee/TFGShinyApp>

Conclusiones