Protocol d'Anàlisi de Dades d'Expressió Gènica amb Shiny

Antonio Rodríguez Gómez

13 de juliol de 2018

Abstract (versió en català)

L'anàlisi de dades d'expressió genètica és un dels grans reptes estadístics per a detectar expressions diferencials entre gens sota una condició donada. La idea principal d'aquest treball, és la creació d'un protocol d'anàlisi interactiu de matrius d'expressió genètica. Es presenten mètodes estadístics com l'anàlisi de la variància (ANOVA). Mètodes de correcció per multiplicitat de contrastos també queden descrits. S'utilitzen mètodes gràfics com els mapes de calor. Tot aquest protocol d'anàlisi s'implementa en una aplicació web desenvolupada amb Shiny.

Abstract (versión en castellano)

El análisis de datos de expresión genética es uno de los grandes retos estadísticos para detectar expresiones diferenciales entre genes mediante una condición establecida. La idea principal de este trabajo, es la creación de un protocolo de análisis interactivo de matrices de expresión genética. Se presentan métodos estadísticos como el análisis de la varianza (ANOVA). Métodos de corrección por multiplicidad de contrastes también quedan descritos. Se utilizan métodos gráficos como los mapas de calor. Todo este protocolo de análisis se implementa en una aplicación web desarrollada con Shiny.

Abstract (English version)

Data gene expression analysis is one of the major statistical challenges to detect differential expressions between genes under a given condition. The main idea is the creation of an interactive analysis protocol of gene expression matrix. Methods are presented for detecting differential expression using statistical hypothesis testing methods including analysis of variance (ANOVA). Methods for multiple testing correction and their application are described. Graphical methods such as heatmaps are used. This analysis protocol is implemented in a web application developed in Shiny.

Índex

1 Introducció

Un dels reptes més grans de la biologia actualment és analitzar els volums massius de dades creats, per exemple, en la seqüenciació de DNA. La gran evolució de les tècniques de recollida de dades biològiques ha fet que sigui necessari el desenvolupament de metodologies eficients a l'hora de tractar i analitzar les dades. La disciplina que recull aquestes metodologies s'anomena bioinformàtica.

La bioinformàtica és un àrea emergent interdisciplinària que s'ocupa de l'aplicació de l'informàtica a la recopilació, emmagatzematge, organització, anàlisis, manipulació, presentació y distribució d'informació relativa a les dades biològiques o mèdiques.

Aquest treball s'ha centrat en l'anàlisis de bases de dades d'expressió genètica a partir de diferents condicions experimentals. Al llarg del treball s'utilitzen técniques per analitzar aquests tipus de dissenys on l'objectiu recau en veure quins gens s'expressen significativament sota condicions experimentals establertes.

Amb aquesta premissa, el treball també s'ha enfocat en crear un aplicatiu web capaç de fer un anàlisi estadístic de les matrius d'expressió gènica. L'aplicatiu ha sigut programat amb R, per mitjà del paquet Shiny. Aquest paquet és capaç d'implementar el codi R de manera interactiva. No només implementa el codi R, sinó també és capaç d'interactuar amb diferents llenguatges com html, css o java. Tot el conjunt de l'aplicació ha sigut organitzada i compartida per mitjà d'un repositori creat en la plataforma GitHub. D'aquesta manera el codi queda a disposició de qualsevol usuari que vulgui utilitzar-lo o consultar els métodes empleats.

1.1 Introducció als conceptes bàsics de la bioinformàtica

Cada organisme es defineix pel seu material genètic, el genoma. La informació genètica la trobem emmagatzemada en una macromolècula anomenada DNA, que es troba al nucli de cada cèl·lula.

Un gen consisteix en un segment de DNA que conté el codi per a la producció d'una proteïna. Una única cadena d'ADN conté milers de gens, cadascun sintetitza una proteïna concreta. Per fer-nos a la idea, els humans tenim al voltant de 20.000 gens. La longitud i seqüència d'un gen determina la grandària i la forma de la proteïna que sintetitza i quina funció tindrà aquesta proteïna dins de l'organisme.

La dotació de gens que presenta una espècie, s'anomena genotip, i l'aparença externa d'un caràcter genètic, l'anomenem fenotip. L'expressió del genotip ve determinat, a més de per la càrrega genètica, per l'ambient i el comportament dels éssers vius. Si un gen no s'expressa en un individu, aquest tindrà el mateix fenotip que un individu que no presenti el gen. Però com podem arribar a obtenir una mesura de l'expressió dels gens? Existeixen tècniques per quantificar-ho? La resposta és sí.

La mesura de l'expressió gènica generalment és dur a terme quantificant els nivells de producte del gen. Una tècnica molt utilitzada de mesura de l'expressió gènica que utilitza ARN missatger és la denominada transcripció inversa, seguida de la reacció en cadena quantitativa de la polimerasa (qPCR [1]). Una de les seves principals característiques és la seva sensibilitat, ja que només necessita una única molècula per iniciar el procés de replicació. A més, és molt robusta gràcies al fet que permet utilitzar diferents productes biològics, com cabells, teixits, mucoses, sang, etc. Aquesta tècnica és fonamental per l'anàlisi de dades d'expressió gènica, perquè per l'obtenció d'aquestes dades es requereix una quantitat suficientment gran de producte biològic que no sempre és de fàcil obtenció, per tant, és important disposar d'una tècnica que faciliti la seva replicació de forma controlada, robusta i eficient.

Centres de genòmica s'encarreguen de fer aquests processos i de retornar els resultats en matrius de dades on es recullen els nivells d'expressió per a cada gen. Per tant, és important fer un bon disseny experimental ja que aquests procesos són costosos i requereixen de temps, a més del biaix estadístic que es pot generar.

1.2 Cas d'estudi: Protocol d'análisi d'un OpenArray

Des de la facultat de veterinaria de l'Universitat Autònoma s'han dut a terme estudis experimentals sobre l'expressió dels gens animals en certes condicions experimentals. L'aplicatiu web ha sigut creat per donar suport estadístic al grup d'investigació de la UAB **Nombre del grupo**.

El cas d'estudi que el treball ha contemplat consisteix en un experiment amb animals, concretament, amb porcins. Durant l'experiment s'administraven diferents tractaments/dietes als porcins. D'aquesta manera es volia veure l'afectació d'alguns tractaments en la regulació intestinal i com afectava al creixement dels porcins.

Les dades utilitzades en aquest treball han sigut proporcionades per **Nombre del grupo** per mitjà de la tecnologia OpenArray. El material biològic que s'ha utilitzat per l'obtenció de les dades, han sigut diferents tipus de teixits del intestí. Els gens van ser escollits amb criteris científics pels investigadors i tenen un significat concret dins del funcionament de la regulació intestinal.

Encara que l'aplicatiu web s'ha creat a partir d'aquest estudi, la idea ha sigut generalitzar el codi per poder utilitzar-ho amb altres dissenys experimentals.

[1] Aquesta tècnica serveix per amplificar un fragment d'ADN i la seva utilitat rau en el fet que després de l'amplificació resulta molt més fàcil identificar material genètic amb una gran precisió.

2 Protocol d'anàlisi

En aquest apartat queden definits els mètodes estadístics utilitzats en el protocol d'anàlisi. Cada mètode ha sigut implementat en l'aplicatiu i més endavant es mostren els resultats del cas d'estudi.

2.1 Anàlisi de la variància (ANOVA)

L'anàlisi de la variància (ANOVA) és el mètode clàssic per comparar mitjanes entre grups dos grups o més. Suposem que tenim N observacions repartides en k grups i definim $n = \frac{N}{k}$. Llavors x_{ij} seria l'individu j corresponent al grup i. En aquest cas assumim que l'estudi és balancejat, és a dir, el nombre d'individus per grup és el mateix. Denotem $\bar{x}.$ com la mitjana de la mostra global, i $\bar{x_i}$ com la mitjana del grup i. Les observacions es poden tornar a escriure com:

$$x_{ij} = \bar{x}. + (\bar{x_i} - \bar{x}.) + (x_{ij} - \bar{x_i})$$

Això ens porta al següent model:

$$x_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

on μ i α_i són la mitjana global i la mitjana del grup i respectivament. S'assumeix que el terme d'error ϵ_{ij} és iid i segueix una distribució normal

$$\epsilon_{ij} \sim \mathcal{N}(\mu, \sigma^2)$$

La hipòtesi nul·la en un model ANOVA és que les mitjanes dels grups són iguals, és a dir:

$$\alpha_1 = \alpha_2 = ... = \alpha_k$$

Si això és cert, el terme d'error per a la diferència de grups queda definit com:

$$\bar{x_i} - \mu \sim \mathcal{N}(0, \frac{\sigma^2}{n} = \bar{\sigma}^2)$$

Es pot mesurar la quantitat total de variabilitat entre observacions sumant els quadrats de les diferències entre cadascun $\bar{x}.$ i x_{ij}:

$$SST(\text{Suma de quadrats totals}) = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}.)^2$$

La variabilitat total es pot desglossar en 2 termes:

1. La variabilitat entre grups:

$$SSG = \sum_{i=1}^{k} n_i(\bar{x_i} - \bar{x}.)^2$$

   amb $k - 1$ graus de llibertat.

2. La variabilitat intra grups:

$$SSE = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x_i})^2$$

   amb $N - k$ graus de llibertat.

Per tant, podem escriure la suma de quadrats totals com:

$$SST = SSG + SSE$$

Si la variabilitat entre grups és gran en relació amb la variabilitat intra grups, llavors les dades suggereixen que les mitjanes de les poblacions són significativament diferents. Si no existeixen diferencies, entre els grups, esperaríem que les mitjanes quadràtiques

$$MSG = \frac{SSG}{(k-1)}$$

$$MSE = \frac{SSE}{(N-k)}$$

siguin similars. El test estadístic ANOVA es defineix com la ràtio entre les dues mitjanes quadràtiques:

$$F = \frac{MSG}{MSE}$$

L'estadístic F segueix una distribució F de Snedecor amb $k - 1$ i $N - k$ graus de llibertat. Si la hipòtesi nul·la és certa, F seria proper a 1. D'altra banda, si la mitjana quadràtica entre grups MSG és gran, suposaria un valor gran de l'estadístic F. Bàsicament, l'ANOVA examina les dues fons de la variància total i mira quina part contribueix més. Per aquest motiu, s'anomena anàlisi de la variància encara que la intenció sigui comparar les mitjanes dels grups.

Hi ha una sèrie de supòsits que s'han de fer abans que s'apliqui l'ANOVA, la desviació en aquests supòsits portaran a resultats que poden ser enganyosos o inexactes. Aquests supòsits inclouen la independència, normalitat i variància constant dels errors. En algunes situacions, hi ha transformacions que poden ser utilitzades per evitar les violacions d'aquests supòsits, com ara la transformació logarítmica de les dades.

2.1.1 Anova per a dissenys desbalancejats

2.2 Correcció per multiplicitat de contrastos

Un problema comú què ens podem trobar a qualsevol investigació és voler comparar més de 2 grups de dades per detectar possibles diferències entre ells. La utilització de models d'ANOVA ens pot permetre detectar diferències, a escala global, entre les mitjanes involucrades, però en moltes ocasions volem detectar les diferències entre grups concrets. Aquest cas només és possible mitjançant l'ús dels Procediments de Comparacions múltiples (PCM).

En aquest treball el nostre interès no és avaluar si un o dos gens concrets s'expressen d'una forma diferencial entre les condicions considerades. Volem veure això a un nivell global i respondre a una pregunta com: Quins gens s'expressen d'una manera diferent (diferencial si utilitzem la literatura biològica) en els grups/tractaments que considerem? L'objectiu és poder contestar aquesta pregunta de manera que puguem controlar les vegades que afirmem expressions diferencials quan realment no la tenen (Error de tipus I).

Si numerem els gens $i = 1, ..., N$ llavors per a l'i-èssim gen estem considerant el següent contrast:

- H_0 : El gen i no té una expressió diferencial entre les condicions considerades.
- H_1: El gen i té una expressió diferencial entre les condicions considerades.

Si plantegem aquest contrast per a cada gen, podem denotar com $G = 1, .., N$ el conjunt d'hipòtesis nul·les que estem avaluant. El número d'hipòtesis que avaluem és conegut a priori, ja que correspon al número de gens que volem avaluar. És important destacar que en els estudis d'investigació en intentar acceptar o rebutjar la hipòtesi nul·la (H_0) es poden cometre dos tipus d'errors:

- Error de tipus I: Rebutjar H_0 quan realment és certa.
- Error de tipus II: No rebutjar H_0 quan realment és falsa.

Imaginem que fem un test contrastant diferencies entre mitjanes, i fixem un nivell de significació $\alpha = 0.05$, i sabem que la hipòtesis nul·la és certa; llavors l'error de tipus I serà exactament el nivell de significació α. Per tant, podem definir la probabilitat de tenir un fals positiu en un test, és a dir, rebutjar H_0 quan realment és certa:

$$P(\text{Fals positiu}) = \alpha$$

$$P(\text{No cometre l'error}) = 1 - \alpha$$

Per tant, si definim m tests d'hipòtesis podem definir la probabilitat d'almenys tenir 1 fals positiu com:

$$P(\text{No cometre l'error en m tests}) = (1 - \alpha)^m$$

$$P(\text{Almenys 1 fals positiu en m tests}) = 1 - (1 - \alpha)^m$$

Per exemple, si tenim 1 test, i fixem $\alpha = 0.05$, la probabilitat d'obtenir almenys 1 fals positiu és de:

$$P(\text{Almenys 1 fals positiu}) = 1 - (1 - 0.05) = 0.05$$

Si ara tenim 50 tests i calculem la mateixa probabilitat:

$$P(\text{Almenys 1 fals positiu}) = 1 - (1 - 0.05)^{50} = 0.92$$

Aquí recau el gran problema de la multiplicitat de contrastos, podem observar que si fem un test moltes vegades, hi ha una inflació en l'error de tipus I.

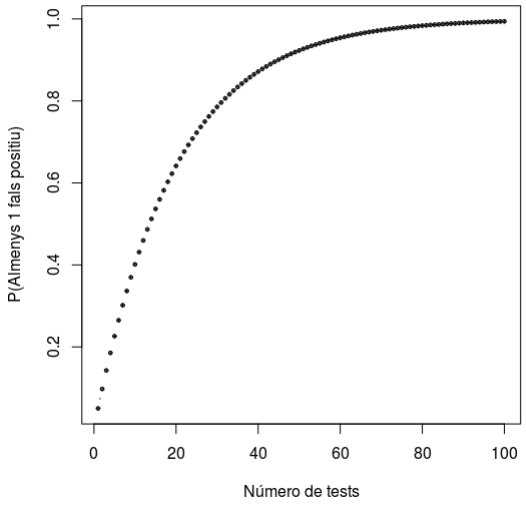

Figura 1: Gràfic contraposant el nombre de tests amb la probabilitat d'obtenir almenys 1 fals positiu. S'observa un increment en la probabilitat quan augmenta el nombre de tests.

En el nostre estudi es important tenir clar aquest problema, ja que si tenim molts gens, i no apliquem una correcció, podem caure en l'error d'afirmar que un gen s'expressa diferencialment quan realment no ho fa.

2.2.1 False discovery Rate

Existeixen molts mètodes per corregir el problema de la multiplicitat de contrastos. El més simple és el métode de Bonferroni, on cada p valor es multiplica pel número de tests realitzats (acotant la probabilitat màxima a 1). És un métode molt conservador i no és el més indicat per al nostre cas d'estudi. Llavors la correcció que hem utilitzat per controlar l'error de tipus I, s'anomena False discovery Rate (FDR).

Los métodos descritos anteriormente se centran en corregir la inflación del error de tipo I (false positive rate), es decir, la probabilidad de rechazar la hipótesis nula siendo esta cierta. Esta aproximación es útil cuando se emplea un número limitado de comparaciones. Para escenarios de large-scale multiple testing como los estudios genómicos, en los que se realizan miles de test de forma simultanea, el resultado de estos métodos es demasiado conservativo e impide que se detecten diferencias reales. Una alternativa es controlar el false discovery rate. El false discovery rate (FDR) se define

como: (todas las definiciones son equivalentes) La proporción esperada de test en los que la hipótesis nula es cierta, de entre todos los test que se han considerado significativos. FDR es la probabilidad de que una hipótesis nula sea cierta habiendo sido rechazada por el test estadístico. De entre todos los test considerados significativos, el FDR es la proporción esperada de esos test para los que la hipótesis nula es verdadera. Es la proporción de test significativos que realmente no lo son.La proporción esperada de falsos positivos de entre todos los test considerados como significativos.

El objetivo de controlar el false discovery rate es establecer un límite de significancia para un conjunto de test tal que, de entre todos los test considerados como significativos, la proporción de hipótesis nulas verdaderas (falsos positivos) no supere un determinado valor. Otra ventaja añadida es su fácil interpretación, por ejemplo, si un estudio publica resultados estadísticamente significativos para un FDR del 10, el lector tiene la seguridad de que, como máximo, un 10 de los resultados considerados como significativos son realmente falsos positivos.

Cuando un investigador emplea un nivel de significancia α, por ejemplo de 0.05, suele esperar cierta seguridad de que solo una pequeña fracción de los test significativos se correspondan con hipótesis nulas verdaderas (falsos positivos). Sin embargo, esto no tiene por qué ser así. La razón por la que un false positive rate bajo no tiene por qué traducirse en una probabilidad baja de hipótesis nulas verdaderas entre los test significativos (false discovery rate) se debe a que esta última depende de la frecuencia con la que la hipótesis nula contrastada es realmente verdadera. Un caso extremo sería el planteado en el ejemplo 1, en el que todas las hipótesis nulas son realmente ciertas y por lo tanto el 100 de los test que resultan significativos son falsos positivos. Así pues, la proporción de falsos positivos (false discovery rate) depende de la cantidad de hipótesis nulas que sean ciertas de entre todas los contrastes.

Los análisis de tipo exploratorio en los que el investigador trata de identificar resultados significativos sin apenas conocimiento previo se caracterizan por una proporción alta de hipótesis nulas falsas. Los análisis que se hacen para confirmar hipótesis, en los que el diseño se ha orientado en base a un conocimiento previo, suelen tener una proporción de hipótesis nulas verdaderas alta. Idealmente, si se conociera de antemano la proporción de hipótesis nulas verdaderas de entre todos los contrastes se podría ajustar con precisión el límite significancia adecuado a cada escenario, sin embargo, esto no ocurre en la realidad.

La primera aproximación para controlar el FDR fue descrita por Benjamini y Hochberg en 1995. Acorde a su publicación, si se desea controlar que en un estudio con n comparaciones el FDR no supere un porcentaje d hay que:

El método propuesto por Benjamini & Hochberg asume a la hora de estimar el número de hipótesis nulas erróneamente consideradas falsas, que todas las hipótesis nulas son ciertas. Como consecuencia, la estimación del FDR está inflada y por lo tanto es conservadora. A continuación se describen métodos más sofisticados que estiman la frecuencia de hipótesis nulas verdaderas a partir de la distribución de los p-values.

An alternative method to create web-based teaching tool applications is provided by Shiny (Chang, Cheng, Allaire, Xie & McPherson, 2015), a recent technology created by RStudio. Shiny is a web application framework for R (R Core Team, 2015) that only requires knowledge in the R programming language. It is not uncommon for instructors to build their own teaching tools via scripts written in R and, as we will show, it is not difficult to convert existing R scripts into Shiny applications, known simply as 'Shiny apps'. With Shiny, instructors can build a teaching tool that is interactive, dynamic, user-friendly, visually appealing, and, with similar functionality to Java/Javascript applets; the only requirement is some familiarity in R.