Análisis funcional de grupos de genes.

Alex Sánchez

6 de febrero de 2017

Índice

1. Introducción

1.1. Análisis de la significación biológica: Expresión diferencial de conjuntos de genes

El análisis de significación biológica de las listas de genes diferencialmente expresados encontrados en la primera fase de este análisis se ha llevado a cabo utilizando el método conocido como *Análisis de sobrerrepresentación ("ORA") o Análisis de Enriquecimiento de Genes ("GEA")* . Esta técnica introducida por [3] analiza si, dado un grupo de genes, representados, por ejemplo, por una categoría de la Gene Ontology, el porcentaje que representa en una lista de genes seleccionados (por ejemplo por estar diferencialmente expresados), es el que correspondería a su tamaño respecto del total de genes analizados o bien es significativamente superior o inferior. En el primer caso suele decirse que la categoría está sobre-representada en la lista o que la lista está enriquecida en esta categoría, de ahí el nombre del método.

El análisis de enriquecimiento ha recibido múltiples críticas:

- Se ha argumentado que tiene poca sensibilidad para identificar los grupos de genes pobremente representados en las listas (por su tamaño no por su significación estadística).
- El hecho de basarse en una lista de genes que ha sido truncada por un punto de corte más o menos arbitrario se ha discutido a menudo puesto que deja a muchos posibles genes interesantes fuera del análisis, por no decir que los que quedan dentro de la lista no reciben ponderación alguna cuando probablemente debería tener más relevancia un gen con un "'fold-change" de 4 y un p-valor de 0.0001 que uno cuyos valores sean 1.3 y 0.002 respectivamente.
- El punto de corte para definir la lista es una de las principales preocupaciones puesto que hay estudios que incluso argumentan que combios en el mismo pueden llevar a cambios en las conclusiones biológicas del estudio. [5]
- Finalmente el análisis de enriquecimiento asume que los genes son independientes entre ellos, lo que es a todas luces una simplificación.

A la vista de las críticas anteriores Mootha et alt. [6] introdujeron el "Gene Set Enrichment Analysis" o GSEA como una alternativa al análisis de enriquecimiento que permitiera soslayar algunos de los problemas descritos. La idea venía a ser que, aunque a veces un gen dado pudiera no estar diferencialmente expresado, un grupo de genes relacionado con una característica médica (enfermedad) o biológica (proceso) podría tender a mostrar más diferencias entre los grupos en estudio que el resto de los genes. Es decir, aunque gen a gen no hubiera expresión diferencial los autores proponian agregar los valores individuales de asociación con la característica en estudio dentro del grupo para decidir si se podía considerar que el grupo en sí estaba relacionado con dicha característica en estudio.

El método de GSEA original se basa en comparar, para cada grupo de genes, la distribución de los estadísticos de test dentro del grupo con la distribución global de dichos estadísticos, es decir la calculada para todos los genes. Para ello utiliza el test de Kolmogorov-Smirnov (K-S test) para calcular una puntuación de enriquecimiento a la que posteriormente asigna un p-valor mediante un cálculo bastante laborioso. El test de K-S es conocido por su poca sensibilidad para detectar diferencias sutiles, lo que, junto con lo costoso del cálculo de los p-valores ha llevado a considerar múltiples alternativas al mismo [4].

En este trabajo aplicaremos diversas aproximaciones que se comentan brevemente a continuación. Éstas oscilan desde métodos que se consideran óptimos por su potencia y flexibilidad como el GSA introducido por Efron y Tibshirani [1] a métodos más sencillos pero también más intuitivos como el propuesto por Irizarry [2].

1.2. Algunos métodos de *Gene Set Enrichment Analysis*

Hay muchos métodos para hacer este tipo de análisis. Lo que es más, en realidad hacen tipos sutilmente distintos de análisis. Por ejemplo se suele diferenciar entre *pruebas competitivas* y pruebas *auto-contenidas*.

- En las pruebas competitivas el objetivo es comparar la expresión diferencial entre el grupo de genes y el resto, es decir ver si los genes del grupo estan más o menos diferencialmente expresados que los que no forman parte de éste.
- En las pruebas auto-contenidas el objetivo es determinar si los genes del grupo estan diferencialmente expresados sin comparar esta expresión diferencial global con la de otros grupos.

Entre estos dos grandes grupos de pruebas se encuentran multitud de variantes de cada una de ellas.

Para acabar de complicar su uso muchos de estos métodos utilizan razonamientos o estadísticos de test relativamente complicados (aunque otros hacen todo lo contrario argumentando que no es necesaria tanta complekidad), lo que, globalmente ha determinado que este tipo de métodos haya sido menos utilizado por la comunidad científica.

En este trabajo se utilizaran algunos métodos ampliamente aceptados como el GSA [1] o métodos de los dos tipos, competitivos y auto-contenidos. Para facilitar la interpretación de los resultados, se utilizaran también algunos métodos más simples como el test de Wilcoxon o la comparación de densidades propuesta por el grupo de R. Irizarry de la Universidad John Hopkins.

1.2.1. GSEA simple 1: Test de Wilcoxon sobre los estadísticos de test

El test de Wilcoxon es una conocida prueba no paramétrica de comparación entre grupos. Su aplicación en GSA consiste en utilizarlo para comparar los valores de los estadístico de test

utilizados para seleccionar genes diferencialmente expresados. Es decir se realiza un test de wilcoxon entre los valores de test dentro del grupo de genes y fuera del grupo. La significación del test puede estimarse mediante una aproximación normal o utilizando un enfoque de permutaciones. La prueba de suma de rangos de Wilcoxon se utiliza para hacer GSA en el popular paquete `limma` de R, aunque aquí se ha implementado manualmente.

1.2.2. GSEA simple 2: Visualización del desplazamiento de los valores de test

Una alternativa al test de Wilcoxon consiste simplemente en comparar las medias de los valores de test en cada grupo, para ver hasta que punto la media de los tests-t dentro del grupo de genes se desplaza hacia la izquierda o la derecha de la media general (que en principio debería estar entorno a cero). Esta comparación puede hacerse mediante un test de permutaciones o visualmente comparando ambas distribuciones, que es el método que se aplicará en este caso.

1.2.3. GSA: *Gene Set Analysis* basado en la maxima diferencia mínima

Los dos métodos anteriores suponen independencia entre las expresiones de los genes pero algunos trabajos han argumentado que esto es una limitación importante. El método de GSA realiza un test competitivo que tiene en cuenta las posibles correlaciones entre genes.

Este enfoque separa los valores de test positivos y negativos obtenidos para los genes de un determinado grupo de genes y, para cada uno de los dos subconjuntos, calcula la suma absoluta dividida por el número total de genes en el conjunto. El estadístico de test para el grupo de genes se define como el máximo de estos dos números ("maxmean"). La significación, como para todos los métodos anteriores, se puede calcular mediante un enfoque de permutación.

1.2.4. ROAST : Gene Set Analysis basado en la rotación

El método ROAST, es similar al GSEA pero utiliza un método conocido como rotación multivariante para calcular un estadístico que permita decidir si un grupo de genes puede considerarse sobre expresado o "down"-regulado. Dos características de este método lo hacen atractivo. Por un lado está basado en la estimación de un modelo lineal (del tipo ANOVA) por lo que puede aplicarse a muchos diseños distintos. Por otro el sistema utilizado para calcular los p-valores -distinto a las permutaciones que utiliza el GSA o el GSEA- permite que sea utilizado en casos en los que el número de muestras es pequeño. AL igual que el GSA el método tiene encuenta, implícitamente, la relación entre los genes, es decir que no son independientes.

1.3. Aplicación

En general la aplicación de cualquiera de los métodos necesita, para cada comparación entre dos o más condiciones experimentales:

- Una matriz de expresión, con una fila por gen, y una columna por muestra.
- Un vector de condiciones experimentales.

- Un vector con los identificadores de los genes.
- Una lista de grupos de genes ("genesets").

Además, necesitamos disponer de la lista de los nombres de los genes en el mismo tipo de símbolo qen que se han codificado los "geneSets". En este caso se ha utilizado el identificador "Entrez" y la lista de genes será la misma para todas las matrices de expresión analizadas.

A partir de dicha información el programa genera una tabla que contiene los "GeneSets" que son declarados diferencialmente expresados (sobre o sub) para un nivel de significación (FDR) dado.

Referencias

[1] Robert Tibshirani Bradley Efron. On testing the significance of sets of genes. 1:107–129.

[2] Rafael A. Irizarry, Chi Wang, Yun Zhou, and Terence P. Speed. Gene set enrichment analysis made simple. *Statistical Methods in Medical Research*, 18(6):565–575, December 2009.

[3] P. Khatri and S. Drăghici. Ontological analysis of gene expression data: current tools, limitations, and problems. *Bioinformatics*, 18:3587–3595, 2005.

[4] Henryk Maciejewski. Gene set analysis methods: statistical models and methodological differences. *Briefings in Bioinformatics*, 15(4):504–518, July 2014.

[5] Kuang-Hung Pan, Chih-Jian Lih, and Stanley N. Cohen. Effects of threshold choice on biological conclusions reached during analysis of gene expression by DNA microarrays. 102(25):8961–8965.

[6] Karl-Fredrik Eriksson Vamsi K Mootha, Cecilia M Lindgren. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. 34:267 – 273.