Tarea 2 Bioinformática - Expresión génica

Manuel Villalobos Cid

28 de octubre de 2015

1. Introducción.

Incluso células que poseen el mismo ADN son diferentes entre sí debido a que presentan distinta expresión genónica. Este proceso es conocido como el **Dogma Central de la Biología Molecular**: un gen expresa su información por transcripción de ADN en ARN mensajero y es traducido a una proteína.

El estudio de expresión génica y análisis de mutaciones permite a los científicos comprender los mecanismos fisiopatológicos de las enfermedades para así determinar sus tratamientos. Esto se logra estableciendo diferencias de expresión entre muestras normales y patológicas. También ha sido utilizado para estudiar características poblacionales o determinar el efecto de medicamentos en células blanco.

Los microarreglos son una de las técnicas mas utilizadas para establecer perfiles de transcripción. Existen múltiples desarrollos comerciales y académicos que presentan características diversas en cobertura, disponibilidad, especificidad y sensibilidad. Los arreglos fabricados por *Affymetrix* son los mayormente utilizados. Sin embargo, también se emplean los construidos por *Agilent* o los de uso académico *CATMA*. Una de las principales fuentes de datos de microarreglos es ***Gene Expression Omnibus* de *NCBI* (GEO)**.

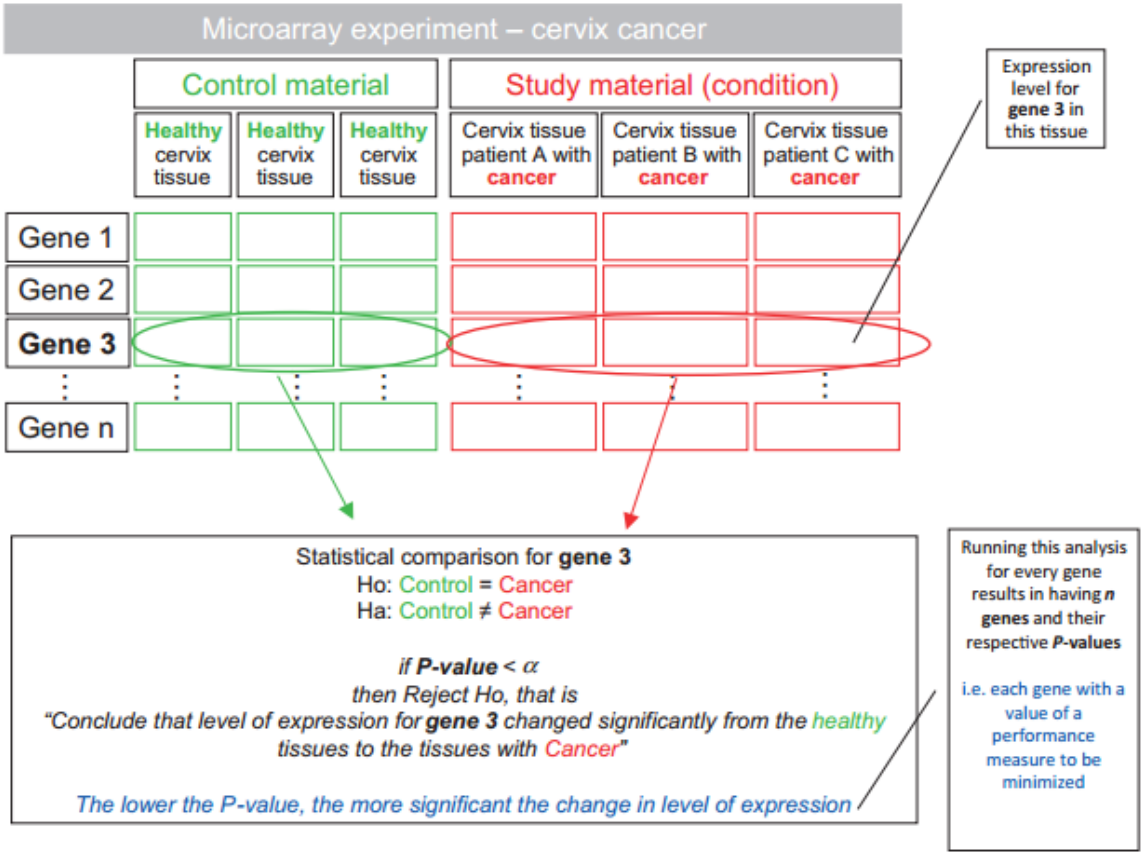

Figure 1.- Esquema de experimento para análisis de expresión génica con microarreglos.

Un juego de datos se caracteriza por tener una alta dimensionalidad: alrededor de 10.000 genes con más de 10 experimentos, además puede presentar ruido o deficiencias provenientes del proceso de hibridación. A raíz de este problema han surgido otras técnicas, como el Análisis en Serie de Expresión de Genes (Serial Analysis of Gene Expression, *SAGE*) o Firma Paralela de Secuenciación Masiva (Massively Parallel Signature Sequencing, *MPSS*). Ambas han demostrado ser mas robustas que el uso de microarreglos, ya que sus resultados no dependen de la selección de la sonda inicial. Recientemente se ha desarrollado una técnica denominada *RNA-seq*, cuyos resultados tampoco dependen de selección de sonda y no posee sesgo producido durante el proceso de hibridación. Sin embargo, aún no existe un conjunto de herramientas definitivas para procesar estos datos y el proceso de preparación de muestras continúa siendo bastante lento.

Para encontrar medidas de similaridad en expresión de genes se utiliza el concepto de distancia. Las más utilizadas corresponden a la Euclidiana y la correlación de *Pearson*. Sobre estas medidas se aplica minería de datos con métodos de clasificación y agrupamiento. Algoritmos basados en *UPGMA*, *NJ*, Mapas Auto-organizados (Self Organizing MAPS, *SOM*), lógica difusa, *SVM* y redes neuronales han sido ampliamente utilizados. Los algoritmos se pueden dividir en métodos planos, jerárquicos, basados en grafos, metaheurísticos y orientados a optimización. Éstos últimos pueden combinar varios de los grupos anteriores.

Esta guía de laboratorio tiene como objetivo que los alumnos se familiaricen con los conceptos relacionados con expresión génica, efectuando una pequeña experiencia en R, usando librerías disponibles por Bioconductor.

2. Actividades prácticas

2.1 Instrucciones

Las actividades prácticas deberán ser efectuadas por cada uno de los alumnos. En grupos de máximo tres estudiantes, desarrollarán un informe siguiendo la metodología de la experiencia anterior, utilizando el formato de **dos columnas**. El plazo máximo de entrega será el día martes 10 de noviembre a las 23:55 horas. El informe deberá ser cargado en formato pdf en la sección específica de Moodle destinada para el Laboratorio de Bioinformática.

2.2 Gene Expression Omnibus: búsqueda, descarga y análisis de conjunto de datos

- Acceda a la base de datos de ***Gene Expression Omnibus* de *NCBI* (GEO)** y descargue el conjunto denominado **GPL1426**. Este conjunto de datos es optativo, ya que pueden usar otros de su interés.
- Resuma brevemente el objetivo del estudio para el cuál se extrajeron los datos.
- Describa y caracterice el conjunto. Comente si es necesario efectuar pre-procesamiento.
- Utilizando R y la librería **Bioconductor**, identifique los genes que permitan diferenciar entre clases: control y no control. Demuestre estadísticamente y gráficamente la expresión diferencial de estos genes. (Incluya métodos de agrupamiento).
- Compare sus resultados con los presentados por las herramientas **Data Analysis Tools** de **GEO**.
- Escriba todos los pasos asociados, algoritmos y parámetros utilizados en su informe.
- Establezca conclusiones e incluya su código como anexo.