Memoria del treball final de màster

Vasyl Druchkiv

Estudiant del Màster de Bioestadística i Bioinformàtica

15 d'Abril 2019

Índice

---

1 Introducció

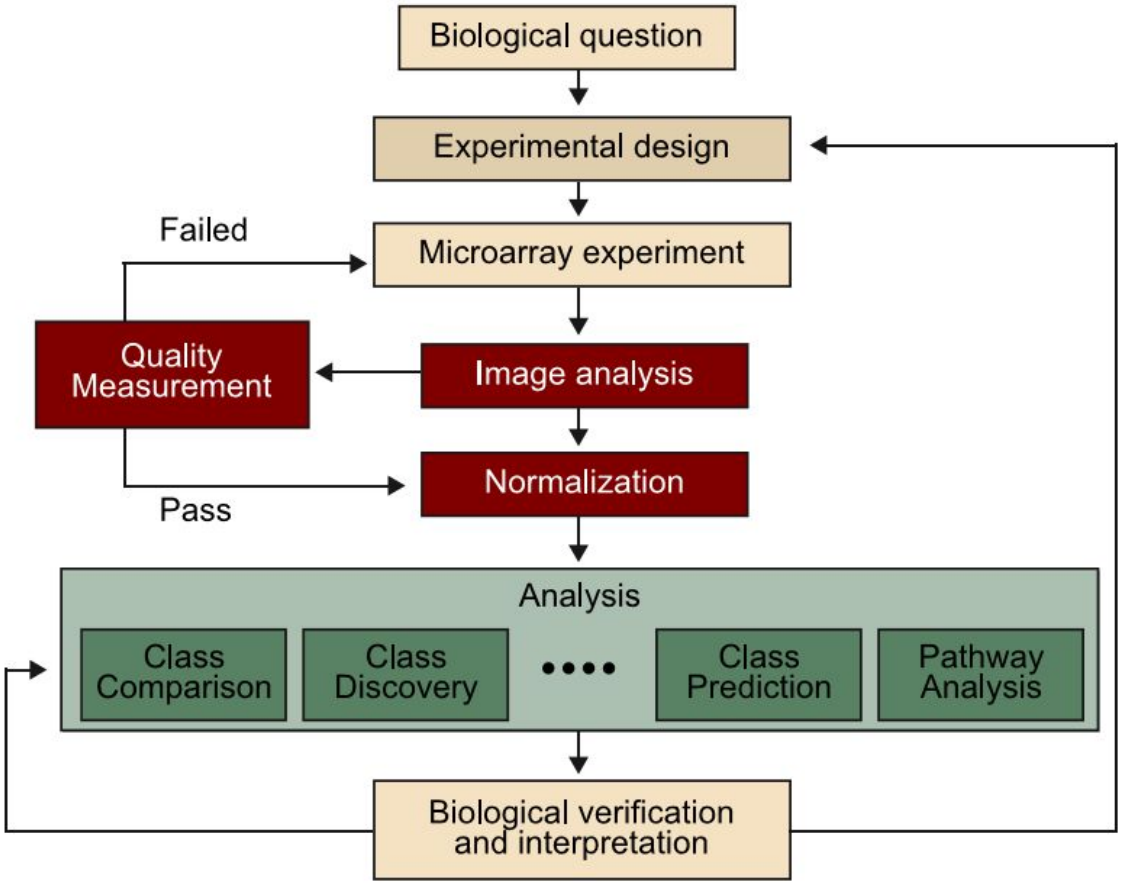

Figure 1: El procès d'anàlisi de microarrays

2 Descripció dels mètodes

3 Anàlisi de les rutes - final del pipeline d'anàlisi d'expressió

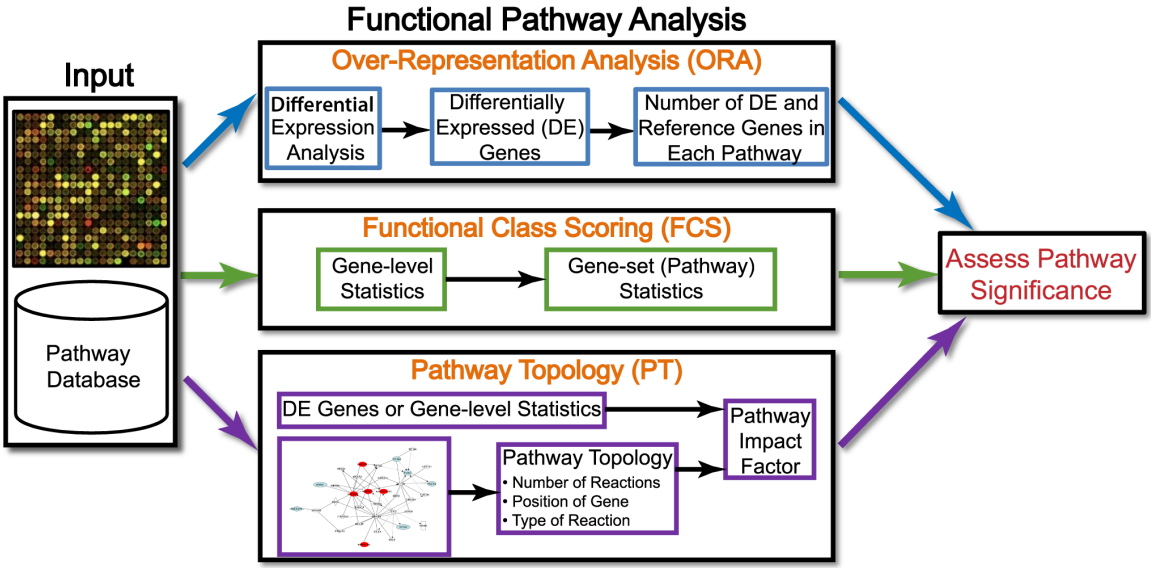

Figure 2: El procès d'anàlisi de les rutes

3.1 ORA

L'anàlisi de sobreexpressió és una tècnica d'identificació de les rutes significativament enriquides en la mostra d'interès.

El paper original què se cita habitualment quan es parla d'anàlisi d'expressió genètica és de [Boyle et al., 2004]. El mètode estadístic descrit consisteix bàsicament en els pasos següents:

1. **De tots els gens de la mostra seleccionar un grup de gens que es considera que són significativament expressats.**

   Els criteris de selecció poden baser-se en *log ratios* o/i en el valor de p provenent d'un test estadístic. *Log ratios* donen la magnitut amb el qual un gen és sobre o sotaexpressats. Les diferències entre els grups però són el resultat d'un procès estochàstic i per tan hem d'intentar de minimitzar el risc de prendre decisions falses. El valor de p representa la probabilitat d'aquest risc i per tant dona certa confiança sobre la significació de les diferències observades.

2. **Determinar si algunes rutes anoten la llista especificada de gens amb la freqüència més alta que un esperaria per casualitat.**

   El test estadístic es basa en la distribució hipergeomètrica:

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{n}}$$

En aquesta equació N és el nombre total de gens en la distribució de fons, M és el nombre de gens dins d'aquesta distribució que són anotats a la ruta d'interès, n és el nombre total en la llista especificada de gens i k és el nombre de gens dins d'aquesta llista que són anotats a la ruta. La distribució de fons pot ser o bé tots els gens en la base de dades d'anotació o bé tots els gens d'experiment.

El valor de P obtingut amb aquesta formula dona la probabilitat de veure el nombre x de gens de la llista relacionats amb la ruta específica en la llista del nombre total de gens n donat la proporció de gens relacionats amb aquesta ruta en la distibucó de fons.

L'aplicació utilitza aquesta idea i calcula una taula amb els camps següents:

- Description. El nom del terme GO;
- GeneRatio. El quocient: $\frac{\text{Nombre dels gens diferencialment expressats que pertanyen al conjunt de gens}}{\text{Nombre total dels gens diferencialment expressats}} = \frac{M}{N}$;
- BgRatio. El quocient: $\frac{\text{Nombre dels gens del conjunt d'interès en la distribució de fons}}{\text{Nombre total dels gens en la distribució de fons}} = \frac{k}{n}$;
- pvalue. Valor de p basat en la distribució hipergeomètrica descrita anteriorment.
- p.adjust. El valor de P ajustat. L'usuari pot seleccionar el mètode d'ajustmanent.

3.2 GSEA

Amb l'anàlisi GSEA podem analitzar els resultats d'un experiment d'expressió per a dos grups. Aquí els gens son ordenats basant-se en la correlació entre la seva expressió i la separació entre les classes. Aquest llistat ordenat L el podem crear utilitzant els *logRatios*.

Donat el conjunt definit dels gens S, que pertanyen per exemple al mateix terme de Gene Ontology, l'objectiu de GSEA és determinar si els membres de S són distribuits aleatoriament en el L o es troben més al cap o a la cua. S'esperaria que els gens relacionats a la separació fenotípica mostraran aquesta última distribució.

L'anàlisi GSEA consisteix en tres pasos:

1. Calculació del puntuació d'enriquement (*ES: Enrichment Score*). La puntuació està calculada anant per la llista i augmentant la suma corrent sempre quan es troba un gen que pertany a S o, al contrari, restant-hi quan el gen no forma part del conjunt S. La puntuació és la desviació màxima del zero observada en aquet camí. L'estadística obtinguda és la estadística de Kolmogorov-Smirnov amb pesos.
2. Estimació del nivell de significació per a la puntuació *ES*. El valor de P nominal es pot obtenir mitjançant o bé la permutació de les classes o bé la permutació de gens, on l'estadística *ES* observada es compara amb la distribució obtinguda amb permutació. A l'aplicació es fa ús de l'útima opció.
3. Calculació del valor de P ajustat. El valor de P nominal s'ajusta per controlar l'error global que es produeix com a resultat de les comparacions múltiples.

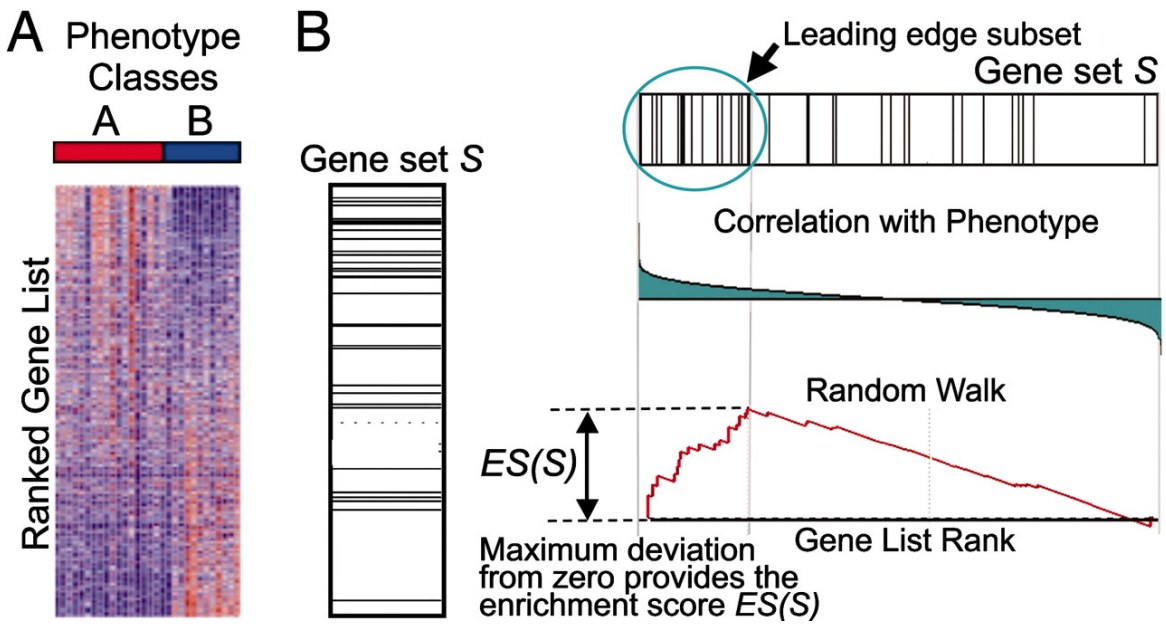

Figure 3: El mètode GSEA

L'aplicació que he desarrolat agafa aquesta idea i calcula la tabla que inclou l'estadístiques següents:

- enrichmentScore. Enrichment score per al conjunt dels gens. Amb altres paraules: el grau amb el qual el conjunt dels gens està sobreexpressat a dalt o a baix del llistat ordenat dels gens en les dades d'expressió.
- NES. Normalized enrichment score. La puntuació per al conjunt de gens després de ser normalitzat tenint en compte tots els conjunts de gens analitzats (la seva mida i la seva correlació amb les dades d'expressió). Aquesta puntuació ajuda a comparar els resultats entre els conjunts de gens.
- pvalue.El valor de p nominal.
- p.adjust. El valor de p ajustat.
- leading_edge
  - Tags. El percentatge de les ocurrències de gens del conjunt específic abans (per als ES positius) o després (per als ES negatius) del cim en la puntuació corrent d'enriquiment. Aquest valor indica el percentatge dels gens que contribueixen a la puntuació d'enriquement.
  - List. El percentatge dels gens en el llistat ordenat de tots els gens abans o després del pic en la puntuació corrent d'enriquiment. Aquest valor ens indica on exactament el pic es produeix.
  - Signal. La fortalesa del senyal d'enriquiment que combina els dos valors anteriors.
- rank. La posició del pic en la llista ordenada dels gens. Els conjunts dels gens més interessants assoleixen el seu màxim o bé al principi o al final de la llista ordenada. Vol dir que tenen aquest valor o bé molt baix o bé molt alt.

4 Instalació

Es pot installar l'aplicació localment utilitzan l'ordre següent de R:

```
devtools::install_github("vdruchkiv/TFM/5_Packages/PathwayApp/PathwayApp")
```

Biblilografia

[Boyle et al., 2004] Boyle, E. I., Weng, S., Gollub, J., Jin, H., Botstein, D., Cherry, J. M., and Sherlock, G. (2004). Go:: Termfinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. *Bioinformatics*, 20(18):3710–3715.