Memoria del treball final de màster

Vasyl Druchkiv

Estudiant del Màster de Bioestadística i Bioinformàtica

15 d'Abril 2019

Índice

---

1 Introducció

2 Descripció dels mètodes

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

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.

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.