PAC0 Definició dels continguts del treball

Vasyl Druchkiv
Estudiant de Màster de Bioestadística i Bioinformàtica

04 de Març 2019

Paraules clau: **expressió gènica; Pathway analisis; R; Bioconductor; ORA; GSEA; ReactomePA; GAGE; Shiny**.

Temàtica de treball

El tema del TFM és la **"Implementació d'una eina en R/Shiny per a l'anàlisi de significació biològica utilitzant l'anàlisi de *Pathways*"**. Es tractarà doncs d'examinar grups de gens (Gene sets) en comptes de només gens individuals. Aquest és l'últim pas del *pipeline* de l'anàlisi de dades d'expressió però, des del meu punt de vista, el més important. Un inconvenient gran de l'anàlisi dels gens individuals és la manca de sentit biològic. Els mètodes actuals com ara microarrays o NGS proporcionen uns llistats extensos de gens diferencialment expressats i és important reduir la complexitat d'una manera intel·ligent agrupant els gens implicats al mateix mecanisme d'acció – el Pathway. Hi ha ja eines bioinformàtiques per dur a terme aquesta tasca. Algunes no necessiten coneixements avançats de programació o estadística perquè ofereixen una interfície fàcil d'usar. Un exemple seria Ingenuity Pathways que és una eina de pagament i de codi tancat. Hi ha altres eines però que d'una banda ofereixen més flexibilitat i, a més a més, són gratuïtes; però d'altra banda són difícils d'usar pels investigadors sense coneixements bioinformàtics/estadístics. La més prometedora d'aquestes eines és el programa R amb Bioconductor - el projecte específic per a bioestadística.

Un *Pathway* és el conjunt de gens relacionats amb una funció biològica i descriu la relació entre els gens. Alguns exemples d'aquestes rutes són les metabòliques, les de traducció de senyal o les de regulació gènica. La necessitat de mirar les rutes biològiques es basa en la idea que els gens produeixen proteïnes, però les proteïnes, encara que individualment fan una tasca específica, actuen en conjunt amb altres proteïnes per assolir un objectiu superior.

L'anàlisi de Pathways no és un concepte nou sinó que ja té una història [Khatri et al., 2012]. La primera generació de l'anàlisi de les rutes és ORA (Over-Representation Analysis), que consisteix en la selecció dels gens diferencialment expressats i basant-se en les agrupacions de gens com ara GO (Gene Ontology) prova estadísticament si una d'aquestes agrupacions està inusualment sobre o sotaexpressada en la mostra. Aquest mètode però té les seves limitacions; en parlaré a l'apartat biològic del treball. Aquestes limitacions van

provocar el desenvolupament de nous mètodes d'anàlisi dels grups de gens, els quals denominem d'acord amb [Khatri et al., 2012] aproximacions FCS (Functional Class Scoring). La idea aquí és que encara que uns grans canvis en l'expressió d'un gen individual poden tenir uns efectes significatius en les rutes biològiques, els canvis més petits però coordinats d'un grup de gens també els pot tenir. Per aquest motiu els mètodes FCS agreguen les estadístiques diferencials en una estadística única al nivell de la ruta.

Hi ha bases de dades que inclouen col·lecciones de rutes i xarxes d'interacció. Algunes d'aquestes bases de dades són KEGG (Kyoto Encyclopedia of Genes and Genomes) [Kanehisa et al., 2004] o Reactome [Croft et al., 2010]. Tant una com l'altra s'utilitzen en els paquets del Bioconductor en R. El primer s'utilitza al paquet GAGE i el segon al paquet ReactomPA. El paquet GAGE té l'avantatge que ofereix estadístiques diverses en funció del disseny experimental i/o mostres de la mida petita [Luo et al., 2009]. Pressento però un possible problema quant a la publicació d'una aplicació basada en GAGE a shinyapps.io, perquè actualment el paquet no permet guardar les imatges de pathways en format d'una imatge .png. El paquet ReactomePA [Yu and He, 2016] permet fer l'anàlisi basant-se tant en ORA com en GSEA (vegeu [Subramanian et al., 2005]. A més a més les imatges de ReactomePA es poden exportar fàcilment.

Problemàtica a resoldre

Amb els paquets mencionats es pot desenvolupar una aplicació amb Shiny que tindrà una interfície fàcil d'usar per les persones sense coneixements bioinformàtics. Es aquí on preveig un repte per a mi perquè encara no tinc coneixements clars sobre el procés del treball amb shiny. A més a més preveig uns problemes en la fase d'implementació de l'aplicació. Aquí tindria dues possibilitats: 1. shinyapps.io i 2. servidor shiny. De fet ja he començat a treballar en el desenvolupament de l'aplicació i n'he creat una petita part que fa referència al gàfic de barres amb les categories de Reactome utilitzant el paquet ReactomePA. He intentat pujar l'aplicació a shinyapps.io i per desgràcia no he tingut èxit. El problema però no és de l'aplicació, que localment funciona bé, sinó de shinyapps.io que no aconsegueix instal·lar el paquet reactome.db. Es veu que el problema és la mida del paquet, que actualment pesa 2.5 gb cosa que causa problemes i produeix un error de *timeout* (vegeu la discussió aquí .

Objectius

L'objectiu principal del treball és la creació d'una aplicació Shiny que permet dur a terme l'anàlisi de Pathway. Per assolir aquest objectiu és necessari fer els passos següents:

- Fase 1. Continguts de l'aplicació
    1. Determinar que és l'anàlisi de Pathways. Per que serveix i què ha de contenir.
    2. Cerca dels paquets de Bioconductor que fan l'anàlisi de Pathways.
    3. Selecció d'un paquet que servirà com a base de l'aplicació.

4. Decisió sobre el format de les dades que l'usuari haurà de pujar a l'aplicació.
5. Decisió sobre l'output de l'aplicació.

- Fase 2. Desenvolupament de l'aplicació Shiny.
  1. Selecció de les dades exemple per dur a terme l'anàlisi sense interfície shiny per comprovar el funcionamet del paquet seleccionat.
  2. Desenvolupament de l'aplicació.
  3. Repetició de l'anàlisi fet en 1 utilitzant l'aplicació.
  4. Selecció de les dades noves per comprovar el funcionament de l'aplicació.
  5. Implementació de l'aplicació: shinyapps.io o servidor?
- Fase 3. Presentació
  1. Finalització de la memòria.
  2. Defensa del treball.

References

[Croft et al., 2010] Croft, D., O'kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G., Jassal, B., et al. (2010). Reactome: a database of reactions, pathways and biological processes. *Nucleic acids research*, 39(suppl_1):D691–D697.

[Kanehisa et al., 2004] Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The kegg resource for deciphering the genome. *Nucleic acids research*, 32(suppl_1):D277–D280.

[Khatri et al., 2012] Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS computational biology*, 8(2):e1002375.

[Luo et al., 2009] Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D., and Woolf, P. J. (2009). Gage: generally applicable gene set enrichment for pathway analysis. *BMC bioinformatics*, 10(1):161.

[Subramanian et al., 2005] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550.

[Yu and He, 2016] Yu, G. and He, Q.-Y. (2016). Reactomepa: an r/bioconductor package for reactome pathway analysis and visualization. *Molecular BioSystems*, 12(2):477–479.