PAC1 Plà de treball

Vasyl Druchkiv

Estudiant de Màster de Bioestadística i Bioinformàtica

18 de Març 2019

Índice

Context i justificació del treball

Descripción general

El treball consistirà en el desenvolupament d'una aplicació per dur a terme l'anàlisi de les rutes (*Pathway analysis*). Amb les rutes entenem un conjunt de gens que actuen junts per dur a terme un procès biològic. Així doncs aquest anàlisi permet donar més sentit a una expressió genètica diferencial entre les proves biològiques d'interès. Recordem que recents avenços tecnològics permeten mesurar els nivells d'expressió en una gran quantitat de gens, cosa que implica una gran quantitat de dades. Al nivell dels gens individuals es poden fer servir mètodes estadístics per comprovar si les diferències en les expressions entre els grups (proves biològiques) són estadísticament significatives. Per dotar encara de més sentit aquesta anàlisi és necessari agregar els resultats al nivell més raonable com ara al nivell de les rutes. Al final el que volem és comprovar si hi ha diferències estadísticament significatives entre les proves no al nivell dels gens perticulars sinó al nivell de les rutes. Tan com en el cas dels gens particulars també en el nivell de les rutes s'han desenvolupat mètodes estadístics específics [Khatri et al., 2012]. En aquest treball vull analitzar quins mètodes són i quins tenen més avantatges que d'altres. A part d'aquest component més biològic i teòric del treball buscaré la possibilitat d'implementar aquests mètodes d'anàlisi en una aplicació intuïtiva i d'un ús fàcil a la qual qualsevol científic que no disposi dels coneixements informàtics suficients per fer aquesta anàlisi podrà accedir gratuïtament. La plataforma que utilitzaré per crear l'aplicació és l'eina Shiny de Rstudio [Chang et al., 2018]. La feina doncs consistirà en la cerca dels paquets de Bioconductor que inclouen els mètodes per l'anàlisi de les rutes, selecció dels paquets més apropiats i la seva integració en una aplicació Shiny amb una interfície atractiva. Ja he fet una cerca previa sobre els paquets-canditats per una applicació. Aquests paquets són: ReactomePA, GAGE, clusterProfiler. Uns dels objectius és però fer una cerca dels paquets més exhaustiva.

Justificació del TFM

La justificació d'aquest tema ve de dues fonts diferents: d'una banda tinc un interès personal i d'altra banda entenc la importància de la meva aportació per a la comunitat científica. El meu interès personal és degut al fet que durant el màster he fet servir àmpliament el programa R però no he arribat a conèixer bé la creació d'una aplicació estadística amb Shiny. Per completar aquesta deficiència i entenent que aquesta eina és útil per al méu desenvolupament professional he buscat el tema que en requeria l'ús. Tot i la importància de l'anàlisi de les rutes, al meu saber encara no existeix cap aplicació Shiny que integri paquets diversos i molt efectius de Bioconductor. L'ús d'aquests paquets requereix coneixements informàtics i estadístics específics i per tant és difícilment accecible per la gran part de la comunitat scientífica. Encara que hi ha ja plataformes gratuïtes que ofereixen l'anàlisi de les rutes [Reimand et al., 2019] crec que val la pena desenvolupar una eina més que seria de codi obert.

Objectius

Objectius generals

1. Identificar els objectius i mètodes de l'anàlisi de les rutes (Bio/Stat)
2. Identificar els paquets de Bioconductor en R que s'aproximin als mètodes (Info)
3. Desenvolupar l'aplicació Shiny amb els paquets escollits per aproximar el resultat als objectius de l'anàlisi de les rutes (Info)

Objectius específics

1. Biologia/Estadística
   (a) Buscar literatura sobre l'anàlisi de rutes
      - Quins mètodes hi ha? Enumerar-los i explicar-los, especialment els tests estadístics.
        - Els tests sobre uns llistats de gens diferencialment expressats → Test de distribució hipergeomètrica?
        - Els tests sobre tots els gens d'experiment → GSEA → Enrichment Score → Permutació de gens/mostres? → Kolmogorov Test?
      - Quines bases de dades es fan servir?
      - Què signifiquen els diagrames més usats en l'anàlisi de rutes?
        - Barplots
        - Enrichment Maps
        - Barplots/Dot plots
        - GSEA plots
        - Altres?
   (b) Identificar les aplicacions existens i investigar què ofereixen
   (c) Analitzar els vignettes dels paquets de Bioconductor i provar-ne el seu ús localment amb R
      - RectomePA
      - GAGE
      - clusterProfiler
      - Altres?
2. Informàtica
   (a) Crear i documentat un protocol (pipeline) de l'anàlisi utilitzant els paquets seleccionats.
   (b) Identificar les dades experimentals per passar-les pel pipeline creat
   (c) Fer proves amb les dades seleccionades
   (d) Fer canvis en el protocol si és necessari
   (e) Integrar el pipeline a l'aplicació Shiny

Enfocament i mètode a seguir

Com es pot deduir a partir dels objectius la feina consistirà d'una banda en l'anàlisi teòrica dels mètodes disponibles actualment per a l'anàlisi de rutes, i d'altra banda en el desenvolupament d'una aplicació que incorporarà aquests mètodes. Es plantegen bàsicament dues estratègies:

1. El mètode seqüencial on primer s'analitza la teoria i després es programa l'aplicació;
2. El mètode simultani on la programació es desenvoluparà alhora de l'anàlisi dels conceptes teòrics (*learning by doing*).

Crec que la segona estratègia és més efectiva perquè la implementació pràctica ajuda a assimilar conceptes teòrics. M'imagino que metodològicament es farà el següent:

1. Trobar un mètode teòric que proporcioni un resultat interessant;
2. Buscar en Bioconductor aquest mètode;
3. Repetir 1 i 2 fins que el conjunt dels mètodes facin l'anàlisi de les rutes complet. Omplir la taula següent de la manera més exhaustiva possible;

Mètode	Paquet Bioconductor	Funció	Observació
GSEA	ReactomePA	gsePathway()	Permutació dels gens (no mostres)
...	...	...	...

4. Quan tots els mètodes estan triats dissenyar un protocol;
5. Aplicar el protocol a les dades independents;
6. Comparar els resultats amb els estudis d'on provenen les dades;
7. Ajustar una darrera vegada el protocol;
8. Desenvolupar l'aplicació

M'agradaria emfatitzar el punt 5 i 6. És essencial trobar les dades que s'utilitzaran per fer les proves durant la fase de desenvolupament de *pipeline*. Les dades han de provenir d'uns resultats ja publicats per poder comparar-los amb els resultats obtinguts amb el programari elaborat. Els dos han de ser similars. Sinó s'han de fer comprovacions del codi i el seu ajustament. Només quan els resultats del pipeline són acceptables es procedirà a desenvolupar l'aplicació.

Planificació amb hits i temporització

Tasques

1. Cerca de la literatura sobre els mètodes de l'anàlisi de les rutes;

2. Relacionar els mètodes trobats en 1 amb els paquets actuals de Bioconductor;

3. Decidir sobre quins resultats són més interessants per a una aplicació Shiny i desenvolupar un protocol d'anàlisi (*pipeline*) que formarà la base de l'aplicació. Documentar el protocol;

4. Buscar 3-5 exemples de dades i fer proves aplicant el protocol i comparant els resultats amb els resultats publicats sobre aquestes dades (si n'hi ha);

5. Fer els últims canvis en el protocol;

6. Dissenyar i programar l'aplicació de Shiny;

7. Publicar l'aplicació en web;

8. Tancar la memòria i fer la presentació per la defensa.

Calendari

Figure 1: Gantt Plot

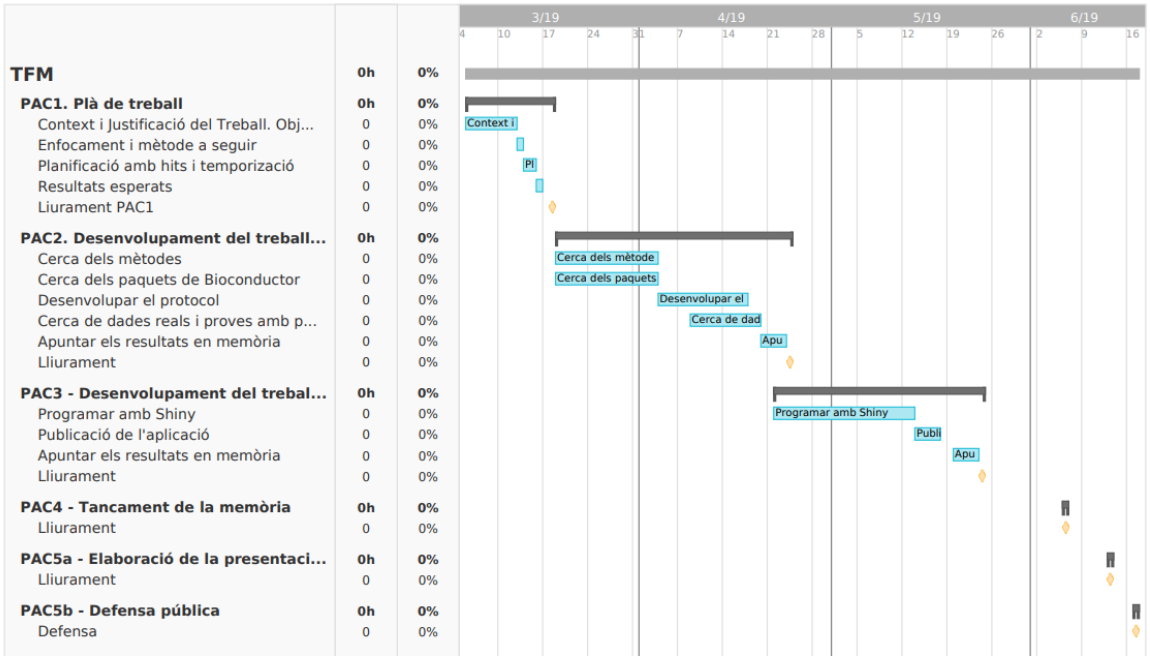


Figure 2: Gantt List

TFM

PAC1. Plà de treball	**92%**		**Start**	**Due**
Context i Justificació del Treball.	100%		Mar 5, 2019	Mar 12, 2019
Enfocament i mètode a seguir	100%		Mar 13, 2019	Mar 13, 2019
Planificació amb hits i	100%		Mar 14, 2019	Yesterday
Resultats esperats. Estructura del	100%		Today	Today
Liurament PAC1			Monday	Monday
PAC2. Desenvolupament del	**0%**		**Start**	**Due**
Cerca dels mètodes	0%		Tuesday	Apr 3, 2019
Cerca dels paquets de	0%		Tuesday	Apr 3, 2019
Desenvolupar el protocol	0%		Apr 4, 2019	Apr 17, 2019
Cerca de dades reals i proves	0%		Apr 9, 2019	Apr 19, 2019
Apuntar els resultats en memòria	0%		Apr 20, 2019	Apr 23, 2019
Lliurament			Apr 24, 2019	Apr 24, 2019
PAC3 - Desenvolupament del	**0%**		**Start**	**Due**
Programar amb Shiny	0%		Apr 22, 2019	May 13, 2019
Publicació de l'aplicació	0%		May 14, 2019	May 17, 2019
Apuntar els resultats en memòria	0%		May 20, 2019	May 23, 2019
Lliurament			May 24, 2019	May 24, 2019
PAC4 - Tancament de la	**0%**		**Start**	**Due**
Lliurament			Jun 6, 2019	Jun 6, 2019
PAC5a - Elaboració de la	**0%**		**Start**	**Due**
Lliurament			Jun 13, 2019	Jun 13, 2019
PAC5b - Defensa pública	**0%**		**Start**	**Due**
Defensa			Jun 17, 2019	Jun 17, 2019

Figure 3: Maig

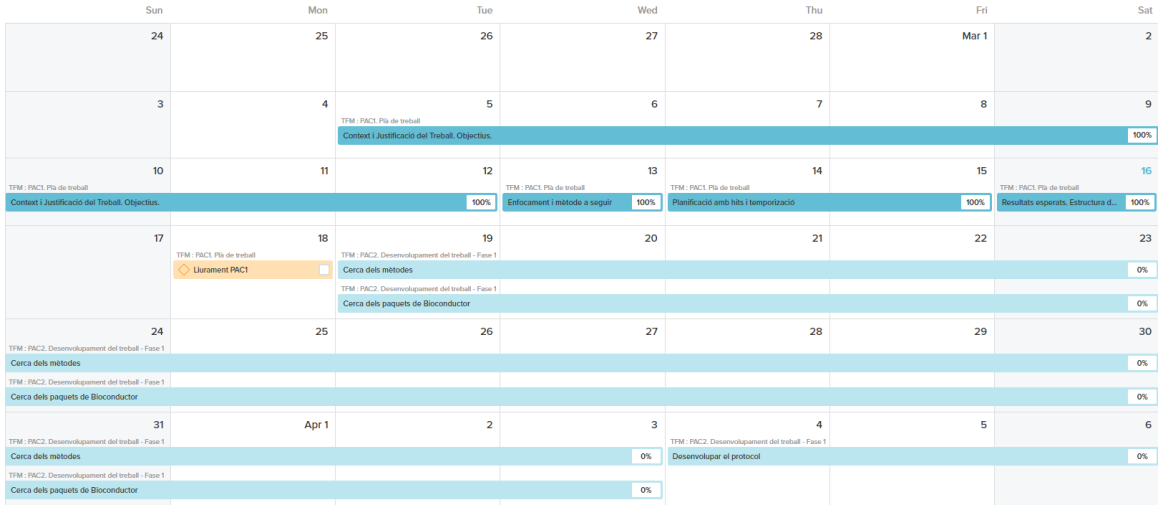


Figure 4: Abril

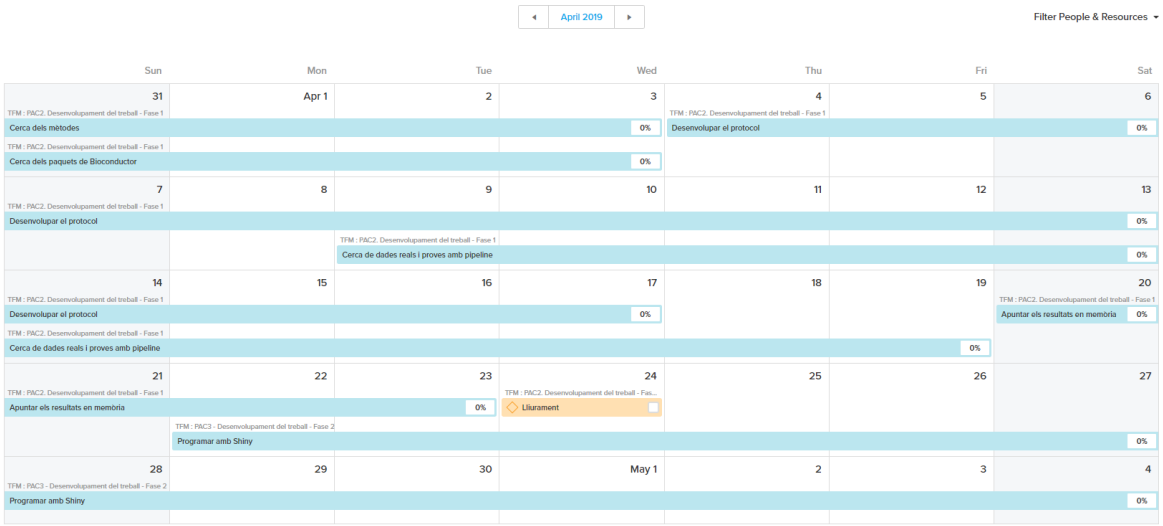
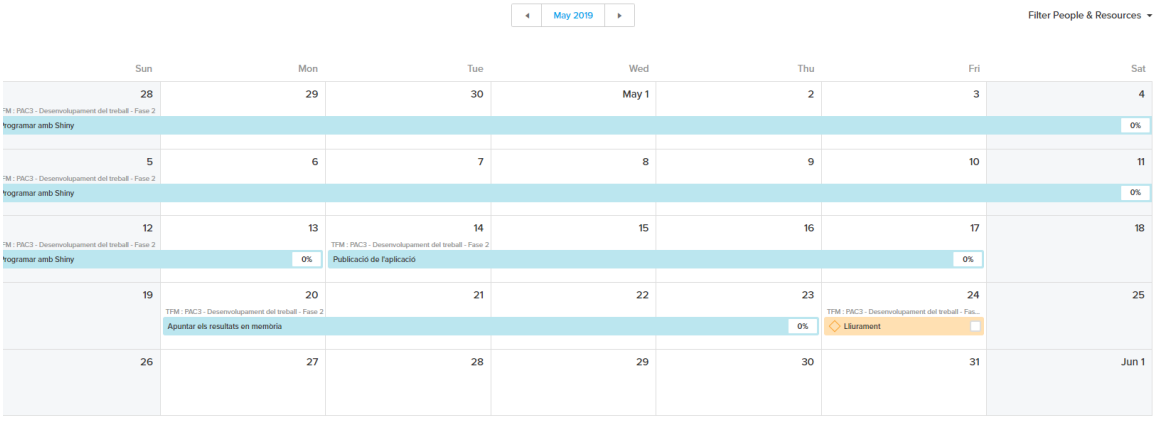


Figure 5: Mai



Hits

Els hits estan definits pel pla docent:

Activitat	Nom d'activitat	Data d'inici	Data d'entrega
PAC0	Definició dels continguts del treball	20/02/19	04/03/2019
PAC1	Pla de treball	05/03/19	18/03/19
PAC2	Desenvolupament del treball - Fase 1	19/03/19	24/04/19
PAC3	Desenvolupament del treball - Fase 2	25/04/19	20/05/19
PAC4	PAC4 - Tancament de la memòria	21/05/19	05/06/19
PAC5a	PAC5a - Elaboració de la presentació	06/06/19	13/06/19
PAC5b	PAC5b - Defensa pública	17/06/19	26/06/19

Anàlisi de riscs

He elegit aquest tema perquè la creació d'una aplicació web amb Shiny és un tema nou per a mi, i és un procediment que m'agradaria aprendre. D'una banda és un desafiament personal per a mi però també representa un risc perquè alguns funcionament del Shiny no seran obvis i necessitaran una cerca intensiva de solucions. L'altre factor del risc podria ser la gran quantitat d'informació sobre el tema de pathways tant en marc teòric com en marc pràctic dels paquets disponibles per fer l'anàlisi. Aquí serà precís prendre les decisions oportunes i ràpides i no desviar l'atenció de l'objectiu principal. Podem resumir els factors del risc de la forma següent:

1. No trobar els paquets adequats;
2. Trobar gran quantitat dels paquets que ofereixen estadístiques diferents. Cosa que dificultaria la selecció de les estadístiques adequades;
3. Problemes amb la creació de l'aplicació amb Shiny perquè serà el primer contacte amb Shiny. Aquí poden aparèixer problemes imprevistos la solució dels quals implicarà una cerca intensiva de solucions en web i fòrums dedicats a Shiny;
4. Dificultats a l'hora de la publicació de l'aplicació. Hi ha diferents opcions de publicació de l'aplicació: amb shinyapps.io i amb servidor Shiny. La primera opció és més preferida perquè no implica ni la creació del servidor ni la seva configuració especial. La creació d'un servidor implicarà costos econòmics que van en direcció contrària a l'objectiu de fer l'aplicació gratuïta. L'opció de shinyapps.io però pot ser inviable perquè es podran produir errors inesperats que necessitaran control sobre el servidor cosa que shinyapps.io no ofereix (vegeuun exemple aquí.
5. Els imprevistos personals.

Resultats esperats

Pla de treball

Amb aquest document es pretén estructurar el projecte, definir-ne els objectius i assignar-ne les tasques de tal manera que l'objectiu es pugui aconseguir en el temps d'entrega establert. Es tracta d'identificar els possibles problemes que poden influir en el compliment temporal de les tasques definides. És la primera valoració del projecte des de perspectives diferents.

Memòria

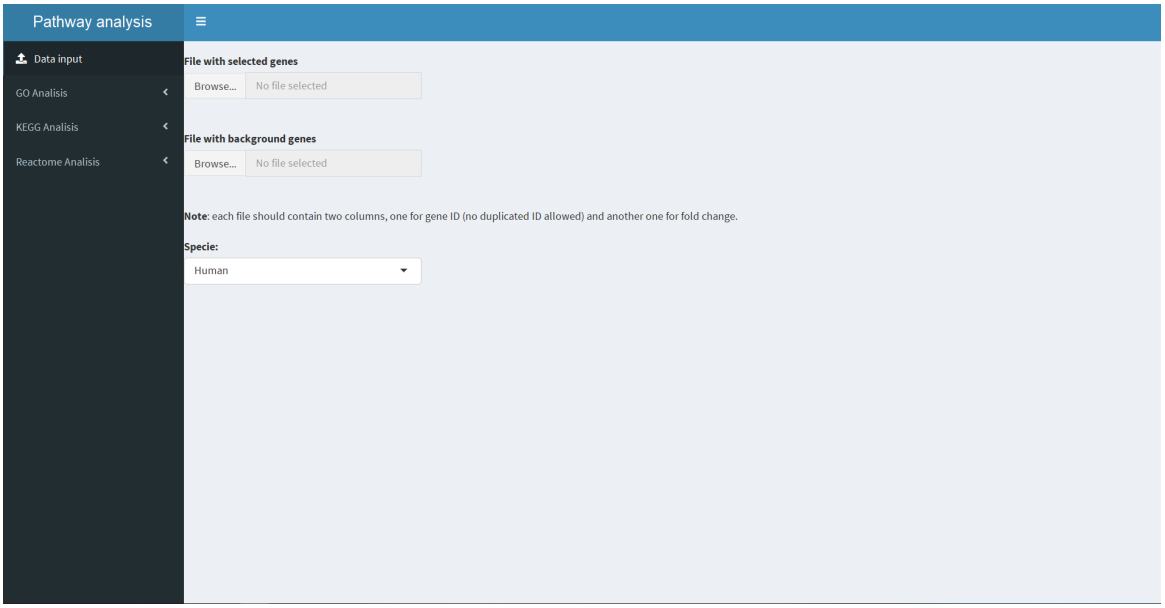
Els mètodes i la seva implementació via programari R han de ser documentats per fer el procés de creació el més reproduïble possible. Amb la memòria es pretén documentar de manera estructurada els passos fets per complir els objectius. També la memòria inclourà l'ús detallat del producte (Aplicació Shiny).

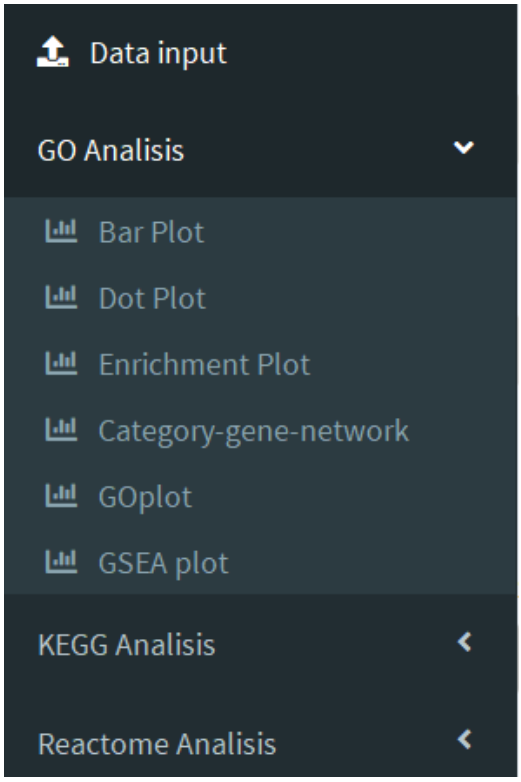
Producte

Amb aquest treball de màster es pretén crear una aplicació Shiny que es basarà en els últims avenços teòrics en l'àmbit de l'anàlisi de les rutes. També haurà de proporcionar una interfície d'ús fàcil i a més a més ser gratuït. Els resultats obtinguts amb l'aplicació hauran de tenir la qualitat suficient per poder ser usats en una publicació científica.

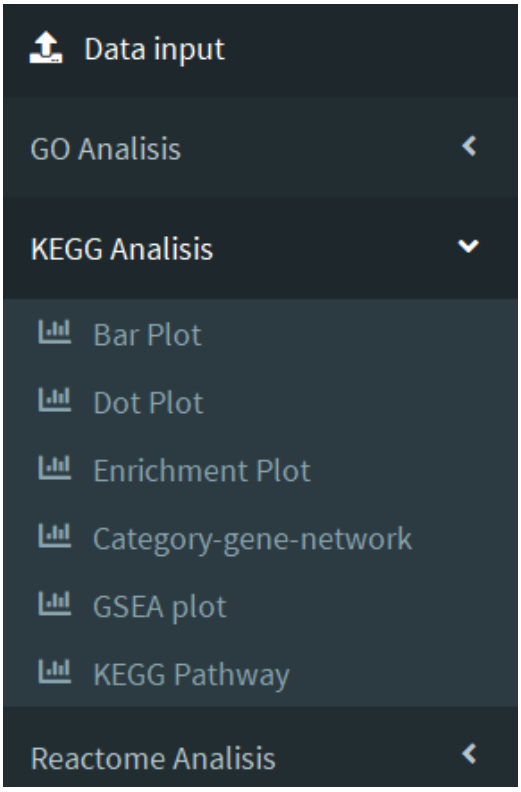
Ja he començat a experimentar amb Shiny i he creat la primera visió de l'aplicació que servirà com a graella:

Figure 6: Disseny de l'aplicació

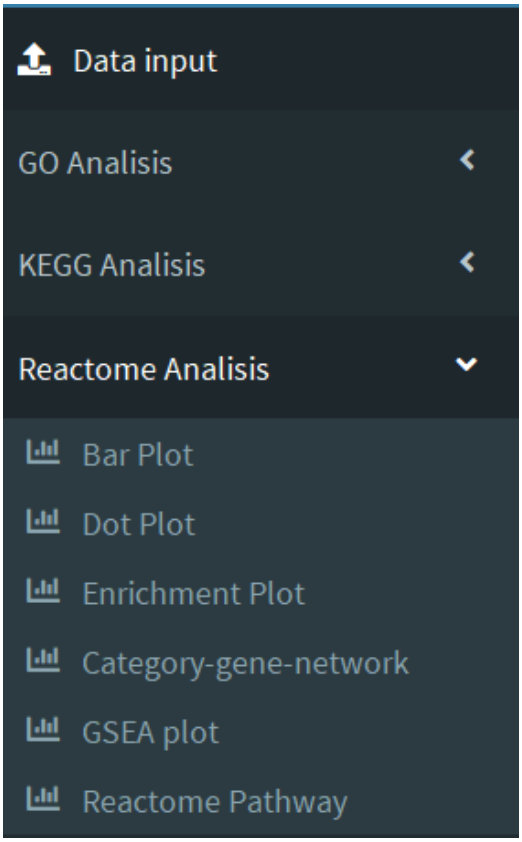




(a) Go Analysis



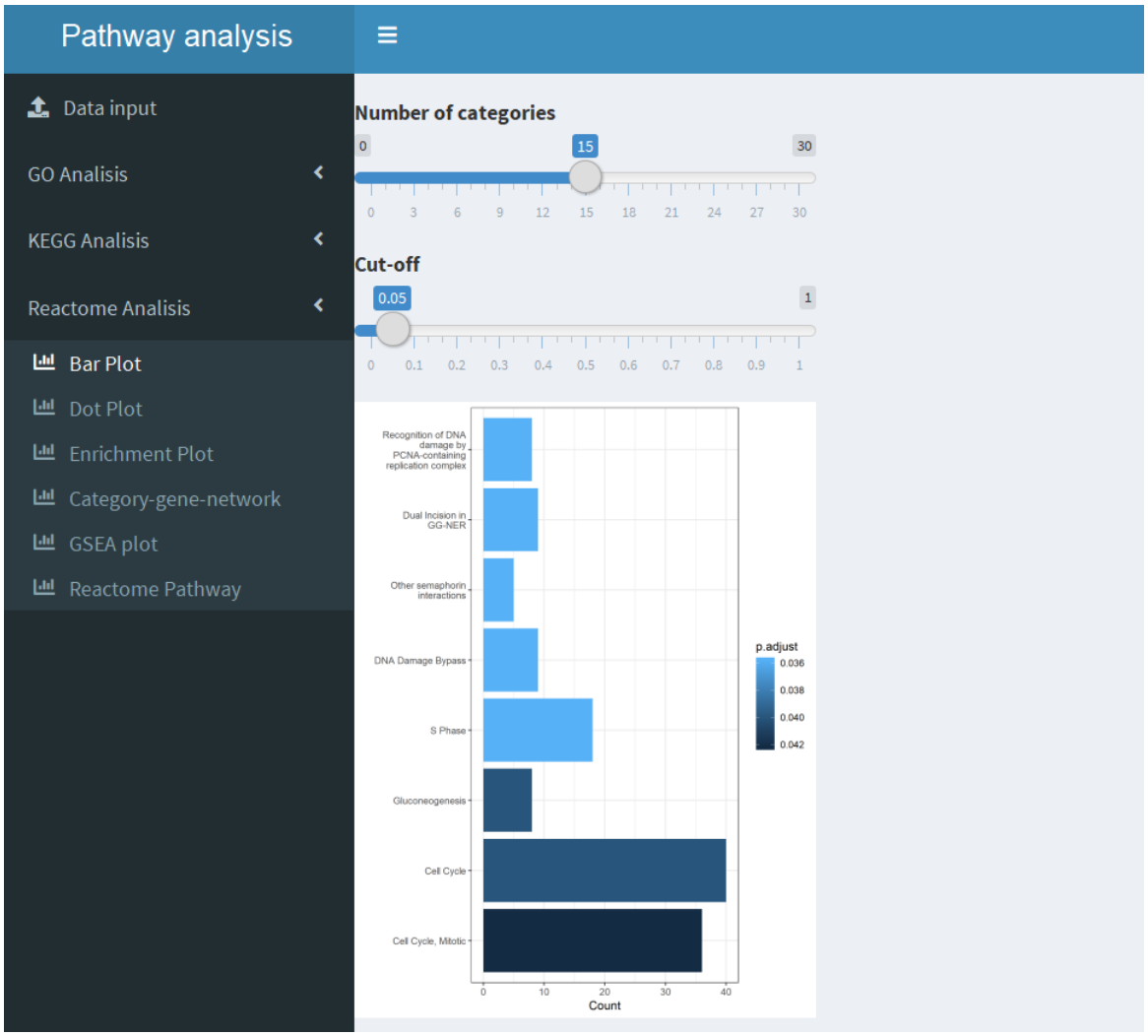
(b) KEGG Analysis



(c) Reactome Analysis

Figure 7: Les anàlisis disponibles

Figure 8: Exemple del disseny



Presentació virtual

El resultat de la feina haurà de ser presentat davant un tribunal que valorarà l'esforç fet i a més la qualitat del producte aconseguit.

Autoavaluació del projecte

A curt termini gairebé mai es pot aconseguir que el resultat sigui perfecte. Aquí s'haurà d'indicar tan les deficiències com les vies de millora sobre el producte presentat.

Estructuració del projecte

L'estructura del projecte està condicionada pel pla docent i inclou els components següents:

- Memòria del projecte
- El producte assolit
- Presentació dels mètodes utilitzats i del producte amb les seves característiques tècniques i pràctiques
- Defensa del Treball davant el tribunal.

La memòria consistirà en una descripció estructurada dels conceptes teòrics relacionats amb l'anàlisi de rutes. També inclourà la descripció de l'aplicació i el protocol del seu ús. Aquí s'indicarà si tots els objectius han estat assolits. També es farà una autoavaluació per comentar els problemes no resolts (si n'hi ha) i les possibilitats de millora.

El producte obtingut es farà accessible on line en una de les vies indicades: o bé via shinyapps.io o bé via un servidor. Aquest producte ha de contenir eines per dur a terme l'anàlisi de les rutes i ha de ser capaç de crear un *output* amb una qualitat apta per a una publicació científica. Que vol dir un rigor metodològic adequat i la qualitat de presentació/visualització dels resultats.

Finalment la memòria i el producte estaran resumits en una presentació amb contingut per a un màxim de 20 minuts d'intervenció oral. Aquí s'inclourà la descripció del *pipeline* elaborat i l'exemple pràctic d'anàlisi amb l'aplicació. S'inclouran les carpetes de pantalla dels components principals de l'aplicació juntament amb les tables i gràfics generats durant l'anàlisi de les dades de prova.

En darrer lloc es farà la defensa pública davant del tribunal.

Durant tot el procés s'hauran d'entregar informes de seguiment com a via per exposar el progrés de treball al professor responsable. Les suggerències del professor s'incorporaran al desenvolupament de l'aplicació.

References

[Chang et al., 2018] Chang, W., Cheng, J., Allaire, J., Xie, Y., and McPherson, J. (2018). *shiny: Web Application Framework for R*. R package version 1.2.0.

[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.

[Reimand et al., 2019] Reimand, J., Isserlin, R., Voisin, V., Kucera, M., Tannus-Lopes, C., Rostamianfar, A., Wadi, L., Meyer, M., Wong, J., Xu, C., et al. (2019). Pathway enrichment analysis and visualization of omics data using g: Profiler, gsea, cytoscape and enrichmentmap. *Nature protocols*, page 1.