



**DESENVOLUPAMENT I ANÀLISI D'UN CONJUNT D'EINES
BIOINFORMÀTIQUES PER AL RASTREIG DE PACIENTS DE LA
BASE DE DADES GEO PER ESTUDIAR I IDENTIFICAR
PATRONS D'EXPRESSIÓ GÈNICA I COMORBIDITATS
ASSOCIADES**

Cinta Gisbert Bouzas

TREBALL FINAL DE GRAU BIOTECNOLOGIA

Tutor acadèmic: Gemma Beltran, departament de Bioquímica i Biotecnologia,
gemma.beltran@urv.cat

En cooperació amb: Anaxomics

Supervisor/s: Xouse Mas, xouse@anaxomics.com

Data de defensa: 13 de juliol de 2020

Jo, Cinta Gisbert Bouzas , amb DNI 47479565-K, sóc coneixedor de la guia de prevenció del plagi a la URV *Prevenció, detecció i tractament del plagi en la docència: guia per a estudiants* (aprovada el juliol 2017) (<http://www.urv.cat/ca/vidacampus/serveis/crai/que-us-oferim/formacio-competencies-nuclears/plagi/>) i afirmo que aquest TFG no constitueixen cap de les conductes considerades com a plagi per la URV.

Tarragona, 10 de juliol de 2020



AGRAÏMENTS

Han estat moltes les persones implicades en major o menor mesura en aquest treball. Malgrat que totes elles tenen i tindran el meu reconeixement, cal que faci un esment especial a :

El meu tutor professional Xouse, per saber motivar-me en cada moment en voler saber més sobre el tema i treballar per implementar solucions innovadores. Per resoldre tots els dubtes que em sorgien, per asseure's amb mi per parlar del treball i discutir sobre nous camins. En resum, per ser una persona tan implicada amb el tema i ajudar-me quan més ho necessitava. Va per ell el meu reconeixement.

Al Guillem, per estar-hi en tot moment, per resoldre cadascun dels meus petits grans dubtes, per passar-se més d'una, de dos i de tres hores senceres parlant-me sobre el tema, per facilitar-me tota la informació que necessitava i per guiar-me cada cop que estava perduda. Perquè sense ell, haguera estat totalment impossible la realització d'aquests treball. Va per ell, el meu agraïment etern.

A la resta de persones d'Anaxomics per comprendre tan bé el fet de la meva inexperiència i per explicar-me moltes coses totalment desconegudes per mi. Per acollir-me en el seu entorn de treball i fer-me sentir que formava part de l'equip. Voldria agrair-los el fet de que hagin perdut un minut del seu temps en fer-me sentir una més. Per ells, el meu agraïment infinit.

A la meva tutora Gemma Beltran, per la seva professionalitat i dedicació. Per agrair-li el temps que ha invertit en revisar el treball i l'interès que ha mostrat en tot moment. Va per ella el meu immens agraïment.

A més a més, també voldria dedicar unes paraules i mostrar el més profund dels meus agraïments a totes aquelles persones de la comunitat científica, ja que fan una tasca importantíssima dins de la societat, un tasca que molt cops no està suficientment reconeguda i la seva feina, esforç i dedicació ens permeten avançar com a societat. A tots aquells investigadors que inverteixen temps en compartir els seus resultats, i permeten la creació de bases de dades tan valuoses com GEO. Perquè sense l'existència d'aquestes dades haguera estat totalment impossible la realització d'aquest treball. Per vosaltres científics, el meu agraïment més entranyable.

Als meus pares i Chema, per aguantar els meus "t'ho puc llegir i hem dius si està bé?", per donar-me suport quan més el necessitava i per fer-me sentir que no estava sola, que ells estaven al meu costat al meu costat en tot moment. Va per ells el més sincer dels meus agraïments.

Índex

AGRAÏMENTS.....	4
1. DADES DEL CENTRE	8
2. RESUM.....	9
3. INTRODUCCIÓ	10
3.1 <i>Real World Data</i> (RWD).....	11
3.2 Base de dades Gene Expression Omnibus (GEO).....	11
3.3 Extracció de dades a partir de la font d'informació inicial	12
4. HIPÒTESIS DE TREBALL.....	14
5. OBJECTIU	14
6. MATERIALS I MÈTODES.....	15
6.1 Disseny experimental	15
6.2 Anàlisi estadístic	18
7. RESULTATS I DISCUSSIÓ	20
7.1 Resultats generats durant la fase Inicial	20
7.2 Resultats generats durant la segona fase	23
8. CONCLUSIONS.....	29
9. BIBLIOGRAFIA.....	30
10. AUTOAVALUACIÓ	31
ANNEX 1: LLISTA D'ABREVIATURES	32

1. DADES DEL CENTRE

Anaxomics ha estat l'empresa bioinformàtica on s'ha desenvolupat el projecte d'investigació. Té la seva seu al centre de Barcelona i hi treballen un total de setze persones conjuntament amb cinc doctorands i els diferents alumnes de pràctiques. És tracta d'una empresa bioinformàtica de l'àmbit privat que modela molecularment la fisiologia humana a través de la biologia de sistemes i intel·ligència artificial. Treballa per la integració de dades de diferents experiments i ofereix experimentació virtual generant resultats ràpids, econòmics i perspicaços.

Basa la seva investigació en descobrir i desenvolupar nous i millors medicaments per millorar la salut humana. Inverteix la majoria dels seus esforços en els experiments d'alt rendiment degut a la capacitat única que tenen de proporcionar grans volums de dades amb el potencial de millorar dràsticament el coneixement que es té sobre els processos biològics. Busca resoldre el problema que tenen la majoria d'investigadors al intentar extraure conclusions dels grans oceans d'informació generats. El meu pas per Anaxomics s'ha basat en treballar amb el projecte de la generació d'una base de dades de pacients a partir de la base dades GEO.

2. RESUM

El que es pretén en aquest treball d'investigació es desenvolupar diverses eines bioinformàtiques que permetin l'estudi de dades procedents de tècniques d'alt rendiment (High-Throughput). Es treballa amb la base de dades GEO (Gene Expression Omnibus) per poder crear una base de dades sòlida amb dades d'expressió gènica de pacients de diferents experiments emmagatzemats a GEO. S'estudia la diferència en el patró d'expressió gènica de gens que codifiquen per un llistat de proteïnes d'interès i es compara entre diferents grups de pacients. S'implementen diferents funcions bioinformàtiques que permeten al rastreig de la base de dades GEO i la generació de resultats realitzant els testos estadístics corresponents per a l'estudi de comorbiditats associades a diferents malalties analitzant els valors d'expressió gènica en diferents pacients.

Paraules clau: Expressió gènica, Gene Expression Omnibus (GEO) , tècniques d'alt rendiment (High-Throughput) , Real World Data (RWD) , eines bioinformàtiques.

3. INTRODUCCIÓ

Les tècniques d'alt rendiment (High-Throughput) permeten extreure informació simultània de milions de gens, proteïnes i/o metabòlits. Aquestes tècniques poden identificar ràpidament compostos actius, anticossos o gens. Els resultats poden ser punts de partida per a la identificació de patrons d'expressió gènica en un determinat grup de pacients o permetre la comparació entre diferents grups. Mentre l'anàlisi clàssica consisteix en mesures directament associades a una sola patologia estudiada, les dades procedents de tècniques de HT poden utilitzar-se com a bons identificadors de pacients per a l'estudi d'aquests, generant resultats estadísticament correctes (Willey *et al.* 2017). A pesar de les limitacions associades a l'anàlisi del gran volum de dades generades per aquestes tècniques, les dades procedents d'estudis de HT són aplicades als estudis clínics actuals.

Es bastant comú que els investigadors depositin les dades de HT dels pacients en bases de dades disponibles per a tots els membres de la comunitat científica. Tot i l'enorme quantitat d'informació generada, en la majoria dels casos, els investigadors centren el seu esforç en l'anàlisi d'experiments individuals, sense fer ús de recursos tant valuosos com tota la informació depositada en les bases de dades existents. Així doncs, l'anàlisi d'expressió genètica es realitza generalment des d'una perspectiva centrada amb els gens associats a una sola malaltia. Tota la informació depositada en les diferents bases de dades permet descobrir connexions entre malalties per similituds en les seves firmes d'expressió gènica (Longley and Adnan 2016). Un exemple d'aquestes bases de dades es GEO. Aquest treball es basa en explorar les dades depositades en la base de dades GEO per poder analitzar-les i extraure'n conclusions. Bàsicament és busca generar coneixement a partir de tota la informació que es té a l'abast en la comunitat científica.

Ja que les dades de HT descriuen als pacients a nivell mundial, l'empremta digital d'altres patologies també podria estar present en aquestes dades. Conseqüentment, es plausible que un cert percentatge de pacients puguin patir comorbiditats associades a la patologia estudiada, per exemple, els pacients estudiats per la malaltia de Alzheimer podrien tindre altres patologies relacionades amb l'envelliment. En la majoria dels casos, aquestes comorbiditats no s'informen en la base de dades (Wise *et al.* 2018).

3.1 Real World Data (RWD)

Real World Data es pot definir en el món científic com a dades derivades d'una sèrie de fonts que estan associades amb els resultats d'una població de pacients heterogènia en el món real. L'agència europea de Medicaments (EMA) ha observat que els resultats obtinguts en estudis d'assajos clínics aleatoris (ECA) no sempre s'ajusten a l'observació de medicaments en estudis observacionals. Aquest fet es provocat principalment pel nombre limitat de pacients, la monitorització excessiva d'aquests i el temps de seguiment limitat. Es bastant comú que certes reaccions adverses als medicament o falta d'eficàcia estiguin ocultes fins que es realitzin els estudis de RWD. Conseqüentment la EMA fomenta la extracció d'informació sobre fàrmacs i patologies de fonts complementàries a ECA per tal de generar millors evidències sobre els tractaments (Makady et al. 2017) .

Els estudis basats amb RWD podrien ajudar a identificar, acceptar o bé descartar hipòtesis, per mostrar-ne de noves sobre un medicament o una patologia, per facilitar la comprensió del perfil adequat d'un pacient, per evitar una possible falta d'eficàcia , etc. Com suggereixen molt autors en els articles més recents , RWD es una font d'informació que podria ajudar a prendre decisions en la indústria farmacèutica. Si més no, cal tenir en compte que les dades procedents de RWD provoquen algunes limitacions en comparació amb els ECA (Wise et al. 2018) . Les poblacions de pacients als estudis de RWD generalment són molt heterogènies, les variables que descriuen pacients de diferents fonts poden no ser les mateixes o no tindre exactament el mateix sentit. Aquest problema apareix inclús per a la variable principal i aquest fet condiciona la interpretació dels resultats. Els estudis de RWD generalment requereixen la normalització dels valors d'acord a les dades d'origen, la inferència de valors perduts i altres estratègies que generalment requereixen un anàlisi estadístic força complicat. Però presenten un nou enfoc en l'anàlisi de dades i fan que els resultats siguin més reals, ja que es treballa amb dades del món real (Longley and Adnan 2016) .

3.2 Base de dades Gene Expression Omnibus (GEO)

Gene Expression Omnibus (GEO) és una base de dades pública internacional de dades genòmiques funcionals que emmagatzema i distribueix lliurament dades d'experiments realitzats en *microarrays*, seqüenciació de pròxima generació i altres formes de dades genòmiques funcionals d'alt rendiment presentades per la comunitat investigadora (<https://www.ncbi.nlm.nih.gov/geo/>) . Aquesta base de dades proporciona un recurs inavaluable de dades genòmiques que poder ser integrades i analitzades per a validar noves hipòtesis i generar coneixement. El metanàlisi de dades d'expressió gènica es una tècnica cada cop més utilitzada en diversos camps d'investigació, ja que permet augmentar la reproductibilitat de l'estudi i descobrir nous biomarcadors (Min, Lee, and Yoon 2017) . Ara bé, la bioinformàtica té un paper molt important en la capacitat de poder integrar aquestes dades de manera que siguin analitzables.

GEO accepta dades basades en matrius i seqüències. Podria considerar-se una font del que es coneix com a dades del món real (RWD) d'acord amb les recomanacions de la Societat Internacional de Farmacologia i investigació de resultats. És una de les majors bases de dades públiques amb dades corresponents a estudis de *microarray* juntament amb Arrayexpress. Proporciona eines per ajudar als usuaris a consultar i descarregar experiments i perfils d'expressió gènica seleccionats. Només aquelles dades que compleixin amb l'estàndard MIAME (informació mínima sobre un experiment de *microarray*) són acceptades en la base de dades. Aquest estàndard descriu la informació mínima que ha d'incloure la descripció d'un estudi de *microarray*. El compliment de l'estàndard MIAME no està relacionat amb el format, sinó amb el contingut proporcionat. Un dels objectius de GEO es assegurar que proporciona suficient informació per permetre la interpretació a nivell general dels experiments proporcionats. D'aquesta manera tots els membres de la comunitat científica comparteixen els experiments realitzats i tenen tota la informació a l'abast per realitzar nous estudis.

Els principals objectius de GEO són entre d'altres proporcionar una base de dades robusta i versàtil en la que poder emmagatzemar, de manera eficient, dades genòmiques funcionals d'alt rendiment. A més, ofereix procediments i formats de presentació simples que acceptin fitxers de dades complets i ben anotats de la comunitat d'investigació. També proporciona mecanismes fàcils d'utilitzar que permeten als usuaris consultar, localitzar, revisar i descarregar estudis i perfils d'expressió gènica d'interès.

La base de dades GEO *Profiles* emmagatzema perfils d'expressió gènica derivats de conjunts de dades GEO seleccionats (<https://www.ncbi.nlm.nih.gov/geoprofiles>). Els perfils GEO es poden buscar utilitzant diferents atributs, incloent paraules claus, símbol de gens, nom de gens, nombre d'accessos de GenBank o perfils marcats com expressats diferencialment. Una eina molt útil per seleccionar pacients d'interès a l'hora de realitzar un estudi *in silico*.

3.3 Extracció de dades a partir de la font d'informació inicial

Anaxomics té un total de 1M de pacients procedents de la BD GEO guardats en un BD local o en diferents carpetes i fitxers. Cada pacient té la seva pròpia carpeta amb un llistat de fitxers que proporcionen informació sobre el pacient. Per aquest estudi s'ha utilitzat el fitxer que conté la informació de l'experiment de *microarray*. El fitxer consta de dues columnes una corresponent al *accession number* de la proteïna, que es l'identificador de la proteïna i l'altra corresponent al valor d'expressió normalitzat utilitzant la normalització *MinMax*. Per tal d'extreure la informació d'interès s'ha dissenyat un script que comprova pacient per pacient l'existència d'aquest fitxer, en cas de que si que el contingui genera un nou fitxer per cadascuna de les diferents proteïnes identificades en el pacient i anota l'identificador del pacient i el valor d'expressió per aquell pacient en el nou fitxer de proteïna. Així doncs al finalitzar aquest estudi

s'ha generat un total de 19540 fitxers corresponents a les diferents proteïnes identificades en tots els pacients . Cada fitxer conté un llistat amb l'identificador de pacients i el valor d'expressió del pacient per a la corresponent proteïna. Amb tota aquesta informació es poden realitzar diversos estudis, com per exemple la identificació de patrons d'expressió d'una proteïna en diferents pacients. Al llarg del treball s'especifiquen tots els processos realitzats, eines bioinformàtiques desenvolupades i es mostren alguns dels resultats obtinguts aplicant les eines bioinformàtiques creades per a l'estudi d'aquestes dades.

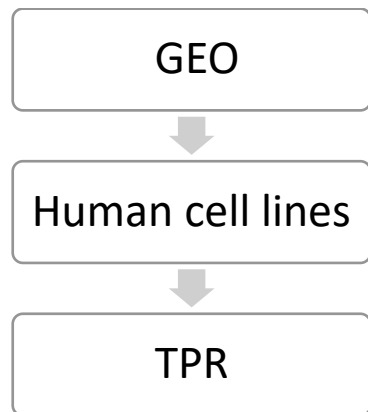


Figura 1. Figura per mostrar que l'extracció de dades inicial és de la base de dades GEO, on es seleccionen aquells experiments que s'han utilitzat cèl·lules humanes i es busca crear la base dades de pacients TPR (The Patient Repository) .

La base de dades GEO conté milions d'experiments amb humans, això permet tenir una base dades amb pacients del món real. Permet millorar els estudis realitzats fins el moment amb la metodologia RCT (*Randomized Controlled Trial*). En molts articles l'EMA ha reconegut el gran poder i aplicabilitat que té la utilització de RWD en assajos clínics.

El que es busca es generar una base dades amb pacients GEO dels que es coneix el seu patró d'expressió gènic. La creació del nou projecte, TPR (*The Patient Repository*), busca generar una base de dades de pacients del món real. Amb dades reals, d'alta qualitat per millorar la salut global. Aquets pacients es poden utilitzar per a la realització de diferents assajos clínics, desenvolupament de medicaments, medicina de precisió o avaluació de riscos entre d'altres.

4. HIPÒTESIS DE TREBALL

Existeixen diferències significatives al estudiar la distribució i mitjana dels valors d'expressió de gens que codifiquen per un llistat de proteïnes efectores d'asma en pacients asmàtics vs pacients aleatoris. Les diferències no són tant significatives al estudiar la distribució i mitjana de valors d'expressió gènica de gens que codifiquen per un llistat de proteïnes aleatòries en pacients asmàtics vs pacients aleatoris. L'anàlisi de la BD GEO permet extraure informació coherent a la bibliogràfica i identificar comorbiditats no detectades fins al moment, generant nou coneixement.

5. OBJECTIU

L'objectiu principal d'aquest treball es disposar d'un conjunt d'eines que permetin rastrejar la base dades GEO i comparar l'expressió diferencial entre dos grups de pacients determinats, per exemple pacients amb asma vs pacients aleatoris.

Aquest treball té també com objectiu secundari la comparació de patrons d'expressió gènica per a l'estudi d'una proteïna d'interès en dos grups de pacients determinats.

Durant la realització d'aquest treball es té com objectiu desenvolupar i analitzar la utilitat de les següents eines bioinformàtiques:

Rastreig de de pacients 1: Una funció per generar un fitxer per proteïna per tots els pacients de GEO que conté l'identificador de pacient i el valor d'expressió gènica (els valors d'expressió utilitzats estan normalitzats amb el *MinMax*) .

Rastreig de de pacients 2: Una funció que retorna el percentatge de pacients amb els valors d'expressió gènica més alts o més baixos (segons s'especifiqui) de la proteïna o llistat de proteïnes que s'especifiquin.

Comparar patrons d'expressió 1: Una funció que retorna la gràfica PDF (*probability density function*) per a l'expressió gènica de gens que codifiquen per un llistat de proteïnes en pacients i teixits específics.

Comparar patrons d'expressió 2: Una funció que retorna la gràfica PDF del valor d'expressió mitjà de 1000 proteïnes aleatòries d'un teixit específic. També hi ha l'opció de que en lloc de mostrar totes les proteïnes només mostri les que ens interessin.

Es treballa per fer un anàlisi exploratori de les dades que hi ha dipositades en la BD GEO ordenar-les, transformar-les, visualitzar-les i intentar modelar-les per tal de ser capaços de generar coneixement a partir de tota la informació emmagatzemada. Demostrar la hipòtesis plantejada a partir del tractament de les dades disponibles.

6. MATERIALS I MÈTODES

6.1 Disseny experimental

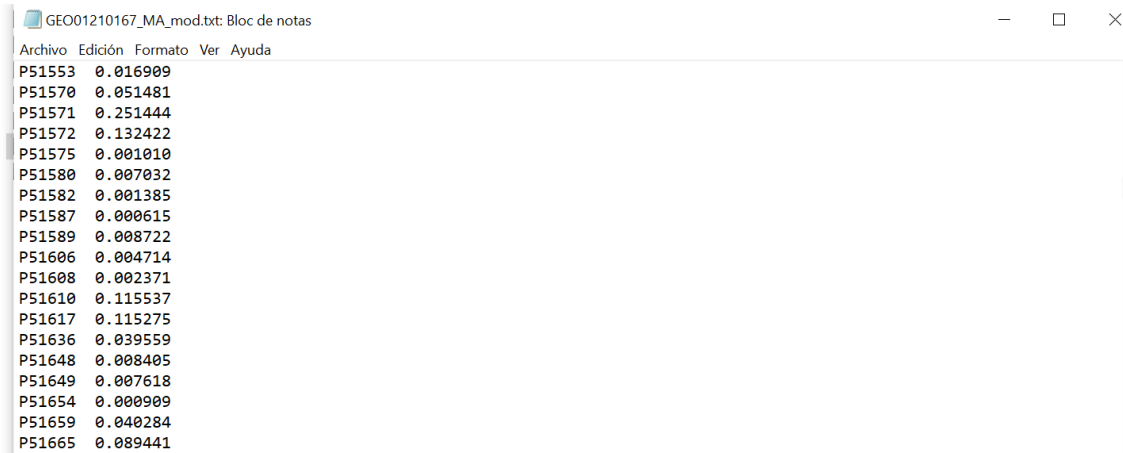
El treball es divideix amb dues fases. Una primera fase inicial de tractament i transformació de les dades i una segona fase d'estudi i anàlisi de les dades generades, un cop ja han estat integrades. La principal font d'informació és la base de dades GEO. S'ha treballat amb un total d'un milió de pacients durant la primera fase del treball per generar les eines bioinformàtiques inicials. L'estudi final s'ha realitzat amb un total de mil pacients aleatoris per posar en pràctica les eines bioinformàtiques desenvolupades i comprovar el seu funcionament. Els pacients utilitzats en l'estudi han estat extrets de la base de dades GEO. Aquests, estan guardats en diferents carpetes, cada carpeta es correspon amb un pacient i conté diferents fitxers, tal i com es mostra en les següents imatges. Cadascun dels fitxers continguts en la carpeta de cada pacient dona informació sobre aquest. A continuació es comenta el contingut dels fitxers i quins són els que s'ha utilitzat per a l'estudi realitzat.

Nombre	Fecha de modificación	Tipo	Tamaño
GEO01210167	03/03/2020 19:17	Carpeta de archivos	
GEO01210168	03/03/2020 19:15	Carpeta de archivos	
GEO01210169	03/03/2020 19:12	Carpeta de archivos	
GEO01210170	03/03/2020 19:10	Carpeta de archivos	
GEO01210171	03/03/2020 19:08	Carpeta de archivos	
GEO01210172	03/03/2020 19:05	Carpeta de archivos	
GEO01210173	03/03/2020 19:02	Carpeta de archivos	
GEO01210174	03/03/2020 19:00	Carpeta de archivos	
GEO01210175	03/03/2020 18:58	Carpeta de archivos	

Imatge 1. Carpeta que conté diferents pacients procedents de la base de dades GEO. Les dades s'han integrat d'aquesta manera per tal de generar una base de dades sòlida en l'empresa i poder fer consultes per futurs estudis in silico.

Nombre	Fecha de modificación	Tipo	Tamaño
GEO01210167.txt	12/11/2019 19:55	Documento de texto	6 KB
GEO01210167_enrich_KEGG.txt	03/03/2020 19:17	Documento de texto	3 KB
GEO01210167_info.txt	26/11/2019 5:10	Documento de texto	1 KB
GEO01210167_labels.txt	11/01/2020 20:36	Documento de texto	1 KB
GEO01210167_MA_logitNorm_4.txt	14/02/2020 0:09	Documento de texto	290 KB
GEO01210167_MA_logitNorm_vers4.txt	14/02/2020 16:09	Documento de texto	290 KB
GEO01210167_MA_minmaxNorm_vers5.txt	15/02/2020 8:07	Documento de texto	290 KB
GEO01210167_MA_mod.txt	22/11/2019 20:10	Documento de texto	332 KB
GEO01210167_MA_raw.txt	12/11/2019 19:55	Documento de texto	1.169 KB

Imatge 2. Contingut de la carpeta corresponent al pacient amb identificador GEO01210167.



Protein ID	Normalized Expression
P51553	0.016909
P51570	0.051481
P51571	0.251444
P51572	0.132422
P51575	0.001010
P51580	0.007032
P51582	0.001385
P51587	0.000615
P51589	0.008722
P51606	0.004714
P51608	0.002371
P51610	0.115537
P51617	0.115275
P51636	0.039559
P51648	0.008405
P51649	0.007618
P51654	0.000909
P51659	0.040284
P51665	0.089441

Imatge 3. Fitxer MA_mod que conté els valors d'expressió gènica procedents de l'experiment de microarray. Els valors estan normalitzats. La primera columna correspon a l'identificador de la proteïna i la segona correspon al valor d'expressió de la proteïna corresponent per al pacient amb l'identificador GEO0120167.

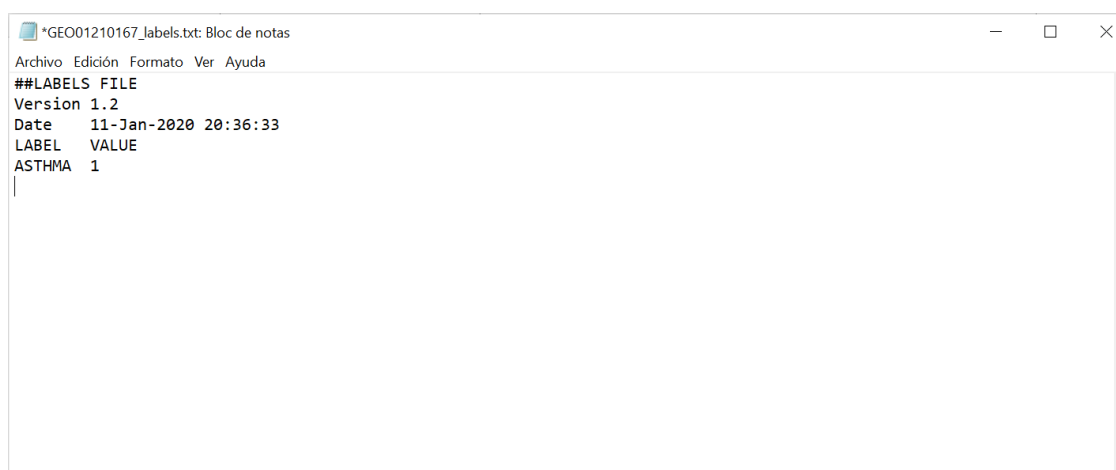
Com es pot observar en les imatges anteriors cada pacient conté un fitxer amb el llistat de proteïnes estudiades en l'experiment de *microarray* realitzat i el valor d'expressió gènica per a cadascuna d'aquestes. Uns dels principals problemes en l'anàlisi de dades GEO és que la gran quantitat d'informació que és té a l'abast suposa un alt cost computacional i per tant, molt temps d'espera per a l'obtenció de resultats. Per exemple, en el cas de que es vulgui fer una consulta sobre el valor d'expressió gènica del gen que codifica per una determinada proteïna en tots els pacients s'hauria de consultar pacient per pacient, el fitxer corresponent i seleccionar la proteïna d'interès i guardar-la en un nou fitxer. Així doncs, s'hauria de recórrer un milió de carpetes per obtenir la informació d'una sola proteïna i repetir el procés per les dinou mil proteïnes analitzades. És per això que el primer que es va plantejar es la generació d'un fitxer per proteïna que contingui dos columnes. Una primera columna amb l'identificador del pacient i una segona amb el valor d'expressió de la proteïna corresponent. D'aquesta manera només caldrà recórrer el milió de pacients un cop i ja es tindrà tots els valors d'expressió de la proteïna d'interès en un sol fitxer. En lloc de recórrer 1M de fitxers, obrint un sol fitxer ja es té a l'abast tota la informació de la proteïna d'interès. Per executar aquesta funció es necessiten 3 paràmetres d'entrada tal i com es mostra en la següent línia de codi:

```
axTprGeneratePatientsProteinExpressionList(patStart,patEnd, NormVersion)
```

Al treballar amb un volum de pacients tant gran és millor agrupar els pacients en diferents grups per tal d'enviar els diferents grups de pacients a les diferents màquines disponibles de l'empresa. D'aquesta manera s'assignen els diferents processadors lliures per executar la funció implementada de forma paral·lela. El primer paràmetre correspon al primer pacient del grup a analitzar i el segon a l'últim pacient del grup amb el que es treballarà. L'últim paràmetre fa

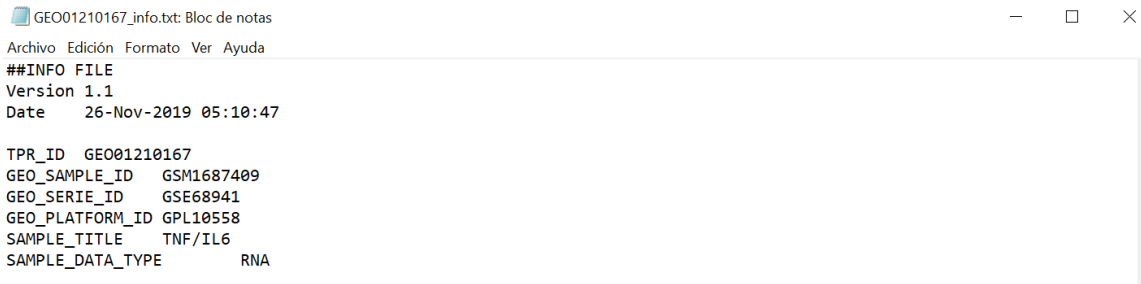
referència al fitxer d'on s'agafaran les dades de valor d'expressió gènica. Hi ha diferents versions de normalització.

Es important especificar també que hi ha una base de dades que permet fer consultes de manera eficient i seleccionar els identificadors dels pacients amb un paràmetre específic, per exemple pacients amb l'etiqueta “asma”. Això es gràcies a la informació que hi ha en el fitxer *ID_Pacient_label*. Tenint en compte l'etiqueta del pacient es poden seleccionar també els pacients dels que es té informació sobre un teixit concret, una línia cel·lular concreta, etc. Podria ser interessant estudiar el patró d'expressió gènica en un teixit específic i comparar-lo entre individus malalts i individus sans. Hi ha molts paràmetres possibles a seleccionar. Un cop es té el llistat dels identificadors dels pacients d'interès s'executen les diferents eines bioinformàtiques per agafar només la informació d'interès. Les imatges adjuntes a continuació mostren el contingut del fitxer label. Per entendre on trobar aquest fitxer revisar la imatge 2.



Imatge 4. Fitxer label que conté informació sobre la label associada al pacient. En aquest cas es pot observar que es tracta d'un pacient amb asma.

A més a més, també es molt important la informació que aporta el fitxer *ID_Pacient_info*. Aquest fitxer ens dona informació sobre l'experiment al que està associat a la base de dades GEO, l'identificador del pacient a la base de dades GEO i la plataforma que s'ha utilitzat per a l'obtenció dels valors d'expressió gènica. Aquest és un paràmetre important a l'hora de fer l'estudi estadístic ja que les dades poden variar si es treballa amb plataformes diferents, poden utilitzar diferents unitats, etc. Per tant els paràmetres d'experiment i plataforma s'han tingut en compte per seleccionar els pacients inclosos en l'estudi. Tots els pacients amb els que s'ha treballat per l'obtenció de resultats han utilitzat la plataforma GPL10558 per a l'extracció de dades en els experiments de *microarray* realitzats.



Imatge 5. Contingut fitxer info del pacient. El GEO_SAMPLE_ID ens dona informació sobre l'ID del pacient en la BD GEO. El GEO_SERIE_ID ens dona informació sobre l'ID de l'experiment al que està associat i el GEO_PLATFORM_ID ens dona informació sobre la plataforma de microarray utilitzada.

Tal i com s'ha especificat anteriorment la segona part de l'estudi es basa amb l'anàlisi de les dades generades. Així doncs per tal d'avaluar la hipòtesi plantejada es treballa amb dos grups de pacients diferents. Pacients asmàtics i pacients sans. L'important però d'aquest estudi no es la malaltia que s'escull per obtenir els resultats, que en aquest cas ha estat l'asma, sinó la versatilitat que tenen el conjunt d'eines bioinformàtiques implementades. Tota aquesta arquitectura permet escollir quins són els pacients d'interès, seleccionar-los i emmagatzemar, transformar i integrar les dades per tal de poder extraure uns resultats consistents i recolzats per un estudi estadístic.

6.2 Anàlisi estadístic

La comparació de grups de pacients s'ha dut a terme mitjançant el test *T-Student* per a variables independents. S'ha utilitzat el *software RCmdr*. Els valors $p < 0.05$ s'han considerat estadísticament significatius.

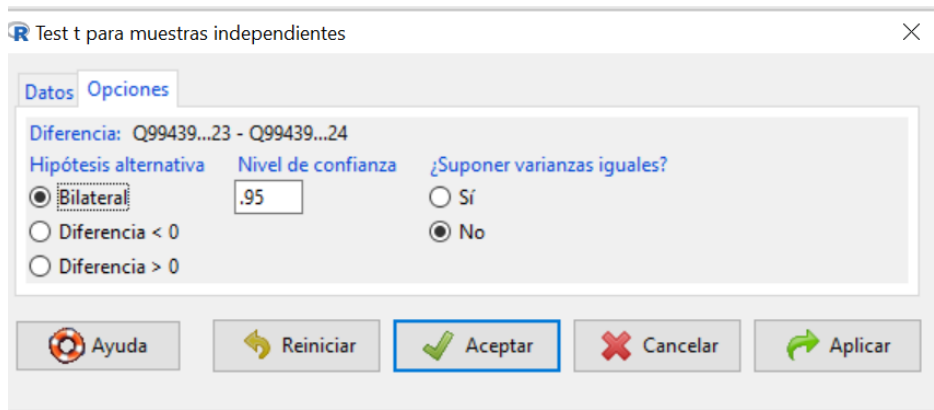
S'ha utilitzat un primer test *F de Fisher* per comprovar si es pot considerar que les variàncies entre les dues columnes de valors d'expressió gènica que s'estan comparant són iguals o diferents. Un cop realitzats els testos per als diferents grups es pot confirmar gràcies als valors de *p-value* obtinguts que la variància entre grups es pot considerar que no és igual. Així doncs, s'ha realitzat un primer estudi de la distribució mitjana dels valors d'expressió gènica de gens que codifiquen per un llistat de proteïnes efectores de l'asma en pacients amb asma vs pacients aleatoris. També s'ha dut a terme un darrer estudi de la distribució mitjana dels valors d'expressió gènica de gens que codifiquen per un llistat de proteïnes aleatòries en pacients amb asma vs pacients aleatoris. S'ha aplicat el test *T-Student*. S'ha treballat amb un nivell de confiança del 95% i per tant s'ha acceptat la hipòtesis alternativa per als *p-value* inferiors a 0.05.

μ_1 : Mitjana valors d'expressió gènica per al grup de pacients amb asma.

μ_2 : Mitjana valors d'expressió gènica per al grup de pacients aleatoris.

Hipòtesi nul·la: $\mu_1 = \mu_2$

Hipòtesi alternativa: $\mu_1 \neq \mu_2$



Imatge 6. Definició dels paràmetres per a la realització del test T per a dues mostres independents. S'ha treballat amb un nivell de confiança del 95%. És treballa amb la bilateral ja que el que ens interessa es saber si hi ha diferències significatives entre les mitjanes.

7. RESULTATS I DISCUSSIÓ

A continuació es mostren les dades generades executant les eines bioinformàtiques desenvolupades al llarg del projecte.

7.1 Resultats generats durant la fase Inicial

Tal i com s'especifica en l'apartat anterior s'han desenvolupat diferents eines bioinformàtiques per al tractament de les dades i la posterior generació de resultats. A continuació es mostren els resultats obtinguts al executar aquestes eines bioinformàtiques incloses en la fase inicial de l'estudi per al tractament i transformació de les dades.

Exemple d'implementació de la funció que genera els fitxers de proteïna presentada en l'apartat anterior :

```
axTprGeneratePatientsProteinExpressionList('GEO000000001','GEO00001000','vers5')
```

Executant aquesta línia de codi s'agafen les dades dels 1000 primers pacients. Des de el pacient amb identificador GEO000000001 fins el pacient amb l'identificador GEO000001000. El fitxer que s'utilitza per agafar les dades dels valors d'expressió gènica és el que conté la versió de normalització 5. Es poden executar moltes línies de codi com aquesta al mateix temps en diferents ordenadors i per tant de forma paral·lela es poden estar analitzant molts pacients. A més a més, per tenir un control dels pacients que ja han sigut revisats s'utilitza un fitxer anomenat *Proteins_LOG_file* on es copia l'identificador del pacient un cop s'han copiat totes les dades d'interès. Així doncs, abans d'analitzar un nou pacient es comprova que aquest no estigui en el fitxer de control. Un cop es finalitza l'execució d'aquesta funció s'obté el contingut que es mostra en les següents imatges.

A0AV96.txt
 A6NDV4.txt
 A8MVS5.txt
 B4DS77.txt
 H3BTG2.txt
 O75689.txt
 P0COL4.txt
 P01375.txt
 P02318.txt
 P52815.txt
 Q9HC82.txt
 Q96JM7.txt
 W5XKT8.txt
 P57077.txt

Imatge 7. Carpeta que conté un fitxer per proteïna. S'ha creat un total de 19540 fitxers, ja que aquest es el nombre de proteïnes totals estudiades en els diferents experiments de microarray emmagatzemats en la BD utilitzada.

P57077.txt: Bloc de notas

Archivo	Edición	Formato	Ver	Ayuda
GE001210398	0.000448			
GE001210399	0.092729			
GE001210400	0.001501			
GE001210401	0.000531			
GE001210402	0.178683			
GE001210403	0.464108			
GE001210404	0.000554			
GE001210405	0.099030			
GE001210406	0.001104			
GE001210407	0.000539			
GE001210408	0.000564			
GE001210409	0.000848			
GE001210410	0.000543			
GE001210411	0.000753			
GE001210412	0.000648			
GE001210413	0.000872			
GE001210414	0.000648			
GE001210415	0.000466			
GE001210416	0.000646			
GE001210417	0.000496			

Imatge 8. Fitxer corresponents a la proteïna P57077 que conté dos columnes. La primera correspon a l'identificador del pacient i la segona correspon al valor d'expressió del pacients per aquesta proteïna.

Com es pot observar en les imatges anteriors la creació d'aquest fitxers fa que els valors d'expressió gènics es presentin d'un forma molt més eficient i accessible. I que per tant consultant un sol fitxer es pot obtenir la informació de tots els pacients emmagatzemats en la base dades per a la proteïna d'interès.

S'ha desenvolupat també una eina bioinformàtica que retorna el percentatge de pacients que tenen per una proteïna, especificada per l'usuari, els valor d'expressió més alts o mes baixos segons s'especifiqui. D'aquest manera per exemple es poden identificar comorbiditats associades a diferents malalties. La funció conté els següents paràmetres d'entrada:

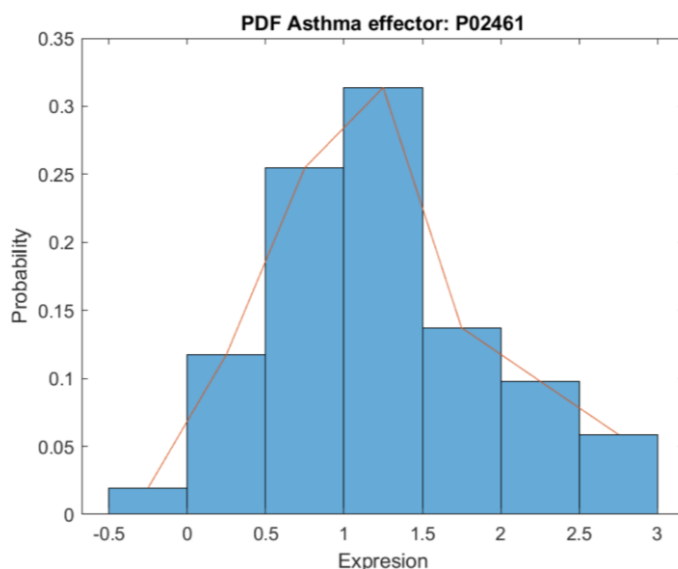
```
axTprGetPatientsWithProts({'A4UGR9'}, '+50', 1)
```

El nom de la funció es `axTprGetPatientsWithProts`, i els possibles paràmetres són els que es troben continguts en els parèntesis. El primer paràmetre fa referència a l'identificador de la proteïna de la que es vol treure la informació, en el cas de l'exemple es consultarà el fitxer que conté els valors d'expressió gènica del gen que codifica per a la proteïna amb l'identificador 'A4UGR9'. El segon paràmetre es correspon amb el percentatge de pacients que es vol que es retorni. Per exemple el valor '+50' retornarà el 50% dels pacients que tenen la proteïna especificada amb valors d'expressió més alts ja que s'utilitza el signe '+'. L'últim paràmetre especificat fa referència a si és vol generar el llistat d'identificadors de pacients en un fitxer d'Excel. Ja que des de l'empresa tot i executar les funcions amb el l'entorn de programació *Matlab* és treballa molt amb excel. Una altra possible entrada d'aquesta funció podria ser:

```
axTprGetPatientsWithProts({'A4UGR9', []})
```

En aquest cas la funció retornaria el 100% del llistat de valors d'expressió gènica del gen que codifica per a la proteïna especificada i al no aplicar el tercer paràmetre no es genera el fitxer Excel. En el cas de que en el segon paràmetre hi hagués especificat un '-25' és retorna el 25% de pacients amb valors d'expressió gènica mes baixos ja que s'utilitza el signe '-'.

Una altra eina bioinformàtica que s'ha desenvolupat durant l'estudi ha estat una funció que genera la gràfica PDF per estudiar el patró d'expressió d'una proteïna en els pacients seleccionats. Podria ser interessant estudiar el patró d'expressió gènica del gen que codifica per a la proteïna efectora d'asma amb ID P02461, tal i com es mostra en el següent gràfic.



Gràfic 1. Gràfic PDF de la proteïna amb identificador P0246. L'eix de les x mostra el rang de valor d'expressió gènica que en aquest cas es troba entre -0.5 i 3. L'eix de les y indica la probabilitat de que un pacient tingui el valor d'expressió gènica determinat. Les dades la gràfica s'han extret d'un grup de 51 pacients asmàtics.

Aquest gràfic podria servir per exemple per comprar-lo amb pacients asmàtics i pacients sans. Permetria veure la diferència de rang d'expressió gènica entre els dos grups de pacients. Veure per exemple, de forma molt clara, que en el grup de pacients asmàtics el valor d'expressió gènica més probable es que es trobi entre el rang d'1 i 1,5 i en canvi en pacients sans el valor que mostri la gràfica PDF generada per aquest grup de pacients.

7.2 Resultats generats durant la segona fase

Tal i com s'especifica en l'apartat de materials i mètodes l'estudi consta de dues parts. Les eines mostrades anteriorment formen part de la fase inicial i seguidament s'especifica la segona part de l'estudi, l'anàlisi de les dades generades en aquesta primera fase. A continuació s'especifica els passos que s'han seguit per obtenir els resultats amb pacients amb asma però es podria estudiar qualsevol paràmetres d'interès i seleccionar diferents grups de pacients a comprar. Cal destacar un cop més que l'objectiu de l'estudi es desenvolupar un conjunt d'eines bioinformàtiques versàtils, totalment parametritzades per assegurar que en qualsevol moment l'usuari pot redefinir els paràmetres per a l'estudi de diferents grups de pacients.

Es va treballar inicialment amb desenvolupar una nova eina bioinformàtica per crear taules de correlació i covariància entre els dos grups de pacients a estudiar, tal i com es mostra en les següents imatges. Però la majoria d'estudis publicats fins el moment confirmen que el coeficient de correlació de *Pearson* no és un bon indicador per estudiar patrons l'expressió de proteïnes. Aquest coeficient entén cada objecte com una variable aleatòria i mesura la semblança entre dos grups calculant la relació lineal entre les distribucions de les variables aleatòries corresponents. És àmpliament utilitzat com a mesura de semblança entre dades d'expressió genètica. Ara bé, hi ha estudis que confirmen que no és robust amb els *outliers*, ja que la correlació està dominada per aquests, en el ca de que existeixen. Així doncs, es va decidir treballar amb l'obtenció dels *p-value* aplicant el test T Student per estudiar el patró d'expressió gènica entre grups.

48x48 cell

	1	2	3	4	5	6	7	8	9	10	11
1 'Proteins'		'Q86V15'	'P42785'	'Q9NXA8'	'O94905'	'Q12980'	'Q659A1'	'Q8N0Z9'	'Q9Y343'	'Q15349'	'O76024'
2 'Q86V15'	1		0.9016	0.5513	0.5651	0.9865	0.8168	0.7796	0.7487	0.9698	0.9573
3 'P42785'	0.9016	1.0000		0.7028	0.7169	0.9253	0.8805	0.8378	0.8402	0.9711	0.9776
4 'Q9NXA8'	0.5513	0.7028	1		0.9967	0.5679	0.8385	0.9378	0.9652	0.6835	0.6870
5 'O94905'	0.5651	0.7169	0.9967	1		0.5805	0.8696	0.9339	0.9684	0.6975	0.7021
6 'Q12980'	0.9865	0.9253	0.5679	0.5805	1		0.8222	0.7911	0.7566	0.9724	0.9761
7 'Q659A1'	0.8168	0.8805	0.8385	0.8696	0.8222	1.0000		0.8942	0.9254	0.8929	0.8963
8 'Q8N0Z9'	0.7796	0.8378	0.9378	0.9339	0.7911	0.8942	1		0.9867	0.8572	0.8555
9 'Q9Y343'	0.7487	0.8402	0.9652	0.9684	0.7566	0.9254	0.9867	1		0.8456	0.8442
10 'Q15349'	0.9698	0.9711	0.6835	0.6975	0.9724	0.8929	0.8572	0.8456	1		0.9898
11 'O76024'	0.9573	0.9776	0.6870	0.7021	0.9761	0.8963	0.8555	0.8442	0.9898	1	
12 'Q5JVG2'	0.9056	0.9258	0.7933	0.8168	0.9071	0.9672	0.9134	0.9176	0.9555	0.9512	
13 'O94855'	0.5749	0.6980	0.8370	0.8781	0.5850	0.9367	0.7876	0.8526	0.6838	0.6989	

Imatge 9. Matriu de correlació entre els valors d'expressió gènica de gens que codifiquen per un llistat de proteïnes totalment aleatòries entre un grup de pacients amb asma i un grup de pacients aleatoris. La matriu es simètrica es a dir hi ha els mateixos valors per sota i sobre de la diagonal.

Per tant, per l'extracció de dades s'ha implementat una nova eina bioinformàtica que consisteix en una funció que al executar-la emmagatzema els valors d'expressió gènica dels gens que codifiquen per un llistat de proteïnes en dos grups de pacients diferents. O bé els valors d'expressió gènica d'un mateix grup de pacients per diferents llistats de proteïnes.

La funció s'executa amb la següent instrucció:

```
PCell=axGetRandomPatientProteinExpressionNOSortedValues('C:\Users\asthma.xlsx', 'C:\Users\IDPatientsAsma.xlsx', 1 , 1)
```

A l'esquerra del signe d'igualtat hi ha el nom de la variable que es correspon amb la matriu que es generà on cada columna representa una proteïna i cada fila un pacient. Així doncs en una fila hi tindrem representat el valor d'expressió gènica del gen que codifica per a la proteïna de la columna corresponent per un determinat pacient. S'ha muntat aquesta arquitectura per tal de generar quatre matrius diferents. Una primera matriu amb els valors d'expressió gènica d'un grup de pacients amb asma per proteïnes efectores de l'asma. Una segona matriu amb els valors d'expressió gènica d'un grup de pacients amb asma per proteïnes totalment aleatòries. Una tercera matriu amb els valors d'expressió gènica d'un grup de pacients aleatoris per proteïnes efectores de l'asma. I una última i quarta matriu on hi ha representats els valors d'expressió gènica d'un grup de pacients aleatoris per a proteïnes aleatòries. Cal especificar que el llistat de proteïnes aleatòries es genera un únic cop i s'utilitza al llarg de l'execució de la funció el mateix llistat de proteïnes aleatòries per tal de que siguin comparables els valors d'expressió entre grups.

El primer paràmetre de la funció implementada fa referència a la ubicació on es troba un fitxer que conté el llistat de proteïnes a estudiar. En aquest cas en el llistat hi ha els identificadors de les proteïnes efectores de l'asma. Ara bé, l'arquitectura està muntada per que en tot moment es pugui escollir quines proteïnes es vol estudiar. En cas de que en aquest paràmetre s'especifiqui un 0 el programa escollirà un llistat de proteïnes aleatòries. El segon paràmetre fa referència a la ubicació del fitxer que conté els identificadors dels pacients que es volen estudiar. En aquest cas es corresponen amb pacients amb asma, però es pot escollir quins són els pacients d'interès. En cas de que en aquest paràmetre s'especifiqui un 0 s'escolliran pacients totalment aleatoris. Els últims dos paràmetres poden tenir el valor 0 o 1. Si en el tercer paràmetre s'especifica un 1 es genera una gràfica PDF per cada proteïna estudiada i s'emmagatzema en la carpeta corresponent. En cas de que s'especifiqui un 0 no es generen els gràfics. Per últim, si s'especifica un 1 en el quart paràmetre la matriu s'emmagatzema en un arxiu Excel, si hi ha un 0 no es generarà aquest fitxer, i s'emmagatzema la informació en una variable del *Matlab*.

52x63 cell	1	2	3	4	5	6	7	8	9	10	11	1
1	'Q96SW2'	'Q9HCE6'	'Q8N4W6'	'Q9UQC2'	'P47890'	'Q86SJ2'	'P60520'	'P01903'	'Q5TA45'	'O15354'	'P09001'	'Q7L7'
2	-1.2875	-0.0962	0.0400	0.4734	1.6276	-1.6664	0.9762	5.7184	-1.4883	2.8454	-2.0254	0.1335
3	-2.2463	0.0045	-0.5074	0.1649	2.0641	-1.4549	0.1914	5.3547	-1.5438	0.7598	-1.9368	1.1852
4	-1.8416	-0.2921	-0.2338	0.4150	0.7393	-0.6752	-0.2514	5.8924	-2.1865	-0.3151	-2.1016	0.8645
5	-1.4488	-0.1253	-0.0963	0.5871	1.7148	-2.2540	-0.7095	5.6437	-2.4077	0.2346	-2.0622	1.2285
6	-1.4549	-0.6993	0.5422	1.4462	0.3143	-1.4864	0.6123	6.4636	-0.6632	0.1263	-1.9791	2.4035
7	-1.1798	-0.1453	-0.1575	0.1785	1.1273	-0.5719	-0.1366	5.9638	-2.2616	1.5757	-1.8671	1.6380
8	-1.1551	-0.7195	0.0015	-0.2099	-0.0805	-1.0983	-0.1007	7.0029	-1.6294	-0.4791	-1.5824	1.5585
9	-1.0802	-0.4034	-0.3162	0.0280	0.4462	-0.5319	-0.0592	6.2977	-2.0382	0.4044	-1.8210	1.0190
10	-1.4522	-0.4022	-0.0483	0.0801	0.2796	-1.0863	0.4064	5.3933	-1.1222	0.0694	-2.0112	1.5590
11	-1.1243	-0.6645	-0.2844	0.0822	1.6397	-0.8066	-0.0375	4.9060	-1.6867	0.0610	-1.6461	1.2895
12	-1.4529	-0.2763	0.3477	-0.0230	-0.4543	-1.7263	0.5379	8.2335	-1.2087	0.6674	-1.6141	1.2865
13	-1.8586	-0.1311	0.2092	0.2910	0.2530	-1.6120	0.2728	6.0846	-1.7968	0.3739	-1.7653	1.0855

Imatge 10. Matriu generada per la funció `axGetRandomPatientProteinExpressionNOSortedValues`. Cada columna representa una proteïna i cada fila un pacient. Aquestes dades s'han extret d'un llistat de 63 proteïnes aleatòries i 51 pacients amb asma.

Un cop generades les 4 matrius, s'ha treballat en implementat una nova funció per tal d'estructurar les dades de manera que puguin ser analitzades per l'entorn de programació Rcmdr d'anàlisi estadístic. Per això s'ha generat l'Excel que es mostra en la següent imatge:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Q96SW2	Q96SW2	Q9HCE6	Q9HCE6	Q9UQC2	Q9UQC2	P60520	P60520	P01903	P01903	Q5TA45	Q5TA45	P09001	P09001	Q5QJ66	Q5QJ66
2	-1.536527	-2.246268	-1.872217	0.004468	-0.569275	0.164883	-1.037743	0.191387	-0.056942	5.354675	-0.680542	-1.543783	1.639672	-1.936778	1.317556	0.1335
3	-0.139	-1.841599	-0.2415	-0.292056	-0.114	0.414968	0.2755	-0.251399	-0.014	5.89236	0.132	-2.186488	0.1815	-2.101584	0.313	0.1335
4	2.147482	-1.448824	3.041863	-0.125259	0.157446	0.58709	0.928689	-0.709465	2.239108	5.643686	0.072749	-2.407706	-0.074258	-2.062223	0.176916	0.1335
5	-0.211	-1.454863	0.61	-0.699325	0.554	1.446223	0.5805	0.612261	0.654	6.463594	0.639	-0.663231	-0.002	-1.979124	0.165	0.1335
6	-0.442286	-1.179773	-1.128382	-0.145307	-0.77742	0.178501	0.018506	-0.136639	-3.031226	5.963775	0.432909	-2.261612	0.997429	-1.867128	0.241572	0.1335
7	-0.158	-1.15506	0.098	-0.719525	-0.203	-0.209869	0.3635	-0.100707	-0.1995	7.00289	0.274	-1.629441	0.3405	-1.58239	-0.057	0.1335
8	2.264917	-1.080191	3.105092	-0.403424	0.794774	0.02803	1.074241	-0.059151	2.887513	6.297706	0.033376	-2.038152	-0.066395	-1.820993	0.520134	0.1335
9	-0.046841	-1.452188	-1.520053	-0.402201	-2.456585	0.080103	-0.167601	0.406426	0.062095	5.393271	0.975554	-1.122244	0.932568	-2.011212	0.761063	0.1335
10	0.563822	-1.124289	0.23847	-0.664478	-0.93131	0.082193	-0.083669	-0.037526	-2.951355	4.905978	-0.211535	-1.686723	0.310759	-1.646129	-0.395782	0.1335
11	-0.555491	-1.85864	0.893453	-0.131129	-0.166108	0.29104	-0.072993	0.272839	-5.371557	6.084573	0.272344	-1.796802	-0.226413	-1.765292	-1.199928	-0.1335
12	-0.063957	-1.13729	-0.142424	1.046891	0.298311	-0.081382	-0.000545	-0.328726	-0.140702	6.63454	-0.424437	-0.884437	-0.025968	-1.421071	0.420236	-0.1335
13	-0.732727	-3.22913	-0.669599	-0.201405	-0.322597	0.602929	-0.391689	-0.442818	0.58476	6.086792	0.947514	-1.375348	-0.058582	-1.865634	0.305398	-0.1335
14	-0.92512	-0.998764	-1.093065	0.114615	-1.377928	-0.212369	-0.337233	-0.547552	-0.273865	5.355595	0.293273	-2.667832	0.423874	-1.631153	0.73162	-0.1335
15	2.200196	-1.015032	2.962179	-0.431827	0.08077	0.665595	0.727946	-0.797397	2.023753	6.197311	0.17542	-0.554828	-0.083803	-1.691284	0.219769	-0.1335
16	2.214132	-1.371399	3.058598	-0.438182	0.660851	-0.00962	1.070493	0.339506	2.921818	5.912745	0.049494	-1.46769	0.031081	-1.555987	0.394763	-0.1335
17	2.079531	-1.351708	3.610326	-0.163081	0.245498	-0.106746	0.480286	0.260973	3.538678	7.562152	0.075598	-2.064118	0.09855	-1.359298	0.134966	-0.1335
18	-0.1185	-1.186514	-0.414	-0.481916	0.464	-0.349804	0.1105	-0.005868	0.178	6.689475	-0.335	-1.855007	0.234	-1.97006	0.484	-0.1335
19	2.302717	-0.98928	2.562464	-0.143017	0.263374	0.499441	0.813022	0.559244	2.77621	5.231389	0.214387	-1.488377	0.024494	-1.605065	0.358157	-0.1335
20	0.11378	-2.084694	-0.164094	0.035737	0.525161	-0.469857	0.075012	-2.701497	5.475656	-0.39865	-1.026013	-0.200579	-1.885217	-1.585977	-0.1335	-0.1335
21	-0.978647	-0.351189	-1.104243	-0.229086	0.111605	-0.425627	-0.17173	0.419745	-2.049789	6.956848	0.767273	-0.936854	0.573225	-0.887909	0.494008	-0.1335
22	-0.21802	-2.49477	-0.61635	-0.747076	0.746606	-0.827641	-0.036061	-0.843201	-4.08862	8.416248	0.200211	-0.822605	0.102605	-0.717637	0.858577	-0.1335

Imatge 11. Excel que conté les dades per aplicar l'anàlisi estadístic i estudiar si hi ha diferències significatives entre el grup de pacients sans i el grup de pacients amb asma en estudiar el patró d'expressió de proteïnes aleatòries. La primera fila correspon al llistat d'identificadors de les proteïnes aleatòries estudiades i s'agrupen les columnes dos a dos. En la primera columna hi ha els valors d'expressió gènica per a pacients asmàtics i el la segona els valors d'expressió gènica per a gens que codifiquen per a la mateixa proteïna en pacients aleatoris i així successivament.

Un cop es tenen les dades estructurades tal i com es mostra en la imatge 11 es pot aplicar l'estudi estadístic. El que s'estudia es si hi ha diferències significatives entre els valors d'expressió gènica que hi ha en les dues columnes que corresponen a la mateixa proteïna però per als dos grups de pacients a estudiar. Per exemple, les dues primeres columnes contenen la primera els valors d'expressió gènica del gen que codifica per a la proteïna amb identificador Q96SW2 per al grup de pacients amb asma i la segona els valors d'expressió gènica per al grup de pacients aleatoris. Es comparen les dues columnes per veure si hi ha diferència significativa en les patrons d'expressió gènica entre grups.

Les següents taules mostren els resultats de l'estudi estadístic realitzat amb les dades explicades anteriorment. En la Taula 1 es mostren els resultats de *p-value* obtinguts en estudiar el valor d'expressió gènica mitjà de gens que codifiquen per un llistat de proteïnes efectores d'asma en pacients en asma i en pacients aleatoris. En la majoria dels casos si que hi ha diferència significativa entre els pacients amb asma i els pacients aleatoris i per tant es confirma la hipòtesi inicial. És a dir la majoria de *p-value* obtinguts són inferiors a 0.05 i per tant és confirma la hipòtesi alternativa; la mitjana entre els dos grups es diferent. Això significa per tant, que efectivament el coneixement que es té sobre els efectors de l'asma en la literatura es pot observar en el patró d'expressió de GEO. És a dir aquelles proteïnes que es consideren efectores d'asma es comporten de forma significativament diferent per als pacients amb asma que al estudiar pacients aleatoris.

Cal tenir en compte però que algunes proteïnes del llistat proposat com efectores del asma no mostren diferències significatives entre els dos grups de pacients estudiats. Les proteïnes que no mostren diferències significatives entre els grups podrien ser interessants per revisar o intentar entendre perquè no apareix aquesta diferència en el patró d'expressió GEO.

És important també confirmar que aquesta diferència significativa no apareix al comparar proteïnes totalment aleatòries entre aquests dos grups de pacients. La segona taula mostra els resultats de *p-value* obtinguts al comprovar si hi ha diferències significatives al comparar els valors d'expressió de proteïnes aleatòries entre els dos grups de pacients. La majoria de proteïnes estudiades no presenten una diferència significativa entre els dos grups de pacients i per tant això reconforta els resultats obtinguts en la primera taula. Aquest fet apunta a que les diferències significatives que s'han trobat entre els dos grups de pacients en la primera taula no són aleatòries ja que al comparar el patró d'expressió entre els dos grups de pacients en proteïnes aleatòries no es troben aquestes diferències significatives en la majoria dels casos. Aquells casos en els que si que es troben diferències significatives al estudiar proteïnes aleatòries serien unes proteïnes interessants a estudiar. És podria buscar si són efectores d'alguna malaltia o en quins processos estan involucrades i podria demostrar comorbiditats associades en l'asma per exemple.

Taula 1. Resultats test T student per a variables independents, bilateral amb nivell de confiança del 95% i suposant variàncies entres grups no iguals. Valors d'expressió de proteïnes efectores de l'asma en pacients aleatoris vs valors d'expressió de proteïnes efectores de l'asma en pacients amb asma.

Proteïna	p-value	Diferencia significativa	Proteïna	p-value	Diferencia significativa
P1	2.265e-16	Si	P20	0.2091	No
P2	1.122e-10	Si	P21	0.6346	No
P3	0.2393	No	P22	0.01811	Si
P4	0.07637	No	P23	0.07839	No
P5	1.346e-7	Si	P24	3.591e-14	Si
P6	0.2686	No	P25	2.41e-6	Si
P7	2.252e-8	Si	P26	0.04408	Si
P8	1.595e-6	Si	P27	0.6348	No
P9	4.565e-4	Si	P28	0.009754	Si
P10	3.474e-10	Si	P29	0.4212	No
P11	0.08017	No	P30	0.002134	Si
P12	0.01309	Si	P31	0.8962	No
P13	8.618e-7	Si	P32	0.03936	Si
P14	2.481e-12	Si	P33	0.3028	No
P15	0.000593	Si	P34	0.0001498	Si
P16	0.004669	Si	P35	2.521e-9	Si
P17	2.069e-13	Si	P36	0.9581	No
P18	0.005439	Si	P37	1.328e-11	Si
P19	0.01717	Si			

Taula 2. Resultats test T Student per a variables independents, bilateral amb nivell de confiança del 95% i suposant variàncies entres grups no iguals. Valors d'expressió de proteïnes aleatòries en pacients aleatoris vs valors d'expressió de proteïnes aleatòries en pacients amb asma.

proteïna	ID proteïna	p-value	Diferencia significativa	proteïna	ID proteïna	p-value	Diferencia significativa
P1	P09001	0.2439	No	P20	P49354	0.6632	No
P2	O00584	0.7053	No	P21	P51970	0.002891	Si
P3	Q96RR4	0.1795	No	P22	O94985	0.054	No
P4	P60520	0.3213	No	P23	Q9Y6M1	0.3178	No
P5	Q567U6	0.15	No	P24	P16083	0.2926	No
P6	P01903	0.3251	No	P25	Q9BZ67	0.9561	No
P7	P24821	0.6652	No	P26	P55347	0.3107	No
P8	Q14162	0.4454	No	P27	Q9NPA5	0.6598	No
P9	Q5TA45	0.3171	No	P28	P14314	2.2e-16	Si
P10	Q5QJE6	0.3113	No	P29	P00813	1.183e-12	Si
P11	Q99439	0.0017	Si	P30	Q9ULE0	0.846	No
P12	Q9Y5V0	0.1102	No	P31	Q7Z2T5	1.215e-4	Si
P13	P23511	0.2995	No	P32	P42898	0.9414	No
P14	O75324	0.006124	Si	P33	Q9BTD3	0.008492	Si
P15	Q9BW27	0.02649	Si	P34	Q9UPY3	0.001112	Si
P16	Q9UQC2	0.32	No	P35	O75094	0.8374	No
P17	Q5VT52	2.578e-6	Si	P36	Q96SW2	0.2775	No
P18	Q12768	0.007827	Si	P37	Q9HCE6	0.3214	No
P19	P31639	0.1729	No				

8. CONCLUSIONS

Tal i com es mostra en els resultats anteriors la majoria de proteïnes efectores d'asma mostren diferències significatives al comparar el patró d'expressió entre els dos grups de pacients estudiats, sans i amb asma. Així doncs es confirma per a la majoria de proteïnes del llistat proposat com efectores de l'asma que realment es comporten de forma significativament diferents amb pacients aleatoris que al estudiar-les amb pacients amb asma. Ara bé dotze de les trenta-set proteïnes proposades com efectores d'asma no presenten diferències significatives entre els dos grups de pacients i per tant podrien ser proteïnes interessants a revisar i intentar entendre perquè no apareixen aquestes diferències en el patró d'expressió.

Pel que fa al segon anàlisi realitzat, on es comprova que realment aquesta diferència significativa no apareix al estudiar proteïnes totalment aleatòries, els resultats són força bons ja que la majoria de proteïnes estudiades no mostren aquest diferència entre grups de pacients. Al comparar el valor d'expressió gènica de gens que codifiquen per un llistat de proteïnes totalment aleatòries entre els dos grups de pacients, vint-i-sis de les trenta-set proteïnes estudiades no tenen uns valors d'expressió gènica significativament diferents. Això mostra que les diferències significatives que s'han trobat en el primer estudi no es troben en la majoria de les proteïnes aleatòries. Ara bé, cal tenir en compte que onze de les trenta-set proteïnes estudiades si que han mostrat diferències significatives entre els dos grups de pacients estudiats i que per tant podrien ser unes bones candidates a estudiar en futurs estudis. Entre la funció que fan aquestes proteïnes si els processos en els que estan implicades podria donar a la llum comorbiditats associades a l'asma que no s'han estudiat fins al moment, o bé confirmar alguns dels estudis ja realitzats.

En general les conclusions extretes en l'estudi realitzat són bones ja que es confirma el coneixement que es té sobre els efectors d'asma en els estudis realitzats fins al moment. I a més es proposa un llistat d'onze proteïnes interessants a estudiar en futurs estudis.

Pel que fa referència a la resta de funcions implementades les conclusions extretes són bones també, ja que la funció de rastreig de pacients que genera un fitxer per proteïna s'ha executat correctament i agilitza de forma molt clara la consulta de dades d'expressió gènica. La implementació de la segona funció de rastreig de pacients que retorna el percentatge de pacients amb els valors d'expressió més alts a més baixos segons s'especifiqui s'implementa també correctament i retorna les resultats desitjats facilitant l'estudi d'aquets i la generació de posteriors resultats.

9. BIBLIOGRAFIA

1. Makady A, de Boer A, Hillege H, Klungel O, Goettsch W. What Is Real-World Data? A Review of Definitions Based on Literature and Stakeholder Interviews. *Value Heal.* 2017;
2. Wise J, Möller A, Christie D, Kalra D, Brodsky E, Georgieva E, et al. The positive impacts of Real-World Data on the challenges facing the evolution of biopharma. *Drug Discovery Today.* 2018.
3. Makady A, Ham R ten, de Boer A, Hillege H, Klungel O, Goettsch W. Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value Heal.* 2017;
4. Toro-Domínguez D, Martorell-Marugán J, López-Domínguez R, García-Moreno A, González-Rumayor V, Alarcón-Riquelme ME, et al. ImaGEO: Integrative gene expression meta-analysis from GEO database. *Bioinformatics.* 2019;
5. Bobolea I, Arismendi E, Valero A, Agustí A. Early life origins of asthma: A review of potential effectors. *Journal of Investigational Allergology and Clinical Immunology.* 2019.
6. Longley PA, Adnan M. Geo-temporal Twitter demographics. *Int J Geogr Inf Sci.* 2016;
7. Reuter JA, Spacek D V., Snyder MP. High-Throughput Sequencing Technologies. *Molecular Cell.* 2015.
8. Min S, Lee B, Yoon S. Deep learning in bioinformatics. *Briefings in bioinformatics.* 2017.
9. Wildey MJ, Haunso A, Tudor M, Webb M, Connick JH. High-Throughput Screening. In: *Annual Reports in Medicinal Chemistry.* 2017.
10. NCBI Gene Expression Omnious. <https://www.ncbi.nlm.nih.gov/geo/> (consulta març 2020)
11. NCBI Gene Expression Profile <https://www.ncbi.nlm.nih.gov/geoprofiles> (consulta març 2020)

10. AUTOAVALUACIÓ

Aquest treball m'ha aportat moltes coses i m'ha fet créixer tant a nivell professional com personal. Al realitzar un treball d'aquestes dimensions on hi dediques tant d'esforç i inverteixes tant temps a la cerca de bons resultats fa que tingues que aprendre també a sobreportar situacions on l'estrès i la frustració dominen la situació. L'important però es poder superar aquestes situacions i entendre que es normal que en algun punt del projecte apareguin aquets sentiments però que tot el treball sempre acaba sortint a la llum i la satisfacció personal que t'aporta és molt superior.

Al llarg de l'estudi realitzat m'he adonat de la gran dificultat que hi ha a l'hora d'intentar extreure uns bons resultats. Tot el tractament previ que hi ha per reordenar les dades, entendre bé el seu origen i intentar generar resultats reals tenint en compte tots els paràmetres.

Gràcies a la realització del treball he pogut posar en pràctica molts dels coneixements adquirits durant el grau. He pogut treballar en tot moment compaginant dos de les meves gran passions, la biotecnologia i la enginyeria informàtica. He sentit molta llibertat a l'hora de poder crear noves solucions a problemes tant presents en la societat actual com es la gestió de la gran quantitat d'informació que s'està generant constantment. M'he sentit molt recolzada per grans professionals en l'àmbit de la bioinformàtica i he sentit que he après molt d'ells.

M'ha donant també una gran satisfacció el fet de veure tot el coneixement que arribem a adquirir al llarg dels anys d'estudi dels graus. I que tot i que ens queda encara molt per aprendre, tenim un bon punt de partida i unes bones bases per emprendre la nostra vida laboral.

Em sento molt afortunada de l'oportunitat que he tingut de treballar en un empresa del sector bioinformàtic i sento que he aprofitat molt bé l'oportunitat. La valoració global de la realització del treball és molt bona. Aquest treball representa molta constància en l'aprenentatge i la gran aplicabilitat i poder incalculable que té la bioinformàtica en l'àmbit de la investigació. Es una mostra de la importància del treball en equip en la comunitat científica.

ANNEX 1: LLISTA D'ABREVIATURES

<i>Real World Data</i>	RWD
<i>Gene Expression Omnibus</i>	GEO
<i>Hight-Throughput</i>	HT
Agència europea de Medicaments	EMA
Assajos clínics aleatoris	ECA
Informació mínima sobre un experiment de microarray	MIAME
<i>The Patient Repository</i>	TPR
Probability Density Funtion	PDF
Base de dades	BD
<i>Randomized Controlled Trial</i>	RCT