Estudio de mecanismos comunes y específicos en patologías asociadas a la Diabetes tipo II (T2D) utilizando aproximaciones de Biología de Sistemas. Construcción de *Diseasome* (*Human Disease Network*)".

M.ª Begoña Hernández Olasagarre
Máster Bioinformática y Bioestadística, 2015-2017
Área Estadística y Bioinformática

Director proyecto: Guerau Fernàndez Isern, PhD
Consultora externa: Susana Kalko, PhD

2016-12-26

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Estudio de mecanismos comunes y específicos en patologías asociadas a la Diabetes tipo II (T2D) utilizando aproximaciones de Biología de Sistemas. Construcción de Diseasome (Human Disease Network)"*
Nombre del autor:	*M.ª Begoña Hernández Olasagarre*
Nombre del consultora:	*Susana Kalko Witruk*
Nombre del PRA:	*Guerau Fernàndez Isern*
Fecha de entrega (mm/aaaa):	12/2016
Titulación::	*2015-2017*
Área del Trabajo Final:	*El nombre de la asignatura de TF*
Idioma del trabajo:	***CASTELLANO***
Palabras clave	*Metabolito, Similaridad & Diseasome.*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados y conclusiones del trabajo.*

El objetivo de este trabajo es el estudio de mecanismos comunes y específicos en patologías asociadas a la Diabetes tipo II (T2D, Diabetes Millitus, non-insulin dependent) utilizando aproximaciones de Biología de Sistemas. Este estudio pretende construir diseasomes basados en genes y en la estructura química de metabolitos.

Se he elegido la enfermedad T2D, por su prevalencia en la sociedad, y porque se quiere profundizar en el estudio de enfermedades que son comorbilidades a la T2D (la comorbilidad describe el efecto de una enfermedad/es en un paciente cuya enfermedad primaria es otra distinta).

Los diseasomes se han generado teniendo en cuenta que sólo hay 41 enfermedades comunes etiquetadas tanto en CytoScape/DisGeNet (base CURATED de genes – enfermedades) como en HMDB (metabolitos – enfermedades).

El diseasome de genes y enfermedades relacionadas con T2D se genera utilizando la aplicación de CytoScape/DisGeNet dando un diseasome con 83 genes únicos.

Para generar el diseasome de metabolitos, no sólo se tienen en cuenta los metabolitos con una relación directa con T2D y sus enfermedades relacionadas sino con aquellos metabolitos que son similares (cutoff > 0.98) a los metabolitos diferenciados. Para ello se realiza un cálculo de similitud entre los primeros metabolitos y la base de datos de HMDB que contiene 40844

metabolitos. En el diseasome aparecen 126 relaciones entre las 41 enfermedades a través de metabolitos comunes.

Aunque uno de los objetivos principales, generar diseasomes, se ha conseguido, no se ha podido profundizar en la aproximación de Biología de Sistemas.

Abstract (in English, 250 words or less):

The objective of this work is the study of common and specific mechanisms in diseases associated to Diabetes type II (T2D, Diabetes Millitus, non-insulin dependent) using approaches of Systems Biology. This study aims to construct gene-based diseasomes and on chemical structure of metabolites.

T2D disease has been chosen because of its prevalence in society, and because we want to go deeper into the study of diseases that are comorbidities to T2D (comorbidity describes the effect of a disease on a patient whose primary disease is a different one).

The diseasomes have been generated taking into account that there are only 41 common diseases labeled in both CytoScape / DisGeNet (CURATED base of genes - diseases) and in HMDB (metabolites – diseases).

The diseasome of genes and diseases related to T2D is generated using the application of CytoScape / DisGeNet giving a diseasome with 83 unique genes.

In order to generate the metabolite diseasome, it are considered metabolites with direct relationship with T2D and related diseases and with the metabolites that are similar (cutoff> 0.98) to the differentiated metabolites. To do this, we performed a similarity calculation between the first metabolites and the HMDB database that containing 40844 metabolites. In the diseasome, there are 126 relationships between the 41 diseases appear.

Although one of the main objectives, to generate diseasomes, has been achieved, it has not been possible to deepen the approach of Systems Biology.

Índice

Lista de figuras y tablas

1.Introducción

1.1. Contexto y justificación del Trabajo

Las enfermedades crónicas no-contagiosas (NCD - Non-communicable chronic diseases), incluyendo diabetes, pre-diabetes, obesidad, hígado graso no-alcohólico, coronarias, etc., son parte importante de los problemas de salud globales de este siglo, y son causadas por complejas interacciones entre los genes y el medio ambiente.

Por otra parte, se ha propuesto, que las NCD deberían ser consideradas como la expresión clínica de un continuo de mecanismos patogénicos complejos que involucran a diversas moléculas, células y tejidos, que son en última instancia, responsable de la presentación fenotípica en los pacientes individuales. Esta premisa permite pensar que la investigación de la patobiología de las NCD debe ser considerada desde una perspectiva integral, multidimensional, como son las aproximaciones de **Systems Biology** & **Network Biomedicine** [1].

En la actualidad, hay una gran variedad de iniciativas "*Whole Diseasomes*" que demuestran la cercanía entre algunas enfermedades en base a diferentes características en común, desde genes [2] a *pathways* [3]. La novedad de este trabajo es la utilización de metabolitos para crear un diseasome que permitirá evaluar las relaciones entre las distintas enfermedades de una manera más robusta y precisa en relación a las *networks* anteriormente propuestas. Bases de datos como **Malacards** [4] o **ImmunoBase** [5] son fuentes de información importantes a fin de relacionar comorbilidades.

Este Trabajo fin de Máster (TFM) se focaliza en la diabetes tipo II (T2D) y en las enfermedades relacionadas con la diabetes. La T2D (también llamada no dependiente de la insulina) se debe al uso ineficaz en el cuerpo de la insulina y es en gran parte el resultado del exceso de peso corporal y la inactividad física. La prevalencia mundial de T2D es de 8.5% en 2014 y el número de personas con diabetes ha aumentado de 108 millones en 1980 a 422 millones en 2014 (generando una elevada presión económica a los estados). Además, para la T2D se conoce que es una causa importante de ceguera, insuficiencia renal, ataques cardíacos, accidente cerebrovascular y amputación de miembros inferiores (comorbilidades) [6]. Conocer el origen común de las enfermedades relacionadas, ayudaría a gestionar de manera más eficaz los tratamientos de las enfermedades relacionadas.

Por otra parte, el TFM pretende ser un proyecto con capacidad para obtener, integrar y analizar conjuntos de datos complejos a partir de múltiples fuentes experimentales en bases de datos abiertas (genómica y metabolómica), utilizando herramientas interdisciplinarias, bajo la aproximación de *Systems Biology*. El resultados de este proyecto tienen el potencial de desentrañar mecanismos comunes, específicos y novedosos que vinculan la presentación clínica de la diabetes y las enfermedades relacionadas, por lo tanto, identificar posibles nuevos objetivos preventivos y terapéuticos.

1.2. Objetivos del Trabajo

El objetivo principal de este proyecto es evaluar la subred de comorbilidades asociadas a la diabetes tipo II (T2D) a partir de datos *metabolómicos* de bases de datos libres (*Malacards* [4], HMDB [7], etc) y compararlo con el generado por genes diferenciados.

1.2.1. Objetivos generales:

Montaje de la red de regulación entre los genes y metabolitos en NCD de la familia de la T2D.

1.2.1.1. Identificación de los genes claves y metabolitos en la patogénesis de los NCD.
1.2.1.2. Identificación de los más importantes *pathways* metabólicos involucrados en NCD.
1.2.1.3. Identificación de los mecanismos patogénicos compartidos entre las diferentes NCD.

1.2.2. Objetivos específicos:

1.2.2.1. Seleccionar y establecer la calidad de las Bases de Datos que se utilicen.
1.2.2.2. Establecer el perfil de expresión génica y de metabolitos para cada enfermedad seleccionada.
1.2.2.3. Generar una red de interacción que contiene tanto datos de metabolitos como de los genes.
1.2.2.4. Analizar las redes de interacción para identificar *hubs*, grupos, nodos de puentes y otros elementos críticos.
1.2.2.5. Realizar el análisis de enriquecimiento de vía en los clústeres identificados en las redes de interacción.
1.2.2.6. Comparar las redes de T2D para identificar módulos compartidos o elementos críticos.
1.2.2.7. Presentar un documento con todo el trabajo realizado.

1.3. Enfoque y método seguido

Es sabido que de todas las ciencias "*ómicas*", la que se dedica al estudio global de los metabolitos (metabolómica) en procesos patológicos tiene el potencial de desentrañar los verdaderos mecanismos que se producen en el organismo, de saber cuáles son las proteínas verdaderamente implicadas de toda la maquinaria funcional de dicho organismo. Pero se trata de técnicas muy complejas y con mucho ruido a nivel de exploración "*non-targeted*", y por ello es de gran utilidad predecir posibles productos y metabolitos utilizando bases de datos de calidad y bien curadas.

La hipótesis de trabajo es que un subconjunto bien controlado de enfermedades puede tener una topología similar como *Diseasomes (human disease network)* basados en los *metabolitos y* permitiría la identificación de nuevos biomarcadores de la enfermedad T2D y las enfermedades relacionadas

a T2D. Se tiene la intención de explorar diferentes niveles de metabolitos químicamente relacionados, pero quizás no detectados en los experimentos anteriores, con el objetivo de ampliar el universo de posibles metabolomas completos para las enfermedades de interés. Los resultados de este proyecto tienen el potencial de desentrañar nuevos mecanismos subyacentes de la presentación clínica de las enfermedades asociadas con la diabetes, y en consecuencia, la identificación de posibles nuevas dianas terapéuticas.

1.4. Planificación del Trabajo.

1.4.1. **Tareas y calendario.** Se adjunta un Diagrama de *Gantt* con el listado de las tareas e hitos que se han programado realizar durante el TFM. El programa de planificación que se utiliza es el paquete de R *DiagrammeR* (Fig1) y organizado en función de la entrega de las PEC.

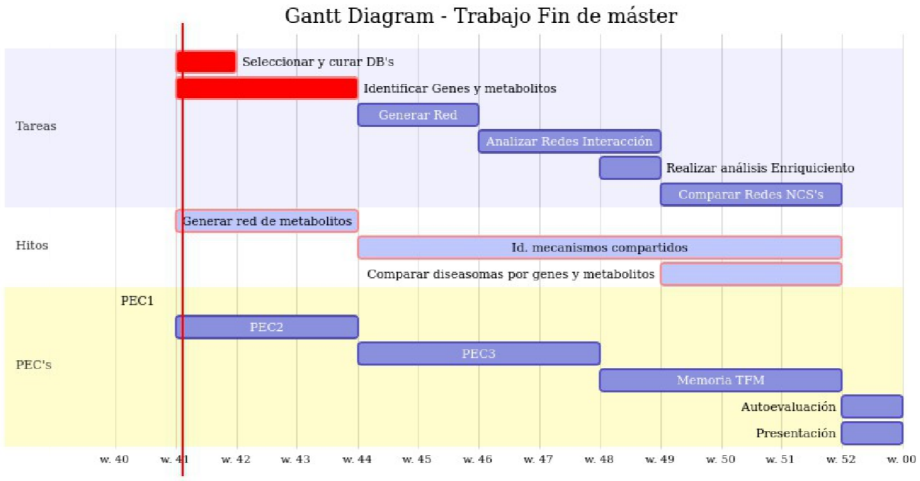

Fig.1: Diagrama de Gantt con el listado de tareas que se programan realizar durante el TFM.

1.4.2. **Hitos.** En este caso los hitos son:

1.4.2.1. Identificación de los genes claves y metabolitos en la patogénesis de los NCD. Generar red de metabolitos (PEC1).
1.4.2.2. Identificación de los mecanismos patogénicos compartidos entre las diferentes NCD (PEC2).
1.4.2.3. Comparar diseasomes de genes y metabolitos. Análisis de enriquecimiento (PEC3)
1.4.2.4. Entrega de memoria del TFM (PEC4)

1.5. Recursos necesarios.

1.5.1. Hardware. Servidor de *10 cores, 2.40GHz, 30 MB CPU cache y 192 GB RAM*
1.5.2. Software:
1.5.2.1. *R* program [8]
1.5.2.2. *Cytoscape* program [9]

1.5.2.3. *Digesnet* program [10]

1.6. Breve sumario de productos obtenidos

1.6.1. Listado, en forma de pares de enfermedades, con el número de metabolitos comunes a ambas enfermedades para generar el diseasome de metabolitos.
1.6.2. Listado, en forma de pares de enfermedades, con el nombre de metabolitos comunes a ambas enfermedades para realizar análisis de enriquecimiento de metabolito a pathway .
1.6.3. Listado, en forma de pares de enfermedades, con el número de genes comunes a ambas enfermedades para generar el diseasome de genes.
1.6.4. Listado, en forma de pares de enfermedades, con el nombre de genes comunes a ambas enfermedades para realizar análisis de enriquecimiento de Gene Ontology.

1.7. Breve descripción de los otros capítulos de la memoria.

El capítulo 2 de la memoria se llama Estructura del proyecto y en el se describe de manera ordenada el trabajo realizado en el Trabajo Fin de Máster. El capítulo 2 es el cuerpo de la memoria. También se describen los inconvenientes surgidos durante el proceso, así como las toma decisiones que se han ido tomado para solventar los problemas.

2. Estructura del proyecto.

2.1. Análisis inicial

2.1.1. Seleccionar y establecer la calidad de las Bases de Datos (DB) que se utilicen

2.1.1.1. *DisGeNet* [10]

Para obtener los diferentes *diseasomes* de genes, se han de encontrar las relaciones de los genes con enfermedades. En esta información se obtiene de la base de datos *DisGeNet* [10].

DisGeNET es una plataforma de descubrimiento de integración de la información sobre las asociaciones entre genes y enfermedades (GDA) de varias fuentes públicas de datos y la literatura. Integra asociaciones entre genes y enfermedades humanas (*GDA*) de diversas bases de datos curadas por expertos y asociaciones derivadas de text-*mining*, incluyendo enfermedades mendelianos, complejas y ambientales. La integración se realiza por medio del mapeo de vocabulario entre los genes y la enfermedad y utilizando el tipo de asociación ontológica descrito por *DisGeNET.* La versión actual (*v4.0 DisGeNET*) contiene 429,036 asociaciones, entre 17.381 genes y 15,093 enfermedades, trastornos y fenotipos clínicos humanos o anormales, y 72,870 asociaciones variante de enfermedad (VDA), entre 46,589 SNPs y 6.356 enfermedades y fenotipos. Dado el gran número de GDAs compilados en *DisGeNET*, también se ha desarrollado una puntuación a fin de clasificar las asociaciones basadas en la evidencia de apoyo.

En este TFM sólo se trabaja con la base de datos denominada CURATED que está compuesta por: 1) UNIPROT: UniProt / SwissProt es una base de datos que contiene información curada sobre secuencia de la proteína, estructura y función (El Consorcio UniProt, 2014). 2) CTDTM: El comparativo Toxicogenómica DatabaseTM contiene información acerca de las relaciones establecidas de forma manual entre genes y enfermedades, con especial atención en la comprensión de los efectos de las sustancias químicas ambientales en la salud humana. 3) CLINVAR: ClinVar es un archivo de acceso libre, pública de los informes de las relaciones entre las variantes y fenotipos de relevancia médica, sin evidencia de respaldo. 4) Orphanet: Orphanet: una enfermedad rara en línea y la base de datos de medicamentos huérfanos (INSERM©1997) es el portal de referencia para la información sobre las enfermedades raras y medicamentos huérfanos, para todos los públicos. 5) GWAS del catálogo: El catálogo NHGRI-EBI GWAS es una calidad controlada, analizada manualmente, colección literatura derivados de todos los estudios publicados en todo el genoma de asociación ensayar al menos 100.000 SNPs y todas las asociaciones SNP-rasgo con los valores de p <1,0 x 10-5. Esto determina que la búsqueda con DisGeNet en este proyecto se realiza sobre 9362 genes, 7607 enfermedades y 32834 GDAs.

2.1.1.2. *Human Metabolome Database (HMDB)*

La selección de los metabolitos se basa en el criterio de que estos presenten una relación evidente con las enfermedades mediante datos clínicos. Esta información está disponible en *Human Metabolome Database (HMDB)* [7]. *S*e ha seleccionado esta DB porque *HMDB* es una base de datos electrónica de libre acceso que contiene información detallada sobre **metabolitos** que se han localizado en el cuerpo humano. La base de datos está diseñada para contener y/o enlazarse con tres tipos de datos: 1) datos químicos, 2) datos clínicos y 3) la biología molecular / bioquímica, toda esta información se guarda en un fichero de estructuras (SDF) [11]. Este fichero de ha descargado para su posterior uso y se comprueba que contiene 42003 estructuras químicas.

La base de datos contiene 42.003 entradas de metabolitos, incluyendo metabolitos solubles en agua y lípidos, así como metabolitos que serían considerados como abundantes (> 1 microM) o relativamente raros (<1 nM). Además, 5.701 secuencias de proteínas están ligadas a estas entradas de metabolitos. Cada entrada *MetaboCard* contiene más de 110 campos de datos con 2/3 de la información que se dedica a datos clínicos químicos / y el otro tercio dedicado a la enzimática o datos bioquímicos. Muchos campos de datos se enlazará con otras bases de datos (*KEGG, PubChem, MetaCyc, ChEBI, PDB, UniProt, y GenBank*) y una variedad de estructura y de visualización vía applets. Cuatro bases de datos adicionales, *DrugBank, T3DB, SMPDB y FooDB* son también parte de la suite HMDB de bases de datos. La consulta de texto compatible con una búsqueda más sofisticada de texto de la parte de texto de *HMDB*. La secuencia de búsqueda permite a los usuarios realizar búsquedas de secuencias *BLAST* de las más de 5.701 secuencias de genes y proteínas contenidas en *HMDB*. Tanto las consultas individuales y múltiples secuencias *BLAST* son compatibles. MS/Search permite a los usuarios enviar archivos de espectros de masas (formato MoverZ) que se registraron en la biblioteca de espectros MS/MS del *HMDB*. Esto permite la identificación de metabolitos a partir de mezclas a través de espectroscopia de MS/MS y espectroscopía de RMN.

En HMDB, además, hay disponible información sobre 397 enfermedades diferentes.

Por otra parte, los cálculos de pruebas para la selección del *package* se realizaron con dos *subsets* de HMDB, de 7 y 672 metabolitos respectivamente. Con los dos *subsets*, se han realizado todos los cálculos para la evaluación de los *packages* e implementación de los algoritmos. Todos los procesos se pudieron completar de manera adecuada.

2.1.2. **Similaridad entre los compuestos: *R package selection*.**

Para poder realizar el *diseasome* de los metabolitos, se utilizan los metabolitos identificados en HMDB en T2D y de las enfermedades relacionadas con T2D que han surgido al analizar los genes. También se tiene en cuenta un *subset* de estructuras química elegidas por su semejanza con los metabolitos relacionados con las enfermedades en HMDB. El estudio de semejanza entre los estructuras químicas relacionadas con la T2D (y las enfermedades relacionadas) y la base de datos completa de HMDB (42003 compuestos) se puede realizar con diferentes *packages* de R. Los *packages* de *R* que se evalúan son: *Eirf* [12], *fmcsR* [13] *Rchemccp* [14] & *ChemmieR* [15].

La selección del *package* de R es fundamental para el resultado de este estudio. A continuación se describen los criterios de calidad que se utilizan para la selección del package para realizar el estudio de similaridad.

2.1.2.1. HMDB suministra en su portal la posibilidad de realizar estudios de similaridad, a través de la applet *ChemQuery:Search by structure* [16] . Se seleccionan dos moléculas de referencia: HMDB00011 & HMDB00153 y se tiene en cuenta los resultados que suministra HMDB para estas dos estructuras: valor de similaridad y orden relativo de las moléculas más similares. No se ha podido averiguar el método de similaridad que se ha utilizado en HMDB.

Para HMDB00011, dentro de las 30 estructuras más similares en la DB completa de HMDB, hay 14 estructuras que se encuentran en el *subset* seleccionados de 672 metabolitos.

Para HMDB00153, dentro de las 30 estructuras más similares en la DB completa de HMDB, hay 18 estructuras que se encuentran en el *subset* seleccionados de 672 metabolitos.

2.1.2.2. Tiempo de cálculo. Los *packages* que supongan un tiempo de cálculo elevado y no permitan avanzar en el proceso de manera adecuada no podrán ser seleccionados, aunque el valor de similaridad/orden sea adecuado. El TFM se encuentra muy ajustado en el tiempo.

2.1.3. Selección del packages más adecuado para el estudio de similaridad.

2.1.3.1. *eiR*: *Accelerated Similarity Searching of Small Molecules* [12].

El paquete *eiR* es útil en la búsqueda de similaridad entre estructuras en conjuntos de datos con moléculas pequeñas y en gran cantidad usando un enfoque de incrustación e indexación (es una versión acelerada). El cálculo tiene en cuenta la distancia que hay entre dos estructuras, y seguidamente lo asocia a la similaridad entre las

estructuras. Las pruebas de este package se realizan teniendo en cuenta las opciones *"ap" atompair y "fp" fingerprint.*

Cuando se evalúa la similaridad de la DB de 672 compuestos, y analizando las dos moléculas de referencia, y considerando las dos opciones (*atompair & fingerprint*), el método queda descartado porque tanto para HMDB000011 como para HMDB00153, el orden de los metabolitos sólo coinciden en los tres y dos primeros, respectivamente. El resto de metabolitos no coinciden con el orden suministrada por *HMDB-ChemQuery*. Eso si, el tiempo de cálculo sería adecuado para este proyecto.

2.1.3.2. *fmcsR*: *Mismatch Tolerant Maximum Common Substructure Searching* [13]

El *package fmcsR* introduce el concepto de Máximo Común eficiente Subestructura (MCS) y lo combina con una estrategia que permite el desajuste de átomos y/o enlaces entre subestructuras compartidas entre dos moléculas pequeñas. Esta manera de trabajar, se utiliza para encontrar compuestos con semejanzas estructurales débiles.

Al realizar el cálculo de similaridad con la DB de 672 compuestos ocurre lo mismo que con el package *eIR*: para HMDB00011 hay sólo cuatro metabolitos que coinciden con el orden suministrada por *HMDB-ChemQuery.* Para el caso de HMDB00153, el orden de los metabolitos las moléculas que son semejantes a las moléculas de referencia tampoco se ordenan como los de referencia. El tiempo de cálculo sería adecuado para el proyecto. Este método queda descartado para el proyecto.

2.1.3.3. *Rchemccp*: *Similarity measures for chemical compounds* [14]

El paquete *Rchemcpp* implementa el graph *kernel* y extensiones del mismo, los núcleos de *Tanimoto*, los núcleos de los gráficos, los núcleos de *Pharmacophore* y *3D* para medir la similitud de las moléculas, en total siete núcleos diferentes. Los resultados de las pruebas realizadas con la DB de 672 compuestos son los siguientes:

a kernelType="sizebased" – tiempo = 50 segundos - orden similaridad de los metabolitos:10/14 HMDB00011 – 12/18 HMDB00153.
b a kernelType="branchingbased" – Los resultados son idénticos a los obtenidos con kenelType="sizedbased".
c a kernelType="spectrum" – tiempo = 9 segundos - orden similaridad de los metabolitos:12/14 HMDB00011 – 15/18 HMDB00153. Los valores de similitud para los metabolitos son más altos que los que suministra *HMDB-ChemQuery*.
d a kernelType="tanimoto" – tiempo = 9 segundos – Para HMDB00011 hay 11 metabolitos con similaridad igual 1. Se visualizan las estructuras y no tiene sentido este resultado: este algoritmo no discrimina entre las estructuras. No se puede tener en cuenta para el proyecto.

e a kernelType="marginalized" – tiempo = 9 segundos – El orden en que se ordenan los metabolitos no tiene sentido.
f a kernelType="minmaxTanimoto" – tiempo = 50 segundos - orden similaridad de los metabolitos:10/14 HMDB00011 – 12/18 HMDB00153.
g a kernelType="2Pspectrum" – tiempo = 42 minutos - orden similaridad de los metabolitos:4/14 HMDB00011 – 5/18 HMDB00153. Tiempo de cálculo muy elevado.
h g a kernelType="3Ptanimoto" – tiempo = 42 minutos – Para las dos moléculas de referencia, el orden en que se ordenan los metabolitos no tiene sentido.
i g a kernelType="triangular" – El programa R "dead" con este cálculo.

2.1.3.4. *ChemmieR*: Cheminformatics Toolkit for R [15]

ChemmineR es un package de quimioinformática para analizar los datos de moléculas pequeñas de fármacos en R. La versión actual contiene funciones para el procesamiento eficiente de un gran número de moléculas pequeñas, predicciones físico-químicas/estructurales, búsqueda de similitud estructural (incluye *eiR* y *fmcsR*), clasificación y agrupación de bibliotecas compuestas con una amplia espectro de algoritmos. Al incluir sólo *eiR y fmcsR* en su algoritmo para calcular la similaridad, este *package* no aporta nada al estudio de similaridad.

Una vez evaluado los diferentes *packages* y métodos, el método elegido para realizar el cálculo de similaridad con la base de datos HMDB completa, tanto por la calidad del cálculo, como por el tiempo de procesamiento, es el que el utiliza el kernel "*spectrum*" del paquete *Rchemcpp*.

2.1.4. Semejanza entre los compuestos de la DB-672 metabolitos.

A modo de ejemplo, al resultado del cálculo se similaridad con el kernel *spectrum* de la DB que contiene 672 compuestos se le realiza un análisis de cluster en función del valor de similaridad y se representa en un *heatmap*, todo ello realizado con el package *Rchemcpp*. El resultado se muestra en la figura 2.

Esta representación visual no se puede mostrar con una DB-HMDB completa.

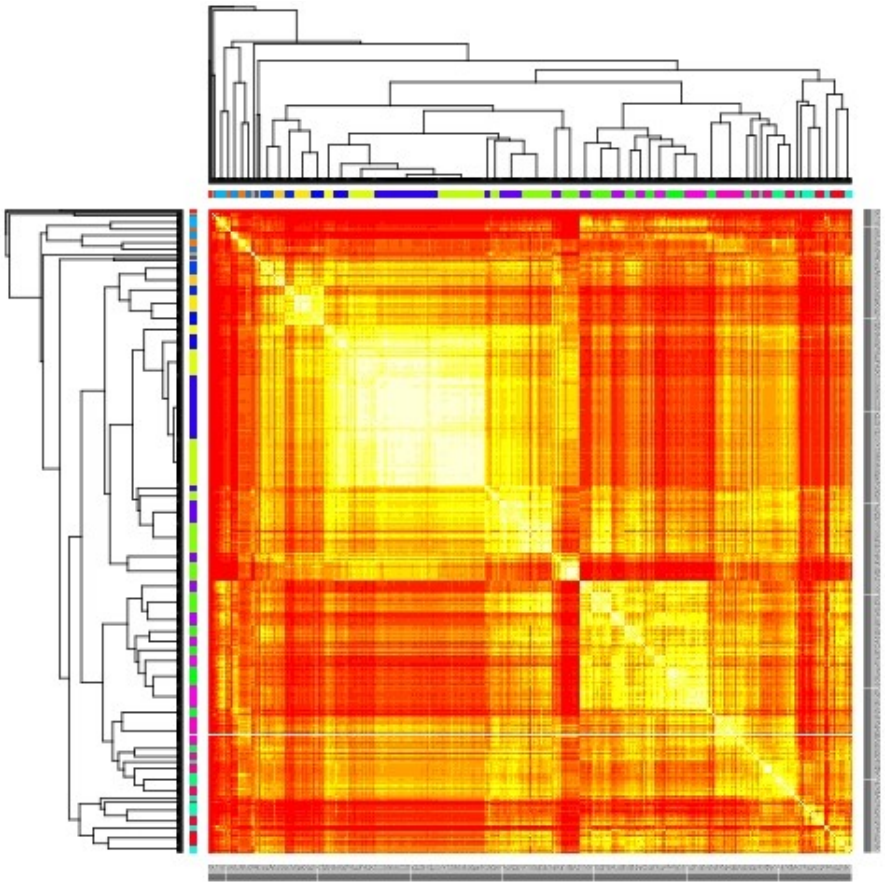

Fig. 2 : HeatMap - Análisis de Cluster vs similaridad (kernel="spectrum" (HMDB - 672 metabolitos)

2.1.5. Cálculo se similaridad de la DB-HMDB completa.

Cuando se intenta realizar el cálculo de similaridad con la DB de HMDB completa aparecen diferentes tipo de problemas. En este apartado se describe los criterios utilizados para tomar decisiones, la decisión que se toma, así como alternativas posibles.

2.1.5.1. En el SDF suministrado por HMDB el campo *name* se encuentra vacío de manera que hay que re-etiquetar el fichero original de HMDB con la nomenclatura de HMDB (ie. HMDB00001). Es la única manera que el package *Rchemcpp* entiende el fichero al cargarlo en R.

2.1.5.2. Al cargar HMDB completa en R, el programa informa de que hay moléculas que no “entiende” y se han de eliminar de la base de datos. Se eliminan de manera automática, de manera que la cantidad de metabolitos en HMDB disminuye hasta 41858 compuestos únicos.

2.1.5.3. Al realizar el cálculo de la matriz de similaridad de HMDB completa, el cálculo se interrumpe de manera abrupta con un mensaje de “memory allocate”. Este cálculo se intenta realizar en tres máquinas de cálculo diferentes. El servidor con más capacidad utilizado para realizar el calculo en un servidor de 10 cores, 2.40GHz, 30 MB CPU

cache y 192 GB RAM. En este servidos, tampoco se termina el cálculo: el mensaje al interrumpirse es el mismo y no hay posibilidad de tener acceso a un servidor con mayores prestaciones.

2.1.5.4. Al no poder calcularse la matriz de similaridad de HMDB completa por si misma, se decide cambiar de estrategia. En este proyecto sólo se calcularán la similaridad de los metabolitos diferenciados relacionados con T2D (alrededor de 400) y la HMDB completa. De esta forma, la matriz de similaridad contendrá aproximadamente de 400x41858 registros (BD-HMDB compuestos únicos) y se disminuye considerablemente la carga de memoria, principal problema en el cálculo de la matriz . En el apartado *Generación de datos* se explica de manera más detallada los números que aparecen.

2.1.5.5. El cálculo de similaridad con 400 compuestos vuelve a interrumpirse con el mismo mensaje. Este resultado es confuso ya que la carga de memoria, principal *handicap* para generar la matriz de similaridad, es similar al ejercicio realizado con la base de datos HMDB de 672 compuestos.

En una de las pruebas para entender que está ocurriendo, se decide imprimir en pantalla el nombre de la molécula responsable de que el cálculo se interrumpa. Realizando este ejercicio se localizan 14 moléculas en HMDB que el package *Rchemcpp* no sabe interpretar, y que tampoco ha detectado en el paso automática de búsqueda de moléculas que no entiende y mencionado anteriormente.

Estas 14 moléculas están etiquetadas como: HMDB02086 - HMDB02087 - HMDB02202 - HMDB02274 - HMDB03429 - HMDB03458 - HMDB06288 - HMDB06316 - *HMDB14370* - HMDB14470 - HMDB15617 - HMDB32287 - HMDB34831 – HMDB37761. Estas moléculas tienen características especiales ya que contienen átomos de cobalto (Co), gadolinium (Gd), hierro (Fe) o fósforo (P). Pero el átomo en si, no es la razón por la que el package *Rchemcpp* se detiene, ya que hay moléculas, ie. HMDB01083 que contiene Co, HMDB014635 contiene P, HMDB00887 contiene Fe y HMDB014678 contiene Gd y *Rchemcpp* lo entiende. No se ha podido averiguar porque estas moléculas no se entienden.

Por otra parte, para descartar que el kernel=”spectrum” sea el causante de no poder leer todas las estructuras, se crea una lista con las 14 moléculas y se realizan tres cálculos con otros kernels = (“sizebased”, “branchingbased”, “2Pspectrum”). En los tres cálculos ocurre lo mismo: el programa R “dead”. De manera que se descarta la posibilidad que sea el kernel el causante de la mala lectura.

Se decide eliminar estos 14 compuestos de la DB- HMDB de manera que ahora hay 41844 compuestos. Se realiza un cálculo de similaridad entre HMDB00001 y la nueva DB-HMDB para comprobar que no hay problemas con las estructuras. El cálculo funciona.

2.1.5.6. Se realiza un *script* prueba inicial para que se genere una matriz de similaridad de 7 estructuras frente a 41844 (7x41844). El cálculo se interrumpe con el mismo mensaje de error de “allocate memory”.

2.1.5.7. Se re-evalúa todo el proceso de cálculo empleado hasta ese momento y se determina que el procedimiento para obtener la matriz de similaridad es realizando el cálculo de manera que se obtenga un vector de 41844 filas x 1 columna y no de 1 fila x 41844 columnas, como era hasta ese momento. Se genera un *script* con siete metabolitos y el cálculo vuelve a interrumpirse. En estos momentos se decide generar diferentes scripts de manera que se obtenga un vector para el conjunto de metabolitos que se está analizando.

2.1.5.8. Cada cálculo de similaridad de cada metabolito con la DB-HMDB requiere un tiempo entre 30-40 minutos. Se pueden lanzar varios cálculos en el servidor de manera simultánea.

2.1.5.9. Control de calidad a la DB- HMDB.

Tal como se ha descrito en los apartados anteriores, la utilización de la DB-HMDB ha generado muchas complicaciones. Se exponen en este apartado, mejoras en su tratamiento para una mejor eficiencia en su uso en un futuro.

a Reducción de la cantidad de estructuras de metabolitos de HMDB. El análisis de similitud empieza buscando unos metabolitos específicos para una enfermedad. Estos metabolitos deben ser analizados gráficamente y en función del tipo de estructura que presentan se pueden eliminar aquellos compuestos que se sabe que los metabolitos de referencia no podrán ser similares. En este ejemplo, en la DB se han encontrado compuestos organometálicos, derivados de ácidos grasos, iones metálicos… estos estructuras se podrían eliminar de HMDB sin afectar a la calidad del estudio de similaridad. Disminuiría los problemas con las estructuras y tiempo de cálculo.

b Reducción de tiempo de cálculo mediante cálculos en paralelo. Se realizó una estimación de 13 meses de cálculo/un nodo para poder tener la matriz de similitud de HMDB completa (40000 x 40000, aprox). También se estimó en 5-6 semanas de cálculo si se utilizara el servidor de cálculo que se ha utilizado todos los procesadores disponibles. La ventaja de tener la matriz completa es que se podrían ampliar/cambiar el nombre de las enfermedades - metabolitos de manera rápida.

c La relación de enfermedades de HMDB no está bien desarrollada. Más de la mitad de las enfermedades de HMDB no están etiquetadas con *ulms o cui.* Un esfuerzo etiquetando las enfermedades aumentaría el número de enfermedades con el que las enfermedades problema podrían, y están, interaccionado.

d Seguir analizando nuevos packages de cálculo de similitud en R.

2.2. Generación de datos.

2.2.1. Listado de genes relacionados con la Diabetes tipo II (T2D) y de las enfermedades relacionadas.

Para obtener los diferentes *diseasomes* de genes, en este caso el de T2D, se ha de poder relacionar genes con enfermedades. En esta información se obtiene de la base de datos DisGeNet [9]. Se selecciona DisGeNet porque integra asociaciones entre genes y enfermedades humanas (GDA) de diversas bases de datos curadas por expertos y asociaciones derivadas de text-mining, incluyendo enfermedades mendelianos, complejas y ambientales.

La información de la base de datos de DisGeNet [9] se ha obtenido utilizando el *plugin* DisGen [9] creado para Cytoscape [10] . Primero se realiza una búsqueda por genes en la base de CURATED y se filtra por la enfermedad que se quiere analizar: Diabetes Mellitus, Non-Insulin-Dependent ulms: C0011860 (T2D). Se selecciona la T2D en la proyección de enfermedad y se obtienen los genes asociados. Una vez obtenidos los genes se vuelve a analizar la proyección de las enfermedades. Al final se genera un fichero con 827 enfermedades relacionadas con T2D a través de los genes que comparten. Este fichero se etiqueta como disgen_t2d.csv (el fichero se entregó en la PEC2). La lista de genes se obtiene de la primera búsqueda con DisGeNet, son 173 y se etiqueta como disgen_t2d_gen.csv (el fichero se entregó en la PEC2).

En paralelo, cuando se ha determinado el número y el nombre de enfermedades comunes presentes en HMDB y en DisGeNet (41) (apartado 2.2.2) se utiliza el *package* (disgenet2r) [18] para seleccionar los genes de cada una de las enfermedades.

Se describe a continuación el *script* utilizado en R para obtener la lista de genes para las 41 enfermedades. Este *script* corresponde a la enfermedad de Chron:

```
library(disgenet2r)
setwd("/home/lucas/bego/Cytoscape/GENES")
diseaseOfinterest <- c("C2675113")
dq <- disgenetDisease(disease = diseaseOfinterest, database = "CURATED", score = c('>', 0))
gene.Chron <- sort(dq@qresult$c2.name)
write.table(gene.Chron, file='gene.Chron.txt', quote = FALSE, sep = " ", col.names = F, row.names = F)
```

Una vez obtenidas las 41 listas de genes para cada enfermedad, se compara cada una de ellas con los genes de T2D para seleccionar los genes comunes y construir una tabla donde estén descritos el nombre de las enfermedades y el número de genes comunes. Con esta tabla se puede presentar el diseasome de genes de T2D y las enfermedades relacionadas.

2.2.2. Listado de enfermedades relacionados con la Diabetes tipo II (T2D) y de las enfermedades relacionadas.

Tal como se ha comentado anteriormente, en HMDB los metabolitos están relacionados con enfermedades por lo que es necesario identificar el conjunto de enfermedades que HMBD tiene clasificadas.

Esta lista no está accesible en el portal de HMDB pero se obtuvo a partir de una petición vía *mail* a su portal. HMDB suministró un documento en formato *excel* con el nombre de 620 enfermedades, pero únicamente 301 de ellas tienen identificación *OMIM* [19], hmdb_deseases.csv (el fichero se entregó en la PEC2) y son las únicas con las que se trabajará.

Por otra parte, el identificador utilizado para enfermedades en la versión actual de DisGeNET es el vocabulario Unificado de Lenguaje Médico System® (UMLS®) por lo que el anterior fichero (disgen_t2d.csv) se manipula con el package *diseasemapping* de R que permite relacionar el identificador de DisGeNet (cui_id) con el identificador de HMDB (omim_id). Esta traducción permite tener una lista de 490 enfermedades únicas relacionadas con T2D con las que trabajar y se ha etiquetado como disgen_t2d_cui_omim.csv (el fichero se entregó en la PEC2).

Finalmente hay que obtener la lista de enfermedades presentes en en DisGeNet (disgen_t2d_cui_omim.csv) y que se encuentran representadas en HMDB (hmdb_deseases.csv). Para ello se manipulan ambos ficheros con los paquetes *dplyr* & *tidyr* de R [17] hasta obtener la lista de que se encuentran en HMDB y están relacionados con T2D a través de genes comunes (DisGeNet). Esta lista se llama disease_list_t2d.csv y presenta inicialmente 61 enfermedades.

Tal como se comenta a continuación, para obtener los metabolitos diferenciados para cada enfermedad, se ha de ir uno a uno. Pero hay algunas enfermedades, aunque estén identificadas, no se encuentran en HMDB. Estas enfermedades son: Down's sydrome, Osteoarthritis, Farber disease, Cerebellar Ataxia, Metabolic Syndrome, Minimal brain dysfunction, Major depressive disorder, Carnitine palmitoyltransfersase I deficiency, Cerebral Malaria, Renal Cell Carcinoma. Por otra parte, una misma enfermedad puede tener para un identificador omim_id (que es único para cada enfermedad) 2 ó 3 identificadores cui diferentes, por lo que disminuye el nombre de enfermedades. Finalmente, hay 41 enfermedades con las que realizar los diferentes *diseasomes* y la lista de enfermedades a analizar es:

Adrenoleukodystrophy – Adrenomyeloneuropathy - Alzheimer's disease - Amyotrophic lateral sclerosis - Asthma – Autism – Beta – thalassemia - Celiac desease - Colorectal cancer - Crohn's disease - Cystic fibrosis - Diabetes mellitus type 1 - Diabetes mellitus type 2 - Eczema - Endometrial cancer - Glucose transporter type 1 deficiency syndrome - Hepatocellular carcinoma – Homocystinuria - Huntington's

disease – Hypertension - Lung Cancer – Malaria – Methionine adenosyl transferase deficiency – Migraine - Multiple myeloma – Obesity – Osteoporosis - Ovarian cancer - Pancreatic cancer - Polycysticovary syndrome - Primary biliary cirrhosis - Prolidase deficiency – Prostate cancer - Renal cell carcinoma - Rett syndrome - Rheumatoid arthritis – Rhinitis – Schizophrenia - Sickle cell anemia - Spina Bifida - Stroke - Wilson's disease.

2.2.3. Listado de estructuras de metabolitos relacionados con la Diabetes de tipo II (T2D) y de las enfermedades relacionadas.

2.2.3.1. Listado de metabolitos diferenciados obtenidos desde el *Browser* de HMDB.

Desde el *browser* se puede acceder a los metabolitos diferenciados que se relacionan con la enfermedad de T2D en ensayos clínicos. En el caso T2D, y se obtiene 43 metabolitos diferenciados. De los 43 metabolitos, únicos son 26, ya que algunos de ellos se han encontrado en diferentes fluidos biológicos (b*lood, urine or Cerebrospinal Fluid*). En la tabla 1 se muestra la relación del nombre químico con la nomenclatura de HMDB.

(R)-3-Hydroxybutyric acid (HMDB00011)	(R)-3-Hydroxyisobutyric acid (HMDB00336)	(S)-3-Hydroxyisobutyric acid (HMDB00023)
1,5-Anhydrosorbitol (HMDB02712)	1-Butanol (HMDB04327)	1-Methylhistidine (HMDB00001)
3-Hydroxybutyric acid (HMDB00357)	3-Methylhistidine (HMDB00479)	4-Heptanone (HMDB04814)
8-Hydroxyguanine (HMDB02032)	Acetoacetic acid (HMDB00060)	Acetone (HMDB01659)
D-Fructose (HMDB00660)	D-Glucose (HMDB00122)	D-Lactic acid (HMDB01311)
Dimethylamine (HMDB00087)	Dodecanedioic acid (HMDB00623)	Estriol(HMDB00153)
Fructosamine (HMDB02030)	Glycerol (HMDB00131)	Hyaluronan (HMDB10366)
L-Carnitine (HMDB00062)	Pyruvaldehyde (HMDB01167)	S-Adenosylmethionine (HMDB01185)
Scyllitol (HMDB06088)	Uric acid (HMDB00289)	

Tabla 1. Listado metabolitos diferenciados presentes en ensayos clínicos de T2D (identificación HMDB id)

De manera individual, y para cada una de las 41 enfermedades relacionadas con T2D, se introduce en el *browser* el nombre de cada enfermedad y se obtiene la lista de metabolitos diferenciales que han sido encontrados de manera experimental. Se unen todos los metabolitos diferenciados de todas las enfermedades y se encuentran que hay 738 metabolitos diferenciales relacionados con T2D no únicos y 419 estructuras únicas. De estos 419 estructuras únicas se realizan

los estudios de similaridad con el algoritmo explicado anteriormente (el fichero se entregó en la PEC3).

2.2.3.2. Listado de metabolitos similares generado a partir de cálculos de similaridad.

Tal como se ha descrito anteriormente, para cada una de las enfermedades relacionadas con T2D y la misma T2D, existe una lista con metabolitos diferenciados asociados a la enfermedad (en concreto, 419). Para cada uno de estos metabolitos se realiza el cálculo de similaridad y se recogen estos valores en un *data-frame* utilizando el programa R [8] de manera individual para cada enfermedad. Estos ficheros se trabajarán con los *packages dplyr & tidyr* de R [17], principalmente, hasta obtener una lista de enfermedades y metabolitos comunes con el que poder dibujar el diseasome de metabolitos de T2D. Se detalla a continuación.

Para cada enfermedad se genera un fichero con número de columnas igual a metabolitos diferenciados encontrados en HMDB y como número de filas el número de metabolitos presentes en HMDB completa (41844 metabolitos, las estructuras de los metabolitos que el *packages Rchemcpp* puede entender). En la Fig. 3 se muestra una visión parcial del *data-frame* (T2D) obtenido para T2D.

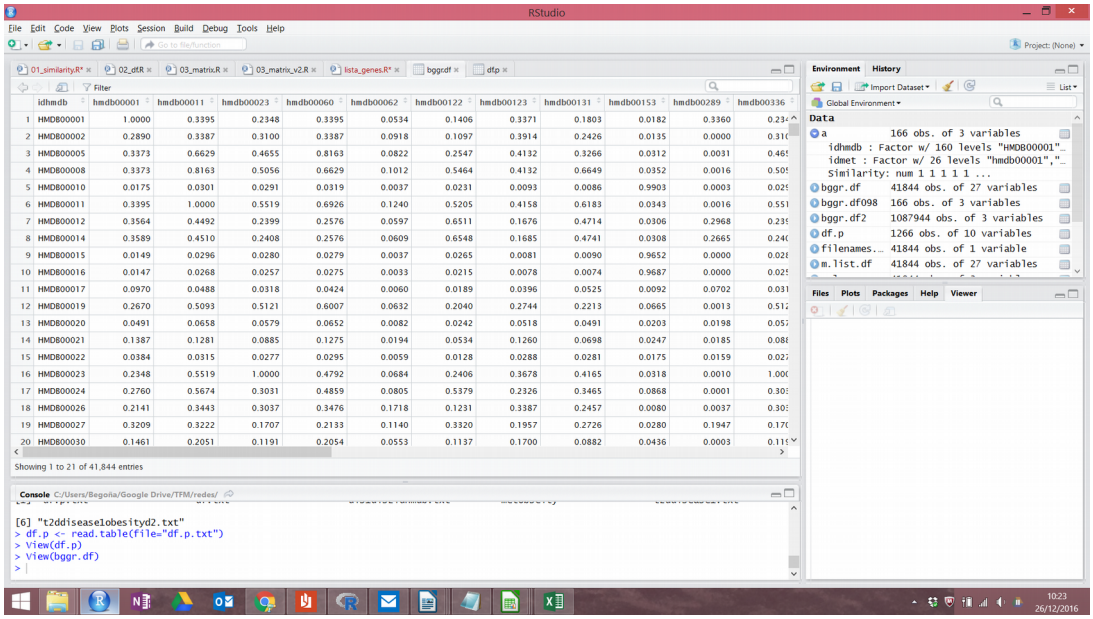

Fig.3: data-frame (T2D) - Similaridad de los metabolitos diferenciados vs HMDB completa.

Antes de seguir avanzando, se intenta crear un *heatmap* con la distribución de los valores de similaridad de los 26 metabolitos diferenciados respecto a la HMDB utilizada. Se utiliza el *package*

ggplot2 pero por problemas de memoria no se crea. El resultado se muestra en la sección 6 (anexos).

Cada una de las *data-frame* (enfermedad) se filtra por el valor de similaridad. En este caso, teniendo en cuenta que al realizar las pruebas con los diferentes *kernels*, los valores de similaridad son más elevados que los que suministra HMDB-ChemQuery, el valor de similaridad por el que se filtra es de 0.98.

A partir del fichero anterior, con los *packages pdlyr & tidyr* se manipulan los *data-frame* (enfermedad) de manera para cada enfermedad se obtenga un *data-frame* con tres columnas con la etiqueta identificativa del metabolito HMDB (idhmdb), la etiqueta identificativa del metabolito diferenciado (idmet) y el valor de similaridad mayor de 0.98 entre los dos metabolitos. En la Fig. 4 se muestra una visión parcial del *data-frame (T2D)* obtenido para T2D.

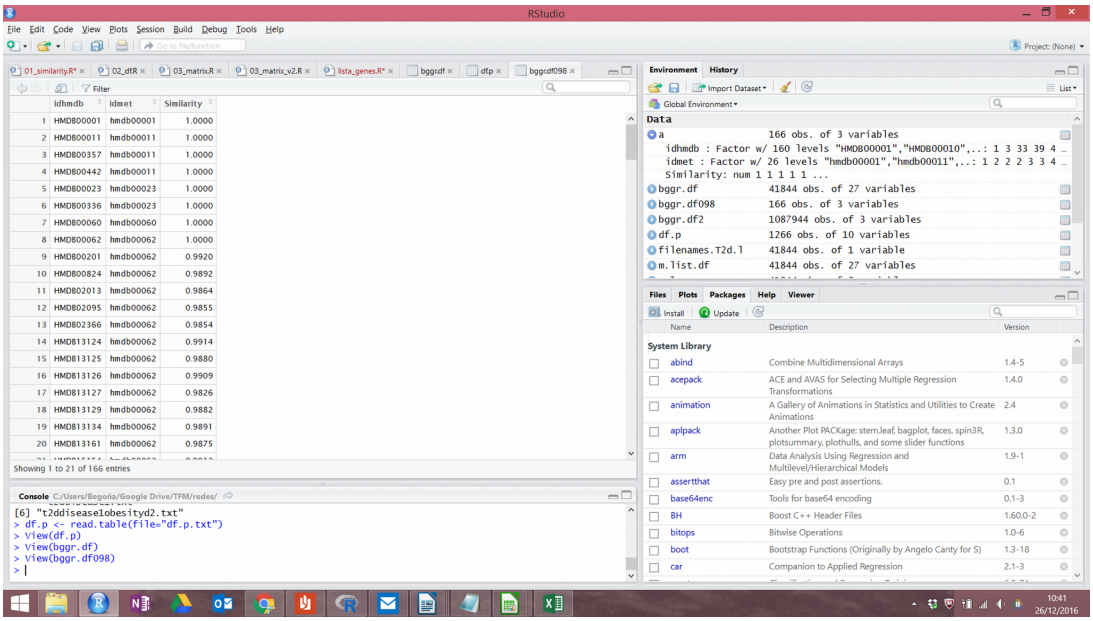

Fig.4: data-frame (T2D): metabolito similar (idhmdb), metabolito diferenciado (idmet), similarity

En este punto, se unen todos los *data-frame* (enfermedad) de todas las enfermedades y se agrupan de manera que se cree una nueva *data-frame (*similaridad*)* donde sólo estén incluidos aquellos metabolitos similares que, por lo menos, compartan dos enfermedades. También se averigua el número de enfermedades que cada metabolito está presente con una similaridad > 0.98, el número máximo de enfermedades donde un metabolito está presente (8) y el nombre de cada enfermedad donde el metabolito similar está presente. El *data-frame* (similaridad > 0.98) tiene 1266 observaciones no únicas. En la Fig. 5 se muestra una visión parcial del *data-frame* obtenido para todas las enfermedades.

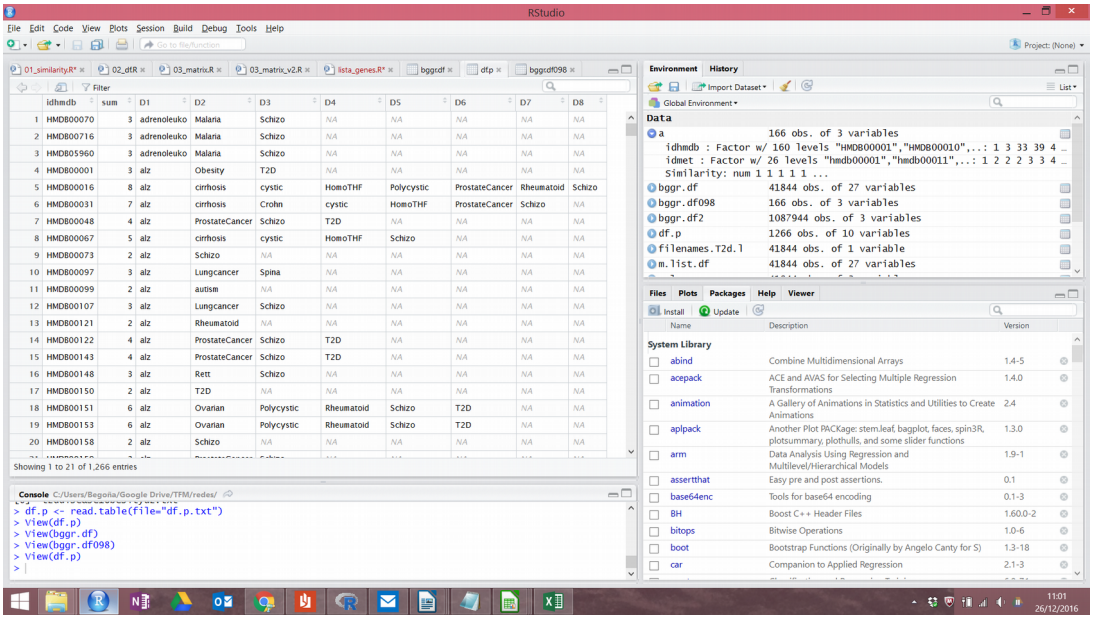

Fig.5: data-frame (similaridad > 0.98): idhmdb, nº total de enfermedades, nombre de enfermedades donde está presente el metabolito similar.

A continuación, el *data-frame* anterior se transforma en un *data-frame* de tres columnas donde las dos primeras columnas son el nombre de dos enfermedades y la tercera, el nombre del metabolito que las relaciona. Este *data-frame* (global) contiene 5494 entradas no únicas. En la Fig. 6 se muestra una visión parcial del *data-frame* (global) obtenido para todas los pares de enfermedades.

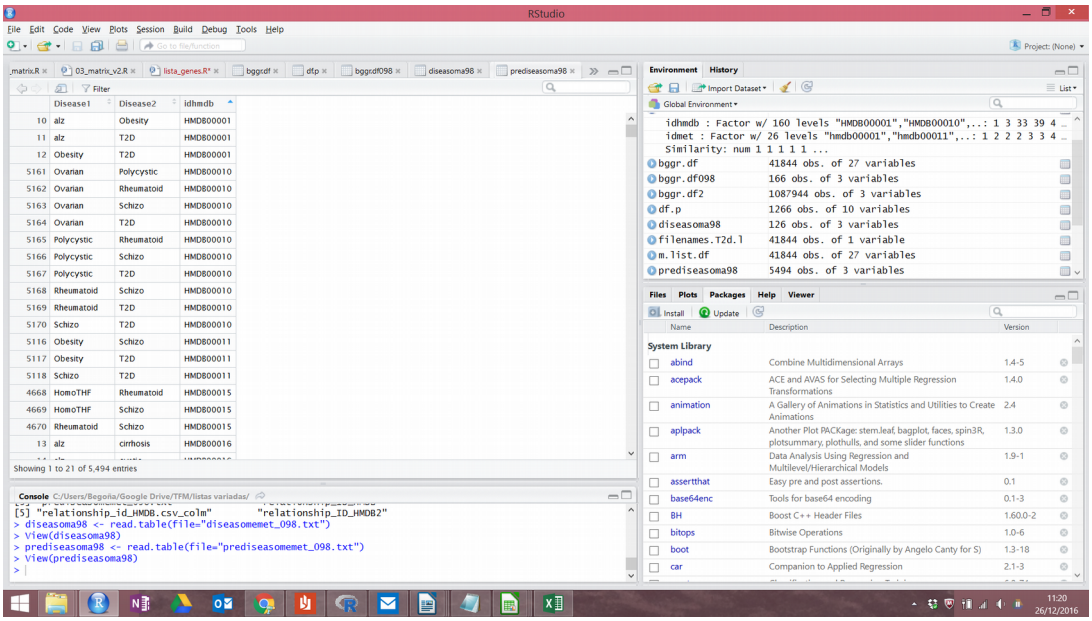

Fig.6: data-frame (global): nombre de enfermedades y nombre del metabolito que relaciona cada par de enfermedades.

Finalmente, se genera el *data-frame* para poder representar el diseasome de metabolitos de T2D y las enfermedades relacionadas. A

partir del anterior *data-frame* (global) se obtiene uno nuevo *data-frame* donde se describe en cada fila el nombre de dos enfermedades que están relacionadas a través de por lo menos un metabolito y el número de metabolitos (similares y diferenciados) que relacionan ese par de enfermedades. En la Fig. 7 se muestra una visión parcial del *data-frame* (diseasome) obtenido para T2D y las enfermedades relacionadas.

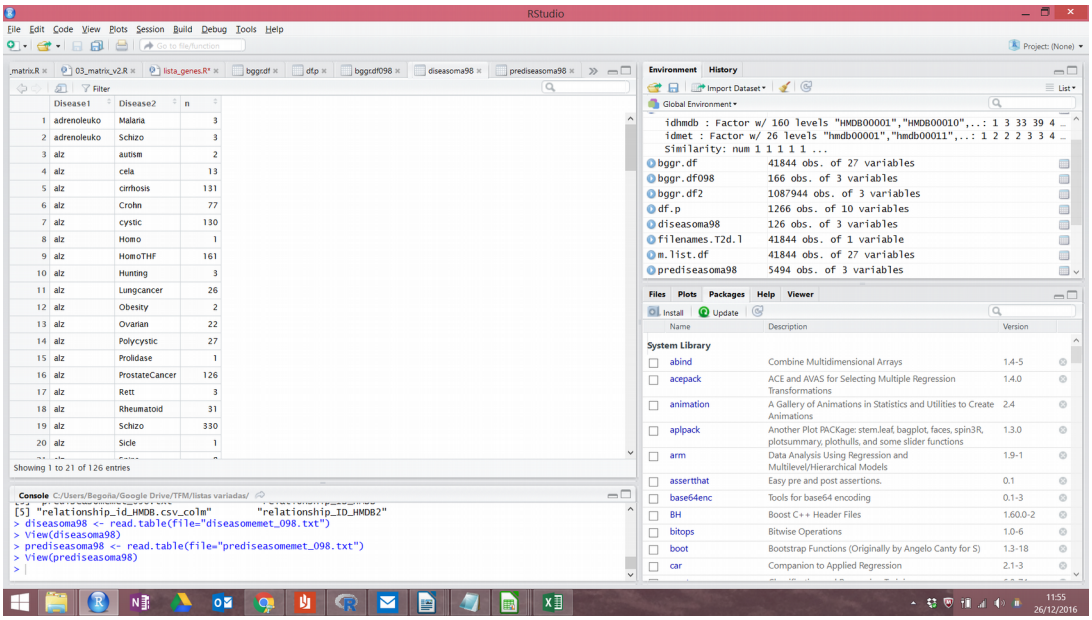

Fig.7: data-frame (diseasome) : nombre de enfermedades, nº metabolitos diferentes que relacionan las dos enfermedades.

Para una similaridad mayor de 0.98 el diseasome de T2D presenta 126 relaciones enfermedad-enfermedad a través de metabolitos comunes.

2.3. Generación de redes de interacción

2.3.1. Para la enfermedad T2D y enfermedades relacionadas – Diseasome de Genes.

A partir de los datos generados en el apartado 2.2.1, data-frame (diseasome) se puede representar con Cytoscape el diseasome de genes de T2D. En la Fig. 8 se muestra el diseasome de genes. Las bolas rosas representan las enfermedades, en el centro se encuentra T2D, y las bolas azules representan los genes. En este caso, no se representan el número de genes comunes entre las dos enfermedades por falta de espacio.

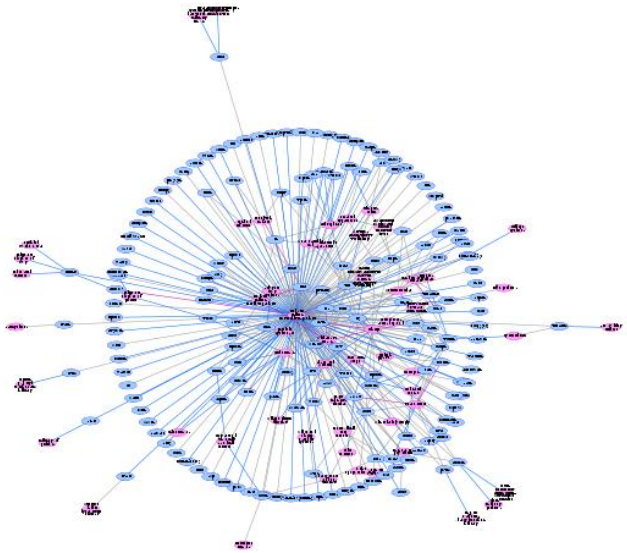

Fig. 8: Diseasome T2D - Genes. Rosa - enfermedades. Azul -genes.

2.3.2. Para las enfermedad T2D y enfermedades relacionadas – Diseasome de metabolitos.

A partir de los datos generados en el apartado 2.2.3.2, data-frame (diseasome) se puede representar con Cytoscape el diseasome de metabolitos de T2D. En la Fig. 9 se muestra el diseasome de metabolismo. Las bolas verdes representan las enfermedades (hubs) y el número que se encuentra encima de las líneas (edges) representan el número de metabolitos comunes entre las dos enfermedades.

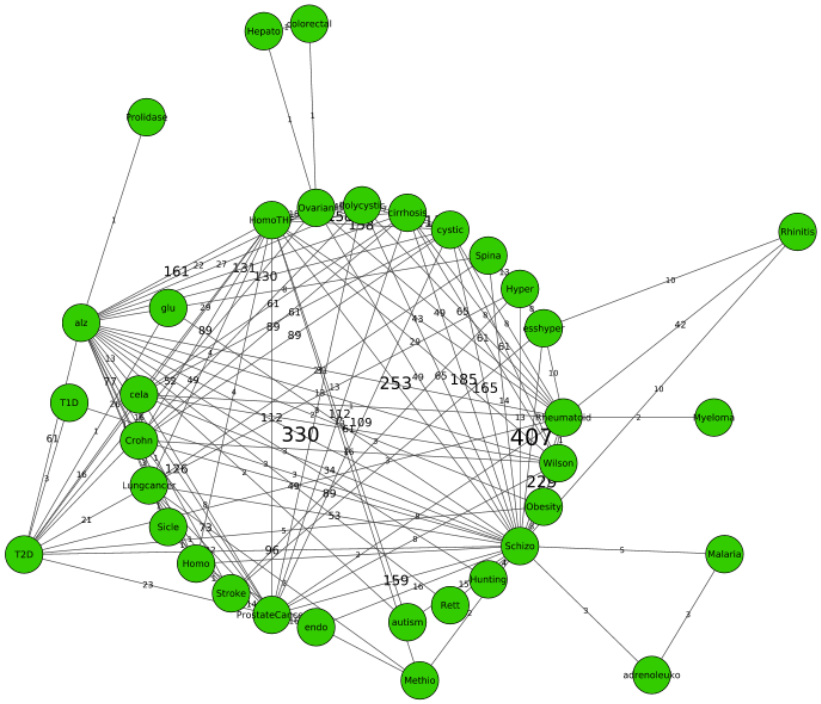

Fig.9: Diseasome T2D - Metabolismo. Hubs - enfermedades. Edges - nº metabolitos comunes entre enfermedades.

2.4. **A partir de los diseasomes**

2.4.1. Análisis de enriquecimiento GO - genes.

Una vez se obtiene el Diseasome de genes también se puede realizar un análisis de términos de Ontología de genes (GO) para cada una de los enfermedades y del Diseasome que se puede establecer entre las 41 enfermedades.

Para cada una de las enfermedades se realiza el estudio de enriquecimiento con la herramienta DAVID 6.8 [20]. Se introduce la lista de genes diferenciados (comunes entre la enfermedad) en el apartado *Functional Annotation Chart,* se selecciona Homo sapiens como especie. Surgen unos primeros resultados bajo el nombre de Annotation Summary Results, se selecciona *GOTERM_BP_DIRECT/Chart* y dentro de options: display FDR (false discovery rate) . Al utilizar FDR, los términos enriquecidos son aquello que tienen un valor de FDR menor o igual a 5 (la escala del término FDR es de 0 a 100, por lo que el valor de 5 es el valor que se considera significativo).

Al realizar el análisis de enriquecimiento, hay algunas enfermedades que no presentan términos enriquecidos: Adrenoleukodystrophy …, y otras enfermedades no se ha terminar el análisis. Se muestra el resultado de 7 enfermedades.

a Alzheimer's Disease: presenta 65 términos enriquecidos con una significancia mayor a 5 (α=5). Se muestran los cinco primeros términos:
GO:0001666~response to hypoxia
GO:0042493~response to drug
GO:0045429~positive regulation of nitric oxide biosynthetic process
GO:0008284~positive regulation of cell proliferation
GO:0043066~negative regulation of apoptotic process

a Obesity: sólo presenta dos términos enriquecidos con una significancia mayor a 5 (α=5) y son:
GO:0006006~glucose metabolic process
GO:0042493~response to drug

a Non-small cell lung carcinoma: presenta 47 términos enriquecidos con una significancia mayor a 5 (α=5). Se muestran los cinco primeros términos:
GO:0008283~cell proliferation
GO:0008284~positive regulation of cell proliferation
GO:0006915~apoptotic process
GO:0001701~in utero embryonic development
GO:0007265~Ras protein signal transduction

a Rheumatoid arthritis: presenta 31 términos enriquecidos con una significancia mayor a 5 (α=5). Se muestran los cinco primeros términos:
GO:0006955~immune response

GO:0006954~inflammatory response
GO:0051092~positive regulation of NF-kappaB transcription factor activity
GO:0045944~positive regulation of transcription from RNA polymerase II promoter
GO:0044130~negative regulation of growth of symbiont in host

a Essential Hypertension:presenta 9 términos enriquecidos con una significancia mayor a 5 (α=5). Se muestran los cinco primeros términos:
GO:0008217~regulation of blood pressure
GO:0007263~nitric oxide mediated signal transduction
GO:0019229~regulation of vasoconstriction
GO:0045909~positive regulation of vasodilation
GO:0002034~regulation of blood vessel size by renin-angiotensin

a Primary biliary cirrhosis: sólo presenta dos términos enriquecidos con una significancia mayor a 5 (α=5)y son:
GO:0019221~cytokine-mediated signaling pathway
GO:0032729~positive regulation of interferon-gamma production

a Diabetes Millitus, non-insulin dependent presenta 14 términos enriquecidos con una significancia mayor a 5 (α=5). Se muestran los cinco primeros términos:
GO:0008283~cell proliferation
GO:0008284~positive regulation of cell proliferation
GO:0006915~apoptotic process
GO:0001701~in utero embryonic development
GO:0007265~Ras protein signal transduction

En el caso del Diseasome de las 7 enfermedades, la búsqueda de ontología se realiza con un número muy pequeño de genes. El resultado determina que NO hay presencia de términos de funciones de ontología de genes que se encuentren significativamente enriquecidas en el Diseasome de estas siete enfermedades. Con más enfermedades hay más posibilidades de obtener genes diferenciales.

2.4.2. Análisis de enriquecimiento pathways – metabolitos.

Se tiene el listado, en forma de pares de enfermedades, con el nombre de metabolitos comunes a ambas enfermedades para realizar análisis de enriquecimiento de metabolito a *pathway*. Se ha localizado una DB denominada IMPaLA (*Integrated Molecular Pathway Level Analysis*) [21] donde se ha pueden realizar análisis de enriquecimiento teniendo en cuenta los metabolitos. En este caso, se pueden realizar los análisis de enriquecimiento porque IMPaLA acepta el identificador de HMDB.

No se presentan resultados porque sólo se han podido realizar análisis de pruebas.

2.4.3. Comparar los diseasomes de genes y metabolitos.

Aunque se tienen los listados para poder realizar el trabajo, no ha sido posible realizar este estudio en el tiempo marcado por el TFM.

3. Conclusiones

3.1. Conclusiones del trabajo

3.1.1. El primer objetivo del trabajo era construir dos redes de interacciones, una basada en genes y otra en metabolitos para la enfermedad T2D y sus enfermedades relacionadas. Este objetivo se ha conseguido.

3.1.2. El segundo objetivo que era realizar un análisis de enriquecimiento para los genes y los metabolitos. Este objetivo no ha sido posible realizarlo de manera completa.

3.1.3. Hay que analizar de manera muy cuidadosa las base de datos con las que se trabaja. Así como ser muy estricto en el análisis de los algoritmos/técnicas que hay que utilizar en la metodología.

3.2. Reflexión crítica sobre el logro de los objetivos.

Este estudio era importante para ampliar los conocimientos de System Biology. Tal como se ha desarrollado el trabajo, el grado de aprendizaje no ha sido satisfactorio.

Por otra parte, la capacidad de programar en R ha mejorado a lo largo de la TFM, así como la capacidad de entender el manejo de los conceptos diseasome, redes de interacción, genes diferenciados, metabolitos diferenciados.

3.3. Reflexión crítica del seguimiento de la planificación y metodología.

3.3.1. Planificación.

La planificación fue pensada para poder conseguir todos los objetivos. Pero no se contempló que el cálculo de la similitud entre los metabolitos tuviera tantos inconvenientes. Ha quedado claro que cuando no se conoce los programas/algoritmos que se han de utilizar para un estudio se ha de marcar un tiempo de pruebas más largo.

3.3.2. Metodología.

La metodología seleccionada ha sido correcta. Se ha trabajado principalmente con R y *bash*. También se realizaron pequeñas pruebas para realizar parte del trabajo con MySQL, pero finalmente no hizo falta.

Por otra parte, se ha trabajado con diferentes bases de datos. Algunas de ellas, por ejemplo, DisGeNet han sustituido a algunas que se habían mencionado al principio del proyecto.

a. ImmunoBase. Esta base de datos no se ha utilizado porque trabaja con enfermedades autoinmunes y T2D no es una enfermedad autoinmune.

b. Malacards. La calidad de esta DB está muy contrastada, pero todo el trabajo de búsqueda de genes diferenciales se ha de realizar a mano. Además, no existe la posibilidad de descargar la información generada en un fichero sino que se ha de realizar copy/paste.

Se ha realizado una comparativa entre los resultados de Malacards y DisGeNet (tabla 2). Se han seleccionado las enfermedades T2D y T1D, se han analizado los genes de cada uno de ellos y finalmente se han localizado los comunes. Los resultados para ambas enfermedades y en cada DB son muy diferentes. Este ejercicio refleja la complejidad en la selección de las BD para realizar cualquier estudio que implique trabajar con bases de datos.

Tal como se ha mencionado al principio del proyecto, se elije DisGeNet para realizar el diseasome de genes por la calidad de la Base de datos y por su facilidad al obtener los resultados.

DB	n.º Genes T2D	n.º Genes T1D	n.º genes compartidos
Malacards	164	147	(45) ABCC8 – ACE – ADIPOQ – ADRB3 - AGT – AGTR1 – AHSG – ALB – APOA1 – APOB – APOC3 – COG2 – DPP4 – EDN1 - G6PC - GCG - GCK - GH1 - GHR - HNF1A - HNF1B - IGFBP1 - LPA - NAMPT - NEUROD1 -NOS3 - NPY - PAX4 - PDX1 - PPARA - PPARG - RBP4 - RETN - SELE - SERPINE1 - SHBG - SLC2A1 - SLC2A2 - SLC2A4 - SLC30A8 - SLC5A2 - SLC5A4 - SST - TNF - VCAM1- WFS1
DisGeNet	173	77	(11) ABCC8 – CAT – GLIS3 – HNF1A – HP – INS – INS-IGF2 – KCNJ11 – NOS3 – PAX4 - TNF

Tabla 2. Comparativa (genes) T2D/T1D : Malacards vs DisGeNet

c. package disgenet2R.

Se ha intentado normalizar los valores de similaridad utilizando los coeficientes de Jaccard, pero el package disgenet2R no ha funcionado correctamente. Este package todavía está “en construcción” de manera que hay que ser prudentes con las utilidades. He escrito a la programadora para solucionar este problema.

3.4. Líneas de trabajo futuro.

3.4.1. En primer lugar se tendría que realizar un análisis de enriquecimiento para los genes comunes que se han encontrado (GO) utilizando DAVID [20] y otras bases de datos que tenga la misma calidad y que se pueda obtener la información de manera no manual.

3.4.2. Realizar un análisis de enriquecimiento para los metabolitos comunes a *pathway*. La página web [21] permite realizar este trabajo ya que tiene como input el identificador de los metabolitos de HMDB. Además, esta página te permite realizar el GO de los genes. Al poder realizar los dos análisis de enriquecimiento en la misma página se asegura, por lo menos, coherencia en los resultados. Habría de asegurase la calidad de la base de datos.

3.4.3. Realizar este estudio con datos experimentales de genómica y metabolómica para la T2D y las enfermedades relacionadas.

4. Glosario

DB	Bases de Datos
FDR	False discovery rate
IMPaLA	Integrated Molecular Pathway Level Analysis
GDA	Asociaciones entre Genes y Enfermedades
HMDB	Human Metabolome Database
NCDs	Non-communicable chronic diseases
GO	Gene Ontology
SDF	Structure Data File
SNP	Polimorfismo en un sólo nucleótido
T2D	Diabetes Millitus, non-insulin dependent
TFM	Trabajo fin de Máster
VDA	Asociaciones Variante de Enfermedad

5. Bibliografía

[1] Albert-László Barabási. "Network Science" . 2007. This book is licensed under a Creative Commons: CC BY-NC-SA 2.0. PDF V26, 05.09.2014

[2] Goh, Kwang-Il, Michael E. Cusick, David Valle, Barton Childs, Marc Vidal, and Albert-László Barabási. 2007. "The Human Disease Network." PNAS 104 (21): 8685-8690

[3] Li Y, Agarwal P. "A Pathway-Based View of Human Diseases and Disease Relationships". 2009 PLoS ONE 4(2): e4346. doi:10.1371/journal.pone.0004346

[4] Malacards (www.malacards.org), 14 de noviembre de 2016

[5] ImmunoBase (https://beta.immunobase.org), 14 de noviembre de 2016.

[6] Global report on diabetes. World Health Organization, Geneva, 2016. http://www.who.int/mediacentre/factsheets/fs312/en, 12 de diciembre de 2016.

[7] HMDB (www.hmdb.ca) , 20 de septiembre de 2016

[8] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

[9] Integration of biological networks and gene expression data using Cytoscape
Cline MS, Smoot M, Cerami E, Kuchinsky A, Landys N, Workman C, Christmas R, Avila-Campilo I, Creech M, Gross B, Hanspers K, Isserlin R, Kelley R, Killcoyne S, Lotia S, Maere S, Morris J, Ono K, Pavlovic V, Pico AR, Vailaya A, Wang PL, Adler A, Conklin BR, Hood L, Kuiper M, Sander C, Schmulevich I, Schwikowski B, Warner GJ, Ideker T, Bader GD
Nat Protoc. 2007;2(10):2366-82
Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks
Rowan Christmas, Iliana Avila-Campillo, Hamid Bolouri, Benno Schwikowski, Mark Anderson, Ryan Kelley, Nerius Landys, Chris Workman, Trey Ideker, Ethan Cerami, Rob Sheridan, Gary D. Bader, and Chris Sander
Am Assoc Cancer Res Educ Book 2005: 12-16
Cytoscape: a software environment for integrated models of biomolecular interaction networks.
Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T.
Genome Research 2003 Nov; 13(11):2498-504

[10] DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database (Oxford). 2015 Apr 15;2015:bav028. doi: 10.1093/database/bav028. Print 2015.
Piñero J1, Queralt-Rosinach N1, Bravo À1, Deu-Pons J1, Bauer-Mehren A1, Baron M1, Sanz F1, Furlong LI2.

[11] http://biotech.fyicenter.com/resource/sdf_format.html, 20 de septiembre de 2016

[12]Accelerated similarity searching and clustering of large compound sets by geometric embedding and locality sensitive hashing.
Cao Y, Jiang T and Girke T (2010). Bioinformatics, 26(7), pp. 953–959.

[13] fmcsR: mismatch tolerant maximum common substructure searching in R.
Wang Y, Backman TWH, Horan K and Girke T (2013). Bioinformatics, 29(21), pp. 2792–2794.

[14] Rchemcpp: a web service for structural analoging in ChEMBL, Drugbank and the Connectivity Map.
G. Klambauer, M. Wischenbart, M. Mahr, T. Unterthiner, A. Mayr, and S. Hochreiter Bioinformatics (2015) Oct 15;31(20):3392-4.
[15] ChemmineR: a compound mining framework for R.
Cao Y, Charisi A, Cheng L, Jiang T and Girke T (2008). Bioinformatics, 24(15), pp. 1733–1734.
[16] http://www.hmdb.ca/structures/search/metabolites/structure#results, 8 de noviembre de 2016.
[17] https://cran.r-project.org/web/packages/dplyr/index.html
dplyr: A Grammar of Data Manipulation
https://cran.r-project.org/web/packages/tidyr/index.h
tml tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions
[18] disgenet2r: An R package to explore the molecular underpinnings of human diseases. Alba Gutierrez-sacristán, Janet Piñero, Nuria Queralt-Rosinach, Emilio Centeno and laura I. Furlong. In preparation (diciembre de 2016).
[19] Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders.
Hamosh A1, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Nucleic Acids Res. 2005 Jan 1;33
[20] [Functional Annotation Bioinformatics Microarray Analysis
https://david.ncifcrf.gov/, 20 de diciembre de 2016.
[21] http://impala.molgen.mpg.de/, 22 de diciembre de 2016.

6. Anexos

Representa el intento fallido de representar los valores de similitud entre los metabolitos diferenciados de T2D y cada uno de los metabolitos que hay presente en HMDB.