

# **UNIVERSIDAD DE MURCIA**



## **TRABAJO FIN DE MÁSTER MÁSTER BIOINFORMÁTICA 2013/2014**

**“Método de clasificación de regiones del cerebro  
según la expresión cuantitativa de grupos  
ontológicos de genes”**

Autor:  
Maria Teresa López Cascales  
48548901H

Directores:  
Faustino Marín San Leandro  
Jesualdo Tomás Fernández Breis

## ÍNDICE

<b>A. RESUMEN .....</b>	<b>pág.</b>	<b>3</b>
<b>B. INTRODUCCIÓN .....</b>	<b>pág.</b>	<b>4</b>
<b>C. MATERIALES Y MÉTODOS .....</b>	<b>pág.</b>	<b>6</b>
C-I. Ontologías		
I.1 Gene Ontology .....	pág.	6
I.2 La anatomía del cerebro representada como ontología de estructuras .....	pág.	6
C-II. Base de datos Allen Brain Atlas.		
II.1 Detección de expresión génica mediante Hibridación in situ.....	pág.	7
II.2 Uso de las API .....	pág.	9
C-III Proyecto Bioconductor .....		
C-IV. BiomaRt.....		
C-V Herramientas estadísticas .....		
A) Normalización .....	pág.	11
B) Clustering .....	pág.	12
C-VI. Método seguido en este trabajo.....		
C-VII Scripts desarrollados para llevar a cabo el método descrito.....		
<b>D. RESULTADOS.....</b>	<b>pág.</b>	<b>14</b>
D-I. Búsqueda de la lista de genes (BiomaRt) .....	pág.	14
D-II. Obtención de la lista de documentos “Allen” para cada gen .....	pág.	14
D-III. Obtención de la lista de valores de expresión/estructura .....	pág.	16
D-IV. Construcción de tabla (data.frame) para cada gen .....	pág.	17
D-V. Agrupación de estructuras atendiendo a los valores de expresión génica .....	pág.	18
V-A) análisis Clustering de k-means .....	pág.	18
V-B) Clustering jerárquico de Ward (para las estructuras cerebrales) .....	pág.	19
V-C) Bioclustering (Heatmap) .....	pág.	19
D-VI Comparación cualitativa de los resultados con la Ontología clásica .....	pág.	20
<b>E. DISCUSIÓN .....</b>	<b>pág.</b>	<b>21</b>
E-I. Validación estadística: Clustering k-means .....	pág.	21
E-II. Validación estadística: Clustering jerárquico y Bioclustering .....	pág.	21
E-III. Validación biológica: “Comparación con la ontología clásica” .....	pág.	23
<b>F. CONCLUSIONES .....</b>	<b>pág.</b>	<b>26</b>
<b>G. REFERENCIAS .....</b>	<b>pág.</b>	<b>27</b>
<b>H. ANEXOS .....</b>	<b>pág.</b>	<b>27</b>

\*\* Para una facilidad de manejo en la lectura y revisión del proyecto, en este trabajo puede dirigirse a cualquier apartado del trabajo pulsando en el índice encima de aquel apartado que quiera revisar, y para volver al índice sólo tiene que señalar la letra donde se encuentra. Además si en el apartado de referencias pulsa el número de la referencia bibliográfica, le dirige a la parte donde la reseña se ha mencionado en el documento.

## A. RESUMEN

La organización estructurada del cerebro desempeña un papel clave en su eficacia funcional. Esta organización es la consecuencia de la identidad molecular única de cada población neuronal establecida gradualmente bajo el control de la expresión génica durante el desarrollo. Actualmente, los estudios basados en criterios tanto moleculares como morfológicos o estructurales, están empezando a revelar cómo los patrones de expresión génica se relacionan con la diferenciación celular y el desarrollo estructural tanto en el espacio como en el tiempo.

En este trabajo mi objetivo es la realización de un estudio sobre la relación entre los valores de expresión génica y la neuroanatomía en el cerebro, considerando múltiples genes pertenecientes a determinadas familias funcionales.

Con este fin, he obtenido dichos valores de expresión génica por cada estructura cerebral o neuroanatómica a partir de los datos de experimentos de hibridación *in situ* ofrecidos por el proyecto "Allen Brain Atlas". He usado esta información para construir tablas que he procesado con técnicas estadísticas de clustering o agrupamiento no jerárquico (*k-means*) y jerárquico.

He obtenido así una clasificación de las estructuras cerebrales atendiendo a sus patrones de expresión génica. Finalmente, para habilitar los resultados de forma visual, he trasladado los clusters o grupos resultantes de estructuras cerebrales sobre un mapa gráfico de un plano sagital medio del cerebro.

Mis resultados muestran que los grupos obtenidos en gran medida se corresponden con los datos ofrecidos con la ontología anatómica clásica del cerebro, pero con ciertas variaciones que podrían ser analizadas en futuros estudios, ya que estas variaciones están basadas estrictamente en criterios moleculares o de expresión génica.

## B. INTRODUCCIÓN

El cerebro es la parte del sistema nervioso central (SNC) que se encuentra dentro del cráneo. Su complejidad abarca muchos aspectos funcionales diferentes relacionados con el control de la mente, los órganos y otros sistemas que componen el organismo, lo cual se refleja en su organización anatómica o estructural, que se divide en una amplia variedad de regiones y subregiones que contienen diferentes tipos celulares o poblaciones de neuronas con características morfológicas, moleculares y conectividad específicas y que son funcionalmente distintas. [1]

Una de las cuestiones básicas de la Neurobiología es responder a la pregunta de "cuántas partes posee el cerebro" y "cómo están articuladas espacialmente unas con otras". Ya que las numerosas regiones del cerebro no existen desde un principio, sino que se van formando progresivamente, se hace preciso estudiar la regionalización durante el desarrollo embrionario, analizando los procesos controlados genéticamente que conducen a la progresiva diferenciación de sus partes.

En la regionalización anatómica clásica, el cerebro se desarrolla de las tres vesículas primarias del tubo neural (primordios de los futuros Romboencéfalo, Mesencéfalo y Prosencéfalo). Los hemisferios cerebrales derivan de una subdivisión del Prosencéfalo denominada Telencéfalo. [2]

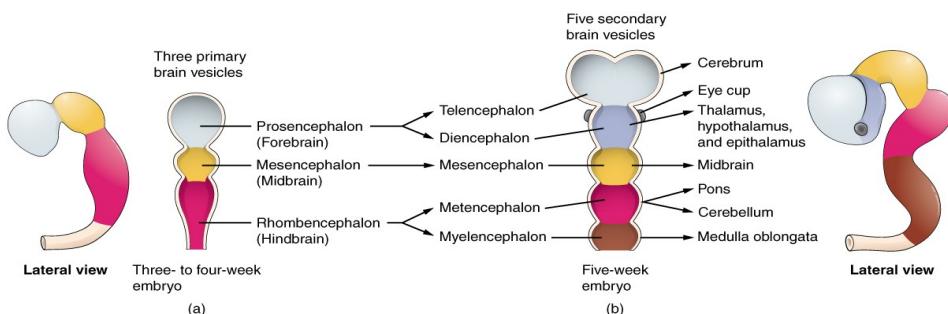


Figura 1: Desarrollo de la regionalización del cerebro en la etapa embrionaria, mediante la subdivisión del tubo neural en vesículas (a) La etapa de vesícula primaria tiene tres regiones, y (b) la etapa de vesícula secundaria tiene cinco regiones [3]

Se reconocen 3 subdivisiones principales en el cerebro adulto:

- **Tronco encefálico o “Brainstem”:** Es la mayor ruta de comunicación entre el Cerebro anterior, la Médula espinal y los Nervios periféricos. Consiste en tres subdivisiones:

1) Mielencéfalo:

- Bulbo raquídeo, Médula oblongada, o “Medulla” .

2) Metencéfalo:

- Protuberancia anular, Puente de Varolio o “Pons”.

- Cerebelo o “Cerebellum”.

3) Mesencéfalo o “Midbrain”

- Diencefalo**

- Tálamo
- Hipotálamo

- Prosencéfalo secundario**

- Telencéfalo
- Porción anterior

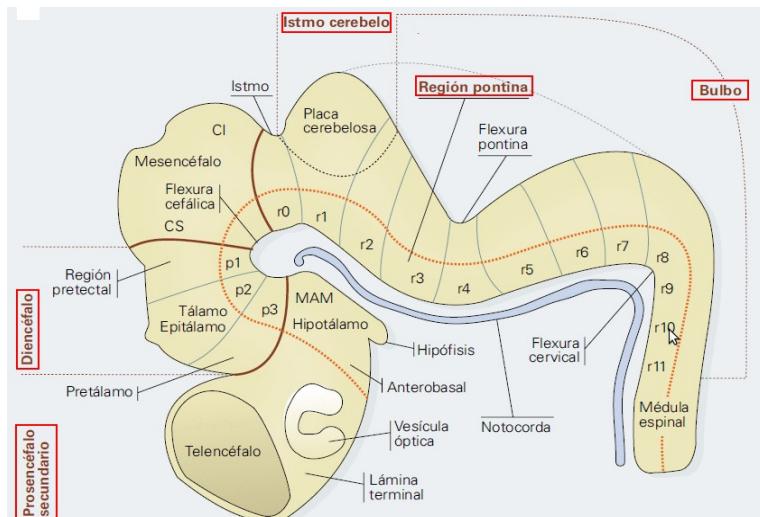


Figura 2: Formación de las vesículas y segmentación (24-36 días de gestación). Segmentación neural [4]

Durante el desarrollo temprano del cerebro, la expresión de diferentes genes marca los límites para definir diferentes regiones dentro de un área más grande. Por ejemplo, la segmentación de la parte posterior del cerebro embrionario contiene doce compartimentos de rostral a caudal. El primer compartimento, al lado del cerebro medio, se llama Istmo. Los restantes once compartimentos se llaman rombómeros (de r1 a r11 en Fig. 2). Los rombómeros 2 a 11 se definen por la expresión de

diferentes genes Hox. El Istmo se define por la expresión del gen FGF8 y el Cerebelo está formado por extensiones del Istmo y el primer rombómero. Las partes media (Mesencéfalo) y anterior (Prosencéfalo) siguen asimismo procesos de regionalización similares.<sup>[5]</sup>

A pesar de los avances de la neurociencia moderna, se carece de una comprensión completa de cómo estos compartimentos cerebrales se especifican o de cómo las diferencias estructurales se pueden traducir en las diversas funciones que desempeña el cerebro.

Históricamente, los proyectos de investigación dedicados al estudio de la morfología y el desarrollo embrionario, han sido estudios a nivel macroscópico, delimitando las regiones que son visibles a simple vista, o microscópicos a nivel cito- e histológico; en estos estudios por limitaciones técnicas no usaban los criterios moleculares de expresión génica. [6]

En proyectos actuales se estudia la “genearquitectura neural”, que se refiere a la descripción de la estructura neural en términos de patrones de expresión génica [4]. Este enfoque anatómico emergente está representado de forma masiva por diversos proyectos en desarrollo ofrecidos por: el Instituto Allen para la Ciencia del Cerebro (<http://mousespinal.brain-map.org>), Eurexpress (<http://www.eurexpress.org>) o GENSAT (<http://www.gensat.org>), entre otras fuentes. Este concepto novedoso, se ha añadido a los conceptos clásicos de quimioarquitectura (definida por la expresión de marcadores de actividad bioquímica) y de citoarquitectura (definida por la morfología celular de las neuronas en la región de estudio) [7]

El presente trabajo intentará resolver la cuestión fundamental de si la regionalización por patrones de expresión génica (considerando diversas familias de genes) nos dará un resultado similar a la regionalización clásica. [8]

Recientemente, también hay otros proyectos de investigación que hacen estudios relacionados analizando la información disponible en el Allen Brain y otras bases de datos, con formas y finalidades diferentes, tales como los grupos de: (1) Montoya Villegas,J.C (2012) que utiliza un grupo discreto de genes para su estudio computacional (in silico) [9], (2) Ko(2013), que hacen agrupación de regiones cerebrales por k-means con grupos reducidos de genes, de neuronas y de glia, trabajando con grupos de pixeles en vez de estructuras cerebrales identificadas [10]; (3) Shuiwang Ji(2013), cuyo grupo usa datos de expresión de genes obtenidos del Allen, usando t-SNE (t-distributed stochastic neighbour embedding) y técnicas de imagen para mapear dichos genes individualmente [11]; o (4) Jeremy A. Miller (2014), que estudia la anatomía y función cerebral a través de procesos transcripcionales del embrión, utilizando técnicas de imagen con MRI (Magnetic resonance imaging) [12].

El método que desarrollaremos aquí, permitirá eventualmente llevar a cabo una nueva aproximación experimental para clasificar y agrupar las regiones del cerebro según su similitud genética. El proyecto está enfocado al uso de herramientas computacionales, que permitan extraer información de niveles de expresión de genes a lo largo de las distintas estructuras cerebrales, a partir de bases de datos que contienen información integrada de miles de genes, obtenida a través de experimentación basada en la técnica de hibridación in situ.

A pesar de que existen varias bases de datos transcripcionales, como la creada por Paxino (Paxino, Assheuer y Mai, 2003), la base de datos procedente del Instituto Allen de Ciencias Cerebrales. proyecto online del “Allen Brain Atlas”(<http://www.brain-map.org/>) es la más completa y además está disponible de forma gratuita para su utilización. Por esta causa, los valores de expresión serán extraídos de la información disponible en este proyecto Allen.

Seguidamente estudiaremos la asociación o clustering de dichas estructuras en base a dichos valores de expresión cuantitativos. Para finalmente comprobar si la clasificación obtenida usando el método propuesto es similar a la ontología neuroanatómica clásica, o se correlaciona con ella, comparando los resultados de distintas familias de genes.

De tal modo que se trata de un estudio a una escala mayor que los análisis de genes individuales o grupos pequeños de ellos, desarrollando una técnica que potencialmente permitirá incluso usar todos los genes del genoma, mediante un método sencillo, y nuevo, ya que no se conoce ningún trabajo que haya enfocado de esta forma.

Ésto servirá para la clasificación de las regiones o poblaciones neuronales del cerebro a través de los datos de expresión de múltiples genes, mediante el cual se podrán usar todos los grupos ontológicos de genes existentes.

## C. MATERIALES Y MÉTODOS

### C-I. Ontologías

El objetivo de una ontología es describir los conceptos dentro de un dominio y la relación que existe entre esos conceptos. Se utilizan para definir vocabularios que puedan entender los ordenadores y que su precisión sea tan específica que sean capaces de permitir diferenciar términos y referenciarlos de manera precisa. [13]

En el trabajo hemos usado dos tipos diferentes de ontologías: C-I.1 Gene Ontology, para genes C-I.2 Ontología para las estructuras del cerebro.

#### C-I.1 Gene Ontology

El proyecto de Gene Ontology (GO) (<http://www.geneontology.org/>) está dirigido a la estandarización de la anotación de productos génicos mediante vocabularios estructurados, y es un esfuerzo de colaboración para hacer frente a la necesidad de descripciones consistentes de productos de genes en diferentes bases de datos; es decir, se basa en la descripción del gen y los atributos del producto genético de cualquier organismo. El proyecto comenzó en 1998 como una colaboración entre tres bases de datos organismo modelo: FlyBase (Drosophila), Saccharomyces (SGD) y genoma de ratón (MGD).[14] [15]

El proyecto GO agrupa realmente tres ontologías que se corresponden con tres aspectos diferentes de la biología de forma independiente, lo cual facilita la realización de consultas uniformes y específicas a través de todos ellos: (1) describen productos de genes en términos de sus procesos biológicos asociados, (2) componentes y localización celular y (3) funciones moleculares.

Todos los términos de GO tienen un nombre y un identificador único de la forma GO:nnnnnnn, la mayoría de ellos con una definición textual, con referencia a la fuente donde fue descrito. Los términos que se consideran obsoletos se marcan como 'obsoleto', pero tanto el término como su identificador se mantienen en la base de datos de GO; por lo general, se añade un comentario que explica su caducidad y se sugiere un término actual para remplazar el término obsoleto.

El punto de partida de este trabajo es hacer uso de la ontología GO para agrupar los genes de interés en familias funcionales. Se podrá usar cualquier término GO válido contenido en la página del proyecto de Gene Ontology. Con estos términos GO estudiaremos después la clasificación obtenida de las subregiones del cerebro según la expresión de los genes que pertenezcan a dichas familias.

#### C-I.2 La anatomía del cerebro representada como ontología de estructuras

La visión actual del cerebro es representarlo como una ontología, se trata de considerarlo como dominios que van subdividiéndose en ramas y hojas. Éstas múltiples subdivisiones estructurales de la ontología permiten que se vayan generando los siguientes niveles (elementos hijos), lo que refleja la subdivisión de la correspondiente estructura en varias subestructuras. [11]

En la imagen se muestra como se dividen las distintas regiones del cerebro siguiendo la ontología clásica de regionalización que se puede ver en una de las ontologías anatómicas del cerebro existentes actualmente, que encontraremos en la página del Allen-Brain <http://atlas.brain-map.org/>. Más adelante iremos explicando que esta imagen la usaremos para la comparación cualitativa con los resultados obtenidos.

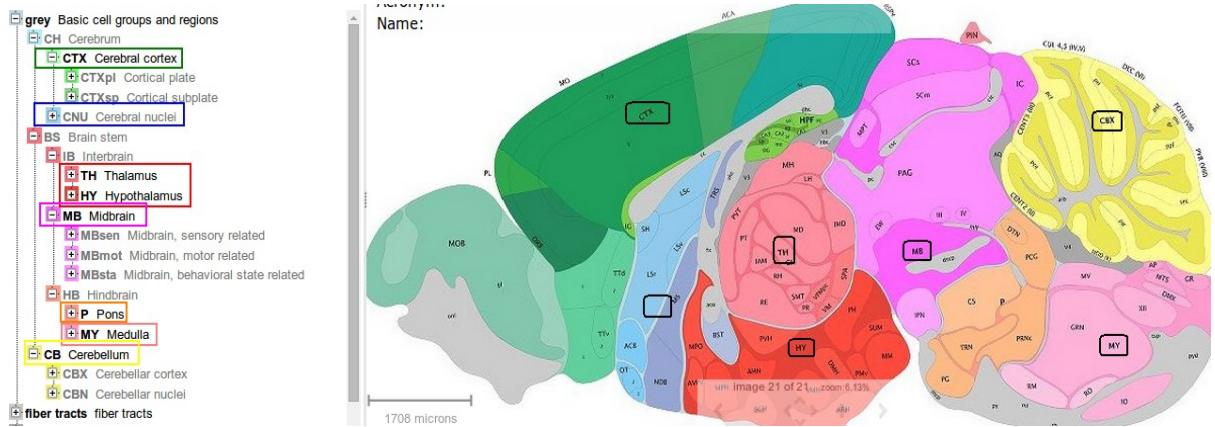


Figura 3: Tomada de la página Allen Brain Atlas, representa un corte sagital del cerebro por su zona media, y como vemos los colores están dados según la subdivisión de las regiones anatómicas clásicas:

- Prosencéfalo que incluye en los tonos verdosos el Telencéfalo con la Corteza, el Bulbo olfatorio y el Hipocampo; y en tonos azulados incluye los Ganglios basales.
- Diencéfalo: en tonos rojizos, dividido en Tálamo (rojo claro), e Hipotálamo (rojo intenso).
- Mesencéfalo: en tono lila.
- Romboencéfalo: tiene tres subdivisiones: en naranja el Puente, en amarillo el Cerebelo y en rosa la Médula oblongata.

## C-II Base de datos Allen Brain Atlas.

El origen del proyecto “Allen Brain” surgió de una serie de encuentros promovidos por Paul G. Allen, el co-fundador de Microsoft, en el 2001. La idea principal es crear un mapa cerebral en base a resultados de experimentos actuales, en función de la actividad de los genes del cerebro. Allen donó 100 millones de dólares para fundar el Instituto Allen.

Este proyecto ofrece principalmente una interfaz para el usuario donde se pueden observar y descargar las fotografías de las secciones cerebrales procesadas con Hibridación in situ (ISH) para los distintos genes; además de las API (Application Programming Interface) destinadas a programadores.

### C-II.1 Detección de expresión génica mediante Hibridación in situ.

En las últimas décadas, los estudios de la organización anatómica del cerebro se han basado principalmente en los patrones de expresión génica, asumiendo que las poblaciones neuronales que comparten la expresión de determinados genes pueden considerarse como derivadas del mismo territorio embrionario y pertenecientes a la misma región anatómico-funcional. Esta perspectiva ha dado lugar al concepto de *gene arquitectura* cerebral, mencionada anteriormente [4].

Una de las herramientas utilizadas, relativamente novedosa, para estudiar la expresión génica de cada gen conocido, es la hibridación in situ (ISH); técnica que estudia la distribución y densidad de un gen o molécula de ARN en una célula o un tejido utilizando una sonda de ADN o ARN de una sola cadena, que se une al tejido diana por complementariedad de bases (específica para cada gen). Aunque la ISH no es una técnica de alto rendimiento como los microarrays (para analizar la expresión diferencial de genes, monitorizando de manera simultánea los niveles de miles de ellos), esta técnica es ampliamente utilizada por su ventaja de permitir observar la morfología o estructura del tejido estudiado.

Multitud de proyectos en la actualidad hacen uso de esta técnica, cuyo objetivo es cubrir el patrón de expresión de ISH del genoma completo en sistemas modelo como el embrión de ratón (Eurexpress) y el cerebro de ratón y el cerebro humano (Allen brain). Así, su utilidad más importante actualmente es estudiar la subdivisión de distintas regiones, mediante el estudio de la expresión de los genes diferentes que se expresan en las diferentes estructuras del tejido.

Esta técnica consiste básicamente en los siguientes pasos: [16] [17]

(1) Diseño del ADN complementario (ADNc) que corresponde a la secuencia de gen dado, y su inserción en un vector plásmido adecuado, (2) Síntesis in vitro de una ribosonda con la RNA polimerasa tomando como molde la secuencia de ADNc, y el uso de una mezcla de nucleótidos que se encuentran marcados con digoxigenina o fluoresceína, (3) Fijación y procesamiento de la muestra biológica (órgano, tejido, embrión, etc) para ser analizados; (4) Hibridación de la ribosonda con muestra biológica, y los lavados posteriores. (5) Recubrimiento de la molécula sonda marcada (digoxigenina o fluoresceína) con un anticuerpo acoplado a una enzima, la fosfatasa alcalina normalmente, o a un marcador fluorescente. (6) Visualización de la actividad enzimática con los sustratos adecuados, o la visualización de la marcador fluorescente, de manera que las células que expresan originalmente el ARNm aparecen marcadas.

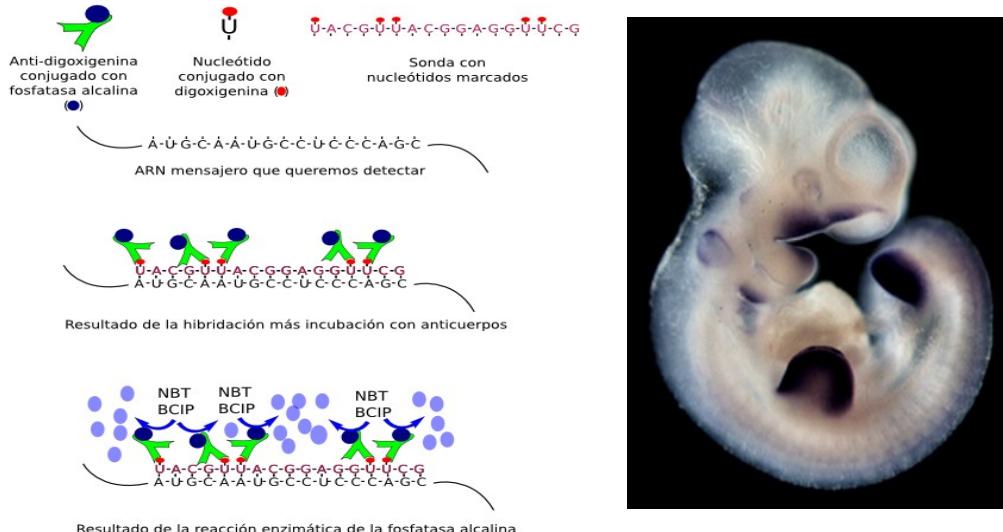


Figura 4: a) A la izquierda observamos la representación del proceso de hibridación in situ: NBT (nitro blue tetrazolium) y el BCIP (5-Bromo-4-chloro-3-indolyl fosfato, sal de toluidina) son los sustratos de la fosfatasa alcalina [18]. b) En la imagen de la derecha observamos la tinción por hibridación in situ en un embrión, representada en color violeta oscuro.

En el proyecto Allen, el resultado de la Hibridación in situ es:

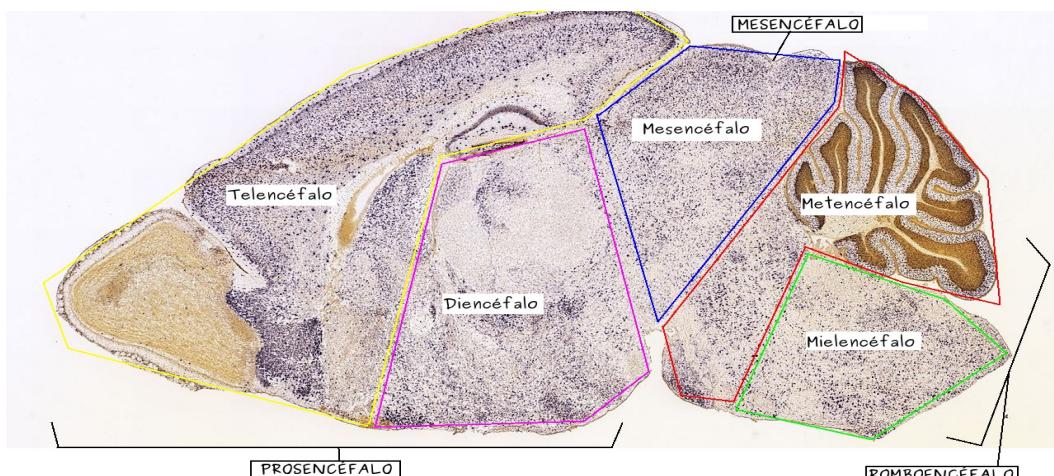


Figura 5: Representa un corte salgital del cerebro, resultado de una tinción por hibridación in situ de la Base de datos del Allen Brain.

Los puntos con una tinción más intensa (violeta oscuro), es debido a que se encuentra un número mayor de células positivas juntas y con mayor nivel de expresión del gen estudiado. Realmente lo que se tiñe es el soma neuronal, lo cual nos permite ver el color violeta. Las zonas sin tinción (color amarillento y marrón, que es debido a una contratinción para poder observar todo el tejido), son células negativas o fibras. En este trabajo estudiaremos cortes sagitales del cerebro de ratón (resultado de dividir el cerebro en dos imágenes especulares).

Posteriormente en el Instituto Allen, usan microscopios robotizados que fotografian un millón de estos cortes histológicos cerebrales, ofreciendo para cada gen buscado la posibilidad de visualizar las fotografías en alta resolución de las secciones procesadas por ISH. El resultado es que estas fotos son usadas para llenar una base de datos que construirá una imagen digital del cerebro en el que se ven finalmente los patrones de expresión génica.

Podremos observar en la página web (<http://mouse.brain-map.org/>) como mínimo un experimento para cada gen, con un corte de cerebro en el plano sagital, mientras que para otros genes hay dos cortes procesados, respectivamente, uno en plano sagital y otro en plano coronal; algunos genes presentan un número mayor de experimentos, debido a que han sido genes de mayor interés para los investigadores, así que se han hecho más experimentos para comprobar la veracidad de los resultados.

### C-II.2 Uso de las API

Gracias a las API(Application Programming Interface), un programador no necesita preocuparse de cómo funciona una aplicación remota ni de la forma en que las funciones fueron implementadas, para poder utilizarla en un programa.

En la página web del Allen, ofrecen para el usuario la interfaz gráfica, que ofrece acceso a los datos publicados a través de una interfaz de API, y todos estos recursos están asociados a una URI, que se puede utilizar para obtener información de forma que se pueda manipular.

Para realizar una minería de la base de datos Allen, he usado parte de las API ofertadas por dicha base de datos, de genes que me interesaban, dentro de las cuales he buscado la información final de valor de expresión génica para toda una familia de genes, en cada una de las estructuras cerebrales.

La API proporciona datos sobre los recursos disponibles en distintos formatos: JSON (.json), formatos de datos XML (.xml) o CSV (.csv). Cuando no se indique el formato en la consulta, el predeterminado es JSON, y en el trabajo será utilizado este tipo de formato.

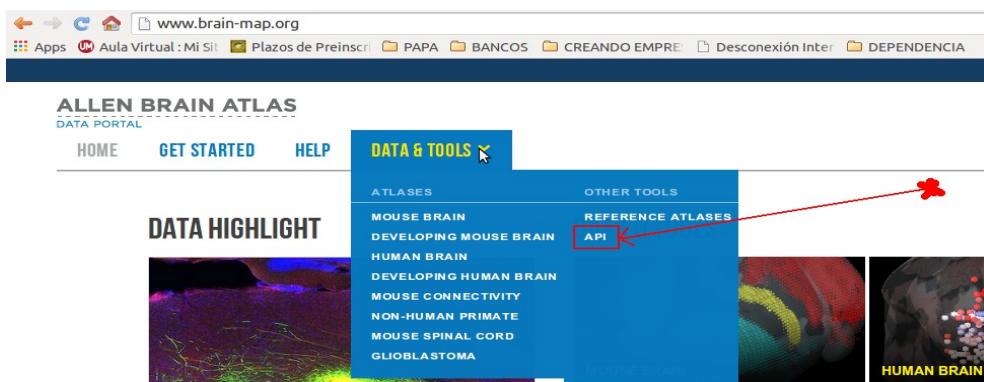


Figura 6: Puedes acceder a todos los tipos de API que se pueden encontrar en la página de Allen Brain Atlas, en el recuadro señalado en rojo.

De la página Allen, obtenemos la ontología de todas las estructuras del cerebro, con formato 1.json en formato de API (El 1 es porque nos estamos refiriendo a las regiones del cerebro, (si fuese 2 nos referiríamos a la médula espinal), recogido en el archivo de texto 'jsonn.txt'.

Ha sido descargada como un archivo .json jerárquicamente estructurado, mediante esta API ([http://api.brain-map.org/api/v2/structure\\_graph\\_download/1.json](http://api.brain-map.org/api/v2/structure_graph_download/1.json))

Nuestro objetivo a partir de esta ontología será la obtención de la información en un documento que contendrá el nombre de todas las estructuras 'hojas' (children: [ ]).

Si tenemos "children: [ ]" → significa que es una hoja.

Si aparece "children: [ ]" → está abierto y eso significa que tiene hijos, y por tanto se trata de una rama.

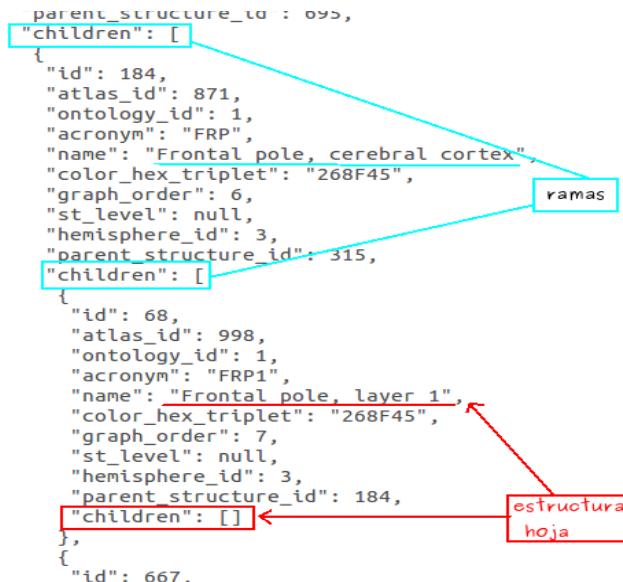


Figura 7: fracción del extracto jsonn.txt que contiene la ontología, y en la cual se puede observar la estructura del fichero.

En nuestro trabajo hemos querido recoger todas los 'nodos hojas' (que son aquellos que no tienen más subdivisiones), porque el conjunto de todas ellas es un conjunto homogéneo y universal que engloba todo el cerebro.

### C-III . Proyecto Bioconductor

Es un proyecto similar a BioPerl (colección de módulos de Perl que facilitan el desarrollo de scripts en Perl para aplicaciones de bioinformática, de software libre apoyado por la Open Bioinformatics Foundation), es un proyecto de código abierto para el análisis de datos Genómicos y de Biología molecular que utiliza en este caso el lenguaje de programación estadística R (<http://www.bioconductor.org>). Bioconductor provee de herramientas para cada paso del proceso de análisis (cargar, preparar y analizar los datos).

En mi proyecto esta herramienta ha servido para usar en R el paquete “BiomaRt”, el cual me ha ofrecido una interfaz con las herramientas de Biomart, o el paquete “limma”, que sirve para el análisis de datos de expresión génica de microarrays, especialmente para el uso de modelos lineales en el análisis de experimentos diseñados y la evaluación de las diferencias de expresión, el cual me ha permitido realizar la normalización de los datos.

### C- IV. BiomaRt

Proporciona una interfaz que comunica con una creciente colección de bases de datos, que se puede ver en <http://www.biomart.org>. Permite la recuperación de grandes cantidades de datos de una manera uniforme, sin la necesidad de conocer los esquemas de bases de datos subyacentes o escribir consultas SQL complejas. Ejemplos de bases de datos BioMart son Ensembl, Uniprot y HapMap. Estas grandes bases de datos ofrecen a los usuarios de BioMart el acceso a un conjunto de datos muy diverso y permiten una gran cantidad de consultas de R.

Las preguntas a biomaRt (biomaRt queries), se componen de 3 componentes principales:

1. Filters (filtros): es un vector que se usa como “input” de la consulta, y que componen la restricción a la consulta.
2. Attributes (atributos): es un vector, que contiene los atributos que se quieren recuperar (por ejemplo: “simbolos de genes”, o “coordenadas cromosómicas”).
3. Values (valores): es un vector de valores de los filtros. El argumento de valores requiere una lista de valores, donde cada posición en la lista corresponde a la posición de los filtros en “filters”.

El formato en el que me guardará los resultados puede ser CSV, HTML, TSV o XLS. En mi trabajo he usado fundamentalmente el formato TSV, que es el que se obtiene al usar la interfaz que tiene BiomaRt con Perl.

The screenshot shows the BiomaRt interface on the Ensembl website. On the left, there are filters for Dataset (Mus musculus genes (GRCm38.p2)), Filters (Gene type: protein\_coding, GO Term Name [e.g. regulation of biological process]: axon guidance), and Attributes (Ensembl Transcript ID, Associated Gene Name). The 'Results' button is highlighted with a pink box. In the center, there's a form to 'Please select columns to be included in the output and' with radio buttons for Features, Homologs, Structures, Variation, Transcript Event, and Sequences. Below it are sections for GENE, EXTERNAL, and PROTEIN DOMAINS AND FAMILIES. To the right, there are buttons for 'Export all results to File' and 'Email notification to'. A pink box highlights the 'CSV' button. At the bottom, a table lists Ensembl Transcript IDs and Associated Gene Names, with a pink box highlighting the first row: ENSMUST00000095987, Neurog2.

Ensembl Transcript ID	Associated Gene Name
ENSMUST00000095987	Neurog2
ENSMUST00000104875	Olftr160
ENSMUST00000032561	Vasp
ENSMUST00000065086	Gas1
ENSMUST00000022262	Fezf2
ENSMUST00000095012	Sema3a
ENSMUST00000125629	Sema3a
ENSMUST00000137798	Sema3a
ENSMUST00000030714	Sema3a
ENSMUST00000030714	Sema3a

Figura 8: Representa la interfaz gráfica de biomaRt:Ensembl, donde se ponen los filtros y atributos que nos interesa encontrar dentro de la base de datos que le especificamos también.

En el presente trabajo esta herramienta se ha usado para obtener las familias de genes de interés, a través de términos de GO válidos. El resultado obtenido se ha utilizado para obtener los documentos del Allen con los experimentos asociados para cada gen. Sin embargo, no ha sido usada como tal, sino a través del script1 que he desarrollado.

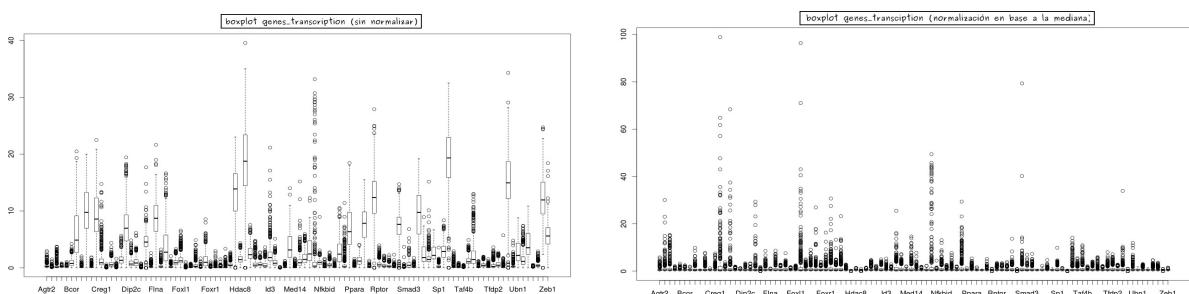
### C-V Herramientas estadísticas:

#### A) Normalización de los datos:

Normalizar es transformar una variable aleatoria que tiene alguna distribución en una nueva variable aleatoria con distribución normal o aproximadamente normal.

La normalización que me interesa realizar, es la que iguala las medianas, pero que conserva las variaciones, por eso usamos el método <normalizeMedianAbsValues>. Así mantiene las variaciones de expresión de cada gen según la estructura, y como hay genes que se expresan más en unas regiones que en otras, se seguirá observando esta diferencia de expresión. Esta normalización se ha aplicado para todos los datos, en un paso justamente anterior a la agrupación de estructuras y genes.

Se ha realizado un diagrama de cajas para observar la expresión génica de la familia de genes estudiada (en este caso *transcription*).



a) Boxplot representando en el eje x los genes, y en el eje y sus respectivos valores de expresión en las diferentes estructuras

b) En una normalización en base a las medianas, se puede ver la variación de expresión de cada uno de los genes intacta.

Figura 9: Representación del diagrama de cajas, sin normalizar y normalizado de los genes de la transcripción

## B) Clustering:

Las técnicas de Clustering (Análisis de Conglomerados) han probado ser de gran utilidad en los análisis de microarrays a la hora de descubrir grupos de genes que intervienen en una misma función celular o que están regulados de la misma manera, así como para la clasificación de pacientes de una enfermedad determinada según sus transcriptomas respectivos [19]. Estas técnicas tienen por objetivo agrupar los objetos de interés en grupos homogéneos (parecidos = mismo grupo) y heterogéneos (diferentes = distinto grupo), de forma que los objetos clasificados en un mismo grupo son similares según los criterios que se establezcan, que en nuestro caso serán las estructuras anatómicas regionalizadas del cerebro. [20]

### 1) Clustering de k-means de las estructuras cerebrales

En ocasiones interesa que el resultado sea una partición del conjunto de los objetos a clasificar de forma que cada objeto pertenece a uno y solo a uno de los grupos. Las técnicas que consiguen este objetivo se denominan no jerárquicas y se ejecutan mediante el llamado Algoritmos de k-medias. [21]

El algoritmo de k-medias produce una partición del conjunto de los n individuos en k grupos, donde k es un valor entero prefijado de antemano, es decir, que el usuario elige los grupos en los que le interesa particionar el conjunto de estructuras cerebrales en este caso.

Esta es la herramienta principal usada en este trabajo, con la que he llegado a los resultados finales, que me permitirán obtener la clasificación de estructuras cerebrales según sus patrones de expresión génica.

### 2) Clustering jerárquico de las estructuras cerebrales .

También es posible producir una estructura de datos en forma jerárquica, produciendo distintas particiones del conjunto de objetos a clasificar, en función del nivel de similitud entre grupos. Estas técnicas se llaman jerárquicas.

Para el agrupamiento jerárquico de las estructuras cerebrales, el Análisis de Clustering Jerárquico comienza separando cada estructura cerebral (hoja) en un clúster por sí mismo. En cada etapa del análisis, el criterio por el que las estructuras del cerebro son separadas, se relaja para enlazar los dos conglomerados más similares hasta que todos las estructuras sean agrupadas en un árbol de clasificación completo (dendograma).

El criterio básico para cualquier agrupación es la distancia. Las estructuras que estén cerca pertenecerían al mismo conglomerado o cluster, y los datos que estén más alejadas pertenecerán a distintos clusters. Encontramos muchas distancias propuestas ("euclidean", "maximum", "manhattan", "canberra", "binary" or "minkowski"). En el proyecto he usado la medida de la distancia Euclídea, debido a que es la más intuitiva y la más utilizada, que calcula la distancia como una línea recta entre dos clusters, siendo una aplicación del Teorema de Pitágoras. Se ha hecho mediante la realización de una matriz de distancias, con la función dist().

La función hclust() de R realiza el Clustering jerárquico utilizando el criterio de agregación especificado, tiene que ser uno de los siguientes: ("ward", "single", "complete", "average", "mcquitty", "median" or "centroid". En el trabajo se ha usado el método ward. [22]

El motivo de haber escogido el método de Ward (Ward.D2 para la versión de Rstudio 3.1.0), es porque este método es un criterio que se aplica en el análisis de agrupamiento jerárquico como método de la varianza mínima de Ward; es un caso especial del enfoque de la función objetivo originalmente presentado por Joe H. Ward, Jr. [23]. Ward, sugirió un procedimiento de agrupamiento de aglomeración jerárquica, donde el criterio para elegir el par de racimos, surgiese de fusionar en cada paso el valor óptimo de una función objetivo, en el que se mantengan las varianzas lo máximo posible.

### 3) BiClustering (Heatmap)

Recientemente, el BiClustering ha sido propuesto como método para descubrir patrones de

comportamiento específico en los que el valor de expresión de un subgrupo de genes evoluciona de la misma forma a lo largo de un subgrupo de condiciones de laboratorio [24]. Se trata de un método de minería de datos que agrupa simultáneamente ambas condiciones (genes y estructuras cerebrales en este caso) como filas y columnas de una matriz. En el trabajo se ha realizado con la técnica de heatmap.

Un “heatmap” es una representación gráfica de los datos, donde los valores individuales están contenidos en una matriz y se representan como colores. En el heatmap los niveles de mayor expresión génica se representan más claros/calientes (blancos/amarillos), y los de menor expresión más oscuros/fríos (rojos). [25]

Las filas (estructuras cerebrales) y las columnas (genes) van a ser agrupadas utilizando la función de cluster `hclust()` con sus opciones por defecto, como la distancia euclídea, y con los datos previamente normalizados. Así podemos ver los genes con mayor expresión dentro de la familia de genes, en los cuadrados que le corresponden con respecto a la estructura que representan.

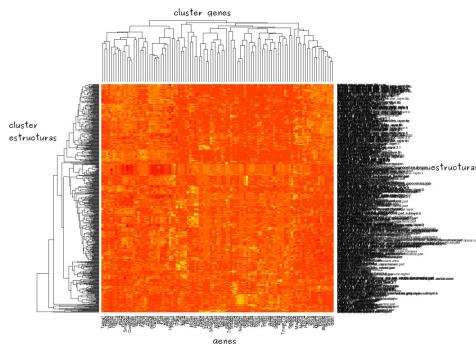
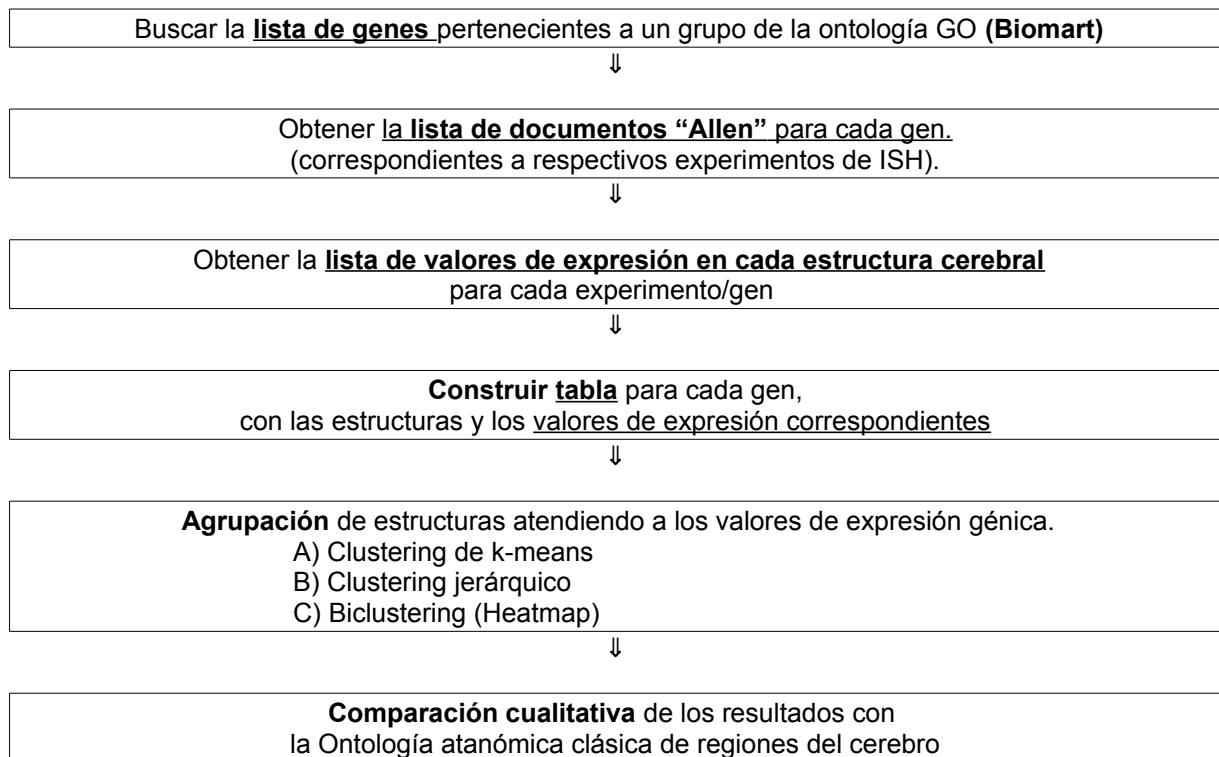


Figura 10: Heatmap de estructuras y genes.

#### C-VI. Método seguido en este trabajo

El esquema de la método propuesto para la clasificación de regiones del cerebro según sus valores de expresión, consistirá en :



## C-VII. Scripts desarrollados para llevar a cabo el método descrito

Todos los scripts serán guardados y ejecutados en la misma carpeta.

→ Script 1: script: step1\_R\_Biomart\_OK.pl

Obtiene la lista de genes a partir del término GO que nos interesa.

Este script de Perl hace uso de una interfaz R (R::Statistics package) y la biblioteca BioMart de R (Durinck et al., 2009) con el fin de conectarse a funciones BioMart (<http://www.biomart.org/>), el cual permite la recuperación de información de múltiples bases de datos biológicas.

→ Script 2: step2\_get\_allen\_exp\_from\_genes\_OK.pl

Obtiene el documento con toda la información de experimentos realizados para cada gen en formato XML, desde la página web Allen Brain Atlas. El programa de Perl, va cogiendo cada uno de los genes de la lista obtenida en el paso anterior, y va guardando ficheros con el nombre 'gen\_nombre del gen', por ejemplo gen\_Gata6. En un paso posterior guarda los ficheros del Allen asociados a cada uno de esos genes con el nombre 'allen\_nombre del gen\_id del experimento', por ejemplo allen\_Gata6\_69289033. (Para reproducir el trabajo con el término GO 'transcription' este script ya está modificado y contiene la lista de genes para hacer que este script funcione)

→ Script 3: step3\_value\_final\_onto1.pl

Este script me ha cosechado todos los 'nodos hojas'. En total son 930 estructuras cerebrales, y se han guardado en el documento jsonn.txt. La ejecución de este script puede ser realizada una sola vez en el proceso, ya que el documento obtenido es único.

→ Script 4: step4\_get\_value\_final\_OK.pl

Como para cada gen, necesito la lista de estructuras cerebrales y la lista de valores de expresión para cada estructura, este script de Perl coge los ficheros de Allen descargados para cada gen (obtenidos con el script 2, y el fichero con solamente los 'nodos hojas' obtenido con el script 3). Entonces con cada gen de interés obtendrá los valores de expresión de las estructuras cerebrales 'hojas' que son válidas.

→ Script 5: step5TFM.R

Utiliza los valores y estructuras obtenidos para cada gen en el paso anterior (step4\_get\_value\_final\_OK.pl), con cada uno de los términos GO, que obtuvimos en los pasos anteriores.

## D. RESULTADOS

### D-I Búsqueda de la lista de genes (Biomart)

Procedimiento mediante el cual utilizo un GO term válido, del proyecto de Gene Ontology (GO), para conseguir la lista de genes perteneciente a dicho término GO.

Los GO term que he usado para este trabajo son 'axon guidance', 'cerebral cortex radially oriented cell migration' y 'transcription'.

Un ejemplo de API de Perl-BiomaRt para sacar el gen es con el GO term *axon guidance* donde se pueden ver las preguntas a BiomaRt, lo muestro a continuación:

```
my $query = BioMart::Query->new('registry'=>$registry, 'virtualSchemaName'=>'default');

$query->setDataset("mmusculus_gene_ensembl");
$query->addFilter("go_parent_name", ["axon guidance"]);
$query->addFilter("biotype", ["protein_coding"]);
$query->addAttribute("external_gene_id");
```

El término GO '*transcription*' ha recibido la calificación "obsolete" atendiendo a una nueva clasificación, así que para hacer posible la reproducibilidad del trabajo con este término, he adjuntado los ficheros obtenidos en la carpeta de anexos)

Para llevar a cabo la finalidad de este paso, se ha desarrollado el siguiente script:

→ Script 1: step1\_R\_Biomart\_OK.pl

## D-II. Obtención de la lista de documentos “Allen” para cada gen

En este segundo paso, el objetivo es obtener información de los experimentos asociados a cada gen de la lista de genes que hemos conseguido en el paso anterior.

Para ello se ha tenido que coger la información de 2 tipos de API para cada gen, una API que contiene todos los experimentos de cada gen, y otra API que contiene la información del experimento de interés elegido para cada gen. Este paso es el más lento, tarda aprox. 1 minuto por cada gen en descargar el documento de la página del Allen.

1) Para obtener la información de todos los experimentos asociados a un gen, y su estructura es la que se muestra en la Figura 11

[http://api.brain-map.org/api/v2/data/SectionDataSet/query.xml?criteria=products%5Bid\\$eq1%5D,genes%5Bacronym\\$eq%27Pdyn%27%5D&include=genes.section\\_images](http://api.brain-map.org/api/v2/data/SectionDataSet/query.xml?criteria=products%5Bid$eq1%5D,genes%5Bacronym$eq%27Pdyn%27%5D&include=genes.section_images)

Cada experimento se reconoce por <section-data-set>.

En este paso, una parte del método ha sido filtrar los datos de tal manera que consiga aquellos resultados que me interesan de cada gen de la lista, por lo que hago una consulta en Perl para ponerle las condiciones que he considerado para la elección del experimento de interés:

Como he querido trabajar con el plano sagital, el cual se corresponde con el número 2, (el plano coronal se corresponde con el 1), una de las condiciones a cumplir es seleccionar aquellos experimentos que se correspondan con <plane-of-section>2</plane-of-section>.

Otra condición que debe efectuarse en todos los experimentos seleccionados, ha sido la elección de experimentos válidos, es decir que en su realización hayan funcionado de forma correcta: <failed>false</failed>. Aquel experimento que cumpla estas condiciones, será el experimento que me interesa de ese gen y por tanto ello guardo la “id” de ese experimento.

El recuadro rojo de la Figura 11. es donde se pone el nombre del gen, que en este caso es Pdyn y la información que se obtiene es la que corresponde a todos los experimentos asociados a ese gen: <section-data-sets>.

```
<Response success="true" start_row="0" num_rows="2" total_rows="2">
  <section-data-sets>
    <section-data-set>
      <blue-channel nil="true"/>
      <delegate>true</delegate>
      <expression>true</expression>
      <failed>false</failed>
      <failed-facet>734881840</failed-facet>
      <green-channel nil="true"/>
      <id>69782969</id> ●
      <name nil="true"/>
      <plane-of-section-id>2</plane-of-section-id> ●
      <qc-date>2009-05-02T22:48:46Z</qc-date>
      <red-channel nil="true"/>
      <reference-space-id>10</reference-space-id>
      <rnameq-design-id nil="true"/>
      <specimen-id>69370910</specimen-id>
      <sphinx-id>23388</sphinx-id>
    </section-data-set>
    <storage-directory>
      /external/aibssan/production32/prod329/image_series_69782969/
    </storage-directory>
    <weight>5470</weight>
  <genes>
    <gene>
      <acronym>Pdyn</acronym>
    </gene>
  </genes>
</Response>
```

Figura 11 :API con todos los experimentos de 1 gen (Pdyn). Se muestra solo el fragmento inicial con el primer experimento (<section-data-set>). Para este gen aparece un segundo experimento no mostrado aquí.

2) Una vez que sabemos el id del experimento que nos interesa, necesitamos filtrar la información para escoger los valores de expresión del gen y la estructura que representa dicho valor. Este paso se realiza procesando la información obtenida con la siguiente API que se muestra en la Figura 12:

[http://api.brain-map.org/api/v2/data/SectionDataSet/query.xml?id=69782969&include=structure\\_unionizes%28structure%29](http://api.brain-map.org/api/v2/data/SectionDataSet/query.xml?id=69782969&include=structure_unionizes%28structure%29)

```

▼<structure-unionizes>
  ▼<structure-unionize>
    <expression-density type="float">0.0113234</expression-density>
    <expression-energy type="float">1.57936</expression-energy>
    <id type="integer">246037469</id>
    <section-data-set-id type="integer">69782969</section-data-set-id>
    <structure-id type="integer">1</structure-id>
    <sum-expressing-pixel-intensity type="float">110457.0</sum-expressing-pixel-intensity>
    <sum-expressing-pixels type="float">791.934</sum-expressing-pixels>
    <sum-pixel-intensity type="float">2176790.0</sum-pixel-intensity>
    <sum-pixels type="float">69938.0</sum-pixels>
    <voxel-energy-cv type="float">1.0154</voxel-energy-cv>
    <voxel-energy-mean type="float">1.57936</voxel-energy-mean>
  ▼<structure>
    <acronym>TMv</acronym>
    <atlas-id type="integer">424</atlas-id>
    <color-hex-triplet>FF4C3E</color-hex-triplet>
    <depth type="integer">8</depth>
    <failed type="boolean">false</failed>
    <failed-facet type="integer">734881840</failed-facet>
    <graph-id type="integer">1</graph-id>
    <graph-order type="integer">682</graph-order>
    <hemisphere-id type="integer">3</hemisphere-id>
    <id type="integer">1</id>
    <name>Tuberomammillary nucleus, ventral part</name>

```

Figura 12: API con el experimento id (69782969) para el gen antes mencionado (Pdyn). Se muestra un fragmento referido a los valores de expresión de este gen concretamente en la estructura cerebral llamada "Tuberomammillary nucleus, ventral part", una de las aproximadamente 2500 estructuras cerebrales incluidas (número variable según cada gen y experimento).

El recuadro azul es donde se pone el número id del experimento, para sacar la información de ese experimento:

El valor de <voxel-energy-mean> se corresponde al valor de expresión del gen en la estructura correspondiente. De los distintos valores de expresión (p.ej.<sum-expressing-pixel-intensity>, <sum-expressing-pixels>, etc.) este es el más significativo, ya que representa la energía de expresión media por unidad volumétrica (voxel). La estructura específica está en el recuadro <name>. Esto quiere decir que para el gen que corresponde al id=69782969, que en este caso es Pdyn, para la estructura Tuberomammillary nucleus, ventral part, le corresponde un valor de expresión de 1.57936.

Para resolver este paso se ha desarrollado un script de Perl.

→ Script 2: step2\_get\_allen\_exp\_from\_genes\_OK.pl

El resultado obtenido se utilizará en el siguiente paso, para obtener la lista de estructuras y de valores.

**D-III Obtención de la lista de valores de expresión/estructura**  
para cada gen de la lista (Allen-Brain)

El paso que sigue a continuación es realizar un nuevo enriquecimiento de datos añadiendo a lo obtenido en el paso anterior información de niveles de expresión por estructura para cada gen en cada experimento. Pero solamente me interesa realizar este paso con todas las estructuras que en la ontología anatómica 'nodos hoja', y por tanto que la suma del conjunto de todas ellas me dé como resultado todas las regiones del cerebro, porque todas están al mismo nivel.

En este paso el objetivo es analizar cada uno de los documentos de experimentos obtenidos con el script 2, de modo que se trata de obtener para cada gen 2 ficheros: un fichero con la lista de estructuras cerebrales, y un fichero con la lista de valores de expresión respectivos de esas estructuras. Estos ficheros serán usados para un proceso posterior.

allen\_Gata6\_69289033 x structure\_allen\_Gata6\_69289033 x values\_allen\_Gata6\_69289033 x

```

1 <Response success='true' start_row='0' num_rows='1' total_rows='1'><section-data-sets>
2   <section-data-set>
3     <blue-channel nil="true"/>
4     <delegate>true</delegate>
5     <expression>true</expression>
6     <failed>false</failed>
7     <failed-facet>734881840</failed-facet>
8     <green-channel nil="true"/>
9     <id>69289033</id>
10    <name nil="true"/>
11    <plane-of-section-id>2</plane-of-section-id>
12    <qc-date>2009-05-02T22:45:51Z</qc-date>
13    <red-channel nil="true"/>
14    <reference-space-id>10</reference-space-id>
15    <rnaseq-design-id nil="true"/>
16    <section-thickness>25</section-thickness>
17    <specimen-id>68857425</specimen-id>
18    <sphinx-id>10835</sphinx-id>
19    <storage-directory>/external/aibssan/production32/prod327/image_series_69289033</st
20    <weight>5470</weight>
21    <structure-unionizes>
22      <structure-unionize>
23
24        <voxel-energy-mean type="float">0.0857515</voxel-energy-mean>
25        <structure>
26          <acronym>TMv</acronym>
27          <atlas-id type="integer">424</atlas-id>
28          <color-hex-triplet>FF4C3E</color-hex-triplet>
29          <depth type="integer">8</depth>
30          <failed type="boolean">false</failed>
31          <failed-facet type="integer">734881840</failed-facet>
32          <graph-id type="integer">1</graph-id>
33          <graph-order type="integer">682</graph-order>
34          <hemisphere-id type="integer">3</hemisphere-id>
35          <id type="integer">1</id>
36          <name>Tuberomammillary nucleus, ventral part</name>
37
38        <voxel-energy-mean type="float">0.00186198</voxel-energy-mean>
39        <structure>
40          <acronym>IG</acronym>
41          <atlas-id type="integer">143</atlas-id>
42          <color-hex-triplet>7ED04B</color-hex-triplet>
43          <depth type="integer">7</depth>
44          <failed type="boolean">false</failed>
45          <failed-facet type="integer">734881840</failed-facet>
46          <graph-id type="integer">1</graph-id>
47          <graph-order type="integer">424</graph-order>
48          <hemisphere-id type="integer">3</hemisphere-id>
49          <id type="integer">19</id>
50          <name>Induseum griseum</name>

```

Figura 13: En la imagen se puede ver el documento del Allen que se obtuvo para el gen Gata6, y he comprobado en las pestañas (Imagen siguiente (Figura 14)) que el valor de la estructuras se corresponden con sus valores de expresión génica (<voxel-energy-mean type="float") correctos, los resultados coinciden, con los documentos obtenidos de forma individual. En recuadro lila veo que las condiciones en las que me ha guardado el experimento son las válidas, y con el recuadro rojo he comprobado que el id del experimento también es correcto.

gen Gata6(transcripción)	
estructuras	valores
1 Tuberomammillary nucleus, ventral part	1 0.0857515
2 Primary somatosensory area, mouth, layer 6b	2 0.0164232
3 Principal sensory nucleus of the trigeminal	3 0.0231656
4 Primary somatosensory area, trunk, layer 6a	4 0.0378809
5 Interfascicular nucleus raphe	5 0.000275838
6 Parataenial nucleus	6 0.01193
7 Superior colliculus, motor related, intermediate white layer	7 0.0529845
8 Induseum griseum	8 0.00186198
9 Anterior amygdalar area	9 0.015338
10 Superior colliculus, motor related, deep gray layer	10 0.0101109

Figura 14: Ejemplo del resultado de valores de expresión obtenidos para cada estructura, en el gen Gata6, que comprobaré con la imagen del documento Allen correspondiente en la imagen anterior (Figura 13) del trabajo. Podemos apreciar para el gen Gata6 de la transcripción por ejemplo, que la estructura 1 cuyo nombre (<name>) es Tuberomammillary nucleus, ventral part, se corresponde con el valor de expresión (<voxel-energy-mean>) 0.0857515 y por ejemplo para la estructura 8 cuyo nombre es Induseum griseum, se corresponde con un valor de expresión de 0.00186198.

No obstante, no se trata de incluir todas las estructuras cerebrales sino sólo aquellas que representan nodos "hoja".

Para resolver este paso he desarrollado 2 scripts de Perl.

- Script 3: step3\_value\_final\_onto1.pl
- Script 4: step4\_get\_value\_final\_OK.pl

Por lo tanto, como resultado de este paso, obtengo para cada gen los ficheros con la lista de estructuras cerebrales “hoja” (por ejemplo structure\_allen\_Gata6\_69289033, o structure\_allen\_Hoxa7\_71891539) y con la lista de sus valores de expresión (por ejemplo values\_allen\_Gata6\_69289033 o values\_allen\_Hoxa7\_71891539).

#### D-IV Construcción de tabla (data.frame) para cada gen, con las estructuras y los valores de expresión

Lo que me interesa en este punto del trabajo, es construir una tabla conjunta que contenga información de todas las estructuras, con respecto a sus valores de expresión correspondiente de cada gen para cada una de ellas. Estos datos los he obtenido en el paso anterior, y son los ficheros que ahora utilizo.

El resultado obtenido es una tabla, representada en la Figura 15, con la lista de estructuras, y el valor de expresión que corresponde para cada estructura dentro de cada gen.

	estructuras	genes			
	row.names	Agtr2	Ankrd1	Arap1	Bcl10
1	Frontal.pole..layer.1	0.87995921	3.66217907	2.107237166	0.89804549
2	Frontal.pole..layer.2.3	1.17373318	2.96837158	1.421463115	0.90888345
3	Primary.motor.area..Layer.1	0.02936245	0.09162670	0.077677146	0.10386854
4	Primary.motor.area..Layer.2.3	0.11399183	0.14860507	0.249424267	0.24863314
5	Primary.motor.area..Layer.5	0.34894734	0.24709931	0.508534009	0.65636365
6	Primary.motor.area..Layer.6a	0.47490676	0.39191965	0.467220561	0.84627908
7	Primary.motor.area..Layer.6b	0.45321050	0.50583414	0.296551270	0.64340451
8	Secondary.motor.area..layer.1	0.32895929	0.45302005	0.282471783	0.13034978
9	Secondary.motor.area..layer.2.3	0.46737897	0.54753272	0.395626934	0.31294633
10	Secondary.motor.area..layer.5	0.62187721	0.67309070	0.561136503	0.63952108

Figura 15: data.frame obtenido con las estructuras y los valores de expresión para cada gen

Aquellos datos que no presentan información disponible, tienen valor NA (not available)[26]. Un gen con muchos NAs no es muy útil, ya que le falta mucha información. Por ello, en este trabajo, la tabla (Figura 15) se ha construido de tal manera que podamos trabajar con el máximo de información disponible, por lo que con el script de R, he eliminado aquellas columnas de genes que superan el cuartil que me interese en número de NAs, es decir, en una familia con un elevado número de genes, elimino el porcentaje de genes que no nos interesa (normalmente el 75%) porque su contenido en NAs es mayor.

Finalmente nos hemos quedado con aquellos genes que tienen una mayor información en el Allen Brain Atlas. Usando como mínimo una cantidad de 50 genes (quedándonos normalmente con el 25% de los genes), a no ser que la familia de genes tenga un número inferior en su totalidad.

Como he comentado, se ha desarrollado un script para ejecutar en R, que nos resolverá este paso.

- Script 5: step5TFM.R

#### D-V. Agrupación de estructuras atendiendo a los valores de expresión génica.

- A) Clustering de k-means
- B) Clustering jerárquico.
- C) Biclustering (Heatmap)

Continuando con las tablas obtenidas mediante el Script 5: step5TFM.R, se han realizado el siguiente paso, normalizando previamente los resultados obtenidos de la última tabla, cuyo objetivo es la

obtención de la agrupación no jerárquica y jerárquica de las estructuras asociadas a sus valores de expresión.

#### V-A Para el agrupamiento no jerárquico: El análisis Clustering de k-means:

En la Figura 16, se representa el resultado que se ha obtenido, el cual es una tabla que contiene en la primera columna las estructuras, en la segunda los grupo cluster asociados a cada estructura, y de la columna 3 a la última cada uno de los genes con sus valores de expresión asociados a cada estructura en los experimentos del Allen.

estructuras		cluster (k-means) k=12	genes
row.names	kc12.cluster	Agtr2	
1 Frontal.pole..layer.1	3	0.87995921	
2 Frontal.pole..layer.2.3	10	1.17373318	
3 Primary.motor.area..Layer.1	7	0.02936245	
4 Primary.motor.area..Layer.2.3	8	0.11399183	
5 Primary.motor.area..Layer.5	3	0.34894734	
6 Primary.motor.area..Layer.6a	3	0.47490676	
7 Primary.motor.area..Layer.6b	3	0.45321050	
8 Secondary.motor.area..layer.1	8	0.32895929	
9 Secondary.motor.area..layer.2.3	3	0.46737897	
10 Secondary.motor.area..layer.5	3	0.62187721	

Figura 16: Agrupamiento no jerárquico (k-means) de estructuras junto con valores de expresión de cada gen.

#### V-B Para el agrupamiento jerárquico de las estructuras, se ha realizado un Clustering jerárquico.

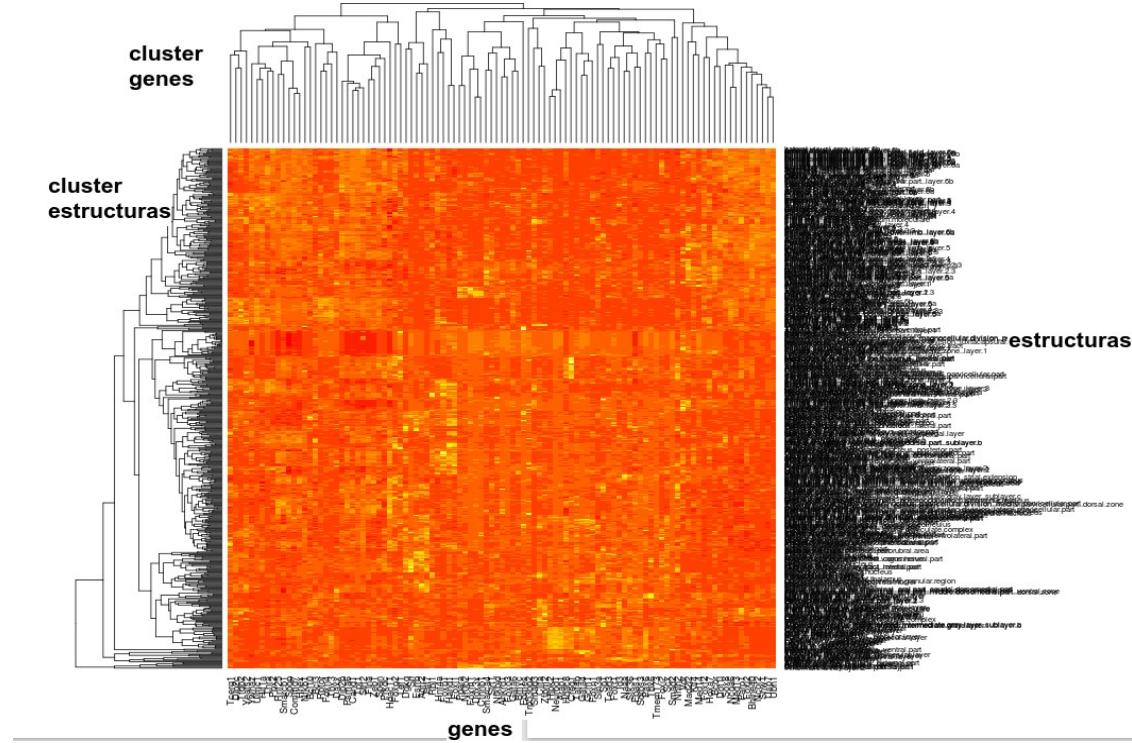
En relación con el Clustering jerárquico, se ha obtenido una tabla similar a la correspondiente al Clustering no jerárquico, pero se ha asignado una agrupación diferente. El producto del uso de esta técnica se puede ver en la Figura 17, que representa el Cluster jerárquico para k=12 por el método ward.D2:

estructuras		cluster jerárquico	genes			578 obser
row.names		numeroclusterward	Agtr2	Ankrd1	Arap1	
1 Frontal.pole..layer.1		1	0.87995921	3.66217907	2.107237166	
2 Frontal.pole..layer.2.3		1	1.17373318	2.96837158	1.421463115	
3 Primary.motor.area..Layer.1		2	0.02936245	0.09162670	0.077677146	
4 Primary.motor.area..Layer.2.3		2	0.11399183	0.14860507	0.249424267	
5 Primary.motor.area..Layer.5		2	0.34894734	0.24709931	0.508534009	
6 Primary.motor.area..Layer.6a		2	0.47490676	0.39191965	0.467220561	
7 Primary.motor.area..Layer.6b		2	0.45321050	0.50583414	0.296551270	
8 Secondary.motor.area..layer.1		2	0.32895929	0.45302005	0.282471783	
9 Secondary.motor.area..layer.2.3		2	0.46737897	0.54753272	0.395626934	
10 Secondary.motor.area..layer.5		2	0.62187721	0.67309070	0.561136503	

Figura 17: Agrupamiento jerárquico (ward) de estructuras junto con valores de expresión de cada gen.

#### V-C Bioclustering → Heatmap

En lo que se corresponde con los genes de la familia GO '*transcription*' por ejemplo, en la imagen, se puede apreciar que en las esquinas superiores se distingue un color amarillo/anaranjado generalizado, por lo que hay varios genes expresándose de forma mayor que en la parte superior central del heatmap, que es más oscura. Además que se expresan en sus estructuras específicas que se corresponden con las superiores, y por el contrario en las estructuras inferiores dichos genes presentan menor expresión.



El resultado obtenido es que la coloración de las estructuras siguiendo un Clustering no jerárquico (k-means) presenta una imagen del cerebro dividida en distintas regiones (grupos de estructuras reunidas en un mismo cluster). La regionalización por Clustering jerárquico es menos significativa para un cluster de k=12.

## E. DISCUSIÓN

En el presente trabajo, he planteado un método de clasificación de las estructuras anatómicas del cerebro, en relación con los patrones de expresión génica. El resultado obtenido me permite hacer una comparación cualitativa con la clasificación clásica de las regiones anatómicas del cerebro. Así mismo, se puede decir que los subconjuntos de genes específicos correspondientes a cada una de las familias, exhiben un claro patrón espacial en los cortes sagitales del cerebro de ratón, ya que la realización en particular del Clustering de k-means, me ha permitido reconocer una correcta clasificación en diferentes regiones fácilmente distinguibles. Esto queda reflejado en las imágenes dibujadas que con estos datos he podido obtener. Estos resultados sugieren que debe existir un control estricto o relación del patrón de expresión génica con respecto a la neuroanatomía.

Desde el punto de vista que se ha desarrollado en esta aproximación experimental, se hace hincapié en las diferencias transcripcionales (es decir, de expresión génica) entre distintas regiones del cerebro, y presumiblemente, el resultado específico de la combinación de patrones de genes específicos pertenecientes a cada una de las familias, lo cual me ha permitido llegar a la conclusión de que el patrón de expresión génica se desarrolla en consonancia con el origen embrionario de las distintas regiones del cerebro, al mismo tiempo que este órgano va desarrollándose en el embrión hasta que se hace adulto, es decir, que las regiones del cerebro se van desarrollando al unísono que los patrones de expresión génica.

### E-I. Validación estadística: Clustering k-means

Podemos observar la unión de distintas estructuras en una misma región, lo que se representa como diferentes regiones cerebrales, debido a que podemos ver diferentes estructuras que expresan patrones similares de expresión génica y que por ello están representados en el mismo clúster y región. Este mapa resultante es el que hemos comparado con la ontología anatómica clásica en el último apartado de esta Discusión.

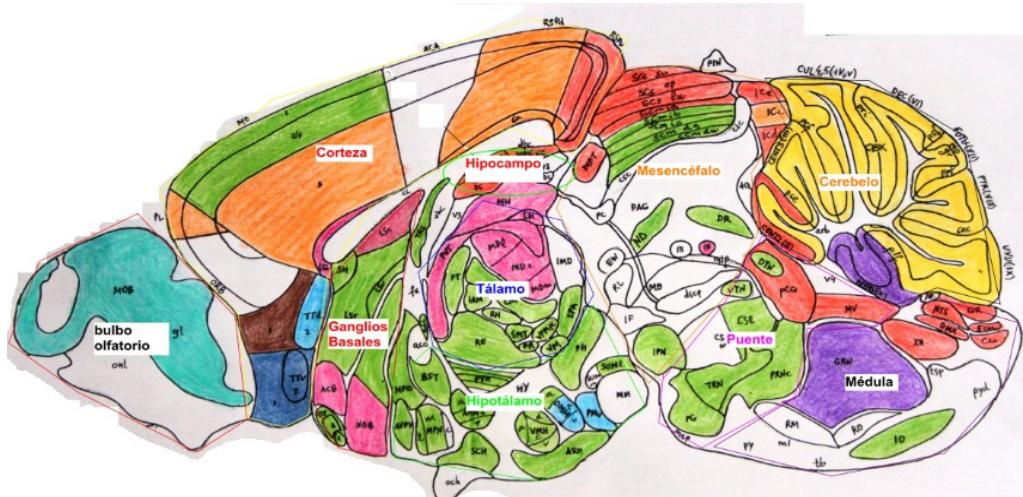


Figura 20: Se corresponde con un plano medio sagital del cerebro, donde se ha realizado un Clustering de k-means de los genes de transcripción, que además se encuentran las partes más importantes del cerebro señaladas, para poder seguir los comentarios durante toda la discusión de los resultados (Imagen a mayor resolución disponible en la carpeta de Anexos, con el nombre de 'regiones\_del\_cerebro').

### E-II. Validación estadística: Clustering jerárquico y Biclustering/Heatmap

#### A) Clustering jerárquico

El dendograma que se muestra a continuación, lo he dividido en 8 grupos (clusters en rojo, el corte lo hace a una altura ~ 90), en 12 grupos (clusters en azul, el corte lo hace a una altura ~ 70) y en 15

grupos (clusters en verde), el corte lo hace a una altura  $\sim 60$ ). El motivo de elegir esta división es para ver la evolución en la agrupación de estructuras teniendo más grupos, es decir, si ciertas estructuras formarían parte de otras regiones diferentes, si permanecerían en el mismo grupo, o si aparecerían nuevas regiones de estructuras agrupadas, teniendo en cuenta la imagen del dendograma.

El resultado final de cada 'nodo hoja' del dendograma se corresponde con cada 'estructura hoja' que queda en el resultado normalizado de las estructuras, y por tanto el conjunto de todas aquellas estructuras que contienen una mayor y más significativa expresión génica. El conjunto final de genes de transcripción se corresponde con 578 estructuras (del total de 930 estructuras 'hojas'), siendo éstas estructuras que presentan valores de expresión para todos y cada uno de los genes. El número de estos genes es 103, después de la selección del cuartil que presentan más información.

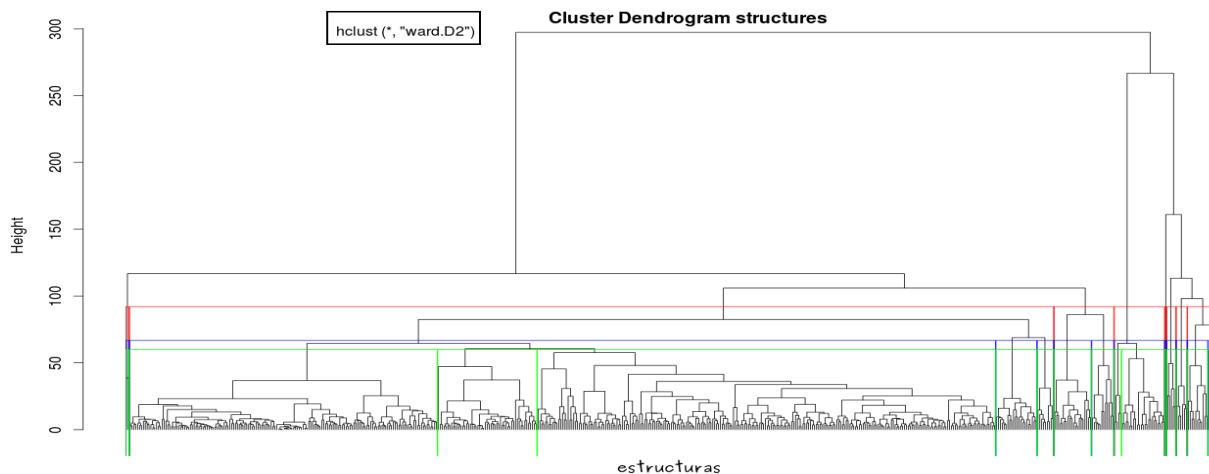


Figura 21: Dendrograma por Clustering jerárquico por el método “ward.D2”

En la Figura 21 se han omitido los nombres de las estructuras en el plot (Labels=FALSE), para facilitar la legibilidad de la figura, ya que debido a la alta cantidad de estructuras no se podría apreciar la estructura del dendograma.

Un ejemplo de la agrupación jerárquica obtenida para cada una de las estructuras con respecto a los valores de expresión génica, lo podemos observar en la Figura 22.

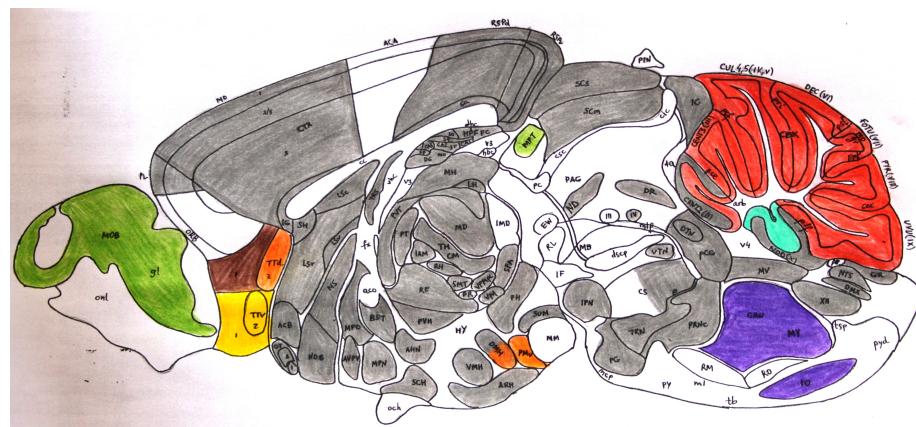


Figura 22: Clustering jerárquico de estructuras para los genes transcripción ( $k=12$ ) por el método Ward.D2

Se observa poca regionalización en este mapa (a diferencia del Clustering de k-means), la región gris se correspondería con el recuadro más grande en color azul del dendograma de la Figura 21, por lo que con la aplicación de este método se necesitaría un número mayor de subdivisiones para observar una mejor regionalización del cerebro. Si observamos una regionalización correcta del Cerebelo (rojo) y de la Médula oblongata (lila).

B) Heatmap:

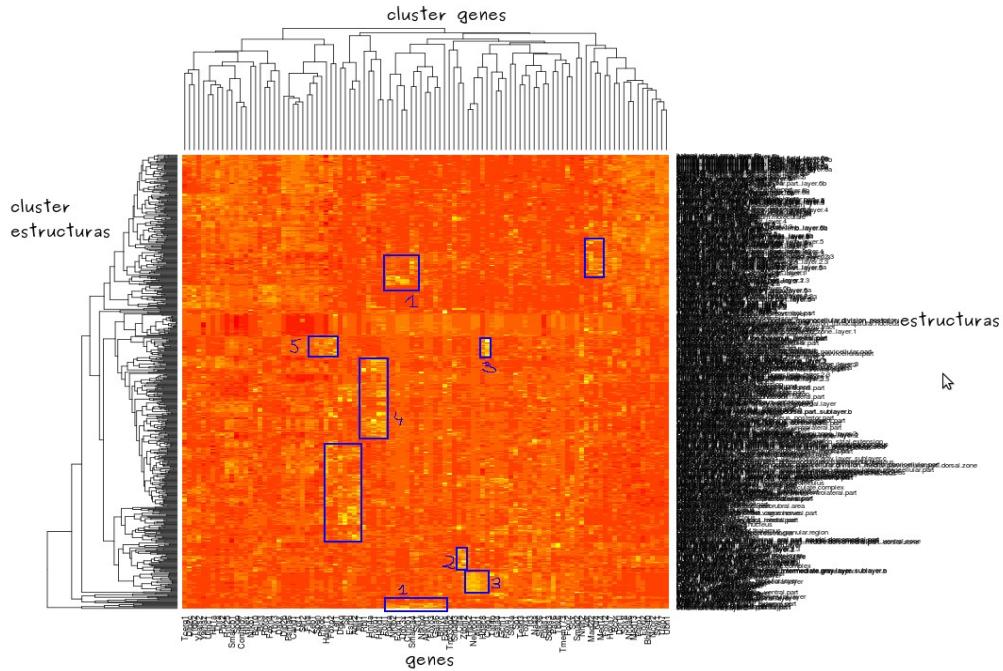


Figura 23: Heatmap de las estructuras y la lista de genes para los genes de la familia de transcripción

En cuadrados azules he señalado las zonas más amarillas según una aproximación visual, y por tanto de mayor expresión, aunque se podrían seleccionar algunas más. Según los grupos que he rodeado en cuadrados azules en el apartado resultados del Heatmap, he podido hacer una aproximación sobre algunos de los genes que se encuentran agrupados en ellos.

grupos heatmap (genes) --- aproximación	
grupo 1:	Ppard, Nfkbid, Foxo3, Gata6, Bmyc
grupo 2:	Ciita, Zfpm2
grupo 3:	Arap1, Hipk2, Neurod1
grupo 4:	Hnf4a, Foxo1, Hand1, Rbl1, Foxl1
grupo 5:	Flna, Zeb1

Figura 24: Aproximación de los grupos de genes que forman parte de cada uno de los cuadrados azules señalados en el heatmap.

Por ejemplo, los genes Foxl1 y Hand1, agrupados según la aproximación de mis resultados de la Figura 24 en el grupo4, están implicados en otro proceso biológico [27]. Una minería buscando asociaciones entre los distintos genes de cada cluster sería una línea de investigación a explorar.

Un paso posterior sería encontrar un método eficaz capaz de definir los grupos de genes y las estructuras en las que se engloban en el heatmap, que como será explicado más adelante, es una de las limitaciones que he tenido.

### E-III. Validación biológica: “Comparación con la ontología clásica”

Quiero hacer un estudio de validación biológica, comparando los resultados obtenidos con la Ontología clásica de regiones del cerebro. Los grupos de la ontología GO que he usado para realizar el experimento son: *cerebral cortex radially oriented cell migration, axon guidance, y transcription*.

A continuación muestro esquemas que se corresponden con respectivos resultados obtenidos al colorear con los grupos cluster producidos por la técnica de k-means comparados con la imagen de la coloración representada por la ontología clásica. Incluyo además diversos ejemplos de imágenes que dan una visión más detallada de los resultados obtenidos sobre el agrupamiento de estructuras cerebrales en base a dichos grupos de genes. Para cada uno de estos ejemplos, comento si reproducen la agrupación de estructuras reconocida en la ontología anatómica clásica; o si por el

contrario implican nuevas clasificaciones o agrupaciones de estructuras no descritas hasta ahora.

En la carpeta de Anexos les adjunto las imágenes originales que tienen una mayor resolución, y así se puede seguir mejor la interpretación de los resultados que a continuación se detallan.

#### **Clasificación según los genes “cerebral cortex”:**

(Anexos: 'cerebral\_cortex\_k8' , 'cerebral\_cortex\_k12')

Aquí he tenido en cuenta el 100% de los genes, debido a que hay muy pocos genes, pero ha salido un resultado bastante correcto que vemos en la Figura 25. Puedo ver que los genes más representativos son aquellos genes específicos de la Corteza, me ofrecen una subdivisión en capas de la Corteza, como era de esperar teniendo en cuenta el rol de estos genes en la regionalización cortical . Pero como son genes poco específicos del resto de zonas del cerebro, se ve un colorido bastante homogéneo fuera dela corteza; apareciendo por ejemplo un cluster (en color amarillo en k=8, en gris en k=12) que se extiende a lo largo de todo el cerebro.

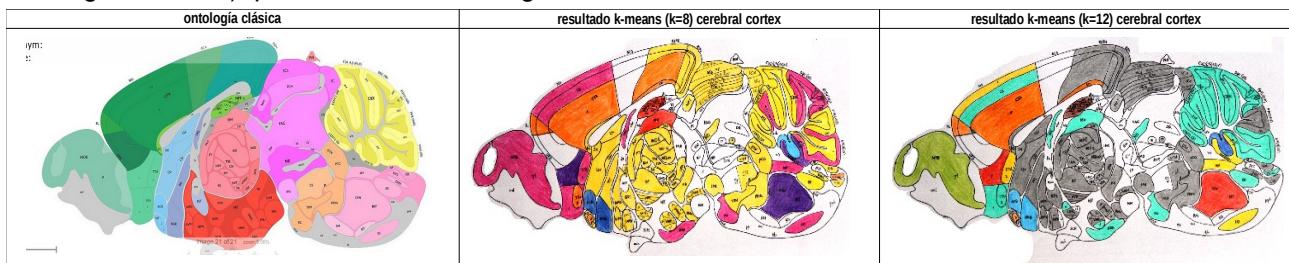


Figura 25: Representa los cortes sagitales del cerebro medio de Ontología clasica, y Clustering k-means ( $k=8$ ) y ( $k=12$ ) de los genes de *cerebral cortex*.

#### **Clasificación según los genes “axon guidance”:**

(Anexos: 'axon\_guidance\_k8' , 'axon\_guidance\_k12')

Teniendo en cuenta el 25% de los genes más representativos, se puede observar la subdivisión en las distintas regiones según se representa en la Figura 26

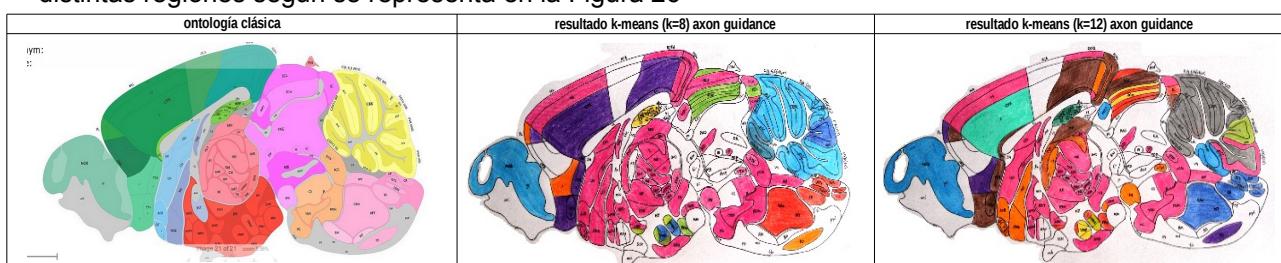


Figura 26: Representa los cortes sagitales del cerebro medio de Ontología clasica, y Clustering k-means ( $k=8$ ) y ( $k=12$ ) de los genes de *axon guidance*.

Lo que más llama la atención de la Figura 26 es la conexión en  $k=12$  de algunos Ganglios basales del cerebro (en azul en el esquema “Ontología clásica”) con el Tálamo (en un color naranja, de aspecto redondeado, en la Figura 27 con más detalle) como si formasen parte de una misma región (todas ellas coloreadas en color rosa en el esquema “resultado k-means ( $k=8$ ) *axon guidance*”). Este mismo resultado también se puede ver en las agrupaciones de estructuras obtenidas a partir de otros grupos ontológicos de genes (coloreadas en amarillo en el esquema “resultado k-means ( $k=8$ ) cerebral-cortex”; en verde en el esquema “resultado k-means ( $k=8$ ) y ( $k=12$ ) transcription”),



Figura 27

En comparación con la ontología clásica puedo apreciar que regiones como el el Diencéfalo y el Romboencéfalo aparecen claramente visibles con esta aproximación de modo similar a la ontología clásica. Se puede deducir que hasta cierto punto las estructuras vecinas y que son de la misma región tienden a quedarse dentro del mismo clúster de expresión génica. No obstante, la zona del Telencéfalo (concretamente la corteza cerebral) es la que más diferenciación interna posee con respecto a la ontología clásica, observándose dentro de ella diversas agrupaciones.

**Clasificación según los genes “transcription”:** (Anexos: 'transcription\_k8' , 'transcription\_k12')  
 Teniendo en cuenta el 25% de los genes más representativos (con mayor número de estructuras valoradas) aparecen los siguientes resultados de la Figura 28

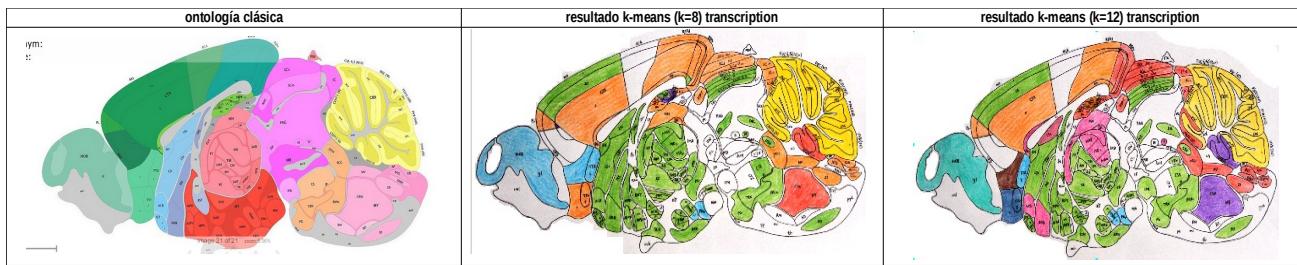


Figura 28: Representa los cortes sagitales del cerebro medio de Ontología clásica, y Clustering k-means ( $k=8$ ) y ( $k=12$ ) de los genes de Transcripción.

En la Figura 29 podemos observar como en la región del Hipocampo de  $k=12$ , habrían dos regiones distintas atendiendo al patrón de expresión génica representadas aquí en colores rojo y naranja.

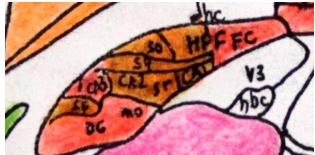
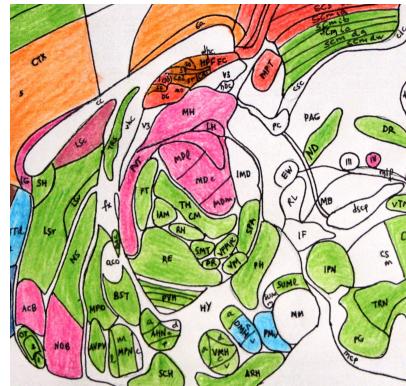


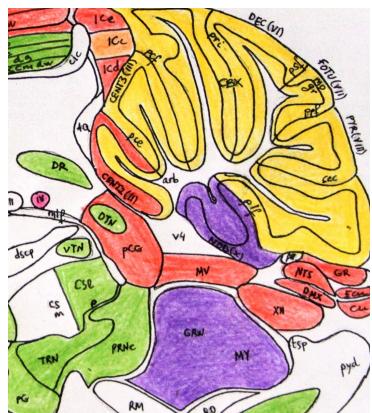
Figura 29 →

En la Figura 30 se ven también zonas representadas de modo homogéneo (color verde), lo cual me indica que partes de los Ganglios basales y del Tálamo, e Hipotálamo se quedan en el mismo clúster. En comparación con la ontología clásica en lo que se refiere a la zona del tálamo se diferencia en 2 regiones (coloreadas en verde y rosa) lo cual sería un ejemplo para buscar posibles implicaciones funcionales o de conectividad de esta regionalización interna del Tálamo.

Figura 30 →



← Figura 31



En la Figura 31: La zona del Cerebelo aparece organizada de forma diferente con respecto a la ontología clásica, con la mayor parte de sus estructuras en el mismo cluster (en color amarillo) mientras que otras de sus estructuras se incluyen en otros clusters (en colores violeta y rojo).

Se puede también observar una clara diferenciación entre la zona del puente y la zona de la médula. Además es destacable en esta zona que ciertas estructuras en  $k=12$  se asocian en un grupo independiente (en color rojo), como si formaran parte de otra subdivisión situada longitudinalmente por encima del Puente y la Médula, bajo el Cerebelo.

Existe por lo tanto cierta correspondencia de mis resultados con las regiones globales de la ontología clásica, en lo que se refiere a Romboencéfalo (exceptuando en particular las regiones excluidas del cluster principal del Cerebelo, y la mencionada columna sobre Puente y Médula).

Comparando los resultados de los anteriores grupos de genes, uno de los resultados más interesantes es que la mayor parte del Diencéfalo (Tálamo y estructuras vecinas) y los diversos componentes del Septum (Ganglios basales) se integran en un grupo común, como comenté anteriormente en los apartados sobre “axon guidance” y “transcription” demostrando una similitud genética entre estas regiones atendiendo a los patrones de expresión de ambas familias génicas y apuntando a la hipótesis un desarrollo embrionario y funcionalidades similares.

Otro de los resultados interesantes es la división de la zona del Hipocampo según los genes de “transcription”, y por último la regionalización en cuatro zonas según estos mismos genes de la parte del Romboencéfalo, donde además de las 3 zonas reconocidas (Cerebelo, Puente y Médula), aparece un nuevo cluster bajo el cerebelo como comentado anteriormente, lo que cual sería una clasificación novedosa que podría explorarse con otras aproximaciones experimentales.

Por lo tanto, para el Clustering no jerárquico (k-means), además de encontrar correlaciones con la ontología neuroanatómica clásica, mi metodología basada estrictamente en un criterio molecular o de expresión génica ha identificado nuevas agrupaciones de estructuras cerebrales, lo cual representaría una estrategia para iniciar aproximaciones experimentales que analicen el posible significado biológico de estas variaciones respecto a la ontología anatómica clásica.

Sin embargo, los resultados obtenidos para el Clustering jerárquico no han sido muy significativos, observando poca regionalización para el grupo de genes de transcripción  $k=12$ , que apreciamos en la Figura 19.b (apartado D-VI de Resultados), Figura 22 (apartado E-II.A de Discusión) y en el anexo 'Cluster\_jerarquico\_k12'. A pesar de esta falta de regionalización, estos resultados también se correlacionan con la ontología clásica, donde las estructuras con patrones de expresión similares se incluyen en el mismo cluster (como observamos en la parte del Cerebelo en rojo y la Médula en lila), por lo que con este método se necesita una mayor subdivisión de clusters ( $k$  mayor) para apreciar la regionalización del cerebro.

Entre las limitaciones asociadas a este trabajo, se han incluido la falta de información completa para cada gen, es decir, la ausencia de parte de las estructuras cerebrales en el fichero descargado del proyecto Allen. Esto ha dado lugar a numerosos valores nulos en las tablas obtenidas (ver apartado D-IV de Métodos) obligándonos a descartar hasta el 75% de los genes con menor información, como hemos comentado anteriormente. También dependemos de la técnica usada por el proyecto Allen para asignar los valores de expresión a cada estructura, basada en un proceso automático superponiendo "grids" de mapas anatómicos y fotografías de las secciones de ISH, lo cual tiene un margen de error en la identificación de los límites precisos de cada estructura cerebral y la consiguiente asignación de valores.

Además de la falta de evidencia para visualizar claramente las estructuras y los genes que pertenecen a cada grupo en el Heatmap, para poder verificar de forma precisa la asociación entre genes y estructuras cerebrales que he mencionado anteriormente en la Figura 24, para que más adelante se pueda hacer una comparación mejor entre esos grupos.

A pesar de ser fundamentalmente un análisis cualitativo, posibles líneas de investigación podrían ser abiertas en un futuro, tales como: (1) realizar un análisis cuantitativo para muchos más términos GO y conjuntos de genes, hasta incluir todo el genoma, (2) se podría estudiar la correspondencia de los grupos de genes agrupados en el Bioclustering, (3) generar un Clustering jerárquico de genes, para poder encontrar algún motivo experimental que asocie dichos grupos de genes y estructuras y (4) realización de estudios posteriores con diversas aproximaciones experimentales (fisiológica, bioquímicas, de conectividad neuronal, etc.) como continuación de los resultados de la regionalización obtenida.

## F. CONCLUSIONES

1. He implementado un método para extraer y procesar información de la expresión génica cuantitativa y diferencial en cada una de las distintas regiones o estructuras cerebrales, a partir del proyecto "Allen Brain Atlas".
2. He usado dicha información para agrupar según criterios estadísticos las regiones o estructuras cerebrales según la similitud de sus patrones de expresión génica, comparando los resultados obtenidos a partir de diversos grupos ontológicos de genes.
3. Las agrupaciones de estructuras obtenidas por Clustering k-means según esta metodología pueden ser correlacionadas con las agrupaciones descritas en la ontología anatómica clásica, aportando así nuestro trabajo una nueva perspectiva basada estrictamente en criterios moleculares. Adicionalmente, nuestro método ha considerado el clustering jerárquico como técnica alternativa para este fin.
4. El hecho de que la neuroanatomía cerebral se corresponda con patrones de expresión génica específicos de modo bastante significativo, me permite finalizar el trabajo con la reflexión de que esta similitud genética que da lugar a la regionalización del cerebro, puede que se encuentre en estrecha relación con el desarrollo embrionario y las funciones del cerebro, lo cual puede ser un ejemplo para comenzar nuevas líneas de investigación a partir de este punto, que resultaría de gran interés para los estudios del funcionamiento y patologías del cerebro.

## G. REFERENCIAS

- [1] Di Salle,F., Duvernoy,H., Rabischong,P. *Atlas of Morphology and Functional Anatomy of the Brain*, (2005)
- [2] Clark,D., Boutros,N., Mendez.M. *The Brain and Behavior - An Introduction to Behavioral Neuroanatomy* (Cambridge University Press, 2005).
- [3] Anatomy & Physiology. OpenStax Coll. 13, 516
- [4] Puelles,L., Martínez,S. *Ontogenia del Sistema nervioso. Master Neurociencia. Chapter 11*, (2012)
- [5] Watson,C., Kirkcaldie,M., Paxinos,G. *The Brain. An introduction to functional neuroanatomy*, (2010).
- [6] Striedter,G.F. *Principles of Brain Evolution* (Sinauer Associates, Sunderland, MA), (2005)
- [7] Puelles,L., F.J. Concept of neural genoarchitecture and its genomic fundament (Review). *Front Neuroanat*, (2012)
- [8] Matinez,S. Mecanismos generales del control molecular de la formación de las regiones del cerebro durante el desarrollo. *Bol. ECEMC Serie VI*, no 1, (2011).
- [9] Montoya Villegas,J.C., Peña,A. Análisis sistémico in silico de la expresión diferencial de genes localizados en la región crítica del síndrome de down (DSCR) en el cerebro humano. *Rev. Med* 15–26, (2012)
- [10] Ko,Y., S.A.A. Cell type-specific genes show striking and distinct. *PNAS* 110, 3095–3100, (2013).
- [11] Ji,S. Computational genetic neuroanatomy of the developing mouse brain: dimensionality reduction, visualization, and clustering. *BMC Bioinformatics* 14, 222, (2013).
- [12] Ding,S-L., Miller,J.A., Sunkin,S. Transcriptional landscape of the prenatal human brain. *Nature* 0, (2014)
- [13] Gómez López,A. Descripción de la ontología Universidad Juan Carlos, (2007) at <https://ai.wu.ac.at/~polleres/teaching/ri2007/alberto.pdf>
- [14] Ashburner,M. et al. Gene ontology: tool for the unification of biology. *The Gene Ontology*. *Nat. Genet.* 25, 25–29, (2000).
- [15] Khatri,P., S.D. *Ontological analysis of gene expression data: current tools, limitations, and open problems*. Bioinforma. Press, (2005).
- [16] Bobrow,M.N, Shaughnessy,K.J, Litt,G.J. Catalyzed reporter deposition, a novel method of signal amplification. *J Immunol Methods* (1991) 137:103–112
- [17] Chao,J., DeBiasio,R., Zhu,Z., Giuliano,K.A, Schmidt,B.F. Immunofluorescence signal amplification by the enzyme-catalyzed deposition of a fluorescent reporter substrate (CARD) (1996). *Cytometry* 23:48–53
- [18] Mejías Pacheco,M., Pombal,D., Molist,P. *Atlas de histología vegetal y animal* <at <http://webs.uvigo.es/mmegias/6-tecnicas/5-hibridacion.php>>
- [19] Baena Diez,J.M, Álvarez Pérez,B. Asociación entre la agrupación (Clustering) de factores de riesgo cardiovascular y el riesgo de enfermedad cardivasculares. *Rev Esp Salud Pública* (2002), 76:7-15
- [20] Corrales,M., Cordero,G., Jiménez,F. Práctica de Minería de datos at [http://rstudio-pubs-static.s3.amazonaws.com/15609\\_4ff4aaf346264aa4b7e0a5650eb8b61f.html](http://rstudio-pubs-static.s3.amazonaws.com/15609_4ff4aaf346264aa4b7e0a5650eb8b61f.html)
- [21] MacQueen,J.B. Some Methods for classification and Analysis of Multivariate Observations (1967). *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*.University of California Press. pp. 281–297.
- [22] Murtagh,F., Legendre,P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? (2013) *Journal of Classification* (in press)
- [23] Ward,J.H., Jr. Hierarchical Grouping to Optimize an Objective Function. *J. Am. Stat. Assoc.* 58, 236 –244 (1963).
- [24] Pontes,B., Rodríguez-Baena, Domingo,S., Díaz-Díaz,N. *Análisis de datos de expresión génica*. Universidad de Sevilla
- [25] Carmona,S. Técnicas para analizar grandes conjuntos de datos y encontrar patrones o comportamientos similares entre ellos.
- [26] Quick-R at <<http://www.statmethods.net/input/missingdata.html>>
- [27] Iascone,M., Ciccone,R. Identification of de novo mutations and rare variants in hypoplastic left heart syndrome. *Clin Genet.* 6:542-54, (2012)
- [28] Netter et al. *Atlas de Anatomía Human*. 5<sup>a</sup>edición, (2011)
- [29] Jagalur,M., C.P. Analyzing *in situ* gene expression in the mouse brain. *BMC Bioinformatics* 8(Suppl 10):S5, (2007).
- [30] Snell R.S. *Clinical neuroanatomy*. (Lippincott Williams & Wilkins, 2001).
- [31] Kerwin,J., Scott,M., Sharpe,J., Puelles,L. 3 dimensional modelling of early human brain development using optical projection tomography. *BMC Neurosci*, (2004).
- [32] Hofmann,H.A. Early developmental patterning sets the stage for brain evolution. *PNAS* 107, 9919–9920, (2010).
- [33] Lein,E.S, Zhao,X., Gage,F.H. Defining a molecular atlas of the hippocampus using ADN microarrays and high-throughput *in situ* hybridization. *J Neurosci.* 24:3879-89, (2004).
- [34] Kandel,E.R. *Principles of Neural Science*, (McGraw-Hill, 2000).

## H. ANEXOS

En la carpeta de Anexos tienen información adicional: (1) Scripts desarrollados (2) Ficheros generados para *Cerebral cortex* y scripts ejecutables (*cerebral cortex*, y *transcription*) (3) Dibujos realizados con mejor resolución.