Universidad del País Vasco

Trabajo Fin de Máster

Estudio de la conectividad estructural y genética del cerebro humano

David Romero Bascones

codirigido por

Dr. Unai Irusta Zarandona

Dr. Jesús María Cortés Díaz

23 de septiembre de 2019

Sólo aquellos que se arriesgan a ir demasiado lejos, llegan a saber hasta dónde pueden llegar.

T.S.Elliot

A la gente de Biocruces, por descubrirme el apasionante mundo de la neurociencia. A Unai, por su excepcional implicación y comprensión más allá del ámbito del proyecto. A mi familia, por no perder nunca la fe. A mi diseñadora y compañera de aventuras favorita, por todo lo que hemos compartido estos últimos años.

Resumen

El cerebro humano, pese a su homogénea apariencia externa, presenta un intrincado patrón de conexiones internas que interconectan las diferentes regiones cerebrales. Estas conexiones, conocidas como tractos neuronales, están formadas por miles de neuronas agrupadas de forma ordenada y son la base de la correcta intercomunicación en un cerebro sano.

Los recientes avances en técnicas de neuroimagen y análisis genético han dado pie a diversos estudios que apuntan a que la genética podría jugar un rol fundamental en la construcción y correcto funcionamiento de los tractos neuronales.

El presente Trabajo Fin de Máster busca avanzar en la comprensión sobre cómo la expresión genética determina o condiciona la interconexión entre las distintas regiones cerebrales. Para ello, se ha llevado a cabo en primer lugar un preprocesado de datos de expresión genética que han permitido su adecuación a los análisis planteados en éste y futuros trabajos.

Los análisis realizados parten de un enfoque basado en la ciencia de redes donde el cerebro es modelado como una red o grafo que es posteriormente analizado desde múltiples vertientes.

En un primer análisis estadístico se ha estudiado si regiones genéticamente similares tienen mayor probabilidad de estar conectadas. Posteriormente, se ha observado en qué medida criterios basados en genética o conectividad agrupan las regiones cerebrales de forma distinta. Finalmente, se ha planteado el uso de técnicas de procesado de señal sobre grafos como una nueva herramienta de análisis. Este enfoque permite obtener una representación frecuencial de los patrones de expresión genética, y ha sido introducido para mejorar la comprensión de dichos patrones.

Abstract

The human brain, despite its external homogeneous appearance, hides a complex intern pattern of connections that interconnect the different brain regions. Underneath the brain surface, thousands of neurons group together in an organized way to form the so-called neural tracts. These connections are the basis of a correct intercommunication in a healthy brain.

Recent progress in both neuroimaging and genetic analysis techniques has given rise to several studies that, all together, point to a fundamental role of genomics in the development and correct functioning of neural tracts.

The aim of this Master's Thesis is to improve the understanding of how genetic expression actually determines the interconnection between brain regions. In this sense, several steps of preprocessing were carried out to adapt genetic expression data for further analysis performed in this and future projects.

Following a network science approach, the brain was modelled as a network or graph that was later studied using different analysis types.

In a first statistical analysis, it was tested whether genetically similar regions are more likely to be connected or not. Later, the way brain regions group together based on different genetic and connectivity criteria was observed. Finally, a novel approach based on graph signal processing techniques was proposed. More specifically, a spectral representation of genetic data was used to better understand gene expression patterns along the brain.

Laburpena

Giza garunak, nahiz eta kanpotik itxura homogeneoa eduki, garun eskualde ezberdinak konektatzen dituen konexio patroi korapilatsua aurkezten du barrutik. Traktu neuronalak deritzen konexio hauek era antolatuan elkartutako milaka neuronaz osatuta daude, eta garun osasuntsu batetan barne komunikazioaren oinarria dira.

Neuroirudi eta analisi genetikoen arloetako azken aurrerapenak ikerketa lan ugari sustatu dituzte. Dirudienez, lortutako emaitzak traktu neuronalen garapen eta funtzionamendu egokietan genetikak garrantzi handia daukala antzematen dute.

Master Amaierako Lan honen helburu nagusia adierazpen genetiko eta garun eskualdeen interkomunikazioaren arteko erlazioaren ulermenean aurrera egitea da. Horretarako, lehenik eta behin, adierazpen genetikoko datu-baseen tratamendu eta egokitzapen prozesu bat jarraitu da. Prozesu honen emaitzak lan honetan egindako analisiak eta baita etorkizuneko beste lan batzuetan egin daitezkeenak posible egin ditu.

Egindako analisi guztiak sare-zientzian oinarritutako ikuspegitik garatu dira. Honen arabera, giza garuna sare edo grafo bat balitz bezala modelatu eta ikertu egin da. Lehen urrats batean, analisi estatistikoa erabili da genetikoki antzekoak diren garun eskualdeak konektatuta egoteko probabilitate handiagoa ote duten aztertzeko. Ondoren, garun eskualdeak genetika edo konektibitatea banaka kontuan hartuta nola elkartzen diren aztertu da. Azkenik, grafoen domeinurako egokitutako seinale tratamendu teknikak erabiliz, gene adierazpen balioen maiztasun irudikapenak kalkulatu dira. Balio hauek aztertuz, adierazpen patroiak hobeto ulertzea izan da helburua.

Índice

Lista de figuras

.

Lista de tablas

Lista de acrónimos

AAL	Automated Anatomical Labeling
ADN	Ácido Desoxirribonucleico
AHBA	Allen Human Brain Atlas
ARNm	Ácido Ribonucleico Mensajero
CGE	Correlated Gene Expression
DAD	Daño Axonal Difuso
DEP	Densidad Espectral de Potencia
IDE	Integrated Development Environment
IRM	Imagen por Resonancia Magnética
ITD	Imagen por Tensor de Difusión
FSL	FMRIB Software Library
GUI	Graphical User Interface
HCP	Human Connectome Project
MNI	Montreal National Institute
ROI	Region Of Interest
SC	Structural Connectivity
SRS	Scaled Robust Sigmoid
SPM	Statistical Parametric Mapping
t-SNE	t-Distributed Stochastic Neighbor Embedding

1 Introducción

El presente Trabajo Fin de Máster ha sido realizado bajo una estancia de cooperación educativa en el Instituto de Investigación Sanitaria **Biocruces Bizkaia**[1]. El centro, creado en 2008, consta de diferentes grupos de investigación centrados en diversas áreas de la salud. En concreto, el trabajo ha sido llevado a cabo dentro del grupo de investigación en **Neuroimagen Computacional**[2], centrado en el procesado de imágenes y señales cerebrales y formado por profesionales de múltiples campos.

La creación de grupos de esta naturaleza resulta de especial interés en neurociencia, una disciplina originalmente ligada a la medicina, que se ha convertido en un campo altamente interdisciplinar donde la aportación de ingenieros, físicos y matemáticos es cada vez mayor.

Dentro de este ámbito, tanto la ingeniería biomédica como las ciencias de la computación adquieren un rol principal en lo relacionado con la adquisición y procesado de datos sobre la actividad y estructura cerebral. El aumento en la capacidad de computo, así como el desarrollo de nuevas técnicas de procesado permiten abordar preguntas que hasta ahora permanecían sin respuesta.

Uno de los grandes problemas no resueltos es comprender el funcionamiento de las redes neuronales que interconectan el cerebro de forma interna. De hecho, el deterioro de las mismas es uno de los principales efectos de enfermedades neurodegenerativas como el Alzheimer [5]. Además, numerosos estudios evidencian la relación existente entre dichas redes neuronales y la genética [6]. Avances en este ámbito han permitido identificar genes relacionados con alteraciones en la conectividad cerebral en patologías como el autismo o la demencia [7, 8]. Pese a todo, la comprensión que tenemos acerca de cómo los genes determinan las conexiones físicas entre regiones cerebrales de un cerebro sano es limitada.

El presente estudio busca profundizar en la relación existente entre genética y conectividad estructural (Structural Connectivity, SC) planteando nuevas técnicas de análisis para la comparación de dos tipos de datos cerebrales: expresión genética y conectividad entre regiones. Para realizar dicho análisis, es de vital importancia el correcto procesado previo de los datos, lo cual representa un componente fundamental dentro del estudio.

[1] https://www.biocrucesbizkaia.org
[2] https://www.biocrucesbizkaia.org/bc5.08

2 Contexto

2.1 Un breve vistazo al cerebro

De entre todos los desafíos científicos a los que se ha enfrentado la humanidad a lo largo de su historia, uno de los más apasionantes y que más enigmas esconde es sin duda el cerebro, el órgano más sofisticado del ser humano. Una compleja estructura donde miles de millones de neuronas interactúan entre sí formando billones de sinapsis responsables de las capacidades cognitivas que nos definen como especie [9, 10].

Desde un punto estrictamente anatómico el cerebro es realmente la parte superior del **encéfalo**, que agrupa las estructuras craneales que forman parte del sistema nervioso central y está formado por las tres estructuras macroscópicas diferenciadas en la figura 2.1: **cerebro**, parte más voluminosa responsable de toda la actividad cognitiva y vital; **cerebelo**, área encargada de integrar las vías motoras y sensitivas; y **tronco cerebral**, que actúa como medio de conexión para conducir los impulsos motores y sensitivos desde o hacia la médula espinal [11].

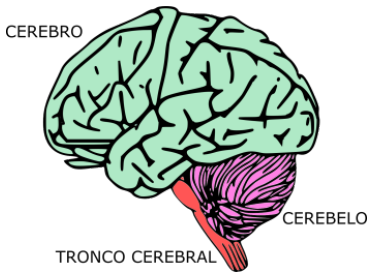


Figura 2.1: Regiones que forman el encéfalo

A nivel microscópico las células encargadas del procesado de información son las neuronas. Éstas están a su vez formadas por diferentes estructuras encargadas de procesar y transmitir las señales sinápticas. El **cuerpo celular** es el centro metabólico que alberga el núcleo que contiene los genes, y es donde se lleva a cabo la síntesis de proteínas. Las **dendritas**, ramificaciones procedentes del cuerpo celular, son las encargadas de la recepción de señales desde otras

neuronas. Los **axones** son estructuras tubulares recubiertas de una membrana llamada mielina y son los responsables de conducir las señales hacia otras neuronas conectadas mediante los **terminales presinápticos** [10], tal y como se muestra en la figura 2.2.

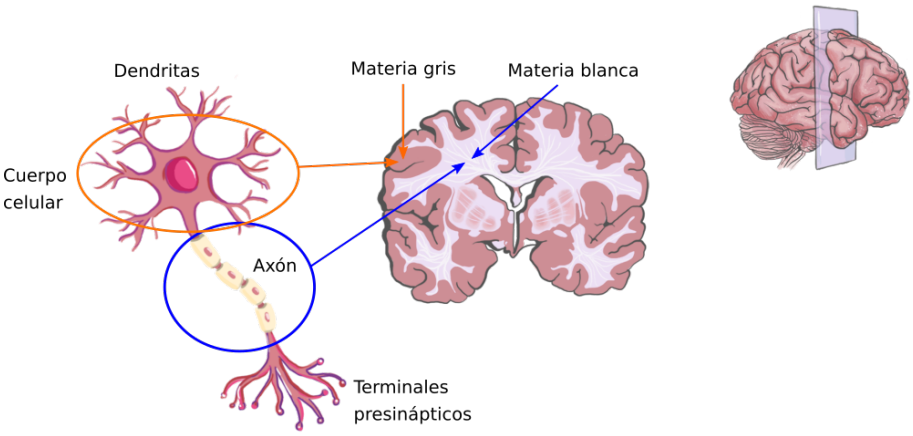


Figura 2.2: Neurona, materia gris y materia blanca

La disposición de las neuronas en el cerebro es tal que los cuerpos celulares están densamente conectados en la corteza cerebral mientras que los axones se agrupan de forma semiordenada en las zonas más interiores del cerebro formando los denominados **tractos neuronales** que actúan como "autopistas de la información" entre las diferentes regiones cerebrales [12].

La mielina que recubre los axones es de color blanco y hace que las zonas con mayor densidad de axones adquieran un color más blanquecino que aquellas con mayor número de cuerpos celulares. Estas diferencias dan lugar a dos regiones bien diferenciadas: **materia gris**, asociada con el procesamiento de la información; y **materia blanca**, relacionada con su transmisión.

Las conexiones que establecen los tractos neuronales que forman la sustancia blanca juegan un rol fundamental en el correcto funcionamiento del cerebro. De hecho, patologías como el daño axonal difuso (DAD) o el Alzheimer están directamente relacionadas con un deterioro en estas conexiones que, en definitiva, implica una desconexión parcial o total de ciertas regiones [13, 5, 14]. Es por ello que resulta de especial interés observar dichos tractos y entender su desarrollo y funcionamiento.

2.2 ¿Cómo vemos el cerebro?

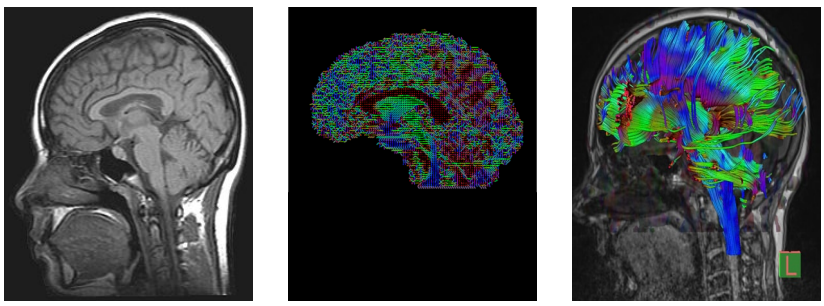
A lo largo de los años, los numerosos avances científicos han ido transformando la forma en la que vemos el cerebro, desde los primeros hallazgos por los egipcios hasta los modernos microscopios capaces de ver neuronas individualmente [15, 16]. A lo largo de ese proceso, el desarrollo de las técnicas de imagen

por resonancia magnética (IRM) [17] supuso una verdadera revolución. La IRM proporciona un método no invasivo que permite obtener imágenes en vivo tanto de la anatomía como de la actividad cerebral, y cambió radicalmente la forma en la que vemos el cerebro. Es por lo tanto una de las mejores herramientas existentes en la actualidad tanto a nivel clínico como científico.

Dentro de las numerosas técnicas existentes, la **Resonancia Magnética Estructural** permite la visualización de las distintas regiones anatómicas que forman el cerebro. Cada una de dichas regiones está compuesta por distintos tejidos, cada uno a su vez con unas propiedades magnéticas diferentes. De este modo, tras aplicar un campo magnético el nivel de señal medido en cada una de las regiones es diferente, pudiendo así generar una imagen en 3D como una matriz de vóxeles (píxeles en 3D) con valores de intensidad diferentes en cada región [18]. La figura 2.3a muestra un corte sagital de dicha matriz, es decir, perpendicularmente al plano horizontal y a la vista de frente. Este tipo de imágenes resultan de especial interés para realizar parcelaciones, y poder dividir el cerebro en diferentes regiones de interés (ROI, Region Of Interest).

Por otro lado, la técnica de **Imagen por Tensor de Difusión** (ITD) permite calcular la orientación en la que se encuentran las fibras que forman la materia blanca. Esto es posible por la tendencia que tienen las moléculas de agua a difundir a lo largo de los axones, generando así una respuesta diferente a campos magnéticos orientados en diferentes direcciones. De esta forma, se obtiene la dirección principal de difusión por cada vóxel, la cual se puede representar codificando las componentes [x y z] a valores RGB tal y como muestra la figura 2.3b. A partir de este tipo de imagen es posible reconstruir los tractos neuronales mediante algoritmos de **tractografía** que trazan las direcciones de difusión a través de los vóxeles [19].

Los resultados de la tractografía (figura 2.3c), permiten conocer qué zonas conecta cada uno de los tractos neuronales. Además, es posible cuantificar estas conexiones mediante métricas de conectividad pudiendo así medir cómo de conectadas están dos regiones desde un punto de vista estructural. Estas medidas permiten en muchas ocasiones diagnosticar patologías relacionadas con un deterioro de la conectividad [14].



(a) Estructural **(b)** Difusión **(c)** Tractografía

Figura 2.3: Diferentes imágenes generadas a partir de IRM. Obtenidas de [1] y [2]

2.3 La importancia de la genética

El ácido desoxirribonucleico (ADN) es el componente fundamental que codifica las instrucciones para el desarrollo y funcionamiento de los organismos vivos. Los segmentos de ADN forman unidades de almacenamiento de información denominadas genes.

Nuestro genoma al completo, formado por más de 20 000 genes, se encuentra almacenado en todas las células que componen nuestro cuerpo. Sin embargo, sólo un porcentaje de los genes se "activan" en cada región y proceso de nuestro organismo. Este mecanismo de activación es conocido como expresión genética. Se trata de un proceso complejo que puede ser dividido en dos fases principales (figura 2.4): **transcripción**, donde los segmentos de ADN son leídos y copiados en ácido ribonucleico mensajero (ARNm); y **traducción**, donde dichas copias son procesadas para la síntesis de proteínas, que son las que realmente desencadenan y regulan los procesos fisiológicos [6].

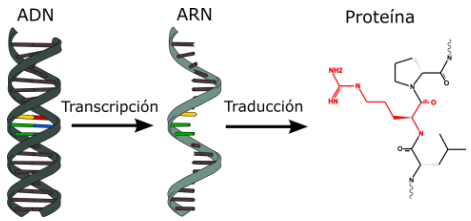


Figura 2.4: Proceso de expresión genética. Obtenida y editada de [3]

Este proceso de expresión genética también regula la fisiología a nivel cerebral por lo que es uno de las herramientas más prometedoras para desentrañar el complejo funcionamiento del cerebro. En lo que respecta a la conectividad estructural, resulta de especial interés conocer de qué manera las conexiones formadas por los tractos neuronales tienen su base en un componente genético.

Debido a la escasez de muestras, hasta hace relativamente poco los estudios se limitaban al estudio de zonas reducidas del cerebro y pequeñas variaciones en el ADN. No obstante, el desarrollo de técnicas de procesado masivo **microarray** o **RNA-Seq** capaces de medir el nivel de expresión del genoma completo ha permitido la creación de bases de datos que abarcan un gran porcentaje del cerebro [20], y que permiten estudiar la conectividad cerebral a mayor escala.

2.4 Resolviendo el puzle

Llegados a este punto tenemos por tanto las herramientas necesarias para conocer las distintas regiones cerebrales y determinar tanto su conectividad como su expresión genética.

El uso combinado de todas ellas permite plantear análisis más ambiciosos

que relacionen el genoma al completo con la conectividad estructural de todo el cerebro. Se busca, por lo tanto, no sólo identificar genes específicos relacionados con ciertas conexiones, sino patrones que expliquen la relación de forma global.

Para ello, es necesario en primer lugar obtener y preprocesar conjuntos de datos (datasets) tanto de expresión genética como de conectividad. A partir de aquí, se busca realizar análisis innovadores que puedan aportar nuevo conocimiento a la literatura existente.

3 Objetivos

3.1 Objetivo principal

El objetivo principal de este trabajo es estudiar la relación entre expresión genética y conectividad estructural del cerebro humano. Se busca analizar dicha relación desde múltiples vertientes. Es posible subdividir todo este proceso en una serie de subobjetivos más específicos.

3.2 Subobjetivos

3.2.1 Preprocesado base de datos genética

Seleccionar la base de datos a emplear y seguir una serie de pasos (filtrado, normalización, etc.) que permitan obtener finalmente una matriz que refleje la expresión genética de cada gen en cada una de las muestras de la base de datos.

3.2.2 Cálculo de conectividad genética y estructural

Seleccionar un atlas que parcele el cerebro en un número reducido de regiones. Para la conectividad genética, asignar las muestras a las diferentes regiones, promediar y obtener una matriz de correlación entre regiones. Para la conectividad estructural obtener métricas de conectividad entre las regiones del atlas.

3.2.3 Análisis

En primer lugar, realizar una primera exploración de los datos generados mediante técnicas de reducción de la dimensionalidad que permitan una visualización gráfica.

En segundo lugar, partiendo de las matrices de conectividad estructural y

genética, hacer uso de técnicas basadas en el análisis de comunidades para relacionar la conectividad estructural y la genética a nivel modular.

Finalmente, plantear un enfoque diferente basado técnicas de procesado de señal sobre grafos para tratar de entender como la expresión genética de cada gen se adapta a grafos basados en la conectividad estructural.

4 Beneficios

El presente estudio, de naturaleza investigadora, presenta beneficios directos en el ámbito científico, así como posibles beneficios futuros desde un punto de vista sanitario y económico.

4.1 Científicos

El trabajo realizado contribuye al avance en la comprensión de la relación existente entre la genética y la conectividad estructural del cerebro desde dos vertientes.

En primer lugar, los resultados del preprocesado de los datos de expresión genética son puestos a disposición de otros investigadores de modo que dichos datos puedan ser usados en futuros estudios.

En segundo lugar, se llevan a cabo análisis hasta ahora no realizados para correlacionar genética y conectividad como son el análisis basado en comunidades o el de procesado de señal sobre grafos. Este último, plantea por primera vez un enfoque de este tipo para el procesado de datos de expresión genética.

4.2 Sanitarios

Al igual que en otros campos de la ciencia, los resultados derivados de la investigación básica no suelen suponer avances inmediatos a nivel clínico. Sin embargo, forman la base de conocimiento sin la cual sería imposible poder desarrollar nuevas técnicas en un futuro. De esta forma, es posible que los resultados derivados de éste y otros estudios ayuden en un futuro a comprender mejor patologías con un componente genético que estén relacionadas con la conectividad estructural del cerebro. Esto podría dar lugar al desarrollo de nuevos tratamientos o a la identificación de biomarcadores genéticos para un diagnóstico precoz de este tipo de enfermedades.

4.3 Económicos

Desde un punto de vista económico el estudio no presenta beneficios monetarios directos o inmediatos. Aun así, existe la posibilidad de que, a largo plazo, los posibles avances en el ámbito clínico derivados de esta línea de investigación redujeran los costes sanitarios de tratamiento y diagnóstico. Es posible incluso que estos avances permitan desarrollar nuevas técnicas de diagnóstico, que debidamente explotadas supongan el desarrollo de nuevos productos y/o soluciones que puedan explotarse económicamente.

5 Estado del Arte

En el presente capítulo se describen en primer lugar los fundamentos matemáticos que permiten modelar y estudiar el cerebro humano como si fuera una red. Posteriormente, se presenta el estado actual de conocimiento sobre la relación entre genética y conectividad estructural cerebral para finalmente recopilar una serie de herramientas software empleadas en el presente proyecto.

5.1 El cerebro como una red

5.1.1 Ciencia de redes

¿ Qué es una red ?

Se denomina sistemas complejos a todos aquellos sistemas con un elevado número de componentes que interactuan entre sí de forma no trivial y que por tanto, resultan especialmente difíciles de modelar. La **ciencia de redes** es el campo científico que estudia aquellos sistemas complejos que pueden ser representados como redes o grafos. El campo tiene su base en el área matemática conocida como **teoría de grafos**, cuyo origen se remonta a 1736 y al célebre matemático Leonard Euler [21].

La representación mediante grafos modela los sistemas en base a nodos que se conectan entre sí mediante enlaces. Los nodos representan los diferentes componentes del sistema y los enlaces definen la relación entre los mismos. Usando este enfoque, es posible modelar sistemas como redes de telecomunicación, redes de carreteras e incluso redes de contactos en redes sociales. Como ejemplo, la figura 5.1 muestra un mapa parcial de lo que era internet en 2005, los nodos representan direcciones IP y los enlaces el retardo entre cada par de direcciones.

El potencial de un enfoque basado en grafos reside en la capacidad de analizar datos y sistemas con topología compleja. Además, el significado de los nodos y enlaces puede ser real e intuitivo como en la red de carreteras o puede también representar conceptos más abstractos como el número de amigos en común en Facebook. De esta forma, es posible utilizar técnicas de este tipo para encontrar patrones y relaciones en sistemas no analizables de otra forma.

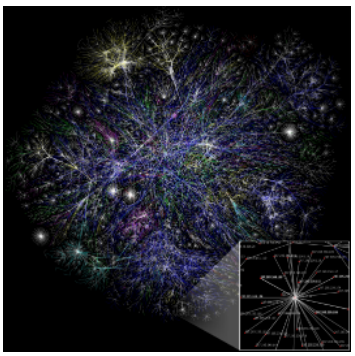
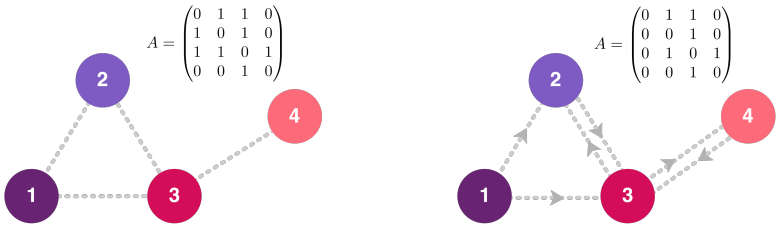


Figura 5.1: Mapa parcial de Internet en 2005, obtenido de [4]

Los grafos se caracterizan mediante su **matriz de adyacencia**, una matriz cuadrada con una fila y columna por cada nodo. En el caso más sencillo, cada elemento de la matriz toma valores de 1 o 0 para indicar la existencia o inexistencia de enlace entre los nodos correspondientes a cada par fila-columna.

La figura 5.2a muestra un grafo sencillo junto a su matriz de adyacencia. Se trata de una **red no dirigida** donde los enlaces son independientes del sentido de conexión entre dos nodos y por tanto, su matriz de adyacencia es simétrica. Por otra parte, existen también **redes dirigidas** donde la conectividad entre nodos sí tiene un sentido pudiendo estar los nodos conectados bidireccionalmente o sólo en una dirección. La figura 5.2b muestra un ejemplo de este tipo donde las conexiones del nodo 1 son únicamente salientes mientras que los enlaces restantes interconectan los demás nodos en ambas direcciones.



(a) Red no dirigida binaria

(b) Red dirigida binaria

Figura 5.2: Ejemplos de red binaria dirigida y no dirigida

Las redes mostradas hasta ahora han sido binarias, sin embargo, en numerosos casos es de interés modelar también la intensidad de los enlaces. La matriz de adyacencia deja de ser binaria y toma valores continuos proporcionales a la intensidad de la conexión del enlace. Un ejemplo de este tipo de redes se muestra en la figura 5.3, donde el enlace entre los nodos 1 y 2 es de mayor intensidad que los demás.

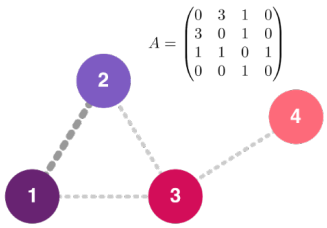


Figura 5.3: Ejemplo de red ponderada no dirigida

Métricas sobre redes

Tras modelar un sistema como una red, es posible analizar ciertas propiedades del mismo estudiando la matriz de adyacencia. En numerosas ocasiones es interesante conocer cómo de interconectados están cada uno de los nodos. Este valor se conoce como **grado**(k_i), y para un nodo i cualquiera se calcula de forma sencilla como la suma de sus valores de adyacencia hacia todos los N nodos existentes:

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{5.1}$$

El grado mide el número de enlaces de un determinado nodo. Cuando se trabaja con matrices de adyacencia no binarias, surge el concepto de **fuerza**, que es el valor resultante de aplicar la misma fórmula para el cálculo del grado con valores no binarios. En este caso no se mide directamente el número de enlaces si no la intensidad total con la que está conectado cierto nodo.

El objetivo final de estas métricas es identificar nodos que centralizan gran parte de las conexiones y que por tanto resultan de especial interés para comprender el comportamiento global de la red.

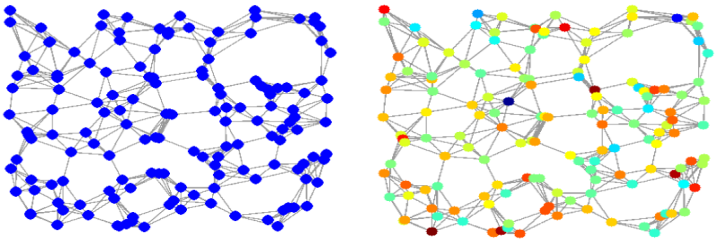
En numerosas ocasiones, las redes que modelan ciertos sistemas presentan una arquitectura muy poco mallada en la que es posible distinguir diferentes grupos de nodos fuertemente conectados internamente pero poco interconectados entre sí. Estos grupos se denominan **modulos o comunidades** y existen numerosas estrategias para su detección [22, 23]. Para poder obtener la partición óptima en comunidades de una red es necesario definir una métrica que mida cómo de modular es la red bajo una determinada asignación de nodos a comunidades. Surge el concepto de **índice de modularidad** introducido en [24] y definido como:

$$Q = \frac{1}{2N} \sum_{\substack{ij \\ mismo \\ módulo}} \left[A_{ij} - \frac{k_i k_j}{2N} \right] \tag{5.2}$$

donde Q toma valores entre -1 y 1 y compara la cantidad de enlaces entre nodos de una misma comunidad respecto a los existentes entre las diferentes comunidades. La ecuación 5.2 suma, por cada enlace dentro de una comunidad, la diferencia entre el valor de dicho enlace (A_{ij}) y el valor esperado si los enlaces se distribuyeran de forma uniforme ($\frac{k_i k_j}{2N}$), donde N es el número total de enlaces y k_i y k_j los degrees de los nodos i y j que forman un enlace. Los métodos de detección de comunidades buscan por lo tanto encontrar la partición de la red que maximice el valor de Q.

Procesado de señal sobre grafos

En numerosos casos es posible modelar sistemas como una red de nodos conectados entre sí pero que toman valores a lo largo del tiempo. El ejemplo más claro de este tipo de sitemas es una red de sensores como la de la figura 5.4a. Cada uno de los nodos representa un sensor. Los nodos se encuentra conectados con sus vecinos más cercanos siendo los pesos de los enlaces inversamente proporcionales a la distancia entre sensores. En este caso, las medidas tomadas por los sensores en un instante de tiempo pueden ser consideradas como una señal que toma valores en los nodos del grafo (ver figura 5.4b). Tenemos por tanto un grafo y una o varias señales que toman valores sobre ese grafo.



(a) Red de sensores **(b)** Medidas de los sensores

Figura 5.4: Grafo y señal de una red de sensores

En este tipo de casos es de gran interés tratar de estudiar las propiedades de dichas señales. Para ello es necesario extender las habituales técnicas de procesado de señal, como la transformada de Fourier o el filtrado, a señales en dominios complejos definidos por grafos. Surge de este modo el campo del procesado de señal sobre grafos. Se trata de un campo relativamente reciente, cuyas bases aún están siendo sentadas y en el que todavía existen numerosas preguntas abiertas, pero que está comenzado a ser usado por la comunidad científica como un nuevo enfoque para analizar problemas relacionados con grafos [25]. El potencial de este enfoque radica en su capacidad para tener en cuenta la estructura del grafo subyacente a la hora de analizar las señales.

Uno de los conceptos más importantes del procesado de señal sobre grafos es la obtención de una representación espectral de las señales. En el estudio de señales convencionales, el análisis tiempo-frecuencia es una herramienta conocida y empleada con éxito desde hace mucho tiempo. Sin embargo, la definición

del concepto de frecuencia en grafos no resulta tan sencillo. ¿Cómo descomponer en bajas y altas frecuencias una señal que varía a lo largo de un grafo?

Tal y como se ha explicado previamente, los grafos quedan definidos por su matriz de adyacencia ($\mathbf{A}$). Esta matriz está formada por los pesos de cada enlace entre dos nodos i y j cualquiera (w_{ij}):

$$(\mathbf{A})_{ij} = \begin{cases} 0 & i = j \\ w_{ij} & i \neq j \end{cases} \qquad i, j = 1 \dots N, \tag{5.3}$$

donde N es el número de nodos. A partir de ella es posible obtener una matriz diagonal $\mathbf{D}$ conocida como matriz de grado del grafo y definida como:

$$(\mathbf{D})_{ij} = \begin{cases} \sum_{k=1}^{N} (\mathbf{A})_{ik} & i = j \\ 0 & i \neq j \end{cases} \qquad i, j = 1 \dots N \tag{5.4}$$

Con estas dos matrices es posible obtener el **Laplaciano** como:

$$\mathbf{L} = \mathbf{D} - \mathbf{A} \tag{5.5}$$

Esta matriz es la base para la definición de la transformada de Fourier sobre grafos. La descomposición de dicha matriz en autovectores ($\boldsymbol{\chi}_0 \dots \boldsymbol{\chi}_{N-1}$) y autovalores ($\lambda_0 \dots \lambda_{N-1}$) permite describir un grafo en base a una serie de modos ortogonales. Los autovalores son análogos a las frecuencias definidas en la transformada de Fourier convencional mientras que los autovectores actúan como las exponenciales complejas y son modos o señales base de diferente frecuencia.

Para ilustrar estas ideas la figura 5.5 muestra cuatro modos distintos resultado de la descomposición en autovectores del grafo del ejemplo anterior. El primero de ellos corresponde a la componente continua. Los dos siguientes son modos de baja frecuencia que presentan patrones de variación lentos a lo largo del grafo. Finalmente se muestra también un modo ligeramente superior cuyos valores varían de forma más rápida. En esta línea, los modos asociados con autovalores mayores representan patrones de variación más rápida.

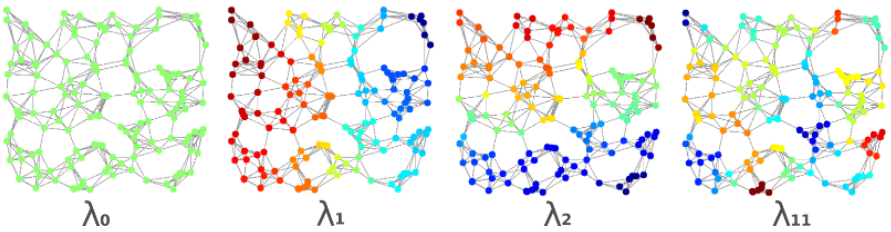
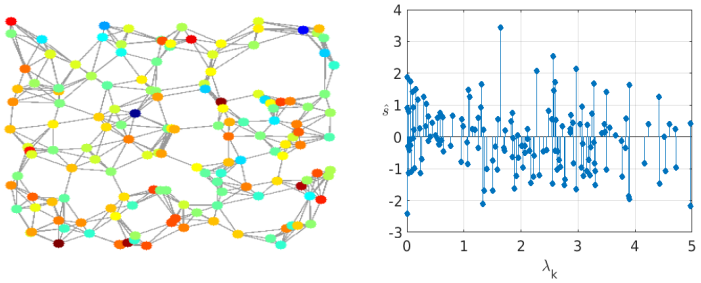


Figura 5.5: Diversos modos de un grafo

A partir de esto es posible obtener una representación espectral de una señal $\boldsymbol{s}$ mediante su descomposición en base a estos modos. Esta operación es realmente la transformada de Fourier aplicada a grafos definida como:

$$\hat{s}(k) = \sum_{i=1}^{N} \boldsymbol{\chi}_k^*(i) \cdot \boldsymbol{s}(i) \tag{5.6}$$

La figura 5.6 muestra una señal junto a su transformada de Fourier. Esta representación espectral permite comprender mejor su patrón y realizar operaciones como filtrado, que en el dominio espacial del grafo serían complicadas de realizar.



(a) Representación espacial **(b)** Representación espectral

Figura 5.6: Señal en grafo

Las posibilidades que ofrece esta metodología ha motivado su aplicación en diversos estudios en el ámbito de la neuroimagen [26, 27]. Sin embargo, los estudios se han centrado en el ámbito de la neuroimagen funcional y no se ha planteado aún su uso en el ámbito de la genómica.

5.1.2 Los nodos del cerebro

Tras plantear la ciencia de redes como herramienta para estudiar el cerebro, resulta lógico plantearse cómo definir los nodos que forman dicha red. Cuando el estudio se realiza a nivel macroscópico los nodos se asocian a regiones cerebrales. El proceso de división del cerebro en regiones, conocido como **parcelación**, no es trivial y puede hacerse en base a criterios de cercanía, arquitectura cerebral o funcionalidad. De este modo, es posible delimitar áreas cerebrales relacionadas con diferentes funciones. Además, la resolución de dichas regiones influye directamente en los posteriores análisis.

El problema radica en que cada cerebro individual tiene una morfología distinta, y por tanto, las regiones resultado de la parcelación no son iguales entre sujetos por lo que no es posible utilizar un único sujeto como referencia. Como solución a este problema surge el concepto de **Atlas**, una imagen del cerebro promedio donde las regiones son definidas en base a cierto esquema de

parcelación. Esta imagen, actúa como plantilla y sirve de referencia para situar dichas regiones en un sujeto individual mediante un proceso de registro.

A lo largo de los años, el avance en la comprensión del cerebro y la mejora de las técnicas de neuroimagen han permitido obtener Atlas con mayor precisión y resolución. Así, de la primera definición de 48 regiones conocidas como áreas de Brodmann (figura 5.7) se ha pasado a Atlas como el Craddock [28], con más de 1000 regiones.

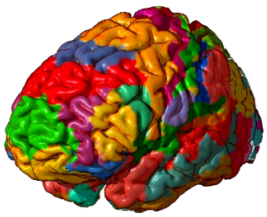


Figura 5.7: Áreas de Brodmann

Para la construcción de los Atlas más modernos se hace uso de imágenes MRI de múltiples sujetos y se obtiene la parcelación común. De este modo el Atlas representa la división promedio de ciertas regiones. La gran mayoría de los Atlas recientes hacen uso de la base de datos de imágenes IRM conocida como Human Connectome Project (HCP) [29], un proyecto internacional que busca mapear con detalle las conexiones cerebrales y que ofrece cientos de imágenes IRM de libre acceso.

Cada imagen IRM individual tiene el centro de coordenadas, orientación y tamaño propios del cerebro del sujeto por lo que se dice que está tomada en **espacio sujeto**. Sin embargo, para combinar imágenes de diferentes sujetos para construir un Atlas común es necesario corregistrar todas las imágenes en un mismo espacio de modo que tengan la misma orientación, centro de coordenadas y tamaño. Este espacio se conoce como **espacio estándar o MNI** que debe su nombre al centro donde fue definido, el Montreal Neurological Institute (MNI). El origen de coordenadas se define en base a puntos anatómicos bien diferenciados y la orientación se define por convención.

El objetivo final es definir la imagen MRI estructural de un cerebro promedio más representativa a lo largo de la población. Para ello, se han de corregistrar las imágenes de múltiples sujetos y promediar el resultado. A medida que se ha ido aumentando el número de sujetos se han creado versiones más refinadas del estándar. La plantilla estándar ICBM 2009c [30] (ver figura 5.8) es la más actual, y se construyó a partir de plantillas anteriores mediante registro no lineal entre sujetos.

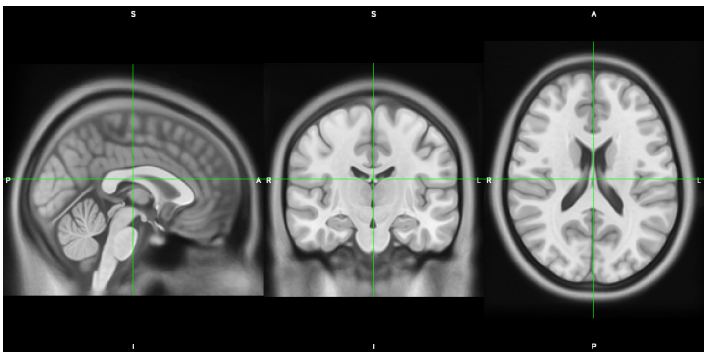


Figura 5.8: Template ICBM 2009c no lineal asimétrico

5.1.3 Los enlaces del cerebro

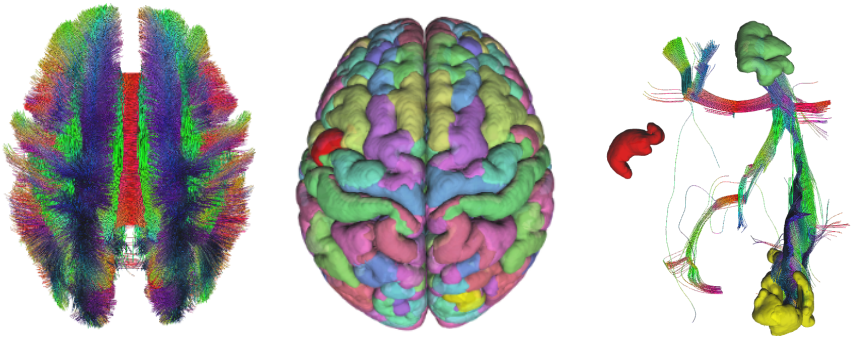
Una vez definidos los nodos de la red como regiones cerebrales es necesario definir los enlaces que interconectan dichos nodos. Los enlaces físicos del cerebro están formados por neuronas. Éstas se agrupan de forma ordenada a lo largo de la materia blanca formando tractos neuronales que conectan las distintas regiones cerebrales. El conjunto de estas conexiones neuronales se denomina **conectoma**. Dado que no todas las regiones se interconectan entre sí de la misma manera es posible caracterizar los enlaces de la red en función de la existencia de conexión entre dos regiones o no.

La única forma de medir estas conexiones de forma no invasiva es mediante **imágenes ITD** y la posterior aplicación de algoritmos de **tractografía**. A grandes rasgos, las imágenes ITD miden la difusión de moléculas de agua en diferentes direcciones y permiten inferir la dirección de las fibras en cada vóxel. Una vez conocidas las direcciones de difusión de cada vóxel, los algoritmos de tractografía tratan de reconstruir las fibras. Para ello, en primer lugar, se definen múltiples semillas en vóxeles aleatorios a lo largo del cerebro y a partir de cada una de ellas, se trazan las fibras siguiendo la dirección de difusión preferente por cada vóxel. El funcionamiento real de los algoritmos más avanzados es más complejo pues se basan en un enfoque probabilístico donde la dirección de las fibras en cada vóxel se representa mediante una distribución de probabilidades y no una única dirección. Además, para que los resultados derivados de estas técnicas reflejen de la manera más feaciente las fibras reales, es necesario imponer ciertas condiciones como detener la creación de fibras en la frontera entre materia blanca y gris, limitar la longitud de las fibras o descartar aquellas con recorridos poco realistas desde un punto de vista anatómico.

Es importante destacar que los resultados derivados de estas técnicas permiten obtener una representación virtual de las conexiones existentes que en diversas situaciones puede diferir del conectoma real de un sujeto.

Una vez llevada a cabo la tractografía se obtienen como resultado un conjunto de fibras que recorren el cerebro siguiendo los caminos más probables marcados por la imagen ITD (figura 5.9a). A partir de estas fibras, superponiendo las regiones de un esquema de parcelación (figura 5.9b), es posible definir como métrica de conectividad estructural el número de fibras que conectan dos

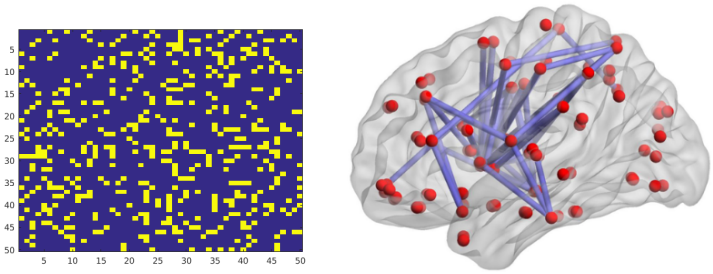
regiones concretas. La figura 5.9c muestra un ejemplo donde sólo se visualizan las fibras que pasan por las regiones amarilla y verde. Se observa cómo existen numerosas fibras que conectan dichas regiones mientras que una región alejada en el otro hemisferio (roja) no está conectada con ninguna de las dos. De este modo se calcula la conectividad entre cada par de regiones.



(a) Tractografía **(b)** Parcelación **(c)** 3 ROIs

Figura 5.9: Tractografía y cálculo de conectividad

Para eliminar la dependencia con el tamaño de la región es común normalizar dicho valor respecto al volumen agregado de las dos regiones involucradas. Estos valores pueden usarse directamente para construir una matriz de adyancencia (figura 5.10a) que defina la red a estudiar y modelar por tanto el conectoma real como una red, tal y como muestra la figura 5.10b.



(a) Matriz de adyacencia **(b)** Red cerebral. Obtenida de [31]

Figura 5.10: Matriz de adyacencia y red asociada

Al igual que con los esquemas de parcelación, es interesante definir una matriz de conectividad representativa de múltiples sujetos. Para ello es común unir los resultados derivados de la tractografía sobre un número lo más elevado posible de sujetos. Han surgido por lo tanto Atlas de tractos que definen el recorrido promedio de cada tracto neuronal. Uno de los más avanzados es el presentado en [32], construido en base a las imágenes ITD de 842 sujetos del proyecto HCP y que cubre toda la materia blanca.

5.2 Genética y conectividad

En los últimos años, gracias a las técnicas de procesado masivo de información de expresión genética basadas en **microarray** se han podido construir bases de datos del nivel de expresión del genoma completo a nivel cerebral. La base de datos más extensa y que mayor número de regiones cubre es el Allen Human Brain Atlas (AHBA) [33].

A día de hoy, más de 30 estudios han analizado estos datos y se comienza a comprender mejor como funciona la expresión genética a nivel cerebral y con qué procesos o patologías están relacionados cada uno de los genes [6].

Una propiedad importante que complica los estudios es la relación existente entre la expresión de muchos genes y la citoarquitectura cerebral. Los niveles de expresión de numerosos genes varían en función del tipo de células. Esto hace que exista una fuerte autocorrelación espacial encontrando valores similares de expresión genética en aquellas zonas más cercanas. A nivel cerebral, existe un gradiente de expresión genética desde el lóbulo frontal hasta el occipital. Esta propiedad complica en gran manera los análisis ya que dos regiones podrían presentar valores de expresión genética sólo por estar relativamente cerca y por lo tanto presentar similar tipo de células. y no por estar conectadas estructural ni funcionalmente [6].

Otro problema de estudios de este tipo es el correcto preprocesado de los datos. El elevado número de pasos y opciones existentes para realizar el análisis dificulta la reproducibilidad de los resultados. En [34] se presenta una guía de buenas prácticas para este tipo de estudios que diversos estudios previos no siempre han llevado a cabo.

En cuanto a la relación existente entre la expresión genética y la conectividad estructural, son varios los estudios realizados hasta hoy. El primer estudio [35] no encontró correlaciones directas entre exprexión genética y conectividad estructural pero sí una leve asociación entre ellas. Conviene destacar que el preprocesado de datos reportado es relativamente simple y se omiten algunos pasos considerados importantes en [34] como el filtrado por intensidad o selección de muestras. En [36] se realiza un estudio bastante más extenso y sí se identifican 11 clusters de genes asociados con la conectividad de regiones concretas. Aún así, aunque se menciona, no se detalla en profundidad el método seguido para eliminar la influencia de la distancia de los resultados.

Más recientemente, en [37] se hace uso de técnicas de Machine Learning para determinar los genes que mejor se correlan con la conectividad de diferentes regiones. Los resultados muestran que la expresión genética codifica de mejor manera la conectividad funcional que la estructural. Tampoco se detalla el método seguido para eliminar la influenccia con la distancia.

En conclusión, los estudios tanto en humanos como en roedores indican que la conectividad estructural debería tener una base genética. Sin embargo, la comprensión que tenemos sobre esta relación es bastante limitada.

5.3 Herramientas de análisis

Dentro del campo de la neurociencia computacional existen numerosas herramientas SW de visualización y análisis que facilitan la labor del investigador y, a menudo, representan contribuciones científicas muy relevantes. Aunque algunas de ellas se describen en el apartado de análisis de alternativas , a continuación se citan de forma expresa aquellas empleadas a lo largo del proyecto.

- **FSL** (FMRIB Software Library) para la visualización y registro de neuroimagen [38, 39, 40],
- **DSI Studio** para tractografía y visualización de imágenes [41].
- **Brain Connectivity Toolbox** para el análisis de comunidades [42].
- **Graph Signal Processing Toolbox** para el análisis de comunidades [43].
- **BrainNet Viewer** para la visualización de nodos y regiones [44].

6 Análisis de alternativas

Para llevar a cabo un proyecto de este tipo es importante escoger las herramientas más adecuadas entre las múltiples alternativas existentes. En concreto, es necesario escoger tanto el entorno de programación como diversas herramientas más específicas en neurociencia empleadas para el procesado y visualización de neuroimagen. En este apartado se detallan las alternativas valoradas así como los criterios que se han tenido en cuenta para escoger entre ellas.

El proceso de selección comienza con la definición de una serie de criterios con un coeficiente de ponderación asociado. A partir de aquí, se asigna un valor entre 0 y 1 a cada par criterio-alternativa. La nota final de cada alternativa se obtiene como la suma de todas las valoraciones ponderadas por los coeficientes de ponderación de cada criterio. Finalmente, se escoge la alternativa que obtiene mayor puntuación.

6.1 Entorno de programación

Es el entorno software donde se van a realizar la mayoría de los análisis. Por ello, es fundamental escoger un entorno versátil que permita programar de forma ágil y sencilla.

6.1.1 Alternativas

Matlab

Es un entorno de programación integrado (IDE, Integrated Development Environment) que cuenta con su propio lenguaje de programación, ambos propietarios y de pago. Está concebido para llevar a cabo cálculos matemáticos de cualquier tipo y optimizado para trabajar con matrices.

Tanto el entorno de desarrollo como el lenguaje en sí son relativamente sencillos, convirtiéndolo en una buena opción como lenguaje de scripting. Sin embar-

go, para el desarrollo de aplicaciones o proyectos software de mayor envergadura puede resultar menos ágil que otros lenguajes. Además, por tratarse de un lenguaje interpretado resulta menos eficiente que lenguajes compilados como C.

Destaca por la gran cantidad y calidad de de los paquetes software especializados, denominados toolbox, que incluyen funciones que permiten trabajar de forma sencilla en campos muy específicos como el diseño de filtros, modelado de sistemas de comunicación, análisis genéticos, etc.

Debido a su sencillez y a la posibilidad de obtener licencias de estudiante es empleado de forma extendida en universidades. Sin embargo, el elevado coste tanto de la versión básica como de los diferentes toolboxes es una de las razones por las que su uso a nivel industrial es menor y se limita a situaciones donde las funcionalidades de un determinado toolbox justifican su coste.

En lo que respecta a neurociencia, existen numerosos toolbox tanto oficiales como desarrollados por terceros enfocados al procesado de neuroimagen. Para los análisis estadísticos el toolbox oficial **Statistics and Machine Learning Toolbox** permite realizar la gran mayoría de análisis convencionales. Además, para el manejo y estudio de grafos redes destaca el toolbox de acceso libre **Brain Connectivity Toolbox**.

Python

Se trata de un lenguaje abierto de propósitio general empleado en múltiples campos, tanto como lenguaje de scripting como para el desarrollo de aplicaciones. La sintaxis es sencilla y, a pesar de estar orientado a objetos, resulta un lenguaje fácil de aprender. Además, el uso de la sangría para agrupar los elementos hace que el código sea en general más compacto y legible.

Su versatilidad y facilidad de uso le ha situado entre los lenguajes más empleados a nivel global. Sin embargo, al igual que Matlab, es un lenguaje interpretado, situándose su rendimiento por debajo de lenguajes compilados como C o C++.

Una de las ventajas de usar Python es su portabilidad, el usuario tiene total libertad para escoger el entorno de desarrollo que mejor se adapte a sus necesidades, pudiendo usar desde un simple bloc de notas hasta IDEs más sofisticados como PyCharm o Spyder.

A nivel científico existen paquetes que cubren la gran mayoría de necesidades como **NumPy** (cálculo matemático), **Matplotlib** (visualización) o **Scikit-learn** (Machine Learning). La calidad así como la facilidad de uso de estos paquetes han convertido a Python en una de los lenguajes más populares en investigación.

Para neurociencia existen también paquetes para el procesado de neuroimagen como **NiPy** y una versión en Python del **Brain Connectivity Toolbox**.

R

Es un lenguaje y entorno de programación orientado al análisis estadístico. Se trata de un software libre desarrollado originalmente para Unix aunque disponible ya para los demás sistemas operativos. Es un lenguaje interpretado y está orientado al trabajo con datos tanto en forma de matriz como en otro tipo de estructuras.

Destaca por la gran variedad librerías de visualización que permiten generar plots y figuras avanzadas de forma sencilla y con alta calidad. Por todo esto, es uno de los lenguajes preferidos en Data Science. A pesar de todo, para ámbitos más específicos como la neuroimagen el número de paquetes existentes es menor al de otros entornos, limitando su uso al análisis estadístico de los datos.

Octave

Se trata de un lenguaje de programación y entorno libres empleados para el cálculo numérico. Es posible emplearlo tanto a través de consola como desde la interfaz de usuario del programa. Representa una alternativa de software libre a Matlab, ya que la sintaxis es prácticamente la misma.

Dada su compatibildad con Matlab, es posible reutilizar código de Matlab con funcionalidades básicas. Sin embargo, en numerosas ocasiones es complicado reimplementar con éxito y rapidez código de Matlab complejo. Además, es algo menos estable que Matlab y con una ayuda menos detallada.

Existen paquetes de funciones diseñados para Octave aunque en mucha menor medida que los existentes para Matlab. Es una alternativa real para el desarrollo aplicaciones sencillas que no requieran toolboxes específicos y donde Matlab quede descartado debido a su alto coste.

La existencia de funcionalidades para neurociencia se limita a la posibilidad de reimplementar con éxito toolboxes desarrolladas para Matlab.

6.1.2 Criterios de selección

Curva de aprendizaje

Dado al elevado número de scripts a realizar, se busca emplear un entorno de programación que permita obtener la mayor velocidad de implementación posibe. Por ello, se ha valorado cuánto tiempo llevaría aprender a usar con agilidad cada uno de los lenguajes teniendo en cuenta la experiencia y conocimientos previos.

Herramientas específicas

Además de las funcionalidades de cálculo básicas, es necesario hacer uso de herramientas más específicas para análisis estadístico, procesado de neuroimagen

o aplicación de técnicas de Machine Learning. Se ha considerado por tanto la cantidad y calidad de las librerías existentes para cada uno de los entornos de programación.

Calidad de visualización

Dada la naturaleza del proyecto, es importante escoger un entorno que permita obtener gráficas y figuras de alta calidad de forma relativamente sencilla. En este criterio se valoran los tipos de gráficas y figuras que pueden generarse con cada uno de los entornos.

Coste

Como en cualquier proyecto de ingeniería, es necesario tener en cuenta el coste de cada una de las alternativas.

6.1.3 Selección

En base a los resultados mostrados en la tabla 6.1, el entorno escogido ha sido Matlab. A pesar de su coste, es el entorno con menor curva de aprendizaje y ofrece una gran variedad de herramientas de visualización y análisis.

Criterio	Ponderación	**Matlab**	Octave	Python	R
Curva de aprendizaje	4/10	10/10	9/10	7/10	6/10
Herramientas específicas	3/10	10/10	8/10	10/10	5/10
Calidad de visualización	2/10	8/10	7/10	8/10	10/10
Coste	1/10	3/10	10/10	10/10	10/10
Total	10/10	**8.7/10**	8.4/10	8.4/10	6.9/10

Tabla 6.1: Criterios de selección del entorno de programación

6.2 Atlas de regiones cerebrales

Uno de los componentes más importantes del proyecto es la construcción y comparación de matrices de conectividad cerebrales. Para obtener dichas matrices es necesario subdividir el cerebro en regiones a partir de un template de referencia conocido como Atlas. El Atlas, es realmente un conjunto de máscaras que definen las regiones sobre el espacio estándar, de modo que sea posible definirlas para un nuevo sujeto mediante un proceso de registro.

6.2.1 Alternativas

Automated Anatomical Labeling (AAL)

Se trata de un Atlas implementado en Matlab dentro del paquete de software de libre acceso SPM (Statistical Parametric Mapping). Fue uno de los primeros Atlas modernos y por tanto, ha sido bastante empleado en la literatura. La versión original del Atlas, publicada en 2002, consta de un total de 90 regiones corticales definidas únicamente en base a criterios anatómicos. En concreto, se definen en base a los surcos que presenta el cerebro [45]. Actualmente se trabaja en la extensión del atlas mediante la definición de un mayor número de regiones.

Glasser

El atlas Glasser, publicado en 2016, tiene una notable resolución ya que consta de 360 regiones corticales [46]. A diferencia del atlas AAL, está basado en imágenes IRM multimodales de un total de 210 sujetos del proyecto HCP [29]. Así, para la delimitación de las regiones, además de propiedades anatómicas, se han empleado criterios relativos a la conectividad funcional y estructural. En consecuencia, las regiones obtenidas no son meras particiones anatómicas si no que son regiones con propiedades independientes.

Craddock

Publicado en 2018, además de un Atlas, presenta una metodología para definir nuevos atlas con un número de regiones deseado en base a imágenes IRM funcionales [28]. Las regiones anatómicas obtenidas son por tanto aquellas que se activan de forma independiente. Su definición se hace en base a clustering espectral pudiendo definir un número de regiones tan elevado como sea necesario. La comparación de las particiones obtenidas con otros Atlas únicamente basados en la anatomía como el AAL evidencia la poca validez de dichos Atlas para definir regiones funcionalmente diferenciadas.

Desikan-Kiliany

Esta basado en las imágenes IRM estructurales de 40 sujetos en las que fueron definidas manualmente un total de 68 regiones corticales [47]. Fue publicado en 2006 y, al igual que el atlas AAL, las regiones fueron definidas en base a características anatómicas. Además de ello, también se define un método para obtener poder definir las regiones del Atlas en nuevos sujetos.

6.2.2 Criterios de selección

Resolución

Es importante emplear un Atlas con un número óptimo de regiones que permita realizar el análisis con el mayor grado de resolución posible sin exceder

la limitación impuesta por la existencia de muestras de expresión genética.

Criterios de definición

Se valoran los parámetros tenidos en cuenta para definir las regiones de cada Atlas. En este proyecto se estudia la conectividad estructural por lo que es interesante emplear un Atlas cuyas regiones hayan sido definidas teniendo en cuenta criterios de conectividad y no solo la anatomía.

Disponibilidad

Se valora que el Atlas sea de acceso libre y que sea posible hacer uso de él de una forma sencilla, en múltiples entornos y sin tener que instalar plataformas software adicionales o llevar a cabo algún tipo de procesado.

6.2.3 Selección

Las valoraciones se muestran en la tabla 6.2. El Atlas escogido finalmente es el **Glasser**. Se trata de un atlas moderno con una resolución suficiente para los análisis a realizar. Además el hecho de que para la definición de las regiones se hayan considerado criterios de conectividad lo hace idóneo para análisis como los realizados en este proyecto.

Criterio	Ponderación	AAL	Desikan	**Glasser**	Craddock
Resolución	4/10	5/10	4/10	9/10	10/10
Criterios de definición	4/10	6/10	8/10	10/10	8/10
Disponibilidad	2/10	7/10	7/10	9/10	6/10
Total	10/10	5.8/10	6.2/10	**9.4/10**	8.4/10

Tabla 6.2: Criterios de selección del Atlas

6.3 Visualización y procesado de neuroimagen

En numerosos puntos del proyecto es fundamental poder visualizar con rapidez imágenes IRM estructurales así como máscaras referentes a Atlas o a determinadas regiones cerebrales. Además, es necesario también tener herramientas que permitan procesar dichas imágenes ya sea para realizar registros o edición de las mismas. Dentro de las múltiples herramientas existentes, se ha decidido valorar aquellas más comunmente empleadas por la comunidad cientifica.

6.3.1 Alternativas

FSL

Es un conjunto de librerías y herramientas para el análisis de imágenes IRM estructurales, funcionales y de difusión desarrolladas por la universidad de Oxford. Se trata de software libre implementado en Linux aunque puede ser usado en Windows mediante máquina virtual. Entre las múltiples opciones que ofrece, contiene herramientas de visualización, segmentación y registro de neuroimagen.

Las herramientas básicas son accesibles mediante una interfaz de usuario gráfica (GUI, Graphical User Interface). Sin embargo, todas las herramientas pueden ser ejecutadas desde línea de comandos o mediante scripts en bash. Este modo, aunque más avanzado, permite automatizar y personalizar las herramientas con una mayor libertad.

Existe gran cantidad de documentación disponible acerca del uso de las diferentes herramientas, incluyendo tanto cursos online como guías de usuario. Esto hace que el aprendizaje sea relativamente fluido y sea una de las herramientas más empleadas a día de hoy.

FreeSurfer

Se trata de un conjunto de software desarrollado por la universidad de Harvard para el procesado y análisis de imágenes IRM. De forma similar a FSL, se trata de software libre que incluye herramientas para el procesado y visualización de neuroimagen.

Se han desarrollado versiones tanto para Linux como para Mac OS y su instalación, aunque larga, es sencilla. Existen bastantes tutoriales online, así como cursos de pago que permiten aprender a usar las diferentes herramientas que, en su mayoría, funcionan en base a comandos.

SPM

SPM es un paquete de software de libre acceso desarrollado por la University College of London y orientado al procesado de neuroimagen en Matlab. En concreto, se centra en el estudio estadístico de imágenes IRM funcionales.

La mayoría de las herramientas funcionan desde una GUI, por lo que es bastante popular entre aquellos neurocientíficos no familiarizados con la programación. Para aprender a usar el software existe un extenso manual de uso así como numerosos tutoriales online.

6.3.2 Criterios de selección

Curva de aprendizaje

Cada una de las alternativas tiene unas particularidades diferentes y requiere por lo tanto de un proceso de aprendizaje. Además de las dificultades intrínsecas de cada una de las alternativas, se ha considerado la experiencia previa del grupo de investigación en el uso de cada una de ellas.

Funcionalidad

Las herramientas que ofrece cada uno de los paquetes software son distintas. En este caso, se busca que la visualización de imágenes IRM estructurales sea sencilla y fluida tanto en 2D como en 3D. Además, es de interés que el proceso de registro entre imágenes sea fiable y sencillo de implementar.

Coste

Es importante considerar el coste de cada una de las opciones. Aunque todas ellas son de acceso libre, SPM funciona sobre Matlab por lo que realmente requiere de un software de pago para su funcionamiento.

6.3.3 Selección

Finalmente, se ha escogido **FSL** por ser la herramienta más versátil de todas y la empleada habitualmente en el laboratorio. Las puntuaciones se muestran en la tabla 6.3.

Criterio	Ponderación	**FSL**	FreeSurfer	SPM
Curva de aprendizaje	4/10	7/10	6/10	7/10
Funcionalidad	4/10	9/10	8/10	6/10
Coste	2/10	10/10	10/10	3/10
Total	10/10	**9.2/10**	8.4/10	5.8/10

Tabla 6.3: Criterios de selección de la herramienta de procesado de neuroimagen

6.4 Software de tractografía

Es necesario seleccionar también el software para llevar a cabo tareas de tractografía y obtener matrices de conectividad estructural. A continuación, se detallan las alternativas valoradas así como el proceso de selección llevado a cabo.

6.4.1 Alternativas

DSI Studio

Es un software gratuito desarrollado para el análisis de imágenes ITD y la aplicación de algoritmos de tractografía. Permite cargar imágenes ITD y reconstruir a partir de ellas los tractos neuronales mediante un algoritmo de tractografía con múltiples parámetros configurables. Es posible definir ROIs específicas sobre las que llevar a cabo dicho proceso.

Por otro lado, funciona también como herramienta de visualización, mostrando las fibras en 3D en diferentes calidades, pudiendo también filtrar o agrupar las fibras según sea necesario. Además es una buena herramienta de visualización de las diferentes ROIs que forman los Atlas. Destaca por la posibilidad de realizar tractografía sobre el total de la matería blanca y obtener matrices de conectividad derivadas.

Probtrackx

Es la herramienta de tractografía incluida en FSL. Aunque es posible configurar ciertos parámetros del algoritmo, en su configuración por defecto, sólo permite obtener 27 tractos neuronales individualmente y no es posible realizar tractografía del cerebro completo. Este funcionamiento es de especial interés para aquellos estudios en los que se busca extraer y comparar entre sujetos diversos tractos neuronales considerados como principales y para los cuales se conocen las regiones que los delimitan. Aun así, aunque no de forma sencilla, es posible extraer nuevos tractos mediante nuevas regiones delimitadoras. Los resultados se pueden visualizar como máscaras en 2D y 3D pero no es posible visualizar las fibras individualmente.

TRACULA

Se trata del algoritmo de tractografía implementado en FreeSurfer mediante el cual es posible obtener únicamente 18 tractos principales. El funcionamiento del algoritmo no permite añadir nuevos tractos ni llevar a cabo tractografía sobre el cerebro completo. Análogamente a Probtrackx, es posible visualizar las máscaras de los tractos obtenidos mediante cualquier herramienta de visualización de máscaras en 2D y 3D.

6.4.2 Criterios de selección

Configuración

Es interesante que el algoritmo de tractografía sea configurable de modo que se puedan definir el número máximo de fibras, la longitud mínima y máxima de las fibras etc.

Visualización

Se valora que sea posible visualizar los tractos obtenidos así como las máscaras empleadas para su delimitación de forma sencilla y con buena calidad.

Conectividad

Para obtener matrices de conectividad cerebral es necesario realizar la tractografía a lo largo de toda la materia blanca y no solo en unos tractos concretos. Además es interesante que la misma herramienta de tractografía permita obtener matrices de conectividad basadas en las fibras generadas.

6.4.3 Selección

Las ponderaciones de cada criterio y los resultados se muestran en la tabla 6.4. La herramienta que más se adapta es DSI Studio ya que es la que mayores opciones de configuración y visualización tiene. Además la simplicidad con la que permite obtener matrices de conectividad es otra ventaja.

Criterio	Ponderación	**DSI Studio**	Probtrackx	Tracula
Configuración	3/10	9/10	7/10	4/10
Visualización	5/10	9/10	7/10	7/10
Conectividad	2/10	9/10	0/10	0/10
Total	10/10	**9/10**	5.6/10	4.7/10

Tabla 6.4: Criterios de selección de la herramienta de tractografía

7 Análisis de riesgos

A pesar de tratarse de un proyecto de investigación, es necesario también considerar los posibles riesgos que pudieran afectar al correcto transcurso del proyecto. Para poder hacer frente a dichos riesgos, es necesario en primer lugar identificarlos y caracterizarlos en base a su probabilidad de ocurrencia y impacto sobre el proyecto. Por ello, se ha asignado a cada uno de los riesgos que se detallan a continuación valores en la escala Bajo-Medio-Alto tanto para la probabilidad de ocurrencia como para el impacto.

Una vez caracterizados los riesgos, es importante plantear medidas para evitar cada uno de ellos y paliar sus posibles efectos en lo que se conoce como plan de contingencia. Dicho plan, detalla los pasos a seguir para solventar o corregir la problemática derivada de cada uno de los riesgos.

Riesgo 1: Errores en las hipótesis de partida

Es posible que alguna de las asunciones sobre las que se sustenta el proyecto no sea correcta. Esto podría implicar que todos los análisis posteriores carecieran de sentido y por lo tanto hubiera que replantear el proyecto de forma global.

Se trata de un riesgo bastante común en líneas de investigación de una naturaleza más exploradora como el presente proyecto, donde se busca avanzar en la comprensión de un tema extenso y complejo y no en la solución a un problema específico.

Probabilidad de ocurrencia: **Media**

Impacto: **Alto**

Plan de contingencia: La mejor forma de evitar este riesgo es planteando las bases del proyecto con cautela. Si aun así se se detectaran debilidades en las hipótesis de partida, se identificará qué partes son válidas y permiten avanzar hacia los objetivos del proyecto, y qué partes no son válidas y habrá que desechar o reformular.

Riesgo 2: Pérdida de datos

Como en cualquier proyecto, existe el riesgo de que ciertos datos del mismo se pierdan y parte de los avances del proyecto desaparezcan. Se trata, por lo tanto, de un riesgo a tener en cuenta cuyo impacto aumenta a medida que se avanza en el proyecto.

Probabilidad de ocurrencia: **Baja**

Impacto: **Alto**

Plan de contingencia: La forma más efectiva de hacer frente a este tipo de riesgo es mediante la paulatina realización de copias de seguridad. Estas copias deben ser actualizadas cada cierto tiempo especialmente en aquellas partes vitales del proyecto como son el informe final o los scripts de análisis más importantes. En esta línea, el uso de repositorios con control de versiones es una alternativa efectiva y muy conveniente.

Riesgo 3: Problemas de personal

Aunque poco frecuente, existe la posibilidad de que alguna de las personas involucradas en el proyecto dejara de formar parte de él voluntaria o involuntariamente. El impacto de dichas bajas estaría relacionado con la duración del periodo de ausencia y el rol de la persona en el proyecto.

Probabilidad de ocurrencia: **Baja**

Impacto: **Medio**

Plan de contingencia: intentar, en la medida de lo posible, que las personas involucradas en el proyecto adquieran un compromiso con el mismo. Si aun así, alguna de las personas no estuviera disponible durante cierto tiempo, deberían buscarse posibles sustitutos o reasignar las competencias de dicha persona a alguna otra capacitada para ello.

Riesgo 4: Retrasos

Las causas que pueden hacer que un proyecto se retrase son múltiples, desde resultados inesperados, problemas personales o una planificación errónea. Si bien los retrasos leves son relativamente frecuentes, un retraso excesivo podría suponer la no finalización del proyecto en plazos y por lo tanto su fracaso. Por ello, es necesario contemplar dicha posibilidad y tratar de minimizar el riesgo derivado.

Probabilidad de ocurrencia: **Media**

Impacto: **Medio**

Plan de contingencia: planificar con la suficiente antelación las diferentes fases del proyecto detallando los objetivos e hitos de cada una de ellas. Es importante que la planificación contemple la posibilidad de incurrir en retrasos inesperados. Además, a lo largo del proyecto, es de suma importancia tratar de adherirse lo máximo posible a la planificación establecida.

Riesgo 5: Desvío del presupuesto

Siempre cabe considerar que por causas de fuerza mayor el proyecto se desvíe del presupuesto inicial. Aun así, en el presente proyecto este riesgo es realmente poco probable debido a las pocas partidas de gastos involucradas. Por dicha razón, además, el impacto sobre el proyecto sería muy pequeño.

Probabilidad de ocurrencia: **Baja**

Impacto: **Bajo**

Plan de contingencia: presupuestar cuidadosamente todos los gastos al inicio del proyecto, incluyendo un pequeño porcentaje para imprevistos. En caso de superar el presupuesto, simplemente reajustar el presupuesto inicial y tratar de evitar gastos que supongan un gran desvío respecto a lo presupuestado originalmente.

Tabla resumen

A modo de resumen, en la tabla 7.1 se muestran ubicados los cinco riesgos principales identificados y descritos en la sección anterior.

		Probabilidad de ocurrencia		
		Bajo	Medio	Alto
Impacto	Bajo	5	-	-
	Medio	3	4	-
	Alto	2	1	-

Tabla 7.1: Análisis de riesgos

8 Descripción de la solución

8.1 Preprocesado de datos genéticos

A día de hoy, el dataset de expresión genética a nivel cerebral más completo que existe es el **Allen Human Brain Atlas (AHBA)** [33], desarrollado por el centro de investigación Allen Institute. A lo largo de un proceso de tres años se midió el nivel de expresión de más de 20 000 genes en un total de 3702 muestras obtenidas de seis cerebros donantes sanos. Sin embargo, esta toma de muestras no fue uniforme, existiendo diferencias entre donantes tanto en el número de muestras como en el de los hemisferios muestreados (ver tabla 8.1).

Id Donante	1009	1012	1015	1016	2001	2002
Muestras	363	529	470	501	946	893
Hemisferio	Izquierdo				Ambos	

Tabla 8.1: Número de muestras y hemisferios de cada donante

Tal y como muestra la figura 8.1, las gran mayoría de las muestras fueron tomadas en la materia gris cubriendo la corteza cerebral, cerebelo, tronco cerebral y diversas estructuras subcorticales.

Los niveles de expresión genética fueron medidos mediante la técnica **microarray**, responsable de la gran resolución del AHBA. Sin embargo, el uso de esta técnica así como el hecho de que las muestras hayan sido obtenidas de múltiples sujetos hace necesario un preprocesado previo de los datos del AHBA. Con este fin, junto con los valores de expresión genética se incluyeron diversos ficheros de anotaciones que resultan de gran utilidad en las diversas etapas de preprocesado.

En el presente proyecto se ha decidido seguir las etapas sugeridas en [34], que propone un pipeline común que permita mejorar la reproducibilidad de los resultados. El objetivo final es doble: obtener una matriz de N genes x 3702 muestras preparada para su uso en diferentes análisis e implementar todas las

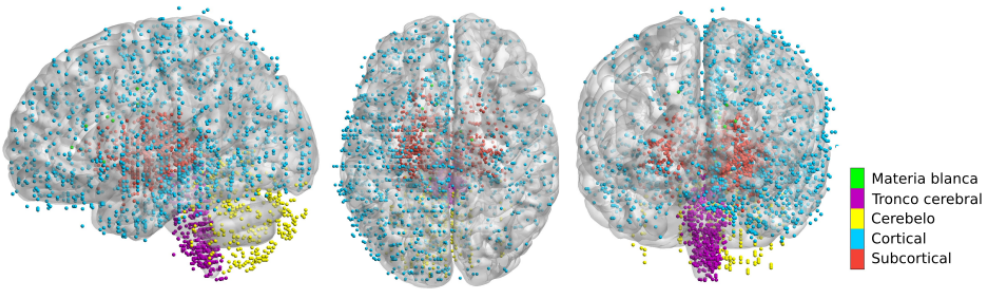


Figura 8.1: Muestras que componen el atlas AHBA

etapas de forma que sea relativamente sencillo volver a preprocesar los datos con diferentes parámetros.

8.1.1 Anotación sonda-gen

El primer paso viene derivado de que la técnica microarray no mide la expresión de cada gen directamente. En realidad, los chips de ADN constan de múltiples sondas de medida que se alinean con diferentes partes de la secuencia de ADN. Cuando se conoce a qué gen pertenece cada parte de dicha secuencia es posible asociar cada sonda con un gen concreto. En la práctica, a la hora de realizar esta asociación pueden darse varios casos distintos:

1. Una sonda asociada con un solo gen
2. Múltiples sondas asociadas con el mismo gen
3. Sondas sin gen asociado

Los chips empleados en el AHBA contienen un total de 58 692 sondas de las cuales sólo unas pocas tienen un único gen asociado. Por ello, es necesario llevar a cabo un proceso de anotación para conocer con qué gen están asociadas cada una de ellas y posteriormente descartar aquellas sin gen asociado.

Uno de los ficheros de anotación incluidos en el dataset proporciona una asignación realizada en el momento de su publicación (2012). Sin embargo, los avances en el campo de la secuenciación genética hacen que dichas anotaciones queden progresivamente obsoletas siendo necesario repetir el proceso con las bases de datos de secuenciación genética más recientes.

La anotación empleada en este proyecto fue realizada previamente al comienzo del mismo mediante el software **Re-Annotator** [48] en enero de 2019. A partir de estos resultados se han descartado 11 147 sondas sin gen asociado quedando un total de 47 545 restantes.

8.1.2 Filtrado basado en intensidad

De los más de 20 000 genes que componen el genoma humano es esperable que sólo una fracción de ellos esté relacionada con la conectividad del cerebro. Por tanto, resulta de especial interés tratar de reducir el número de genes a aquellos que presenten un nivel de expresión relevante a nivel cerebral.

En esta línea, otro de los ficheros de anotación que forma parte del dataset asigna a cada uno de los pares sonda-muestra un valor binario que indica si el nivel de expresión medido excede el nivel de ruido de forma estadísticamente significativa. Estas anotaciones permiten clasificar las sondas en base a su nivel de señal.

Tratando de eliminar aquellas sondas con un nivel de expresión reducido se ha llevado a cabo un **filtrado basado en intensidad** mediante el cual todas aquellas sondas que no superan el nivel de ruido en al menos el 50 % de las muestras son descartadas.

La figura 8.2 muestra cómo existen múltiples sondas con un nivel de expresión muy bajo. En concreto, el 31 % de las sondas (14 796) no alcanzan el umbral fijado y han sido excluidas quedándonos finalmente con 32 749.

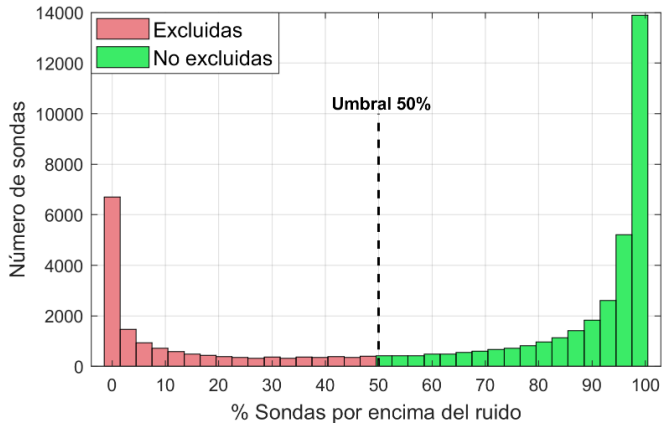


Figura 8.2: Distribución de sondas para el filtrado basado en intensidad

8.1.3 Selección de sondas

Tras haber descartado las sondas sin gen asociado o con bajos niveles de expresión, todavía es necesario resolver la problemática presentada en el apartado 8.1.1 donde múltiples sondas apuntan a un único gen. Idealmente los niveles de expresión de sondas asociadas con el mismo gen deberían ser similares pero, debido a las múltiples fuentes de error esto no está asegurado.

Para tratar de mitigar estos errores, durante la creación del AHBA también se midieron los niveles de expresión de 17 769 genes en 301 de las 3702 muestras

mediante la técnica **RNA-Seq**. Este método resulta más preciso que microarray pero no permite el procesado masivo. Aun así, es posible utilizar estos valores como medidas de control en las muestras analizadas tanto mediante microarray como RNA-Seq.

A pesar de que la técnica RNA-Seq ofrece valores de sólo 17 769 genes, los resultados obtenidos en [34] muestran que los genes restantes no están especialmente asociados con procesos fisiológicos cerebrales. Por ello, se ha decidido excluir todas aquellas sondas asociadas con dichos genes.

Para las sondas restantes, la idea es comparar los valores obtenidos por cada sonda de microarray con los obtenidos mediante RNA-Seq y seleccionar las sondas que presenten una mayor correlación. Esta correlación se ha medido mediante el coeficiente de correlación de Spearman [49]:

$$\rho_{X,Y} = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)}, \tag{8.1}$$

donde d_i es la diferencia entre los estadísticos de orden i de las variables X e Y y N es el número de muestras.

Tal y como muestra la figura 8.3, en torno al 51 % de las sondas muestran una correlación menor de 0.2 y solo el 10 % supera el valor de 0.5. Se han excluido por tanto todas aquellas sondas con un valor inferior a 0.2 quedándonos con un total de 13974 sondas. Para los casos en los que múltiples sondas están asociadas con un mismo gen se han elegido las de mayor correlación, llegando finalmente al número definitivo de 8068 genes que serán empleados en siguientes análisis.

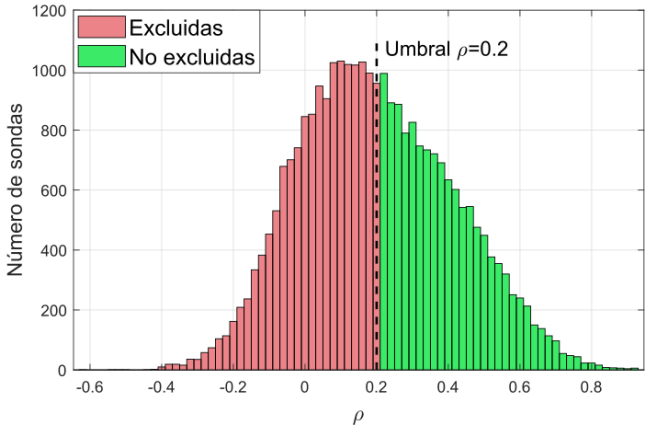


Figura 8.3: Correlación entre microarray y RNA-Seq

8.1.4 Normalización

Pese a que los datos del AHBA ya han pasado por un proceso de normalización previo, existe aún una varianza considerable entre sujetos, estando las

muestras de cada sujeto más correladas entre sí [34]. Esto, supone un claro problema para estudios en los que se requiere utilizar las muestras de todos los sujetos como si pertenecieran a un mismo cerebro.

El objetivo es normalizar el valor de expresión de cada gen a lo largo de todas las muestras de forma independiente para cada sujeto. Para ello, tratando de eliminar también la influencia de outliers, se ha empleado una variante del comunmente empleado z-score denominada Scaled Robust Sigmoid (SRS) [50] definida como

$$x`[i] = \frac{1}{1 + e^{-\frac{x[i] - \langle x \rangle}{\sigma}}} \quad i = 1, 2, \ldots N, \tag{8.2}$$

donde x es el vector con los valores de expresión de un gen en las N muestras de cada donante, y $\langle x \rangle$ y σ son respectivamente la mediana y la desviación estándar de dicho vector, y x' son los valores normalizados entre 0 y 1 . Finalmente, se obtienen los valores x_{norm} tras modificar el intervalo para que quede exactamente entre -0.5 y 0.5 mediante:

$$x_{norm}[i] = \frac{x'[i] - min(x')}{max(x') - min(x')} - 0.5 \quad i = 1, 2, \ldots N \tag{8.3}$$

8.1.5 Visualización de los datos

Tras haber preprocesado los datos y antes de pasar a análisis relacionados con la conectividad estructural, resulta interesante ver los datos de forma gráfica. Todas las muestras del AHBA tienen una región anatómica asociada en un fichero de anotación. Además, resultados previos como los publicados en [36, 35, 51] demuestran que las regiones anatómicas diferenciadas (cortex, cerebelo, tronco cerebral y estructuras subcorticales) muestran perfiles de expresión genética muy diferenciados.

Para dicha visualización se ha hecho uso de la técnica de reducción de la dimensionalidad no lineal conocida como **t-SNE** (t-Distributed Stochastic Neighbor Embedding) [52]. Partiendo de la matriz de 8068 genes x 3702 muestras, el objetivo es reducir los 8068 genes a únicamente dos dimensiones que puedan ser visualizadas en un scatterplot. Haciendo uso de las anotaciones para asignar colores en función de la región anatómica es posible representar los datos como en la figura 8.4.

Tal y como se esperaba, se aprecian claras diferencias entre regiones. Las muestras en la corteza cerebral, pese a su gran número, se agrupan de forma separada a las demás regiones. De forma similar, las muestras tanto del cerebelo como del tronco cerebral también forman clusters relativamente compactos. En cuanto a las muestras subcorticales, se observa una diferenciación pero también cierto solape especialmente con el tronco cerebral. Este solape puede ser debido a que ciertas muestras categorizadas como subcorticales pertenecen realmente

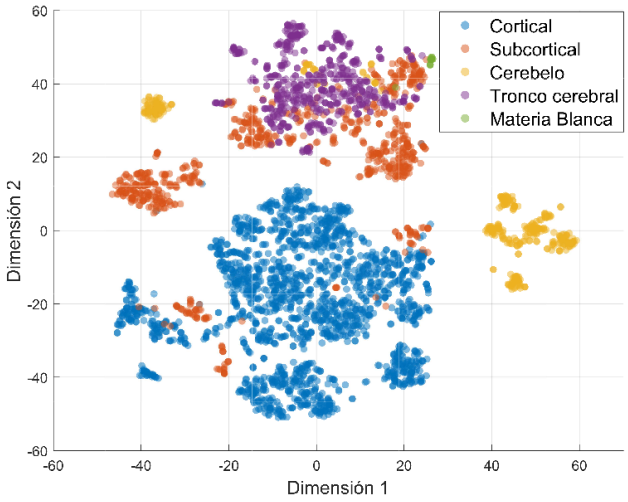


Figura 8.4: Visualización de las muestras mediante t-SNE

a la parte superior del tronco cerebral y forman realmente parte de la misma estructura anatómica. En cuanto a la materia blanca, el reducido número de muestras no permite determinar su agrupación.

Otra forma de evaluar la similitud genética entre muestras es mediante medidas de correlación genéticas o CGE (Correlated Gene Expression) [6]. El perfil genético de cada muestra es un vector de 1 x 8068 con los valores de expresión genética de cada gen. Se busca conocer cómo de correlados se encuentran los valores de expresión genética entre cada par de muestras. Para ello, se emplea comunmente el **coeficiente de correlación de Pearson** [53],

$$\rho_{X,Y} = \frac{E\big[(X-\mu_x)(Y-\mu_y)\big]}{\sigma_x \sigma_y}, \tag{8.4}$$

donde X e Y son dos variables aleatorias con medias son μ_x y μ_y y desviaciones estándar σ_x y σ_y

En la práctica, podemos estimar el coeficiente de correlación para dos poblaciones de muestras como:

$$r_{xy} = \frac{\sum_{i=1}^{8068}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{8068}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{8068}(y_i-\bar{y})^2}}, \tag{8.5}$$

donde x e y son cada uno vectores de 1 x 8068 con los valores de expresión de todos los genes en dos muestras cualquiera y $\bar{x}$ e $\bar{y}$ son sus respectivas medias.

El coeficiente toma valores entre -1 y 1, indicando 1 una correlación total, -1 una correlación total en sentido inverso y 0 no correlación. En este caso concreto, valores positivos indican que dos muestras tienen muchos genes en común con un nivel de expresión similar respecto a la media. Por el contrario, valores negativos

indicarían que existen numerosos genes con un nivel de expresión inferior a la media en una muestra y superior a la media en otra. Valores cercanos a 0 reflejarían la inexistencia de una relación entre muestras.

El resultado de la correlación de Pearson para todas las muestras de un donante se muestra en la figura 8.5. En dicha figura se han reordenado las muestras en función de las mismas anotaciones utilizadas para la visualización con t-SNE. De forma similar a como ocurría con t-SNE, las muestras tanto de la corteza como del cerebelo forman clusters independientes. Las muestras subcorticales y las del tronco cerebral forman a su vez un sólo cluster, no estando demasiado diferenciadas entre sí.

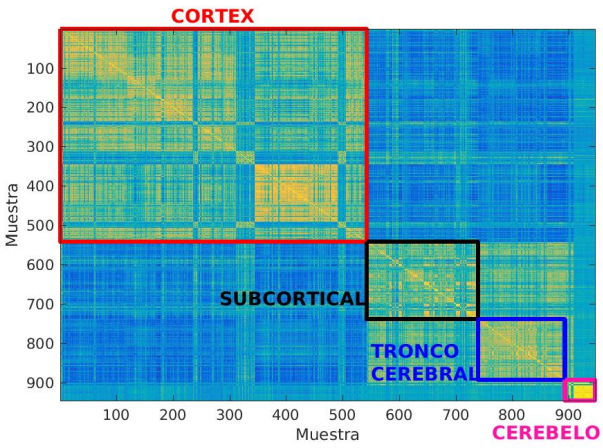


Figura 8.5: Comunidades basadas basadas en CGE del donante 2001

8.2 Obtención matrices de conectividad

El objetivo de este apartado es llegar a dos matrices con valores de conectividad genética y estructural comparables entre sí. Debido a las limitaciones de resolución de las técnicas de tractografía, no es posible obtener valores de conectividad estructural fiables a nivel de muestra (vóxel), por lo que es necesario hacer uso de un atlas de parcelaciones. En lugar de trabajar directamente con las 3702 muestras, se trata de subdividir el cerebro en diferentes regiones, asignar las muestras a su región correspondiente y trabajar con los valores de conectividad y expresión genética de cada región completa. Para ello, es necesario seguir una serie de pasos que permitan construir dichas matrices.

8.2.1 Selección del atlas

Tras observar los muy diferentes perfiles genéticos existentes entre las áreas corticales, subcorticales, cerebelo y tronco cerebral, se ha decidido focalizar el análisis sólo en la corteza cerebral. Se busca tratar de simplificar el problema

y evitar la problemática que supondría analizar de forma conjunta áreas tan diferentes. De este modo, de las 3702 muestras iniciales se ha pasado a trabajar con las 1950 pertenecientes a la corteza cerebral.

El atlas que se ha empleado es el denominado **Glasser** [46], que se muestra en la figura 8.6. Consta de 360 regiones que lo hacen uno de los atlas con mayor nivel de resolución a nivel cortical. Está basado en las imágenes MRI multimodales de un total de 210 sujetos obtenidos del Human Connectome Project (HCP) [29] y para la delimitación de las regiones se han combinado propiedades relativas a la arquitectura cerebral, conectividad funcional y conectividad estructural.

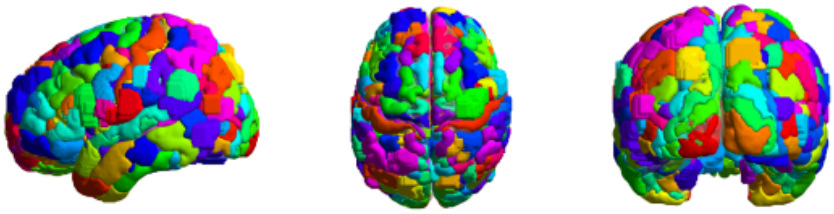


Figura 8.6: Atlas Glasser en 3D

8.2.2 Genética

Asignación de muestras a regiones

Una vez definidas las regiones a analizar en base al atlas, es necesario calcular el valor medio de expresión genética de cada gen en cada una de las regiones. Para ello, es necesario en primer lugar conocer a qué región pertenece cada una de las muestras. Para realizar estas asignaciones, cada una de las muestras tiene asociadas unas coordenadas tanto en el espacio sujeto como en el espacio estándar MNI. Además, también están disponibles imágenes T1 de cada uno de los donantes.

Dado que el atlas se encuentra en el espacio MNI, la opción más sencilla sería usar las coordenadas MNI directamente. Sin embargo, es más preciso utilizar las coordenadas en en espacio sujeto registrando previamente el atlas a dicho espacio [34]. El proceso a seguir para cada donante es el siguiente:

1. Registrar T1 donante a espacio MNI y obtener matriz de transformación
2. A partir de dicha matriz obtener la matriz de transformación inversa
3. Registrar el atlas de espacio MNI a espacio sujeto
4. Mapear las muestras a regiones usando las coordenadas en espacio sujeto

Cada una de las imágenes T1 de cada donante tienen dimensiones y centro de coordenadas diferentes y no se encuentran registradas al espacio MNI. La

figura 8.7 muestra el caso para el donante 2001, en rojo se muestra la imagen T1 original, claramente no registrada a espacio estándar.

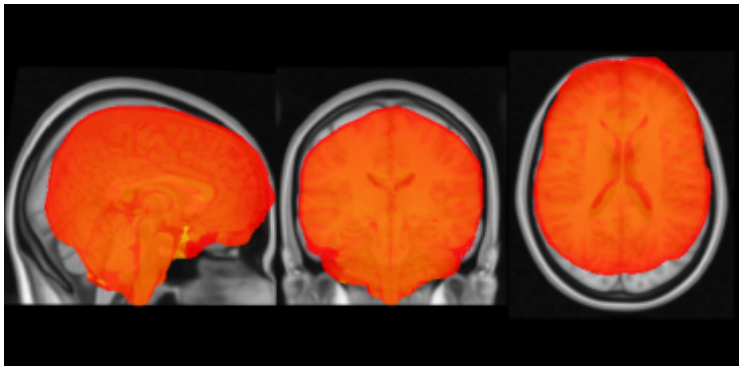


Figura 8.7: Imágen T1 del donante 2001 no registrada a espacio estándar

Para registrar la imagen a espacio estándar se ha hecho uso de la herramienta FLIRT (FMRIB's Linear Image Registration Tool) para registro lineal de imágenes [40]. En concreto, se han registrado las T1 de cada donante al template MNI 152 de un milímetro de resolución mediante un modelo de 12 parámetros.

Como resultado del proceso de registro, se obtiene una matriz de transformación de espacio sujeto a espacio estándar. Esta matriz ha sido posteriormente invertida para obtener la matriz de transformación desde espacio estándar a espacio sujeto.

Por medio de esta última matriz es posible registrar el atlas al espacio de cada donante individual. Para el registro se ha usado el mismo modelo de 12 parámetros con nearest neighbor como método de interpolación. La figura 8.8 muestra el resultado para el donante 2001. Las 360 regiones del Atlas se muestran con colores aleatorios.

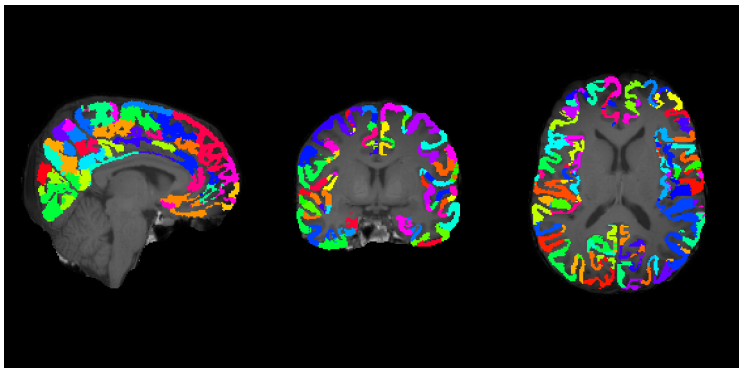


Figura 8.8: Atlas Glasser en espacio sujeto para el donante 2001

A partir de aquí, es posible pasar al mapeado de muestras a regiones. Dado que es posible que ciertas muestras no caigan directamente en una región el

proceso de asignación de muestras se ha realizado individualmente por cada donante mediante un algoritmo iterativo. En primer lugar, se comprueba si el vóxel asignado a la muestra pertenece a alguna región y si es así se le asigna dicha región. En el caso contrario, en cada posterior iteración se amplía en un vóxel por dimension el rango de búsqueda y se asigna la región con más vóxeles en dicho rango de búsqueda. Se define una distancia límite de asignación de 6 vóxeles a partir de la cual la muestra es descartada.

La figura 8.9 muestra el resultado del proceso de asignación. De las 360 regiones definidas por el atlas, 40 quedan sin ninguna muestra y son por tanto excluidas de posteriores análisis. La gran mayoría de regiones tienen asignadas un número reducido de muestras y sólo unas pocas contienen muchas muestras. Esto es probablemente debido a la alta resolución del Atlas, donde la mayoría de regiones son relativamente pequeñas.

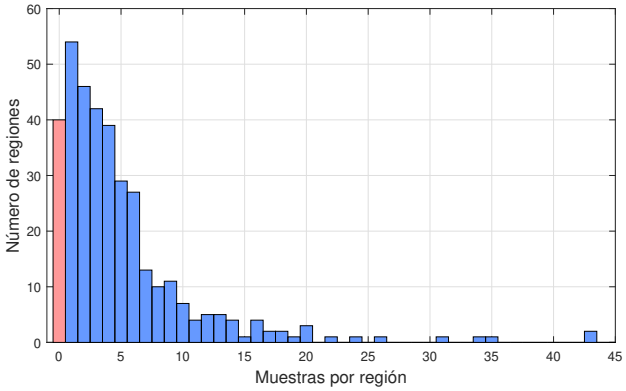


Figura 8.9: Número de regiones y muestras por región

Correlated Gene Expression

Una vez realizada la asignación de muestras, se ha calculado el valor de expresión genética de cada gen en cada región. Para ello, se han promediado los valores de expresión de todas las muestras asignadas a cada región de forma individual para cada donante. Finalmente, promediando entre donantes, se obtiene la matriz final de expresión genética de 8068 genes x 320 regiones.

A partir de los valores de expresión por región, la matriz de conectividad genética se ha obtenido mediante el coeficiente de correlación de Pearson (equación 8.5) entre el valor de expresión de todos los genes por cada par de regiones. De este modo, se obtiene una matriz como la que aparece en la figura 8.10, en la que se visualiza cómo de correlacionadas están dos regiones o muestras a nivel genético.

A simple vista se observan cómo existen regiones con perfil genético similar. También se aprecian dos líneas diagonales relacionadas con que la expresión genética tiende a ser similar entre regiones simétricas en ambos hemisferios.

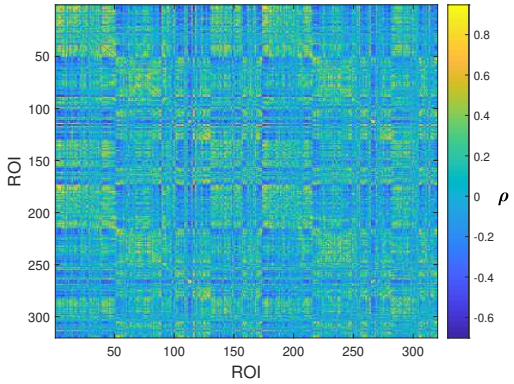


Figura 8.10: Matriz de conectividad genética para el atlas Glasser

8.2.3 Estructural

Para la obtención de valores de conectividad estructural es necesario hacer uso de los resultados de tractografía. Esta técnica permite, a partir de imágenes por tensor de difusión, reconstruir de forma virtual las fibras que recorren la materia blanca y conectan las diferentes regiones cerebrales. Superponiendo las regiones de un atlas sobre dichas fibras es posible obtener métricas de conectividad como el número de fibras o su longitud media.

En este proyecto se ha hecho uso de la matriz de conectividad empleada en [27]. La matriz fue construida a partir de 56 sujetos del proyecto HCP [29]. La métrica empleada para medir la conectividad entre dos regiones es el número de fibras que conectan dichas regiones dividido entre el volumen de ambas regiones. La matriz resultante se muestra en la figura 8.11, donde los valores están en escala logarítmica.

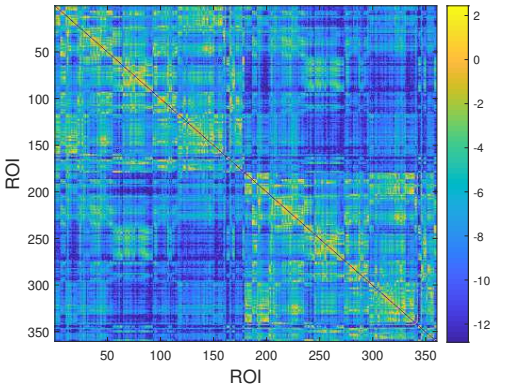


Figura 8.11: Matriz de conectividad estructural para el atlas Glasser

Observando la imagen, se aprecia cómo la conectividad intra-hemisferio es mayor a la inter-hemisferio. Además, aunque en menor medida que en la conectividad genética, también se aprecia la similaridad entre zonas simétricas

de ambos hemisferios. En la matriz de la figura aparecen las 360 regiones que componen el atlas Glasser completo, sin embargo para el análisis posterior sólo se emplearán las 320 para las que se tienen valores de expresión genética.

8.3 Análisis de relación Genómica Estructural

Tras haber construido las dos matrices de 320x320 referentes a la conectividad genética y estructural es posible plantear análisis que traten de relacionarlas directamente.

8.3.1 Conectado vs no conectado

El análisis más simple es comparar los valores de conectividad genética de las regiones conectadas estructuralmente con las que no (valores de conectividad nulos). La figura 8.12 muestra las funciones de densidad de probabilidad (FDP) de las dos distribuciones.

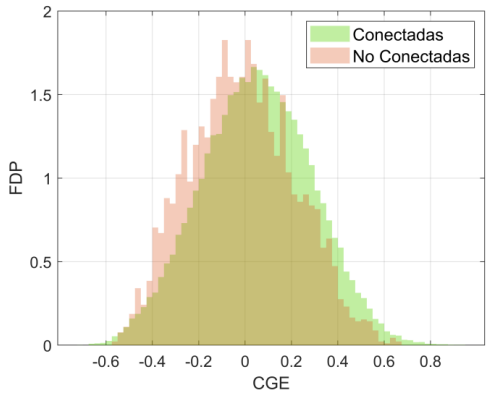


Figura 8.12: Valores CGE de zonas conectadas vs no conectadas

En promedio, las regiones conectadas estructuralmente muestran una mayor similitud genética. Dado que realmente, lo que tratamos de hacer es inferencia estadística, es necesario conocer la significancia estadística de las diferencias observadas. Dado el carácter gaussiano de ambas, es posible hacer uso de la prueba paramétrica t de Student [54]. Se trata de un test de contraste de hipótesis que nos permite conocer la probabilidad de que la hipótesis nula se cierta, es decir, de que ambas distribuciones pertenezcan a la misma distribución.

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_x^2}{N_x} + \frac{\sigma_y^2}{N_y}}} \tag{8.6}$$

El valor t de la ecuación 8.6 se obtiene a partir de las medias ($\bar{x}$, $\bar{y}$), varianzas

(σ_x^2 , σ_y^2) y número de muestras de ambas poblaciones (N_x, N_y). A partir del valor t es posible obtener la probabilidad de obtener las distribuciones bajo la hipótesis nula conocido como valor p.

Por otro lado, además de la significancia estadística, es importante conocer el tamaño del efecto observado, que permite cuantificar cómo de grandes son las diferencias observadas más allá de su significancia. En este caso, se ha hecho uso de la métrica conocida como d de Cohen [55], calculada a partir de la ecuación 8.7. Esta métrica toma valores entre 0 y 1 de forma proporcional a la magnitud del efecto observado.

$$d = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{(N_x-1)\sigma_x^2+(N_y-1)\sigma_y^2}{N_x+N_y-2}}} \tag{8.7}$$

Los resultados se muestran en la tabla 8.2 donde x e y son respectivamente los valores para los pares de regiones conectadas y no conectadas. Debido al elevado número de valores comparados los resultados presentan una alta significancia estadística ($p << 0.05$). Sin embargo, el tamaño del efecto puede considerarse relativamente pequeño ($d < 0.3$).

$\bar{x}$	$\bar{y}$	σ_x^2	σ_y^2	N_x	N_y	valor p	d de Cohen
0.051	-0.014	0.056	0.051	47404	3636	$5.11 \cdot 10^{-57}$	0.274

Tabla 8.2: Significancia estadística y tamaño del efecto

8.3.2 Correlación directa

En un segundo paso, es posible correlacionar directamente los valores de conectividad estructural y genética de cada par de regiones. El objetivo es ver si, dentro de las zonas conectadas aquellas con mayor valor de conectividad estructural tienen también mayor similitud genética. Esta medida se ha llevado a cabo mediante el coeficiente de correlación de Pearson. La figura 8.13 presenta un esquema visual de la correlación del valor de conectividad genético y estructural entre dos regiones.

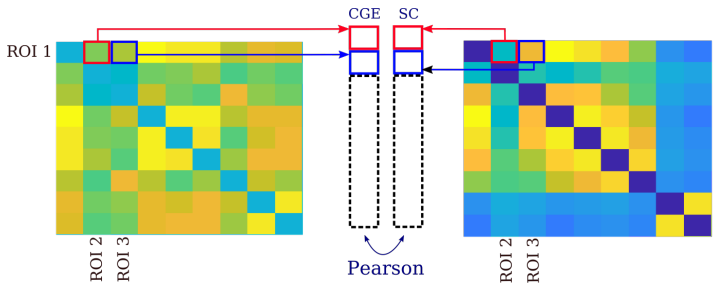


Figura 8.13: Ejemplo de correlación directa entre dos regiones

Resulta importante destacar que la operación se ha llevado a cabo sólo para los pares de regiones con valor de conectividad estructural no nulo. La figura 8.14 muestra los valores de conectividad estructural y genética para todos los pares de regiones. El coeficiente de correlación de Pearson es de 0.26 lo cual indica que existe una correlación pero no demasiado elevada.

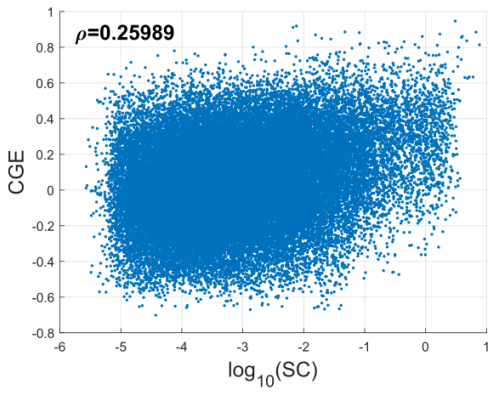


Figura 8.14: Correlación genética - estructural por cada par de regiones

8.3.3 Correlación modular

Otra forma sencilla de correlacionar la genética con la estructura se muestra en la figura 8.15. En este caso, utilizamos la correlación de Pearson para correlacionar la conectividad de cada región con todas las demás. Se obtiene por lo tanto un valor para cada región, indicando cómo de correlada está toda su conectividad.

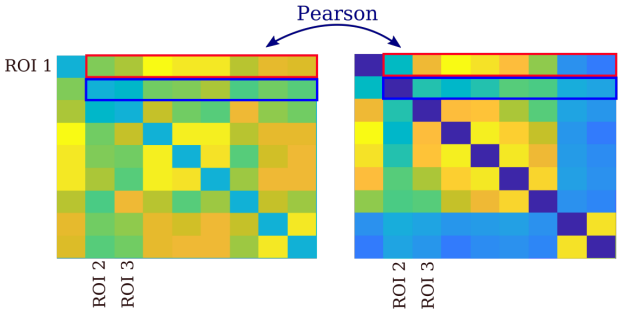


Figura 8.15: Ejemplo de correlación modular entre dos regiones

Con los valores obtenidos para cada una de las regiones se puede construir un histograma como el mostrado en la figura 8.16 donde se aprecia cómo la correlación media se sitúa claramente por encima de 0.

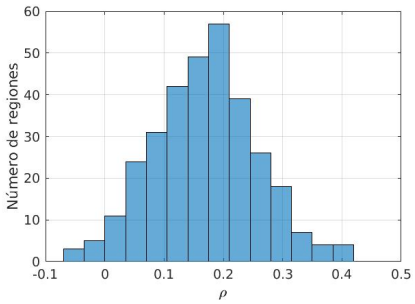


Figura 8.16: Histograma valores correlación modular

8.3.4 Corrección por distancia

Tal y como se mencionaba en el apartado 5.2, la expresión genética a nivel cerebral presenta una alta autocorrelación espacial, estando las zonas más cercanas más correlacionadas genéticamente. Este efecto se muestra en la figura 8.17, donde el scatterplot azul representa los valores de correlación genética (CGE) en función de la distancia euclídea entre cada par de regiones (d). La media de dichos valores, en color rojo, refleja claramente la existencia de un tendencia exponencial decreciente, dándose los mayores valores de correlación entre regiones muy cercanas entre sí.

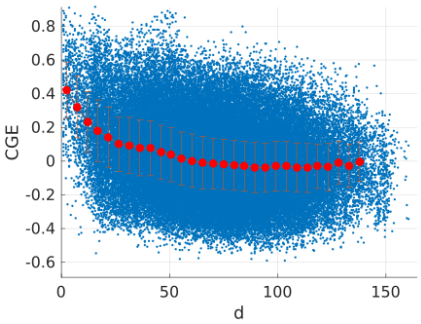


Figura 8.17: CGE en función de la distancia

Esta dependencia implica que los valores de CGE están en gran medida condicionados por la distancia entre regiones. Así, ciertas regiones podrían presentar altos valores de correlación genética sólo por estar cerca y no por estar conectadas estructuralmente. De este modo las correlaciones observadas en el apartado anterior podrían ser explicadas en base a la distancia entre regiones y no a una dependencia real entre genética y conectividad estructural.

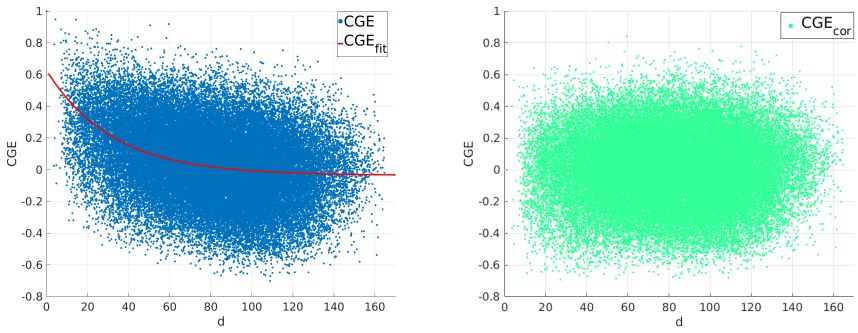
Una forma de hacer frente a esta problemática es mediante la corrección de los valores de CGE por distancia. El objetivo es ajustar los valores de CGE y distancia (d) mediante modelo exponencial de un término:

$$CGE_{fit} = ae^{-b\dot{d}} + c \tag{8.8}$$

donde a, b y c son los coeficientes a determinar que definen el modelo. El ajuste se ha realizado mediante el método de mínimos cuadrados no lineal. Se ha obtenido así la curva que se muestra en la figura 8.18a, donde $a = 0.662$, $b = 0.031$ y $c = -0.036$. Tras el ajuste se obtienen los valores CGE corregidos (CGE_{cor}) como la diferencia entre los valores originales y los obtenidos del modelo:

$$CGE_{cor} = CGE - CGE_{fit} \tag{8.9}$$

La figura 8.18b muestra los valores corregidos en los que la dependencia con la distancia se ha eliminado.



(a) CGE original y ajuste **(b)** CGE corregido

Figura 8.18: Corrección por distancia de CGE

Una vez corregidos los valores, es posible realizar los mismos análisis basados en correlaciones realizados anteriormente. El primera análisis se basaba en comparar los valores de CGE para zonas conectadas con los de aquellas zonas no conectadas. Se obtenían por tanto dos distribuciones sobre las que hacer contraste de hipótesis. Las distribuciones obtenidas tras corregir por distancia los valores de CGE se muestran en la figura 8.19.

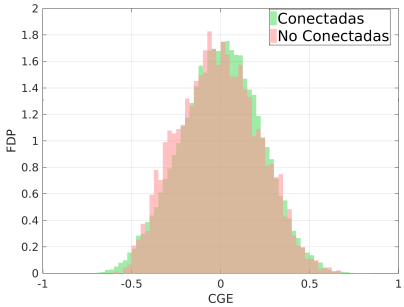


Figura 8.19: Regiones conectadas vs no conectadas tras corrección

La tabla 8.3 muestra los resultados del análisis estadístico. Si bien la prueba

t-student muestra que los resultados son estadísticamente significativos, el tamaño del efecto observado, medido en base a la d de Cohen, es muy pequeño ($d = 0.059$). Se observa por tanto un efecto mucho menor al observado sin la corrección ($d = 0.274$).

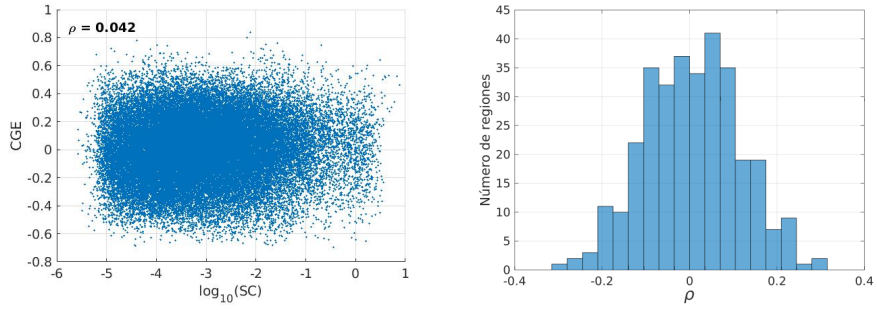
$\bar{x}$	$\bar{y}$	σ_x^2	σ_y^2	N_x	N_y	valor p	d de Cohen
0.001	-0.012	0.050	0.049	47404	3636	$5.89 \cdot 10^{-4}$	0.059

Tabla 8.3: Significancia estadística y tamaño del efecto

El segundo análisis encontraba que existía una correlación directa entre los valores de CGE y SC de $\rho = 0.26$. Sin embargo, la figura 8.20a muestra como, tras corregir por distancia los valores de CGE, la correlación observada previamente desaparece ($\rho = 0.04$).

Por otro lado, el resultado de repetir el análisis de correlación modular tras la corrección se muestra en la figura 8.20b. Se observa como en este caso la distribución obtenida está prácticamente centrada en 0 por lo que tampoco es posible determinar que la genética y la conectividad estructural estén correlacionadas modularmente.

Estos resultados apuntan a que la proximidad entre muestras es el factor explicativo de las correlaciones observadas inicialmente. Aun así, aunque la existencia de una relación exponencial decreciente entre CGE y distancia existe, la dispersión de los valores es muy grande y el modelo de corrección se limita a ajustarse a su media. Esto sugiere que podría ser un método demasiado estricto que además de corregir el efecto elimina posibles relaciones subyacentes entre la genética y la conectividad estructural.



(a) Correlación directa **(b)** Correlación modular

Figura 8.20: Correlaciones tras corrección

8.3.5 Cross-Modularity

Los métodos basados en correlaciones son un buen punto de partida por tratarse de un enfoque sencillo de interpretar e implementar. Sin embargo, es

interesante buscar métodos alternativos de mayor complejidad que traten el problema desde un ángulo distinto.

En esta línea, una opción es tratar de determinar hasta qué punto la genómica y la conectividad estructural comparten algún tipo de estructura modular, es decir, si las regiones cerebrales se agrupan en comunidades de forma similar tanto para la genómica como para la conectividad estructural.

La idea es hacer uso de técnicas de clustering para agrupar las regiones en diferentes módulos y relacionar la genómica y la conectividad estructural a partir de los módulos obtenidos.

El primer objetivo es encontrar el número de módulos en los cuales la conectividad estructural y la conectividad genómica derivada están más relacionadas. La figura 8.21 muestra un esquema del análisis. La idea es, a partir de la matriz de conectividad estructural, ir agrupando iterativamente las regiones cerebrales en un número creciente de módulos mediante **clustering jerárquico**. En cada iteración se obtiene una asignación de regiones a módulos que se emplea para reordenar también la matriz de conectividad genética.

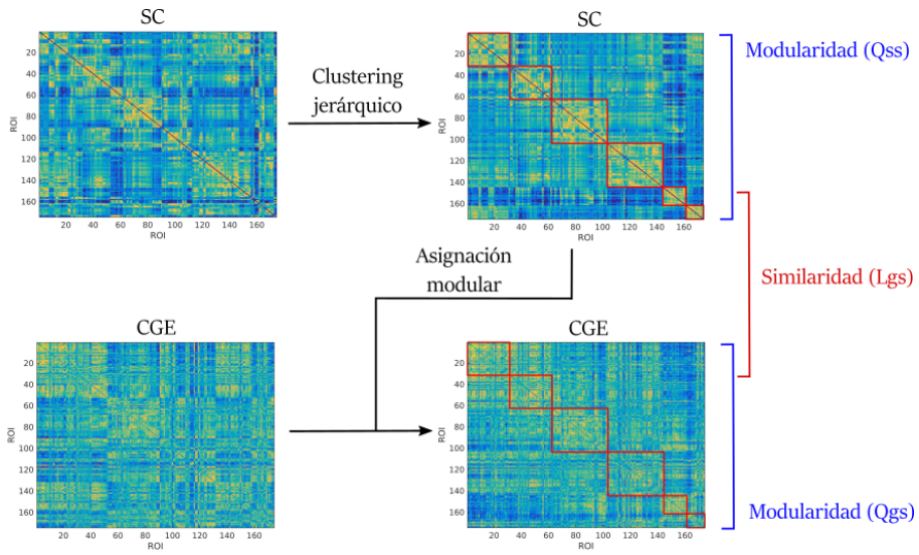


Figura 8.21: Esquema análisis cross-modularity

A partir de aquí se calcula en primer lugar la modularidad de la partición estructural obtenida (Q_{ss}) mediante la ecuación 5.2 descrita en el apartado 5.1.1. Esta métrica permite medir cómo de independientes son las comunidades obtenidas. Posteriormente, se reordena la matriz de conectividad genética en base a la partición estructural y se mide también su modularidad (Q_{gs}) . Este valor permite conocer en qué grado la genética se adapta a la partición estructural. Un valor alto indicaría que tanto la genética como la conectividad estructural forman comunidades muy similares de forma independiente. Finalmente, se umbralizan los valores de ambas matrices y se calcula la similitud módulo a módulo mediante el coeficiente Sørensen-Dice, que para dos sets de datos cualquiera X

e Y se obtiene como:

$$L = 2\frac{|X \cap Y|}{|X| \cup |Y|} \tag{8.10}$$

Los valores obtenidos se promedian para obtener la similitud promedio entre módulos (L_{gs}).

Para poder agrupar estos tres valores se hace uso de la métrica definida como **cross-modularity**, introducida y empleada con éxito en [56] para relacionar la conectividad estructural y funcional. Se calcula como:

$$X_{gs} = \left(Q_{ss}Q_{gs}L_{gs}\right)^{1/3} \tag{8.11}$$

El proceso descrito hasta ahora partía de la matriz de conectividad estructural. Sin embargo, es posible también realizar el proceso contrario, realizando clustering sobre la matriz genética y reordenando posteriormente la matriz de conectividad estructural en base a la partición obtenida. Se obtiene otro índice de cross-modularity (ecuación 8.12) donde Q_{gg} es la modularidad de la partición genética obtenida, Q_{sg} la modularidad de la matriz estructural para dicha partición genética y L_{sg} la similitud media de los módulos obtenidos.

$$X_{sg} = \left(Q_{gg}Q_{sg}L_{sg}\right)^{1/3} \tag{8.12}$$

El análisis se basa por tanto en calcular todas estas métricas para un número creciente de módulos y valorar los resultados. El punto para el cual las métricas de cross-modularidad presentan un máximo es aquel en el cual la estructura y la genética están más relacionadas. Este punto se obtiene buscando el máximo del promedio entre ambas métricas ($(X_{gs} + X_{sg})/2$).

El proceso de clustering consta de tres etapas:

1. **Construcción de matrices de disimilaridad:** las matrices de CGE y SC de las que se parte toman valores mayores cuando la relación entre dos regiones es mayor. Sin embargo, para la aplicación de las técnicas de clustering, es necesario modificar dichas matrices para obtener valores de distancia o disimilaridad que sean mayores cuando la diferencia entre regiones aumente. Para la matriz de SC se ha obtenido las distancia entre dos regiones i y j como:

   $$d_{ij} = 1 - \frac{SC_{ij}}{max(SC)} \tag{8.13}$$

   Para la matriz de CGE, tras probar diferentes combinaciones, se ha optado por emplear la distancia coseno entre los vectores de CGE de cada par de regiones (x_i y x_j) definida como:

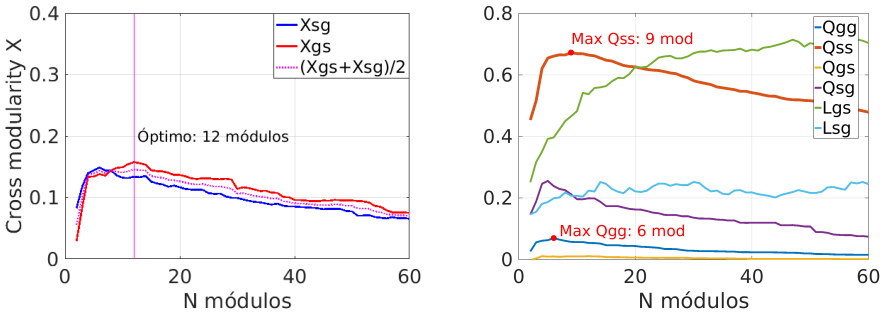
$$d_{ij} = 1 - \frac{x_i x'_j}{\sqrt{(x_i x'_i)(x_j x'_j)}} \tag{8.14}$$

2. **Clustering jerárquico:** Se construye un árbol de agrupamiento jerárquico a partir partir de las métricas de distancia obtenidas en el paso anterior. Partiendo de las 320 regiones, en cada nivel el algoritmo agrupa entre sí el par de regiones con menor distancia entre sí. Este proceso se repite de forma iterativa hasta llegar a agrupar todas las regiones en 2 clusters.

   Para este proceso existen diferentes métodos. Tras probar diferentes opciones se ha optado por emplear el método **Ward** para los valores de distancia genética y **Average** para los de conectividad estructural tal y como están implementados en Matlab [57].

3. **Dendrograma:** se visualiza el árbol generado en el paso previo y se extrae la partición del nivel del árbol que corresponda al número de comunidades de cada iteración.

Todo el proceso descrito hasta ahora se ha llevado a cabo para un número de módulos en el rango [2,60]. Los resultados obtenidos se muestran en la figura 8.22.



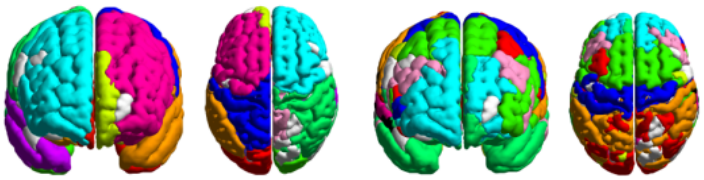
(a) Cross modularity **(b)** Correlación modular

Figura 8.22: Cross-modularity cerebro completo

En primer lugar, se observa cómo ambas métricas de cross-modularity siguen un patrón similar situándose el número óptimo de módulos en 12. Ése es por tanto el punto en el cual, según las métricas empleadas, la estructura modular genética y estructural del cerebro están más relacionadas. A pesar de que los valores no son demasiado altos, es un buen indicador para comprender, en cuanto al número de módulos, en qué orden de magnitud existe una mayor relación.

Respecto a las modularidades individuales, la figura 8.22b muestra cómo la modularidad estructural (Q_{ss}) es mucho mayor que la genética (Q_{gg}). Esto es derivado probablemente de que la conectividad genética, basada en correlaciones, presenta una red totalmente mallada entre sí. La conectividad estructural, sin embargo, sí presenta pares de regiones no conectadas para los cuales su conectividad es nula, aumentando así el índice de modularidad.

Es interesante comparar visualmente las estructuras modulares obtenidas partiendo tanto desde la genética como desde la conectividad estructural. La figura 8.23 muestra la partición en 12 módulos de ambas matrices. Se usan colores aleatorios para diferenciar los diferentes clusters encontrados. Las regiones de color blanco son aquellas para las que no existían muestras genéticas y no han sido incluidas en el estudio.



(a) Conectividad estructural **(b)** Genética

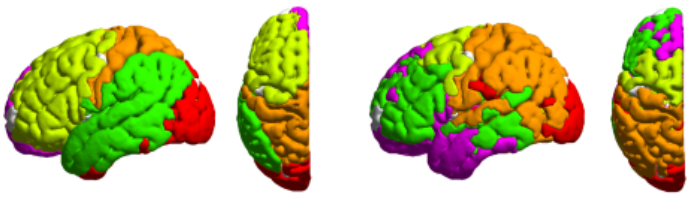
Figura 8.23: Partición en 12 módulos del cerebro completo

Se observa con claridad cómo los módulos obtenidos a partir de la conectividad estructural se agrupan de forma ordenada en un sólo hemisferio, no existiendo módulos que contengan regiones de ambos. Una posible explicación es la dificultad de las técnicas de ITD para captar aquellas fibras que conectan ambos hemisferios. Dichas fibras siguen un recorrido complejo de giros entrelazadonse con otras fibras siendo más difíciles de reconstruir mediante tractografía. La genética, sin embargo, muestra módulos menos definidos y que sí abarcan ambos hemisferios. Los módulos se generan de forma relativamente lateralmente simétrica desde el lóbulo frontal hacia el occipital. Este patrón cuadra con el conocido gradiente de variación en la citoarquitectura cerebral, en base al cual las composición celular del cerebro varía progresivamente siguiendo la dirección frontal a occipital. Esto hace que aquellas zonas más alejadas presenten menor similitud celular y por tanto expresión genética diferenciada [6].

Vista la gran diferencia existente en la generación de comunidades entre genética y estructura ,y tratando de simplificar el problema, se ha realizado el mismo análisis únicamente para el hemisferio izquierdo. Éste es es el hemisferio para el que se tienen más muestras genéticas por lo que se pasa a trabajar con 174 de las 180 regiones definidas para dicho hemisferio en el Atlas Glasser.

Tras repetir el análisis, el número óptimo de módulos obtenido es de 5, lo cual entra dentro de lo esperado teniendo en cuenta que el valor óptimo para el cerebro completo era de 12. Los 5 módulos se muestran en la figura 8.24. Tal y como cabía esperar cuando se reduce el estudio a un sólo hemisferio los módulos resultan mucho más similares entre sí. Aun así, de forma similar al caso anterior, los módulos estructurales resultan ser algo más compactos que los genéticos.

Hasta ahora se ha desarrollado el análisis de forma global sobre todos los módulos. No obstante, es posible descomponer los valores de similaridad promedio (L_{gs} y L_{sg}) en los respectivos valores de similaridad módulo a módulo. El objetivo es identificar aquellos módulos que presentan mayor similitud. El análisis se ha centrado en el valor promedio L_{gs}. La figura 8.25 muestra los valores de



(a) Conectividad estructural **(b)** Genética

Figura 8.24: Partición en 5 módulos del hemisferio izquierdo

similaridad tanto módulo a módulo como promedio a medida que aumentamos el número de módulos. El tamaño de los puntos es proporcional al número de regiones involucradas en cada módulo.

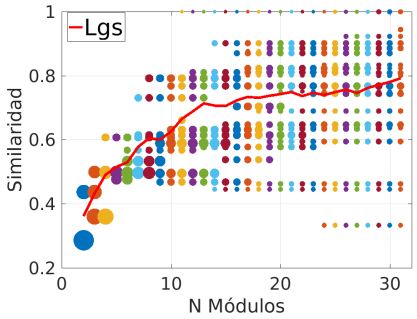


Figura 8.25: Similaridad módulo a módulo en el hemisferio izquierdo

Se observa cómo, a medida que se aumenta el número de comunidades los módulos se van subdividiendo en módulos más pequeños. También se aprecia que existen diferencias significativas en los valores de similaridad de cada uno de los módulos que sugieren que existen diferencias significativas entre regiones. Se ha decidido escoger la partición en 20 módulos y considerar aquellos con valores de similaridad superiores a 0.6 como módulos altamente similares. La figura 8.26 muestra dichos módulos con colores aleatorios. Visualmente se observa que se trata de módulos poco distribuidos y con una aparente predominancia en los lóbulos occipital y parietal.

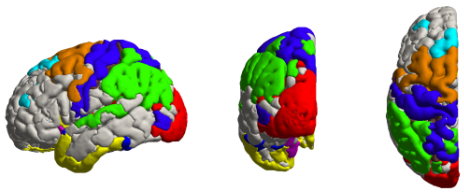


Figura 8.26: Módulos altamente similares

8.3.6 Procesado de señal sobre grafos

La idea de emplear el procesado de señal sobre grafos surge de los diversos trabajos recientes que ya emplean un enfoque de este estilo para relacionar la conectividad estructural y la funcional. En todos ellos se construye un grafo basado en la conectividad estructural y se consideran los valores de activación funcional como señales que toman valores sobre ese grafo. El problema de este proyecto, aunque distinto, puede también abordarse desde una metodología parecida. La idea fundamental es que la expresión genética de cada gen puede considerarse como una señal sobre el grafo de conectividad estructural. Tenemos por tanto un grafo y 8068 señales que toman valores sobre él. A partir de aquí, es posible obtener una representación espectral de cada una de las señales y aplicar distintas técnicas de procesado en dominio espectral.

El primer paso es obtener los modos base del grafo estructural para poder posteriormente calcular la Transformada de Fourier sobre Grafos de cada una de las señales. Partiendo de la matriz de SC y siguiendo el procedimiento descrito en [27] se obtienen 320 modos. La figura 8.27 muestra una selección de los modos.

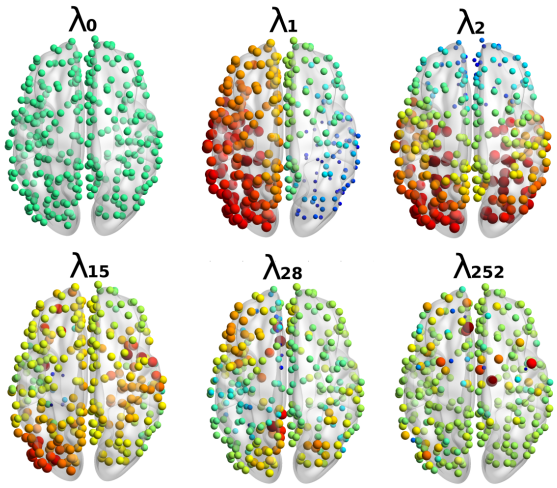


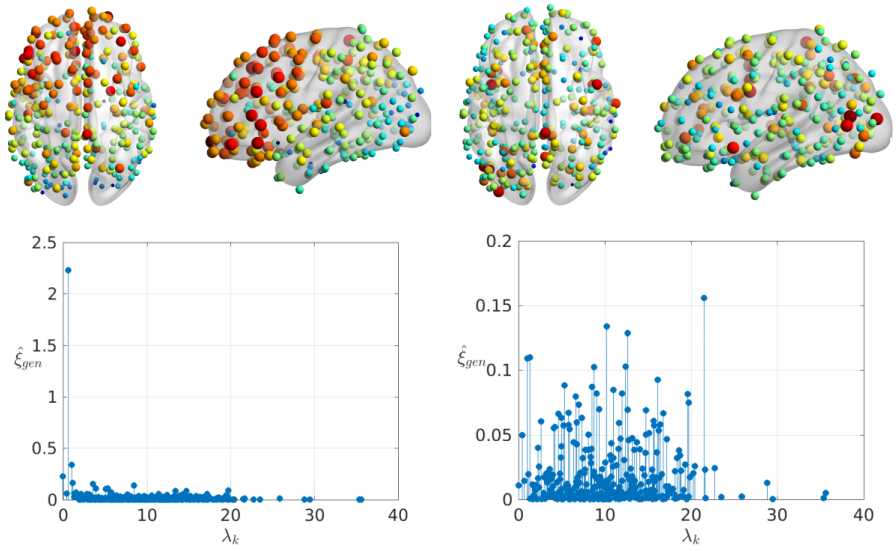
Figura 8.27: Modos del cerebro humano

El primer modo representa la componente continua. De forma muy visual, los modos de menor frecuencia siguen patrones de variación lenta mientras que, a medida que subimos en frecuencia, los valores cambian más rápidamente de un nodo a otro. Intuitivamente, una respuesta frecuencial de baja frecuencia indica un mayor acoplo de la señal con el grafo, es decir, que la señal varía de forma progresiva a lo largo de los nodos.

Partiendo de estos modos se calcula la transformada de Fourier de cada una de las señales de expresión genética de los 8068 genes. A partir transformada de Fourier de un gen n cualquiera ($\hat{s}_{gen_n}(k)$) es posible obtener su densidad espectral de potencia (DEP) como:

$$\hat{\xi}_{gen_n}(k) = |\hat{s}_{gen_n}(k)|^2 \tag{8.15}$$

Se obtiene así una representación frecuencial de la expresión genética de cada gen a lo largo de la corteza cerebral. Estas representaciones permiten analizar la expresión genética de cada gen desde un nuevo punto de vista. Como ejemplo, la figura 8.28 muestra la expresión genética y la DEP de dos genes distintos. El primero de ellos es un gen cuya expresión parece seguir un patrón desde el lóbulo frontal al occipital, probablemente debido a que la expresión de este gen está relacionada con determinado tipo de células más presentes en el lóbulo frontal. La representación espectral capta el patrón de forma muy clara, estando la gran mayoría de la potencia concentrada en el modo λ_2 mostrado en la figura anterior. Este modo sigue un patrón de variación lenta que se asemeja mucho con el del gen, captando lógicamente la mayoría de la potencia. El segundo gen, por el contrario, presenta un patrón de expresión muy poco localizado que espectralmente se refleja en una DEP con la potencia distribuida a lo largo de numerosos modos. Esto indica que la transformada de Fourier es una buena herramienta para comprender los patrones de expresión de los diferentes genes.



(a) Gen A **(b)** Gen B

Figura 8.28: Representación espacial y espectral de dos genes

Visto lo distintos que pueden ser los patrones de expresión de cada uno de los genes resulta interesante averiguar si existe algún tipo de patrón mayoritario. Se calcula por lo tanto la DEP promedio a lo largo de los 8068 genes analizados (ver ecuación 8.16). El resultado se muestra en la figura 8.29. Los valores de expresión genética han sido normalizados respecto a zero por lo que no presentan componente continua. Por este motivo y para poder visualizar los datos en escala logarítmica en la figura se omite el autovalor correspondiente a $\lambda_0 = 0$.

$$\hat{\xi}(k) = \frac{1}{8068} \sum_{n=1}^{8068} \hat{\xi}_{gen_n}(k) \tag{8.16}$$

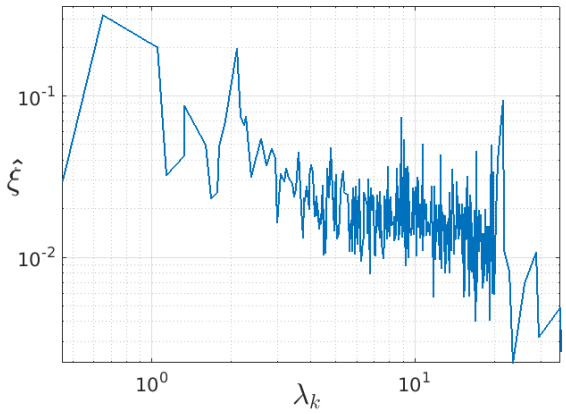


Figura 8.29: DEP media de los genes

Se observa cómo la distribución de potencia no es uniforme y existe una tendencia decreciente mediante la cual las componentes de baja frecuencia resultan más importantes que las de alta frecuencia. Concretamente los modos λ_2 y λ_3 son los que mayor potencia capturan. A partir de esta representación espectral puede interpretarse que la expresión genética se adapta relativamente bien al grafo estructural ya que la existencia de fuertes componentes de baja frecuencia indica que muchos genes no varían de forma brusca de un nodo a otro.

Es interesante comparar estos resultados con los obtenidos en [27] para relacionar la conectividad estructural y la funcional. Los autores de ese estudio llegan a una gráfica similar que muestra cómo la activación funcional está fuertemente ligada a la estructura. La distribución obtenida en ese caso muestra una diferencia entre bandas de baja y alta frecuencia aún más acusada que la obtenida en este proyecto. A pesar de tratarse de conceptos diferentes, esta comparación es un buen indicador de que la expresión genética, al menos en promedio, se adapta al grafo estructural pero no de una forma total.

Finalmente, replicando las ideas de [27], es interesante comparar las regiones cerebrales en cuanto a distribución espectral de potencia. Se busca conocer qué regiones muestran un perfil de expresión genética más homogéneo, es decir, aquellas en las que la mayoría de los genes varían de forma gradual entre esa región y las adyacentes.

El primer paso es separar la señal de cada gen n en dos señales de baja y alta frecuencia (s_{low_n} y s_{high_n}). Esto se lleva a cabo mediante filtrado espectral y posterior transformada inversa de Fourier. El filtrado se ha realizado de forma simple escogiendo los primeros 60 valores del espectro para la señal de baja frecuencia y los siguientes para la de alta.

A partir de estas señales se puede definir un índice de acoplamiento por cada región como:

$$\Delta(k) = \frac{\sum_{n=1}^{8068} s^2_{low_n}}{\sum_{n=1}^{8068} s^2_{high_n}} \tag{8.17}$$

Para cada región k, el índice compara el ratio entre la suma de la potencia de la señal de baja frecuencia respecto a la de alta. La idea es que, aquellas regiones donde la expresión genética es más homogénea tendrán un índice superior.

Tras calcular el índice, es posible graficarlo como en la figura 8.30 y analizar los resultados. Según escala de color empleada el rojo representa aquellas regiones con alto índice de acoplamiento y el azul aquellas más desacopladas. El tamaño de los nodos es proporcional al módulo del índice de acoplamiento tras pasarlo a escala logarítmica.

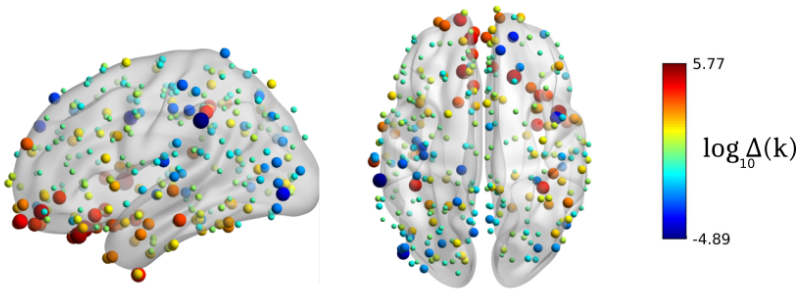


Figura 8.30: Índice de acoplo

El patrón emergente dista bastante de ser aleatorio. Aparentemente, el lóbulo frontal muestra un índice de acoplo significativamente mayor al de la parte occipital. Esto puede ser debido a que un alto porcentaje de los genes estudiados se expresan de forma más homogénea en estas regiones.

8.4 Resumen de resultados

Dada la diversidad de métodos empleados, a continuación se detallan brevemente los principales resultados derivados de los análisis realizados:

1. Las zonas cerebrales muy diferenciadas (cerebelo, cortex, estructuras subcorticales y tronco cerebral) muestran perfiles genéticos muy distintos.
2. Los primeros análisis basados en correlaciones muestran la existencia de una correlación entre expresión genética y conectividad estructural.
3. Esta correlación desaparece si se corrigen los resultados teniendo en cuenta la distancia entre regiones.

4. Las regiones se agrupan en comunidades de forma distinta para la conectividad estructural y genética. En el caso de la conectividad estructural, los módulos se generan preferentemente en un único hemisferio. La genética, sin embargo, deriva en módulos que incluyen ambos hemisferios recorriendo el cortex desde el lóbulo frontal al occipital.

5. El número de módulos que maximiza la relación estructura-genética es de aproximadamente 12 para el cerebro completo y 5 para un solo hemisferio.

6. No todos los módulos son igual de similares en cuanto a conectividad estructural y genética. Los de mayor similaridad son módulos poco distribuidos y aparentemente situados en los lóbulos parietal y occipital.

7. Se plantea la posibilidad de emplear el procesado de señal sobre grafos como un nuevo enfoque. Se demuestra que la representación espectral de los genes es una herramienta válida para clasificar y comprender sus patrones de expresión.

8. La expresión genética, en promedio, presenta un espectro de baja frecuencia que se adapta relativamente bien al grafo estructural. Los modos predominantes del espectro se corresponden con gradientes de variación frontal-occipital.

9. La expresión genética promedio parece ser más homogénea en las regiones del lóbulo frontal.

9 Metodología

En el presente apartado se describe la metodología llevada a cabo para la realización del proyecto. En primer lugar, se detallan los recursos humanos y materiales involucrados en el proyecto. Posteriormente, se describen las diferentes fases del proyecto, divididas en paquetes de trabajo. Se detallan a su vez las fechas de inicio y final, recursos empleados y entregables de cada uno de los paquetes. Finalmente, para tener una visión global del proyecto, se listan los diferentes hitos y entregables y se muestran junto con las diferentes paquetes de trabajo en un Diagrama de Gantt.

9.1 Recursos humanos

En la tabla 9.1 se listan los integrantes del equipo de trabajo del proyecto, indicando tanto su puesto como el rol que han tenido en el proyecto.

Código	Puesto	Nombre-Apellidos	Rol
RH1	Ingeniero junior	David Romero Bascones	Autor del proyecto
RH2	Ingeniero senior	Unai Irusta Zarandona	Codirector del proyecto
RH3	Investigador senior	Jesús Cortes Díaz	Codirector del proyecto

Tabla 9.1: Equipo de trabajo

9.2 Recursos materiales

Los recursos hardware y software empleados en el proyecto se detallan en la tabla 9.2. Los códigos de identificación comienzan por la letra H o S dependiendo de si se trata de recursos hardware o software.

Código	Recurso	Uso
H1	Ordenador de mesa HP	Trabajo en Biocruces
H2	Ordenador portátil HP	Trabajo en remoto
H3	Cluster de computación	Simulaciones exigentes
H4	Impresora	Impresión de bibliografía
S1	Matlab 2018b incluyendo *toolboxes*	Análisis y procesado general
S2	FSL	Procesado de neuroimagen
S3	DSI Studio	Tractografía y cálculo de conectividad
S4	LaTeX	Redacción del proyecto

Tabla 9.2: Recursos hardware (H) y software (S)

9.3 Paquetes de trabajo

A continuación se detallan cada uno de los paquetes de trabajo en los que se ha dividido el proyecto.

PT0. Gestión y supervisión

Reuniones periódicas entre el autor y los supervisores para analizar el transcurso del proyecto.

Duración: todo el proyecto
Recursos humanos: RH1, RH2 y RH3 (50 horas cada uno)
Recursos materiales: H2 (50 horas)

PT1. Preparación del proyecto

Establecer los objetivos y la planificación básica del proyecto. Además, realizar los trámites administrativos necesarios: formalización del convenio de cooperación educativa y solicitud de la tarjeta y permisos de acceso a Biocruces.

Fecha de inicio: 2019-02-20
Fecha final: 2019-03-01
Recursos humanos: RH1 (20 horas), RH2 (6 horas) y RH3 (20 horas)
Recursos materiales: H2 (10 horas)

PT2. Estado del arte y familiarización con las herramientas

Adquisición de conocimientos básicos relacionados con genética y neurociencia. Revisión bibliográfica del estado del arte referente a la conectividad genética y estructural del cerebro humano. Familiarización con las herramientas software a emplear en el proyecto.

Fecha inicio: 2019-03-04
Fecha final: 2019-03-27
Recursos humanos: RH1 (108 horas)
Recursos materiales: H1 (100 horas), H3 (20 horas), H4, S1 (40 horas), S2 y S3

PT3. Procesado de datos genéticos

Descarga y preprocesado de la base de datos de expresión genética mediane la implementación de un pipeline completo.

Fecha de inicio: 2019-03-28
Fecha final: 2019-04-17
Recursos humanos: RH1 (90 horas)
Recursos materiales: H1 (90 horas), S1 (90 horas) y S2
Entregable: Datos genéticos preprocesados

PT4. Cálculo de matrices de conectividad

Obtención de matriz de conectividad genética en base a los datos preprocesados anteriormente. Obtención de matriz de conectividad estructural.

Fecha de inicio: 2019-04-18
Fecha final: 2019-05-15
Recursos humanos: RH1 (90 horas)
Recursos materiales: H1 (90 horas), H3 (20 horas), S1 (60 horas), S2 y S3
Entregable: Matrices de conectividad

PT5. Análisis 1 (Correlaciones)

Análisis estadístico basado en correlaciones para tratar de relacionar la conectividad genética y estructural.

Fecha de inicio: 2019-05-16
Fecha final: 2019-05-28
Recursos humanos: RH1 (60 horas)
Recursos materiales: H1 (60 horas) y S1 (60 horas)
Entregable: Resultados del análisis 1

PT6. Análisis 2 (Cross-modularity)

Comprensión de la metodología de análisis basad

Fecha de inicio: 2019-05-29
Fecha final: 2019-06-14
Recursos humanos: RH1 (90 horas)
Recursos materiales: H1 (90 horas), H3 (20 horas), S1 (90 horas) y S3 (60 horas)
Entregable: Resultados del análisis 2

PT7. Análisis 3 (Grafos)

Comprensión de la teoría matemática del procesado de señal sobre grafos. Adaptación de los datos a la metodología. Diseño y ejecución de análisis.

Fecha de inicio: 2019-06-17
Fecha final: 2019-07-05
Recursos humanos: RH1 (90 horas)
Recursos materiales: H1 (90 horas), S1 (90 horas) y S3 (60 horas)
Entregable: Resultados del análisis 3

PT8. Documentación y redacción

Preparación de la memoria y presentación del trabajo.

Fecha de inicio: 2019-07-08
Fecha final: 2019-09-15
Recursos humanos: RH1 (150 horas) y RH2 (60 horas)
Recursos materiales: H2 (150 horas), S2 (150 horas), S6 (120 horas)
Entregable: Memoria

9.4 Hitos y entregables

Los hitos representan puntos de especial relevancia en el transcurso del proyecto y marcan la correcta finalización de las fases más importantes del mismo (ver tabla 9.3).

Los entregables, listados en la tabla 9.4, son elementos tangibles obtenidos como resultado de los paquetes de trabajo más relevantes.

Código	Descripción	Fecha
M1	Comienzo del proyecto	2019/02/20
M2	Datos genéticos preprocesados	2019/04/17
M3	Datos preparados para análisis	2019/05/15
M5	Análisis 1 finalizado	2019/05/28
M5	Análisis 2 finalizado	2019/06/14
M5	Análisis 3 finalizado	2019/07/05
M6	Fin del proyecto	2019/10/08

Tabla 9.3: Hitos

Código	Descripcióin	Fecha
E1	Datos genéticos preprocesados	2019/04/17
E2	Matrices de conectividad	2019/05/15
E3	Resultados análisis 1	2019/05/28
E4	Resultados análisis 2	2019/06/14
E5	Resultados análisis 3	2019/07/05
E6	Memoria	2019/09/15

Tabla 9.4: Entregables

9.5 Diagrama de Gantt

La figura 9.1 muestra un diagrama de Gantt con los paquetes de trabajo e hitos del proyecto.

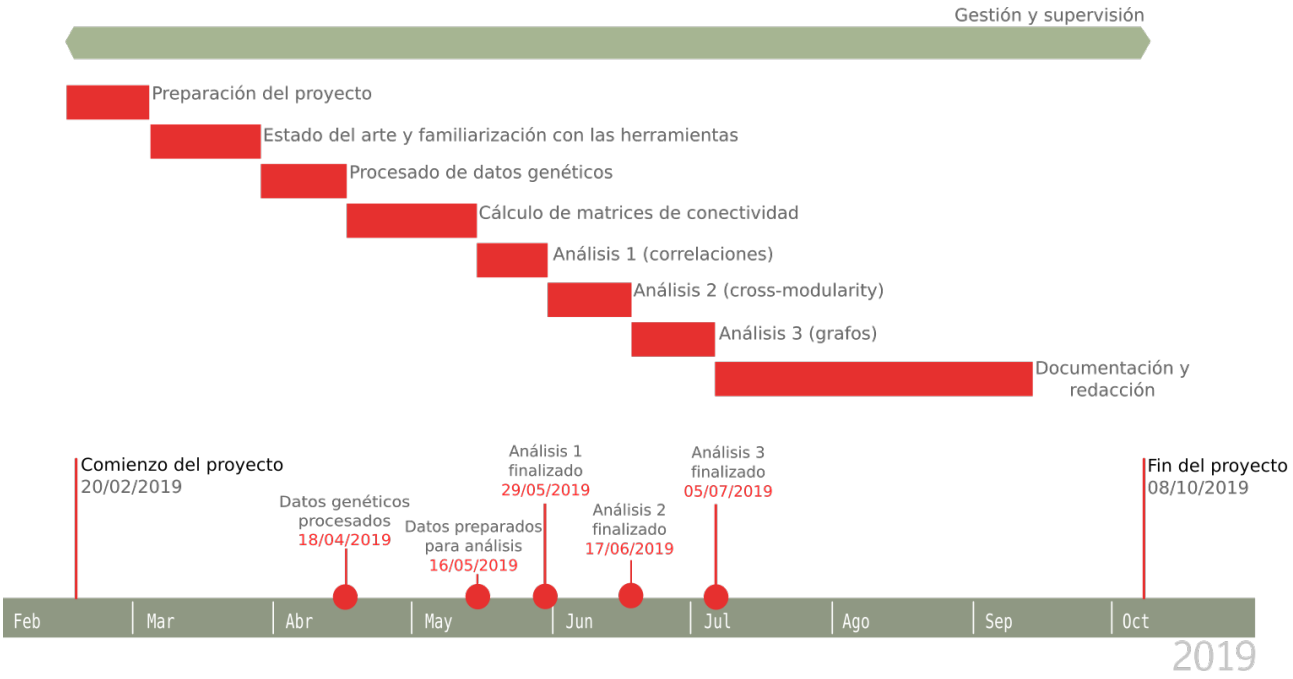


Figura 9.1: Diagrama de Gantt

10 Descargo de gastos

En este apartado se detallan los costes del proyecto. Se han dividido entre las siguientes partidas de gasto: horas internas, amortizaciones, subcontrataciones y gastos. Finalmente se calcula el cose total del proyecto como la suma de todas las partidas de gasto.

- **Horas internas**

  Hacen referencia a los recursos humanos del proyecto. El coste de cada uno de los integrantes del proyecto se obtiene como el producto entre el coste por hora y el número de horas de trabajo. Los resultados se muestran en la tabla 10.1.

Código	Coste (€/h)	Número de horas	Coste(€)
RH1	30	748	22440
RH2	60	116	6960
RH3	60	70	4200
		Subtotal	33600 €

Tabla 10.1: Gastos horas internas

- **Amortizaciones**

  Se trata de los recursos materiales empleados que no están destinados únicamente para este proyecto, si no que tienen una vida útil mayor y por lo tanto, el coste asignado al proyecto debe ajustarse en función de las horas de uso. El coste se calcula como el producto entre el coste inicial y el porcentaje de la vida útil empleada en el proyecto:

$$\text{Coste} = \text{Coste inicial} \cdot \frac{\text{Número de horas}}{\text{Vida Útil}} \tag{10.1}$$

  La tabla 10.2 muestra los diferentes recursos junto a sus gastos de amortización. Los recursos relativos a software libre (S2, S3 y S4) no han sido incluidos dado que su coste es nulo.

Código	Coste inicial (€)	Vida útil (h)	Número de horas	Coste (€)
H1	1500	4500	520	173
H2	700	3500	210	42
S1	2500	2000	340	425
			Subtotal	640 €

Tabla 10.2: Amortizaciones

- **Subcontrataciones**

  En el proyecto no se han llevado a cabo subcontrataciones por lo que el gasto de esta partida ha sido nulo.

- **Gastos**

  Son aquellos gastos que deben ser atribuidos en su totalidad al proyecto. En esta partida se encuentran los gastos de oficina estimados a lo largo del proyecto (luz, limpieza, ...), donde se incluyen los gastos de la impresora. Además, se consideran aquí los gastos de transporte para el desplazamiento diario a Biocruces. Los detalles de esta partida se muestran en la tabla 10.3.

Recurso	Coste(€)
Gastos de oficina	150
Transporte	150
Subtotal	300 €

Tabla 10.3: Gastos

- **Resumen de costes**

  La tabla 10.4 muestra el resumen de las distintas partidas de gastos. Teniendo en cuenta todas las partes del presupuesto, los costes totales para el desarrollo del proyecto ascienden a un total de **treinta y cuatro mil quinientos cuarenta euros**.

Resumen de costes	
Recursos humanos	33600 €
Amortizaciones	640 €
Subcontrataciones	0 €
Gastos	300 €
Total	34540€

Tabla 10.4: Resumen de costes

11 Conclusiones

A lo largo del presente proyecto de investigación se han realizado diferentes análisis con el objetivo de relacionar la expresión genética con la forma en la cual regiones cerebrales están conectadas.

En primer lugar, se ha procesado y adecuado para los análisis la base de datos de expresión genética más extensa que existe, el AHBA. Los resultados de este trabajo son la base de éste y otros futuros proyectos del grupo de investigación. En una primera visualización de los datos se han reproducido importantes resultados previos en los que se muestra que las regiones anatómicas muy diferenciadas (cerebelo, corteza cerebral y estructuras subcorticales) presentan perfiles genéticos muy distintos.

Centrando el estudio en la corteza cerebral, se ha realizado un análisis estadístico basado en correlaciones donde se han obtenido correlaciones no muy elevadas entre genética y conectividad estructural. Sin embargo, el hecho de que regiones cercanas presentan una mayor similitud genética sólo por tener una composición celular similar podría distorsionar los resultados y se han corregido los valores teniendo en cuenta la distancia entre regiones. Si bien tras la corrección las correlaciones se reducen mucho, el modelo empleado para la corrección no presenta un ajuste perfecto a los datos por lo que cabe plantearse si no es un método de corrección demasiado estricto. En futuros estudios sería interesante emplear otras alternativas para la corrección por distancia como son el uso de modelos nulos.

Posteriormente, se ha relacionado la genética y la conectividad a nivel modular mediante algoritmos de agrupamiento jerárquico y medidas de modularidad. Se ha observado cómo la genética tiende a agrupar las regiones en sentido frontal-occipital mientras que la conectividad forma módulos dentro de cada hemisferio individualmente. La subdivisión del cerebro en 12 módulos maximiza la relación entre conectividad estructural y genética. Estos módulos a su vez presentan valores de similitud más elevados en regiones parietal y occipital. No obstante, se ha observado que la metodología empleada es muy sensible a variaciones en la configuración de parámetros por lo que deberían realizarse más simulaciones y estudiar con mayor profundidad la consistencia de los resultados.

Finalmente, se han aplicado técnicas procesado de señal sobre grafos no antes empleadas en el ámbito de la genética. Concretamente, se ha comprobado que la

transformada de Fourier sobre grafos es una herramienta válida para entender los patrones de expresión genética. Los resultados demuestran que una gran mayoría de genes siguen un patrón de expresión frontal-occipital correspondiente con el conocido gradiente de variación en la composición celular cortical. A partir de aquí, se ha encontrando que regiones cercanas al lóbulo frontal presentan una expresión genética más homogénea. De todas formas, haría falta más trabajo para comprender las posibles razones que podrían explicar este último punto. El trabajo por lo tanto abre una prometedora línea que cabe esperar sea seguida en posteriores investigaciones.

Fuentes de Información

[1] [Online]. Available: https://commons.wikimedia.org/wiki/File:MRI_brain.jpg

[2] [Online]. Available: https://scanexpert.ro/wp-content/uploads/2017/07/tractografia-rmn-rezonanta-magnetica-nucleara-brasov-7-300x300.jpg

[3] [Online]. Available: https://commons.wikimedia.org/wiki/File:Difference_DNA_RNA-ES.svg

[4] "Partial map of the internet based on the january 15, 2005 data found on opte.org."

[5] M. Pievani, N. Filippini, M. P. van den Heuvel, S. F. Cappa, and G. B. Frisoni, "Brain connectivity in neurodegenerative diseases—from phenotype to proteinopathy," *Nature Reviews Neurology*, vol. 10, no. 11, pp. 620–633, oct 2014.

[6] A. Fornito, A. Arnatkevičiūtė, and B. D. Fulcher, "Bridging the gap between transcriptome and connectome," *Trends Cognit. Sci.*, vol. 23, pp. 34–50, 2019.

[7] N. Jahanshad *et al.*, "Genome-wide scan of healthy human connectome discovers SPON1 gene variant influencing dementia severity," *Proceedings of the National Academy of Sciences*, vol. 110, no. 12, pp. 4768–4773, mar 2013.

[8] J. D. Rudie *et al.*, "Autism-associated promoter variant in MET impacts functional and structural brain networks," *Neuron*, vol. 75, no. 5, pp. 904–915, sep 2012.

[9] "Scale of the human brain," 2015. [Online]. Available: https://aiimpacts.org/scale-of-the-human-brain/

[10] E. R. Kandel, J. H. Schwartz, T. M. Jessell, D. of Biochemistry, M. B. T. Jessell, S. Siegelbaum, and A. Hudspeth, *Principles of neural science*. McGraw-hill New York, 2000, vol. 4.

[11] I. P. Pérez, *El mapa del cerebro: un paseo anatómico por la máquina de pensar*, ser. Neurociencia y Psicología, E. E. País, Ed. EMSE EDAPP S.L., 2018.

[12] "Fsl course. oxford centre for functional mri of the brain." [Online]. Available: https://fsl.fmrib.ox.ac.uk/fslcourse/lectures/intro.pdf

[13] J. Lafuente Sánchez, "Daño axonal difuso: Importancia de su diagnóstico en neuropatología forense," *Cuadernos de Medicina Forense*, no. 41, pp. 173–182, 2005.

[14] M. Daianu, N. Jahanshad, T. M. Nir, A. W. Toga, C. R. Jack, M. W. Weiner, and f. t. A. D. Paul M. Thompson, "Breakdown of brain connectivity between normal aging and alzheimer's disease: A structural k-core network analysis," *Brain Connectivity*, vol. 3, no. 4, pp. 407–422, aug 2013.

[15] A. M. Araguz, C. B. Martínez, M. T. E. Mansour, and J. M. M. Martínez, "Neurociencia en el egipto faraónico y en la escuela de alejandría," *Revista de Neurología*, vol. 34, no. 12, p. 1183, 2002.

[16] "Neuroscience in pictures: the best images of the year." [Online]. Available: http://theconversation.com/neuroscience-in-pictures-the-best-images-of-the-year-89077

[17] H. Y. Carr, "Free Precession Techniques in Nuclear Magnetic Resonance." Ph.D. dissertation, HARVARD UNIVERSITY., 1953.

[18] M. Symms, "A review of structural magnetic resonance neuroimaging," *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 75, no. 9, pp. 1235–1244, sep 2004.

[19] P. D. Bhattacharya, *Diffusion MRI: Theory, methods, and applications*. Elsevier, 2012.

[20] J. M. Keil, A. Qalieh, and K. Y. Kwan, "Brain transcriptome databases: A user's guide," *The Journal of Neuroscience*, vol. 38, no. 10, pp. 2399–2412, feb 2018.

[21] A.-L. Barabási *et al.*, *Network science*. Cambridge university press, 2016.

[22] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, oct 2008.

[23] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, feb 2010.

[24] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, may 2006.

[25] A. Ortega, P. Frossard, J. Kovacevic, J. M. F. Moura, and P. Vandergheynst, "Graph signal processing: Overview, challenges, and applications," *Proceedings of the IEEE*, vol. 106, no. 5, pp. 808–828, may 2018.

[26] W. Huang, T. A. W. Bolton, J. D. Medaglia, D. S. Bassett, A. Ribeiro, and D. V. D. Ville, "Graph signal processing of human brain imaging data," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2018.

[27] M. G. Preti and D. Van De Ville, "Decoupling of brain function from structure reveals regional behavioral specialization in humans," *arXiv preprint arXiv:1905.07813*, 2019.

[28] R. C. Craddock, G. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg, "A whole brain fMRI atlas generated via spatially constrained spectral clustering," *Human Brain Mapping*, vol. 33, no. 8, pp. 1914–1928, jul 2011.

[29] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, W.-M. H. Consortium *et al.*, "The wu-minn human connectome project: an overview," *Neuroimage*, vol. 80, pp. 62–79, 2013.

[30] V. Fonov, A. C. Evans, K. Botteron, C. R. Almli, R. C. McKinstry, and D. L. Collins, "Unbiased average age-appropriate atlases for pediatric studies," *NeuroImage*, vol. 54, no. 1, pp. 313–327, jan 2011.

[31] [Online]. Available: https://commons.wikimedia.org/wiki/File:Brain_network.png

[32] F.-C. Yeh, S. Panesar, D. Fernandes, A. Meola, M. Yoshino, J. C. Fernandez-Miranda, J. M. Vettel, and T. Verstynen, "Population-averaged atlas of the macroscale human structural connectome and its network topology," *NeuroImage*, vol. 178, pp. 57–68, 2018.

[33] M. J. Hawrylycz *et al.*, "An anatomically comprehensive atlas of the adult human brain transcriptome," *Nature*, vol. 489, no. 7416, p. 391, 2012.

[34] A. Arnatkevic Iūtė, B. D. Fulcher, and A. Fornito, "A practical guide to linking brain-wide gene expression and neuroimaging data." *NeuroImage*, vol. 189, pp. 353–367, Apr. 2019.

[35] P. Goel, A. Kuceyeski, E. LoCastro, and A. Raj, "Spatial patterns of genome-wide expression profiles reflect anatomic and fiber connectivity architecture of healthy human brain." *Human brain mapping*, vol. 35, pp. 4204–4218, Aug. 2014.

[36] M. Forest, Y. Iturria-Medina, J. S. Goldman, C. L. Kleinman, A. Lovato, K. Oros Klein, A. Evans, A. Ciampi, A. Labbe, and C. M. T. Greenwood, "Gene networks show associations with seed region connectivity." *Human brain mapping*, vol. 38, pp. 3126–3140, Jun. 2017.

[37] M. A. Bertolero, A. S. Blevins, G. L. Baum, R. C. Gur, R. E. Gur, D. R. Roalf, T. D. Satterthwaite, and D. S. Bassett, "The network architecture of the human brain is modularly encoded in the genome," *arXiv preprint arXiv:1905.07606*, 2019.

[38] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith, "Fsl," *Neuroimage*, vol. 62, no. 2, pp. 782–790, 2012.

[39] P. McCarthy, "Fsleyes," 2019.

[40] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved optimization for the robust and accurate linear registration and motion correction of brain images," *NeuroImage*, vol. 17, no. 2, pp. 825–841, oct 2002.

[41] F.-C. Yeh, T. D. Verstynen, Y. Wang, J. C. Fernández-Miranda, and W.-Y. I. Tseng, "Deterministic diffusion fiber tracking improved by quantitative anisotropy," *PLoS ONE*, vol. 8, no. 11, p. e80713, nov 2013.

[42] M. Rubinov and O. Sporns, "Complex network measures of brain connectivity: Uses and interpretations," *NeuroImage*, vol. 52, no. 3, pp. 1059–1069, sep 2010.

[43] N. Perraudin, J. Paratte, D. Shuman, L. Martin, V. Kalofolias, P. Vandergheynst, and D. K. Hammond, "GSPBOX: A toolbox for signal processing on graphs," *ArXiv e-prints*, Aug. 2014.

[44] M. Xia, J. Wang, and Y. He, "Brainnet viewer: a network visualization tool for human brain connectomics," *PloS one*, vol. 8, no. 7, p. e68910, 2013.

[45] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot, "Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain," *NeuroImage*, vol. 15, no. 1, pp. 273–289, jan 2002.

[46] M. F. Glasser *et al.*, "A multi-modal parcellation of human cerebral cortex," *Nature*, vol. 536, no. 7615, pp. 171–178, jul 2016.

[47] R. S. Desikan *et al.*, "An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest," *NeuroImage*, vol. 31, no. 3, pp. 968–980, jul 2006.

[48] J. Arloth, D. M. Bader, S. Röh, and A. Altmann, "Re-annotator: Annotation pipeline for microarray probe sequences," *PLOS ONE*, vol. 10, no. 10, p. e0139516, oct 2015.

[49] C. Spearman, "The proof and measurement of association between two things," *American journal of Psychology*, vol. 15, no. 1, pp. 72–101, 1904.

[50] B. D. Fulcher, M. A. Little, and N. S. Jones, "Highly comparative time-series analysis: the empirical structure of time series and their methods," *Journal of The Royal Society Interface*, vol. 10, no. 83, pp. 20 130 048–20 130 048, apr 2013.

[51] A. Patania, P. Selvaggi, M. Veronese, O. DiPasquale, P. Expert, and G. Petri, "Topological gene-expression networks recapitulate brain anatomy and function," nov 2018.

[52] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. [Online]. Available: http://www.jmlr.org/papers/v9/vandermaaten08a.html

[53] J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing*. Springer, 2009, pp. 37–40.

[54] Student, "The probable error of a mean," *Biometrika*, pp. 1–25, 1908.

[55] J. Cohen, *Statistical power analysis for the behavioral sciences*. Routledge, 2013.

[56] I. Diez, P. Bonifazi, I. Escudero, B. Mateos, M. A. Muñoz, S. Stramaglia, and J. M. Cortes, "A novel brain partition highlights the modular skeleton shared by structure and function," *Scientific reports*, vol. 5, p. 10532, 2015.

[57] MathWorks, "pdist function help." [Online]. Available: https://es.mathworks.com/help/stats/pdist.html