UNIVERSIDAD DE CHILE
FACULTAD DE CIENCIAS FÍSICAS Y MATEMÁTICAS
DEPARTAMENTO DE INGENIERÍA QUÍMICA Y BIOTECNOLOGÍA

MODELAMIENTO Y ESTUDIO DE LA RED DE INTERACCIONES PROTEICAS DEL COMPLEJO NRC/MASC

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO CIVIL EN BIOTECNOLOGÍA

JAIME ALBERTO CAMPOS VALENZUELA

PROFESOR GUÍA:
JOSÉ CRISTIAN SALGADO HERRERA

MIEMBROS DE LA COMISIÓN:
ZIOMARA GERDTZEN HAKIM
ORIANA SALAZAR AGUIRRE

SANTIAGO DE CHILE
ABRIL 2010

ESTE TRABAJO HA SIDO FINANCIADO EN PARTE POR EL PROYECTO FONDECYT 11080016

Resumen

La presente memoria tiene por objetivo principal el introducir nuevas metodologías en la inferencia de interacciones interproteicas y aplicar aquellas relaciones putativas en el estudio de la estructura receptora NRC/MASC. Esto con el objeto de aumentar nuestro conocimiento sobre el sistema sináptico y levantar nuevas hipótesis acerca la relación de la organización de la membrana postsinaptica con el gatillamiento de enfermedad cognitivas. Lo anterior con el fin de desarrollar nuevas terapias que permitan un ataque al mecanismo de las diversas enfermedades y no sólo a las expresiones de ellas.

Para el desarrollo de los objetivos se propuso un protocolo nuevo, en donde se unen dos metodologías novedosas. En primer lugar la aplicación del clasificador Naïve-Bayes para inferir interacciones interproteicas en el ser humano, logrando de esa forma obtener una red de interacción más amplia y con un parámetro de confianza para cada uno de sus elementos. En segundo lugar, un estudio sistémico de la unidad NRC/MASC, buscando comprender el funcionamiento de ella a través de su red de interacciones (utilizando allí la integración de la red inferida con otras redes entregadas por proyectos anteriores), con lo cual se define un modelo funcional completamente nuevo y que permite estudiar cómo afecta a la unidad el desarrollo de las enfermedades cognitivas.

Se logró a través del proyecto el calcular la red de interacciones interproteicas para el ser humano, se concluyó que la red obtenida era comparable con las entregadas por otros estudios, tanto en la calidad de sus datos como en la cantidad de elementos de ésta. Luego de la combinación de las redes se entregó un conjunto de interacciones interproteicas de la NRC/MASC con un número de interconexiones similar a los de trabajos anteriores. Finalmente se definió las unidades funcionales del complejo NRC/MASC y se realizaron diversas conjeturas sobre el funcionamiento del complejo como un sistema integrado y de cómo se ve modificado cuando alguna enfermedad cognitiva está presente.

De allí se obtuvo una organización totalmente nueva, que pasa desde un modelo tipo caja negra a uno con mayor complejidad entre los elementos pertenecientes y que permite comprender a la unidad NRC/MASC como un sistema de múltiples vías de señalización que trabajan de forma paralela.

Índice general

Índice de Figuras

Capítulo I

Introducción

La presente memoria tiene por finalidad el señalar los elementos del trabajo realizado en el transcurso del año 2009; es así como en este capítulo se presentarán los principales elementos que permitieron desarrollar el tema: la motivación del proyecto, el estado de arte de las metodologías utilizadas en la implementación del proyecto, los objetivos propuestos y las metodologías utilizadas en el proyecto. Además de presentar en los capítulos siguientes los resultados logrados, en conjunto con las discusiones realizadas y las conclusiones que han sido posible obtener a través del desarrollo de este estudio.

I.1. Motivación

La motivación de este proyecto nace, en primer lugar, de la comprensión del rol que juega el neurotransmisor glutamato en el desarrollo de enfermedades neurológicas tales como la esquizofrenia, Alzheimer, autismo y depresión, entre otras. El papel preponderante de este neurotransmisor lo convierte en un candidato ideal de estudio para comprender el desarrollo de las diversas enfermedades señaladas, pues, si se lograra modelar su orgánica (la lógica interna de funcionamiento), esto permitiría el desarrollo de nuevos tratamientos para enfermedades tan complejas como la esquizofrenia y el autismo.

Con el fin de lograr este objetivo se ha iniciado desde el comienzo de la década actual un estudio de la unidad NRC/MASC (N-methyl-D-aspartate receptor complex/MAGUK[1] associated signalling complex), el cual es un complejo receptor de glutamato. Se le ha señalado como el principal actor en el gatillamiento de las enfermedades antes señaladas. El enfoque de este estudio es novedoso, pues implica el estudio del sistema cognitivo como una red de interacciones dentro de la neurona y no sólo como la red de interacciones interneuronal, el cual es el paradigma clásico.

Por otro lado, este proyecto busca nuevas metodologías que permitan un modelamiento

[1]Membrane-Associated Guanylate Kinase

novedoso de la unidad NRC/MASC, el cual entregue información antes no conocida y que finalmente logre darle una vuelta de tuerca a la compresión actual de la unidad NRC/MASC. Es con este fin que se aplicará un algoritmo de clasificación para poder obtener un set de interacciones interproteicas (Protein-Protein Interaction: PPI) putativas, las que permitan finalmente definir una nueva configuración interna del complejo NRC/MASC.

I.2. Descripción del proyecto

Para lograr establecer el nuevo modelo del complejo NRC/MASC este proyecto de memoria se dividió en 2 fases principales:

Inferencia de Interacciones Interproteicas: La primera parte del proyecto buscó entregar una red de interacciones interproteicas inferidas. Es decir, interacciones obtenidas no a través de la literatura, sino por la integración de variadas bases de datos -sin redundancia entre ellos- que nos entreguen información sobre la posibilidad de que una PPI en particular se lleve a cabo.

Modelo y estudio de la unidad NRC/MASC: Se utilizó las PPIs obtenidas en la primera parte y se definió el conjunto de interacciones que pertenecen a la unidad NRC/MASC. Fue posible modelar esta unidad en función de sus interacciones internas; obteniendo a través de diversas metodologías las subunidades funcionales de ella. Finalmente se obtuvo una comprensión nueva del funcionamiento de la unidad y de cómo se ve afectada cuando una enfermedad neurológica está presente.

La primera fase integró 4 bases de datos: (a detallar en las secciones siguientes) Interacciones Ortológicas, Matrices de Coexpresión, Funciones Biológicas Compartidas y Pares de Dominio Enriquecido. Las cuales no entregan en principio información sobre posibles PPIs, pero a través de una serie de hipótesis (ver sección I.3.3) y su agregación a través del algoritmo de Naïve-Bayes, fue posible transformar información diversa en interacciones interproteicas con un parámetro probabilístico. Este parámetro señala la probabilidad de que una interacción en particular sea un verdadero positivo o un verdadero negativo. Lo vital de esta primera fase fue la ampliación del conocimiento que se tiene sobre la red de PPIs del ser humano, aumentar el número de conexiones a través de ciertas hipótesis, para que posteriormente se seleccionó aquellas con un más alto grado de certeza y permitió realizar la segunda parte del proyecto.

Por otro lado, la segunda fase utilizó los estudios sobre la proteómica de la postsinápsis (Postsynaptic proteomic: PSP) y en particular de la densidad postsináptica (Postsynaptic density: PSD) para definir las proteínas que pertenecen a la unidad y con ello las PPIs que señalan la estructura de la misma. Luego se aplicó un algoritmo de clustering sobre la red para así obtener las subunidades que la conforman. En función de la calidad de las PPIs putativas calculadas se definieron subunidades altamente determinadas y con funciones particulares, lo que permitió proponer nuevas tesis sobre la orgánica y funcionamiento de la unidad.

I.3. Antecedentes

I.3.1. Antecedentes Generales

Con el desarrollo en la última década de las técnicas de *high-throughput* en la biología molecular -tales como los *microarrays* de RNAm y de proteínas, espectrometría de masa y *two-hybrid*- se ha desarrollado un nuevo campo dentro de la biotecnología llamado "biología de sistemas". La cual tiene por énfasis el comprender y modelar los procesos biológicos como sistemas complejos, utilizando con este fin tanto técnicas experimentales (*high-throughput*) como técnicas teóricas (*in silico*). Es con ello que se ha logrado desarrollar nuevas perspectivas de investigación en el último tiempo, tales como la genómica, transcriptómica, proteomica y metabolómica, las cuales tienen como fin último el clasificar y comprender los procesos que son llevados a cabo por los distintos componentes, ya sean éstos genes, transcripciones, proteínas o metabolitos, generando finalmente redes de componentes celulares.

Dentro de este ámbito se ha iniciado la aplicación de metodologías novedosas para la biología, tal como es el uso de algoritmos clasificadores. Estos han sido aplicados a diferentes componentes moleculares, principalmente relaciones binarias. A raíz de ello se ha desarrollado la inferencia de redes de regulación génica a través de distintas metodologías -tales como las redes bayesianas, booleanas y sistema de ecuaciones lineales- rompiendo con este nuevo enfoque la metodología "clásica" de recreación basada en la minería de datos a través de la literatura. El paso siguiente ha sido el estudiar la aplicación de estas metodologías en la predicción de interacciones proteicas y con ello enfocar el estudio en el funcionamiento de los complejos biológicos.

Así, el conjunto de procedimientos utilizados para definir la proteómica del ser humano, en particular son de altísimo interés los trabajos realizados para estudiar y definir la proteómica de la densidad postsináptica. Además se ha puesto énfasis en la serie de estudios sobre la aplicación de algoritmos de clasificación en la inferencia de interacciones interproteicas, un área de estudio bastante nueva y que está mostrando sus frutos.

A continuación se presenta el estado del arte para las dos fases del proyecto, señaladas anteriormente, y su aplicación al trabajo aquí realizado.

I.3.2. Estudio de la unidad NRC/MASC

El tratamiento de enfermedades neurológicas, como la esquizofrenia, retardo mental, Alzheimer y otras, ha tenido un desarrollo lento y dificultoso; en la actualidad no existen curas para aquellas afecciones y los tratamientos disponibles están enfocados principalmente en la inhibición de la expresión de los endofenotipos[2] producidos por ellas.

Es por ello que un cambio de paradigma en la elaboración de nuevas terapias se hace

[2]Síntoma (psiquiátrico) cuya fuente es del tipo genético

necesario, terapias que estén enfocadas en atacar la raíz de los engranajes que producen estas enfermedades y no únicamente las expresiones de ellas. Bajo esa lógica se han desarrollado estudios que buscan aquellos genes responsables del gatillamiento de la enfermedad [21, 22]. Así se ha logrado definir -a grandes rasgos- los mecanismos moleculares involucrados en la existencia de las ya precisadas enfermedades, en particular la esquizofrenia.

De allí, en diversos estudios [9, 18] se ha planteado el enlace existente entre la neurotransmisión de glutamato y la expresión de enfermedades neurológicas, en particular, la esquizofrenia. De estos estudios se plantea la tesis de que este tipo de enfermedad no puede ser estudiado de manera exclusivamente molecular, es decir, estudiar los genes principales que las gatillarían, sino que se debe estudiar el sistema receptor en su conjunto. Comprender la orgánica de la recepción como un total, tal como señala Lang *et al.* [18]: "..., avances moleculares sugieren la existencia de una relación compleja entre receptores, quinasas, proteínas y hormonas, la cual está involucrada en la esquizofrenia. En una hipótesis unificadora, diferentes cascadas se unen, logrando finalmente el desarrollo de los síntomas de los desordenes esquizofrénicos.[3]"

Es por ello que en este trabajo se recoge la tesis planteada por Chaudhary *et al.* [4], en donde se señala una aproximación postgenómica al problema: se trata de utilizar la proteómica como enlace entre la genómica, la fisiología y la patología. Es decir, a través del estudio proteómico desarrollar modelos sistémicos de los mecanismos postsinápticos que permitan entender la génesis de estas enfermedades.

La importancia de los receptores de glutamato se basa en su rol como receptores del código neuronal e iniciadores de cambios en el largo de plazo de la estructura y función de la sinápsis [12, 9]. Además existen 3 características particulares de los receptores que los convierten en interesantes candidatos a estudio [15]:

- Están físicamente unidos a una multitud de proteínas, formando complejos mediadores e iniciadores de señales.
- La proteómica de la postsinápsis ha mostrado que muchas de las proteínas pertenecientes a ella están involucradas en enfermedades humanas.
- A través de la manipulación genética en el ratón común (*Mus musculus*), se ha logrado diversificar los antagonistas de las proteínas de la PSD, logrando con ello estudiar la funcionalidad de genes particulares para ciertos estados alterados.

Receptores de glutamato

El glutamato es emitido desde la terminal presináptica en respuestas de potenciales activadores, siendo entonces transmitido a través del espacio sináptico hacia el terminal postsináptico, excitando receptores ubicados en la membrana postsináptica. Existen principalmente 3

[3]However, molecular findings suggest that a complex interplay between receptors, kinases, proteins and hormones is involved in schizophrenia. In a unifying hypothesis, different cascades merge into another that ultimately lead to the development of symptoms adherent to schizophrenic disorders.

tipos de receptores de glutamato, definidos inicialmente por el antagonista de glutamato que reciben:

N-methyl-D-aspartic acid receptor (NMDA): Este receptor es del tipo ionotrópico (forma canales iónicos) y su principal rol es el de iniciar la vía de señalización por transducción de la señal entrante.

α-amino-3-hydroxy-5-methyl-4-isoxazolepropionic acid (AMPA): También es un receptor ionotrópico, siendo su rol principal el de mediar en la despolarización de la membrana, la cual es necesaria para iniciar el potencial de acción (también llamado impulso eléctrico) en la neurona receptora.

Metabotropic glutamate receptor (mGluR): A diferencia de los otros receptores, el mGluR es un receptor metabotrópico, ligado a proteínas G. Este tipo de receptor, al igual que el NMDA-R, tiene por principal misión el iniciar vías de señalización. Al no tener canales iónicos la transducción de la señal se lleva a cabo a través de vías de señalización mediadas por proteínas G.

En la topología de la terminal postsináptica el receptor AMPA-R se encuentra desligado de los otros 2 receptores, pues ellos se encuentran enlazados por proteínas "andamio", estando localizados en complejos proteícos junto con proteínas señalizadoras, de regulación y otras.

Todos estos receptores se encuentran dentro de la llamada proteómica postsináptica, la cual es el conjunto de proteínas pertenecientes a la terminal postsináptica. Su estudio total se hace necesario para comprender los mecanismos inherentes a la recepción e iniciación de señales sinápticas.

La proteomica y la densidad postsináptica

Se ha logrado definir y describir sus componentes, en Collins *et al.* [5] y Pocklington *et al.* [27], a través de diversas metodologías y agregando los resultados obtenidos por distintos autores. De estos estudios se logró definir que la PSP contiene $\sim$ 1800 proteínas [36], estando integrada por varios complejos proteicos. El más grande de ellos es la densidad postsináptica (PSD), una estructura densa que se ubica bajo la membrana postsináptica, tal como se observa en la Figura I.1, estando compuesta por $\sim$ 1124 proteínas.

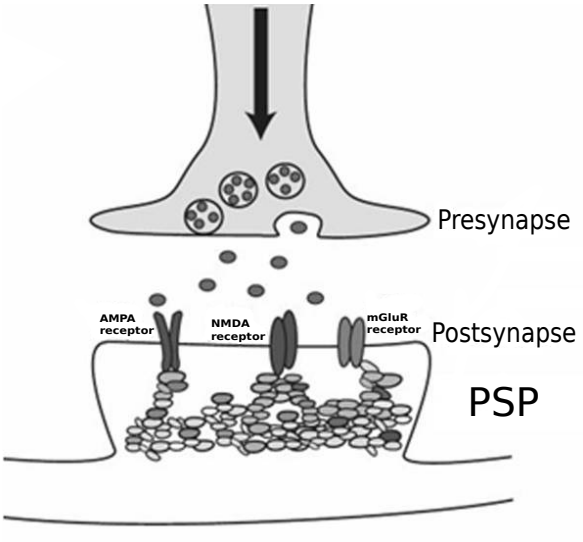

Figura I.1: Diagrama de la terminal postsináptica. Obtenido de Laumonnier *et al.* [19]

Las proteínas que componen a la PSD pertenecen a diferentes clases, representando una diversidad de funciones biológicas, las cuales se pueden observar en el anexo F de este trabajo. Para una lista completa de la PSP ver Laumonnier *et al.* [19].

Tal como señala Pocklington *et al.* [28], la PSD tiene un enriquecimiento en dominios de interacción interproteica, tales como PDZ y SH3, los cuales funcionan como proteínas "andamio" y unen complejos proteicos, tales como los encontrados en la densidad postsináptica.

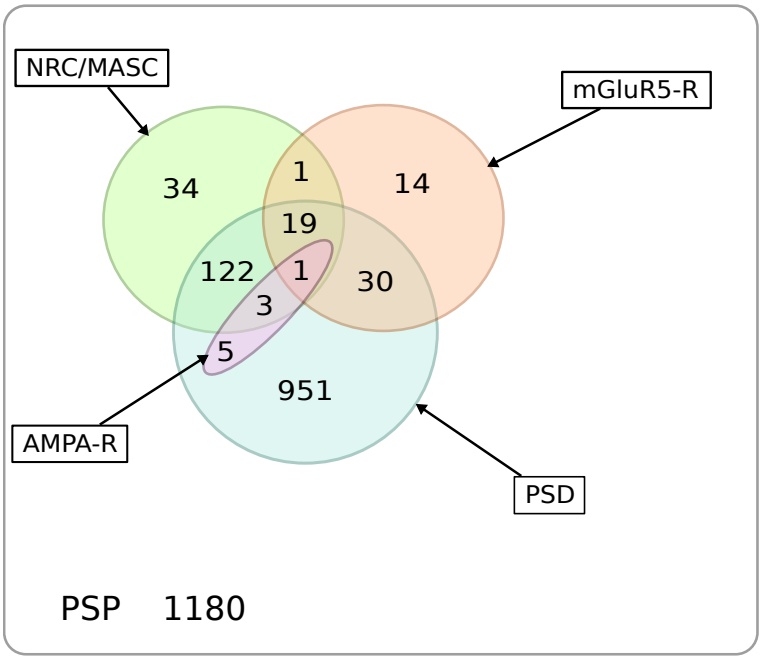

Figura I.2: Diagrama de Venn del número de proteínas que componen la PSP y sus complejos

Complejos receptores de glutamato

Los complejos receptores de glutamato son subconjuntos de la PSP; existe además un alto grado de solapamiento de las proteínas participantes en ellos, tal como se puede observar en la Figura I.2. Existe también un gran número de estos receptores por terminal postsináptica. La purificación de los complejos receptores de la NMDA (NRC) o de las proteínas MAGUK[4] -las cuales están directamente unidas al NRC- entregó un total de 185 proteínas. A éstas se les ha denominado como el complejo NRC/MASC[5], pues ambas purificaciones entregaron resultados comparables [5].

Dentro del NRC/MASC se ubican las subunidades NMDA y mGluR, mientras que las subunidades AMPA-R se encuentran en complejos separados y más pequeños. A través del aislamiento por afinidad se ha obtenido que los receptores NMDA y mGluR se encuentran unidos por varias proteínas andamio. Por otro lado y tal como se observa en la diagrama de Venn de la Figura I.2, los complejos se encuentran introducidos parcialmente en la PSD, considerandola como un complejo "supermolecular".

El complejo NRC/MASC es la subunidad más estudiada de la PSP, pues se le considera un prototipo del funcionamiento global de la PSP. El rol fisiológico del NRC/MASC ha sido estudiado utilizando ratones con *knockouts* e intervención farmacológica (utilizando antagonistas de glutamato). Con ello se ha obtenido que se necesitan más de 40 proteínas del NRC para llevar a cabo el proceso de convertir patrones neuronales en cambios duraderos de la función neuronal; parecido número es necesario para la permitir la plasticidad del comportamiento, en casos tales como el aprendizaje o el miedo condicionado [27, 28].

Estructura y funcionalidad del NRC/MASC

Además de la acumulación de conocimiento sobre la genética del ratón y datos fenotípicos sobre el complejo NRC/MASC, se ha utilizado las interacciones interproteicas dentro del complejo para generar redes funcionales [27]. De allí se han logrado obtener variados resultados, entre ellos destacan que el número medio de proteínas separando cualquier par es de 3.3, siendo éste muy bajo, ante lo cual se ha discutido la alta intraconectividad del complejo y se ha sugerido que el complejo consiste en una red con múltiples clusters, en vez de una red lineal, medianamente conectada.

Para estudiar la posible estructura de la red se aplicó el algoritmo desarrollado por Newman y Girvan [24], en donde se llevó a cabo una nueva metodología para realizar clustering sobre grafos en función de su conectividad.

El procedimiento del algoritmo se basa en obtener el parámetro *betweenness*, el que cuantifica cuán vital es un vértice en la conexión total del grafo. Este se obtiene -de forma general- realizando un conteo de veces que un vértice fue ocupado para conectar dos nodos cualquiera

[4]Membrane-associated guanylate kinase
[5]MAGUK-associated signaling complexes

(para todas los pares de nodos posibles), tal como se puede observar en la Figura I.3.

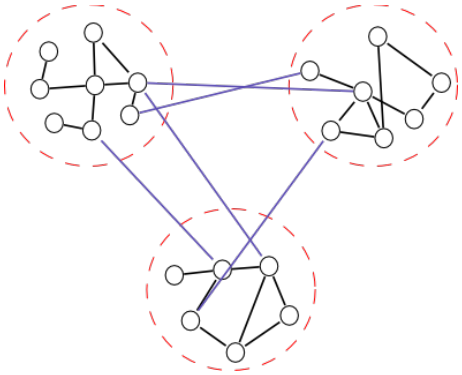

Figura I.3: Diagrama del clustering de grafos por conectividad

Luego se realiza un clustering jerárquico utilizando el parámetro *betweenness*, partiendo desde el vértice con mayor *betweenness*.

Luego se selecciona un corte en la árbol jerarquico fundamentado en el parámetro modularidad[6] (Q), el que señala cuán independiente e intraconectando están los clusters obtenidos. En la Figura I.4 se puede observar el orden jerárquico y la línea punteada señala el corte con el máximo Q posible. Al eliminar los vértices que son cortados por la línea se obtiene los clusters del grafo.

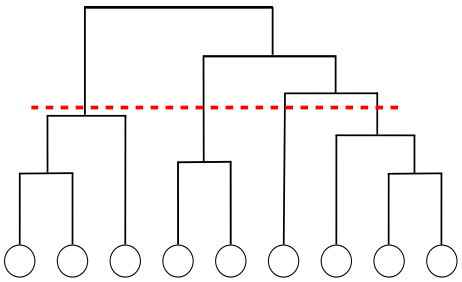

Figura I.4: Diagrama de la división jerárquica de la red y corte en función de Q.

A través de este procedimiento Pocklington *et al.* lograron definir una organización del complejo NRC/MASC; de allí se obtuvo que éste está consituido por 13 *clusters*, cado uno caracterizado funcional y fenotípicamente. Éstos pueden ser observados en la Figura I.5, en donde los clusters están jerarquizados en 3 capas: la primera es llamada "capa receptora" (*input*), en donde están los representados los receptores de glutamato (NMDA-R y mGluR); la segunda es la "capa de procesamiento" (*processing*), donde están las principales proteínas señalizadoras. Finalmente la "capa de salida" (*output*), la cual tiene como característica principal ser un conjunto de proteínas ejecutoras y de conducción de vías de señalización, tales como la vía ERK/MAPK.

[6] *Modularity*

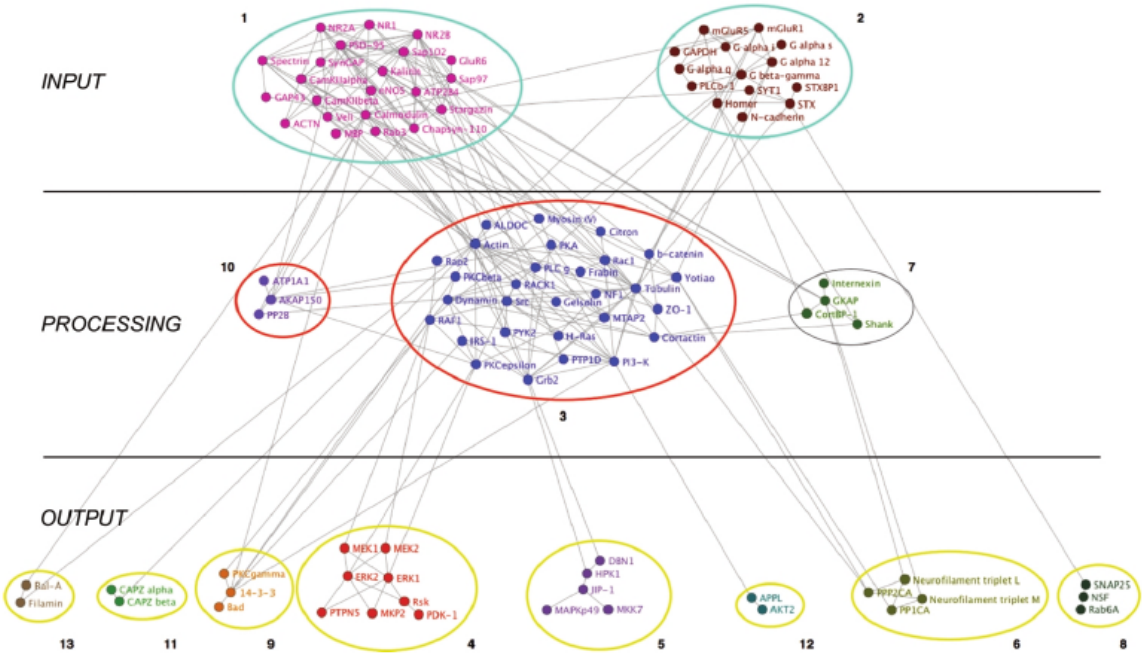

Figura I.5: Diagrama de la organización del complejo NRC/MASC según Pocklington *et al.*

El modelo del complejo se presenta en la Figura I.6, en donde se puede observar como los módulos receptores -ionotrópicos (1) y metabotrópico (2)- en conjunto con sus proteínas asociadas reciben la señalan neuronal, para luego pasar a un gran módulo de procesamiento de información (3) , el cual distribuye las señales a las redes de modulación de señal (amarillo) y mecanismos efectores (verde), las cuales regulan la salida funcional (flecha azul).

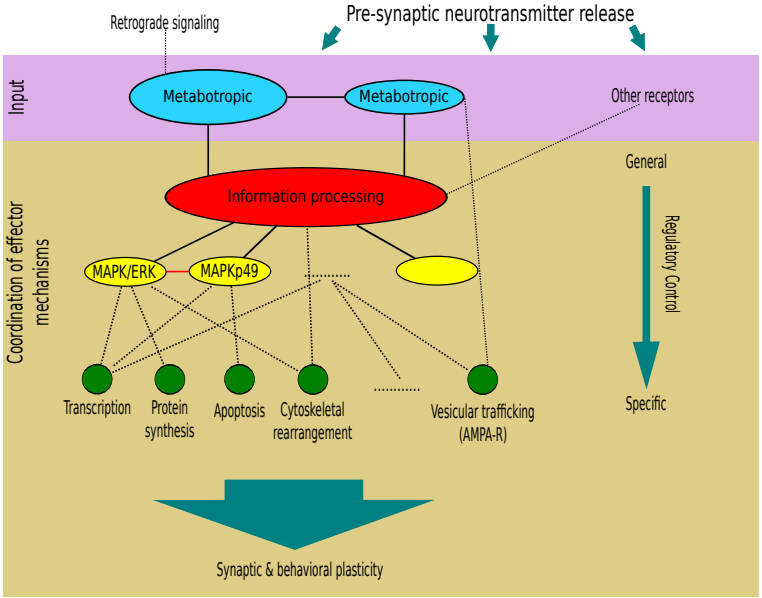

Figura I.6: Organización funcional del complejo NRC/MASC, basado en Pocklington *et al.* [28].

Las mutaciones en humanos y ratones que producen cambios en la plasticidad sináptica y de aprendizaje fueron mapeados en la red, los cuales señalaron una distribución fenotípica en los módulos de la red, demostrando con ello el carácter estructurado del complejo y dándole un primer enfoque a la comprensión funcional de éste.

Siguiendo con la descripción del proyecto planteada con anterioridad, es necesario señalar los antecedentes vitales para la obtención de interacciones interproteicas putativas.

I.3.3. Inferencia de Interacciones Interproteicas

La obtención de las redes de interacción interproteicas ha sido vista con especial interés al establecerse las PPIs como las bases de los complejos y vías que llevan a cabo los procesos biológicos, considerando como interacción no sólo la unión o acoplado, sino todo tipo de relación interproteica (las cuales se definen en la sección I.3.3). De allí que su conocimiento sea el primer paso para descubrir los mecanismos que subyacen en las distintas funciones biológicas, además de poder asignar en una primera instancia funciones biológicas a proteínas desconocidas.

A lo largo del último tiempo se han aplicado acercamientos experimentales *high-throughput* con el fin de determinar PPIs a escala de la proteómica total de una especie. Estos se tratan preferentemente de sistemas *two-hybrid* (Y2H) y de espectrometría de masas. Estos sistemas presentan variados problemas tanto en su aplicación como en la calidad de los resultados [16, 4, 29]. Así por ejemplo, tal como señala Jansen *et al.* [16] muchas veces los sets de datos de interacciones están incompletos o son contradictorios. Por otro lado, en el contexto de la proteómica total del individuo, estos errores se ven potenciados, pues el número de proteínas que no interaccionan son muchísimo más que los que sí interaccionan. De allí la necesidad de buscar nuevas vías para aumentar la fidelidad de los resultados y a la vez la cantidad de éstos.

Métodos de inferencia

Para subsanar las inexactitudes de los métodos tradicionales se ha considerado la integración de variadas bases de datos, entregando todas ellas algún grado de certeza sobre la interacción de 2 proteínas, permitiendo de esa manera descubrir interacciones a través de un método más robusto. Con ello se lograría disminuir el número de falsos positivos del set de datos de interacciones y por lo tanto la calidad final de éste.

Así por ejemplo, muchas proteínas interactuantes son coexpresadas, y por otro lado, proteínas que se encuentran en los mismos complejos tienen una probable interacción en función de la especificidad del complejo. Bajo esta lógica se propuso utilizar tanto los datos directos de interacciones como los indirectos, e integrarlos en un modelo de aprendizaje.

Se ha demostrado que estos procedimientos *in silico* han logrado mejorar la predicción de

las interacciones interproteicas, siendo comparadas con metodologías exclusivamente directas. Esta mejora no se basa únicamente en el encuentro de interacciones nuevas, sino que también en la capacidad de estratificar las interacciones candidatas a través de un parámetro de confianza.

A pesar que todos los procedimientos de inferencia tienen por elemento común el utilizar algoritmos de clasificación, estos tienen 3 puntos de divergencia importantes:

Set de datos *Gold Standard*: Es el set de datos utilizado para el aprendizaje y testeo del algoritmo. Se han ocupado 3 tipos distintos de sets. El primero es utilizado para predecir interacciones físicas y tiene como base de datos el entregado por DIP [38]. El segundo es ocupado para definir interacciones de forma más amplia, en donde las proteínas se consideran interactuantes, incluso si sólo interaccionan a través de una tercera proteína. Para ello se ocupa la base de datos MIPS [14], siendo utilizado el tercer tipo de set de datos para inferir redes de rutas (*pathways*), utilizando para ello la base de datos KEGG [17].

Codificación de las características ocupadas para la predicción: Fundamentalmente se ocupa dos tipos distintos de características: el primero utiliza un estilo detallado, en donde cada uno de los experimentos es considerado separadamente, mientras el segundo tipo, definido como estilo resumido, agrega experimentos similares para entregar un único resultado (cómo es el caso del clasificador de Naïve-Bayes [13], en donde se agregan todos las distintas bases de datos).

Métodos de Clasificación: Se ha desarrollado una gran cantidad de métodos de clasificación, ellos se pueden observar en la Tabla I.1, en donde se describen en función del set de datos del Gold Standart que utilizan y el tipo de codificación de las características a utilizar por cada uno de ellos. Así, y señalando los principales métodos, se tiene: el método de *Logistic regression* (LR) ha sido usado para estimar la probabilidad posterior (O_{post}) de que un par de proteínas interactúe de forma directa, utilizando set de datos obtenidos a través de métodos *high-throughput*. En la predicción de interacciones de co-expresión se ha aplicado el algortimo de Naïve-Bayes, utilizando características agregadas de diversa procedencia. Finalmente, el árbol de decisión (DT) utiliza una codificación detallada de los set de datos, sin agregarlas, para obtener relaciones co-expresadas.

Tabla I.1: Métodos de clasificación de interacciones interproteicas

	Tipo Predicción		
Tipo Codificación	Co-complejo	Interacción Física	Ruta
Agregado	Naïve-Bayes, LR, Random Forest	Logistic Regression	Bayesian Statistic Scoring
Detallado	Árbol de decisión		Kernel method

En particular para este proyecto es esencial el modelo de inferencia basado en el algoritmo de Naïve-Bayes.

Clasificador Naïve-Bayes

Recientemente se ha comenzado el uso de algoritmos de aprendizaje capaces de utilizar características tanto genéticas como experimentales. Los clasificadores bayesianos tienen una base probabilística y poseen la capacidad de integrar una gran gama de datos heterogéneos, lo cual lo diferencia de otros clasificadores como *random forest* y *logistic regression*, tal como se señaló con anterioridad.

Ciertas características hacen único al modelo de clasificación por Bayes. Puede integrar tipos datos tan disímiles como valores de expresión de microarrays, valores de categorización del Gene Ontology y valores ortológicos booleanos. Cada uno de los valores crudos es transformado y uniformado a valores probabilísticos a través del cálculo del *likelihood ratio* (LR). Además, cada fuente de datos es automáticamente ponderada a través de la confiabilidad de sus datos. Datos ausentes son tolerables para el algoritmo. Por otro lado, tal como señala Xia *et al.* [40], el modelo bayesiano es un algoritmo rápido y simple, además de no necesitar la normalización de sus fuentes de datos. Todas estas características, además de sus excelentes resultados [16, 30, 40, 32, 2], lo presentan como un método de inferencia idóneo para la obtención de redes de interacción nóveles.

La metodología para utilizar el clasificador se inicia, como cualquier otro algoritmo de clasificación, con la definición de los set de datos *gold standard*. En este caso es necesario definir un *gold standard* positivo (GSP) y otro negativo (GSN). El positivo contiene todas aquellas interacciones obtenidas a través de métodos físicos (*i.e.* sin predecir), mientras que el *gold standart* negativo tiene aquellas interacciones que no se llevan a cabo. Los *gold standard* deben tener ciertas características indispensables, tales como la independencia con las otras fuentes de datos utilizadas en el clasificador; además deben ser lo suficientemente grandes para ser estadísticamente confiables. Finalmente, deben estar libres de cualquier sesgo sistemático.

Tal como se señaló anteriormente, para calcular el poder predictivo o nivel de confianza de un par de genes se define el parámetro *likekihood ratio* (LR). Para ejemplificar el funcionamiento del LR se debe considerar un par de proteínas o clases, f, expresada en término binarios ("presente"/"ausente"). El likelihood ratio $LR(f)$ es definido como la fracción de la GSP que tiene aquella característica f dividido por la fracción de la GSN que la tienen.

Por otro lado, una característica fundamental del LR es que para dos características o clases f_1 y f_2 con evidencia independiente, el likelihood ratio de la combinación es simplemente $LR(f_1, f_2) = LR(f_1)LR(f_2)$. Para cuando no haya independencia esta igualdad no es tal. De allí nace una de las hipótesis más exigentes del clasificador Naïve-Bayes, la independencia de su evidencia.

El método de obtención del likelihood ratio a través de Bayes se inicia con la definición de `positivo` cuando 2 proteínas interactúan -verdaderos positivos- mientras que `negativo` cuando no lo hacen. Considerando la totalidad de los pares que interactúan dentro del total de los posibles pares, se define que la probabilidad *a priori*, es decir, sin evidencia, de encontrar

un par positivo es:

$$O_{prior} = \frac{P(positivo)}{P(negativo)}$$

Donde $P(positivo)$ es la probabilidad de obtener un par positivo de todo el conjunto de posibles pares y $P(negativo)$ la probabilidad de encontrar uno negativo.

Mientras que la probabilidad posterior es la probabilidad de obtener un positivo cuando se considera la evidencia:

$$O_{posterior} = \frac{P(Positivo|evidencia_1, evidencia_2, \ldots, evidencia_n)}{P(Negativo|evidencia_1, evidencia_2, \ldots, evidencia_n)}$$

Cabe mencionar que tanto O_{prior} y O_{post} son parámetros probabilísticos, así sus valores pueden ir desde 0 hasta ∞.

La evidencia son las bases de datos utilizados para realizar la inferencia. Así el *likelihood ratio* queda definido como:

$$LR(evidencia_1, \ldots, evidencia_n) = \frac{P(evidencia_1, evidencia_2, \ldots, evidencia_n|Positivo)}{P(evidencia_1, evidencia_2, \ldots, evidencia_n|Negativo)}$$

El cual se relaciona con las probabilidades a priori y posterior a través del teorema de Bayes:

$$O_{posterior} = L(evidencia_1, \ldots, evidencia_n) * O_{prior}$$

Para el caso en que se tiene N evidencias independientes, el likelihood ratio es:

$$LR(evidencia_1, \ldots, evidencia_N) = \prod_{i=N}^{N} LR(evidencia_i) = \prod_{i=N}^{N} \frac{P(evidencia_i|Positivo)}{P(evidencia_i|Negativo)}$$

Definida en este trabajo como LR_{comp}:

$$LR_{comp} = \prod_{i=N}^{N} LR(evidencia_i) \tag{I.1}$$

El likelihood ratio puede ser calculado para una evidencia a través de su discretización, creando clases en función de la evidencia, para luego multiplicar los distintos LR obtenidos de distintas evidencias para un par de proteínas en particular. En el siguiente diagrama se muestra en forma general la aplicación de esta metodología en la obtención de LRs para las distintas bases de datos utilizadas.

Como se puede observar en la Figura I.7 de un set de datos se obtienen una serie de posibles interacciones, todas ellas además con un coeficientes K_i, el cual será utilizado posteriormente para definir las distintas clases. Como ejemplo se puede señalar que para la base de datos de coexpresión de genes, se calculó el coeficiente de correlación de Pearson para cada par posible de genes. Esto de acuerdo con el valor del coeficiente se definieron una serie de clases, para luego calcular el LR de cada una de ellas. La división de la clases se debe realizar utilizando

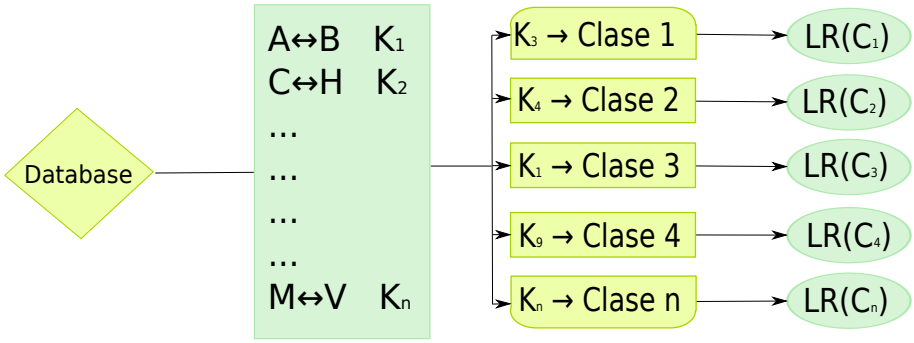

Figura I.7: Diagrama de la obtención del LR para una set de datos.

un parámetro que permita jerarquizar los LRs, para así obtener conjuntos específicos de interacciones con un altos LRs.

Finalmente se predice que un par de proteínas interaccionan si su *likelihood ratio* combinado (LR_{comp}) es mayor a un LR de corte, $LR > LR_{cut}$, mientras que si es menor no interacciona. Se calcula el LR_{cut}, el cual se define de tal forma que $O_{posterior} > 1$, lo que significa que hay más de un 50 % de posibilidad que sea un verdadero positivo a un verdadero negativo. Para esto además se necesita el valor de O_{prior}, siendo ella calculada utilizando la GSP, así se obtiene la probabilidad de sacar 2 proteínas que interactuen entre todo el conjunto posible de pares de proteínas. Si en el GSP hay 100 proteínas, con 2000 interacciones obtenidas a través de diversas metodologías, entonces el conjunto de pares posibles será $\frac{100*99}{2} = 4950$, y la probabilidad de sacar un par que interacciones es:

$$O_{prior} = \frac{2000}{4950} = 0.40404$$

De lo cual se obtiene que:

$$LR_{cut} = \frac{1}{O_{prior}} = 2,465$$

El paso siguiente al uso del algoritmo es la implementación y uso de las distintas bases de datos.

Bases de datos integradas

Las bases de datos integradas a través de Naïve-Bayes deben tener como principal cualidad el ser totalmente independientes, es por ello que Rhodes *et al.* [30] utilizó las siguientes bases de datos en su inferencia:

Interacciones por Ortología Si un par de proteínas en un organismo distinto al humano interaccionan, entonces existe la posibilidad que un par de proteínas, los cuales son su símil ortológico en el ser humano, también interaccionen. Es por ello que se toman las interacciones interproteicas de 3 especies: *S. cerevisiae*, *C. elegans* y *D. melanogaster*.

Estas interacciones son integradas y jerarquizadas en función de parámetros internos de los estudios de donde se obtuvieron a través del mapeo ortológico.

Matrices de Coexpresión Se considera que un par de proteínas que interaccionan entre sí pueden tener un perfil de expresión similar. Ante esta hipótesis se estudia la expresión de los genes a través de *microarrays*, en donde se compara la similitud de la expresión a través del coeficiente de correlación de Pearson, que sirve para definar las distintas clases.

Proceso Biológico Compartido Si dos proteínas se encuentran en el mismo proceso biológico, entonces existe la probabilidad de que ellas interaccionen. Es más, si aquel proceso es muy pequeño -respecto al número de proteínas que lo integran- entonces la probabilidad de interacción aumenta. Es por ello que se realiza una búsqueda para cada par de proteínas del proceso biológico más pequeño (*Smallest Shared Biological Process*: SSBP) en que ambas estén involucradas, para luego estratificar en función del número de elementos distintos de aquel proceso.

Dominios Proteicos Como la interacción entre dominios proteicos implica una relación física entre las proteínas, es posible encontrar nuevas interacciones si se tuvieran un compendio de dominios interactuantes. Es por ello que utilizando las relaciones del GSP es posible crear ese compendio a través de un parámetro D, el cual cuantifica la interacción entre 2 dominios. Luego, es posible definir nuevas interacción basadas en los dominios de las proteínas y clasificándolas a través del parámetro D.

La obtención del LR difiere para cada base de datos, pues para calcularlo se redefine en función de la evidencia utilizadas, en este caso las clases:

$$P(evidencia|Positivo) = P(Clase|GSP)$$

Es decir, la probabilidad que una interacción caiga en la clase, tal que pertenezcan al GSP -mismo caso para el GSN-. Esta probabilidad se puede calcular definiendo $P(Clase)$ como el cociente entre las interacciones pertenecientes a la clase y el total de interacciones posibles. Este conjunto "posibles" puede ser el punto cruz (cruza entre todos los elementos del conjunto) entre todos lo elementos de la red (menos las interaccionen consigo mismo y las relaciones inversas, que para este caso da lo mismo), lo que ejemplificado gráficamente se podría definir como se muestra en la Figura I.8:

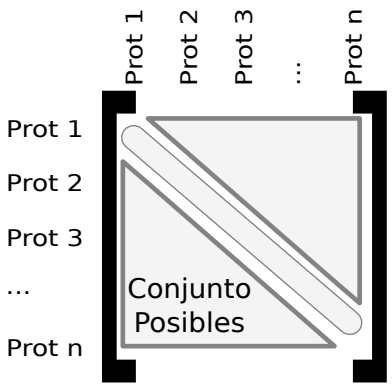

Figura I.8: Diagrama del conjunto "posibles"

Para integrar la condición "tal que pertenezca al GSP", se debe realizar una intersección entre los elementos de la clase y el GSP, así sólo se escogerán los elementos de la clase que a la vez pertenezcan al GSP. Para luego simplemente calcular el cocientes antes señalado.

Con ello se obtendrá una red de interacciones interproteicas, las cuales estarán cuantificadas por el parámetro estadístico LR. Este resultado debe ser agregado con las demás redes obtenidas en diferentes estudios.

Agregación de redes de interacción

Las redes resultantes del trabajo presente y los trabajos de Rhodes *et al.*y Xia *et al.* fueron integradas para definir las interacciones pertenecientes al NRC/MASC, la única condición para que una para de proteínas estuviera allí es que su O_{post} fuera mayor a 1. Esto por considerarse condición básica de un 50 % de certeza de que el par interaccionase.

I.4. Objetivos

El objetivo principal de este trabajo de memoria es realizar un modelo novedoso de la unidad NRC/MASC, utilizando para ello la inferencia de interacciones interproteicas del ser humano. Se espera lograr con esto un nuevo enfoque en la comprensión de la organización sistémica de la unidad y la relación de ésta con la aparición de enfermedades neurológicas.

Por otro lado los objetivos secundarios son:

- Obtención de la red de PPIs utilizando el clasificador Naîve-Bayes, utilizando para ello bases de datos diversas.
- Comparación y agregación de la red de PPIs obtenidas con redes inferidas por otros autores.

- Definición de las relaciones al interior de la unidad NRC/MASC y comparación con lo obtenido a través de la literatura.
- Clustering de la unidad NRC/MASC en función de sus relaciones interproteicas, inferencia de funcionalidades de cada cluster.
- Estudiar aplicaciones de modificaciones en los componentes de la unidad por existencia de diversas enfermedades, para obtener conclusiones sobre el rol jugado por cada componente en presencia de las enfermedades

Capítulo II

Metodología

A continuación se señalan los distintos protocolos ocupados en las secciones de este proyecto:

II.1. Inferencia de PPIs

El protocolo de la primera etapa del proyecto -la inferencia de las interacciones interproteicas- fue realizado utilizando como antecedente el trabajo de Rhodes *et al.* [30], el cual a su vez se basa en lo presentado por Jansen *et al.* [16], en donde se presenta por primera vez la idea de utilizar el clasificador Naïve-Bayes para inferir PPIs.

Tal como se señaló en los antecedentes (ver Tabla I.1) el tipo de interacción interproteica inferida por el algortimo de Naïve-Bayes es amplia. Señalando con esto que no sólo se considerarán pares de proteínas que interaccionan físicamente de manera directa, sino pares que interacciones incluso a través de terceros o en co-complejos.

II.1.1. Protocolo General

En los antecedentes se comentó la lógica detrás de la inferencia de PPIs. En particular el protocolo general para realizarlo se puede observar en la Figura II.1:

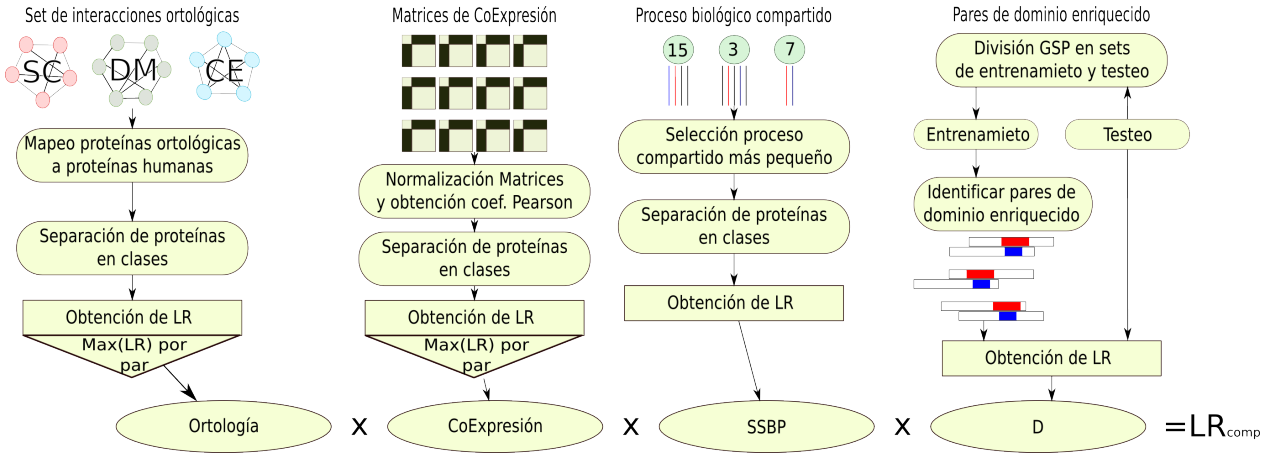

Figura II.1: Diagrama general de inferencia de PPIs

En el diagrama expuesto se puede observar los pasos generales para cada una de las implementaciones de las distintas bases de datos utilizadas. A continuación, y en forma somera, se presentan los pasos a seguir para el procesamiento de las bases de datos:

Interacciones Ortológicas. Luego de la descarga e implementación de los distintos sets de datos -cada uno representando las interacciones en una especie- se procedió a mapear las interacciones a proteínas humanas. A la vez en que se definieron y calcularon los distintos parámetros utilizados para jerarquizar -separar en clases- los sets de datos recientemente creados. Finalmente y gracias las clases generadas, es posible calcular el likelihood ratio para cada par de proteínas y luego unir los 3 sets de datos creados. Como existe la probabilidad de que una interacción esté en más de un set, se seleccionada el valor máximo de LR entre todos los set para un par, creando así la base de datos de interacción humanas obtenidas por ortología.

Matrices de CoExpresión. El primer paso para su implementación es la normalización de los microarrays, así se eliminan entradas poco específicas o desconocidas -genes no identificados o sequencias con muchos blancos-. La obtención del coeficiente de Pearson es el parámetro que permitirá realizar la separación de cada set de datos en clases. Luego de ello se integraron los set de datos resultantes de la misma forma que para el caso de las interacciones ortológicas, es decir, seleccionando el mayor LR.

Procesos Biológicos Compartidos. La selección del proceso biológico compartido más pequeño se logró definiendo todos los procesos compartidos por cada par de proteínas, para luego seleccionar aquel proceso con menos integrantes. Luego cada par fue parametrizado a través del coeficiente SSBP, el cual denota el tamaño de elementos del proceso biológico compartido más pequeño. Así, con el parámetro jerarquizante ya calculado, es posible crear las clases dentro del set de datos, desde donde se calcula de forma inmediata -sin necesidad de agregar distintos sets como en los casos anteriores- el LR de los pares de proteínas.

Pares de Dominio Enriquecidos. El protocolo de este set de datos se dividió en dos partes. La primera fue la deducción de los pares de dominio en conjunto con el parámetro D -cuantificación de la probabilidad de unión a través de los dominios- para luego estratificar a través de este parámetro.

Para lograr obtener el LR de cada clase fue necesario testear los resultados con el GSP y GSN. Como existe una fuerte probabilidad de intersección entre set obtenidos y la GSP, pues de allí se obtiene los pares de dominios enriquecidos, se divide la GSP en 3 partes, definiéndose así 3 sets de datos, cada uno como la suma de 2 de aquellos tercios y sirviendo el tercero como set de testeo. Con esto se logró eliminar la posibilidad de traslape entre los sets.

Finalmente la integración de los set de datos calculados se obtuvo simplemente multiplicando los LR obtenidas para cada par de proteína en cada etapa. Esto gracias a la hipótesis de Naïve-Bayes presentada en la Introducción.

Además, se debe señalar que las bases de datos utilizadas y desarrolladas fueron implementadas en PostgreSQL, mientras que para la programación numérica fue llevada a cabo mediante el software R-project. Los resultados de estos procedimientos se pueden encontrar íntegros en los anexos de la memoria.

II.1.2. GSP y GSN

Para desarrollar la aplicación del clasificador de Naïve Bayes en la inferencia de PPIs fue necesario definir en primer lugar 2 conjuntos estándar:

GSP: *Gold Standar Positive*, se refiere a aquel conjunto de PPIs que se tiene certeza, a través de evidencia experimental, que interaccionan.

GSN: *Gold Standar Negative*, es el conjunto de PPIs que se sabe con certeza que no interactuan.

El GSP fue obtenido de la base de datos HPRD [26], del sitio se descargó la base actualizada[1] de PPIs conocidas a través de la literatura. Por otro lado, y tal como se señaló en la Introducción, el GSN fue creado a través del cruce entre las proteínas pertenecientes a la membrana plasmática y al nucleo. Para llevar a cabo la definición de la GSN se descargó desde la página del proyecto GO [23] el listado de proteínas[2] que pertenecieran al nucleo: `GO:0005634` y a la membrana plasmática: `GO:0005886`. Luego se llevó a cabo un cruce entre las proteínas pertenecen a cada set de datos, para asó definir los pares proteicos que integran la GSN, siendo cada par integrado por una proteína del núcleo y otra de la membrana plasmática.

[1] 6 de Junio del 2009

[2] Fueron obtenidas con fecha Agosto 2009

Se debe señalar que cada una de estas bases fue normalizada, así sólo se consideraron los elementos de la GSN que tuviesen el *id* Swissprot; para el caso de la GSN se consideró únicamente aquellos elementos que tuvieran evidencia distinta al IEA[3].

II.1.3. Interacciones Ortológicas

Para la inferencia de las interacciones interproteicas por ortología se obtuvieron en primer lugar las bases de datos de las relaciones interproteicas en 3 especies diferentes al humano. Así es como se bajaron las bases de datos correspondientes a *Saccharomyces cerevisiae*, *Caenorhabditis elegans* y *Drosophila melanogaster*. Todas ellas fueron obtenidas a través de la base de datos DIP [38].

En el caso de la *Saccharomyces cerevisiae* se descargó la base de datos actualizada[4] para esta especie, la cual integra los distintos set de datos entregados por diferentes estudios; por otro lado, para *C. elegans* y *D. melanogaster* se utilizaron bases de datos desactualizadas pertenecientes a trabajos particulares. En el caso de *C. Elegans* se utilizó la base de datos de Li *et al* [20] y para *D. melanogaster* se ocupó la de Giot *et al* [8]. Ambas bases de datos son *high throughput datasets*, es decir, entregan las relaciones interproteicas para todo el genoma y se ocuparon por la necesidad de utilizar parámetros existentes exclusivamente en aquellos set de datos.

El siguiente paso para conseguir las interacciones probables por ortología es obtener los archivos de mapeo ortológicos, los cuales fueron descargados del sitio Inparanoid [34], el cual entregó las relaciones ortológicas entre una proteína humana y sus posibles pares en otra especie. Cabe destacar que se entrega además un puntaje para cada par ortológico, el cual en esta ocasión no fue utilizado en la inferencia, pero se espera que en el futuro sea considerado.

Así con un par de proteínas interactuantes (i, j) en la proteómica de las especies *Saccharomyces cerevisiae*, *Caenorhabditis elegans* y *Drosophila melanogaster* se lleva a cabo el mapeo ortológico para cada proteína, tal que, $i \longrightarrow i^*$ y $j \longrightarrow j^*$. En donde, tanto i^* como j^* se encuentran en la proteómica del *Homo Sapiens*, resultando de ello la interacción putativa (i^*, j^*).

Se debe señalar que existió un problema con los identificadores utilizados, pues en el caso de la base de datos Inparanoid, ésta utiliza los *ids* de Ensembl [7], mientras que las demás bases de datos utilizan los *ids* Swissprot/Uniprot, lo que requirió mapear las proteínas entregadas por Inparanoid. El efecto de ello fue la ampliación de las PPIs, ya que una proteína en Ensembl puede mapear a varias en Swissprot/Uniprot [37], lo que significó un mayor número de PPIs obtenidas. El mapeo se realizó a través del sitio Biomart [25] el que permite un mapeo diverso.

Luego del mapeo desde las proteínas de los pares proteicos para cada especie a proteínas humanas, se generaron distintas clases para cada set de datos, así siguiendo la metodología

[3]Inferred from Electronic Annotation, evidencia no certera
[4]14 de Octubre del 2008

propuesta por Rhodes *et al.* se consideraron distintos criterios para cada especie:

***S. cerevisiae*:**

$$Dataset \begin{cases} E_V > 1 \longmapsto \text{Clase } I \\ E_V = 1 \begin{cases} N = 1 \longmapsto \text{Clase } II \\ N \in]1, 28[\longmapsto \text{Clase } III \\ N > 28 \longmapsto \text{Clase } IV \end{cases} \end{cases}$$

Con E_V: número de líneas de evidencia independientes de una PPI, es decir, el número de publicaciones de donde se obtiene esta interacción. El número de publicaciones se obtiene de la columna "publication id" entregado por la base de datos de PPIs para *S. Cerevisiae*.

Y N el número de PPIs humanas que entrega una PPI de la especie inicial; se obtiene a través del identificador de PPIs en la especie: "interaction id".

Al existir un solapamiento entre las distintas clases -ciertos PPIs pertenecientes a más de una clase, esto como resultado del proceso de mapeo, a través del cual 2 PPIs diferentes pueden mapear al mismo PPI final- se decidió eliminar las repeticiones pertenecientes a las clases mayores, así si una PPI se presentaba en las clases I,II y IV, se eliminaba de las clases II y IV.

***D. melanogaster*:**

$$Dataset \begin{cases} Conf \geq 0.55 \longmapsto \text{Clase } I \\ Conf < 0.55 \longmapsto \text{Clase } II \end{cases}$$ Siendo el $Conf$[5] un parámetro de confianza entregado por el estudio antes citado. En este caso, similar al anterior, se eliminó las repeticiones de la clase mayor (clase II).

***C. elegans*:** Para este caso no se llevó a cabo ninguna separación, principalmente por el pequeño tamaño final del set de datos.

Para poder aplicar finalmente el clasificador de Naïve-Bayes a las clases obtenidas, se debió definir el conjunto "posibles"para el set de datos de cada especies. En este caso se consideraron como "posibles"aquellas PPIs resultantes de todas las combinaciones binarias entre las proteínas pertenecientes a cada set de datos que se pudieron hacer.
Para obtener los elementos de cada clase perteneciente al GSP o GSN, se intersectó cada una de las clases con el GSP o GSN, obteniéndose así todos los parámetros necesarios para entregar los LR de cada clase, los cuales se presentan con detalle en Resultados.

II.1.4. Matrices de Coexpresión

El primer paso para realizar la inferencia de PPIs a través de las matrices de coexpresión fue la selección de los estudios de *microarray* según Rhodes *et al.* que entregaron los resultados más significativos. El identificador de los sets de datos es:

[5] *Confidence Value*

- Chen *et al.* - Liver [3]
- Rosenwald *et al.* - Lymphoma [31]
- Segal *et al.* - Sarcoma [33]
- Su *et al.* - MultiCancer [35]
- VantVeer *et al.* - Breast [39]

De todos estos estudios se utilizaron los 4 últimos, pues no fue posible obtener los materiales suplementarios para Chen *et al.* Luego, el resto de los set de datos fueron buscados dentro de la base de dato Oncomine [6], desde donde se pueden realizar diversas consultas sobre los resultados entregados por estos experimentos. Además de lo anterior, allí se entrega la localización del material suplementario, el cual fue descargado para llevar a cabo la implementación.

Una vez obtenidos los materiales se realizó la normalización de las matrices de expresión. Esto consistió de los siguientes pasos generales:

1. Identificar las muestras que tuviesen un *p-value* mayor al aceptado (0.005) como `NaN`[6].
2. Se eliminaron todos los sondas que tuviera un 50 % o más de `NaN`.
3. Agregar todas las copias de una misma sonda.
4. Eliminar aquellas sondas que apunten a genes desconocidos o poco confiables, es decir, que apunten a secuencias que no han sido definidas o que apunte a un gran número de ellas y por lo tanto no se pueda realizar una relación directa entre la sonda y un gen en particular.

Las dos primeras etapas fueron realizadas en R-project, mientras que el estudio y selección por identificador fue llevado a cabo en PostgreSQL. Además cada uno de estos pasos varió fuertemente para los distintos set de datos, razón por la cual se detallará el procedimiento ocupado en el procesamiento de cada estudio en la sección siguiente.

Luego de normalizar las distintas matrices, se calculó la matriz de correlación del set de datos definido. Así, para cada par de genes (i, j) se calcula el coeficiente de Pearson ,$r := cor(i, j)$, de ellos. Ello a través de la función `cor` perteneciente al paquete `stats`, implementado en R. Obteniéndose así, una matriz triangular inferior (pues la diagonal es la corelaccion de un gen con sí mismo y la matriz triangular superior es igual a la inferior) con los coeficientes de todas las combinaciones de genes presentes.

Se exportó la matriz en forma de tabla, para su ingreso inmediato a PostgreSQL, en donde se realizaron los siguientes procedimientos: en primer lugar se dividieron las relaciones de correlación en función del coeficiente de Pearson, esto para definir las clases del set de datos.

[6]Not A Number

Al mismo tiempo se mapeo desde el identificador utilizado por los estudios -generalmente UniGene- a SwissProt/UniProt, lo cual modificó el número de relaciones.

Las clases definidas en la división de la tabla se presentan de la siguiente forma:

$$
Dataset \begin{cases}
\mathrm{R} \geq -1 & \longmapsto \text{Clase } I \\
\mathrm{R} \geq -0.9 & \longmapsto \text{Clase } II \\
\mathrm{R} \geq -0.8 & \longmapsto \text{Clase } III \\
\mathrm{R} \geq -0.7 & \longmapsto \text{Clase } IV \\
\mathrm{R} \geq -0.6 & \longmapsto \text{Clase } V \\
\mathrm{R} \geq -0.5 & \longmapsto \text{Clase } VI \\
\mathrm{R} \geq -0.4 & \longmapsto \text{Clase } VII \\
\mathrm{R} \geq -0.3 & \longmapsto \text{Clase } VIII \\
\mathrm{R} \geq -0.2 & \longmapsto \text{Clase } IX \\
\mathrm{R} \geq -0.1 & \longmapsto \text{Clase } X \\
\mathrm{R} \geq 0 & \longmapsto \text{Clase } XI \\
\mathrm{R} \geq 0.1 & \longmapsto \text{Clase } XII \\
\mathrm{R} \geq 0.2 & \longmapsto \text{Clase } XIII \\
\mathrm{R} \geq 0.3 & \longmapsto \text{Clase } XIV \\
\mathrm{R} \geq 0.4 & \longmapsto \text{Clase } XV \\
\mathrm{R} \geq 0.5 & \longmapsto \text{Clase } XVI \\
\mathrm{R} \geq 0.6 & \longmapsto \text{Clase } XVII \\
\mathrm{R} \geq 0.7 & \longmapsto \text{Clase } XVIII \\
\mathrm{R} \geq 0.8 & \longmapsto \text{Clase } XIX \\
\mathrm{R} \geq 0.9 & \longmapsto \text{Clase } XX \\
\mathrm{R} \geq 1 & \longmapsto \text{Clase } XXI
\end{cases}
$$

Con ello definido se calcularon los parámetro últimos (Pr(R|GSP) y Pr(R|GSN)) con el fin de obtener el parámetro LR para cada clase.

En este caso el conjunto "posible" -necesario para el cálculo de las probabilidades- está definido por la suma de los pares que pertenecen a cada clase, esto porque el set de datos es un cruce entre todos los genes, generando así todos los pares posibles. Luego se calculó el número de pares que intersectaban con el GSP o GSN para cada clase, tal como se mencionó en los antecedentes, teniendo así todos los parámetros requeridos.

A continuación se detalla las modificaciones particulares del protocolo para cada set de datos utilizados en esta fase.

Rosenwal – Liver El set de datos no tuvo particularidades en la eliminación de sondas por la existencia de una gran cantidad de `NaN`, ésta fue llevada a cabo de inmediato.

Por otro lado, la eliminación de los nombres de los genes se realizó seleccionando aquellas sondas que no tuvieran un identificador EMBL, por ser este un identificador conocido y completo. Siendo posteriormente mapeados al identificador SwissProt/Uniprot.

Finalmente se agregaron aquellas copias del mismo probe, para luego aplicar las etapas de obtención del coeficiente de correlación y separación de clases.

Segal – Sarcoma En este caso se realizó una normalización de los set de datos entregados, pues el set de datos referente a las muestras con sarcoma tenía muestras adicionales al set de datos con melanoma. Luego de eliminar las entradas adicionales, se procedió a procesar los datos presentes en el set. Se debe aclarar que este set de datos, además de incluir el valor de la señal presenta también el *p-value* y una bandera (*call*) con los posibles valores *present/absent*. Así se eliminaron aquellas señales -marcándolos como `NaN`- cuyos *p-value* fueran mayores a 0.05 y que su bandera fuera *absent*. Luego de lo cual se eliminaron aquellos genes que tuvieran un 50 % o más de `NaN`.

Por otro lado, se eliminaron aquellos genes cuyos identificadores señalaran una sonda inexacta o que la señal obtenida no fuera posible de identificar con un solo transcripto [10], además de eliminar todos los genes cuyo identificador no perteneciera a EMBL. Luego, se llevaron a cabo los pasos de agregación y obtención de la matriz de correlación sin ninguna variación en particular.

Su – MultiCancer El protocolo aplicado a este set de datos fue obtenido del trabajo de Su *et al.* [35], en donde se señala que sólo se aceptaron las sondas que tuvieran a lo menos una señal mayor o igual a 200, lo cual fue la primera condición de eliminación de genes realizada. Además, se aplicó la función logaritmo natural sobre el set de datos, esto para normalizar los valores, tal como se señala en el *paper* de Su *et al.*.

Al ser el microarray de Su *et al.* el mismo utilizado por Segal *et al.*, un microarray Affimetrix U95, se llevó a cabo el mismo procedimiento para eliminar las sondas con identificadores incorrectos o transcritos múltiples.

Vant Veer – Breast Cancer El primer paso llevado en este caso fue la eliminación de las sondas que tuviesen por *Gene description*[7]: ESTs[8], Contig o NM, pues no enlazaban a un gen conocido. Además se eliminaron aquellas sondas que no tuvieran un nombre para la sequencia blanco.

De forma parecida a lo realizado para Su *et al.* se aplicó un par de condiciones para aceptar las señales de una sonda, siendo las condiciones especificadas en los trabajos. En el caso del set de datos de Vant Veer *et al.* se estipuló que las muestras de los genes debiera haber por lo menos 5 *p-values* menores a 0.01 entre las muestras de cada gen y por lo menos la señal fuese el doble[9] de diferencia entre la muestra del tejido normal y del modificado. Luego se utilizó el protocolo estándar expresado con anterioridad.

[7]Última columna del set de datos entregados por Vant Veer *et al.*

[8]Expressed sequence tag.

[9]Se consideró el doble o la mitad.

Agregación Finalmente se debe señalar que el procedimiento de agregación fue realizado a través de dos posibilidades:

- Si el número de copias era menor o igual a 4, simplemente se obtenía la media para cada muestra.
- Si el número de copias era mayor a 4 se aplicaba el algoritmo de *bootstraping*[10] para promedios.

Esta diferencia se debió a que con una población de 4 o menos el *bootstraping* no queda distribuido normalmente, esto por el bajo número de posibilidades de combinaciones existentes.

II.1.5. Funciones biológicas compartidas

El primer paso llevado a cabo para desarrollar esta inferencia fue obtener la base datos ontológica (*Gene Ontology* [23]) para toda la proteómica del ser humano. En ella se señalan las asignaciones de las proteínas a los diversos procesos biológicos; en este caso, y a diferencia de lo realizado en la definición de la GSN, se utilizarán aquellas asignaciones provenientes de la evidencia IEA, esto para aumentar la diversidad de interacciones a trabajar.

Luego y tal como se señaló en los antecedentes del proyecto, la inferencia está dividida en 3 fases:

Identificación procesos compartidos: La identificación se realizó haciendo un cruce entre todas las proteínas representadas en la base de datos de GO, teniendo como restricción de la selección el que cada par de proteínas debiesen compartir un proceso biológico. Además se eliminaron las interacciones de las proteínas consigo mismas.

Número de participantes en los procesos: Se llevó a cabo una contabilización de las proteínas que están asignados a cada proceso biológico desde la base de datos de GO.

Identificar el proceso con menos participantes para cada par: Para realizar la selección se cruzaron las tablas creadas anteriormente. Así se logró crear una tabla en donde para cada par de proteínas se tenía, además del proceso biológico en que participaban, el número de participantes totales. Finalmente se seleccionó el proceso con menos participantes para cada par de proteínas.

Luego de obtener el SSBP (ver Antecedente I.3.3)para cada par de proteínas, éstas fueron separadas en diversas clases, esto según el tamaño del SSBP. Así las clases que se utilizaron fueron:

[10]El algoritmo se muestra en Apéndice A

$$Dataset \begin{cases} \text{SSBP} \leq 5 & \longmapsto \text{Clase } I \\ \text{SSBP} \leq 10 & \longmapsto \text{Clase } II \\ \text{SSBP} \leq 50 & \longmapsto \text{Clase } III \\ \text{SSBP} \leq 100 & \longmapsto \text{Clase } IV \\ \text{SSBP} \leq 500 & \longmapsto \text{Clase } V \\ \text{SSBP} \leq 1000 & \longmapsto \text{Clase } VI \\ \text{SSBP} \leq 2000 & \longmapsto \text{Clase } VII \end{cases}$$

Se debe señalar además que cada PPI pertenece a una sola clase, las clases mostradas tienen como límite superior el valor que se señala -$\leq$ 500 por ejemplo- y como límite inferior el inicio de la clase siguiente. Al igual que en las demás secciones de la inferencia, en ésta se definió el conjunto de "posibles"como aquella combinación binaria entre todos las proteínas pertenecientes al set de datos final.

Así, luego de definir el conjunto, se realizó el cruce de tablas para obtener aquellos elementos de las distintas clases que pertenecieran ademas al GSP y GSN, siendo éstos los parámetros finales para la obtención del LR de cada clase.

II.1.6. Pares de dominios enriquecidos

En la última fase de inferencia de las PPIs fue necesario en primer lugar obtener la base de datos INTERPRO, la cual fue obtenida de la aplicación BioMart; ésta entregó los distintos dominios y familias para las proteínas pertenecientes al GSP. Luego se dividió la base de datos GSP en 3 partes de igual tamaño, esto para aplicar la inferencia -por pares de dominios enriquecidos ya explicada en los antecedentes.

De allí se definieron 3 sets de datos cada uno como la suma de 2 de los 3 tercios de la GSP, lo cual se muestra en la Tabla II.1:

Tabla II.1: Definición de los sets de datos para la inferencia por pares de dominios enriquecidos

Dataset	Secciones
1	GSP_1 y GSP_2
2	GSP_1 y GSP_3
3	GSP_2 y GSP_3

Para cada uno de los sets de datos se obtuvo el parámetro D[11], definido anteriormente por Rhodes *et al.* [30], el cual está definidos por:

$$D = \frac{Pr(d_i : d_j|GSP)}{Pr(d_i|GSP) \cdot Pr(d_j|GSP)}$$

[11]Domain Enrichment Ratio

$$d_i : d_j \geq 3$$

En donde d_i y d_j son 2 dominios proteicos, $d_i : d_j$ es una PPI, en donde una proteína tiene el dominio d_i y la otra el dominio d_j. Además se utiliza la restricción propuesta por Rhodes *et al.*, en donde el número de PPIs en que los dominios estén presentes debe ser mayor o igual a 3.

Este parámetro fue diseñado de tal manera que mida la coocurrencia de d_i y d_j. Así si d_i y d_j ocurren de manera azarosa el valor de D tenderá a 1 (*ie.* $Pr(d_i : d_j|GSP) \approx Pr(d_i|GSP) \cdot Pr(d_j|GSP)$), mientras que si ellos están relacionados y por lo tanto coocurren el valor de D será grande (*ie.* $Pr(d_i : d_j|GSP) \gg Pr(d_i|GSP) \cdot Pr(d_j|GSP)$).

Para obtener el parámetro D fue necesario en primer lugar mapear todas las proteínas interactuantes de cada set de datos con sus dominios y familias. Luego, se calculó D para cada par $d_i : d_j$ presente en el set de datos. Una vez realizado lo anterior se procedió a mapear el set de datos obtenido, pasando de dominios y familias a proteínas.

Se debe señalar que fue muy común encontrar un PPI específico con más de un parámetro D, por las diversas combinaciones de sus dominios y/o familias. Ante esto se seleccionó el parámetro D mayor para cada PPI.

El paso siguiente fue definir las clases en que se separó cada set de datos obtenido, en particular las clases fueron definidas según el parámetro D y éstas fueron:

$$Dataset \begin{cases} \text{D} \geq 2 & \longmapsto \text{Clase } I \\ \text{D} \geq 5 & \longmapsto \text{Clase } II \\ \text{D} \geq 10 & \longmapsto \text{Clase } III \end{cases}$$

Luego de definir las clases para cada set de datos se testeó los D encontrados comparándolos con el tercio de GSP no utilizado. Además de la comparación contra el tercio restante se comparó con el GSN, para poder obtener los parámetros necesarios para calcular el LR de cada clase.

Se debe señalar finalmente que las intersecciones de cada set de datos con la GSP y GSN fueron agregadas al final, es decir, se ponderaron los valores obtenidos para cada PPI, lográndose así los resultados se que se mostrarán en la sección siguiente.

II.1.7. Unión de set de datos

Luego de obtener el LR para cada clase de los distintos set de datos, se procedió a unir los sets para formar una única red de PPIs. Para realizar esto se integraron todos los pares a un solo set, mientras que para aquellos pares de PPIs que aparecieran más de una vez, por haber pertenecido a más de un set de datos, se calculó su LR_{comp}. Esto permite lograr un set de datos con pares únicos y que representen los LR obtenidos en todas las instancias de

inferencia.

Esto permite lograr un set de datos con pares únicos y que representen los LR obtenidos en todas las instancias de inferencia.

II.2. Comparación y unión de redes de PPIs

El set de datos producido a través de la inferencia de interacciones de este proyecto fue designado como el set de Campos *et al.*, esto para diferenciarlo de los otros sets entregados por 2 estudios distintos: el de Rhodes *et al.* [30] y el de Xia *et al.* [40].

Mientras que el set de datos de este estudio utiliza los identificadores SwissProt/Uniprot, para los otros 2 sets se usaron los identificadores UniGene, razón por la cual -y sólo para la comparación de los sets- se mapeo el set de Campos *et al.* a UniGene, pudiendo así comparar los distintos set de datos.

Los 3 sets fueron implementados en PostgreSQL, desde donde se pudo importar las intersecciones y excepciones entre ellos a R-project, desde allí se realizó el análisis estadístico que se señala en Resultados.

En último lugar se agregaron los 3 sets para estudiar la amplitud de estos resultados, llevándose a cabo de idéntica forma a lo realizado con la unión de las distintas fuentes de inferencia, es decir, definiendo el parámetro LR_{comp}. Con la única reserva que sólo se consideraron las interacciones que tuviera un O_{post} mayor a 1.

II.3. Nueva red de PPIs del complejo NRC/MASC

Para definir la nueva red de pares proteicos del complejo NRC/MASC, se utilizó el trabajo de Collins *et al.* [5], en donde se definió las proteínas que integraban el complejo. Así, y realizando una selección de aquellos pares de proteínas en donde ambas pertenecían al complejo, fue posible obtener las PPIs inferidas por este proyecto y los trabajos de Rhodes *et al.* y Xia *et al.* que integraban la red del NRC/MASC.

El material suplementario utilizado para la definición de los sets de datos fueron los entregados por Pocklington *et al.* [28], esto por la accesibilidad de éste y por contener los resultados obtenidos por Collins *et al.*. El material suplementario de Pocklington *et al.*incluyó además los resultados obtenidos por ellos y que fueron importantes en la segunda fase del proyecto.

El conjunto de proteínas pertenecientes al NRC/MASC fue identificadas por Pocklington *et al.* con el *id* SwissProt/UniProt, ante lo cual y para disminuir el error fueron mapeadas a UniGene, para poder realizar la consulta a la red entregada por Rhodes *et al.* y por Xia *et al.*. Por otro lado, para seleccionar las interacciones entregadas por este trabajo se utilizó el

identificador SwissProt/UniProt.

Algunas particularidades de este proceso fueron:

- No se encontró identificador en UniGene para la proteína Neurofilament triplet M[12], razón por la cual no fue posible encontrar PPIs que ella integrara en los resultados de Rhodes *et al.* y Xia*et al.*. Por otra parte, sí fue posible encontrar una interacción en el set de datos obtenido por este proyecto.
- La proteína Calmodulin[13] tiene 3 genes asociados y por lo tanto 3 identificadores UniGene:
  - CAM1
  - CAM2
  - CAM3

  Al encontrarse interacciones de un gen con los 3 genes de la Calmodulin se decidió escoger la interacción mayor (LR más grande), ya que se consideró a la Calmodulin como una unidad. Esto mismo ocurrió con la proteína HSPA1, la cual tiene 2 genes asociados, aplicándose el mismo proceso que para la Calmodulin.
- Dentro de las proteínas pertenecientes al NRC/MASC se señalan 5 proteínas compuestas, las cuales debieron ser identificadas posteriormente de forma manual, pues la inferencia sólo incluye proteínas, no agregados.

Con todo esto se llevó a cabo la creación de la nueva red de PPIs del complejo NRC/MASC.

II.3.1. Clustering de la red del complejo NRC/MASC

El clustering de la red putitativa de la unidad NRC/MASC fue segmentada a través del uso del algoritmo de *clustering* desarrollado por Newman y Girvan [24], el cual fue descrito en I.3.2.

El algoritmo ha sido implementado en el paquete **iGraph** para R-project, siendo factible su uso de forma inmediata y sin mayores manipulaciones. Tal como se señaló en los antecedentes, el algoritmo entrega un parámetro cuantificador Q, el cual fue utilizado para optimizar el *clustering* de la red. Se diseñó un programa que seleccionara el punto de corte del grafo jerárquico de tal forma que maximizara el valor de Q, siendo el programa incluido en Apéndice B .

[12]Uniprot: P07197
[13]Uniprot:P62158

II.3.2. Estudio y modelamiento de la red

Una vez obtenidos los distintos módulos que conforman el complejo, se procedió a investigar las proteínas que los componen y su posible función biólogica. Para ello se realizó una búsqueda en la literatura pertinente [1], así como en diversos estudios [27, 28].

Por otro lado para el estudio de las modificaciones producidas por las enfermedades y las proteínas vitales para la plasticidad sináptica y del comportamiento, se utilizó la base de datos curada por Pocklington *et al.* [28], siendo ella realizada a través de la minería en la literatura pertinente. En donde se recopiló una gran cantidad de antecedentes que relacionan modificación de una proteína (*knockout*/ mutación) con la aparición de una enfermedad.

Capítulo III

Resultados

Se presentan a continuación los resultados obtenidos en las distintas fases del proyecto:

III.1. Inferencia de PPIs

III.1.1. GSP y GSN

Para el GSP se obtuvó los siguientes elementos, presentados en la Tabla III.1:

Tabla III.1: Elementos pertenecientes al GSP

Número de proteínas	8482
Número PPIs	33865

El número de proteinas pertenecientes a la membrana plasmática y el núcleo se señalan en la Tabla III.2:

Tabla III.2: Elementos pertenecientes la membrana plasmática y el núcleo

Proteínas membrana plasmática	2024
Proteínas núcleo	2665

Antes de realizar el cruce de las proteínas se eliminaron 175 proteínas que se hallaron en ambas unidades, además se encontró que 1262 pares de proteínas estaban en la GSN, por lo cual fueron eliminados.

Con lo que finalmente se obtuvo el GSN compuesto de los elementos señalados en la Tabla III.3:

Tabla III.3: Elementos pertenecientes al GSN

Número de proteínas	4339
Número PPIs	4602748

III.1.2. Interacciones Ortológicas

De las bases de datos para las interacciones de cada especie se obtuvieron los siguientes números de proteínas e interacciones, mostrados en la Tabla III.4:

Tabla III.4: Número de proteínas e interacciones para cada especie

Especie	Número PPIs	Número de Proteínas
S. Cerevisiae	17570	8159
C. Elegans	2421	2064
D. Melanogaster	440	863

Luego del mapeo por ortología se obtuvieron las siguientes cantidades de interacciones y PPIs humanas, presentadas en la Tabla III.5:

Tabla III.5: Número de proteínas y PPIs humanas como resultado del mapeo para cada especie

Especie	Clase	Número PPIs	Número de Proteínas
S. Cerevisiae	Todos	15425	2599
	I	2436	1263
	II	2168	955
	III	2935	409
	IV	7886	1919
D. Melanogaster	Todos	462	299
	I	89	102
	II	373	247
C. Elegans	Todos	1005	637

El detalle de las tablas correspondientes se pueden ver en Apéndice C .1, por otro lado, se presenta a continuación los *likelihood ratio* (LR) obtenidos para las distintas clases de cada especie y el LR para todo el set de datos de cada una de las especies.

En la Figura III.1 se muestra el LR para cada clase (según lo definido en la sección II.1.3) obtenida para el set de datos de *S. cerevisiae*, además del LR calculado para el total del set (columna 'Todos').

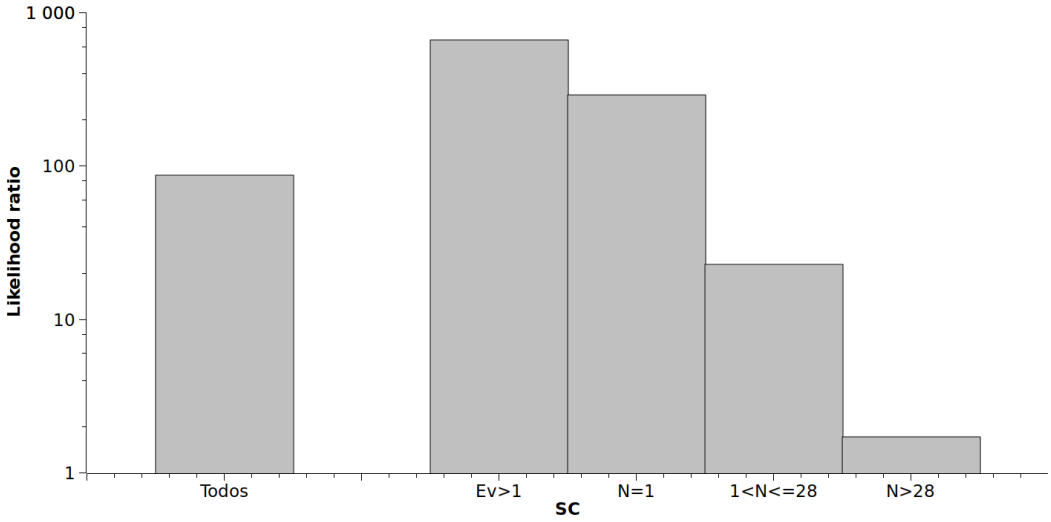

Figura III.1: Likelihood ratio para clases de *S. cerevisiae*

Se presenta en la Figura III.2 el LR calculado para cada clase y para el set total (columna 'Todos') perteneciente al set de datos de *D. melanogastar*.

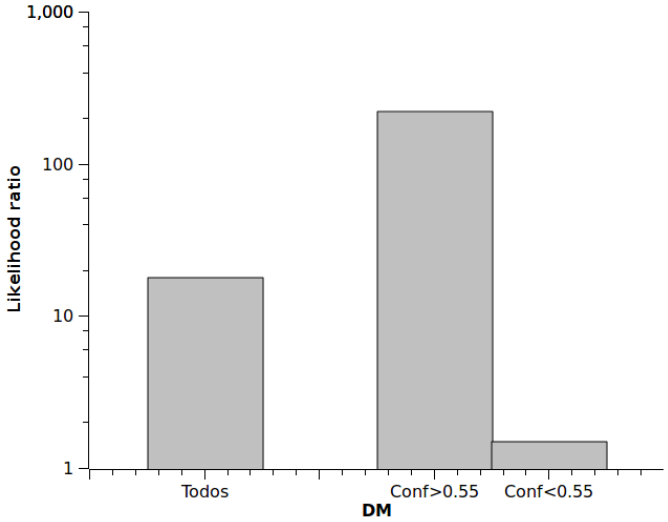

Figura III.2: Likelihood ratio para clases de *D. melanogastar*

Finalmente en la Figura III.3 se presenta el LR para todo el set de datos de *C. elegans*.

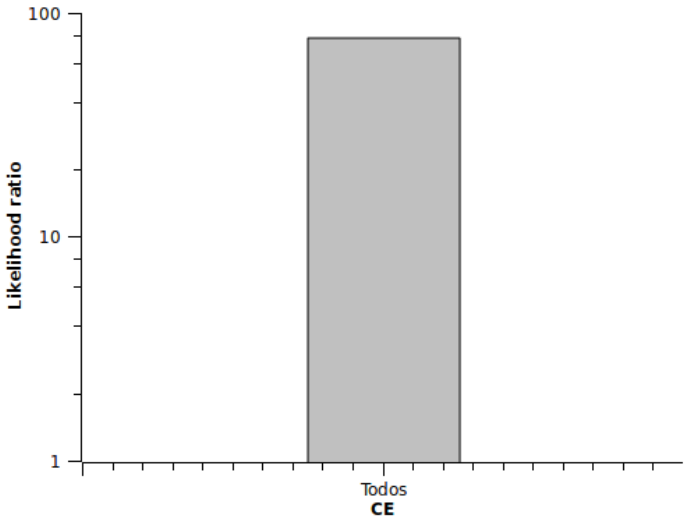

Figura III.3: Likelihood ratio para clases de *C. elegans*

III.1.3. Matrices de coexpresión

De los set de datos entregados por los estudios señalados anteriormente se muestra en la Tabla III.6 el número de elementos en ellos y los finalmente entregados por este trabajo a través del mapeo y aplicación de las restricciones mencionadas en Metodología:

Tabla III.6: Elementos entregados por los estudios de *microarray* y resultados obtenidos por mapeo

	Datos entrada		Resultados	
Set de datos	Muestras	Sondas	Genes	Pares
Rosenwald - Lynphoma	293	7399	2181	2377290
Segal - Sarcoma	81	12533	3609	4706136
Su - MultiCancer	174	12627	4986	12427605
Vant Veer - Breast	117	24186	2063	2126953

Luego fue posible realizar la separación en las distintas clases y calcular el LR de cada una de ellas, las cuales se muestran en forma resumida a continuación, Figura III.4. Mientras que se pueden ver las distintas clases obtenidas de cada set de datos y el LR entregado para cada una de ellas en Apéndice VI.2 y VI.3.

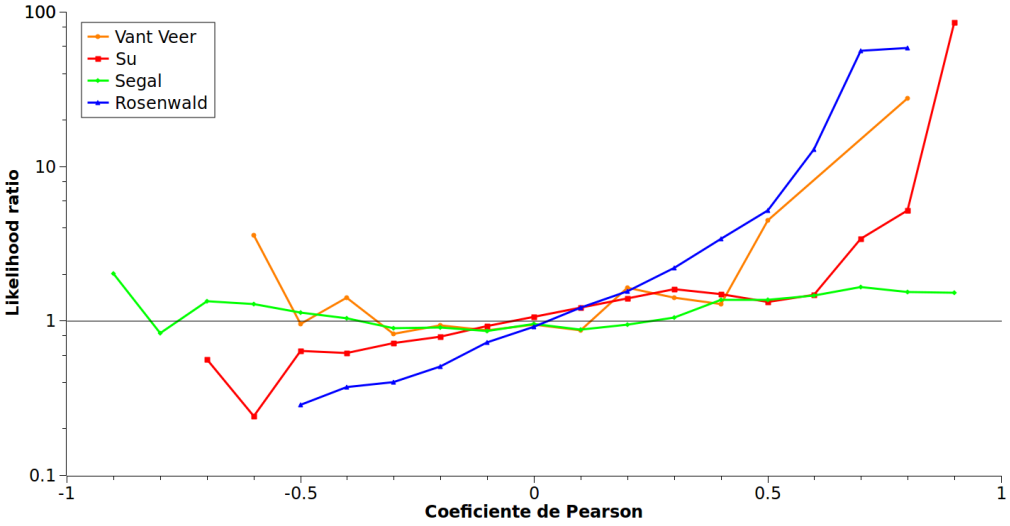

Figura III.4: LR para las distintas clases obtenidas por coexpresión

III.1.4. Funciones biológicas compartidas

La principal base de datos que fue necesaria obtener para la realización de esta etapa fue la base de datos de las anotaciones para productos de genes del ser humano entregado por GO[1] [23]. De esta base datos se obtuvieron los elementos presentados en la Tabla III.7:

Tabla III.7: Elementos obtenidos de la base de datos GO

Anotaciones GO	
58324	Anotaciones
14902	Proteínas
5260	Procesos Biológicos

Aplicando entonces los procedimientos señalados tanto en los antecedentes como en el protocolo, sección II.1.5, se lograron obtener las 7 distintas clases antes mencionadas. De ellas se calcularon los LR para cada clase, que son presentados en la Figura III.5:

[1]Obtenida con fecha Agosto del 2009

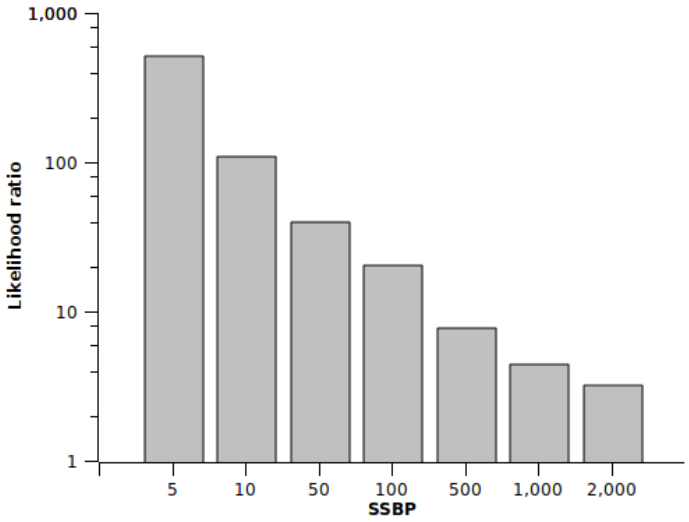

Figura III.5: LR obtenido para las distintas clases del parámetro SSBP

III.1.5. Pares de dominios enriquecidos

La base Interpro entregó los elementos mostrados en la Tabla III.1.5, utilizando como base las proteínas pertenecientes a la GSP:

Tabla III.8: Elementos obtenidos desde la base de datos Interpro

Base de Datos Interpro	
27463	Relaciones
8008	Proteínas
5247	Dominios y Familias

Por otro lado los sets de datos creados desde la GSP tuvieron las dimensiones presentadas en la Tabla III.9:

Tabla III.9: Dimensiones de los sets de datos definidos desde la GSP

Dataset	Secciones	N°PPIs
1	GSP_1 y GSP_2	22577
2	GSP_1 y GSP_3	22577
3	GSP_2 y GSP_3	22576

De allí se aplicó el mapeo a los dominios y familias correspondientes, los cuales, junto al parámetro D obtenido para cada par, se entregan en el anexo del proyecto. Luego de la

selección del mayor D para cada par de proteínas mapeado, se pudo obtener el LR para cada clase como se señala en la Figura III.6:

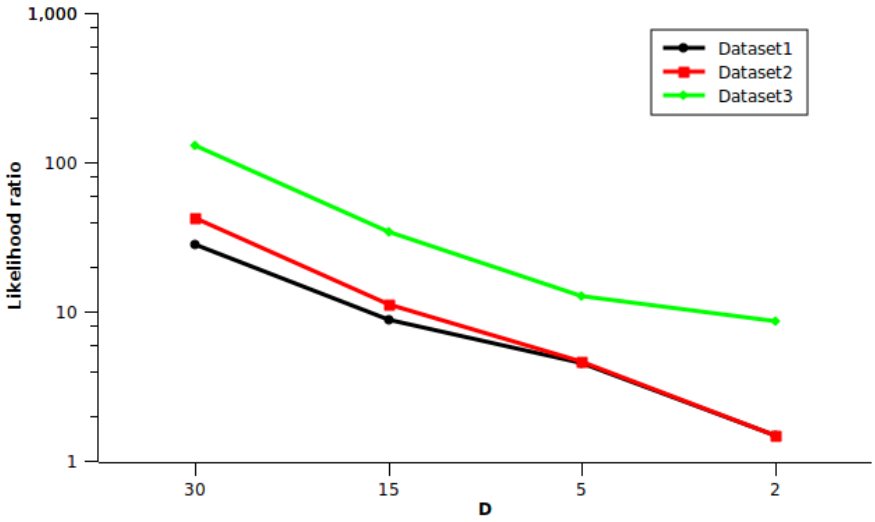

Figura III.6: LR obtenido para los distintos set de datos de D

III.1.6. Integración set de datos

La integración de los sets de datos[2] entregó los resultados presentados la Tabla III.10:

Tabla III.10: Elementos resultantes de la integración de los set de datos

28079358	Pares de Proteínas
16920	Proteinas

Además se realizó el histograma del LR entregado para todos los pares de interacciones inferidos en la Figura III.7:

[2]Ocupando el identificador SwissProt/UniProt

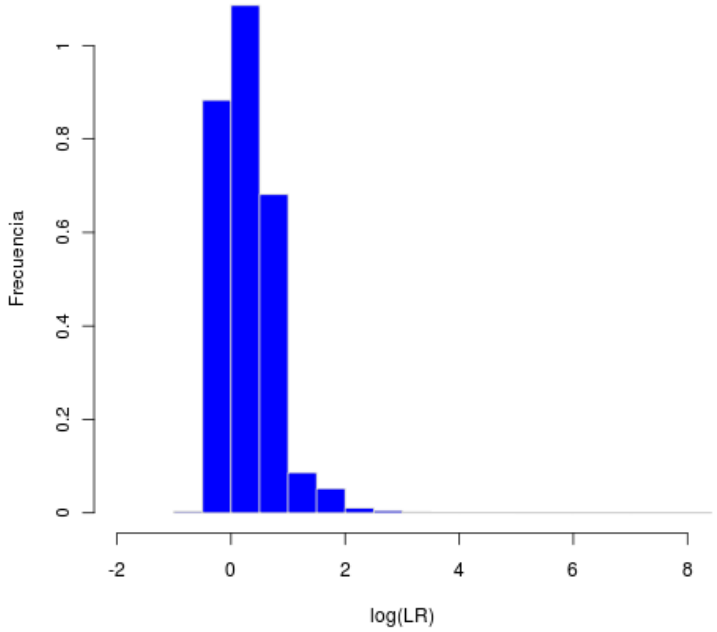

Figura III.7: Histograma $log_{10}(LR)$ del set de datos obtenido
Frecuencias divididas por 10^7

Se muestran a continuación la segmentación del gráfico anterior, separado para valores de LR menores, ver Figura III.8, y mayores de 10000 en la Figura III.9.

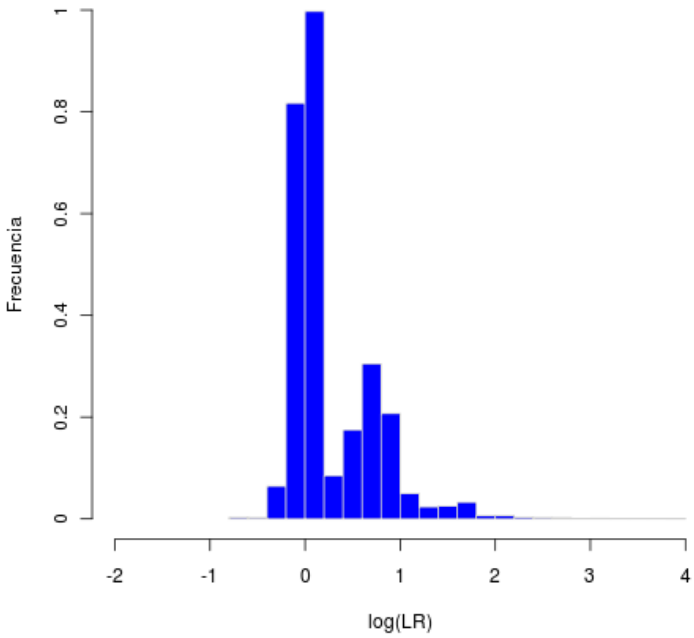

Figura III.8: Histograma $log_{10}(LR) < 4$ del set de datos obtenido
Frecuencias divididas por 10^7

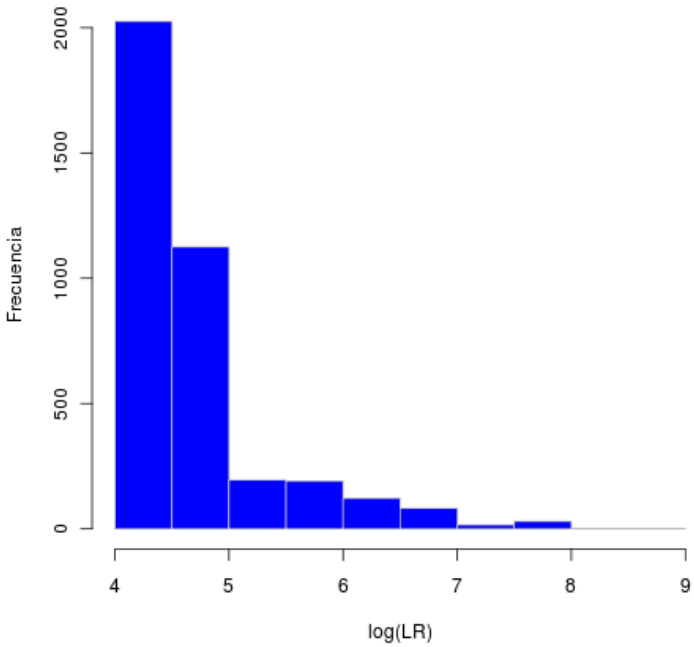

Figura III.9: Histograma $log_{10}(LR) > 4$ del set de datos obtenido

Se realizó además un pequeño estudio de las 5 interacciones con mayor LR, para comparar si existe evidencia de su interacción en la literatura, citando los identificadores de los estudios en la columna 'Estudio', lo cual se puede ver en la Tabla III.11.

Tabla III.11: Principales interacciones inferidas

Proteina 1	Proteina 2	LR	Sets de datos	Literatura	Estudio
P35249	P40938	573.075.310	4	1	EBI-1169825
P43246	P52701	230.292.575	4	4	EBI-944708
					EBI-1164258
					EBI-1164270
					EBI-1164907
P00505	P17174	85.681.573	4	0	
P33992	P49736	84.003.324	4	1	EBI-375112
P68032	P68133	53.930.861	4	0	

III.2. Comparación sets de PPIs inferidas

Los 3 set de datos de las interacciones interproteicas inferidas por este trabajo, Rhodes *et al.* y Xia *et al.* tuvieron las dimensiones[3] presentadas en la Tabla III.12:

Tabla III.12: Comparación de resultados de interacciones de los distintos proyectos de inferencia

Estudio	PPIs	Proteinas	LR_{cut}
Campos	19964	14857	1062
Rhodes	38379	5791	382
Xia	61032	17086	920

En la Figura III.10 se muestra el número de PPIs intersectadas entre los distintos set de datos, esto para $O_{post} \geq 1$.

[3]Utilizando el identificador UniGene

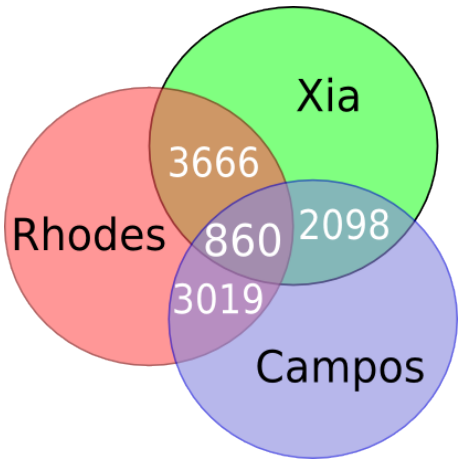

Figura III.10: Número de PPIs intersectando para los distintos sets de datos

Además se muestran a continuación la distribución de los O_{post} para los distintos estudios en la Figura III.11, además, se anexan en Apéndice D los histogramas individuales de los estudios.

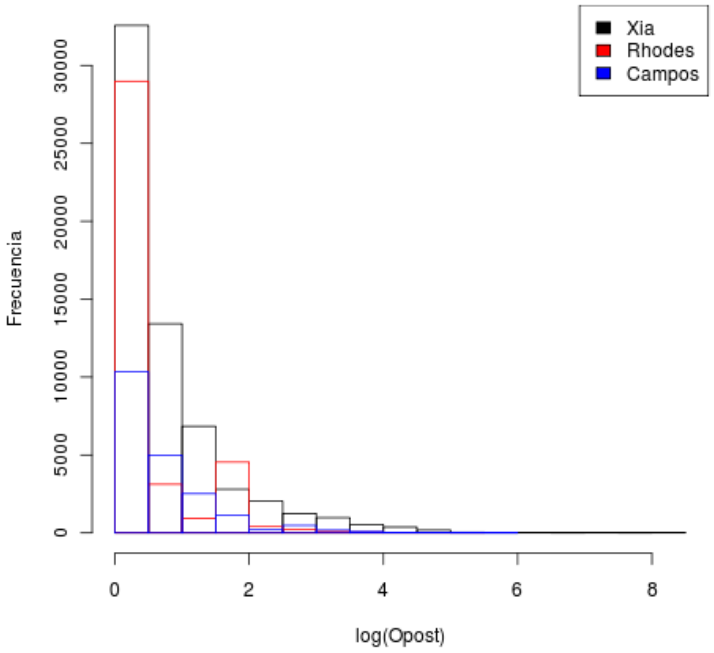

Figura III.11: Histograma O_{post} para los tres estudios

III.3. Red de interacciones del complejo NRC/MASC

La integración de los tres sets de datos para las proteínas pertenecientes al complejo NRC/MASC arrojó los resultados señalados en la Tabla III.13:

Tabla III.13: Elementos de la unidad NRC/MASC inferidos en los distintos estudios

Estudio	PPIs	Proteínas
Campos	135	85
Rhodes	121	84
Xia	132	108

Además se muestra el número de PPIs y el porcentaje de proteínas[4], en la Tabla III.14, relacionadas con alguna interacción por la integración de los set de datos inferidos y su comparación con lo logrado por Pocklington *et al.* a través de la minería de dato de literatura.

Tabla III.14: Datos de la red NRC/MASC entregados por inferencia y por Pocklington *et al.*

	PPIs	%Proteinas
Integración	296	0.7027
Pocklington	299	0.56757

De la inferencia de PPIs se definieron 6 subunidades proteicas, todas ellas definidas por la independencia de sus interacción, es decir, no se lleva a cabo ninguna interacción entre ellas, éstas se muestran en la Tabla III.15:

Tabla III.15: Subunidades independientes de la unidad NRC/MASC

Subunidad	Proteínas	PPIs
1	125	259
2	4	3
3	3	2
4	12	15
5	7	8
6	5	3

Comparando las subunidades mayores obtenidas por este estudio y el realizado por Pocklington *et al.* se obtienen los resultados generales señalados en la Figura III.12:

[4]De un total de 222 proteínas

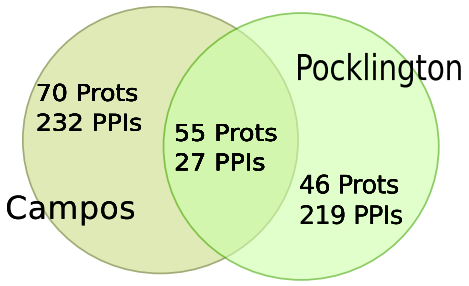

Figura III.12: Comparación de las proteínas y PPIs pertenecientes a la subunidad principal definidos por este trabajo y Pocklington *et al.*

Se seleccionó la subunidad mayor -subunidad 1- para realizar el estudio funcional, siguiendo el protocolo planteado por Rhodes *et al.*, así, al aplicar el algoritmo de *clustering* sobre la subunidad se logró definir 12 conjuntos proteicos, todos ellos con fuertes características individuales. El valor del parámetro Q, el cual -y tal como se mencionó en los antecedentes cuantifica la intraconexión de los módulos- fue de 0.698 sobre 1.

En la Figura III.13 se muestra la relación entre estos 12 clusters (un diagrama más detallado se encuentra en Apéndice VI.3).

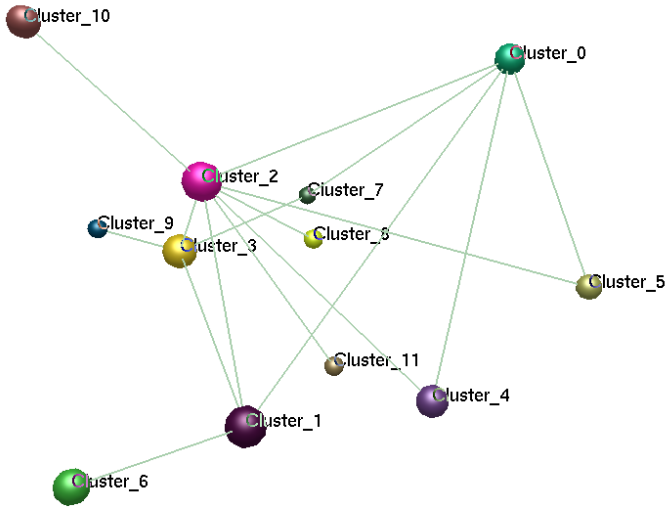

Figura III.13: Estructura de las interacciones en la red NRC/MASC

III.3.1. *Clusters* del complejo NRC/MASC

Se muestra a continuación en la Figura III.14 un diagrama general de la organización de la subunidad mayor de la unidad NRC/MASC señalando las proteínas pertenecientes a ella

y luego se muestra la estructura y las principales características de los *clusters*:

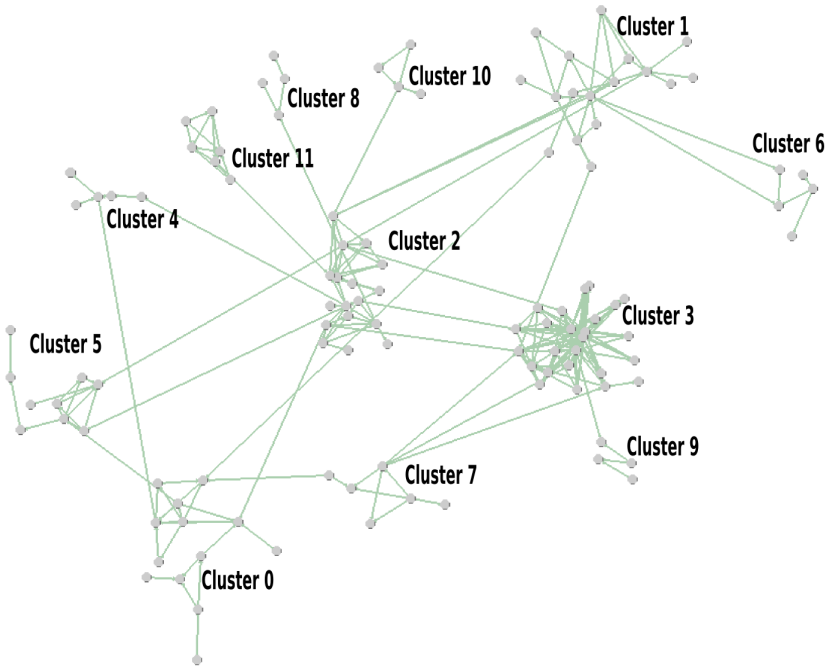

Figura III.14: Diagrama general de la organización del complejo NRC/MASC

Cluster 0

Este *cluster* contiene de las 6 proteínas que conforman el complejo PKA, el cual pertenece a la familia de las Ser/Thr quinasas y es molecularmente dependiente del cAMP. Estas características lo hace un complejo integrador de señales. La presencia de otras proteínas (3) pertenecientes a la subfamilia de la Ser/Thr quinasa refuerza la funcionalidad planteada para este *cluster*. Por otro lado, las demás proteínas tienen diversas funciones, por lo cual no se les puede dar un rol en el *cluster*. Además, este módulo tiene una alta conectividad con los demás *clusters* -5 conexiones- en comparación con el promedio del sistema.

En último lugar se presenta en la Figura III.15 un diagrama del Cluster 0:

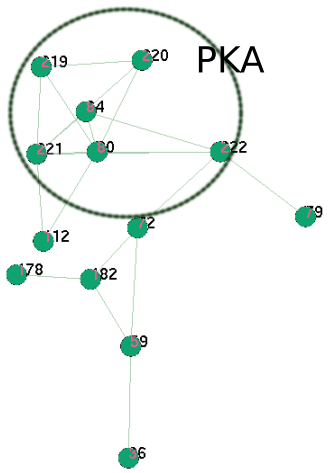

Figura III.15: Diagrama del *Cluster* 0

***Cluster* 1**

Las proteínas pertenecientes a este módulo pertenecen casi en su totalidad -a excepción de 2: CamKIIalpha y PP2B- a las siguientes familias:

- Adhesión celular y citoesqueleto
- Proteínas G y moduladores
- Moléculas y enzimas señalizadoras
- Vesículas sinápticas /Proteínas transportadoras

En este *cluster* se encuentran proteínas pertenecientes a la familia ARP[5], la cual tiene un rol en la regulación de la polimerización de las redes de actina. Además la presencia de las proteínas $CAPZ - \alpha$ y $CAPZ - \beta$ -captadoras de Ca^{+} para el rápido crecimiento de los filamentos de actina- y de la proteína actina señala el rol fundamental en el crecimiento celular.

En segundo lugar, las proteínas G miembros de familia *Ras small GTPase*, lo que implica una función en la señalización interna y en especial en el tráfico vesicular. El tercer grupo, que se relaciona con los dos primeros, es el de las proteínas con función motora. Las 3 proteínas pertenecientes al grupo tienen una función estructural junto con miosina V, la cual se acopla a los filamentos de actina y se mueve por ellos, permitiendo así el movimiento vesicular(o mantención en el espacio).

El último grupo está conformado por moléculas y proteínas señalizadoras, las cuales tienen diversas funciones: la proteína PI3-K[6] tiene el dominio *Phosphoinositide 3-kinase, ras-*

[5]Actin-Related Proteins
[6]UniProt:P42336

binding[7], el cual permite la fosforilización de ciertos sitios como efector de la vía de RAS. Además se encuentra la proteína ARF3, la cual tiene el mismo domino -*Small GTP-binding protein*- que las proteínas relacionadas con RAS.

Cluster 2

El *cluster* tiene como principales componentes a proteínas receptoras de glutamato y transportadoras de vesículas sinápticas. Dentro de las proteínas receptoras se encuentran las proteínas receptoras metabotrópicas y las ionotrópicas -pertenecientes a los complejos mGlur y NMDA-R respectivamente- conformando los principales receptores del NRC/MASC. Además se encuentra presete la proteína HOMER, la cual a través de su dominio EVH1 interactúa con las mGlur, PI_3, TRPC y PLCβ.

El otro gran conjunto es el de las proteínas transportadoras de vesículas sinápticas, las cuales presentan una gran diversidad en sus dominios. Las funciones de estas proteínas son:

Dymanin: Responsable de la endocitosis de las vesículas sinápticas a través de la formación de microtubulos.

NSF: ATPasa necesaria para la fusión de vesículas.

SNAP25: Se asocia con proteínas involucradas en la anclaje vesicular, permite la unión entre la vesícula y la membrana plasmática.

STX: Asociado al anclaje vesicular y principalmente a la exocitosis de los neurotransmisores de la vesícula.

SYT1: Se supone que tiene un rol regulatorio en el anclaje vesicular.

Cluster 3

Aquí se encuentran la mayoría de las proteínas que pertenecen los proteínas compuestas G, las cuales funcionan como moduladores de las señales captadas por el complejo mGluR. Las proteínas G aquí encontradas son:

- $G_{\alpha q}$
- $G_{\alpha i}$
- $G_{\beta\gamma}$

Finalmente se presenta en la Figura III.16diagrama del *cluster* y las proteínas G allí encontradas:

[7]InterPro: IPR000341

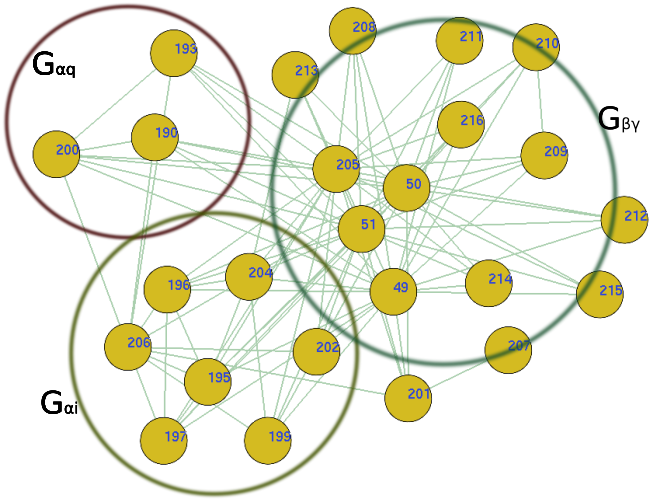

Figura III.16: *Cluster* 3 y complejos proteicos G

Cluster 4

Este *cluster* tiene los siguientes compuestos proteicos:

14-3-3: Proteína transductora de diversas vías de señales.

PPP2CA: Proteína fosfatasa, puede activar vías de detención del crecimiento celular y la activación de la apoptosis.

Además se encuentra la proteína PP5, que regula la mitosis; se presenta en la Figura III.17 el diagrama del *cluster* con sus complejos proteicos.

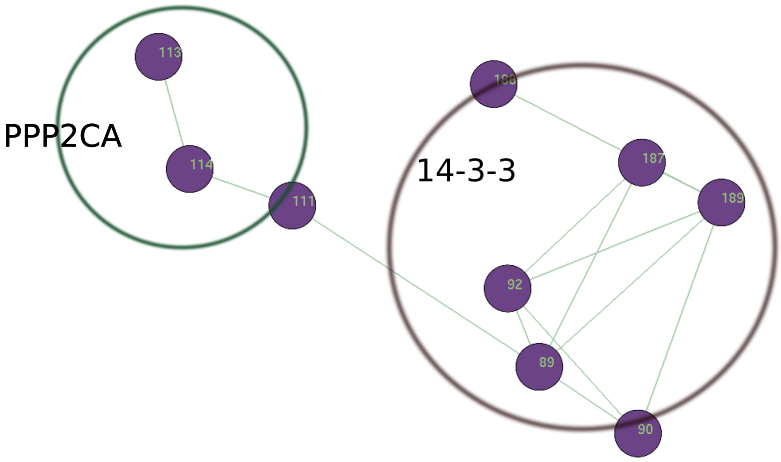

Figura III.17: *Cluster* 4 y sus proteínas compuestas

Cluster 5

Este módulo está compuesto por proteínas que juegan un rol en la regulación y/o ejecución de modificaciones en el crecimiento y en la forma de la célula, tales como DBN1 y Karilin. Otros integrantes son proteínas cuya función biológica es la regulación a diversos procesos biológicos, como se puede apreciar en la diversidad de dominios (ver Material Suplementario) de las proteínas IRS-1 y AKT2. Otra funcionalidad del *cluster* estaría en la regulación de la vía RAS con H-RAS y NF1.

Cluster 6

Las proteínas pertenecientes a este *cluster*, el cual sólo está conectado al *cluster* 1, tienen 2 familias principales: proteínas de adhesión celular y proteínas motoras. Las 3 proteínas pertenecientes a la primera familia funcionan como proteínas de anclaje de los filamentos de actina a una serie de estructuras intracelulares. Por otro lado, las otras 2 proteínas tienen un función estructural con la Miosina, de igual forma que ciertas proteínas del *cluster* 1.

Cluster 7

El conjunto 7 tiene casi en su mayoría proteínas pertenecientes a la vía de señalización MAPK/ERK. Esta vía regula una serie de procesos biológicos como el crecimiento, diferenciación, transcripción, duplicación, etc. Se ha visto además que tiene una fuerte relación con plasticidad estructural, regulación de los receptores AMPA y síntesis proteíca (Thomas and Huganir, 2004 -citado de Pocklinton).

Cluster 8

El *cluster* número 8 está conformado (3 de 4 proteínas) por enzimas de la familia PKC, la cual fosforila específicamente Ser/Thr. Funciona como modulador de señales, pues la activación de estas enzimas pueden dependen de señales que aumenten el Ca^+ y fosfolípidos. Sus objetivos son regular una gran variedad de proteínas y procesos.

Cluster 9

Las principales proteínas del *cluster* 9 son la proteína Grb2[8], enlace fundamental entre los factores de crecimiento externo y la vía de señalización RAS, y quinasas con los dominios SH2 y SH3, lo que las hace moduladoras de señales en cascada de diversas vías.

[8] *Growth factor receptor-bound protein 2*

***Cluster* 10**

Este *cluster* contiene únicamente 4 proteínas, 3 de las cuales son ATPasas transportadoras de iones: la proteína ATP2B4 transporta Ca^+ fuera de la célula, mientras las proteínas ATP1A1 y ATP1A3 median el transporte inverso de Na^+ y K^+, ingresando potasio a la célula. Por otro lado la cuarta proteína -la Calretinina- tiene como única función conocida la unión al ion calcio.

***Cluster* 11**

Todas las proteínas del *cluster* 11 tienen un dominio PDZ, lo que significa que se tratan de proteínas de transmembrana. Todas las proteínas, a excepción de una, tienen un dominio SH3, lo cual revela una funcionalidad en las vías regulatorias localizadas río abajo. Por otro lado se ha obtenido que la funcionalidad de la mayoría de estas proteínas es el ser proteínas andamios[9].

III.4. Influencia de las interacciones en la generación de enfermedades

Para señalar cúales son las proteínas afectas en la estructura del complejo NRC/MASC en los sistemas en donde se ha gatillado una enfermedades cognitiva o se ha visto modificado su plasticidad sináptica y de comportamiento se presentan a continuación una serie de diagramas, en donde la mutación de una proteína se representa a través de la coloración de la proteína o , en caso que el afectado sea un complejo proteico, se muestra un círculo colorado en su representación.

Para la coloración se utilizó la información recopilada por Pocklington *et al.* [28] y Grant *et al.* [11], la cual fue obtenida a través de la minería en la literatura pertinente. Buscándose aquellos estudios en donde se presetaran polimorfismos en proteínas asociadas al complejo NRC/MASC y que presentaran un fenotipo relacionado con la capacidad cognitiva del individuo. Se busca con estos diagramas encontrar una correlación entre los fenotipos y los módulos y vías de señalización, en donde la modificación de estos últimos supone el gatillamiento de las enfermedades. Además, para diferenciar los elementos y módulos del diagrama se debe comparar con la Figura III.14 presentada anteriormente.

En primer lugar se muestra en la Figura III.18 aquellas proteínas y complejos que se ha encontrado alguna mutación en personas con bipolaridad.

[9]Scaffold proteins

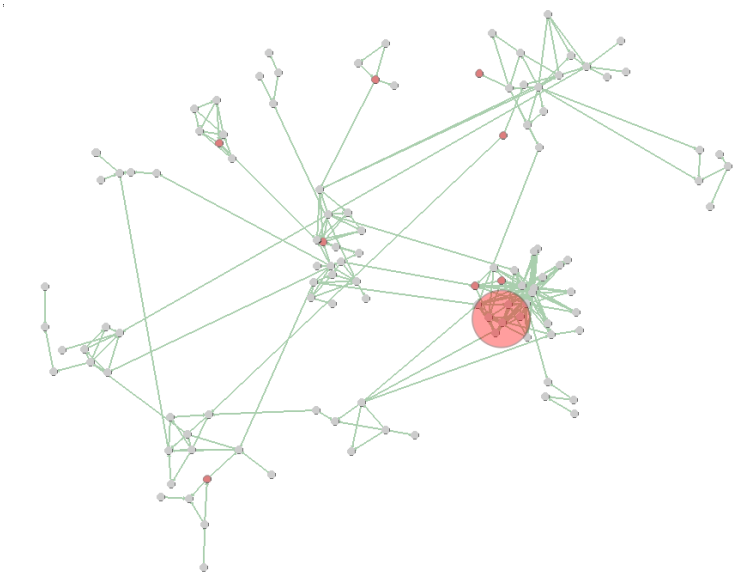

Figura III.18: Efectos de la bipolaridad en el complejo NRC/MASC

Luego se muestra en la Figura III.19 aquellas proteínas y complejos que se ha encontrado alguna mutación en personas con esquizofrenia.

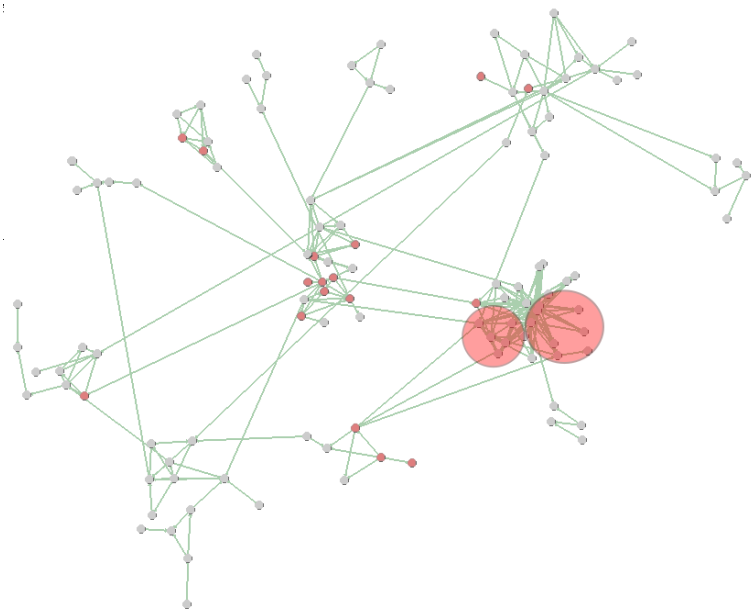

Figura III.19: Efectos de la esquizofrenia en el complejo NRC/MASC

Además, se señala en la Figura III.20 aquellas proteínas y complejos que se ha encontrado alguna mutación en personas con retardo mental.

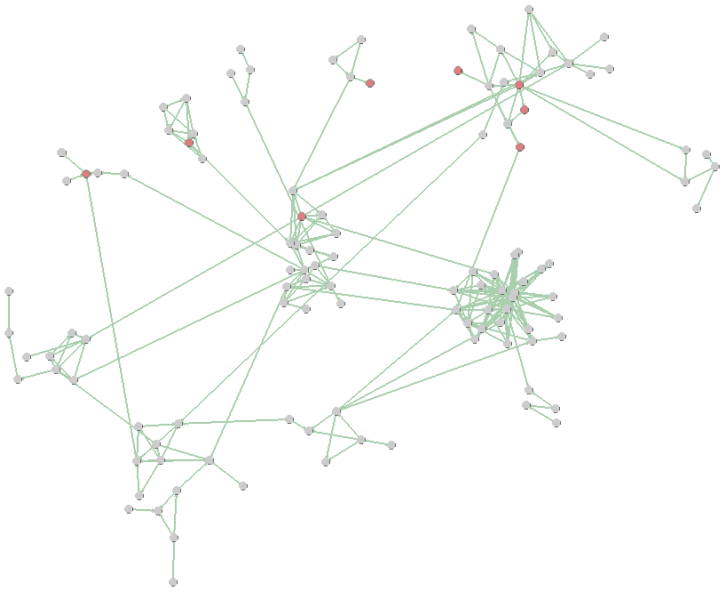

Figura III.20: Efectos del retardo en el complejo NRC/MASC

A continuación se muestra en la Figura III.21 aquellas proteínas y complejos que se ha encontrado alguna mutación en personas con diversas enfermedades cognitivas no presentadas en las figuras anteriores y que pueden ser consultadas en el Material Suplementario de Grant *et al.* [11].

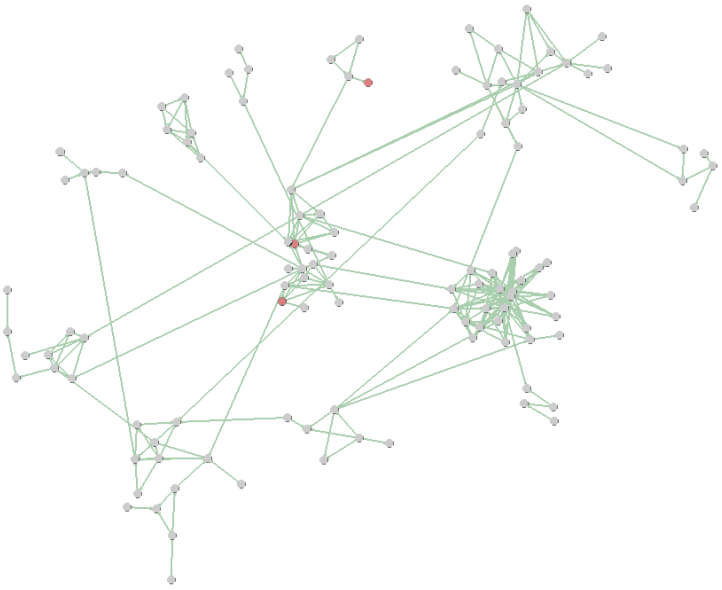

Figura III.21: Efectos de otras enfermedades en el complejo NRC/MASC

Se presenta en la Figura III.22 aquellas proteínas y complejos en que se ha encontrado alguna mutación en individuos con la plasticidad sináptica modificada.

Finalmente se muestra en la Figura III.23 aquellas proteínas y complejos en que se ha

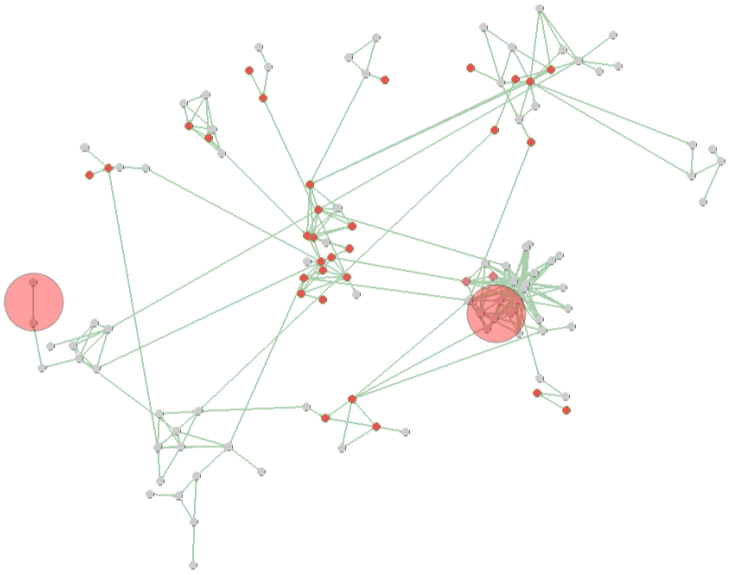

Figura III.22: Proteínas fundamentales en la plasticidad sináptica en el complejo NRC/MASC

encontrado alguna mutación en individuos con la plasticidad del aprendizaje modificada.

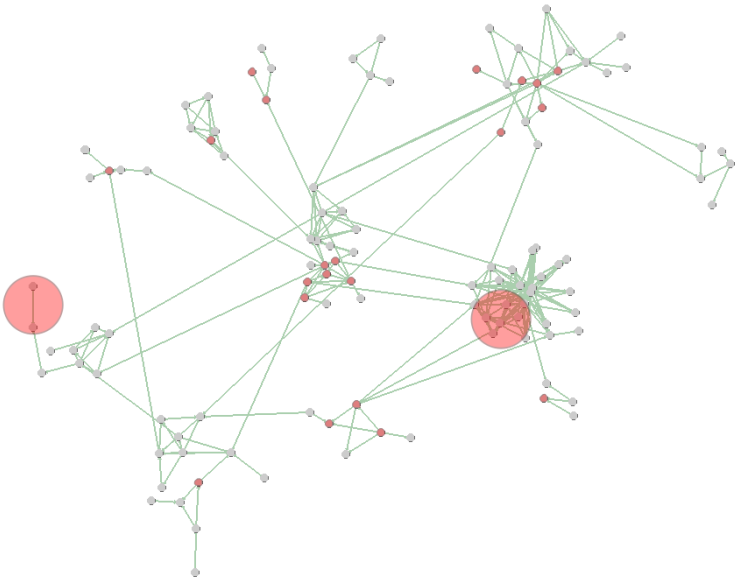

Figura III.23: Proteínas fundamentales en la plasticidad del aprendizaje en el complejo NRC/MASC

Capítulo IV

Discusión

En el presente capítulo se señalarán las discusiones llevadas a cabo sobre el desarrollo y los resultados obtenidos para las fases del proyecto de memoria. Es así como a continuación se presentarán los elementos más relevantes sobre la inferencia de la red de PPIs.

IV.1. Inferencia red de interacciones interproteicas

El primer elemento a considerar son los sets *gold standard* definidos en este proyecto. Ellos presentaron varias modificaciones en comparación con los utilizados por Rhodes *et al.* [30], así se tuvo un aumento tanto en el número de proteínas participantes como en el número de interacciones reportadas. Esto se explica fácilmente por la diferencia del tiempo -por lo menos 3 años- en que las bases de datos, desde donde se obtuvieron, fueron consultadas. Por ejemplo, la diferencia para el caso del GSP se tiene una relación de aproximadamente 4 interacciones para cada proteína, mientras que para Rhodes *et al.* la relación es de 2,12 interacciones para cada proteína.

Por otro lado, el solapamiento (intersección entre los dos conjuntos) entre el set GSP y el GSN, en este estudio, fue menor al 4 %, lo que significó una alta calidad de los datos y que la independecia se lograra con una modificación mínima del set. El uso de éstos como *gold standard* es adecuado, pues cabe recordar que la independencia de los sets de datos, en especial de los *gold standard*, es fundamental.

IV.1.1. Implementación de las bases de datos

La existencia de diversas dificultades y logros respecto a la implementación y procesamiento de las 4 bases de datos ocupadas, obligan a realizar una discusión particular sobre cada una de ellas.

Interacciones Ortológicas

Lo que primero llamó la atención de esta sección del proyecto fue la variación en la cantidad de proteínas e interacciones pertenecientes a los distintos set de datos ortológicos. El mapeo modificó los sets de datos en todas las secciones del proyecto, pero fue en ésta en donde la variación fue muy substancial, pues, y tal como se puede observar en la subsección III.1.2, el número de proteínas se redujo en aproximadamente un 30 % en los 3 set de datos. En cambio el número de interacciones se mantuvo casi constante, a excepción del set de datos del *C. elegans*, en donde existió una reducción del 50 %.

Mientras que era esperable una variación del número de integrantes de los distintos set de datos por el mapeo de las proteínas no se esperaba que el grado de variación fuese tan alto (se debe recordar que se mapeó en tres ocasiones consecutivas los sets de datos). La comparación de los resultados actuales y los entregados por Rhodes *et al.* muestra que la modificación, variación en el número de integrantes de los sets de datos, fue más alta en el presente estudio. Se considera que ello fue provocado por el mejor curamiento de las bases de datos utilizadas en este estudio en comparación con las ocupadas por Rhodes *et al.*. Así, muchos identificadores de proteínas que antes pertenecían a las bases de datos ahora ya no existen, esto por haber sido identificadas posteriormente como variaciones de proteínas ya existentes o porque sus identificadores fueron modificados para distinguir sus particularidades. Esto sucederá no solamente en esta parte del proyecto, sino que en cualquier punto donde se necesite mapear identificadores.

Finalmente los resultados entregados señalan, en primer lugar, que la estratificación de los sets de datos a través del uso de clases logra efectivamente diferenciar elementos con mayor probabilidad de interacción. Esto se puede observar en el gráfico III.1, en donde la diferencia del *likelihood ratio* obtenido para las distintas clases es gigantesco, con una diferencia de casi mil veces entre las clases con mayor y menor LR. Esta alta diferencia se mantiene incluso entre la clase con mayor LR y el LR del total del set de datos, con lo cual queda respaldada la hipótesis en que se basaron estas separaciones. Además, y como era de esperar, entre más alto el LR obtenido para una clase, más específica es (menor cantidad de elementos contiene).

Todo ello sucede en los dos sets de datos que fueron separados en clases, mientras que en el set de *C. elegans*, en donde no se llevó a cabo estratificación alguna, el resultado del *likelihood ratio* total fue similares al de las mejores clases de los otros set de datos. Esto permitió el definir interacciones con una alta probabilidad de interacción, a pesar de no haber realizado la separación del set de datos.

Los LR obtenidos en esta sección fueron bastante altos, permitiendo con ello que la mayoría de las interacciones que pertenecen a las clases superiores tener un O_{post} cercano a 0.5.

Matrices de coexpresión

De los 5 grupos de datos de *microarrays* considerados inicialmente para integrar al proceso de inferencia, sólo 4 fueron exitosamente incluidos, esto por ser imposible encontrar el material suplementario -las sondas utilizadas en los experimentos- necesario para implementar el set de datos del estudio de Chen *et al.*. De la implementación de los otros 4 sets finalmente sólo se consideraron 2 en la integración final de los likelihood ratios. Esto sucedió por la calidad de los resultados obtenidos y que se pueden observar en la Figura III.4, en donde solamente 2 curvas -Rosenwald *et al.* y Su *et al.*- tienen un crecimiento estrictamente positivo. Esto por la hipótesis en que se basó la implementación de las matrices de coexpresión, en donde a mayor r -correlación positiva en la expresión entre 2 genes- debiera obtenerse un alto LR. Mientras que la curva de Segal *et al.* tienen un comportamiento de su coeficiente de correlación cercano a 1, pero sin patrón alguno. En el caso de la curva del estudio de Vant Veer *et al.* que para los valores de r cercanos a 1 tiene crecimiento constante, no es fiable por su comportamiento para valores negativos.

Lo anterior se explica por la alta eliminación de datos de los *microarrays* en cuestión, así, las restricciones impuestas para mapear las sondas a genes fueron bastantes restrictivas. como es el caso del set de Vant Veer *et al.*, en donde se mapeo un número de genes equivalente al 8 % de las sondas iniciales. Por otro lado, las restricciones para aceptar una sonda como válida también se aplicaron de manera rígida, en especial para el set de Segal *et al.*.

Las fuertes restricciones fueron impuestas para garantizar que los resultados obtenidos fuesen de una alta calidad y así disminuir la presencia de falsos negativos en los sets obtenidos. Los *likelihood ratios* de los sets considerados fueron bastante altos, en especial si se les compara con los obtenidos por Rhodes *et al.*.

Funciones biológicas compartidas

Los resultados obtenidos para la inferencia por funciones biológicas compartidas entregó resultados bastante positivos. En primer lugar, la hipótesis fue fuertemente respaldada, pues la diferencia entre los likelihood ratios calculados para las distintas clases varía de forma exponencial. Además se mantuvo la tesis de que existe una fuerte correlación entre la especificidad de una clase -en función del porcentaje de interacciones verdaderas positivas en ella- y el bajo número de elementos en ella. Así, para estos resultados se puede ver que existe una diferencia de 2 órdenes de magnitud en el número de elementos entre la primera y la última clase, apoyando esto que el número de PPIs interactuantes -alto LR- es mucho menor al número de no interactuantes. La metodología utilizada en la manipulación de esta base de datos no entregó dificultades excepcionales, es más, este fue uno de los procesamientos más simples del proyecto.

Pares de dominio enriquecido

De forma parecida a los resultados logrados a través de las funciones biológicas compartidas, los likelihood ratios obtenidos a través de esta metodología permitieron confirmar la hipótesis (ver I.3.3) de esta sección, además de lograr definir conjunto de interacciones con una alta probabilidad de ser verdaderos positivos. El único punto que llama la atención de estos resultados fue el hecho que uno de los set de datos definidos tuviese una distribución de likelihood ratios mucho mayor a los otros dos. Al ser utilizada la misma metodología en el tratamiento de todos los sets y al ser estos definidos a través de selección azarosa no se encuentra ningún argumento que explique el por qué de estos resultados. Lo único presumible es que se produjera una cierta conjunción de elementos que permitieran el cálculo de LR mayores, pues a pesar de la notoria diferencia la curva de likelihood ratio tiene la misma forma que para los otros sets.

IV.1.2. Integración de los sets de datos

La integración de los sets permitió definir una base de datos de más de 28 millones de interacciones putativas, con casi 20000 proteínas. El histograma de estas interacciones señala que la gran mayoría ($> 95\,\%$) de las interacciones tiene un $LR_{comp} < 100$. Por otro lado, la distribución de estas interacciones, aunque decreciente, no presenta una curva definida. Así por ejemplo, el número de interacciones que tienen los menores LR no muestran un curva decreciente como era de esperar, sino que tienen una alta variación, comprendiéndose ello por la forma en que el LR_{comp} fue obtenido.

Adicionalmente se debe discutir los resultados presentados en el cuadro III.1.6, los cuales apoyan el concepto de likelihood ratio. Las interacciones allí presentadas tienen un valor de LR_{comp} en extremo alto, lo cual, según la definición de likelihood ratio, se puede describir también como:

$$P(Positivo) \gg P(Negativo)$$

Esto indica que la posibilidad que sea un verdadero positivo es mucho mayor a que sea un verdadero negativo, lo cual se ve reflejado en el número de publicaciones en donde estas interacciones han sido demostradas: sólo 3 interacciones (RFC3-RFC4, MSH2-MSH6 y MCM2-MCM5) tienen una fuente en la literatura, mientras que las otras 2 (AATM-AATC y ACTC-ACTS) son desconocidas hasta ahora. No se ha encontrado diferencia alguna entre los pares con o sin evidencia en la literatura. Ya que todos los pares tienen como elementos proteínas de características similares, pertenecientes a los mismos procesos biológicos. Se propone que la única razón para que no se hayan reportado todas las PPIs aquí señaladas es la falta de un experimento enfocado en su procesos y/o el componente celular que los cobija y que con el avanzar de los métodos de detección, en conjunto con la confianza de estos, sea altamente probable (no hay que olvidar el carácter probabilístico) encontrar evidencia empírica de su existencia.

Con todo lo cual el objetivo de esta fase se ha logrado, pues se definieron interacciones in-

terproteicas putativas con un parámetro probabilístico, siendo su conocimiento previo posible o no.

IV.2. Modelo de la red de la NRC/MASC

IV.2.1. Comparación sets de PPIs inferidas

Los distintos sets -Campos, Rhodes y Xia-, presentaron ciertas diferencias en lo que al número de sus integrantes se refiere, en particular el número de PPIs. Así se tiene que el set de datos de Xia *et al.* es muchísimo más grande que el de obtenido en este trabajo y casi el doble de Rhodes *et al.*. Se considera que esto tiene relación a la metodología ocupada en aquel estudio, en donde el número de bases de datos integradas fue 7, mientras que para los otros 2 estudios sólo se ocuparon 4 bases de datos. Esta gran diversidad de información permitió una mayor amplitud en los datos calculados.

Por otro lado, algo parecido ocurrió respecto al número de proteínas que integran los sets de datos, Rhodes *et al.* tuvo un número mucho más bajo que los otros dos sets, esto por ser el estudio de Rhodes *et al.* 2 años más viejo que el de Xia *et al.* y 4 años que el aquí presentado. Lo anterior se puede comprobar gracias al parámetro L_{cut} (Tabla III.12). Allí se observa que el perteneciente a este estudio y al estudio de Xia *et al.* es bastante mayor que el de Rhodes *et al.*, esto por la relación directa que existe entre el número de elementos de la GSP y el L_{cut}, presentado en los antecedentes.

Finalmente, el histograma III.11 muestra la distribución de los PPIs para cada set. En una primera inspección todos los sets se distribuyen de manera parecida, con un decaimiento exponencial. El principal punto de interés de este gráfico es la diferencia existente entre la distribución de los elementos aquí obtenidos y los presentados por Rhodes *et al.*, pues aunque ambos sets fueron obtenidos a través de la misma metodología, la distribución de los LR de este estudiode es mucho más suave que la de Rhodes *et al.*. En este caso, luego de una alta frecuencia en el intervalo de interacciones con $LR_{comp} < 5$ cae fuertemente para los demás intervalos, mientras que para los LR calculados se mantiene una distribución baja pero constante en los distintos intervalos. Este resultado permitiendo hacer la suposición que la calidad de los resultados -en función del LR_{comp} calculado- es mejor el set de datos obtenido en este trabajo, por tener una distribución con más peso para los valores altos de LR.

IV.2.2. Red de interacciones de la NRC/MASC

La integración de los 3 sets de datos entregó una red con una cantidad de PPIs pertenecientes a la NRC/MASC parecida a la obtenida por Pocklington *et al.*, aunque con un mayor número de proteínas integrantes. Lo cual permitiría pensar que la red inferida sería menos robusta (en relación a la conexión interna), pero al comparar los datos de la subunidad independiente mayor para cada una de las redes, se tiene que la red putativa entrega

una subunidad con 125 proteínas y 259 PPIs, mientras que la red obtenida a través de la literatura entrega una subunidad defina a través de 246 PPIs y 101 proteínas, tal como se puede apreciar en el diagrama III.12. Ante esto se podría considerar que se ha obtenido una red más grande que la entregada por Pocklington *et al.*, pero con el mismo nivel de intraconectividad. Sin embargo, considerando únicamente la subunidad mayor de cada estudio, se encontró evidencia contraria a lo señalado anteriormente a través del parámetro conectividad media de cada red. La red obtenida tuvo un conectividad media de 4,4 conexiones por cada proteína, mientras que el de Pocklington *et al.* fue de 4.8 conexiones por cada proteína, lo cual señala una mayor interconexión en la segunda.

La aplicación del clustering sobre la subunidad principal entregó 12 conjuntos proteícos, cada uno fuertemente intraconectado. Al observar la relación entre los distintos clusters en la Figura III.13 se pueden apreciar las características principales del modelo de la NRC/MASC entregadas por este proyecto:

Dentro de la organización general de la unidad, salta a la vista que un *cluster* en particular esté tan conectado con los demás. Mientras que el promedio de conexión de los clusters es de 2.67 el *cluster* 2 esta relacionado con 8 módulos, lo cual indica el rol preponderante que ha de tener este *cluster* en la transducción de las señales dentro de la unidad.

Por otro lado se puede observar que la gran mayoría de los módulos tienen una disposición paralela, es decir, no se forman series de clusters conectados, sino que la gran mayoría de éstos se sitúan alrededor del *cluster* 2, formando comúnmente una conexión adicional con otro conjunto. Finalmente, sólo 3 conjuntos proteicos no tienen relación alguna con el *cluster* 2, lo interesante de éstos es que su número de conexiones es en extremo bajo, dos de ellos tienen una única relación, mientras que el tercero dos. Esto permite suponer que el rol jugado por estos módulos debiera ser parecido a lo señalado por Pocklington *et al.* para aquellos clusters definidos como outputs, por estar anexados a un cluster modulador y probablemente servir como efector de las señales transducidas.

Con los resultados de esta primera inspección de la organización de la unidad disponible se procedió a estudiar la composición de uno de estos conjuntos.

IV.2.3. Clusters de la NRC/MASC

Para iniciar la discusión respecto a la división funcional de la NRC/MASC, se debe considerar ,en principio, el módulo central de la unidad, el Cluster 2. En él se encuentran las principales proteínas receptoras, ellas pertenecen a las unidades receptoras NMDA-R y mGluR. A diferencia de lo señalado por Pocklington *et al.*, en donde estas dos unidades se encontraban en módulos diferentes, en este estudio encontramos ambas unidades unidas, esto se puede considerar tanto como una falta de calidad en los resultados o como información novedosa respecto a la organización de la unidad. Si consideramos todos los elementos del Cluster 2 nos encontramos con una alta especificidad, pues además de los receptores de glutamato el otro gran grupo presente es el de las proteínas involucradas en el transporte de vesículas sinápticas, logrando con esto definir al Cluster 2, como un módulo exclusivo de

proteínas ligadas a la recepción de señales. Esto diferencia a los resultados señalados por Pocklington *et al.*, en donde los clusters receptores -1 y 2- contienen una gran cantidad de proteínas anexas y cuyo rol dentro del conjunto no ha sido esclarecido.

El siguiente punto de discusión de la orgánica de la NRC/MASC es el rol jugado por los módulos relacionados con el Cluster 2. La relación más simple encontrada en el diagrama es la que incluye los Clusters 8, 10 y 11, los cuales sólo tienen relación con el Cluster 2. Así tenemos que el Cluster 8 está conformado por proteínas pertenecientes al PKC[1] y las del Cluster 10 está conformado principalmente por ATPasas. Mientras que en último lugar el Cluster 11 tiene como función principal el servir como un complejo andamio, que una las distintas subunidades del complejo NRC/MASC. Al tener esa función es comprensible su relación directa con los complejos receptores.

Lo interesante de estos módulos es que -además de ser transductores de señales o efectores- no habían sido definidos anteriormente Es decir, en los resultados de Pocklington *et al.* no se había logrado definir clusters donde proteínas con las mismas características, y de forma tan exclusiva, estuviesen juntos. Aventurarse a señalar cuales podrían ser las particularidades de estos módulos, además de los ya señalados, no es posible con los resultados obtenidos hasta ahora, pues sería necesario un estudio a detalle de las relaciones de estos clusters con unidades fuera del NRC/MASC. Otro conjunto de clusters que han de ser estudiados con especial interés es el que reune a los Cluster 0, 1 y 3, por el alto número de conexiones que presentan con otros módulos. El Cluster 0, tal como se señaló en Resultados, está compuesto por el compuesto proteico PKA[2] y proteínas quinasas, además de otras proteínas con diversas funciones. Al ser el compuesto PKA un complejo fosforalizante regulado por cAMP, es comprensible su relación con el *cluster* que contiene al receptor mGluR -río arriba- y los demás *clusters* efectores -río abajo-. Es por ello que este módulo tendría un rol parecido al Cluster 3 definido por Pocklington *et al.*, modulando las señalas de los receptores y transduciéndolas a los módulos ejecutores.

Por otra parte, el Cluster 1 tiene dos tipos principales de proteínas: aquellas destinadas a funciones estructurales y las GTPasas señalizadoras (de la subfamilia de las small RAS proteins). Al igual que el Cluster 0 este módulo tendría por fin el transducir señales que permiten el crecimiento, diferenciación y otros tipos de cambios en la célula. Ante esto se puede suponer que estamos frente a un conjunto tanto efector -en lo referente a la estructura de la célula- como modulador gracias al conjunto de proteínas RAS. Finalmente el Cluster 3 al estar compuesto casi en su totalidad por proteínas G, es de esperar que sea la vía de modulación principal del mGluR, pues como fue mencionado en los antecedentes, al no crear un canal ionotrópico, el mGluR utiliza cascadas señalizadoras GTPasas como vías de señalización.

Es por ello que se ha definido a los Clusters 0 y 3 como secundarios o "tranductores de señal", tomando un rol parecido al del Cluster 3 de Pocklington *et al.*, pero de una manera más estructurada. Cluster 1 se debiera incluir en el tercer nivel, el de efectores o transductores extracelular de señales, esto por su fuerte rol en las modificaciones estructurales de la célula.

[1]Protein kinase C
[2]Protein kinase A

Como último punto quedan los módulos con baja conectividad, éstos son los Clusters 4, 5, 6, 7 y 9. A continuación se lleva a cabo una corta discusión sobre las principales características y se presentan las hipótesis sobre su función en la organización del complejo NRC/MASC.

Cluster 4: Tiene 2 compuestos proteícos que tiene como principal funcionalidad el transducir señales y ser efectores de ellas. Así, su unión al Cluster 2 y 0 tendría a fortalecer la hipótesis del paralelismo en la unidad. El recibir y transducir señales, además de activar diversas señales e iniciar ciertas funciones ubicaría a este *cluster* en una zona intermedia, entre la capa transductora y efectora definidas por Pocklington *et al.* Además su participamiento en el crecimiento y muerte celular le dan un rol específico en la NRC/MASC.

Cluster 5: Este conjunto proteíco al igual que el Cluster 4, tiene características tanto efectores como moduladoras. Al no existir una especificidad muy alta, pues a pesar de la alta presencia de la familia proteínas G y moduladores no permite definir una hipótesis funcional clara para el *cluster*. Pero sí se puede localizarlo en una zona transductora.

Cluster 6: La principal característica de este módulo es que está unido exclusivamente al Cluster 1, lo que permite suponer que este *cluster* debiese tener una funcionalidad efectora. Sus proteínas han sido clasificadas como estructurales, pues ellas tienen un activo rol en las modificaciones en las fibras de actina. Estos elementos señalan la fuerte relación entre el Cluster 1 y el 6, siendo el último un efector especializado del primero.

Cluster 7: Lo que primero salta a la vista de este módulo es que está relacionado con 2 clusters, siendo ninguno de ellos el Cluster 2. La gran especificidad de este módulo en la transducción y regulación de vías de señalización, tanto en elementos río arriba -como el receptor AMPA- como río abajo. Un punto importante a considerar en el futuro es la relación de estos 3 clusters como un complejo modulador de señales dentro de la NRC/MASC.

Cluster 9: Es un *cluster* con un fuerte presencia de proteínas de señalización, tanto internas como externas. Al igual que el Cluster 6, el Cluster 9 sólo está unido a otro módulo, el cual es distinto al Cluster 2. En este caso es el Cluster 3, lo cual permitiría a este módulo es transducir señalas moduladas por el Cluster 3 a diversas vías de señalización en cascada.

Respecto a la especificidad encontrada en los clusters definidos se debe comentar que aquellos resultados alentaron el desarrollo del proyecto, pues se esperaba como resultado inicial una cierta particulatización de los módulos obtenidos a través del algoritmo, ya que ello significaba que el algoritmo efectivamente había logrado separar las distintas proteínas en función de sus características funcionales. Además cabe señalar que las interacciones obtenidas a través de Naïve-Bayes integran una gran gama de características de las distintas proteínas, es por ello que se considera que el algoritmo logró separar funcionalmente las proteínas, a la vez que la relación entre los módulos está definida por la agregación de las características de las proteínas dentro de un *cluster*.

A continuación se presentan los diagramas representando la organización hipotetizada del complejo NRC/MASC y las relaciones entre los módulos a través de los resultados aquí obtenidos , Figura IV.1, y los entregados por Pocklington *et al.* [28], Figura IV.2.

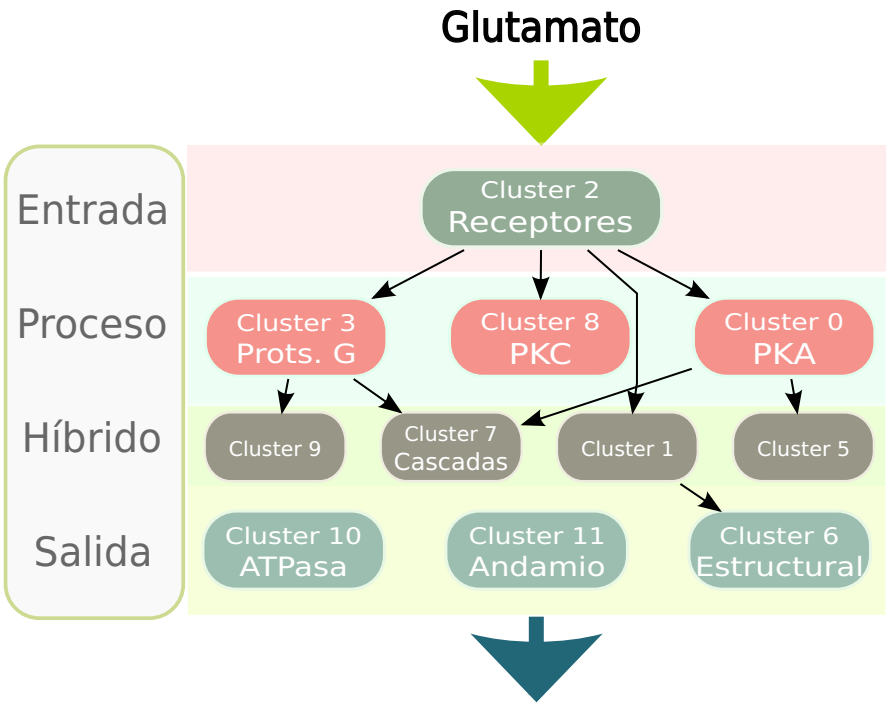

Figura IV.1: Diagrama del modelo final para la NRC/MASC

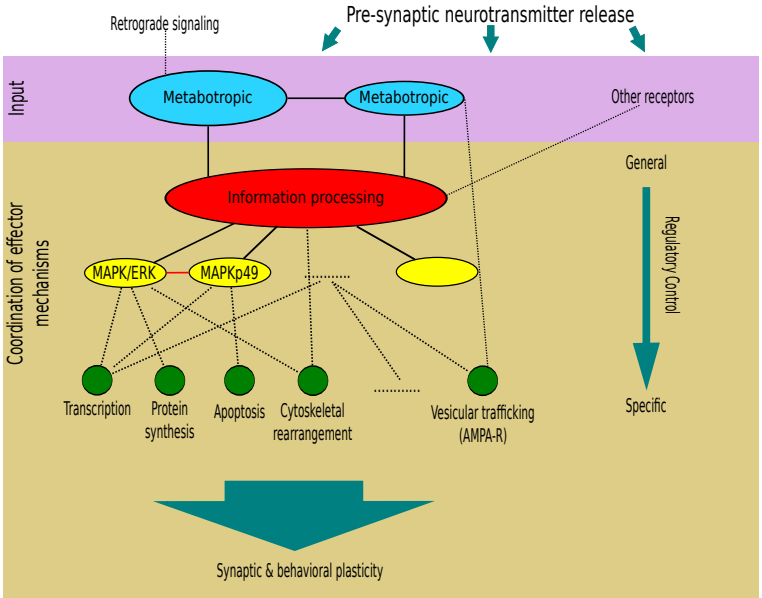

Figura IV.2: Organización funcional de la NRC/MASC, basado en Pocklington *et al.* [28].

En el diagrama IV.1 se presenta el modelo funcional de la NRC/MASC propuesto en este trabajo. Allí las vías de señalización y los efectores tienen una organización paralela, es decir, no existe una matriz organizadora de todas las señales recibidas, la cual después de

modular las señales, las envía a los distintos *clusters* efectores. Aquí se ha podido dilucidar una serie de vías y conjuntos organizadores independientes. Mientras que otros elementos siguen teniendo una mecanismo desconocido, como es el caso de los módulos efectores 10 y 11. En estos casos su única relación es con el *cluster* receptor, siendo esperable entonces una conexión adicional con un *cluster* regulador; sin embargo aunque no se descarta que sólo sea necesaria la relación aquí señalada, pues es esperable, por ejemplo, que el Cluster 11 -*cluster* andamio- no necesite más conexiones.

La tesis de este proyecto sobre el funcionamiento de la NRC/MASC es que las distintas vías se dividen en 2 tipos: aquellas que están integradas por las 3 capas definidas por Pocklington *et al.*, siendo éstas entrada, procesamiento y salida, con la capacidad de relacionarse fuertemente con otras vías de señalización a través de los *clusters* transductores. El segundo tipo es una vía más simple, en donde desde el módulo receptor se pasa la señal a un módulo híbrido, es decir, tanto modulador de señales como efector. A pesar que en el diagrama IV.1 no se señalen todas las relaciones obtenidas por el proceso de *clustering*, se debe recordar que sólo 3 módulos no tienen relación con el *cluster* receptor. Un punto que todavía se debe dilucidar es el que trata las relaciones entre las vías paralelas, pues aunque se han obtenido ciertas conexiones intermódulos, se sospecha que el complejo tendría una conectividad interna robusta, basado además en los resultados expuestos por Pocklington *et al.* [28].

Cabe finalmente mencionar que todos aquellos módulos pertenecientes a la capa de salida e híbrida tienen expresiones funcionales dentro de la célula y solamente se han omitido sus flechas de efecto por considerarse que contaminarían de sobremanera el diagrama, haciéndolo más confuso.

Lo presentando modifica la comprensión que se tenía sobre el complejo NRC/MASC. El modelo basado en una organización "humita", tal como se describe en Pocklinton *et al.* [28], Figura IV.2, da paso a una organización paralela, en donde el gran *cluster* modulador ha sido divido funcionalmente -con una alta especialización- permitiendo entonces modelar las vías de señalización de una forma más clara y sin una "caja negra" que module y transduzca las diversas señales internas del complejo. Además de entregar un primer acercamiento a cómo estas vías se organizan, al definir ciertas subestructuras dentro del complejo, como lo son las uniones de los Cluster 1-6, 3-9 y 0-3-7. No se ha llevado a cabo un proceso de designación definitivo de los distintos módulos, pues se considera que no existe una particularización en tal alto grado que permita definir completamente los *clusters* aquí presentados.

Por otro lado se considera que esta organización mucho más estática que la considerada anteriormente aunque limite la plasticidad del sistema, y por lo tanto se pueda suponer que una afección en alguno de sus elementos no podrá ser compensada por los demás elementos del sistema, entrega una particularización de las subunidades -módulos- del sistema y las vías que en él se encuentran. Se hace dificultoso, tal como se mencionó anteriormente, designar inmediatamente los roles de las vías y clusters obtenidos más allá de las inferencias obvias, pero asientan las bases para estudios en el futuro enfocados en estas proto-vías de señalización.

IV.2.4. Modificaciones en la red en individuos con enfermedades cognitivas y variaciones en las plasticidades neurológicas

Los resultados de este trabajo muestran que existen ciertos conjuntos proteicos en donde estas modificaciones afectan de mayor manera, si se realiza un catastro por enfermedad y variaciones en las plasticidades se tiene:

Bipolaridad: La minería en la literatura realizada por Grant *et al.* [11] aplicada al modelo aquí obtenido permitió obtener que los principales clusters en ser afectados cuando esta enfermedad está presente son el Cluster 1 y el Cluster 3, mientras que otros 4 módulos son afectados en menor grado. Esto permitiría suponer que la enfermedad es proclive a atacar aquellos nodos moduladores por sobre otros, pero mayor información es necesaria.

Esquizofrenia: A través de una simple inspección es fácil de ver que el principal *cluster* afectado por esta enfermedad es el módulo receptor. Mientras que otros 4 *clusters* son afectados potentemente por ella. Nuevamente los Clusters 1 y 3 aparecen, pero además de ellos están los Clusters 7 y 11. Que el módulo 11 se vea afectado no presenta mayor novedad, pues ya fue discutida su cercanía con los complejos receptores de glutamato. Bajo esa misma lógica es interesante notar como el Cluster 7, cuya probable relación con el Cluster 3 también se subrayó, es afectado de manera poderosa. Se plantea la tesis que el tipo de afección que gatilla ambas enfermedades, la bipolaridad y la esquizofrenia, es el mismo, es decir, la variación de la capacidad moduladora del sistema. Siendo esta proposición novedosa, pues anteriormente sólo se había definido una cierta especifidad en los módulos afectados cuando la esquizofrenia está presente.

Retardo Mental: El Cluster 1 se vió afectado en aquellos individuos con retardo, siendo otros módulos modificados en menor manera. Es interesante notar cómo el Cluster 1 es el mayor afectado dentro de la totalidad de enfermedades neurológicas estudiadas hasta ahora. Hipotetizando respecto a esta relación, se puede suponer que el rol fundamental del Cluster 1 en el crecimiento y organización celular fue lo que permitió esta comunión entre el *cluster* y el retardo mental. Así, según lo señalado por Pocklington *et al.* [28], el retardo -según sus resultados- se encuentra repartido en toda la unidad, y apoya la hipótesis de que el retardo afecta a todo el complejo. Esto puede ser comprendido desde otro prisma gracias a los nuevos resultados, el retardo no afectaría a la unidad en su conjunto, sino a los elementos que la organizan y dan una estructura.

Plasticidad Sináptica: Las proteínas cuya modificación producen variaciones importantes en la plasticidad sináptica se encuentran repartidas entre casi todos los módulos, en especial el Cluster 2, en donde todas, a excepción de tres proteínas, se ven involucradas. Los módulos afectados son de variados tipos, tanto efectores como moduladores. Por otro lado, los tres *clusters* que no se vieron afectados -0, 5 y 6- tiene una funcionalidad estructural y/o motora dentro del complejo, lo que nos permitiría señalar que se ha logrado una división funcional adecuada de los módulos.

Plasticidad del Aprendizaje: Las proteínas vitales para la plasticidad del aprendizaje, se encuentran dispersas en un patrón parecido al definido en el punto anterior. Permitiendo

suponer un solapamiento entre las redes fundamentales de ambas capacidades plásticas dentro de la NRC/MASC.

Se puede concluir que las proteínas efectadas en mayor medida se encuentran principalmente en el *cluster* receptor, así como en los módulos transductores de señales. Lo que permitiría suponer un modelo de interacción en donde los elementos río arriba de las vías de señalización son afectados en mayor medida, derivando esto en una modificación global del complejo. Las proteínas vitales para el buen funcionamiento de las plasticidades se hacen presente en un número mucho mayor que las proteínas afectadas cuando se ha gatillado alguna enfermedad neurológica, lo que permite pensar sobre la sensibilidad de estas capacidades neurológicas. Siendo además los clusters iniciales importantísimos en el buen funcionamiento del complejo.

Por otro lado, el alto número de proteínas involucradas en donde las enfermedades están presentes -y no la modificación de un único gen como es el paradigma molecular clásico- que afectan los *clusters* iniciales del flujo de señales en el complejo, permitiría hipotetizar sobre -y tal como se mencionó en Pocklington *et al.*- que las enfermedades neurológicas son a fin de cuentas, un cúmulo de modificaciones en la red del complejo. De allí que comprender la orgánica del complejo se vuelva una tarea fundamental si se desea atacar a las enfermedades neurológicas. Lográndose a través de este trabajo un primer paso en la definición de la funcionalidad de los módulos que componen la unidad, ya que anteriormente no había una definición particular de los distintos elementos de la unidad.

Capítulo V

Conclusiones

Del trabajo realizado en este proyecto de memoria se logró, como resultados finales y generales, definir una nueva red de interacciones interproteicas para el ser humano, además de un nuevo modelo funcional para el complejo receptor NRC/MASC.

Según lo señalado en los objetivos primarios y secundarios del proyecto se puede decir que éstos fueron logrado en su totalidad. Así, la obtención de la red de PPIs, se pudo llevar a cabo tanto en la inferencia a través de la aplicación del clasificador de Naïve-Bayes, como en la integración posterior de los sets de datos de Campos *et al.*, Rhodes *et al.* y Xia *et al.* Los *likelihood ratios* calculados y la amplitud de los resultados fueron los esperados, esto a pesar de ciertas dificultades acaecidas durante la inferencia de la red de PPIs, en especial los resultados fuera de lo esperado en el procesamiento de las matrices de co-expresión. Además de lo anterior, se debe mencionar que al utilizar bases de datos recientes los resultados se consideran más robustos, esto por el mayor grado de confianza de los distintos elementos, a raíz de la mejor curación de las bases de datos.

Aunque en un primer momento era esperable que aquellas PPIs con mayores valores de LR estuvieran documentadas en la literatura, se pudo comprobar que sólo la mitad tenía alguna referencia experimental. Este permite reforzar el carácter putativo de la red definida, en donde las interacciones con un alto LR no necesariamente son conocidas y aún persiste la probabilidad de que sean un verdadero negativo (aunque en extremo baja), así como se mantiene que las interacciones con un LR sólo un poco mayor al LR_{cut} sean probables verdaderas positivos y han de considerarse, para efectos prácticos, así. Esperándose en el futuro encontrar evidencia empírica de su existencia. De todo lo anterior es importante recalcar la importancia de las hipótesis que se utilizaron para desarrollar este trabajo, desde la independencia de las bases de datos hasta el carácter probabilístico de las redes utilizadas.

La agregación de las distintas redes de PPIs y la aplicación de restricciones sobre el parámetro O_{post} entregó un número de interacciones tres veces más grande que el número de interacciones en el GSP. demostrando con ello el poder de los clasificadores de ampliar el universo de interacciones consideradas con un alto grado de confianza. Las diferencias de las redes calculadas por los distintos estudios, aunque importantes, tuvieron una base

metodológica similar -tal como se señaló en la discusión- y no significaron una diferencia en la calidad de las redes.

A pesar que para la definición de las interacciones pertenecientes a la NRC/MASC fue hecha utilizando estudios proteómicos con 3 años de edad, esto se consideró una obligación -al igual que la utilización del algoritmo de *clustering* de Newman- para poder discutir las diferencias y similitudes entre los resultados aquí logrados y los entregados por Pocklington *et al.*, en el futuro se abre la puerta a añadir modificaciones a las distintas etapas, pues se tendrá un *background* teórico más amplio.

El punto central del trabajo -más allá de los resultados particulares de los *clusters* y su especificidad funcional- fue el modelo de la NRC/MASC obtenido. El modelo modifica totalmente la lógica que se concebía anteriormente al funcionamiento del complejo y permite un enfoque particular de las vías de señalización. Este tipo de avance evidencia el objetivo del proyecto, el cual, más que entregar un resultado final monolítico, buscaba señalar nuevas vías y obtener información (modelos) novel a través de los sets de datos conocidos, pero dándole una vuelta de tuerca a la utilización de ellos. Así se ha dejado atrás un modelo que se fundaba en un elemento "caja negra", en donde es en extremo difícil comprender de forma detallada el funcionamiento de la unidad como un todo, y se ha obtenido un sistema complejo basado en el paralelismo de las vías. Las cuales nos han permitido suponer que existe una un sistema maleable pero definido que transduce las distintas señales, todo ello con el fin de segmentar las vías de señalización y hacer particulatizar las subunidades de las densidad postsináptica.

Por otro lado, se debe mencionar que el estudio realizado sobre la incidencia de las distintas enfermedades y afecciones entregó sólo resultados superficiales, desde el comienzo del proyecto se propuso un estudio más profundo de las expresiones de las enfermedades, pero no fue posible de realizar, pues para llevarlo a cabo es necesario definir un nuevo enfoque y nuevas herramientas, las cuales no pudieron ser definidas durante la realización de esta memoria. De todas formas el modelo obtenido permite iniciar suposiciones acerca la real orgánica del sistema y plantea un punto de inflexión sobre los conocimientos existentes sobre la densidad postsináptica. Pues hay que recordar, que el fin último del proyecto es el abrir las puertas a nuevas metodologías en la lucha contra las enfermedades neurológicas y afecciones en las plasticidad neuronal. Lográndolo esto a través de encontrar las vías y elementos precisos que al ser afectados gatillan algunas de las variaciones cognitivas antes señaladas.

Se han propuesto varios puntos adicionales para lograr un estudio más acabado, los cuales no fueron considerados en los objetivos por el alto costo en horas hombre que implicarían. Así, y en primer lugar, sería necesario aplicar una serie de estudios sobre la calidad de los resultados aquí obtenidos. Por ejemplo, una comparación estadística sobre la red de PPIs inferida, en comparación con la GSP y otras bases de datos de interacciones reconocidas. Esto para demostrar la robustez de los resultados. Otro elemento que entregaría un alto número de parámetros que enriquecerían la discusión es el estudio estadístico de los *clusters*. Pocklington *et al.* realizó un profundo estudio estadístico sobre la especificidad de los clusters. Trabajo que permitiría en entendimiento mayor del modelo propuesto. Además de ampliar

el estudio a la capacidad de transmitir, la unidad NRC/MASC, variaciones LTD[1] y LTP[2], elementos centrales en la sinapsis humana.

Se ha comparado los resultados entregados por este estudio y el de Pocklington *et al.* en donde a lo largo del mismo protocolo se han definidos conjuntos proteicos completamente diferente, obteniendo este trabajo módulos y una red de interacciones con mayor especificidad que en el caso de Pocklington *et al.* Esto a pesar de utilizar ambos estudios una cantidad de datos prácticamente igual (en número de interacciones y de proteínas integrantes), lo que permite suponer que la diferencia radicó en la calidad de las interacciones utilizadas en el proceso de *clustering*. Con lo cual se apoya la tesis inicial que una inferencia a través de la combinación de bases de datos totalmente independientes permite disminuir la incertidumbre de las interacciones interproteicas.

Considerando todos los elementos antes mencionados, se debe concluir que el proyecto de memoria entregó resultados de gran calidad y totalmente novedosos, los que contribuirán al conocimiento sobre el complejo NRC/MASC, y por sobre todo, ha generado una vía en el desarrollo para nuevas terapias que logran eliminar las enfermedades cognitivas aquí señaladas. La herramienta generada entrega un primer esbozo de los mecanismo internos de las distintas enfermedades y desórdenes.

Una última consideración es que los resultados entregados son aproximaciones primeras, las cuales deben ser cuantificadas a través de estudios con un mayor grado de certeza en el futuro. Como se ha hecho mención en todo el trabajo, este modelo tiene una base probabilística, y a pesar que entrega primeras impresiones fundamentales y se ha logrado definir las directrices del trabajo futuro, se deben considerar estos resultados como preliminares y no conclusivos de una veta de investigación que recién comienza.

[1]Long-Term Depression
[2]Long-Term Potentiation

Capítulo VI

Glosario

Algoritmo clasificador: es un algoritmo que separa los distintos elementos de un conjunto en clases en función de sus atributos.

Clasificador Naïve-Bayes: algoritmo clasificador probabilístico con una fuerte hipótesis de independencia (de allí el término *naïve*), que jerarquiza utilizando como base el teorema de Bayes.

***Clustering*:** es un proceso de clasificación en donde los componentes son asignados a subconjuntos en función de sus particularidades, así en un subconjunto (*cluster*) todos los elementos son similares de acuerdo a alguna condición (distancia). Los algoritmos de clustering más conocidos son los de partición (en especial el *k-means*) y los jerarquizantes. Este proceso también puede ser aplicados a grafos.

Complejo NRC/MASC: complejo proteico perteneciente a proteómica postsináptica, en donde se encuentra el complejo receptor de glutamato NMDA-R, también llamado NRC, y las proteínas MAGUK. Fundamental en la plasticidad sináptica y del...

Densidad Postsináptica: conjunto de proteínas ubicado debajo de la membrana postsináptica, todo ello en la terminal postináptica. Está altamente enriquecida con dominios de interacción interproteicas.

Glutamato: principal neurotransmisor, fundamental en la plasticidad sináptica y por lo tanto en los procesos cognitivos.

Interacción interproteica: relación entre dos proteínas, dependiendo del tipo de clasificador a utilizar será el nivel de la relación para aceptar la interacción. Así en el grado más estricto sólo se considerará la interacción física.

Interacciones ortológicas: interacciones interproteicas en dos especies distintas, las cuales se derivaron de una interacción en particular en un ancestro común.

Plasticidad sináptica: capacidad de modificar la fuerza de la conexión, sinapsis, entre dos neuronas.

Plasticidad del comportamiento: capacidad de la red de modificar el comportamiento de individuo, enfocado en la capacidad de aprendizaje de éste.

Proteómica Postsináptica: Conjunto de proteínas que se encuentran en la terminal postsináptica.

Receptores de glutamato: conjunto de receptores ubicados en la terminal postsináptica, estos se dividen en función del antagonista que se acopla al receptor y del tipo de señal que éste inicia. Así dos de ellos forman canales ionotrópicos (teniendo como antagonistas NMDA y AMPA respectivamente) y uno está unido a proteínas G, a través de las cuales se transduce la señal.

SSBP: *Smallest shared biological process*, se refiere al proceso en común más pequeño en que dos proteínas se ven involucradas.

Referencias

[1] ALBERTS, B., JOHNSON, A., LEWIS, J., RAFF, M., ROBERTS, K. y WALTER, P., 2002, Molecular Biology of the Cell, fourth edition. Garland Science.

[2] BROWNE, F., WANG, H., ZHENG, H. y AZUAJE, F., 2009, GRIP: A web-based system for constructing Gold Standard datasets for protein-protein interaction prediction. Source Code Biol Med, 4:2.

[3] CHEN, X., CHEUNG, S. T., SO, S., FAN, S. T., BARRY, C., HIGGINS, J., LAI, K.-M., JI, J., DUDOIT, S., NG, I. O., VAN DE RIJN, M., BOTSTEIN, D. y BROWN, P. O., 2002, Gene Expression Patterns in Human Liver Cancers. Mol. Biol. Cell, 13(6):1929–1939.

[4] CHOUDHARY, J. y GRANT, S., 2004, Proteomics in postgenomic neuroscience: the end of the beginning. Nature Neuroscience, 7(5):440–445.

[5] COLLINS, M. O., HUSI, H., YU, L., BRANDON, J. M., ANDERSON, C. N., BLACKSTOCK, W. P., CHOUDHARY, J. S. y GRANT, S. G., 2006, Molecular characterization and comparison of the components and multiprotein complexes in the postsynaptic proteome. J Neurochem, 97 Suppl 1:16–23.

[6] COMPENDIA BIOSCIENCE, Oncomine. `http://www.oncomine.org`, [consulta: diciembre 2009].

[7] EMBL - EBI y WELLCOME TRUST SANGER INSTITUTE, Ensembl. `http://www.ensembl.org`, [consulta: agosto 2009].

[8] GIOT, L., BADER, J. S., BROUWER, C., CHAUDHURI, A., KUANG, B., LI, Y., HAO, Y. L., OOI, C. E., GODWIN, B., VITOLS, E., VIJAYADAMODAR, G., POCHART, P., MACHINENI, H., WELSH, M., KONG, Y., ZERHUSEN, B., MALCOLM, R., VARRONE, Z., COLLIS, A., MINTO, M., BURGESS, S., MCDANIEL, L., STIMPSON, E., SPRIGGS, F., WILLIAMS, J., NEURATH, K., IOIME, N., AGEE, M., VOSS, E., FURTAK, K., RENZULLI, R., AANENSEN, N., CARROLLA, S., BICKELHAUPT, E., LAZOVATSKY, Y., DASILVA, A., ZHONG, J., STANYON, C. A., FINLEY, J., R. L., WHITE, K. P., BRAVERMAN, M., JARVIE, T., GOLD, S., LEACH, M., KNIGHT, J., SHIMKETS, R. A., MCKENNA, M. P., CHANT, J. y ROTHBERG, J. M., 2003, A Protein Interaction Map of Drosophila melanogaster. Science, 302(5651):1727–1736.

[9] GOFF, D. C. y COYLE, J. T., 2001, The emerging role of glutamate in the pathophysiology and treatment of schizophrenia. Am J Psychiatry, 158(9):1367–77.

[10] GÖHLMANN, H. y TALLOEN, W., 2009, Gene Expression Studies Using Affymetrix Microarrays. 1ª edición, Chapman & Hall.

[11] GRANT, S. G., MARSHALL, M. C., PAGE, K. L., CUMISKEY, M. A. y ARMSTRONG, J. D., 2005, Synapse proteomics of multiprotein complexes: en route from genes to nervous system diseases. Hum Mol Genet, 14 Spec No. 2:R225–34.

[12] GRANT, S. G. y O'DELL, T. J., 2001, Multiprotein complex signaling and the plasticity problem. Curr Opin Neurobiol, 11(3):363–8.

[13] GUAN, Y., MYERS, C. L., LU, R., LEMISCHKA, I. R., BULT, C. J. y TROYANSKAYA, O. G., 2008, A genomewide functional network for the laboratory mouse. PLoS Comput Biol, 4(9):e1000165.

[14] HELMHOLTZ ZENTRUM MÜNCHEN, Munich Information Center for Protein Sequences. `http://www.mips.helmholtz-muenchen.de`, [consulta: enero 2010].

[15] HUSI, H., WARD, M. A., CHOUDHARY, J. S., BLACKSTOCK, W. P. y GRANT, S. G., 2000, Proteomic analysis of NMDA receptor-adhesion protein signaling complexes. Nat Neurosci, 3(7):661–9.

[16] JANSEN, R., YU, H., GREENBAUM, D., KLUGER, Y., KROGAN, N. J., CHUNG, S., EMILI, A., SNYDER, M., GREENBLATT, J. F. y GERSTEIN, M., 2003, A Bayesian Networks Approach for Predicting Protein-Protein Interactions from Genomic Data. Science, 302(5644):449–453.

[17] KANEHISA LABORATORIES, Kyoto Encyclopedia of Genes and Genomes. `http://www.genome.jp/kegg/`, [consulta: enero 2010].

[18] LANG, U., PULS, I., MÜLLER, D. y STRUTZ-SEEBOHM, N., 2007, Molecular mechanisms of schizophrenia. Cellular Physiology and Biochemistry, 20(6):687–702.

[19] LAUMONNIER, F., CUTHBERT, P. C. y GRANT, S. G., 2007, The role of neuronal complexes in human X-linked brain diseases. Am J Hum Genet, 80(2):205–20.

[20] LI, S., ARMSTRONG, C. M., BERTIN, N., GE, H., MILSTEIN, S., BOXEM, M., VIDALAIN, P.-O., HAN, J.-D. J., CHESNEAU, A., HAO, T., GOLDBERG, D. S., LI, N., MARTINEZ, M., RUAL, J.-F., LAMESCH, P., XU, L., TEWARI, M., WONG, S. L., ZHANG, L. V., BERRIZ, G. F., JACOTOT, L., VAGLIO, P., REBOUL, J., HIROZANE-KISHIKAWA, T., LI, Q., GABEL, H. W., ELEWA, A., BAUMGARTNER, B., ROSE, D. J., YU, H., BOSAK, S., SEQUERRA, R., FRASER, A., MANGO, S. E., SAXTON, W. M., STROME, S., VAN DEN HEUVEL, S., PIANO, F., VANDENHAUTE, J., SARDET, C., GERSTEIN, M., DOUCETTE-STAMM, L., GUNSALUS, K. C., HARPER, J. W., CUSICK, M. E., ROTH, F. P., HILL, D. E. y VIDAL, M., 2004, A Map of the Interactome Network of the Metazoan C. elegans. Science, 303(5657):540–543.

[21] MATSUZAKI, S. y TOHYAMA, M., 2007, Molecular mechanism of schizophrenia with reference to disrupted-in-schizophrenia 1 (DISC1). Neurochem Int, 51(2-4):165–72.

[22] MINORETTI, P., POLITI, P., COEN, E., DI VITO, C., BERTONA, M., BIANCHI, M. y EMANUELE, E., 2006, The T393C polymorphism of the GNAS1 gene is associated with deficit schizophrenia in an Italian population sample. Neurosci Lett, 397(1-2):159–63.

[23] NATIONAL HUMAN GENOME RESEARCH INSTITUTE, Gene ontology. `http://www.geneontology.org/`, [consulta: diciembre 2009].

[24] NEWMAN, M. E. y GIRVAN, M., 2004, Finding and evaluating community structure in networks. Phys Rev E Stat Nonlin Soft Matter Phys, 69(2 Pt 2):026113.

[25] ONTARIO INSTITUTE FOR CANCER RESEARCH y EUROPEAN BIOINFORMATICS INSTITUTE, BioMart. `http://www.biomart.org/`, [consulta: enero 2010].

[26] PANDEYLAB y INSTITUTE OF BIOINFORMATICS, Human protein reference database. `www.hprd.org/`, [consulta: agosto 2009].

[27] POCKLINGTON, A. J., ARMSTRONG, J. D. y GRANT, S. G., 2006, Organization of brain complexity–synapse proteome form and function. Brief Funct Genomic Proteomic, 5(1):66–73.

[28] POCKLINGTON, A. J., CUMISKEY, M., ARMSTRONG, J. D. y GRANT, S. G., 2006, The proteomes of neurotransmitter receptor complexes form modular networks with distributed functionality underlying plasticity and behaviour. Mol Syst Biol, 2:2006.0023.

[29] QI, Y., BAR-JOSEPH, Z. y KLEIN-SEETHARAMAN, J., 2006, Evaluation of different biological data and computational classification methods for use in protein interaction prediction. Proteins: Structure, Function, and Bioinformatics, 63(3):490–500.

[30] RHODES, D., TOMLINS, S., VARAMBALLY, S., MAHAVISNO, V., BARRETTE, T., KALYANA-SUNDARAM, S., GHOSH, D., PANDEY, A. y CHINNAIYAN, A., 2005, Probabilistic model of the human protein-protein interaction network. Nat Biotech, 23(8):951–959.

[31] ROSENWALD, A., WRIGHT, G., CHAN, W. C., CONNORS, J. M., CAMPO, E., FISHER, R. I., GASCOYNE, R. D., MULLER-HERMELINK, H. K., SMELAND, E. B., GILTNANE, J. M. y HURT, E. M., 2002, The Use of Molecular Profiling to Predict Survival after Chemotherapy for Diffuse Large-B-Cell Lymphoma. N Engl J Med, 346(25):1937–1947.

[32] SCOTT, M. S. y BARTON, G. J., 2007, Probabilistic prediction and ranking of human protein-protein interactions. BMC Bioinformatics, 8:239.

[33] SEGAL, N. H., PAVLIDIS, P., NOBLE, W. S., ANTONESCU, C. R., VIALE, A., WESLEY, U. V., BUSAM, K., GALLARDO, H., DESANTIS, D., BRENNAN, M. F., CORDON-CARDO, C., WOLCHOK, J. D. y HOUGHTON, A. N., 2003, Classification of Clear-Cell Sarcoma as a Subtype of Melanoma by Genomic Profiling. J Clin Oncol, 21(9):1775–1781.

[34] STOCKHOLM BIOINFORMATICS CENTRE, Inparanoid: Eukaryotic ortholog groups. `http://inparanoid.sbc.su.se/`, [consulta: agosto 2009].

[35] SU, A. I., WELSH, J. B., SAPINOSO, L. M., KERN, S. G., DIMITROV, P., LAPP, H., SCHULTZ, P. G., POWELL, S. M., MOSKALUK, C. A., FRIERSON, J., HENRY F. y HAMPTON, G. M., 2001, Molecular Classification of Human Carcinomas by Use of Gene Expression Signatures. Cancer Res, 61(20):7388–7393.

[36] THE WELLCOME TRUST MEDICAL LIBRARY y DOLAN DNA LEARNING CENTER, Genes2Cognition. `http://www.genes2cognition.org/`, [consulta: febrero 2010].

[37] UNIPROT CONSORTIUM, Uniprot. `http://www.uniprot.org/`, [consulta: febrero 2010].

[38] UNIVERSITY OF CALIFORNIA, Dataset of interacting proteins. `http://dip.doe-mbi.ucla.edu`, [consulta: agosto 2009].

[39] VAN 'T VEER, L., DAI, H., VAN DE VIJVER, M., HE, Y. y HART, A., 2002, Gene expression profiling predicts clinical outcome of breast cancer. Nature, 415(6871):530–536.

[40] XIA, K., DONG, D. y HAN, J.-D., 2006, Intnetdb v1.0: an integrated protein-protein interaction network database generated by a probabilistic model. BMC Bioinformatics, 7(1):508.

Apéndices

A . Algoritmo de *Bootstraping*

El algoritmo de *bootstraping* tiene por fin poder calificar cuantitativamente estadistísticos sobre una población. En particular para este proyecto es necesario la aplicación de este algoritmo para las medias aritméticas. El algoritmo permitirá obtener una estimación del error estándar del estadístico en cuestión y tal como se mencionó en este caso se aplicará para estimar el error del promedio.

El procedimiento utilizado en este caso en particular es expuesto en forma general en el siguiente diagrama de la Figura VI.1.

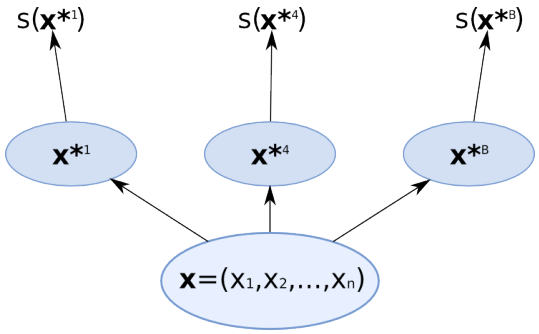

Figura VI.1: Diagrama del mecanismo de *bootstraping*

Para estimar el error estándar se selecciona en primer lugar distintas muestras aleatorias con reemplazo (x^{*b}) desde el vector de muestras X. Cada una de estas muestras contiene n elementos y tal como se ha señalado, éstas pueden estar conformadas por cualquier combinación de elementos de X con reemplazo (repetición).

Una vez definido las B muestras se procede a obtener el estadístico, en este caso la media, de cada una de las muestras $s(x^{*b})$.

El error estándar estimado se define como:

$$\widehat{se}_{boot} = \left\{ \frac{\sum_{b=1}^{B} \left[s(x^{*b}) - s(\cdot) \right]^2}{B-1} \right\}^{\frac{1}{2}} \tag{VI.1}$$

En donde

$$s(\cdot) = \sum_{b=1}^{B} \frac{s(x^{*b})}{B} \tag{VI.2}$$

Entre más grande B se vuelva, ie. $B \to \infty$, la aproximación tenderá al error estándar real.

Con ello lo obtiene una distribución de la media de las distintas muestras aleatorias, con ello se calcula tanto la media como la desviación estándar de ella. Así pues, se definirán los límites sobre los cuales los elementos son estadísticamente significativos. Los que para este proyecto fueron definidos como aquellos que estén en el 95 % de la distribución:

$$\text{Límite derecho} = \bar{x} + 1.96\sigma$$

$$\text{Límite izquierdo} = \bar{x} - 1.96\sigma$$

A continuación se muestra el código escrito para realizar este procedimiento en el programa R-project:

```
boot.mean<-function(x){
l<-length(x)
boot.final<-numeric(l)

boot.sample<-numeric(1000)
for (h in 1:1000){
boot.sample[h]<-mean((x[sample(l,l,replace=T)]),na.rm=T)
}
mean<-mean(boot.sample,na.rm=T)
sd<-sd(boot.sample,na.rm=T)
li<-mean-1.86*sd
ld<-mean+1.86*sd
boot.final<-mean((x[x>=li & x<=ld]),na.rm=T)
if (is.na(boot.final)) boot.final<-mean

return(boot.final)
}
```

B . *Clustering* de la unidad NRC/MASC

Se muestra a continuación el código implementado para optimizar el *clustering* en función de Q:

```
library("igraph")
library("RPostgreSQL")
con<-dbConnect("PostgreSQL",dbname="clustering")
query<-dbGetQuery(con,"Select p_1,p_2 from infe_1")
prots<-dbGetQuery(con,"Select p_1 from infe_1 union select p_2 from infe_1")
grafo<-graph.data.frame(query,directed=F,vertices=prots)
cluster<-edge.betweenness.community(grafo,directed=F)
mod<-array(NA,dim=119)

for (i in 1:119){
comunidad<-community.to.membership(grafo,cluster[["merges"]],i)
mod[i]<-modularity(grafo,comunidad$membership)}
max<-max(mod)
n_max<-which.max(mod)
comunidad<-community.to.membership(grafo,cluster[["merges"]],n_max)
modu<-modularity(grafo,comunidad$membership)
clases<-cbind(prots[[1]],comunidad$membership)
write.table(clases,file="clases.tab",quote=F,col.names=F,row.names=F,sep="\t")
```

C . Resultados Inferencia Interacciones

C .1. Inferencia por Ortología

Tabla VI.1: Clases y LR para la inferencia por ortología

Dataset	Class	GSP	GSN	Total	Pr(CL|GSP)	Pr(CL|GSN)	LR
S. Cerevisiae	Todos	466	257	15425	0.238	0.00274	86.841
	$E_V > 1$	276	20	2436	0.14096	0.00021	660.927
	$N = 1$	97	16	2168	0.04954	0.00017	290.353
	$N > 28$	1	28	2935	0.00051	0.0003	1.71
	$1 < N <= 28$	92	193	7886	0.04699	0.00206	22.83
	posibles	1958	93775				
D. Melanogastar	Todos	24	13	462	0.14634	0.00824	17.764
	Conf > 0.55	22	1	89	0.13415	0.00063	222.824
	Conf < 0.55	2	13	373	0.0122	0.00824	1.48
	posible	164	1578				
C. Elegans	Todos	37	10	1005	0.16372	0.00212	77.078
	posible	226	4708				

C .2. Inferencia por CoExpresión

Tabla VI.2: Clases y LR para la inferencia por coexpresión

Dataset	R≥	GSP	GSN	TOTAL	Pr(R|GSP)	Pr(R|GSN)	LR
Rosenwald	-1	0	0	0	0	0	0
	-0.9	0	0	0	0	0	0
	-0.8	0	0	0	0	0	0
	-0.7	0	2	6	0	0.00001	0
	-0.6	0	27	184	0	0.00017	0
	-0.5	4	546	4237	0.00096	0.00335	0.2861
	-0.4	34	3562	32650	0.00816	0.02188	0.37276
	-0.3	134	13105	125786	0.03215	0.08051	0.39932
	-0.2	346	26817	291872	0.08301	0.16475	0.50387
	-0.1	650	35236	454523	0.15595	0.21648	0.7204
	0	768	32931	505976	0.18426	0.20231	0.91076
	0.1	754	24369	426155	0.1809	0.14971	1.20832
	0.2	584	14730	287132	0.14012	0.0905	1.54832
	0.3	408	7291	163018	0.09789	0.04479	2.18536
	0.4	259	2979	78187	0.06214	0.0183	3.3953
	0.5	126	950	30425	0.03023	0.00584	5.1796
	0.6	67	203	8438	0.01607	0.00125	12.88926
	0.7	30	21	1367	0.0072	0.00013	55.78935
	0.8	3	2	106	0.00072	0.00001	58.57881
	0.9	1	0	9	0.00024	0	0
	1	0	0	0	0	0	0
	SUM	4168	162771	2410071			
Segal	-1	0	5	89	0	0.00002	0
	-0.9	5	212	5091	0.0018	0.0009	2.00768
	-0.8	13	1341	33955	0.00468	0.00567	0.82523
	-0.7	52	3332	88125	0.01871	0.01408	1.32849
	-0.6	83	5507	149588	0.02987	0.02328	1.28299
	-0.5	121	9151	231666	0.04354	0.03868	1.12558
	-0.4	183	15025	364668	0.06585	0.06351	1.0368
	-0.3	240	22961	585678	0.08636	0.09706	0.88978
	-0.2	332	31201	780415	0.11947	0.13189	0.90579
	-0.1	338	33561	839004	0.12163	0.14187	0.85732
	0	339	30442	768331	0.12199	0.12868	0.94795
	0.1	256	24886	630550	0.09212	0.1052	0.87568
	0.2	207	18619	475775	0.07449	0.07871	0.9464
	0.3	173	13999	343822	0.06225	0.05918	1.05198
	0.4	163	10204	248597	0.05865	0.04313	1.35981
	0.5	119	7451	184663	0.04282	0.0315	1.35954
	0.6	82	4813	130593	0.02951	0.02035	1.4503
	0.7	51	2626	77364	0.01835	0.0111	1.65324
	0.8	18	1003	30454	0.00648	0.00424	1.52768
	0.9	4	225	7559	0.00144	0.00095	1.51334
	1	0	0	0	0	0	0
	SUM	2779	236564	5975987			

Tabla VI.3: Clases y LR para la inferencia por coexpresión

Dataset	R≥	GSP	GSN	TOTAL	Pr(R|GSP)	Pr(R|GSN)	LR
Su	-1	0	0	0	0	0	0
	-0.9	0	0	7	0	0	0
	-0.8	0	36	610	0	0.00005	0
	-0.7	2	329	6196	0.00024	0.00043	0.56032
	-0.6	5	1910	33855	0.0006	0.00248	0.24129
	-0.5	50	7219	130292	0.00599	0.00939	0.6384
	-0.4	145	21652	439706	0.01738	0.02815	0.61726
	-0.3	474	60850	1307521	0.05681	0.07912	0.71799
	-0.2	1173	136983	3016973	0.14058	0.17811	0.78928
	-0.1	1924	192523	4439238	0.23058	0.25033	0.92113
	0	1866	163256	3863757	0.22363	0.21227	1.05352
	0.1	1260	95745	2396940	0.15101	0.12449	1.21298
	0.2	702	46488	1206886	0.08413	0.06045	1.39186
	0.3	376	21583	565372	0.04506	0.02806	1.60575
	0.4	174	10799	284076	0.02085	0.01404	1.48514
	0.5	81	5646	153646	0.00971	0.00734	1.32234
	0.6	43	2692	72649	0.00515	0.0035	1.47229
	0.7	39	1058	28110	0.00467	0.00138	3.39766
	0.8	17	303	7630	0.00204	0.00039	5.17139
	0.9	13	14	847	0.00156	0.00002	85.5886
	1	0	0	0	0	0	0
	SUM	8344	769086	17954311			
Vant	-1	0	0	2	0	0	0
	-0.9	1	3	486	0.00365	0.00013	0
	-0.8	0	16	1637	0	0.00071	0
	-0.7	0	51	2821	0	0.00226	0
	-0.6	3	69	5603	0.01095	0.00306	3.57997
	-0.5	2	173	13892	0.0073	0.00767	0.9519
	-0.4	9	524	36101	0.03285	0.02323	1.41423
	-0.3	14	1411	91430	0.05109	0.06254	0.81698
	-0.2	38	3356	219472	0.13869	0.14875	0.93233
	-0.1	58	5535	380718	0.21168	0.24533	0.86282
	0	64	5573	380189	0.23358	0.24702	0.94558
	0.1	35	3326	222076	0.12774	0.14742	0.86647
	0.2	28	1418	101328	0.10219	0.06285	1.62588
	0.3	11	640	44961	0.04015	0.02837	1.41521
	0.4	4	257	19865	0.0146	0.01139	1.28155
	0.5	6	111	8626	0.0219	0.00492	4.45078
	0.6	0	70	3912	0	0.0031	0
	0.7	0	24	1891	0	0.00106	0
	0.8	1	3	548	0.00365	0.00013	27.44647
	0.9	0	1	67	0	0.00004	0
	1	0	0	0	0	0	0
	SUM	274	22561	1535625			

C .3. Funciones biológicas compartidas

Tabla VI.4: Clases y `LR` para la inferencia por proceso biológico compartido

SSBP ≤	GSP	GSN	Total	Pr(CL|GSP)	Pr(CL|GSN)	LR
5	1174	300	10314	0.03832	0.00007	516.693
10	627	760	15535	0.02046	0.00019	108.928
50	2474	8201	155953	0.08075	0.00203	39.831
100	1210	7836	207487	0.03949	0.00194	20.388
500	3824	65582	1795498	0.12481	0.01621	7.699
1000	911	27270	1813052	0.02973	0.00674	4.411
2000	1166	48431	1788306	0.03806	0.01197	3.179
posibles	30639	4045389				

C .4. Pares de dominio enriquecido

Tabla VI.5: Clases y `LR` para la inferencia por dominio enriquecido, *dataset* 1

Dataset 1						
D >	GSP	GSN	Total	Pr(CL|GSP)	Pr(CL|GSN)	LR
30	1236	9228	321659	0.15685	0.00554	28.306
15	594	14334	399599	0.07538	0.00861	8.758
5	1409	66255	1587564	0.17881	0.03979	4.494
2	1185	168940	2604635	0.15038	0.10145	1.482
posibles	7880	1665312				

Tabla VI.6: Clases y `LR` para la inferencia por dominio enriquecido, *dataset* 2

Dataset 2						
D >	GSP	GSN	Total	Pr(CL|GSP)	Pr(CL|GSN)	LR
30	1858	8072	266134	0.21174	0.00499	42.430
15	754	12586	322714	0.08593	0.00778	11.043
5	1694	67072	1494354	0.19305	0.04147	4.656
2	1476	183061	2941520	0.16821	0.11317	1.486
posibles	8775	1617554				

Tabla VI.7: Clases y LR para la inferencia por dominio enriquecido, *dataset* 3

Dataset 3						
D >	GSP	GSN	Total	Pr(CL|GSP)	Pr(CL|GSN)	LR
30	756	1026	55429	0.08084	0.00062	130.196
15	410	2139	76518	0.04384	0.00129	33.868
5	852	11865	353786	0.0911	0.00718	12.688
2	1547	31637	734241	0.16542	0.01915	8.640
posibles	9352	1652446				

D . Histogramas de distribución de los *likelihood ratios* obtenidos y los pertenecientes a los estudios de Rhodes *et al.* y Xia *et al.*

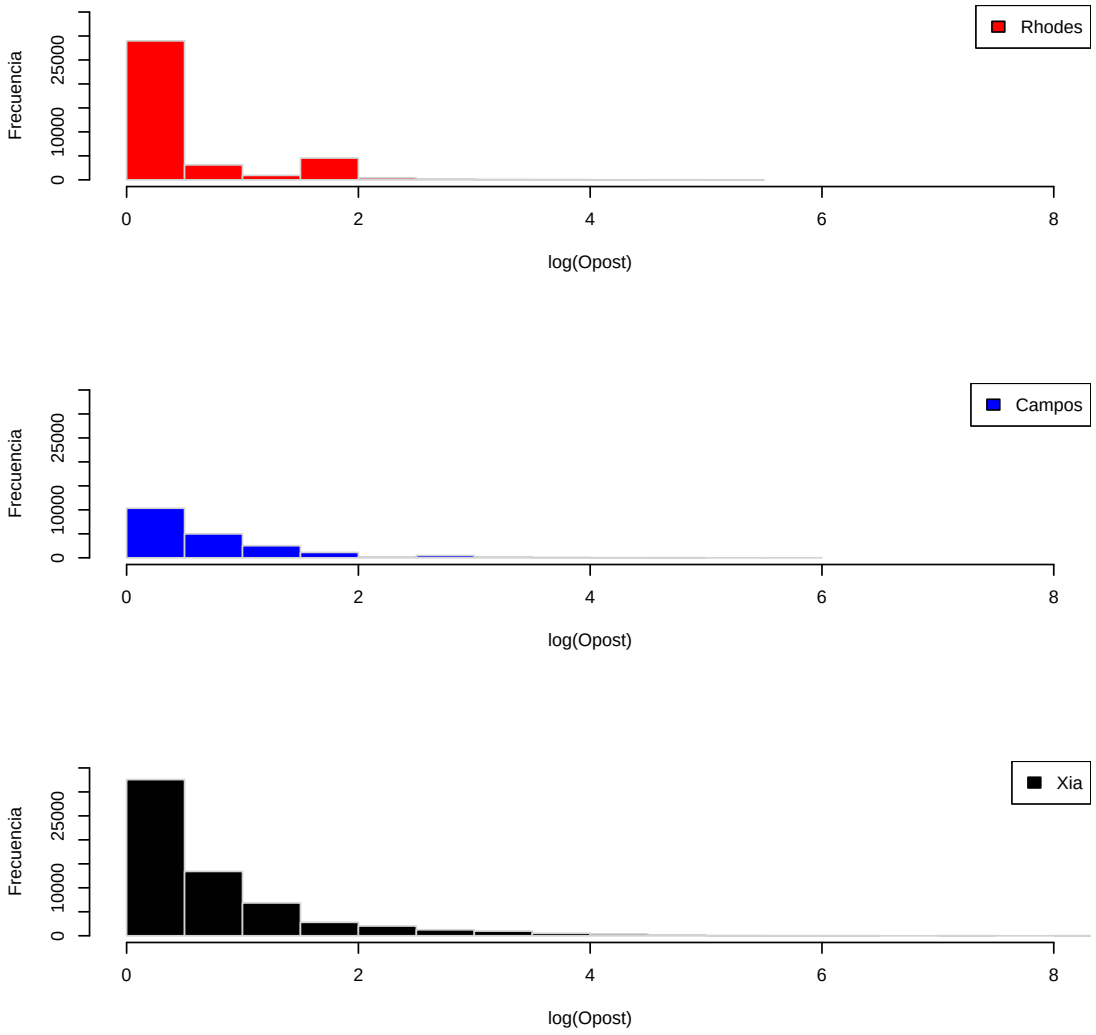

Figura VI.2: Histograma de los LR para cada estudio en separado

E . Red de interacciones del complejo NRC/MASC

El diagrama detallado de las proteínas y los clusters se muestra a continuación:

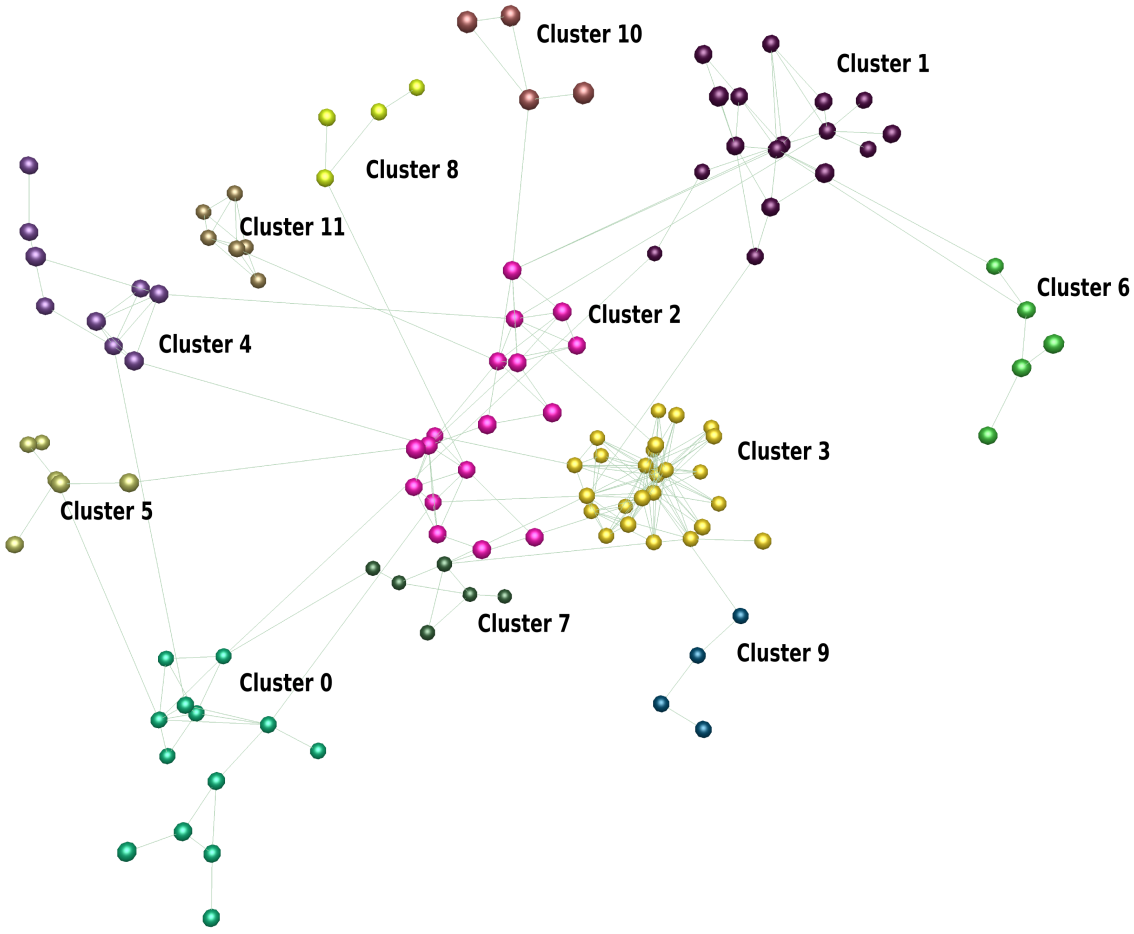

Figura VI.3: Diagrama detallado de los *clusters* obtenidos

F . Clusters obtenidos del complejo NRC/MASC

Tabla VI.8: Proteínas pertenecientes al Cluster 0

Índice	Nombre	Familia	Subfamilia
36	CSE1L	Cell Adhesion and Cytoskeletal	Other Cytoskeletal Proteins
59	Ran	G-proteins and Modulators	G-proteins
72	GSK3 beta	Kinases	Ser/Thr Kinases
79	PDK-1	Kinases	Ser/Thr Kinases
80	PKA-R2b	Kinases	Ser/Thr Kinases
84	PRKACB	Kinases	Ser/Thr Kinases
112	PPP1CC	Protein Phosphatases	Protein Phosphatases
178	SLC25A5	Synaptic Vesicles / Protein Transport	Transporters
182	D-Prohibitin	Transcription and Translation	Transcription Elements
219	PKA-R1a-a	Kinases	Ser/Thr Kinases
220	PKA-R1a/b	Kinases	Ser/Thr Kinases
221	PKA-R2a	Kinases	Ser/Thr Kinases
222	PRKACA	Kinases	Ser/Thr Kinases

Tabla VI.9: Proteínas pertenecientes al Cluster 1

Índice	Nombre	Familia	Subfamilia
13	Actin	Cell Adhesion and Cytoskeletal	Actin / ARP
14	ARPC2	Cell Adhesion and Cytoskeletal	Actin / ARP
15	ARPC3	Cell Adhesion and Cytoskeletal	Actin / ARP
16	ARPC4	Cell Adhesion and Cytoskeletal	Actin / ARP
32	CAPZ alpha	Cell Adhesion and Cytoskeletal	Other Cytoskeletal Proteins
33	CAPZ beta	Cell Adhesion and Cytoskeletal	Other Cytoskeletal Proteins
53	Rab2	G-proteins and Modulators	G-proteins
56	Rab6A	G-proteins and Modulators	G-proteins
57	Rac1	G-proteins and Modulators	G-proteins
68	CamKIIalpha	Kinases	Ser/Thr Kinases
110	PP2B	Protein Phosphatases	Protein Phosphatases
135	PI3-K	Signalling molecules and Enzymes	Other enzymes
140	ARF3	Signalling molecules and Enzymes	Other signalling molecules
143	Calmodulin	Signalling molecules and Enzymes	Other signalling molecules
159	MYH10	Synaptic Vesicles / Protein Transport	Motor Proteins
162	MYH9	Synaptic Vesicles / Protein Transport	Motor Proteins
165	Myosin (V)	Synaptic Vesicles / Protein Transport	Motor Proteins

Tabla VI.10: Proteínas pertenecientes al Cluster 2

Índice	Nombre	Familia	Subfamilia
2	GluR6	Channels and Receptors	Glutamate Receptors
3	mGluR1	Channels and Receptors	Glutamate Receptors
4	mGluR5	Channels and Receptors	Glutamate Receptors
5	NR1	Channels and Receptors	Glutamate Receptors
6	NR2A	Channels and Receptors	Glutamate Receptors
7	NR2B	Channels and Receptors	Glutamate Receptors
31	Bassoon	Cell Adhesion and Cytoskeletal	Other Cytoskeletal Proteins
54	Rab3	G-proteins and Modulators	G-proteins
69	CamKIIbeta	Kinases	Ser/Thr Kinases
103	Homer	MAGUKs / Adaptors / Scaffolders	PDZ-domain containing scaffolders
166	Dynamin	Synaptic Vesicles / Protein Transport	Synaptic vesicle
168	NSF	Synaptic Vesicles / Protein Transport	Synaptic vesicle
169	SNAP25	Synaptic Vesicles / Protein Transport	Synaptic vesicle
171	STX	Synaptic Vesicles / Protein Transport	Synaptic vesicle
172	STXBP1	Synaptic Vesicles / Protein Transport	Synaptic vesicle
173	Synaptogyrin	Synaptic Vesicles / Protein Transport	Synaptic vesicle
174	SYT1	Synaptic Vesicles / Protein Transport	Synaptic vesicle

Tabla VI.11: Proteínas pertenecientes al Cluster 3

Índice	Nombre	Familia	Subfamilia
49	GNB1	G-proteins and Modulators	G-proteins
50	GNB2	G-proteins and Modulators	G-proteins
51	GNB4	G-proteins and Modulators	G-proteins
190	GNA11	G-proteins and Modulators	G-proteins
193	GNA14	G-proteins and Modulators	G-proteins
195	GNAI1	G-proteins and Modulators	G-proteins
196	GNAI2	G-proteins and Modulators	G-proteins
197	GNAI3	G-proteins and Modulators	G-proteins
199	GNAO1	G-proteins and Modulators	G-proteins
200	GNAQ	G-proteins and Modulators	G-proteins
201	GNAT1	G-proteins and Modulators	G-proteins
202	GNAT2	G-proteins and Modulators	G-proteins
204	GNAZ	G-proteins and Modulators	G-proteins
205	GNB3	G-proteins and Modulators	G-proteins
206	GNB5	G-proteins and Modulators	G-proteins
207	GNG1	G-proteins and Modulators	G-proteins
208	GNG10	G-proteins and Modulators	G-proteins
209	GNG11	G-proteins and Modulators	G-proteins
210	GNG12	G-proteins and Modulators	G-proteins
211	GNG13	G-proteins and Modulators	G-proteins
212	GNG2	G-proteins and Modulators	G-proteins
213	GNG3	G-proteins and Modulators	G-proteins
214	GNG4	G-proteins and Modulators	G-proteins
215	GNG5	G-proteins and Modulators	G-proteins
216	GNG7	G-proteins and Modulators	G-proteins

Tabla VI.12: Proteínas pertenecientes al Cluster 4

Índice	Nombre	Familia	Subfamilia
89	14-3-3epsilon	MAGUKs / Adaptors / Scaffolders	14-3-3
90	14-3-3eta	MAGUKs / Adaptors / Scaffolders	14-3-3
92	14-3-3zeta	MAGUKs / Adaptors / Scaffolders	14-3-3
111	PP5	Protein Phosphatases	Protein Phosphatases
113	PPP2CA	Protein Phosphatases	Protein Phosphatases
114	PPP2R1A	Protein Phosphatases	Protein Phosphatases
187	14-3-3beta/alpha	MAGUKs / Adaptors / Scaffolders	14-3-3
188	14-3-3sigma	MAGUKs / Adaptors / Scaffolders	14-3-3
189	14-3-3theta/tau	MAGUKs / Adaptors / Scaffolders	14-3-3

Tabla VI.13: Proteínas pertenecientes al Cluster 5

Índice	Nombre	Familia	Subfamilia
37	DBN1	Cell Adhesion and Cytoskeletal	Other Cytoskeletal Proteins
52	H-Ras	G-proteins and Modulators	G-proteins
64	Kalirin	G-proteins and Modulators	Modulators
65	NF1	G-proteins and Modulators	Modulators
67	AKT2	Kinases	Ser/Thr Kinases
149	IRS-1	Signalling molecules and Enzymes	Other signalling molecules

Tabla VI.14: Proteínas pertenecientes al Cluster 6

Índice	Nombre	Familia	Subfamilia
17	ACTN	Cell Adhesion and Cytoskeletal	actinin
18	ACTN3	Cell Adhesion and Cytoskeletal	actinin
19	ACTN4	Cell Adhesion and Cytoskeletal	actinin
160	MYH11	Synaptic Vesicles / Protein Transport	Motor Proteins
161	MYH7	Synaptic Vesicles / Protein Transport	Motor Proteins

Tabla VI.15: Proteínas pertenecientes al Cluster 7

Índice	Nombre	Familia	Subfamilia
70	ERK1	Kinases	Ser/Thr Kinases
71	ERK2	Kinases	Ser/Thr Kinases
75	MAPKp49	Kinases	Ser/Thr Kinases
76	MEK1	Kinases	Ser/Thr Kinases
77	MEK2	Kinases	Ser/Thr Kinases
109	MKP2	Protein Phosphatases	Protein Phosphatases

Tabla VI.16: Proteínas pertenecientes al Cluster 8

Índice	Nombre	Familia	Subfamilia
81	PKCbeta	Kinases	Ser/Thr Kinases
82	PKCepsilon	Kinases	Ser/Thr Kinases
83	PKCgamma	Kinases	Ser/Thr Kinases
85	RAF1	Kinases	Ser/Thr Kinases

Tabla VI.17: Proteínas pertenecientes al Cluster 9

Índice	Nombre	Familia	Subfamilia
34	Cortactin	Cell Adhesion and Cytoskeletal	Other Cytoskeletal Proteins
87	PYK2	Kinases	Tyr Kinase
88	Src	Kinases	Tyr Kinase
96	Grb2	MAGUKs / Adaptors / Scaffolders	non-PDZ-domain containg scaffolders

Tabla VI.18: Proteínas pertenecientes al Cluster 10

Índice	Nombre	Familia	Subfamilia
1	ATP2B4	Channels and Receptors	Ca2+-ATPases
8	ATP1A1	Channels and Receptors	NA+/K+-ATPases
9	ATP1A3	Channels and Receptors	NA+/K+-ATPases
144	Calretinin	Signalling molecules and Enzymes	Other signalling molecules

Tabla VI.19: Proteínas pertenecientes al Cluster 11

Índice	Nombre	Familia	Subfamilia
28	Veli	Cell Adhesion and Cytoskeletal	Other Cell Adhesion Molecules
100	Chapsyn-110	MAGUKs / Adaptors / Scaffolders	PDZ-domain containing scaffolders
101	DLGH2	MAGUKs / Adaptors / Scaffolders	PDZ-domain containing scaffolders
102	DLGH3	MAGUKs / Adaptors / Scaffolders	PDZ-domain containing scaffolders
104	PSD-95	MAGUKs / Adaptors / Scaffolders	PDZ-domain containing scaffolders
106	Sap97	MAGUKs / Adaptors / Scaffolders	PDZ-domain containing scaffolders

G . Material Suplementario

Se anexa junto a este informe un DVD con el material suplementario producido en el desarrollo del proyecto. En él se incluyen las bases de datos utilizadas tanto en la aplicación del clasificador Naïve-Bayes como en el proceso de clustering de la red de PPIs inferidas del complejo NRC/MASC.

Además se incluyen: una hoja de cálculos llamada `fase_1.ods`, formato del OpenOffice Suite, con los resultados obtenidos en la primera etapa del proyecto, los que fueron incluidos en Apéndice C , y el archivo `campos_lr_final.tab.tar.gz`, una tabla con las interacciones obtenidas en la primera parte de este estudio.

Las bases de datos fueron organizadas de la siguiente forma:

Inferencia: En esta carpeta van las bases de datos referentes a la primera parte del proyecto. En cada uno de los subdirectorios se encuentran las bases de datos para cada sección de la etapa de inferencia.

- **Parte 1:** Se refiere a la inferencia por interacciones ortológicas. Tres bases de datos correspondientes a cada especie y una final donde se agregan todos los resultados -`p1_final.sql.gz`- en formato comprimido. Además de incluir las tablas finales de cada sección en un archivo que lleva el nombre de la base de datos con extensión `.tab`
- **Parte 2:** Son las bases de datos utilizadas en la sección de matrices de coexpresión. Se adjuntan 4 bases de datos, cada una para un *microarray* en particular y una base de datos donde ellas se unieron, `p2.dump.out.gz`. Además se incluye la tabla final de la sección, `p2_final.sql`, para su rápida implementación.
- **Parte 3:** Los resultados finales de la sección de funciones biológicas compartidas se presentan en el archivo `p3_final.sql` y las distintas tablas utilizadas en el archivo `p3.dump.out.gz`, ambas listas para ser implementadas en PostgreSQL.
- **Parte 4:** Al igual que para los resultados para la Parte 3, los resultados obtenidos a través del SSBP se anexan como un archivo con la tabla final en formato SQL, `p4_final.sql`, y un archivo con todas las tablas utilicidas `p4.dump.out.gz`.

Clustering: La base de datos obtenida por el clustering se encuentra en esta carpeta -archivo clustering.sql.gz- además se encuentra un archivo -`inferidos_uno.layout`- la subunidad principal del complejo NRC/MASC obtenida. Este archivo puede ser abierto simplemente como un archivo de texto o como diagrama por el programa BioLayout.