Análisis del transcriptoma temporal de *Drosophila melanogaster* en busca de genes sinápticos

Flavio Pazos
Tesis de Maestría en Bioinformática | PEDECIBA
Orientadores: Rafael Cantera y Gustavo Guerberoff
Departamento de Biología del Neurodesarrollo
Instituto de Investigaciones Biológicas Clemente Estable

A Lucía

INDICE

Prólogo

Hasta donde llega nuestro actual conocimiento, debemos pensar que el extremo de las ramitas de la arborescencia [axonal] no se continúa, sino que apenas contacta con la sustancia de la dendrita o el cuerpo celular en que incide. Tal conexión especial de una célula nerviosa con otra podría ser llamada "sinapsis".

Sir Charles Sherrington introdujo el término sinapsis hace casi 120 años, al escribir el capítulo dedicado al Sistema Nervioso para el Manual de Fisiología de Michael Foster[1]. El término, que deriva de la palabra griega para "*unión*"[2], fue introducido como concepto fisiológico, pero no fue sino hasta varias décadas después que tuvieron lugar las investigaciones que condujeron al descubrimiento de la sinapsis como entidad estructural y sitio de contacto entre las neuronas [3]. Sherrington recibiría años más tarde el premio Nobel por su trabajo acerca del funcionamiento de las neuronas.

Unos treinta años después, analizando tejido cerebral mediante técnicas de centrifugación recientemente desarrolladas, Whittaker y De Robertis, en paralelo, consiguieron aislar sinapsis del resto de los componentes celulares, abriendo así la puerta al análisis bioquímico de sus componentes[4]. Hoy sabemos que la sinapsis es la estructura a través de la cual las neuronas contactan entre si para transmitir información, y que al hacerlo, la sinapsis la filtra, la integra y la modifica, actuando como reguladora fundamental del flujo de información en los circuitos neuronales [5]. La sinapsis tiene por lo tanto una importancia central para comprender el funcionamiento del sistema nervioso, tanto en condiciones normales como patológicas.

De modo muy esquemático, podríamos decir que una sinapsis química típica[1] es una unión celular especializada, que posee la maquinaria molecular necesaria para la liberación de neurotransmisores del lado pre-sináptico, en oposición a los receptores de esos neurotransmisores presentes en el lado post-sináptico. Los elementos pre y post sinápticos se encuentran separados por la hendidura sináptica, un espacio en el cual los neurotransmisores son liberados por la terminal pre-sináptica. [5]

[1] Existen dos tipos de sinapsis; la neuronal y la inmunológica. A su vez, la sinapsis neuronal puede ser química o eléctrica. En esta tesis nos referiremos exclusivamente a la sinapsis neuronal química.

Las estructura presentes del lado pre-sináptico incluyen a las vesículas sinápticas, que almacenan neurotransmisores en su interior y que son mantenidas en la cercanía de la hendidura sináptica. Cuando un potencial de acción alcanza la zona de la sinapsis, se activan los canales iónicos que permiten el flujo de iones a través de la membrana y provocan su despolarización, lo que a su vez induce a las vesículas sinápticas a fundir sus membranas con la membrana de la célula, volcando así su contenido de neurotransmisores en la hendidura sináptica por exocitosis.

Una vez liberadas, las moléculas de neurotransmisor son captadas por receptores específicos acoplados a la membrana post-sináptica, que activados tras esa unión con los neurotransmisores, pueden despolarizar o hiperpolarizar la membrana celular post-sináptica. De darse esa despolarización, la misma comienza entonces a transmitirse por la membrana celular hasta llegar, eventualmente, a una nueva sinapsis. Por otro lado, una vez liberadas de los receptores, las moléculas de neurotransmisor son re-captadas por la membrana pre-sináptica mediante un mecanismo molecular que permite su reciclaje. Mediante ese mecanismo, los neurotransmisores se reincorporan a las vesículas sinápticas, que comienzan a acumularse para próximos ciclos de transmisión.

Es este un esquema muy general de la sinapsis química, del que se encuentran muchísimas variantes a distintos niveles. Todas las estructuras nombradas (canales, vesículas, receptores, etc.) están conformadas por una variedad de proteínas que requieren, para su correcto ensamblaje y funcionamiento, de la expresión de un conjunto de genes que no se conoce en su totalidad. No hay buenos catálogos que listen estos genes cuya expresión es necesaria para la sinaptogénesis. A grandes rasgos, se han ensayado dos estrategias para abordar este problema:

-elaborar catálogos de genes a partir de datos experimentales, por ejemplo, buscando mutaciones con pérdida de función que produzcan fenotipos anormales del sistema nervioso[6–8]

-elaborar catálogos de genes que compartan patrones temporales de expresión, en los que luego se determinan eventuales enriquecimientos estadísticos en grupos funcionales. Sin embargo, hasta ahora, los grupos funcionales enriquecidos son muy generales y carecen de poder predictivo [9–11]

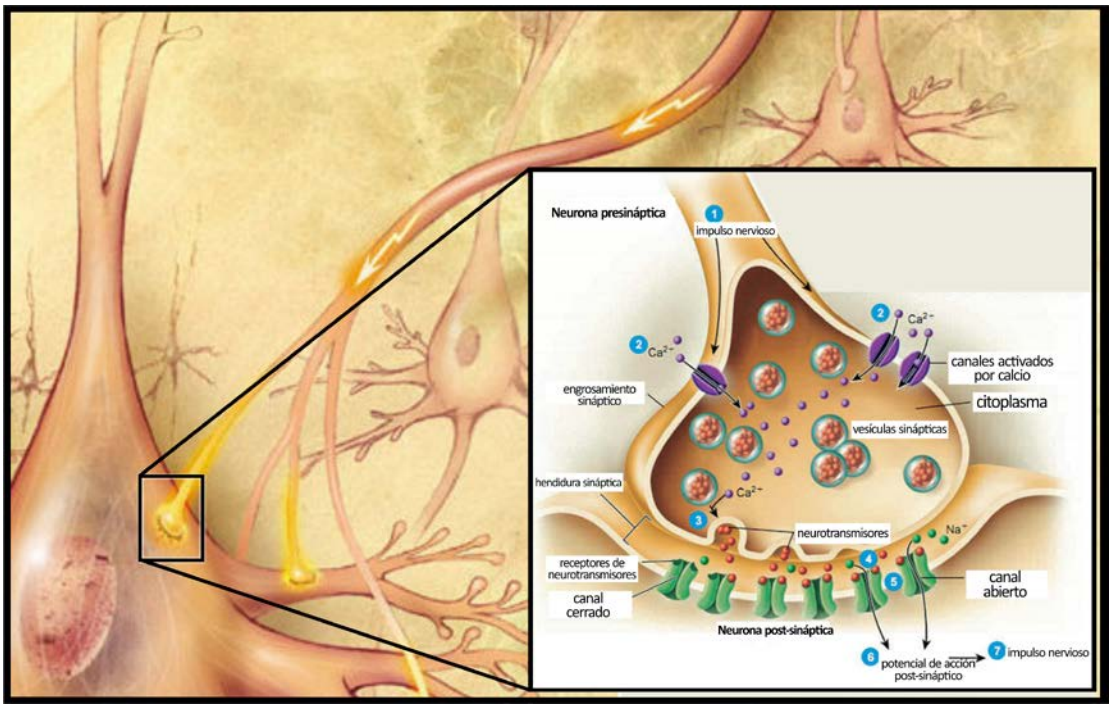


Figura 1 – Organización general de una sinapsis química tipo. *La figura del fondo representa una vista general de varias sinapsis (resaltadas en amarillo) sobre el cuerpo y las ramificaciones de una neurona. Una sola neurona puede llegar a recibir más de 10.000 sinapsis. En el cuadro que representa una ampliación de la figura del fondo, pueden apreciarse algunos detalles de la subestructura de la sinapsis. Todas las estructuras representadas, así como otras que no se detallan en la figura, requieren para su correcta constitución y funcionamiento, de la expresión de un conjunto de genes que no se conoce en su totalidad. En círculos azules se enumeran los principales sucesos necesarios para la transmisión del potencial de acción desde una neurona a la otra; (1) Arribo del impulso nervioso al engrosamiento sináptico, (2) apertura de los canales de calcio, (3a) unión del Ca++ a las vesículas sinápticas, (3b) liberación de los neurotransmisores hacia la hendidura sináptica, (4) unión de los neurotransmisores a los receptores, (5) entrada de Na+ al canal abierto, (6, 7) generación del potencial de acción en la neurona post-sináptica*

Nuestro conocimiento sobre la sinapsis, así como sobre su funcionamiento, su plasticidad y su mantenimiento mejoraría mucho si supiésemos cuales son todos estos genes. El objetivo principal de esta tesis es obtener una lista tentativa de estos genes, a los cuales llamaremos "genes sinápticos", analizando perfiles temporales de transcripción génica mediante una combinación de métodos de aprendizaje automático supervisado.

INTRODUCCIÓN

La genómica funcional

La estrategia tradicional para asignarle función a un gen implica análisis extensivos basados en bioquímica, genética, electrofisiología y otros estudios. Este tipo de estrategias experimentales, si bien pueden llegar a resultados contundentes, suelen demandar mucho tiempo y recursos. Por otro lado, en las últimas dos décadas, con el auge de los experimentos de microarreglos de ADN *(DNA microarrays)* y contando con cada vez más genomas secuenciados, y más datos de expresión a escala genómica, comenzaron a ser posibles abordajes alternativos [12, 13].

Los primeros trabajos en los que se utilizaron éste tipo de abordajes analizaban, mediante diversas técnicas estadísticas, datos de expresión obtenidos mediante miroarreglos en varios organismos. Esos trabajos pusieron de manifiesto que genes con funciones similares frecuentemente muestran patrones de expresión parecidos, lo que a su vez sugiere una relación funcional entre aquellos genes que fluctúan en paralelo [9, 10, 14–19]. Esas observaciones abonaron el uso de técnicas de aprendizaje automático para asignarle tentativamente nuevas funciones a genes basándose en sus perfiles de expresión, tanto en microorganismos como en insectos, plantas y mamíferos [20–24].

A medida que fue creciendo el volumen y la calidad de la información disponible respecto a la expresión de genes en distintas condiciones

experimentales, los métodos de aprendizaje automático comenzaron a ser cada vez más utilizados para predecir posibles funciones de la proteína codificada por un gen.

Las nuevas tecnologías de secuenciación

En los últimos años ha habido grandes avances en los métodos de secuenciación de ADN y ARN. Las nuevas tecnologías de secuenciación masiva (NGST) han resuelto algunas limitaciones importantes que tienen los microarreglos. como la baja precisión al medir transcriptos poco abundantes o la necesidad de contar con sondas específicamente diseñadas para detectar cada transcripto [25] . Más aun, se ha demostrado que en comparación a los microarreglos los datos de secuenciación por RNA-seq son altamente replicables y con poca variación técnica, y que en términos de performance general, RNA-seq debería ser la técnica de elección cuando se requiere una estimación precisa de los niveles absolutos de transcriptos presentes[25, 26].

Por otro lado, el descenso de los costos de secuenciación permitió un aumento exponencial en la disponibilidad de series temporales de datos de expresión génica (en adelante STDEGs) [27]. Las STDEGs se pueden usar para caracterizar funciones de genes específicos, relaciones entre genes, su regulación, coordinación e incluso implicaciones clínicas de su expresión diferencial[28]. Se han desarrollado métodos computacionales para representar, agrupar y clasificar este tipo de datos, e integrarlos con otros datos "ómicos"[28–30]. La calidad de los datos usados al aplicar estos métodos es uno de los aspectos críticos para su capacidad predictiva. Sin embargo, y hasta donde sabemos, aun no se han publicado estudios que predigan función génica basándose en perfiles de expresión obtenidos a través de estas nuevas tecnologías de secuenciación.

Varios estudios en base a STDEGs han mostrado la existencia de expresión coordinada de "clusters" de cientos de genes a lo largo del desarrollo, así como la correlación temporal entre los perfiles de expresión de algunos grupos (*clusters*) con hitos importantes del desarrollo morfológico y funcional [9–11, 31]. Hasta

ahora, esta correlación se limita a aspectos muy generales, como la gastrulación o la formación de la musculatura [10, 11]. Sin embargo, la resolución temporal de los datos disponibles y el conocimiento detallado de los procesos que llevan a la diferenciación del sistema nervioso en *Drosophila* [32] sugieren que se podrían analizar aspectos mucho más específicos, como el ensamblaje de la sinapsis neuronal.

Drosophila como organismo modelo en neurobiología

La mosca de la fruta *Drosophila melanogaster* es uno de los animales más utilizados como organismo modelo por la investigación biológica moderna. Fue trabajando con *Drosophila*, por ejemplo, que Thomas Morgan descubrió hace más de 100 años, el primer gen [33], siendo así la primer persona en vincular la herencia de un carácter fenotípico particular con un cromosoma específico. Desde entonces, Drosophila se convirtió en un organismo modelo excepcionalmente útil para abordar los más diversos problemas de la biología moderna.

Una de las ventajas de *Drosophila* es que su genoma es mucho menos complejo que el de mamíferos, tiene un tamaño de casi la vigésima parte del genoma humano y está organizado en solo 4 cromosomas [34], no obstante lo cual, contiene una cantidad de genes que es aproximadamente la mitad de genes que tiene *Homo sapiens.* Más aun, casi el 75% de los genes humanos relacionados con alguna enfermedad tienen genes homólogos en *Drosophila* [35]. En parte por esa misma razón, las herramientas de experimentación genética que se han desarrollado para trabajar con la mosca de la fruta no tienen parangón en otros organismos modelo[34]. Como resultado de esto último, el genoma de *Drosophila* es uno de los mejor anotados[36], lo cual es una indudable ventaja a la hora de entrenar un algoritmo de aprendizaje supervisado (ver más adelante).

Además de haber facilitado la identificación de varios mecanismos celulares fundamentales y de haber sentado las bases de la Genética hay varios motivos que hacen de *Drosophila* un sistema particularmente atractivo para la neurobiología [37]. En primer lugar, hay significativas similitudes entre nuestro cerebro y el de la

mosca [38]. Además, hay varias perturbaciones de la conducta (consecuencia de funciones cerebrales alteradas) que están bien caracterizadas y son fácilmente identificables, cuantificables y trazables genéticamente. Por esas razones, entre otras, *Drosophila* es utilizada como organismo modelo en el estudio de numerosas enfermedades neurodegenerativas humanas, como el Parkinson, el Alzheimer, la enfermedad de Huntington [39] y la esclerosis lateral amiotrófica [40] .

Finalmente, y de particular relevancia para esta tesis, *Drosophila* ha sido fundamental para la comprensión de cómo los genes controlan el desarrollo del sistema nervioso y el funcionamiento de la sinapsis [5, 41, 42]. El alto grado de conservación entre los genes que codifican proteínas necesarias para la exocitosis, la endocitosis y otras funciones básicas de la sinapsis tanto en mosca de la fruta como en mamíferos, sugiere que la organización básica de la sinapsis se completó tempranamente en la evolución de los metazoos [43]. Este hecho indicaría que se ha tolerado poca diversificación proteica durante la evolución de la transmisión sináptica, por lo que el estudio de la sinaptogénesis en *Drosophila* puede aportar conocimientos neurobiológicos fundamentales que nos ayuden a comprender nuestro propio sistema nervioso[44]. Por otro lado, la mayor parte de la complejidad transcripcional de *Drosophila* se da en el tejido nervioso, y un pequeño grupo de genes, en su mayoría expresados específicamente en las neuronas, tiene el potencial de codificar miles de transcriptos diferentes cada uno, mediante el uso de promotores y empalmes de ARN alternativos. [45]

Teniendo en cuenta, además de todo lo ya señalado, su bajo costo de mantenimiento y la existencia de una considerable y productiva comunidad científica dedicada a la neurobiología en *Drosophila*, se puede afirmar que la mosca de la fruta es no solamente una buena elección para el desarrollo de esta tesis, sino que el catálogo resultante de la misma podrá ser utilizado por esa comunidad y facilitará el avance de nuestro conocimiento en este campo.

Desarrollo de *Drosophila* y sinaptogénesis

Se denomina sinaptogénesis a una fase intensa y relativamente acotada en el tiempo de formación masiva de sinapsis. ¿Se puede esperar alguna característica particular en los perfiles temporales de transcripción de los genes necesarios para la sinaptogénesis? La sinaptogénesis forma parte de un proceso más general: el desarrollo del sistema nervioso. Para poder formar sinapsis funcionales, el organismo debe haber atravesado previamente una serie de etapas y formado una serie de estructuras anatómicas. En particular, la sinaptogénesis solo es posible durante una etapa avanzada de la diferenciación neuronal, el proceso de especificación por el cual la progenie resultante de la proliferación de células madres neurales se convierte en células neuronales [46], lo cual incluye el crecimiento axonal [47, 48] y el crecimiento y la ramificación de las dendritas [49, 50]. Por lo tanto, para responder a nuestra pregunta, debemos considerar algunos aspectos más generales del desarrollo de *Drosophila*, y en particular, del desarrollo de su sistema nervioso.

Antes de eclosionar como organismo adulto, *Drosophila* atraviesa dos etapas de desarrollo que pueden abarcar poco más de una semana. La primera de esas etapas es el desarrollo embrionario, que en condiciones estandard de cultivo dura casi exactamente 24 hs. La segunda es la etapa de desarrollo postembrionario, que comprende a su vez dos estadíos. El primero de ellos es el estadío larval, de una duración de entre 3 y 4 días. Le sigue el estadío pupal, que insume otros 3 a 4 días. Más tarde, y si todo transcurrió con normalidad, emerge la mosca, un organismo adulto que puede vivir más de 90 días. *Drosophila* pasa entonces por cuatro formas de organización corporal a lo largo de su ciclo vital; embrionaria, larval, pupal y adulta [32]. Sin embargo, solo en dos de esas formas, la larval y la adulta, el organismo se mueve, se alimenta e interactua con el entorno, actividades para las que necesita de un sistema nervioso completamente funcional.[32]

En cuanto al desarrollo del sistema nervioso, sabemos que durante el breve período de 24 hs del desarrollo embrionario, el organismo es capaz de, comenzando como una sola célula, desarrollar el sistema nervioso con el que la larva comenzará su vida activa y controlará su musculatura para moverse, buscar

comida y comer, una actividad a la que está casi exclusivamente dedicada. Por lo tanto, los genes necesarios para la formación del sistema nervioso con el que nace la larva deben ser expresados en la etapa embrionaria. Sin embargo, esa expresión solo es posible luego de que se hayan completado algunas etapas previas que son imprescindibles, como la definición de los ejes corporales y las cavidades internas, pues de no haberse llevado a cabo con normalidad, no solo harían imposible construir un sistema nervioso sino que directamente haría al organismo inviable.

Un poco más adelante, hacia el final de la etapa larval, los genes sinápticos deberían bajar sus niveles de expresión, puesto que la actividad sináptica y el número de sinapsis se reducirán mucho cuando el sistema nervioso larval comience, poco después, la metamorfosis que lo transformará en el sistema nervioso del adulto. En efecto, cuando el organismo entra a la etapa pupal, atraviesa una reorganización radical de sus tejidos, durante la cual buena parte de los mismos son desmontados. Esa reorganización incluye al SN, que debe pasar de controlar los tejidos de una larva, que se arrastraba en busca de comida, a controlar los de un organismo adulto que camina, corre, salta, vuela y posee una conducta mucho más compleja. La reorganización del SN se inicia a poco de comenzar la etapa pupal, un período durante el cual se espera observar otro aumento en la expresión de los genes necesarios para la sinaptogénesis.

En resumen, y sin olvidar que se trata de una generalización, se puede esperar que los niveles de expresión de los genes necesarios para la sinaptogénesis suban al final de la etapa embrionaria, bajen al final de la etapa larval y vuelvan a subir al inicio de la etapa pupal, en coincidiencia con las dos olas de sinaptogénesis masivas recién nombradas.

Aprendizaje automático

El aprendizaje automático es una rama de la inteligencia artificial que tiene como objetivo desarrollar técnicas que permitan a un computador aprender. Aquí se entiende que aprender es la capacidad de decidir cuándo algo pertenece a una categoría de cosas y cuando no. Se trata de crear programas, o algoritmos, que

sean capaces de generalizar un conocimiento, a partir de información suministrada en forma de ejemplos. Los métodos de aprendizaje automático pueden dividirse en métodos supervisados y no-supervisados [51]. Los métodos no supervisados aprenden sin nada que les enseñe y dependen únicamente de la medida de similitud que se haya definido para formar grupos (*clusterizar)*, con los ejemplos que le fueron proporcionados. Las características según las cuales los ejemplos son agrupados son llamadas "variables predictivas". Una vez definida la distancia, esto es, una vez definido qué quiere decir que dos ejemplos se parezcan entre sí, los distintos algoritmos buscan formar grupos en los que los ejemplos de un grupo se parezca más entre sí que a los ejemplos que pertenecen a otros grupos. Las variables predictivas según las cuales se pueden agrupar genes pueden ser sus perfiles de expresión, sus secuencias de ADN, sus interacciones genéticas, etc. El conocimiento biológico solo se usa después de que los algoritmos han formado grupos de genes, para caracterizarlos y establecer hipótesis acerca de las posibles funciones de genes desconocidos[16].

Es muy difícil establecer un criterio objetivo según el cual evaluar la fuerza predictiva de los grupos de genes que forma un método de aprendizaje no supervisado, por lo cual las posibilidades para estimar su calidad o para comparar los resultados de distintos agrupamientos son muy limitadas [20]. De todos modos, debido a que inicialmente se conocía muy poco sobre el repertorio completo de patrones de expresión, para los primeros experimentos se prefirieron métodos no supervisados [15].

Por otro lado, un método de aprendizaje supervisado intenta aprender a reconocer automáticamente una categoría dada de ejemplos a partir de una muestra de entrenamiento, en la cual se indica cuales de los ejemplos pertenecen a esa categoría y cuales no. Al algoritmo entrenado y ajustado se le llama modelo predictivo, o clasificador, y es luego utilizado para establecer la probabilidad de que un nuevo ejemplo pertenezca a la categoría en cuestión. Una característica importante de los métodos supervisados es que las clasificaciones que hace el clasificador, también llamadas predicciones, pueden ser objetivamente evaluadas poniéndolas a prueba con un pequeño grupo de nuevos ejemplos de categoría conocida [20].

Una posible aplicación de estos métodos es entrenarlos para que aprendan a reconocer qué genes tienen una función biológica dada y qué genes no. Se ha generado una profusa bibliografía que recoge abordajes de este tipo, informando sobre estudios en los cuales se entrenaron algoritmos de aprendizaje automático para asignar tentativamente funciones biológicas a genes o para definir grupos funcionales de genes. [12, 13, 15–17, 20, 22, 24, 36].

Construcción de la muestra de entrenamiento

La calidad predictiva de un clasificador depende fuertemente del modo en que se haya construido la muestra de entrenamiento. Cuando el objetivo es entrenar a un clasificador para aprender a reconocer qué genes cumplen determinada función biológica, puede ser relativamente sencillo encontrar ejemplos positivos, esto es, genes para los cuales esa función está bien establecida, pero es más difícil encontrar ejemplos negativos, o sea, genes de los que se pueda asegurar que no son importantes para esa función. Tal vez por esa razón, una estrategia relativamente común es considerar a todos los ejemplos no-positivos como ejemplos negativos. Sin embargo, esta estrategia puede dar lugar a un gran desbalance entre la cantidad de casos positivos y negativos, lo cual genera serios problemas de ajuste del modelo [52]. Elegir aleatoriamente un subconjunto de muestras no-positivas podría remediar este problema[53]. Sin embargo, como un gen puede tener más de una función, hacer una selección aleatoria de todos los genes fuera de la clase funcional objetivo tampoco es lo más apropiado, pues el procedimiento no minimiza las chances de incluir en la muestra de entrenamiento falsos negativos[52].

La base de datos Gene Ontology (GO) es una valiosa fuente de información ordenada sobre la función de las proteínas, estructurada como un grafo acíclico direccionado y organizado en tres ontologías; “Función Molecular”, “Procesos Biológicos” y “Componentes Celulares” [54]. GO es ampliamente utilizada como fuente de ejemplos positivos y negativos para entrenar este tipo de clasificadores. En general, una vez que se ha elegido la función biológica a predecir, todos los

genes que se encuentran asociados a esa función en GO son automáticamente incluidos como ejemplos positivos en la muestra de entrenamiento, cuyos ejemplos negativos son luego seleccionados entre el resto de los genes. Esta aproximación tiene la ventaja de permitir construir clasificadores para muchas funciones biológicas de manera automática, pero tiene algunas desventajas importantes; no minimiza las chances de incluir falsos negativos en la muestra de entrenamiento e implica renunciar a la posibilidad de efectuar análisis de enriquecimiento funcional contra la base de datos GO para evaluar los resultados de las clasificaciones.

Tomando en cuenta estas limitaciones, decidimos abordar el problema de la construcción de la muestra de entrenamiento de un modo alternativo. Por un lado, seleccionamos los ejemplos positivos uno a uno, a partir de una exhaustiva revisión bibliográfica. Por otro lado, seleccionamos los ejemplos negativos a partir de criterios biológicos objetivamente cuantificables, buscando minimizar la probabilidad de incluir falsos negativos en la muestra de entrenamiento.

Número esperado de genes sinápticos.

Un aspecto importante al planificar un estudio de aprendizaje automático es hacer una estimación de la cantidad de genes que tienen la función que se intentará predecir. En esta tesis llamamos "genes sinápticos" a todos aquellos genes necesarios para el ensamblaje, el funcionamiento, la plasticidad y el mantenimiento de la sinapsis. Una parte de esos genes ya es conocida y otra parte resta por ser descubierta. ¿Cuál es el número esperado de genes sinápticos?

En las últimas décadas se han descubierto unos pocos cientos de genes sinápticos en *Drosophila*, mediante *screens* genéticos de mutantes (entre otros, [6, 7, 55–61]) y otros abordajes experimentales (ver por ejemplo[5]). Sin embargo, estas estrategias demandan mucho tiempo y recursos y existe un amplio consenso en cuanto a que solo se ha identificado una fracción de los genes sinápticos de *Drosophila* (ver por ejemplo [42, 62, 63]). Para comenzar a acotar esa cantidad, podemos proponer un límite inferior y un límite superior a esa cantidad.

La cantidad de proteínas (y genes) necesarias para la formación y el funcionamiento de un subdominio celular (un organelo) depende en parte de su tamaño y complejidad. Uno de los organelos celulares más pequeños y simples, la vesícula sináptica, es una esfera de unos 40 nanómetros de diámetro, formada por una sola membrana y con una composición molecular de no más de 50 proteínas distintas[64]. Las vesículas sinápticas, aun cuando son de una importancia fundamental para el funcionamiento de la sinapsis, son solo uno de los muchos componentes de la sinapsis, por lo cual la cantidad total de genes sinápticos debería exceder con creces los 50, tal vez en un orden de magnitud o más.

Al otro extremo del amplio espectro de tamaño y complejidad de los organelos celulares, las sinapsis y las mitocondrias son organelos mucho más grandes y complejos que las vesículas. Ambos están compuestos de dos membranas con diferente composición molecular y función, tienen más componentes, mucho mayor tamaño que las vesículas y formas muy elaboradas. Un estudio reciente concluye que aproximadamente unos 1100 genes nucleares son necesarios para la formación y el funcionamiento de las mitocondrias en el cerebro, la médula espinal, el hígado y otros tejidos (14 en total), de los cuales un tercio codifica proteínas necesarias para todas las mitocondrias, independientemente del tejido y el resto codifica proteínas distintas en las mitocondrias de cada uno de los tejidos estudiados [65]. También se ha sugerido esta combinación para la sinapsis y se espera que dentro del total de los genes sinápticos habrá un conjunto de genes "básicos", que se expresan en todas las sinapsis y un número de genes "específicos", necesarios exclusivamente para ciertos tipos de sinapsis (Por ejemplo, se sabe que las sinapsis que usan glutamato como transmisor contienen ciertas proteínas necesarias para la transmisión glutamatérgica, que no existen en sinapsis que usan otros neurotransmisores).

Algunos estudios proteómicos de sinapsis [66–68] han revelado cientos de proteínas y que su especificidad, esto es, su carácter excitatorio, inhibitorio o modulatorio, se establece a través del reclutamiento y ensamblaje de complejos proteicos particulares[69].

Actualmente Gene Ontology[54] provee la lista más confiable de genes con importancia para la sinapsis. Gene Ontology recoge de FlyBase las anotaciones para *Drosophila.* Existen además algunos catálogos que listan "genes sinápticos"

(Tabla 1) reuniendo anotaciones y predicciones de diversas fuentes. Mencionaremos estos tres; SynDB[70], SynaptomeDB[71] y SynSysNet[2].

Base de Datos	***D. melanogaster***	***H. sapiens***
SynDb	1.073	3.249
SynaptomeDB	-	1.886
SynSysNet		1.028
Gene Ontology	456	

Tabla 1. Número de genes sinápticos *listado en las tres bases de datos especializadas en genes sinápticos, y en Gene Ontology, a mayo de 2015.*

Cierta proporción de estos genes fue incorporada a estos catálogos en base a trabajos experimentales en distintos organismos, incluyendo microscopía electrónica y electrofisiología para evaluar la ultraestructura y la transmisión sináptica en mutantes, pero la mayoría de ellos han sido anotados en base a estudios proteómicos de componentes sinápticos (por ej; vesículas sinápticas, densidad presináptica [72, 73]), homología de secuencias entre especies, dominios proteicos conservados u otros abordajes bioinformáticos.

Por lo tanto, los datos disponibles sugieren que el número esperado de genes necesarios para el ensamblaje y funcionamiento de la sinapsis en *Drosophila melanogaster* probablemente se sitúa alrededor de los mil genes.

[2]http://bioinformatics.charite.de/synsysnet/index.php?site=faq

HIPÓTESIS:

Es posible determinar si un gen es relevante para la sinaptogénesis a partir de su perfil temporal de transcripción.

OBJETIVOS GENERALES:

1. Investigar la hipótesis principal analizando un transcriptoma temporal de *Drosophila melanogaster*.
2. Adquirir dominio en el uso de diversos métodos estadísticos y computacionales que permitan investigar la hipótesis principal.

OBJETIVOS ESPECÍFICOS:

1. Elaborar un catálogo de genes con muy alta probabilidad de ser "genes sinápticos", que aún no hayan sido asociados experimentalmente a esa función.
2. Delinear un abordaje general que permita elaborar catálogos de este tipo para otros procesos biológicos.
3. Adquirir formación en distintos métodos estadísticos y computacionales.
4. Adquirir formación acerca de la sinaptogénesis y el neurodesarrollo de *Drosophila.*

RESULTADOS ESPERADOS

1. Un catálogo de genes con muy alta probabilidad de tener relevancia para el ensamblaje y funcionamiento de la sinapsis neuronal.
2. Un esquema metodológico que permita elaborar catálogos de genes con muy alta probabilidad de ser relevantes para funciones biológicas con las que aun no han sido asociados, a partir de sus perfiles temporales de transcripción.
3. Formación en el uso de distintos métodos estadísticos y computacionales.
4. Formación en neurodesarrollo en *Drosophila*, en particular sinaptogénesis.

MÉTODOS

Datos utilizados

Utilizamos los datos del transcriptoma temporal de *Drosophila melanogaster* publicado por el Proyecto MODENCODE en 2012 [74]. En particular, tomamos los datos de la "Tabla suplementaria 9", que recojen los niveles de ARN-seq poliA (esto es, el ARN que será efectivamente traducido a proteínas) de los 15.398 genes del genoma. Los datos representan la cantidad de transcriptos en muestras biológicas tomadas en 30 momentos del ciclo vital del organismo, incluyendo todas sus fases de desarrollo. Cada muestra consiste del ARN total aislado de 30 animales completos. Los niveles de transcripción están dados en valores de FPKM (Fragments Per Kilobase of transcript per Million fragments mapped), un método de cuantificar el nivel de transcripción normalizado por el largo de cada gen y el número de "reads" detectados en cada muestra.

Las muestras de esta serie no están igualmente espaciadas en el tiempo. En primer lugar, la serie de muestras incluye los niveles de expresión correspondientes a 12 intervalos, de 2 horas cada uno, que abarcan todo el desarrollo embrionario. Luego comprende 6 momentos de la etapa larval, con 6 muestras separadas entre sí por intervalos de tiempo diferentes, de entre 3 y 24 horas cada uno, y por último 6 muestras tomadas en distintos momentos de la etapa pupal, separadas entre sí por intervalos de 12 (las tres primeras) o 24 (las tres siguientes) horas. El set de datos original también incluye tres muestras tomadas en la etapa adulta, separadas además por sexo. Esas muestras no fueron consideradas para la tarea de clasificación, aunque sí fueron utilizadas, como se explica más adelante, para seleccionar genes de la muestra de entrenamiento. En la Figura 2 se presenta un esquema del diseño de muestreo de los datos.

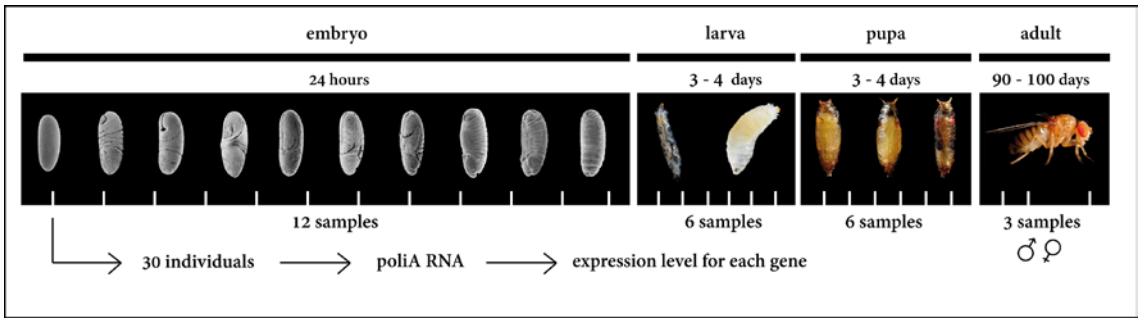


Figura 2. Diseño de muestreo del transcriptoma temporal de Drosophila utilizado en este estudio. *El esquema representa los momentos del ciclo vital de Drosophila en los que se cuantificaron los niveles de RNA-seq de todos los genes de su genoma. Las tres muestras correspondientes a la etapa adulta están diferencias por sexo.*

Pre-procesamiento de los datos originales

El transcriptoma publicado por MODENCODE consiste en los perfiles de transcripción de 15.398 genes. De ese conjunto original de datos, excluimos los 1.756 genes que tienen niveles de transcripción por encima de cero solamente en las muestras correspondientes al estadio adulto. La razón para esa exclusión es que nuestro objetivo era buscar genes que se expresen en el embrión y en la pupa, dos estadios del desarrollo en los que tienen lugar períodos de intensa sinaptogénesis [32]. De este modo, de nuestro conjunto original de datos, los genes que se expresan en algún momento del desarrollo son 13.642. Eliminamos asimismo los 489 genes que usamos para construir la muestra de entrenamiento (ver más adelante), por lo cual el conjunto de genes que intentaremos clasificar entre sinápticos y no-sinápticos es de 13.153 genes.

Decidimos normalizar cada perfil de transcripción dividiéndolo por el máximo valor que alcanza en la serie, obteniendo así para cada gen una serie de valores que oscilan entre 0 y 1. Esta normalización se basa en la asunción de que la forma del perfil temporal de expresión de un gen es mucho más informativa acerca de la función de ese gen que sus niveles absolutos de expresión. Si bien la cantidad absoluta de transcriptos está correlacionada con la cantidad de copias de

ARN que serán traducidas a proteína, lo que más nos interesa es discernir cuando esa cantidad de proteína sube o baja.

Los perfiles de transcripción absolutos y normalizados de los 13.153 genes que fueron clasificados en "sinápticos" o "no sinápticos" se muestran en las figuras **3A** y **3B** respectivamente.

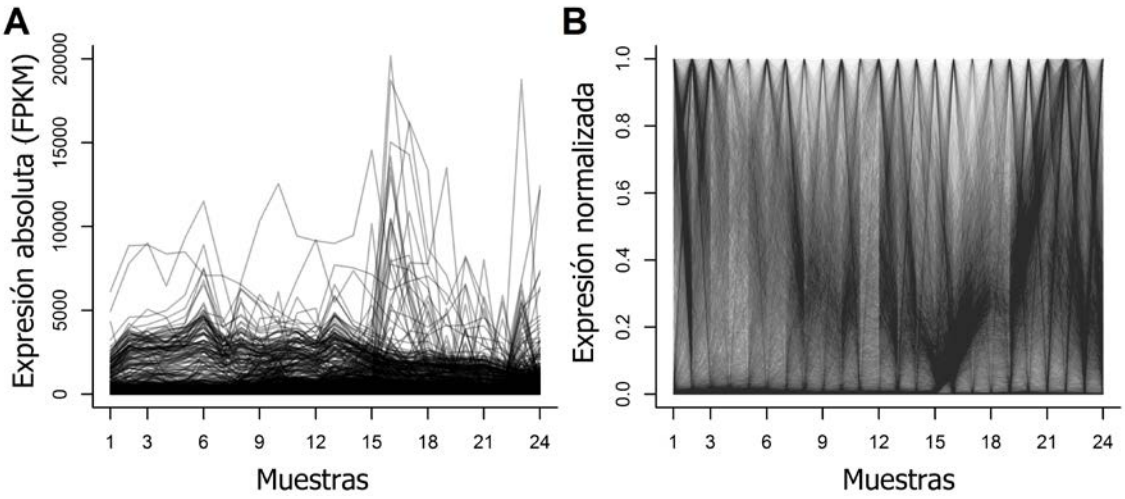


Figura 3 – Perfiles temporales de transcripción a ser clasificados. *Perfiles temporales de transcripción de los 13.642 genes de Drosophila que muestran algún valor de transcripción mayor a cero en al menos una de las muestras que corresponden a las etapas embrionaria, larval o pupal. Los gráficos están construidos a partir de los valores de transcripción absoluta (en A), expresados en FPKM, o con los valores normalizados entre cero y uno (en B). Datos originales publicados en Graveley et al., 2011 y adaptados como se explica en Métodos.*

Algoritmos de aprendizaje utilizados

A la hora de abordar un problema de clasificación, la elección del algoritmo de aprendizaje es en sí misma una decisión crítica, ya que no hay un algoritmo que sea superior a los demás en todas las situaciones e incluso aquellos que tienen en promedio la mejor *performance*, pueden desempeñarse muy mal en ciertos problemas o métricas [75].

Trabajar con varios modelos predictivos y combinar sus clasificaciones de distintas maneras es una metodología que, como se ha demostrado[13, 76], puede mejorar la *performance* de la clasificación. Existe gran diversidad de estas "técnicas de ensamble". En nuestro caso, y dado los buenos resultados en experimentos preliminares, decidimos aplicar un abordaje relativamente sencillo, consistente en ajustar tres algoritmos bien conocidos por su buen desempeño general y luego simplemente intersectar sus predicciones.

Los algoritmos elegidos fueron K - Vecinos más Cercanos (o k-NN, del inglés *k-Nearest Neighbours*), Bosques Aleatorios (o más comúnmente Random Forest) y Máquinas de Vectores Soporte (o SVM, del inglés Support Vector Machine), por ser tres de los métodos que tienen los mejores desempeños en promedio y porque han dado buenos resultados en una amplia gama de estudios de datos biológicos [75].

Utilizamos estos tres algoritmos para predecir si, dado su perfil temporal de transcripción $\{\mathbf{x}_i = x_i^j: j=1, 2,..., 24\}$, un gen **i** es sináptico ($y_i = 1$) o no lo es ($y_i=0$). Estos algoritmos nos permitieron estimar, a partir de la muestra de entrenamiento, la probabilidad de que dado su perfil de transcripción, determinado gen pertenezca a la categoría "sináptico", esto es, permiten calcular $\mathbf{P}(y_i = 1| \mathbf{x}_i)$

En general, la regla de decisión es que el gen **i** es clasificado como sináptico si la probabilidad estimada es mayor a cierto umbral, que por defecto se fija en 0,5. Luego, para evaluar la bondad del clasificador, se puede estimar la tasa de error mediante la técnica de validación cruzada. El entrenamiento de cada uno de los algoritmos implica el ajuste de dos parámetros, como se explica más adelante. Todos los cálculos fueron efectuados en R [77], haciendo uso de los paquetes *randomForest* [78], *e1071* [79] y *class*[80].

Random Forest. Se trata de un método de agregación basado en árboles de clasificación, que mejora los resultados de los mismos al introducir dos pasos de aleatorización: un cierto número **B** de muestras "*bootstrap"* tomadas de la muestra de entrenamiento y la elección aleatoria de las **m** variables predictivas. La estimación de $\mathbf{P}(y_i = 1| \mathbf{x}_i)$ se obtiene al agregar las predicciones de cada uno de los **B** árboles del "bosque" [81]. Construimos 500 árboles de clasificación y usamos 4 variables en cada partición de los árboles.

Máquinas de Vectores Soporte. Es un método que busca el hiper-plano que separe de forma óptima a los ejemplos de la muestra de entrenamiento según la clase a la que pertenecen (en nuestro caso, genes sinápticos o no sinápticos). Se busca maximizar el margen de separación entre ambas clases, ciñéndose a cierta función de costo **C** por violación del margen. Usando un *kernel* radial [51, 82] aumentamos las dimensiones del espacio en que habitan nuestros datos mejorando la separación entre ambas clases. El radio de la deformación que produce el kernel es controlado por el parámetro γ, el exponente del kernel radial. Ajustamos los parámetros C y γ mediante una búsqueda en grilla, obteniendo la menor tasa de error por validación cruzada décuple, 3.3%, con valores de C=2.2 y γ =0.02.

k-NN. En k-vecinos más cercanos, se asigna un objeto a una clase determinada según cual sea la clase más común entre sus *k* vecinos más cercanos. Los parámetros a ajustar son la medida de distancia a utilizar y el número **k** de vecinos a considerar. Utilizando distancia euclideana e incrementando uno a uno el número de vecinos considerados, observamos que luego de 7 vecinos las estimaciones de la tasa de error por validación cruzada alcanzan valores estables. Sin embargo, utilizamos k = 25, ya que ello nos permitió modificar con mayor libertad el umbral de clasificación (ver más adelante) Ese es el error que muestra la figura 5 para k-NN.

Construcción de la muestra de entrenamiento

Genes sinápticos. Qué se considera un "gen sináptico" depende de qué se considere una "sinapsis", un asunto que, sorprendentemente, no se encuentra aun saldado en la bibliografía científica (un problema semántico que se considera en [41]). En este trabajo, optamos por usar la palabra "sinapsis" para referir al contacto especializado entre dos células, también llamado a veces "zona activa", formado por una zona pre-sináptica asociada a una acumulación (o "cluster") de vesículas sinápticas y una zona post-sináptica, ambas bien delimitadas y observables por microscopía electrónica, separadas por una hendidura sináptica

en la cual la zona pre-sináptica libera neurotransmisores tras el arribo de un potencial de acción [41].

Tras una exhaustiva revisión bibliográfica, elaboramos una lista de 92 "genes sinápticos", cuyo rol en el ensamblaje y/o el funcionamiento de la sinapsis está apoyado por evidencia experimental fuerte e inequívoca. A modo de ejemplo, consideramos sinápticos aquellos genes cuyas mutaciones con falta de función impiden total o parcialmente la normal formación y función de las sinapsis, es decir, genes para los cuales un fenotipo mutante fue analizado a nivel sináptico con métodos de morfología, fisiología, microscopía electrónica, bioquímica, etc.

Genes no-sinápticos. Para elaborar la lista de "genes no-sinápticos", establecimos dos criterios biológicos y seleccionamos luego a todos aquellos genes que cumpliesen con al menos uno de esos criterios. Los criterios que establecimos son;

- mostrar, durante el tercer estadio larval, un nivel de expresión en el SNC muy bajo respecto al nivel de expresión promedio en todo el cuerpo.
- mostrar, durante la vida adulta, una expresión con un sesgo sexual extremo.

El primer criterio de selección se basa en que el tercer estadio larval es un período de rápido crecimiento y de sinaptogénesis intensa, por lo cual se puede asumir que aquellos genes que se estén expresando a niveles extremadamente bajos en el SNC cuando al mismo tiempo se están expresando a niveles considerables en el resto del cuerpo, probablemente sean de poca relevancia para el ensamblaje y el funcionamiento de la sinapsis. Por otro lado, podemos asumir que, a nuestros efectos, la sinapsis es esencialmente igual entre machos y hembras, por lo cual podemos esperar que aquellos genes que en la etapa adulta del organismo no se expresan en uno de los sexos pero sí lo hacen en el otro, también sean genes de poca relevancia para la sinapsis.

Para determinar qué genes cumplen con la primer condición utilizamos los datos de expresión tejido–específica disponibles en FlyAtlas [83]. Encontramos que de los 13.642 genes que se expresan en algún momento del desarrollo, los genes que muestran una relación menor a 0,05 entre su expresión en el SNC y su expresión en el cuerpo completo durante el tercer estadio larval son 352.

Para determinar qué genes cumplen con la segunda condición, analizamos las

muestras del transcriptoma temporal de MODENCODE que corresponden a la vida adulta y que fueron preparadas y secuenciadas por separado para machos y hembras[74]. Encontramos que hay 45 genes que, en las tres muestras que corresponden a la etapa adulta del transcriptoma, tienen transcripción nula en uno de los sexos y mayor a 25 FPKM en el otro.

De este modo, obtuvimos una lista de 397 genes no-sinápticos, que forman, junto a los 92 genes sinápticos seleccionados a partir de la bibliografía, nuestra muestra de entrenamiento de 489 genes. Es de suma importancia destacar que, ni los genes sinápticos ni los no-sinápticos, fueron incluidos en esas categorías teniendo en cuenta sus perfiles de expresión. Esto es importante porque es en función de esos perfiles de expresión que luego clasificamos el resto del genoma.

Otro aspecto importante respecto a la construcción de nuestra muestra de entrenamiento es que la inclusión de cada uno de los genes que la componen fue hecha de manera independiente a su anotación en Gene Ontology. En efecto, dejamos fuera de la misma a la mayoría de los genes que ya estaban asociados a la sinapsis en Gene Ontology. Incluimos solamente 83 de los 456 genes que, a julio de 2014, estaban anotados con algún término GO asociado a sinapsis. Por otro lado, incluimos 9 genes cuyo rol fundamental para la sinapsis está apoyado por fuerte evidencia experimental, pero que sin embargo aun no estaban anotados como sinápticos en Gene Ontology. Por otro lado, los genes no sinápticos fueron seleccionados a partir de dos criterios independientes, sin considerar sus eventuales anotaciones GO. Esta independencia de la muestra de entrenamiento respecto a Gene Ontology es muy importante, pues nos permitió luego evaluar el enriquecimiento funcional de los catálogos obtenidos por aprendizaje automático respecto a la base de datos de Gene Ontology sin problemas de circularidad.

Estimación de la verdadera tasa de error

Para evaluar el desempeño de cada clasificador, llevamos a cabo un procedimiento de validación cruzada décupla sobre la muestra de entrenamiento. El procedimiento de validación cruzada consiste en separar aleatoriamente la

muestra de entrenamiento en 10 subconjuntos del mismo tamaño. Luego, cada clasificador es entrenado con 9 de los subconjuntos y a continuación se clasifica el subconjunto restante. Como se conoce la clase a la cual pertenece cada uno de los ejemplos de la muestra de entrenamiento, la tasa de error del clasificador entrenado se puede determinar directamente, contando la cantidad de errores que cometió al clasificar el subconjunto de la muestra de entrenamiento que fue dejado fuera para entrenar al clasificador. El procedimiento se repite 10 veces, dejando fuera un subconjunto distinto cada vez. Luego, como estimación de la verdadera tasa de error del clasificador, se calcula el promedio de las 10 tasas de error obtenidas por validación cruzada.

Incremento secuencial del umbral de clasificación.

En lugar de fijar un umbral de clasificación que generase directamente un catálogo de genes del tamaño estimado previamente, decidimos incrementarlo secuencialmente partiendo de su valor por defecto. Esto es, fuimos aumentando gradualmente el mínimo de probabilidad de ser sináptico que un gen debe tener para ser clasificado como tal. Este umbral de clasificación tiene un valor por defecto de 0,5 y decidimos considerar los catálogos generados cuando ese umbral es aumentado sucesivamente a 0,6; 0,7; 0,8; y 0,9. Este procedimiento dio lugar a una serie de catálogos cada vez más pequeños de genes que son, de acuerdo a cada modelo ajustado, “genes sinápticos”. El procedimiento dio lugar también, a una serie decreciente de catálogos consenso. Llamamos catálogo consenso a la intersección entre las clasificaciones de los tres algoritmos, esto es, la lista de genes que fueron clasificados como sinápticos por los tres algoritmos a la vez para cada valor de umbral.

Caracterización biológica de los catálogos

Decidimos investigar si el aumento del umbral de clasificación, que da lugar a nuestra serie de catálogos, va acompañado de un aumento en la calidad biológica de los mismos. Para ello, evaluamos dos características en cada uno de los catálogos: el enriquecimiento funcional en genes anotados con términos GO asociados a sinapsis y el enriquecimiento en genes con expresión diferencialmente alta en el SNC relativa a otros tejidos.

Análisis de enriquecimiento funcional

Para evaluar la calidad de los catálogos producidos por nuestro abordaje con criterios biológicos, determinamos su eventual enriquecimiento funcional en términos GO. El análisis de enriquecimiento funcional es un método estándar para evaluar listas de genes. Dada una lista de genes del mismo tamaño que la lista en estudio, pero generada al azar, se puede determinar cual es la cantidad más probable de genes de esa lista que estarán anotados con cierto término GO. Luego se determina cuántos términos de la lista en estudio están efectivamente anotados con ese término GO. Si la lista en estudio contiene una cantidad de genes anotados con ese término que es significativamente superior a la que cabría esperar en una lista del mismo tamaño, pero generada al azar, se dice que la lista está enriquecida en ese término GO.

Más formalmente, el enriquecimiento **E** de un catálogo **C** en determinado término GO **x** se define [84] como;

E =(b/n) / (B/N),con;

N el número total de genes de la lista de referencia,

B el número total de genes de la lista de referencia anotados con **x**,

n el número de genes en **C** y

b el número de genes en **C** anotados con **x**.

En nuestro estudio, solo tuvimos en cuenta los valores de enriquecimiento con un p-valor asociado menor a 10^{-4} y un q-valor FDR menor a 10^{-3}. El q-valor FDR es la corrección del p-valor para testeo múltiple usando el método de Benjamini y Hochberg.

Existe una gran variedad de herramientas en línea para realizar análisis de enriquecimiento funcional, que introducen diversas modificaciones al esquema general arriba esbozado, para optimizar los resultados en distintas circunstancias o modelos biológicos [85]. En nuestro caso, utilizamos la plataforma GOrilla [84], una herramienta que tiene la particularidad de permitir la comparación de las anotaciones GO de un catálogo de genes respecto a las anotaciones GO de una lista de genes provista por quien hace el análisis, y no solamente respecto a la anotación del genoma completo. En nuestro caso, esa lista de referencia contra la cual comparamos las anotaciones GO de nuestros catálogos está compuesta por los 13.153 genes que quedaron luego de excluir de nuestro set inicial a los genes que pertenecen a la muestra de entrenamiento y a los que no mostraban transcripción durante las etapas del desarrollo.

Genes con expresión diferencial tejido-específica

El otro método que utilizamos para evaluar la relevancia biológica de nuestros catálogos consiste en evaluar la cantidad de genes con expresión diferencial tejido-específica (GEDTEs) que los componen. La hipótesis es que un catálogo de genes sinápticos estará enriquecido en genes que se expresan preferencialemente en el SNC. Para poner a prueba esta hipótesis, creamos listas de GEDTEs en diferentes tejidos, para lo cual usamos datos de expresión por tejido también publicados por modENCODE y disponibles en FlyBase [3] [86].

Asumiendo entonces que los genes sinápticos se expresan en el SNC (el tejido que tiene mayor cantidad de sinapsis por unidad de volumen) a mayores niveles que en otros tejidos con pocas o ninguna sinapsis, creamos listas de GEDTEs en cinco tejidos; el sistema nervioso central, las glándulas salivales, el cuerpo graso (tejido adiposo), el sistema digestivo y la carcaza. Las glandulas salivales y el cuerpo graso, en tanto no están innervadas, no poseen sinapsis, mientras que el intestino y la carcaza sí estan innervados, pero poseen muchísimas menos sinapsis que el SNC. Mediante la interfase disponible en FlyBase, seleccionamos los genes que durante el tercer estadío larval tienen una expresión al menos moderada en uno de estos tejidos y una expresión muy baja en los otros cuatro tejidos

[3] http://flybase.org/static_pages/rna-seq/rna-seq_search.html

considerados. A través de este procedimiento, obtuvimos cinco listas de genes que muestran una expresión diferencialmente alta en cada uno de estos cinco tejidos.

La tabla 2 muestra la cantidad y proporción de GEDTEs en cada tejido presente en el set inicial de genes a clasificar.

Tejido	**#GTEDEs**	**%**
Carcasa	155	23
Sistema Nervioso Central	165	24
Sistema Digestivo	257	37
Cuerpo graso	82	12
Glándula salival	27	4
Total	**686**	**100**

Tabla 2. Cantidad y porporción de GTEDEs en el set de genes a clasificar. *La tabla resume la cantidad de genes con expresión diferencial tejido específica, en cada uno de los cinco tejidos considerados, que se encuentra en el set inicial de genes a clasificar. La tercera columna muestra la proporción del total de GTEDES que corresponde a cada uno de los cinco tejidos.*

Finalmente evaluamos cómo cambian estas proporciones de GEDTEs en nuestros sucesivos catálogos, en el entendido de que un aumento en la proporción de GEDTEs en el SNC sería un buen indicador de la calidad de los catálogos desde el puto de vista de este criterio biológico.

RESULTADOS

Muestra de entrenamiento

Para entrenar a nuestros clasificadores, definimos las etiquetas "gen sináptico" y "gen no sináptico" (ver Métodos). La lista completa de los 92 genes sinápticos de la muestra de entrenamiento, con las correspondientes referencias bibliográficas que justifican su inclusión en esa lista, se presenta en el **Anexo 1.** Los perfiles de expresión normalizados de los 92 genes sinápticos se muestran en la **Figura 4A.** El **Anexo 2** contiene la lista de los 397 genes que cumplen con al menos uno de los criterios biológicos que establecimos para incluir a un gen en la muestra de entrenamiento como gen no sináptico (ver Métodos). La **Figura 4B** muestra los perfiles de expresión normalizados de los genes no sinápticos de la muestra de entrenamiento.

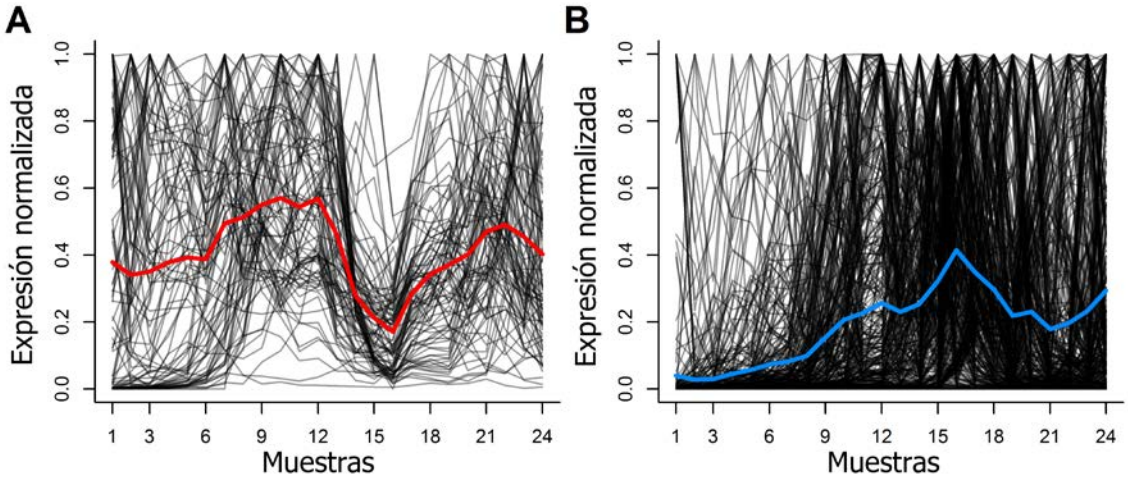


Figura 4. Perfiles de transcripción de los genes que forman parte de la muestra de entrenamiento. *(A) Perfiles de transcripción normalizados de los 92 "genes sinápticos". Es clara la correspondencia entre lo que cabría esperar del perfil de expresión promedio para los genes sinápticos de Drosophila y el promedio de los perfiles de transcripción realmente observados. En el ciclo de vida de Drosophila, un primer período de sinaptogénesis masiva tiene lugar durante la segunda mitad de la etapa embrional (muestras 7 a 12) y un segundo período de*

sinaptogénesis tiene lugar en la pupa (muestras 19 a 24), durante el cual se forma el cerebro adulto. Por otro lado, entre ambas olas de sinaptogénesis, tiene lugar un período de desensamblaje masivo de sinapsis. La línea roja, que representa el promedio de transcripción de los 92 genes sinápticos de la muestra de entrenamiento, se ajusta muy bien a éstos eventos. (B) Perfiles de transcripción de los 397 genes no-sinápticos de la muestra de entrenamiento. La línea azul corresponde al promedio de los 397 perfiles de transcripción. Ambos gráficos fueron construidos con valores obtenidos al normalizar entre 0 y 1 los datos valores originales en FPKM, publicados por Graveley et al., en 2011.

Ajuste de los clasificadores

La **Figura 5** muestra el error promedio y la dispersión para los tres clasificadores luego de la validación cruzada décupla sobre la muestra de entrenamiento. Los tres clasificadores alcanzaron tasas de error promedio inferiores al 5%.

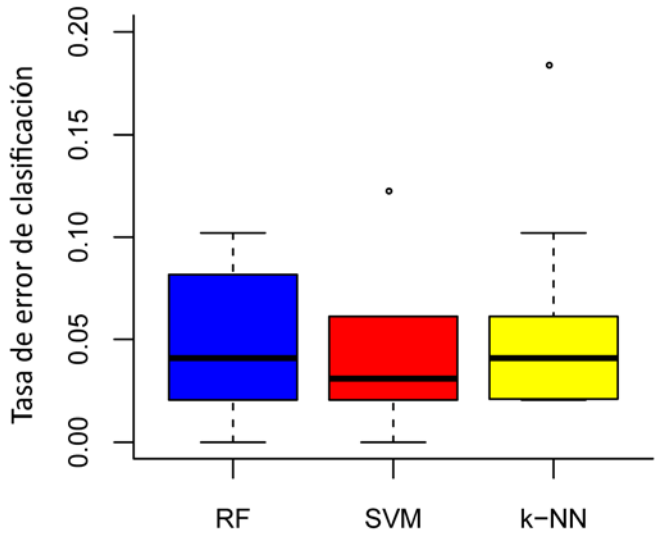


Figura 5 – Tasa de error de los clasificadores. *Diagrama de caja de las tasas de error de clasificación de cada uno de los tres clasificadores ajustados; Random Forests (RF), Support Vector Machine (SVM) y Vecinos*

más cercanos (k-NN), calculadas por validación cruzada décupla tal como se describió en Métodos. En cada caja, la línea negra horizontal representa la mediana y los puntos fuera de las cajas corresponden a valores atípicos.

Otro método estándar para estimar la bondad de un clasificador es el cálculo del área bajo su curva ROC (acrónimo de *Receiver Operating Characteristic*). Para construir la curva ROC, se calculan mediante validación cruzada décupla, las tasas de verdaderos positivos y de falsos positivos para cada valor de umbral de clasificación y luego se grafica la relación entre ambas tasas. El área bajo la curva de este gráfico se considera un estimador del desempeño del clasificador. Un área bajo la curva igual a 1 corresponde a una clasificación perfecta, y un área bajo la curva igual a 0,5 corresponde a una clasificación hecha al azar. El área bajo la curva ROC de los tres clasificadores se muestra en la Figura 6.

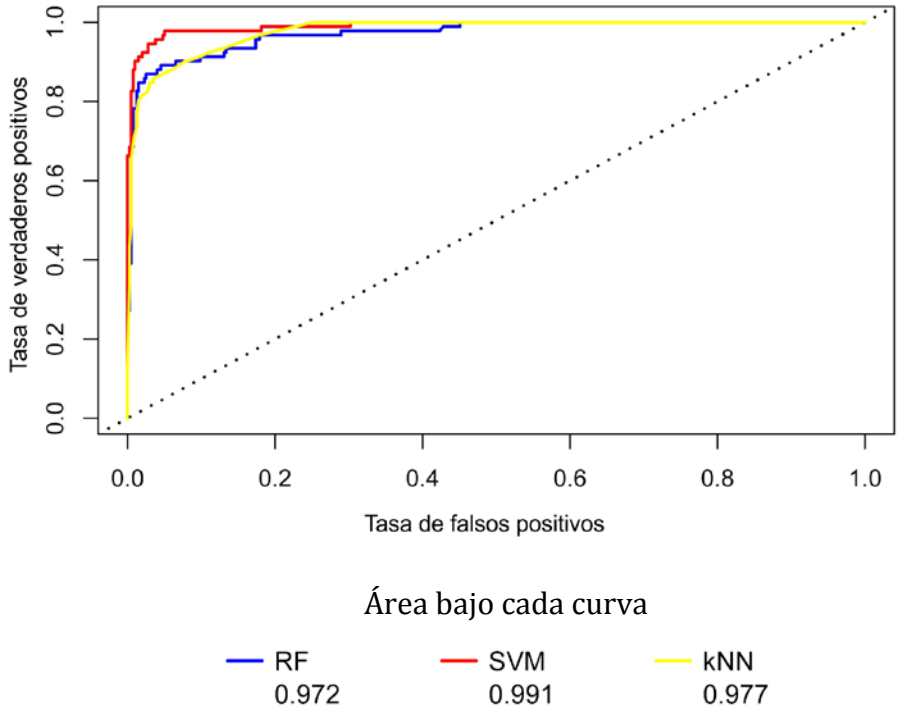


Figura 6. Curvas ROC de cada clasificador. *El gráfico muestra la curva ROC de cada clasificador. Los valores de las respectivas áreas bajo la curva se muestran debajo del gráfico. Los valores alcanzados son excelentes en los tres casos.*

Clasificación inicial de los tres clasificadores

Una vez ajustados los tres clasificadores, clasificamos con ellos nuestro set inicial de genes, obteniendo tres catálogos de genes tentativamente sinápticos. La cantidad de genes que recibieron la etiqueta de "gen sináptico" tras éstas clasificaciones iniciales se muestra en la **Figura 7**. Las tres clasificaciones muestran un alto grado de coincidencia entre sí. El modelo con la clasificación más divergente respecto a las otras es k-NN, y aun así, su clasificación muestra más de un 83% de coincidencia con las otras dos. Los genes clasificados como sinápticos por los tres modelos constituyen nuestro "catálogo consenso primario", que está formado por 4.872 genes. Este catálogo contiene muchos más genes de los que, según discutimos en la introducción, cabría esperar para un catálogo de genes sinápticos.

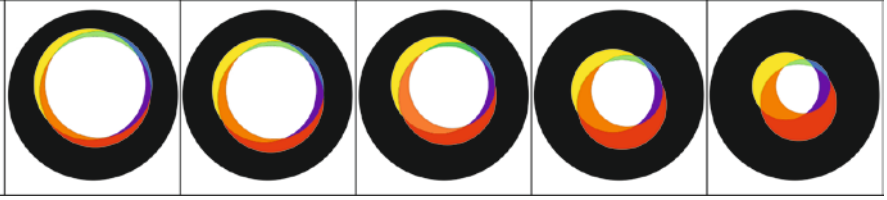
Umbral	0,5	0,6	0,7	0,8	0,9
kNN	5.944	5.239	4.673	3.434	2.114
RF	5.452	4.598	3.662	2.529	1.363
SVM	5.886	5.421	4.779	3.986	2.731
kNN∩RF∩SVM	4.872	4.047	3.203	2.095	988
kNN RF SVM kNN∩RF∩SVM					

Figura 7 - Genes clasificados como sinápticos por cada modelo según el umbral de clasificación. *Cada columna corresponde al umbral que la probabilidad de ser sináptico asignada a un gen debe sobrepasar para ser efectivamente clasificado como sináptico. Cada fila corresponde a uno de los tres clasificadores o alguna de sus intersecciones. La última fila muestra el número de genes clasificado como sináptico, para cada valor de umbral, por los tres clasificadores a la vez. Los círculos de los diagramas de Venn del panel inferior tienen un área proporcional al número de genes que cada uno de ellos representa. Los 13.153 genes a*

clasificar están representados por el área en negro, y el número de genes clasificados como sinápticos por cada uno de los modelos, o por sus posibles intersecciones, está representado según el código de colores que se muestra a su izquierda.

Incremento secuencial del umbral de clasificación

Con el fin de reducir el tamaño de nuestro catálogo de genes tentativamente sinápticos, y al mismo tiempo aumentar la probabilidad de que el catálogo final esté compuesto por genes sinápticos, aumentamos secuencialmente el umbral de clasificación de nuestros clasificadores. De ese modo, fuimos excluyendo secuencialmente a aquellos genes cuyas clasificaciones tenían mayor probabilidad de ser falsos positivos (ver Métodos). Mediante este procedimiento generamos una serie de catálogos de tamaño gradualmente decreciente, para los cuales una caracterización biológica hecha *a posteriori* podría además aportar apoyo adicional. Las columnas 2 a 5 de la Figura 7 muestran el número de genes clasificados como sinápticos por cada modelo al aumentar el umbral de clasificación, así como el grado de coincidencia entre las distintas clasificaciones.

Análisis de enriquecimiento funcional

Para caracterizar biológicamente la serie de catálogos obtenida, evaluamos su enriquecimiento funcional en términos GO relacionados a sinapsis. En primer lugar, encontramos que todos nuestros catálogos estaban enriquecidos en varios de esos términos. Más importante aun, encontramos que el enriquecimiento funcional de los catálogos producidos por cualquiera de los clasificadores, aumenta al aumentar el umbral de clasificación. Esta relación directa entre enriquecimiento funcional y umbral de clasificación se aprecia claramente en la Figura 8. En la figura también se aprecia como, dados un umbral de clasificación y un término GO, el enriquecimiento del catálogo consenso siempre es mayor al de los catálogos generados por cada clasificador por separado.

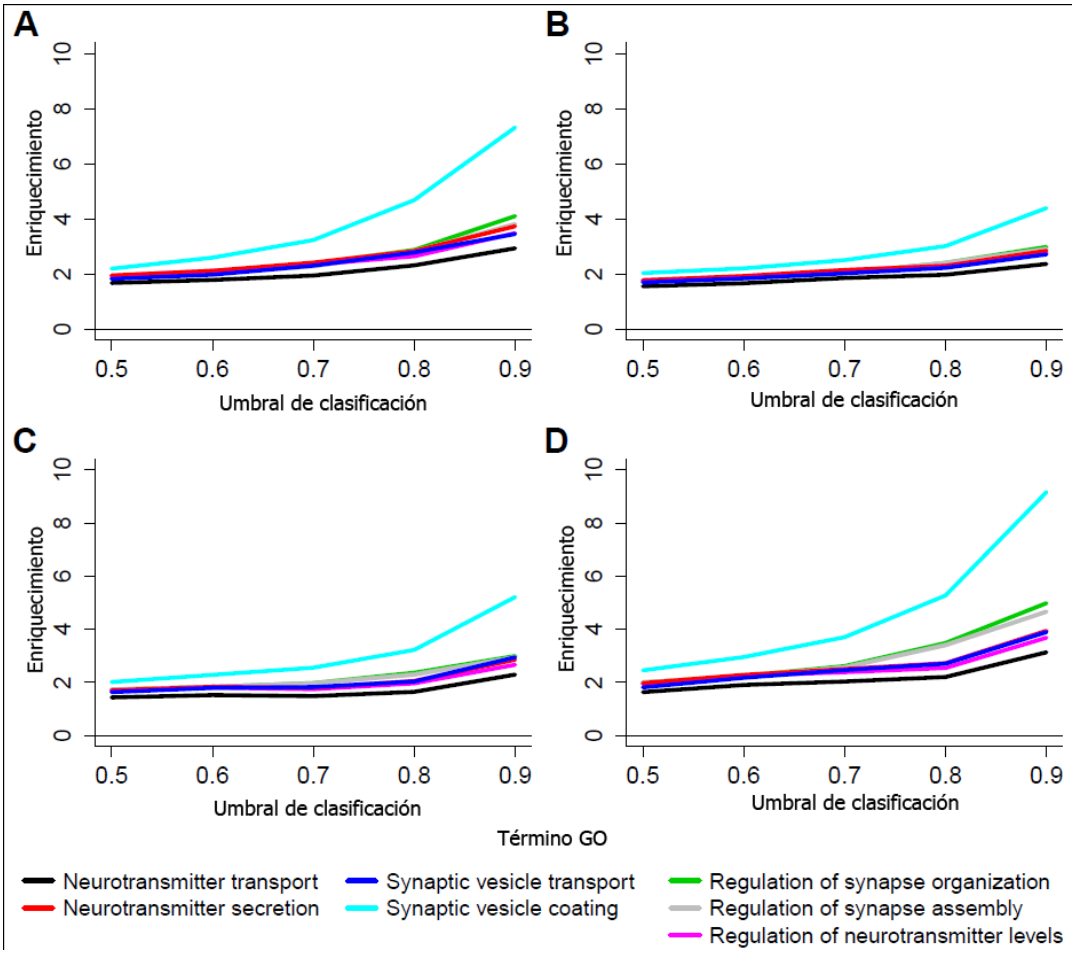


Figura 8 – Enriquecimiento funcional según el umbral de clasificación. *La figura muestra el enriquecimiento funcional en una selección representativa de términos GO asociados a sinapsis, de cada uno de los catálogos producidos al aumentar el umbral de clasificación de cada modelo, así como para la serie de catálogos consenso. En todos los casos, el aumento del umbral de clasificación resulta en un incremento del enriquecimiento funcional, aumento que se acentúa al considerar los catálogos consenso. Este demuestra que la calidad de la predicción mejora al considerar la intersección de las clasificaciones. A; catálogos generados por k-NN para cada umbral de clasificación, B; catálogos generados por Random Forest para cada umbral de clasificación, C; catálogos generados por SVM para cada umbral de clasificación, D; catálogos de genes clasificados como sinápticos por los tres clasificadores a la vez, para cada umbral de clasificación. Todos los valores de enriquecimiento tienen un p-valor asociado menor a 10^{-4} y un q-valor menor a 10^{-3}.*

Genes con expresión diferencial tejido-específica (GEDTEs)

Entre los 13.153 genes que fueron clasificados en este estudio, 686 mostraron una expresión diferencial tejido-específica (tal como fue definida en Métodos, la lista completa se muestra en el Anexo 3). Los porcentajes de GEDTEs en cada tejido que encontramos en el set inicial de genes a clasificar se muestran en la Tabla 2. Dado que la inmensa mayoría de las sinapsis son formadas por neuronas dentro del SNC, es de esperar que en un catálogo de genes sinápticos, los GEDTEs que corresponden al SNC estén sobre-representados. Por otro lado, podríamos esperar que nuestros catálogos tengan una sub-representación de los GEDTEs que corresponden a las glándulas salivares o al cuerpo graso, ya que se trata de tejidos en los cuales no se forman sinapsis.

La Figura 9 muestra el porcentaje de GEDTEs encontrados en cada tejido en cada uno de nuestros catálogos. En todos los catálogos hay una marcada sobre-representación de los GEDTEs del SNC, que pasan de ser un 24% de los GEDTEs presentes en el set inicial a clasificar, a ser más de un 80% de los GEDTEs presentes en el catálogo inicial. Esa cantidad desproporcionalmente alta de GEDTEs se incrementa al aumentar el umbral de clasificación. Al mismo tiempo, los catálogos muestran una sub-representación de los GEDTEs de tejidos sin sinapsis (glándulas salivares y cuerpo graso) o de tejidos con muy pocas sinapsis (sistema digestivo y carcaza), que también se acentúa al aumentar el umbral de clasificación.

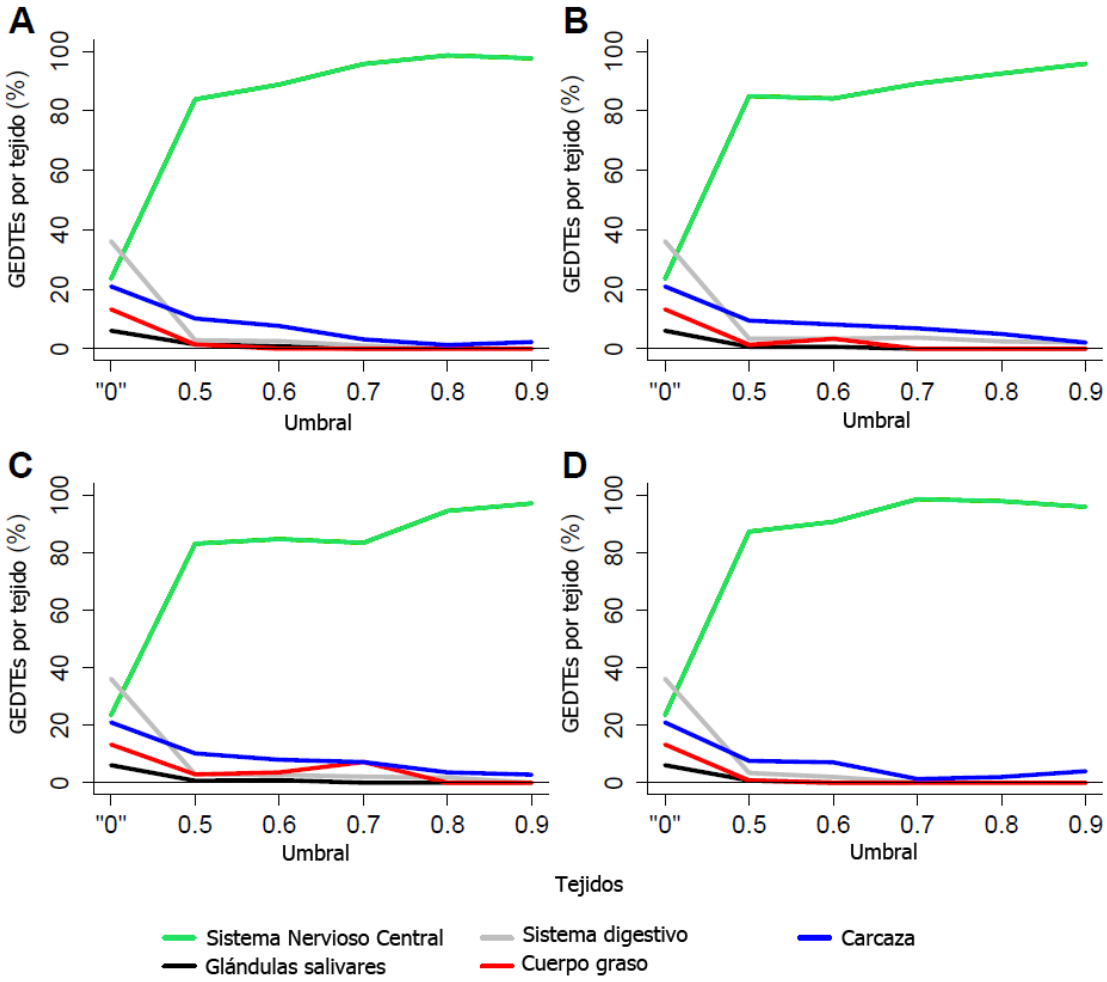


Figura 9 - Relación entre el porcentaje de GEDTEs y el umbral de clasificación. *Los paneles A-D muestran la relación entre expresión diferencial por tejido, método y umbrales de clasificación. Sea cual sea el método de clasificación, en todos los catálogos están sobre-representados los genes que se expresan mucho más en el SNC que en otros tejidos con menos sinapsis. Nótese que el aumento en el umbral de clasificación solo resulta en un aumento de la proporción de genes con expresión diferencial en el SNC y que la proporción de genes con expresión diferencial en otros tejidos disminuye en todos los casos. A: catálogos generados por k-NN, B; catálogos generados por Random Forest, C; catálogos generados por SVM, D; catálogos consenso. El eje horizontal muestra los umbrales de clasificación, con el "0" representando al set inicial de genes a clasificar. Cada línea de color corresponde a uno de los cinco tejidos analizados, según el código de colores en la parte inferior de la figura.*

Catalogo final de genes de *Drosophila* tentativamente sinápticos

El catálogo consenso que corresponde a un umbral de clasificación de 0,9 contiene 988 genes que, según los tres modelos ajustados, tienen una probabilidad mayor a 0,9 de estar involucrados en el ensamblaje o el funcionamiento de la sinapsis. De todos los catálogos obtenidos, éste es el que muestra el mayor enriquecimiento funcional en términos GO-BP relacionados a la sinapsis, así como la mayor proporción de GEDTEs en el SNC.

El enriquecimiento funcional de éste catálogo se explica por la presencia en el mismo de 95 genes que, a julio de 2014 (fecha en que se llevaron a cabo éstos análisis), ya habían sido anotados como relevantes para la sinapsis en Gene Ontology. Excluyendo esos 95 genes, obtuvimos nuestro catálogo final de 893 genes tentativamente sinápticos, disponible en el Anexo 4.

Caracterizaciones adicionales de nuestro catálogo final

Genes de nuestro catálogo final con homólogos humanos previamente descriptos como genes sinápticos

De los 893 genes de *Drosophila* que conforman nuestro catálogo final, hay 607 que tiene al menos un homólogo en *Homo sapiens* [87] de los cuales 66 genes (11%) tienen al menos un homólogo humano con al menos una anotación relacionada a la sinapsis [88]. La Tabla 3 muestra esos 66 genes de *Drosophila* de nuestro catálogo final, que aún no han sido definidos como sinápticos en *Drosophila*, pero que tienen homólogos en *H. sapiens* que sí han sido definidos como sinápticos.

Como modo de estimar la significancia estadística de este resultado decidimos construir una lista de 988 genes de *Drosophila,* tomados al azar de nuestro catálogo inicial de 13.153 genes a clasificar y luego determinar, tras haber excluido de la lista a todos aquellos genes con alguna anotación relacionada a la sinapsis, cuántos de esos genes poseen al menos un homólogo humano con al menos una anotación relacionada a sinapsis. Repetimos este procedimiento tres

veces, obteniendo en las tres ocasiones el mismo resultado; 25 genes. Nuestro catálogo final incluye entonces 2,6 veces más genes con homólogos humanos sinápticos que lo que cabría esperar por azar.

CG	Símbolo	Prob.	Homólogo	HGI	Término GO (en Homo sapiens)
CG7392	Cka	1.00	STRN	ENSG00000115808	CC: postsynaptic membrane
CG9634	NA	1.00	MME	ENSG00000196549	CC: synaptic vesicle
CG8529	Dyb	0.99	DTNA	ENSG00000134769	BP: synaptic transmission
CG7023	Usp12-46	0.99	USP46	ENSG00000109189	BP: regulation of synaptic transmission
CG16973	msn	0.99	MINK1	ENSG00000141503	BP: synaptic transmission
CG6593	Pp1alpha-96A	0.99	PPP1CA	ENSG00000172531	CC: dendritic spine
CG15112	ena	0.99	ENAH	ENSG00000154380	CC: synapse
CG10545	Gbeta13F	0.99	GNB2	ENSG00000172354	BP: synaptic transmission
CG14991	Fit1	0.99	FERMT1	ENSG00000101311	CC: synapse
CG1651	Ank	0.99	ANK2	ENSG00000145362	CC: postsynaptic membrane
CG32717	sdt	0.99	MPP4	ENSG00000082126	CC: presynaptic membrane
CG5248	loco	0.99	RGS14	ENSG00000169220	BP: long-term synaptic potentiation
CG7147	kuz	0.99	ADAM10	ENSG00000137845	CC: postsynaptic density
CG10566	NA	0.99	ICA1	ENSG00000003147	CC: synaptic vesicle membrane
CG32264	NA	0.99	PHACTR1	ENSG00000112137	CC: synapse
CG13830	NA	0.99	SPOCK2	ENSG00000107742	BP: synapse assembly
CG10637	Nak	0.98	AAK1	ENSG00000115977	CC: terminal bouton (of the axon)
CG3269	Rab2	0.98	RAB14	ENSG00000119396	BP: neurotransmitter secretion
CG10011	NA	0.98	ANK2	ENSG00000145362	CC: postsynaptic membrane
CG8440	Lis-1	0.98	PAFAH1B1	ENSG00000007168	BP: synaptic transmission
CG5650	Pp1-87B	0.98	PPP1CA	ENSG00000172531	CC: dendritic spine
CG10579	Eip63E	0.98	CDK16	ENSG00000102225	CC: synaptic vesicle
CG10538	CdGAPr	0.98	ARHGAP32	ENSG00000134909	CC: postsynaptic membrane
CG7535	GluClalpha	0.98	CHRNA5	ENSG00000169684	BP: synaptic transmission
CG8726	NA	0.98	PXK	ENSG00000168297	BP: regulation of synaptic transmission
CG30389	NA	0.98	TMEM57	ENSG00000204178	CC: synapse part
CG7546	NA	0.98	BAG6	ENSG00000204463	BP: synaptonemal complex assembly
CG1506	Ac3	0.98	ADCY3	ENSG00000138031	BP: synaptic transmission
CG6214	MRP	0.98	ABCC8	ENSG00000006071	BP: synaptic transmission
CG4574	Plc21C	0.98	PLCB1	ENSG00000182621	BP: synaptic transmission
CG11734	HERC2	0.98	HERC1	ENSG00000103657	MF: neurotrans:Na symporter activity
CG42829	CadN2	0.97	CDH2	ENSG00000170558	CC: synapse
CG6383	crb	0.97	DNER	ENSG00000187957	BP: synapse assembly
CG1862	Ephrin	0.97	EFNB1	ENSG00000090776	CC: synapse
CG7100	Cadherin-N	0.97	CDH1	ENSG00000039068	BP: synapse assembly
CG8948	Graf	0.97	OPHN1	ENSG00000079482	BP: synaptic vesicle endocytosis
CG9361	Task7	0.97	KCNK3	ENSG00000171303	BP: synaptic transmission
CG6998	ctp	0.97	DYNLL2	ENSG00000121083	BP: synaptic target recognition
CG18455	Optix	0.97	SIX1	ENSG00000126778	BP: reg. of synaptic growth at nj

CG5912	arr	0.96	LRP6	ENSG00000070018	BP: synaptic transmission CC: synapse
CG8261	Ggamma1	0.96	GNG10	ENSG00000242616	BP: synaptic transmission
CG2849	Rala	0.96	RIT2	ENSG00000152214	BP: synaptic transmission
CG32217	Su(Tpl)	0.96	MARVELD2	ENSG00000152939	MF: neurotrans:Na symporter activity
CG17336	Lcch3	0.96	GABRB1	ENSG00000163288	BP: synaptic transmission
CG15274	GABA-B-R1	0.96	GABBR1	ENSG00000204681	BP: synaptic transmission
CG4244	Su(dx)	0.96	NEDD4	ENSG00000069869	BP: regulation of synapse organization
CG16757	Spn	0.96	PPP1R9A	ENSG00000158528	CC: synapse
CG4625	Dhap-at	0.96	GNPAT	ENSG00000116906	BP: synapse assembly
CG32434	siz	0.96	IQSEC3	ENSG00000120645	CC: inhibitory synapse
CG9491	Gef26	0.95	RAPGEF2	ENSG00000109756	BP: regulation of synaptic plasticity
CG42314	PMCA	0.95	ATP2B2	ENSG00000157087	BP: regulation of synaptic plasticity
CG7223	htl	0.95	FGFR2	ENSG00000066468	BP: synaptic vesicle transport
CG9375	Ras 85D	0.95	HRAS	ENSG00000174775	BP: long-term synaptic plasticity
CG30388	Magi	0.95	MAGI2	ENSG00000187391	CC: synapse
CG11958	Cnx99A	0.95	CANX	ENSG00000127022	BP: synaptic vesicle endocytosis
CG9985	sktl	0.94	PIP5K1C	ENSG00000186111	BP: synaptic vesicle exo and endocitosis
CG8726	NA	0.94	KCNK18	ENSG00000186795	BP: synaptic transmission
CG7641	Nca	0.94	NCALD	ENSG00000104490	BP. synaptic transmission
CG8394	VGAT	0.94	SLC32A1	ENSG00000101438	BP: synaptic transmission
CG3585	Rbcn-3A	0.94	DMXL2	ENSG00000104093	CC: synaptic vesicle
CG7558	Arp3	0.94	ACTR3	ENSG00000115091	CC: excitatory synapse
CG14145	Blos2	0.94	BLOC1S2	ENSG00000196072	BP: synaptic vesicle transport
CG1407	NA	0.94	ZDHHC15	ENSG00000102383	BP: synaptic vesicle maturation
CG31196	14-3-3epsilon	0.94	YWHAE	ENSG00000128245	BP: regulation of synaptic plasticity
CG16928	mre11	0.94	MRE11A	ENSG00000020922	CC: synapsis
CG8705	pnut	0.92	SEPT5	ENSG00000184702	BP: synaptic vesicle targeting

Tabla 3. Genes de Drosophila de nuestro catálogo final, con homólogo humano anotado como sináptico *– Las columnas de la tabla muestran respectivamente; el símbolo de anotación de cada gen, el símbolo común de cada gen, el promedio entre la probabilidad de ser sináptico que le fue adjudicado por los tres clasificadores, el correspondiente homólogo humano sináptico y el término GO asociado a sinapsis con el que ese homólogo está anotado.*

Comparación con una lista de proteínas sinápticas de rata.

En un artículo científico recientemente publicado [89], Wilhelm y colaboradores seleccionaron un conjunto de proteínas de rata para las que ha sido bien establecida una localización en la sinapsis. Debido a que el objetivo de ese trabajo fue obtener una reconstrucción tridimensional de una sinapsis "promedio",

la lista de proteínas seleccionadas incluye no solo al tipo de proteínas que han sido definidas como sinápticas en esta tesis, sino también a proteínas mucho más ubicuas. La lista de Wilhelm contiene por ejemplo proteínas de citoesqueleto o de mitocondria, que no solo forman parte de la sinapsis sino también de una gran variedad de estructuras sub-celulares. No obstante, consideramos que una comparación de esta lista con nuestro catálogo final podría ser un buen test adicional de la calidad de nuestro catálogo. Con ese fin, primero, "tradujimos" esta lista de proteínas de rata a una lista de genes de *Drosophila*. Luego de excluir de esa lista aquellas que no cumpliesen con nuestra definición de "proteína sináptica" (por ejemplo, proteínas de citoesqueleto) o cuyos genes codificantes no tuviesen homólogos en *Drosophila*, obtuvimos una lista de 53 genes de *Drosophila*. Estos 53 genes de *Drosophila* son genes cuyos homólogos en rata tienen una función sináptica (tal como la definimos en esta tesis) bien establecida experimentalmente.

De esos 53 genes, 14 habían sido incluidos en nuestra lista de entrenamiento. De los 39 genes restantes, 28 ya estaban anotados con al menos un término GO asociado a sinapsis, por lo cual habían sido selectivamente excluidos de nuestro catálogo final. Son entonces 11 los genes de la lista de Wilhelm "traducida", que no están en nuestra muestra de entrenamiento ni están anotados como sinápticos en *Drosophila.* Encontramos que 9 de esos 11 genes pertenecen a nuestro catálogo consenso inicial, y que 5 pertenecen a nuestro catálogo final de genes potencialmente sinápticos. Estos resultados se resumen en la Tabla 4.

Proteína (rata)	homologo en Drosophila	Catálogo inicial	Catálogo final
VGlut 1/2	CG10069	si	si
calmodulin	CG8472	si	no
NSF	CG31495	si	no
AP-2 mu2	CG10637	si	si
SGIP1	CG8176	si	si
endophilin II	CG9834	si	si
Hsc70	CG8937	no	no
PIPK Ig	CG9985	si	si
Vti1a	CG3279	si	no
VAMP4	CG1599	si	no
calbindin	CG6702	no	no

Tabla 4. Genes de la lista de Wilhelm et al. *[89]* ***que no están en nuestra muestra de entrenamiento ni están anotados como sinápticos en Drosophila.*** *La primera columna muestra el nombre de la proteína de rata cuyos homólogos en Drosophila no están anotados como sinápticos en la mosca ni fueron incluidos en nuestra muestra de entrenamiento. La segunda columna muestra el nombre del correspondiente gen homólogo en Drosophila. La tercera columna indica si ese gen de Drosophila fue clasificado como sináptico con una probabilidad mayor a 0,5 por los tres algoritmos. La cuarta columna indica si el gen forma parte del catálogo final, por haber sido clasificado como sináptico por los tres algoritmos con una probabilidad mayor a 0,9.*

DISCUSIÓN

Uno de los principales objetivos de la Genómica Funcional es adjudicar nuevas funciones a genes a partir de la información almacenada en las grandes bases de datos biológicos [90]. Hacer predicciones acerca de las posibles funciones de un gen de modo verificable es además una manera de maximizar la utilidad de ese enorme volumen de información disponible. En este trabajo hemos demostrado que es posible obtener catálogos de genes enriquecidos en genes de importancia para la sinapsis neuronal analizando un transcriptoma temporal del desarrollo y de cuerpo completo, mediante una combinación de algoritmos de aprendizaje supervisado y un abordaje bioinformático original.

Hace unos 25 años, los primeros estudios que analizaban datos de expresión genómica con métodos de aprendizaje automático sugerían una relación funcional entre aquellos genes que muestran patrones de expresión similares. Por ejemplo, la *clusterización* de genes de levadura en base a las similitudes en sus patrones de expresión dio lugar a la definición de grupos de genes que comparten importantes similitudes funcionales [15, 90]. También se demostró una correlación entre patrones de expresión y función biológica en *Drosophila* y en humanos [10, 11, 18,

91], aunque en éste último caso esa correlación fue menos evidente dada la mucho más incompleta anotación del genoma de *H. sapiens* [14]. Por otro lado, una clara correspondencia entre grupos funcionales de genes con patrones de expresión específicos fue demostrada en ratas [16]. Desde entonces, el uso de métodos de aprendizaje automático para asignarle funciones a genes, al menos de modo tentativo, y basándose no solamente en sus patrones de expresión sino también en interacciones proteína-proteína, similitudes estructurales o de secuencia, ha conducido a una amplia diversidad de estrategias y a una profusa bibliografía. [20–24, 28].

El antecedente más directo de éste trabajo es un estudio de Yan y colaboradores [90] publicado en el año 2010. En ese estudio, se entrenaron clasificadores específicos por función basados en *Random Forest* para predecir asociaciones entre términos GO y genes de *Drosophila*. Para entrenar los clasificadores se usaron redes de interacción y conservación de dominios de proteínas, perfiles de expresión e interacciones genéticas y similitudes en secuencias como variables predictivas. En lo que tiene que ver con el ensamblaje y el funcionamiento de la sinapsis, el estudio predice términos GO que incluyen la palabra “sinapsis”, “sináptico” o “neurotransmisor” solamente para 31 genes. Dada la cantidad esperada de genes sinápticos, una lista tan pequeña tiene escasa utilidad, más allá de su eventual calidad.

Nuestro estudio difiere de éste y otros estudios previos en varios aspectos importantes. Hasta donde sabemos, es el primero en aplicar algoritmos de aprendizaje automático para predecir funciones de genes usando datos transcriptómicos obtenidos mediante las nuevas tecnologías de secuenciación masiva. Además, el transcriptoma que aquí utilizamos presenta varias ventajas respecto a otros sets de datos de RNAseq disponibles de *Drosophila*. En tanto la formación del cerebro en el embrión de la mosca es un proceso muy rápido y los perfiles de transcripción muestran una buena correlación temporal con las secuencia de procesos biológicos que ocurren en ese período, este set de datos ofrece una clara ventaja para la definición de genes potencialmente sinápticos. La característica clave que convierte a este set de datos en el mejor disponible para nuestro abordaje, es que abarca varias etapas del ciclo vital de la mosca que son relevantes para nuestra clasificación. Las etapas contenidas en el set de datos

utilizado incluyen un momento en el que no existen sinapsis en el organismo, dos períodos de intensa sinaptogénesis y una etapa de desensamblaje masivo de sinapsis, todo lo cual mejora el potencial de los algoritmos para distinguir genes sinápticos.

Otro aspecto novedoso de nuestro estudio radica en que la muestra de entrenamiento fue construida de modo tal que difiere sustancialmente del conjunto de genes que están anotados con algún término GO asociado a la sinapsis. Esa estrategia evitó problemas de circularidad a la hora de evaluar el resultado de la clasificación mediante análisis de enriquecimiento funcional. Dado que el genoma de *Drosophila* es uno de los mejor anotados [92, 93], lo anterior implica una enorme ventaja respecto a otros estudios en los que el desempeño de los clasificadores solo pudo ser evaluado mediante validación cruzada sobre la muestra de entrenamiento, curvas ROC, o por medio de una revisión de la literatura relativa a los genes con las mejores predicciones. Es importante resaltar que nuestra muestra de entrenamiento solamente incluye como genes sinápticos a genes cuya importancia para la formación y funcionamiento de la sinapsis había sido experimental e inequívocamente demostrada. Tampoco incluimos genes con funciones más generales (como por ejemplo "guía de los axones" o "unión neuromuscular") para hacer nuestro análisis más neutral respecto a las diferencias existentes entre distintos organismos en cuanto a la morfología de sus dendritas y sus terminales axonales.

Otra característica importante de nuestro estudio es que seguimos un procedimiento que, al entrenar tres algoritmos distintos y considerar la intersección de sus clasificaciones, mejora el resultado de la clasificación [13, 76]. La ventaja de este procedimiento queda ilustrada por el hecho de que dado un umbral de clasificación, el enriquecimiento funcional en términos GO asociados a sinapsis del catálogo consenso siempre es mayor al enriquecimiento de los catálogos producidos por cualquiera de los tres algoritmos por separado. Este procedimiento tiene la ventaja adicional de reducir la probabilidad de incluir falsos positivos en nuestro catálogo final. En tanto nuestro objetivo no es la obtención de un catálogo exhaustivo de todos los genes sinápticos, sino la obtención de un catálogo de genes con alta probabilidad de ser sinápticos, es más conveniente disminuir el número de falsos positivos aun cuando el costo sea aumentar el

número de falsos negativos. En este contexto, cabe resaltar las bajas tasas de error alcanzadas por los tres algoritmos y sus excelentes desempeños, estimados mediante el área bajo sus curvas ROC.

Finalmente, vale la pena remarcar el hecho de que un 11% de los 893 genes de *Drosophila* que conforman nuestro catálogo final, son genes cuya importancia para la sinapsis ya ha sido bien establecida en humanos. La relevancia de este resultado se desprende del hecho de que esos genes sinápticos humanos son homólogos de genes de *Drosophila* que pertenecen a un catálogo del cual cualquier gen que estuviese anotado como sináptico había sido selectivamente excluido. Consideramos el hecho de que nuestro catálogo final incluya tres veces más de lo esperado a genes con homólogos humanos sinápticos, como una indicación adicional de la buena calidad de la predicción. Por otro lado, nuestro método clasifica como sinápticos a 9 de 11 genes de importancia para la sinapsis en ratas [89] pero cuya importancia en la sinapsis de *Drosophila* aun no ha sido establecida. Teniendo en cuenta el alto grado de conservación funcional existente entre genes homólogos de estas especies, creemos que estas coincidencias representan un fuerte argumento a favor de la conveniencia de nuestro abordaje y de la calidad predictiva de nuestro catálogo final.

La fuerte correlación entre umbral de clasificación, enriquecimiento funcional y proporción de GEDTEs en el SNC, junto a la observación de que un 11% de los genes de *Drosophila* de nuestro catálogo final tienen homólogos en humanos con función sináptica ya descrita, sugieren muy fuertemente que nuestro catálogo final está muy enriquecido en genes de importancia para el ensamblaje y el funcionamiento de la sinapsis de *Drosophila*, pero que aun no han sido reconocidos como tales.

CONCLUSIONES

La utilidad del abordaje utilizado en este estudio radica en que reduce el número de genes a ensayar experimentalmente a la hora de determinar qué genes del genóma de un organismo son importantes para una función biológica determinada. Es por ello que hacemos disponible esta lista de 893 genes potencialmente sinápticos, en el entendido de que la misma facilitará su estudio mediante silenciamiento genético, análisis de mutantes, ensayos de conducta u otros protocolos tradicionales. Es altamente probable que el estudio experimental de estos genes conduzca a la identificación de nuevos genes de importancia para el funcionamiento normal de la sinápsis.

REFERENCIAS

1. Sherrington CS: **Part III - Nervous System**. In *Foster's Textbook of Physiology*. Séptima edición. Londres: Macmillan; 1897.

2. Real Academia Española: *Diccionario de La Lengua Española*. Vigésima segunda edición.; 2001.

3. Cowan WM, Kandel ER: **A brief history of synapses and synaptic transmission**. In *Synapses*. London: Johns Hopkins Univ. Press; 2001:1–87.

4. Rose S: *The Making Of Memory: From Molecules to Mind*. Random House; 2012.

5. Prokop A, Meinertzhagen IA: **Development and structure of synaptic contacts in *Drosophila*.** *Semin Cell Dev Biol* 2006, **17**.

6. Kraut R, Menon K, Zinn K: **A gain-of-function screen for genes controlling motor axon guidance and synaptogenesis in *Drosophila*.** *Curr Biol CB* 2001, **11**:417–430.

7. Valakh V, Naylor SA, Berns DS, DiAntonio A: **A large-scale RNAi screen identifies functional classes of genes shaping synaptic development and maintenance.** *Dev Biol* 2012, **366**:163–171.

8. Sieburth D, Ch'ng Q, Dybbs M, Tavazoie M, Kennedy S, Wang D, Dupuy D, Rual J-F, Hill DE, Vidal M, Ruvkun G, Kaplan JM: **Systematic analysis of genes required for synapse structure and function.** *Nature* 2005, **436**:510–517.

9. Arbeitman MN, Furlong EEM, Imam F, Johnson E, Null BH, Baker BS, Krasnow MA, Scott MP, Davis RW, White KP: **Gene expression during the life cycle of *Drosophila melanogaster*.** *Science* 2002, **297**:2270–2275.

10. Hooper SD, Boue S, Krause R, Jensen LJ, Mason CE, Ghanim M, White KP, Furlong EEM, Bork P: **Identification of tightly regulated groups of genes during *Drosophila melanogaster* embryogenesis.** *Mol Syst Biol* 2007, **3**.

11. Papatsenko I, Levine M, Papatsenko D: **Temporal waves of coherent gene expression during *Drosophila* embryogenesis.** *Bioinforma Oxf Engl* 2010, **26**:2731–2736.

12. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D: **A combined algorithm for genome-wide prediction of protein function.** *Nature* 1999, **402**:83–86.

13. Schietgat L, Vens C, Struyf J, Blockeel H, Kocev D, Dzeroski S: **Predicting gene function using hierarchical multi-label decision tree ensembles.** *BMC Bioinformatics* 2010, **11**.
14. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278**:680–686.
15. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proc Natl Acad Sci* 1998, **95**:14863–14868.
16. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, Furey TS, Ares MJ, Haussler D: **Knowledge-based analysis of microarray gene expression data by using support vector machines.** *Proc Natl Acad Sci U S A* 2000, **97**:262–267.
17. Wen X, Fuhrman S, Michaels GS, Carr DB, Smith S, Barker JL, Somogyi R: **Large-scale temporal gene expression mapping of central nervous system development.** *Proc Natl Acad Sci U S A* 1998, **95**:334–339.
18. Spellman PT, Rubin GM: **Evidence for large domains of similarly expressed genes in the *Drosophila* genome.** *J Biol* 2002, **1**.
19. White KP, Rifkin SA, Hurban P, Hogness DS: **Microarray analysis of *Drosophila* development during metamorphosis.** *Science* 1999, **286**:2179–2184.
20. Hvidsten TR, Komorowski J, Sandvik AK, Laegreid A: **Predicting gene function from gene expressions and ontologies.** *Pac Symp Biocomput Pac Symp Biocomput* 2001.
21. Lukashin AV, Fuchs R: **Analysis of temporal gene expression profiles: clustering by simulated annealing and determining the optimal number of clusters.** *Bioinforma Oxf Engl* 2001, **17**:405–414.
22. Lagreid A, Hvidsten TR, Midelfart H, Komorowski J, Sandvik AK: **Predicting gene ontology biological process from temporal gene expression patterns.** *Genome Res* 2003, **13**:965–979.
23. Zhang W, Morris QD, Chang R, Shai O, Bakowski MA, Mitsakakis N, Mohammad N, Robinson MD, Zirngibl R, Somogyi E, Laurin N, Eftekharpour E, Sat E, Grigull J, Pan Q, Peng W-T, Krogan N, Greenblatt J, Fehlings M, van der Kooy D, Aubin J, Bruneau BG, Rossant J, Blencowe BJ, Frey BJ, Hughes TR: **The functional landscape of mouse gene expression.** *J Biol* 2004, **3**.
24. Lan H, Carson R, Provart NJ, Bonner AJ: **Combining classifiers to predict gene**

function in *Arabidopsis thaliana* using large-scale gene expression measurements. *BMC Bioinformatics* 2007, **8**.
25. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays.** *Genome Res* 2008, **18**:1509–1517.
26. Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P: **Estimating accuracy of RNA-Seq and microarrays with proteomics.** *BMC Genomics* 2009, **10**.
27. Barrett T, Troup DB, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Muertter RN, Holko M, Ayanbule O, Yefanov A, Soboleva A: **NCBI GEO: archive for functional genomics data sets--10 years on.** *Nucleic Acids Res* 2011, **39**(Database issue):D1005–1010.
28. Bar-Joseph Z, Gitter A, Simon I: **Studying and modelling dynamic biological processes using time-series gene expression data.** *Nat Rev Genet* 2012, **13**:552–564.
29. Song JJ, Lee H-J, Morris JS, Kang S: **Clustering of time-course gene expression data using functional data analysis**. *Comput Biol Chem* 2007, **31**:265–274.
30. Yuan M, Kendziorski C: **Hidden Markov Models for Microarray Time Course Data in Multiple Biological Conditions**. *J Am Stat Assoc* 2006, **101**:1323–1332.
31. Levin M, Hashimshony T, Wagner F, Yanai I: **Developmental milestones punctuate gene expression in the *Caenorhabditis* embryo.** *Dev Cell* 2012, **22**:1101–1108.
32. Technau GM: *Brain Development in Drosophila Melanogaster*. Springer Science+Business Media; 2009.
33. Morgan TH: **Sex Limited Inheritance in *Drosophila***. *Science* 1910, **32**:120–122.
34. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, Kellis M, Lai EC, Lieb JD, MacAlpine DM, Micklem G, Piano F, Snyder M, Stein L, White KP, Waterston RH: **Unlocking the secrets of the genome**. *Nature* 2009, **459**:927–930.
35. Reiter LT, Potocki L, Chien S, Gribskov M, Bier E: **A Systematic Analysis of Human Disease-Associated Gene Sequences In *Drosophila melanogaster***. *Genome Res* 2001, **11**:1114–1125.

36. Costello JC, Dalkilic MM, Beason SM, Gehlhausen JR, Patwardhan R, Middha S, Eads BD, Andrews JR: **Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function.** *Genome Biol* 2009, **10**.
37. Rieder LE, Larschan EN: **Wisdom from the fly**. *Trends Genet* 2014, **30**:479–481.
38. Thor S: **The genetics of brain development: Conserved programs in flies and mice**. *Neuron* , **15**:975–977.
39. Nichols CD: ***Drosophila melanogaster* neurobiology, neuropharmacology, and how the fly can inform central nervous system drug discovery**. *Pharmacol Ther* 2006, **112**:677–700.
40. Gama Sosa M, De Gasperi R, Elder G: **Modeling human neurodegenerative diseases in transgenic systems**. *Hum Genet* 2012, **131**:535–563.
41. Collins CA, DiAntonio A: **Synaptic development: insights from *Drosophila*.** *Curr Opin Neurobiol* 2007, **17**.
42. Frank CA, Wang X, Collins CA, Rodal AA, Yuan Q, Verstreken P, Dickman DK: **New approaches for studying synaptic development, function, and plasticity using *Drosophila* as a model system.** *J Neurosci Off J Soc Neurosci* 2013, **33**:17560–17568.
43. Littleton JT, Ganetzky B: **Ion channels and synaptic organization: analysis of the *Drosophila* genome.** *Neuron* 2000, **26**.
44. Lloyd TE, Verstreken P, Ostrin EJ, Phillippi A, Lichtarge O, Bellen HJ: **A genome-wide search for synaptic vesicle cycle proteins in *Drosophila*.** *Neuron* 2000, **26**.
45. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S, Suzuki AM, Wan KH, Yu C, Zhang D, Carlson JW, Cherbas L, Eads BD, Miller D, Mockaitis K, Roberts J, Davis CA, Frise E, Hammonds AS, Olson S, Shenker S, Sturgill D, Samsonova AA, Weiszmann R, Robinson G, Hernandez J, Andrews J, et al.: **Diversity and dynamics of the *Drosophila* transcriptome**. *Nature* 2014, **512**:393–399.
46. Soustelle L, Giangrande A: **Glial differentiation and the Gcm pathway.** *Neuron Glia Biol* 2007, **3**.
47. Prokop A, Beaven R, Qu Y, Sanchez-Soriano N: **Using fly genetics to dissect the cytoskeletal machinery of neurons during axonal growth and maintenance.** *J Cell Sci* 2013, **126**(Pt 11):2331–2341.

48. Rolls MM: **Neuronal polarity in *Drosophila*: sorting out axons and dendrites.** *Dev Neurobiol* 2011, **71**:419–429.
49. Gao FB, Brenman JE, Jan LY, Jan YN: **Genes regulating dendritic outgrowth, branching, and routing in *Drosophila*.** *Genes Dev* 1999, **13**:2549–2561.
50. Jan Y-N, Jan LY: **Branching out: mechanisms of dendritic arborization.** *Nat Rev Neurosci* 2010, **11**:316–328.
51. Hastie T, Tibshirani R, Friedman JH: *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York: Springer; 2009.
52. Zhao X-M, Wang Y, Chen L, Aihara K: **Gene function prediction using labeled and unlabeled data.** *BMC Bioinformatics* 2008, **9**.
53. Mitsakakis N, Razak Z, Escobar M, Westwood JT: **Prediction of *Drosophila melanogaster* gene function using Support Vector Machines.** *BioData Min* 2013, **6**.
54. Consortium TGO: **Gene Ontology: tool for the unification of biology.** *Nat Genet* 2000, **25**:25 – 29.
55. Thomas JB, Wyman RJ: **Mutations altering synaptic connectivity between identified neurons in *Drosophila*.** *J Neurosci Off J Soc Neurosci* 1984, **4**:530–538.
56. Kopczynski CC, Davis GW, Goodman CS: **A neural tetraspanin, encoded by late bloomer, that facilitates synapse formation.** *Science* 1996, **271**:1867–1870.
57. Gorczyca M, Popova E, Jia XX, Budnik V: **The gene mod(mdg4) affects synapse specificity and structure in *Drosophila*.** *J Neurobiol* 1999, **39**:447–460.
58. Wan HI, DiAntonio A, Fetter RD, Bergstrom K, Strauss R, Goodman CS: **Highwire regulates synaptic growth in *Drosophila*.** *Neuron* 2000, **26**:313–329.
59. Featherstone DE, Broadie K: **Surprises from *Drosophila*: genetic mechanisms of synaptic development and plasticity.** *Brain Res Bull* 2000, **53**:501–511.
60. Rieckhof GE, Yoshihara M, Guan Z, Littleton JT: **Presynaptic N-type calcium channels regulate synaptic growth.** *J Biol Chem* 2003, **278**:41099–41108.
61. Long AA, Mahapatra CT, Woodruff EA 3rd, Rohrbough J, Leung H-T, Shino S, An L, Doerge RW, Metzstein MM, Pak WL, Broadie K: **The nonsense-mediated decay pathway maintains synapse architecture and synaptic vesicle cycle efficacy.** *J Cell Sci* 2010, **123**(Pt 19):3303–3315.
62. Broadie K, Baumgartner S, Prokop A: **Extracellular matrix and its receptors**

in *Drosophila* neural development. *Dev Neurobiol* 2011, **71**:1102–1130.
63. Sigrist SJ, Schmitz D: **Structural and functional plasticity of the cytoplasmic active zone.** *Curr Opin Neurobiol* 2011, **21**:144–150.
64. Sudhof TC: **The synaptic vesicle cycle.** *Annu Rev Neurosci* 2004, **27**:509–547.
65. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK: **A mitochondrial protein compendium elucidates complex I disease biology.** *Cell* 2008, **134**:112–123.
66. Grant SGN: **The synapse proteome and phosphoproteome: a new paradigm for synapse biology.** *Biochem Soc Trans* 2006, **34**(Pt 1).
67. Jordan BA, Fernholz BD, Boussac M, Xu C, Grigorean G, Ziff EB, Neubert TA: **Identification and verification of novel rodent postsynaptic density proteins.** *Mol Cell Proteomics MCP* 2004, **3**:857–871.
68. Li K wan, Hornshaw MP, van Minnen J, Smalla K-H, Gundelfinger ED, Smit AB: **Organelle proteomics of rat synaptic proteins: correlation-profiling by isotope-coded affinity tagging in conjunction with liquid chromatography-tandem mass spectrometry to reveal post-synaptic density specific proteins.** *J Proteome Res* 2005, **4**:725–733.
69. Garner CC, Waites CL, Ziv NE: **Synapse development: still looking for the forest, still lost in the trees.** *Cell Tissue Res* 2006, **326**:249–262.
70. Zhang W, Zhang Y, Zheng H, Zhang C, Xiong W, Olyarchuk JG, Walker M, Xu W, Zhao M, Zhao S, Zhou Z, Wei L: **SynDB: a Synapse protein DataBase based on synapse ontology.** *Nucleic Acids Res* 2007, **35**(Database issue):D737–741.
71. Pirooznia M, Wang T, Avramopoulos D, Valle D, Thomas G, Huganir RL, Goes FS, Potash JB, Zandi PP: **SynaptomeDB: an ontology-based knowledgebase for synaptic genes.** *Bioinforma Oxf Engl* 2012, **28**:897–899.
72. Morciano M, Beckhaus T, Karas M, Zimmermann H, Volknandt W: **The proteome of the presynaptic active zone: from docked synaptic vesicles to adhesion molecules and maxi-channels**. *J Neurochem* 2009, **108**:662–675.
73. Coughenour HD, Spaulding RS, Thompson CM: **The synaptic vesicle proteome: A comparative study in membrane protein identification**. *PROTEOMICS* 2004, **4**:3141–3155.
74. Graveley BR, Brooks AN, Carlson JW, Duff MO, Landolin JM, Yang L, Artieri CG,

van Baren MJ, Boley N, Booth BW, Brown JB, Cherbas L, Davis CA, Dobin A, Li R, Lin W, Malone JH, Mattiuzzo NR, Miller D, Sturgill D, Tuch BB, Zaleski C, Zhang D, Blanchette M, Dudoit S, Eads B, Green RE, Hammonds A, Jiang L, Kapranov P, et al.: **The developmental transcriptome of *Drosophila melanogaster***. *Nature* 2011, **471**:473–479.

75. Caruana R, Niculescu-Mizil A: **An empirical comparison of supervised learning algorithms**. In *Proceedings of the 23rd international conference on Machine learning*. Pittsburgh, Pennsylvania: ACM; 2006:161–168.

76. Vinayagam A, Konig R, Moormann J, Schubert F, Eils R, Glatting K-H, Suhai S: **Applying Support Vector Machines for Gene Ontology based gene function prediction.** *BMC Bioinformatics* 2004, **5**.

77. R Development Core Team: **R Development Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.** .

78. Wiener M. LA: **Classification and Regression by randomForest**. *R News* 2002, **2**:18–22.

79. Friedrich Leisch, Andreas Weingessel, Kurt Hornik, Evgenia Dimitriadou, David Meyer: **e1071: Misc Functions of the Department of Statistics (e1071), TU Wien.** *R package version 1.6-1* 2012.

80. Venables, W. N., Ripley, B. D.: *Modern Applied Statistics with S*. 4th edition. New York: Springer; 2002.

81. Breiman L: **Random Forests**. *Mach Learn* 2001, **45**:5–32.

82. Vapnik V: *Statistical Learning Theory*. Wiley; 1998.

83. Chintapalli VR, Wang J, Dow JAT: **Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease.** *Nat Genet* 2007, **39**:715–720.

84. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z: **GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists.** *BMC Bioinformatics* 2009, **10**.

85. Huang DW, Sherman BT, Lempicki RA: **Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists.** *Nucleic Acids Res* 2009, **37**.

86. Gelbart WM, Emmert DB: **FlyBase High Throughput Expression Pattern**

Data Beta Version. 2010.
87. Lyne R, Smith R, Rutherford K, Wakeling M, Varley A, Guillier F, Janssens H, Ji W, Mclaren P, North P, Rana D, Riley T, Sullivan J, Watkins X, Woodbridge M, Lilley K, Russell S, Ashburner M, Mizuguchi K, Micklem G: **FlyMine: an integrated database for *Drosophila* and *Anopheles* genomics**. *Genome Biol* 2007, **8**:R129.
88. Lyne M, Smith RN, Lyne R, Aleksic J, Hu F, Kalderimis A, Stepan R, Micklem G: **metabolicMine: an integrated genomics, genetics and proteomics data warehouse for common metabolic disease research**. *Database* 2013, **2013**.
89. Wilhelm BG, Mandad S, Truckenbrodt S, Krohnert K, Schafer C, Rammner B, Koo SJ, Classen GA, Krauss M, Haucke V, Urlaub H, Rizzoli SO: **Composition of isolated synaptic boutons reveals the amounts of vesicle trafficking proteins.** *Science* 2014, **344**:1023–1028.
90. Yan H, Venkatesan K, Beaver JE, Klitgord N, Yildirim MA, Hao T, Hill DE, Cusick ME, Perrimon N, Roth FP, Vidal M: **A genome-wide gene function prediction resource for *Drosophila melanogaster*.** *PloS One* 2010, **5**.
91. Weber CC, Hurst LD: **Support for multiple classes of local expression clusters in *Drosophila melanogaster*, but no evidence for gene order conservation.** *Genome Biol* 2011, **12**.
92. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF, George RA, Lewis SE, Richards S, Ashburner M, Henderson SN, Sutton GG, Wortman JR, Yandell MD, Zhang Q, Chen LX, Brandon RC, Rogers YH, Blazej RG, Champe M, Pfeiffer BD, Wan KH, Doyle C, Baxter EG, Helt G, Nelson CR, et al.: **The genome sequence of *Drosophila melanogaster*.** *Science* 2000, **287**:2185–2195.
93. McQuilton P, St Pierre SE, Thurmond J: **FlyBase 101--the basics of navigating FlyBase.** *Nucleic Acids Res* 2012, **40**(Database issue):D706–714.