Predicción de función de genes mediante aprendizaje automático, con énfasis en el estudio de los patrones de ubicación de grupos funcionales de genes.

MSc. Flavio Pazos Obregón

Tesis de Doctorado en Ciencias Biológicas - PEDECIBA

Orientadores :

Dr. Rafael Cantera,

Departamento de Biología del Neurodesarrollo.

Instituto de Investigaciones Biológicas Clemente Estable, IIBCE, MEC.

Dr. Gustavo Guerberoff

Laboratorio de Probabilidad y Estadística del Instituto de Matemática y Estadística "Prof. Ing. Rafael Laguardia" de la Facultad de Ingeniería – UDELAR

Dr. Patricio Yankilevich

Plataforma de Bioinformática

Instituto de Investigación en Biomedicina de Buenos Aires (IBioBA) – CONICET

Partner Institute of the Max Planck Society

a mis hijos

Resumen

El trabajo aquí reunido tuvo como objetivo principal avanzar en el campo de la predicción de función de genes mediante aprendizaje automático. El Capítulo I recoge resultados obtenidos al mejorar un modelo de aprendizaje automático supervisado que predice función sináptica en genes de *Drosophila melanogaster* mediante mejoras en la estrategia de entrenamiento, trabajo que dio lugar a una publicación científica (Pazos Obregón et al. 2019). El Capítulo II presenta Cluster Locator, una herramienta en línea que permite el análisis estadístico de los patrones de distribución de listas de genes a lo largo del genoma, trabajo que también dio lugar a una publicación (Pazos Obregón et al. 2018). Luego se presentan los resultados obtenidos con Cluster Locator buscando probar la existencia de grupos de genes con la misma función y con patrones de distribución similares en distintos organismos, así como de grupos de genes del mismo organismo con funciones y perfiles de agrupamiento similares. Finalmente, el Capítulo III reúne resultados obtenidos al entrenar modelos de aprendizaje automático para predecir nuevas funciones de genes en 5 organismos modelo (*Homo sapiens, Mus musculus, Drosophila melanogaster, Caenorhabditis elegans* y *Saccharomyces cerevisiae)* utilizando como únicas variables predictivas ciertas características derivadas de la ubicación de los genes en el genoma.

Índice resumido

Introducción **pág. 4**

- Predicción de función de genes. pág. 4
- Aprendizaje automático. pág. 7
- Agrupamiento de genes funcionalmente relacionados. pág. 9

Objetivos **pág. 13**

Capítulo I - Predicción de función sináptica en genes de *Drosophila melanogaster* **pág. 14**

1.1 Antecedentes. pág. 14

1.2 Evaluación de nuestra predicción original y propuestas para mejorarla. pág. 15

1.3 Un catálogo mejorado de genes potencialmente sinápticos. pág. 23

1.4 Discusión. pág. 24

Capítulo II - Patrones de distribución de grupos funcionales de genes **pág. 26**

2.1 Antecedentes. pág. 26

2.2 Cluster Locator, una herramienta para el análisis del agrupamiento de genes. pág. 27

2.3 Patrones de distribución de grupos funcionales de genes en 5 organismos. pág. 35

2.4 Discusión. pág. 46

Capítulo III - Predicción de función de genes a partir de su ubicación **pág. 50**

3.1 - La ubicación relativa de un gen como variable predictiva de sus funciones. pág. 50

3.2 - Análisis de enriquecimiento funcional local. pág. 55

3.3 - Predicción de función de genes a partir del enriquecimiento funcional local. pág. 59

3.4 - Discusión. pág. 64

Conclusiones generales **pág. 67**

Métodos **pág. 68**

Colaboraciones **pág. 80**

Financiación **pág. 81**

Referencias **pág. 82**

Anexos

A1 - An improved catalogue of putative synaptic genes defined exclusively by temporal. transcription profiles through an ensemble machine learning approach.

A2 - Cluster Locator, online analysis and visualization of gene clustering.

A3 - Precisión, sensibilidad y score F1 de los modelos predictivos según el umbral de clasificación.

Introducción

Predicción de función de genes

Se puede entender a los seres vivos como el resultado de la interacción dinámica entre el ambiente en el que viven y la carga de información heredada que portan. Una parte importante de esa información heredada está codificada en los genes y por esa razón uno de los principales objetivos de la biología moderna es determinar sus funciones, esto es, el rol que juegan en la vida de los organismos los productos de su expresión. Determinar la función de un gen en particular puede ser de enorme relevancia en campos tan distintos como la ecología, la psiquiatría, la salud, la producción de alimentos o de energía y en general, para cualquier forma de conocimiento que se beneficie del entendimiento de los procesos biológicos.

La gran mayoría de los genes conocidos no tiene ninguna función asignada (Zerbino et al. 2018). Además, como a muchos de los genes mejor estudiados ya se les han asignado varias funciones, lo más probable es que aquellos genes que tienen una sola función asignada, tengan otras aún por descubrirse (Rubin y Green 2013). Existen distintas maneras de asignarle función a un gen, que se pueden dividir en dos grandes familias de métodos: experimentales o computacionales (Ashburner et al. 2000). Aunque esta distinción es cada vez más difusa, los métodos experimentales tienen en común que involucran algún tipo de manipulación del objeto de estudio en sí. Ejemplos de métodos experimentales para la determinación de la función de un gen son los ensayos genéticos, bioquímicos o fisiológicos con especímenes que portan mutaciones que afectan la expresión de uno o más genes, usando por ejemplo técnicas de biología molecular como el doble híbrido de levadura o el silenciamiento génico. Durante décadas, una estrategia muy productiva fue generar miles de mutantes por medio de rayos X o por administración de una sustancia mutagénica, para luego identificar, por medio de microscopía u otros medios, cuáles de esos mutantes han sufrido una disminución o pérdida de la función estudiada (Hartwell, Culotti, y Reid 1970; Nüsslein-Volhard y Wieschaus 1980). En general, este tipo de estudios requiere mucho dinero, tiempo y personal altamente calificado, además de tener asociados riesgos para la salud y el ambiente.

Por otro lado, los métodos computacionales asignan funciones a genes analizando datos preexistentes, a partir de los cuales se realiza algún tipo de cálculo para determinar si hay aspectos de la biología de un gen de función desconocida que se asemejen a los de otros, de los que ya se conoce su función. Uno de los primeros métodos computacionales mediante los cuales se comenzó a asignar funciones a genes se basa en algoritmos de alineamiento de secuencias para la

identificación de secuencias homólogas en bases de datos. Si la secuencia nucleotídica de un gen de función desconocida es suficientemente parecida a la de otro gen con función conocida, esta función es asignada al primer gen (Altschul et al. 1990).

Sin embargo, se considera que la evidencia experimental directa es necesaria (y suficiente) para probar que cierto gen tiene una determinada función en cierto organismo. Cuando una función es asignada a un gen a través de métodos computacionales, se dice que esa función está "predicha" hasta que la misma sea confirmada experimentalmente. Estas predicciones, si bien deben ser validadas experimentalmente, tienen la ventaja de hacer más eficiente la investigación tradicional al dirigir la experimentación a listas de genes con alta probabilidad de ser necesarios para una función biológica particular, ahorrando tiempo y recursos.

Una de las fronteras que actualmente tiene el avance del entendimiento biológico está dada por el hecho de que determinar experimentalmente las funciones de los genes conocidos, por la propia naturaleza de los métodos experimentales, tomaría siglos. Además, si se tiene en cuenta que por el abaratamiento de las tecnologías de secuenciación masiva la cantidad de genes y genomas conocidos crece cada vez más rápido, en las próximas décadas el tiempo que hará falta para determinar experimentalmente las funciones de los genes conocidos solo crecerá (ver Figura I.1).

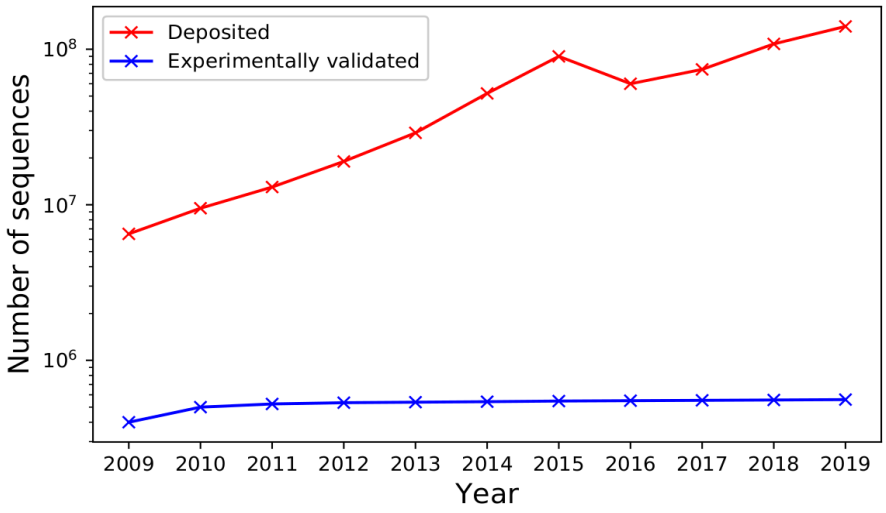


Figura I.1. Cantidad de secuencias depositadas (rojo) y validadas experimentalmente (azul) a lo largo de la ultima década en UniProt. La caída observada entre 2015 y 2016 se debe a una limpieza masiva de secuencias redundantes llevada a cabo por curadores humanos. Notar que la escala del eje de las ordenadas *es logarítmica (tomada de Bonneta et al., 2020).*

En este contexto, el campo de la predicción computacional de funciones de genes ha crecido mucho (ver las revisiones de (Bernardes y Pedreira 2013; Libbrecht y Noble 2015; Zhou et al. 2019; Zhao et al. 2020; Bonetta y Valentino 2020). La Figura I.2 , ilustrativa del crecimiento que ha tenido el área, muestra la cantidad de artículos científicos publicados por año, a lo largo de la última década, que abordan la predicción de función de genes basándose en *Gene Ontology* (*Gene Ontology* es una de varias formas de representar las funciones de los genes y se describe en detalle más adelante).

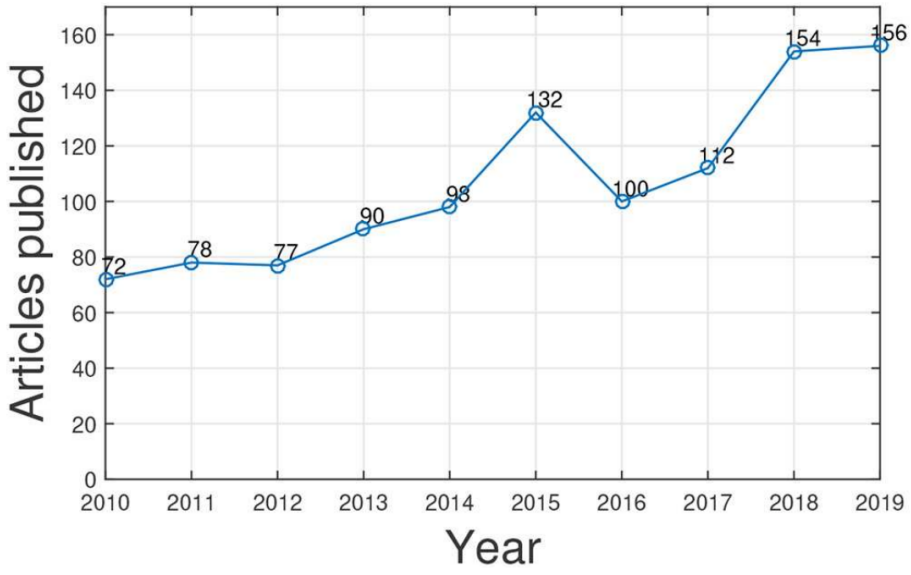


Figura I.2 - Cantidad de artículos científicos publicados por año a lo largo de la última década que abordan la predicción de función de genes basada en Gene Ontology. Tomada de (Zhao et al. 2020)

Dada esa proliferación de abordajes, que combinan distintos algoritmos de aprendizaje, esquemas de entrenamiento, arquitecturas de modelo y variables predictivas (entre otras cosas), surge la necesidad de evaluar y comparar sus desempeños. Ese es el objetivo de los Critical Assessment of Function Annotation, conocidos como los "desafíos CAFA". Los desafíos CAFA son experimentos diseñados para obtener una evaluación general de la diversidad de métodos computacionales utilizados en la predicción de función de proteínas (Radivojac et al. 2013a; Jiang et al. 2016; Zhou et al. 2019). Estas evaluaciones, cuya cuarta edición está en marcha, tienen un formato de competencia, en la que los equipos de investigadores participantes reciben miles de secuencias de proteínas sin función conocida y deben predecir sus funciones en un tiempo

predeterminado. Luego, se deja pasar cierto período durante el cual se acumulan nuevas funciones asignadas experimentalmente, que son utilizadas para evaluar y comparar las predicciones hechas por cada método. En su tercera edición (Zhou et al. 2019), participaron de esta competencia 167 investigadores que totalizan 123 filiaciones en distintas instituciones: 65 en Europa, 36 en Estados Unidos y 22 en Asia. Es de notar la ausencia de instituciones latinoamericanas.

En estas competencias, la mayoría de los abordajes que obtienen los mejores resultados se basan en los principios del aprendizaje automático, tema que tratamos en la siguiente sección.

Aprendizaje automático

El aprendizaje automático, término acuñado hace más de 60 años (Samuel 1959), es una rama de la inteligencia artificial que desarrolla algoritmos capaces de generalizar conocimiento a partir de información no estructurada, suministrada en forma de ejemplos. Su objetivo es que un computador aprenda, mejorando su desempeño con la experiencia. Una de sus aplicaciones más habituales es la de decidir cuándo algo pertenece a una categoría de cosas y cuándo no. Las categorías pueden estar pre-establecidas, en cuyo caso se habla de aprendizaje supervisado, o pueden no estarlo, hablándose entonces de aprendizaje no supervisado (Hastie, Tibshirani, y Friedman 2009; Li, Wu, y Ngom 2018).

Un escenario típico en el que se aplica aprendizaje automático supervisado incluye algún tipo de información de salida que se desea predecir, basándose en un conjunto de variables predictivas (para una definición más formal ver sección Métodos). La información de salida puede ser cuantitativa (por ejemplo: un tamaño) o categórica (por ejemplo: sano/enfermo). A partir de una muestra de entrenamiento, esto es, una serie de ejemplos o casos de los que se conoce tanto el valor de salida como el valor de las variables predictivas, se entrena un modelo para que aprenda a predecir el valor de salida de nuevos casos de los que solo se conoce el valor de las variables predictivas.

Si las categorías (también llamadas etiquetas) a las que pertenece cada caso no están pre-establecidas, se pueden aplicar abordajes de aprendizaje no supervisado, que, mediante diversos enfoques, agrupan los casos en grupos, o *clusters*, que intentan minimizar alguna medida de heterogeneidad, y de esta manera, inferir la cantidad de categorías representadas en los datos de entrenamiento (Hastie, Tibshirani, y Friedman 2009; Li, Wu, y Ngom 2018).

Dada su utilidad para el análisis y la interpretación de grandes volúmenes de datos, el aprendizaje automático ha tenido un auge muy importante en los últimos años, coincidente con la

expansión acelerada del poder de cómputo y de la cantidad de datos disponibles en las más diversas áreas. Ilustrativo de este auge es la cantidad de revisiones publicadas durante los primeros 5 meses de este año acerca de la aplicación del aprendizaje automático en diversos campos. Estos trabajos abordan temas tan dispares como el diagnóstico médico y la planificación de políticas sanitarias (Schwalbe y Wahl 2020), la agricultura (Paul et al. 2020), las políticas educativas (Alenezi y Faisal 2020), el urbanismo (Ullah et al. 2020), el reconocimiento facial (Fathima y Vaidehi 2020), la gestión de transporte de carga (Barua, Zou, y Zhou 2020), la detección de correos electrónicos no deseados (Gangavarapu, Jaidhar, y Chanduka 2020), o la conducción de vehículos (Elamrani Abou Elassad et al. 2020).

Esta verdadera oleada de aplicaciones del aprendizaje automático incluye a las ciencias naturales. Según PubMed [1], durante los primeros 5 meses de este año se publicaron 398 revisiones que abordan su uso en distintos problemas de física, química y biología. Cada uno de estos trabajos revisa abordajes que incluyen aprendizaje automático en temas como la física de partículas (Larkoski, Moult, y Nachman 2020), la física de materiales (Liu et al. 2020) la mecánica de fluidos (Brunton, Noack, y Koumoutsakos 2020), las simulaciones moleculares (Noé et al. 2020), el plegamiento y dinámica de proteínas (Noé, De Fabritiis, y Clementi 2020), el diseño de anticuerpos (Graves et al. 2020), la oncología básica y clínica (Shimizu y Nakayama 2020), el estudio de enfermedades autoinmunes (Stafford et al. 2020) o el estudio de los arrecifes de coral (Raphael et al. 2020).

La genética y la genómica tampoco han sido ajenas a este fenómeno y existen numerosos problemas propios de estas disciplinas en los que se ha recurrido al aprendizaje automático (revisados en (Libbrecht y Noble 2015; Nicholls et al. 2020; Nguyen y Wang 2020). Una de estas aplicaciones es la predicción de función de genes, o de los productos de su expresión. La predicción de función de genes mediante técnicas de aprendizaje automático es un área en la que existe profusa bibliografía, revisada en (Bernardes y Pedreira 2013; Radivojac et al. 2013a; Zhou et al. 2019; Zhao et al. 2020; Bonetta y Valentino 2020). Como se mencionó más arriba, según los pocos estudios comparativos disponibles acerca de los métodos computacionales de predicción de función de genes (Jiang et al. 2016; Zhou et al. 2019) los abordajes que logran los mejores desempeños son los que se basan en los principios del aprendizaje automático.

El aprendizaje profundo (Deep learning) es una clase particular dentro de los métodos de aprendizaje automático que se ha desarrollado en las últimas décadas y que, por diversos factores,

1 - PubMed es un motor de búsqueda de libre acceso que permite consultar los contenidos de alrededor de 4.800 revistas científicas.

en los últimos años comenzó a obtener resultados excelentes en problemas de las más diversas áreas (Webb 2018; Zou et al. 2019; Shrestha y Mahmood 2019). Uno de los aspectos clave del aprendizaje profundo es que las capas de representación no necesitan ser diseñadas por un experto, sino que son aprendidas automáticamente a partir de los datos. En algunos campos, particularmente en aquellos vinculados al procesamiento de imágenes, el aprendizaje profundo ha obtenido resultados sorprendentes. Originalmente este proyecto se proponía utilizar aprendizaje profundo para predecir funciones de genes.

Sin embargo, a pesar de que en años recientes ha habido un aumento en el uso de técnicas de aprendizaje profundo para la predicción de función de genes, los métodos que usan técnicas clásicas de aprendizaje automático siguen obteniendo mejores desempeños (Bonetta y Valentino 2020). Más aun, entre los diez mejores métodos para la predicción de función de genes según la última edición de las competencias CAFA se encuentran métodos que ni siquiera utilizan aprendizaje automático, pero no aparece ningún método que utilice aprendizaje profundo (Zhou et al. 2019). Es probable que este mal desempeño se deba a una de las principales limitaciones de los modelos basados en aprendizaje profundo: necesitan ser entrenados con cantidades enormes de ejemplos de cada una de las clases que se quiere predecir (Shrestha y Mahmood 2019). Al mismo tiempo, una de las principales dificultades inherentes a la predicción de función de genes, es que se suele disponer de muy pocos ejemplos de cada una de las clases que se quiere predecir (Zhao et al. 2020). Por esta razón, en este proyecto de doctorado decidimos prescindir del aprendizaje profundo.

Nuestro proyecto original proponía además hacer énfasis en el estudio de los patrones de distribución de grupos de genes funcionalmente relacionados. La razón es que, en organismos eucariotas, la ubicación de un gen en el genoma al que pertenece ha sido muy poco explorada como variable predictiva de sus funciones. Esto a pesar de que, como veremos en el siguiente apartado, muchos resultados indican que la ubicación de un gen no es independiente de su función.

Agrupamiento de genes funcionalmente relacionados

Los "operones" son grupos de genes funcionalmente relacionados y co-regulados, ubicados uno al lado del otro en el genoma. Los operones son muy comunes en organismos procariotas y raros en organismos eucariotas. El término “operón” fue acuñado hace 60 años por François Jacob y Jacques Monod, quienes descubrieron y caracterizaron el operón *lac* en *Escherichia coli (Jacob y Monod 1961)*. El operón *lac* confiere la habilidad de utilizar la lactosa como fuente de energía, cuando la misma se encuentra en el medio. Los genes que forman operones clásicos no solo tienen

funciones relacionadas y factores reguladores en común, sino que también son transcriptos como un único mARN policistrónico a partir de un solo promotor. Los operones más conservados suelen agrupar genes que codifican complejos proteicos, asegurando así proporciones óptimas entre sus componentes (Price, Arkin, y Alm 2006).

Los operones se han comportado como estructuras dinámicas que se forman y se desensamblan a lo largo de la evolución de las especies. En procariotas, el modo más probable por el cual se han originado los operones es a partir de la transferencia horizontal de material genético (Rocha 2008). Por otro lado, la pérdida de estas estructuras puede ocurrir por deleción de uno o múltiples genes, o por inserciones de genes no relacionados funcionalmente que dividan al operón (Osbourn y Field 2009).

Hasta fines del siglo XX se solía asumir que en los genomas eucariotas los genes estaban distribuidos al azar. Sin embargo, en ese entonces se comenzó a comprender que la ubicación de un gen en un genoma eucariota no es independiente de su función y que los agrupamientos locales de genes funcionalmente relacionados son mucho más frecuentes de lo que se pensaba (Hurst, Williams, y Pál 2002; Yanai, Mellor, y DeLisi 2002; Lee y Sonnhammer 2003). Al mismo tiempo, diversas formas de agrupamiento local de genes con expresión similar o coordinada también fueron documentadas en eucariotas, incluyendo especie de levaduras, hongos, insectos, vertebrados y plantas (Eisen et al. 1998; Niehrs y Pollet 1999; Cohen et al. 2000; Hurst, Pal, y Lercher 2004).

De aquí en más, para referirnos a los distintos tipos de grupos de genes funcionalmente relacionados que se encuentran agrupados de algún modo en el genoma, utilizaremos el término *cluster*. La existencia extendida filogenéticamente de estos *clusters* indica que los mismos probablemente confieran alguna ventaja selectiva y que existe algún mecanismo evolutivo que promueva su formación y/o mantenimiento. La ventaja más clara que podría representar el agrupamiento de los genes relacionados funcionalmente es la simplificación de su regulación común, puesto que su vecindad les permitiría compartir elementos regulatorios (Hurst, Pal, y Lercher 2004; Osbourn y Field 2009).

Los *clusters* de genes eucariotas, que van desde pequeños clusters de unos pocos genes a grandes clusters que abarcan cientos de ellos, difieren del operón procariota en numerosos aspectos (Hurst, Williams, y Pál 2002; Sémon y Duret 2006). En primer lugar, a pesar de que en eucariotas los genes en clusters también pueden estar sujetos a una regulación común, no son transcriptos como mARNs policistrónicos, tal como sucede en procariotas. En segundo lugar, distintos estudios han sugerido diversos mecanismos responsables de la formación de estos *clusters* de genes y a pesar de algunas excepciones en hongos, la transferencia horizontal de material genético no es el

mecanismo más común por el cual se han formado (Hurst, Pal, y Lercher 2004; Fukuoka, Inaoka, y Kohane 2004).

El mecanismo más simple que puede dar lugar a la formación de *clusters* es la duplicación de genes, lo cual da lugar a dos genes en tándem. Este tipo de *clusters* de genes homólogos es muy frecuente en organismos eucariotas (Lee y Sonnhammer 2003). Aun sin ser homólogos, los genes en un *cluster* también pueden pertenecer a una misma vía metabólica, donde cada gen codifica un producto que interviene en uno de los pasos enzimáticos de esa vía. Los productos de los genes en *clusters* también pueden formar redes de interacción, en las que las proteínas codificadas interaccionan formando proteínas multiméricas o sirviendo como ligandos y receptores en cascadas de señalización (Osbourn y Field 2009; Nützmann, Scazzocchio, y Osbourn 2018). El agrupamiento de genes que codifican complejos proteicos puede asegurar proporciones óptimas entre los distintos componentes y así un mejor funcionamiento del complejo (Rocha 2008; Price, Arkin, y Alm 2006).

Como ninguno de los *clusters* descubiertos hasta ahora está muy conservado entre diversos taxones, lo más probable parece ser que estos agrupamientos reflejen las particularidades de cada especie o grupo de especies. Un argumento a favor de esta hipótesis es que, en plantas, la mayoría de los metabolitos producidos por vías metabólicas de genes agrupados están implicados en la defensa del organismo, cuyos mecanismos pueden variar mucho aún entre especies muy cercanas (Sue, Nakamura, y Nomura 2011). Para algunos de estos *clusters* hay incluso evidencia de convergencia evolutiva, en la que el mismo grupo de genes ha terminado agrupado en distintas especies de manera independiente (Takos et al. 2011; Dutartre, Hilliou, y Feyereisen 2012).

Sea cual sea el mecanismo de formación de un *cluster*, debe haber una presión selectiva que lo mantenga. Esa presión selectiva podría ser ejercida por la expresión génica coordinada y se cree que es ese el mecanismo más común que explica el agrupamiento de genes. Se ha observado, en múltiples taxones, que los genes vecinos tienden a estar co-expresados más de lo esperable por azar (Purmann et al. 2007; Fukuoka, Inaoka, y Kohane 2004; Lercher, Blumenthal, y Hurst 2003; Blumenthal et al. 2002; Boutanaev et al. 2002; Williams y Bowles 2004; Dávila López, Martínez Guerra, y Samuelsson 2010), un efecto que a menudo es más pronunciado en pares de genes orientados bidireccionalmente, en los que el promotor está cerca de ambos genes (Wei et al. 2011; Williams y Bowles 2004; Dávila López, Martínez Guerra, y Samuelsson 2010; Cohen et al. 2000). A una escala mayor, los genes expresados en la mayoría de los tejidos (*house keeping genes*) y los genes con alta expresión, también tienden a agruparse localmente en *clusters* de decenas de genes (Caron et al. 2001; Versteeg et al. 2003; Lercher, Blumenthal, y Hurst 2003; Weber y Hurst 2011).

En la misma dirección, se ha establecido que el perfil de expresión de un gen no es independiente del perfil de expresión de sus genes vecinos (Ghanbarian y Hurst 2015). En todos los organismos eucariotas estudiados a la fecha se ha encontrado que genes con perfiles de expresión similar tienden a agruparse de distintas maneras en el genoma (Hurst, Williams, y Pál 2002; Lee y Sonnhammer 2003; Ghanbarian y Hurst 2015; Caron et al. 2001; Spellman y Rubin 2002; Sémon y Duret 2006; Michalak 2008). Generalmente se entiende que el perfil de expresión de un gen está vinculado a sus funciones (Cantera et al. 2014), por lo que la tendencia a agruparse de los genes con expresión similar se puede pensar como otro aspecto del mismo fenómeno: los genes funcionalmente relacionados tienden a agruparse.

Este fenómeno se ha confirmado con el advenimiento de nuevas tecnologías que permiten estimar la distancia en el espacio tridimensional a la que se encuentran los genes en el núcleo celular (Dekker et al. 2002; Lieberman-Aiden et al. 2009). Los genes funcionalmente relacionados también tienden a agruparse en el espacio tridimensional (Tuller et al. 2009b; Diament, Pinter, y Tuller 2014; Karathia et al. 2016; Thévenin et al. 2014).

En resumen, los resultados disponibles muestran que la ubicación de un gen en el genoma al que pertenece no es azarosa y que los clusters de genes funcionalmente relacionados son inherentes a los genomas eucariotas y sus programas de regulación. Por otro lado, se observan grandes discrepancias en cuanto al tamaño y la ubicación de estos clusters, incluso en estudios de la misma especie.

Objetivos

Objetivos generales

- Avanzar en el campo de la predicción de función de genes mediante aprendizaje automático.
- Mejorar la comprensión de los patrones de distribución de grupos funcionales de genes en cinco genomas eucariotas.
- Investigar, en cinco genomas eucariotas, el potencial que tiene la ubicación de un gen para predecir su función.

Objetivos específicos

1 - Mejorar un modelo que predice función sináptica en genes de *Drosophila melanogaster*.
2 - Desarrollar una herramienta que permita analizar estadísticamente los patrones de distribución de grupos funcionales de genes.
3 - Buscar grupos de genes con la misma función y con patrones de distribución similar en distintos organismos.
4 - Buscar, en un mismo organismo, grupos de genes con funciones similares y patrones de distribución similar.
5 - Crear mapas genómicos de enriquecimiento funcional local.
6 - Implementar modelos de aprendizaje automático que predigan funciones de genes a partir de variables derivadas de su ubicación.

Capítulo I

Predicción de función sináptica en genes de *Drosophila melanogaster*

1.1 - Antecedentes.

1.2 - Evaluación de nuestra predicción original y propuestas para mejorarla.

- Nuevos genes sinápticos.
- Flujo de trabajo para obtener un nuevo modelo mejorado.
- Nuevo esquema de entrenamiento.
- Evaluación de los nuevos clasificadores.

1.3 - Un catálogo mejorado de genes potencialmente sinápticos.

1.4 - Discusión.

1.1 - Antecedentes

En todos los organismos con sistema nervioso las neuronas se comunican entre sí medainte una estructura altamente especializada llamada sinapsis neuronal. La sinapsis es fundamental para nuestro entendimiento acerca del aprendizaje, la memoria y otras funciones cerebrales. El ensamblaje y el funcionamiento de la sinapsis neuronal requieren de la expresión coordinada de una cantidad aún no conocida de genes que, por simplicidad, de ahora en adelante llamaremos “genes sinápticos”. Existe amplio consenso respecto a que sólo se ha identificado una pequeña fracción del total de genes sinápticos (Frank et al. 2013; Laßek, Weingarten, y Volknandt 2015). Debido a la gran conservación evolutiva que se ha observado entre los genes sinápticos conocidos, el conocimiento obtenido al estudiar *Drosophila*, así como otros organismos modelo, es muy relevante para otras especies, incluyendo seres humanos (Laßek, Weingarten, y Volknandt 2015; Burkhardt 2015).

En un trabajo publicado en el año 2015 y antecedente directo de esta tesis (Pazos Obregón et al. 2015) nos introdujimos en el área de la predicción de función de genes implementando un modelo *ensemble* de aprendizaje automático que asignó una probabilidad de ser un “gen sináptico” a cada gen codificante de proteína (GCP) de *Drosophila melanogaster*. Las características a partir de las cuales nuestro modelo infirió esas probabilidades fueron los niveles de transcripción de todos los GCP en 24 momentos del desarrollo, datos que fuesen publicados por el proyecto ModENCODE

algunos años antes (Graveley et al. 2011). Hasta donde sabemos, ese sigue siendo el único estudio publicado que predice función de genes basándose exclusivamente en perfiles temporales de transcripción obtenidos mediante tecnologías de secuenciación de nueva generación.

Para entrenar ese primer modelo definimos una muestra de entrenamiento compuesta de ejemplos positivos (genes sinápticos conocidos) y ejemplos negativos (genes que no son importantes para la sinapsis, seleccionados mediante criterios que se detallan en la sección Métodos). Con esa muestra entrenamos tres algoritmos de aprendizaje automático: kNN (Altman 1992), Random Forests (Breiman 2001) y SVM (Vapnik 2000), cuyos resultados fueron intersecados. Esos algoritmos fueron seleccionados luego de haber obtenido, durante estudios preliminares, resultados similares con estos y otros algoritmos, por ser ampliamente usados y por estar entre los algoritmos que obtienen, en promedio, los mejores resultados al ser aplicados a datos biológicos (Caruana y Niculescu-Mizil 2006; Fernández-Delgado et al. 2014).

Cuando se entrenan algoritmos de aprendizaje para llevar a cabo tareas de clasificación binaria, como en este trabajo, los mismos adjudican a cada caso de clase desconocida, una probabilidad de pertenecer a la clase positiva, que aquí es la clase “genes sinápticos”. El umbral de clasificación por defecto es 0.5, lo cual implica que todo nuevo caso al cual el modelo asigne una probabilidad mayor a 0.5 será clasificado como perteneciente a la clase positiva. Lo habitual es fijar el umbral de clasificación maximizando alguna medida de performance del modelo, para lo cual se utiliza una muestra de prueba, la cual es un conjunto de ejemplos de clase conocida pero que no fueron usados para entrenar el modelo y fueron justamente reservados para luego poder evaluarlo. En nuestro trabajo, sin embargo, fijamos el umbral de clasificación de manera tal que el número de genes que recibiesen la clasificación “sináptico” fuese similar al número de genes sinápticos que en ese entonces se estimaba que faltaban por descubrirse. Mediante este procedimiento obtuvimos un catálogo de 988 genes que postulamos estaba muy enriquecido en genes para los cuales en el futuro podría descubrirse una función sináptica (Pazos Obregón et al. 2015).

1.2 - Evaluación de nuestra predicción original y propuestas para mejorarla

Nuevos genes sinápticos

Desde la publicación de nuestro catálogo original en agosto de 2015 hasta principios del año 2019, científicos de distintos laboratorios alrededor del mundo identificaron, mediante diversas aproximaciones experimentales, 79 genes que se ajustan a la definición que usamos para elaborar

nuestra primera lista de genes sinápticos en 2015 y que de ahora en adelante llamaremos “nuevos genes sinápticos” (NGS). La lista de los 79 NGS junto a las referencias que aportan las evidencias experimentales que justifican la asignación de función sináptica a cada uno de ellos se puede encontrar en: https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-019-6380-z#Sec15

Decidimos usar esta lista de 79 NGS para evaluar empíricamente la calidad del catálogo de genes potencialmente sinápticos que habíamos publicado en 2015. Observamos que algo más de un tercio de los 79 NGS (28 genes) forman parte de ese catálogo. Dicho de otro modo, uno de cada tres genes a los cuales se les descubrió una función sináptica luego de la publicación de nuestro catálogo original estaba incluido en el mismo.

Para determinar si esta proporción es estadísticamente significativa efectuamos un análisis de enriquecimiento (Boyle et al. 2004). El análisis de enriquecimiento, cuyos detalles se encuentran en la sección Métodos, se suele utilizar para evaluar la sobrerrepresentación de cierta característica en una lista de genes. En este caso la característica cuya eventual sobrerrepresentación evaluamos es la de ser un NGS. Encontramos que nuestro catálogo original presenta un enriquecimiento en NGS de 4.38, con un p valor asociado $< 10^{-10}$.

	Modelo original	Modelo mejorado	Nuevo modelo
Muestra de entrenamiento	original	original	actualizada
Esquema de entrenamiento	original	nuevo	nuevo
Umbral de clasificación	0.9	0.9	0.95
Genes sobre el umbral	988	192	601
Enriquecimiento en NGS	4.38	6.07	-
P-valor	4E-11	1E-04	-

Tabla 1.1 – Comparación de los distintos modelos. *Comparación de los resultados obtenidos con el modelo original publicado en 2015 (columna 1), con un modelo mejorado, que fue entrenado con la muestra original, pero implementando el nuevo esquema de entrenamiento (columna 2) y con un modelo nuevo, que fue entrenado con el nuevo esquema de entrenamiento y con la muestra de entrenamiento actualizada (columna 3). El enriquecimiento en NGS encontrado en el catálogo obtenido con el nuevo esquema de entrenamiento (columna 2) es un 38% más alto que el encontrado en el catálogo original (columna 1), a pesar de que ambos modelos fueron*

entrenados exactamente con los mismos genes. La muestra de entrenamiento que se usó para entrenar al nuevo modelo (columna 3) incluye a los 79 NGS, por lo que el enriquecimiento en NGS del catálogo resultante no se puede calcular.

Una manera de interpretar los resultados en la Tabla 1.1 es la siguiente: imaginemos que se comenzara a intentar dilucidar experimentalmente el eventual carácter sináptico de todos los genes del genoma, uno tras otro. Es de esperar que pasado cierto tiempo se habría descubierto cierta cantidad de nuevos genes sinápticos. Si nos restringiésemos a testear exclusivamente los genes de nuestro catálogo, nos llevaría menos de la cuarta parte del tiempo descubrir esa misma cantidad de genes sinápticos. Este resultado nos animó a intentar mejorar nuestro modelo en aras de obtener una mejor predicción.

Flujo de trabajo para obtener un nuevo modelo mejorado

Tras evaluar nuestra predicción original nos propusimos utilizar los 79 NGS para obtener una nueva predicción, mejorada. La Figura 1.1 esquematiza el flujo de trabajo que diseñamos con ese fin. En primer lugar, propusimos implementar algunas mejoras al modelo de aprendizaje modificando la estrategia de entrenamiento original. Los 79 NGS nos permitirían evaluar esa nueva estrategia de entrenamiento, constatando si los catálogos resultantes tras su implementación estaban o no más enriquecidos en NGS. Si la nueva estrategia de entrenamiento efectivamente mejoraba los resultados obtenidos, el paso siguiente sería incorporar a la muestra de entrenamiento los 79 NGS. Con una mejor estrategia de entrenamiento validada y con una nueva muestra de entrenamiento ampliada, obtendríamos un nuevo modelo que, cabría esperar, tendría mayor poder predictivo.

Nueva estrategia de entrenamiento

Las mejoras a la estrategia de entrenamiento original se esquematizan en la Figura 1.2 y se detallan en la sección Métodos. Brevemente, se sub-muestreó 5 veces la muestra de entrenamiento original, dejando fuera cada vez una quinta parte diferente de los ejemplos positivos y de los ejemplos negativos. Mediante este procedimiento se obtuvieron 5 muestras de entrenamiento más pequeñas que la original y parcialmente diferentes entre sí. Utilizando los ejemplos positivos y negativos dejados fuera en cada sub-muestreo, se crearon 5 muestras de prueba complementarias a cada una de las 5 muestras de entrenamiento. Con cada muestra de entrenamiento se entrenó a tres

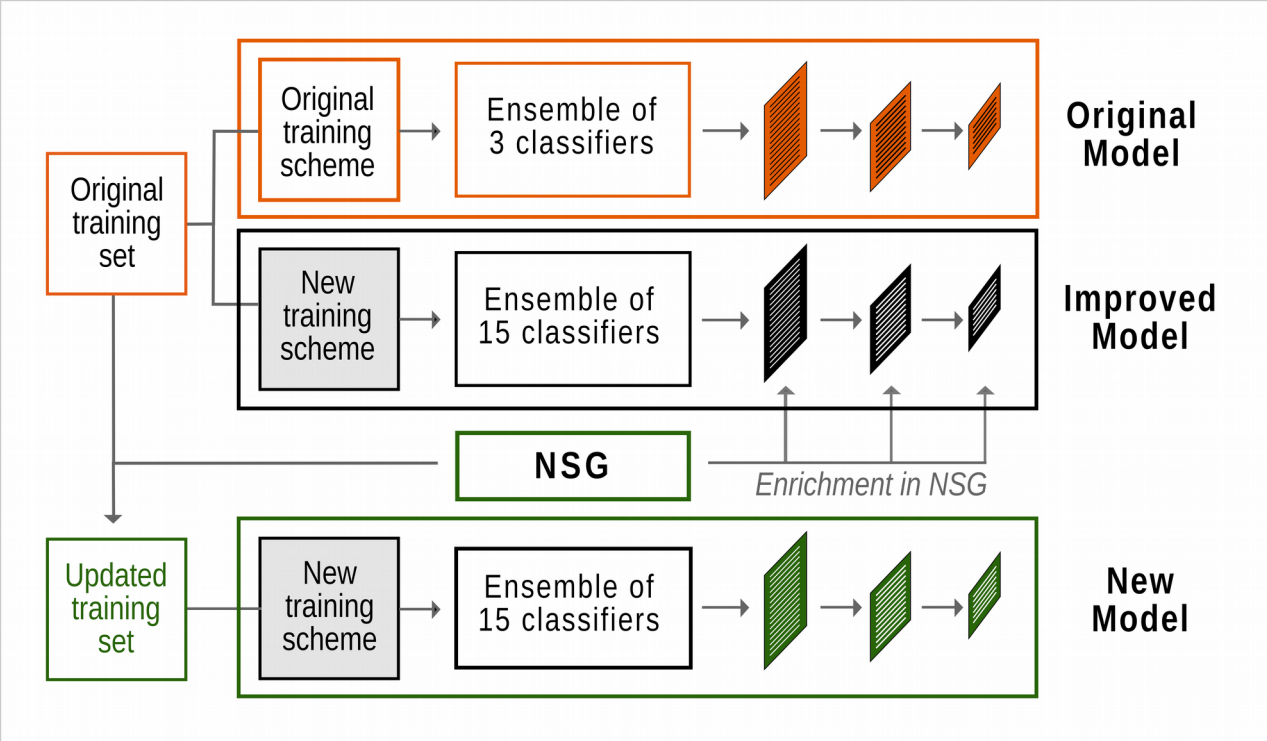


Figura 1.1 – Flujo de trabajo para obtener un catálogo mejorado de genes potencialmente sinápticos. *Utilizando la muestra de entrenamiento original entrenamos dos modelos, uno mediante la estrategia de entrenamiento original y otro mediante la nueva estrategia de entrenamiento (ver Fig. 1.2). Luego comparamos el enriquecimiento en NGS (ver Métodos) en los catálogos resultantes de cada modelo. Tras comprobar que la nueva estrategia de entrenamiento da lugar a un catálogo más enriquecido en NGS (Fig. 1.3), incorporamos los 79 NGS a la muestra de entrenamiento. Con esta nueva muestra de entrenamiento ampliada entrenamos un nuevo modelo mediante la nueva estrategia de entrenamiento, obteniendo un catálogo mejorado de genes potencialmente sinápticos (tomada de Pazos Obregón et al., 2019).*

algoritmos: kNN, SVM y Random Forest, obteniendo 15 clasificadores. Los hiperparámetros de cada clasificador se determinaron por búsqueda en grilla combinada con validación cruzada décuple sobre la propia muestra de entrenamiento. Cada muestra de prueba se usó para evaluar independientemente al clasificador que se entrenó con la muestra de entrenamiento complementaria, determinando su precisión, el área bajo su curva ROC y su F1. Finalmente se fue aumentando el umbral de clasificación de los clasificadores y se intersecaron los catálogos resultantes. Este nuevo esquema de entrenamiento busca aliviar un probable sesgo de nuestro modelo original, causado por el tamaño relativamente pequeño de la muestra de entrenamiento (Dietterich 2000).

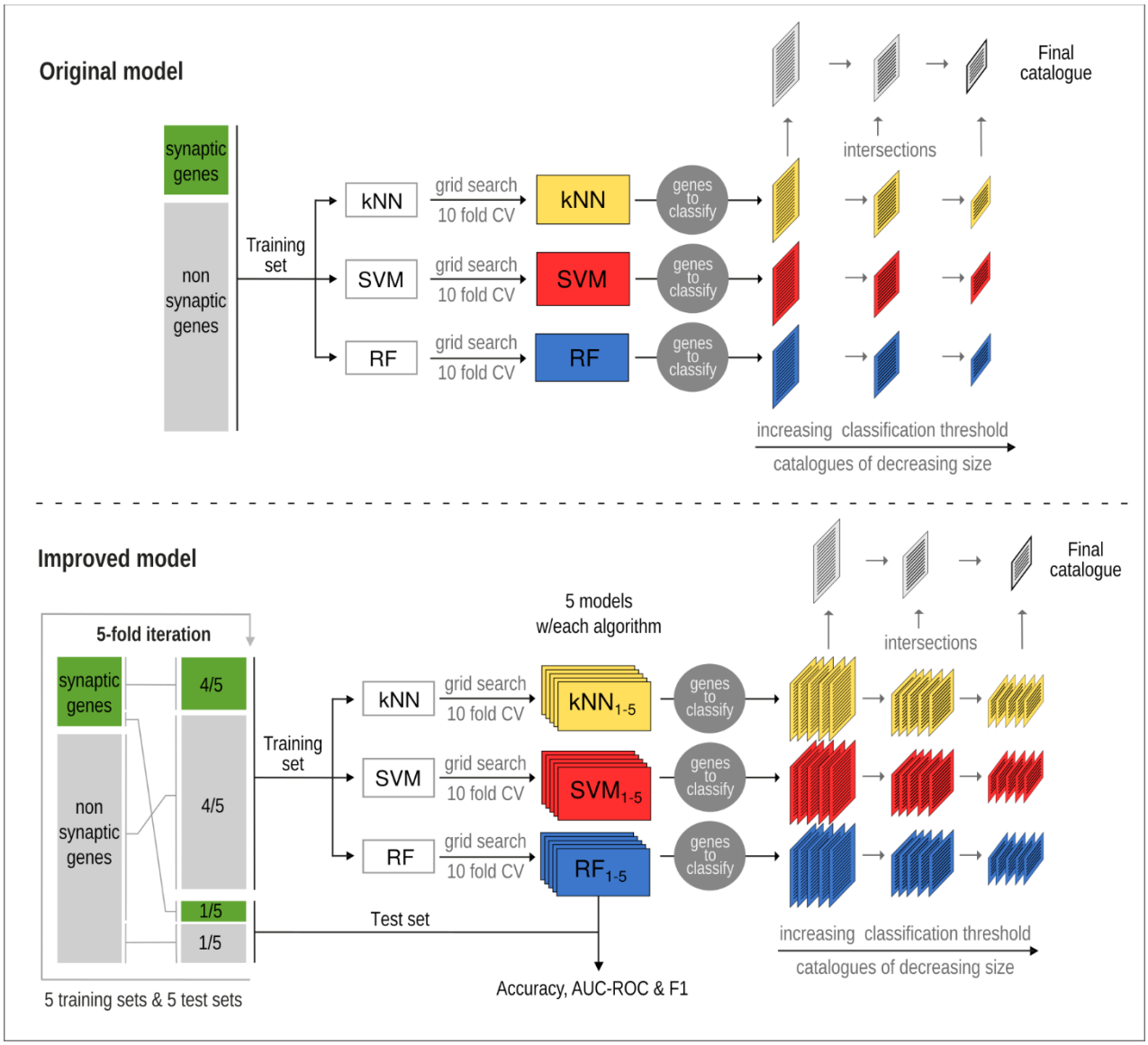


Figura 1.2 – Esquema de entrenamiento original y nuevo esquema de entrenamiento propuesto. *Arriba, esquema de entrenamiento original. Con la totalidad de la muestra de entrenamiento original entrenamos tres algoritmos: kNN, SVM y Random Forest. Los hiperparámetros de cada clasificador se determinaron por búsqueda en grilla combinada con validación cruzada décuple sobre la propia muestra de entrenamiento. El umbral de clasificación de los clasificadores se fue aumentando y los catálogos resultantes se intersecaron hasta obtener un catálogo final de un tamaño definido de antemano según estimaciones biológicas. Abajo, la estrategia de entrenamiento mejorada. En primer lugar, se sub-muestreó 5 veces la muestra de entrenamiento original, dejando fuera cada vez una quinta parte diferente de los ejemplos positivos y de los ejemplos negativos. Mediante este procedimiento se obtuvieron 5 muestras de*

entrenamiento más pequeñas que la original y parcialmente diferentes entre sí. Utilizando los ejemplos positivos y negativos dejados fuera en cada sub-muestreo se crearon 5 muestras de prueba complementarias de las 5 muestras de entrenamiento. Con cada muestra de entrenamiento se entrenó a tres algoritmos: kNN, SVM y Random Forest, obteniendo 15 clasificadores. Los hiperparámetros de cada clasificador se determinaron por búsqueda en grilla combinada con validación cruzada décuple sobre la propia muestra de entrenamiento. Cada muestra de prueba se usó para evaluar independientemente al clasificador que se entrenó con la muestra de entrenamiento complementaria, determinando su precisión, el área bajo su curva ROC y su F1. Finalmente se fue aumentando el umbral de clasificación de los clasificadores y se intersectaron los catálogos resultantes. Tomada de Pazos Obregón et al., 2019.

Para evaluar si la nueva estrategia de entrenamiento realmente mejoraba el poder predictivo del modelo resultante, la aplicamos para entrenar un modelo utilizando la muestra de entrenamiento original y comparando los resultados obtenidos con los del modelo original. Como muestra la Figura 1.3, los cambios propuestos resultaron en una mejora del poder predictivo, medido como enriquecimiento en NGS. Esta mejora se observa tanto cuando se considera cada algoritmo de clasificación por separado (Fig. 1.3A-C), así como cuando se considera la intersección de los tres algoritmos (Fig. 1.3D). Asimismo, para cada algoritmo, el enriquecimiento en NGS de la intersección de los cinco clasificadores entrenados con las cinco muestras de entrenamiento más pequeñas siempre es mayor que el enriquecimiento en NGS del clasificador que se obtiene entrenando una sola vez con la muestra de entrenamiento completa.

Evaluación de los nuevos clasificadores

Tras demostrar que si en 2015 hubiésemos aplicado el nuevo esquema de entrenamiento hubiésemos obtenido catálogos más enriquecidos en NGS, incorporamos estos últimos a la muestra de entrenamiento y repetimos todo el procedimiento descripto anteriormente, obteniendo 15 nuevos clasificadores. Evaluamos la capacidad predictiva de cada clasificador con la muestra de prueba complementaria a la muestra de entrenamiento con la que se lo entrenó (ver Fig. 1.2), con la cual calculamos la precisión, el *score* F1 y el área bajo la curva ROC (ver Métodos).

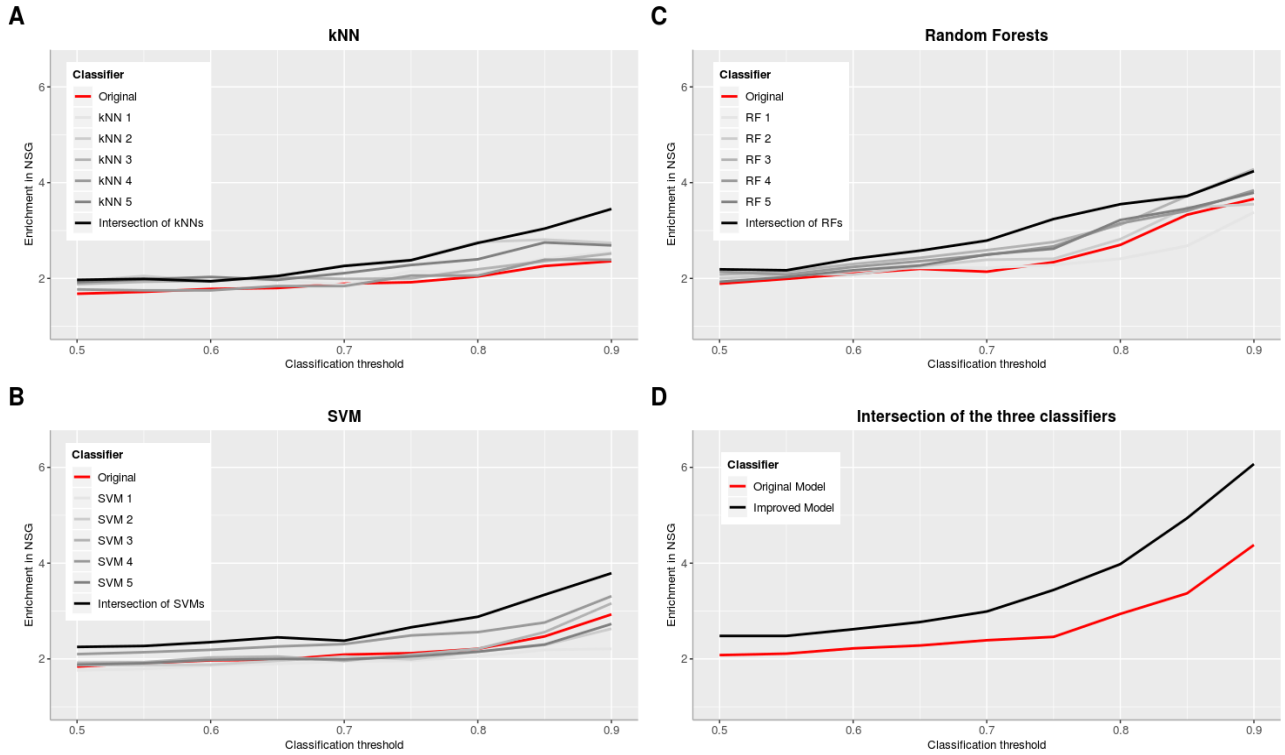


Figura 1.3 – Comparación entre los resultados del modelo original y del modelo mejorado. *Los resultados de ambos modelos se compararon determinando el enriquecimiento en NGS (ver Métodos) de los catálogos obtenidos por uno y otro modelo en función del umbral de clasificación. En A se muestran los resultados de los catálogos obtenidos por kNN, en B por SVM, en C por Random Forest y en D por la intersección de los tres algoritmos. En cada panel, la línea roja representa el enriquecimiento en NGS de los catálogos resultantes del modelo original, las líneas grises, de los catálogos resultantes de los cinco modelos obtenidos con cada sub-muestreo de la muestra de entrenamiento original y la línea negra, del catálogo resultante de intersectar estos últimos. Tomada de Pazos Obregón et al., 2019.*

La Tabla 1.2 recoge los resultados obtenidos tras esta evaluación, así como los valores reportados por otros colegas al evaluar modelos entrenados para predecir otras funciones biológicas (Kacsoh, Greene, y Bosco 2017; Kerepesi et al. 2018; Moore et al. 2019). Los trabajos con los que comparamos nuestros resultados fueron seleccionados por ser recientes, haber sido publicados en buenas revistas y ser similares metodológicamente en cuanto a intentar predecir funciones biológicas entendidas en sentido amplio.

		Accuracy	F1	AU ROC
kNN	Mean	0,93	0,88	0,97
	SD	0,03	0,05	0,02
SVM	Mean	0,93	0,89	0,97
	SD	0,02	0,04	0,01
RF	Mean	0,95	0,91	0,97
	SD	0,03	0,03	0,01
Kerepesi et al. 2018		-	-	0,93
Kacsoh et al. 2017		-	-	0,81
Moore et al. 2019		-	-	0,87

Tabla 1.2. Evaluación de los 15 clasificadores. *Con cada una de las 5 muestras de entrenamiento que resultaron de sub-muestrear la muestra de entrenamiento original se entrenó a 3 algoritmos de clasificación, obteniendo 15 clasificadores. Evaluamos la performance de cada uno de estos clasificadores usando una muestra de prueba conformada por genes que no habían sido usados en el entrenamiento del clasificador evaluado. La tabla muestra los promedios y desvíos estándar de la precisión, el score F1 y el área bajo la curva ROC de los cinco clasificadores entrenados por cada algoritmo. Las últimas 3 filas muestran el área bajo la curva obtenida por trabajos de otros colegas al predecir otras funciones biológicas mediante aprendizaje automático.*

1.3 - Un catálogo mejorado de genes potencialmente sinápticos

Finalmente, entrenamos un nuevo modelo incorporando los 79 NGS a la muestra de entrenamiento e implementando la nueva estrategia de entrenamiento. En nuestro trabajo original incluimos en el catálogo final solo aquellos genes que habían recibido una probabilidad de ser sinápticos mayor a 0,9 de parte de los tres algoritmos. Este umbral de clasificación se estableció para obtener un catálogo final de cierto tamaño prefijado (aproximadamente unos mil genes). Para el nuevo catálogo mejorado el umbral se fijó en 0,95 ya que ahora quedan menos genes sinápticos desconocidos y deseamos entonces obtener un catálogo más pequeño. El catálogo que resulta de interesecar todos los clasificadores con un umbral de 0,95 tiene 601 genes.

Para estimar la calidad del nuevo catálogo determinamos su enriquecimiento en genes asociados a términos GO relacionados a la sinapsis neuronal. Pudimos hacer esto porque construimos nuestra muestra de entrenamiento sin tener en cuenta las anotaciones de Gene Ontology de los genes incluidos. Encontramos que entre los 601 genes a los cuales los 15 clasificadores asignaron una probabilidad de ser sinápticos mayor a 0,95 había 83 genes que ya tenían alguna anotación en Gene Ontology que los relacionaba a la sinapsis.

Para determinar si esta cantidad refleja un enriquecimiento estadísticamente significativo debemos quitar de la lista de referencia (ver Métodos) todos los genes que estuviesen en la muestra de entrenamiento y que tuviesen alguna anotación GO relacionada a sinapsis. Los resultados del análisis, realizado con la herramienta Gorilla (Eden et al. 2009) se muestran en la Tabla 1.3. Luego de excluir del catálogo final los 83 genes que ya tenían alguna anotación relacionada a sinapsis, el nuevo catálogo tiene 518 genes y se puede encontrar aquí:
http://synapticgenes.bnd.edu.uy.

Nos proponemos re-entrenar nuestro modelo cada cierto tiempo, incorporando a la muestra de entrenamiento los NGS que se vayan descubriendo. Esto dará lugar a catálogos actualizados de genes potencialmente sinápticos que iremos subiendo en el sitio web mencionado.

Término GO	Descripción	p valor	FDR q valor	Enriquecimiento
GO:0050807	regulation of synapse organization	1.66E-12	2.03E-10	4.67
GO:0051963	regulation of synapse assembly	1.38E-10	1.27E-8	4.87
GO:0008582	regulation of synaptic growth at neuromuscular junction	2.21E-9	1.59E-7	4.65
GO:0016080	synaptic vesicle targeting	7.01E-4	1.25E-2	8.67

Tabla 1.3 – Enriquecimiento en términos GO relacionados a sinapsis encontrado en nuestro nuevo catálogo. *La primera y la segunda columna muestran el identificador del término GO enriquecido y su descripción. La tercer y cuarta columna muestran el p-valor asociado a ese enriquecimiento y su corrección por la tasa de falsos positivos. La última columna muestra el enriquecimiento funcional encontrado.*

1.4 - Discusión

La hipótesis subyacente a nuestro abordaje es que la transcripción de un gen a lo largo del tiempo es informativa respecto a la función que los productos de esa transcripción tienen en la biología del organismo (Cantera et al. 2014). Por ejemplo, si en ciertos tejidos al menos una de las funciones que tiene cierta proteína es importante para el ensamblaje de las sinapsis, debería ser más probable que en momentos de sinaptogénesis masiva aumente el nivel de transcripción del gen correspondiente. Igualmente, sería razonable esperar que en momentos de desensamblaje masivo de sinapsis, como ocurre en ciertas etapas del desarrollo de *Drosophila*, el nivel de transcripción de ese gen disminuya. Esta correlación positiva y lineal es más probable durante el desarrollo de un organismo como *Drosophila*, donde la rapidez del desarrollo del sistema nervioso no daría cabida a otros mecanismos de regulación más sofisticados.

Sin embargo, se conocen muchos factores y mecanismos biológicos que distorsionan la simpleza de esta correlación. En general los genes tienen más de una función y no siempre se podrá discernir cuál de esas funciones explica cierta subida o bajada de su nivel de transcripción. La correlación entre nivel de transcripción y disponibilidad de proteína no es siempre la misma entre distintos genes, ni entre distintas células, ni entre distintos estados fisiológicos. Más aun, hay genes que pueden ser importantes para cierto proceso biológico cuyo rol es reprimir a otros genes, por lo que su propio nivel de transcripción bajará cuando tenga lugar el proceso para el que es importante.

Por todos estos motivos, sería un error esperar que nuestro catálogo de genes, definidos exclusivamente a partir de un transcriptoma temporal de cuerpo completo, contuviese a todos los genes sinápticos por descubrir. Un catálogo como el nuestro, por fuerza tendría genes que no son sinápticos y dejaría fuera a otros que sí lo son. Pero sí hipotetizamos que nuestro catálogo estaría enriquecido en genes cuya importancia para la sinapsis aún estaba por descubrirse. El conjunto de genes cuya función sináptica fue descubierta luego de la publicación de nuestro primer catálogo nos permitió poner a prueba esa hipótesis. El enriquecimiento en nuevos genes sinápticos que encontramos en nuestro primer catálogo representa un claro soporte experimental para esa hipótesis.

También es interesante notar que ninguno de los 79 NGS había sido incluido en la lista que definimos *a priori* como "no sinápticos" para entrenar los algoritmos, lo cual constituye una validación inequívoca de los criterios usados para su selección.

Es de notar que aún cuando el modelo original y su versión mejorada se basan en los mismos tres algoritmos y que fueron entrenados con exactamente el mismo grupo de genes, el

enriquecimiento en NGS encontrado en el catálogo que resulta de la versión mejorada sea un 38% más alto (4,38 vs. 6,07, ver tabla 1.1). Interpretamos esto como una clara demostración de que el nuevo esquema de entrenamiento (única diferencia entre ambos modelos) realmente mejora el poder predictivo de nuestro abordaje. Una posible explicación de este efecto es que el nuevo esquema de entrenamiento compensaría un probable sesgo de nuestro modelo original, debido a una muestra de entrenamiento relativamente pequeña, incrementando su capacidad de generalización (Hastie, Tibshirani, y Friedman 2009). Dado que probablemente aún quedan cientos de nuevos genes sinápticos por descubrir, esta es una característica importante.

El principal resultado de este trabajo es un catálogo de genes que, según indican todos los análisis efectuados, está muy enriquecido en genes sinápticos aun por descubrirse. Este catálogo, que además será periódicamente actualizado, ha sido puesto a disposición de la comunidad científica (Pazos Obregón et al. 2019). Creemos que su uso facilitará la identificación de genes importantes para el ensamblaje y el funcionamiento de la sinapsis mediante distintos abordajes experimentales. Además, proponemos aquí un esquema de entrenamiento mediante el cual se obtuvieron mejores resultados partiendo de los mismos datos, lo cual puede significar un aporte importante en la predicción de otras funciones biológicas, sobre todo teniendo en cuenta que en la predicción funcional el habitual tamaño reducido de las muestras de entrenamiento es un problema crítico.

Una limitación importante de nuestro modelo es que se basa en una definición *ad hoc* de la función biológica que se predice, lo cual dificulta la comparación de nuestros resultados con los obtenidos por otros estudios que predicen función de genes.

Por otro lado, los resultados obtenidos nos permiten concluir que es posible predecir al menos ciertas funciones de genes basándose exclusivamente en perfiles temporales de transcripción. Con la creciente disponibilidad de transcriptomas de células individuales, abordajes de este tipo son promisorios.

Capítulo II

Patrones de distribución de grupos funcionales de genes

2.1 – Antecedentes.

2.2 - Cluster Locator, una herramienta para el análisis del agrupamiento de genes.

- Características generales.
- Abordaje estadístico.

2.3 - Patrones de distribución de grupos funcionales de genes en 5 organismos.

- Ontología Génica y perfiles de agrupamiento.
- Grupos de genes en distintos organismos con la misma función y con perfiles de agrupamiento similar.
- Grupos de genes con funciones similares y perfiles de agrupamiento similar.

2.4 - Discusión.

2.1 Antecedentes

En los genomas eucariotas los genes no están distribuidos al azar (Feuerborn y Cook 2015; Hurst, Pal, y Lercher 2004). Por el contrario, en todos los organismos eucariotas estudiados a la fecha se han documentado correlaciones entre la ubicación de los genes y su expresión, sus funciones u otros rasgos cuantitativos sujetos a presión evolutiva (De y Babu 2010; Ghanbarian y Hurst 2015). Este tipo de correlación se observó por primera vez en la levadura *Saccharomyces cerevisae* (Eisen et al. 1998) y luego en nematodos, moscas, ratones, humanos y otros organismos (Michalak 2008).

Numerosos estudios han encontrado *clusters* de genes que son co-expresados y que comparten función biológica, o *clusters* de genes funcionalmente relacionados que comparten vecindad en el genoma o *clusters* de genes cercanos con patrones de expresión o funciones similares (Lee y Sonnhammer 2003; G. Yi, Sze, y Thon 2007; Tuller et al. 2009b; Thévenin et al. 2014; Tiirikka, Siermala, y Vihinen 2014; Corrales-Berjano et al. 2017; Reimegård et al. 2017). Lamentablemente, todos estos estudios utilizan diferentes definiciones para el término "*cluster*", lo cual dificulta enormemente la comparación y sistematización de los resultados. Tampoco es

homogéneo el uso que se le da al concepto de “función de un gen” o de “grupo funcional de genes”.

En lo últimos años, gracias a los continuos avances en la anotación de los genomas y a la creciente disponibilidad de transcriptomas obtenidos en diversos organismos y bajo distintas condiciones, se ha vuelto relativamente sencillo elaborar listas de genes con la misma función biológica o con patrones de expresión similares. Con este tipo de datos se puede estudiar sistemáticamente la forma en que grupos de genes co-expresados o con la misma función se distribuyen en diferentes organismos.

En este contexto y en el marco de este proyecto de doctorado, en el cual nos propusimos investigar el potencial de la localización de un gen para inferir sus funciones biológicas, nos planteamos las siguientes preguntas: ¿Existen grupos de genes con la misma función que tienen patrones de distribución similar en distintos organismos? ¿En un organismo dado, los grupos de genes con funciones similares, tienen patrones de distribución similares?

Cuando comenzamos a estudiar la bibliografía buscando responder estas preguntas, encontramos que no existían herramientas que permitiesen comparar sistemáticamente el modo en que distintos grupos de genes se distribuyen a lo largo de los genomas. Si bien se habían desarrollado algunas herramientas que podían resultar de cierta ayuda (G. Yi, Sze, y Thon 2007; Yu et al. 2012; Aboukhalil, Fendler, y Atwal 2013; Dottorini et al. 2013), las mismas no habían sido desarrolladas específicamente con ese fin, solo podían manejar datos de algunos organismos y no estaban disponibles en línea ni luego de ser solicitadas a sus autores. Para cubrir esta carencia desarrollamos "Cluster Locator" (Pazos Obregón et al. 2018), junto a otros investigadores del Departamento de Biología del Neurodesarrollo del Instituto de Investigaciones Biológicas Clemente Estable y de la Plataforma de Análisis del Genoma y de la Unidad de Genómica Funcional del CicBiogune del País Vasco.

2.2 Cluster Locator, una herramienta para el análisis de clusters de genes

Características generales

Cluster Locator es una herramienta en línea de uso libre y gratuito que se encuentra disponible aquí: http://clusterlocator.bnd.edu.uy y que fue publicada como *Application Note* por la revista Bioinformatics (Pazos Obregón et al. 2018). Cluster Locator localiza, cuantifica y muestra los *clusters* formados por los genes de una lista provista por el usuario y lleva a cabo un análisis estadístico que permite estimar que tan alejada de lo esperable por azar se encuentra la cantidad de

genes que se encontró formando *clusters*. Para usar la herramienta el usuario solo debe proveer la lista de genes de interés y seleccionar el valor de un parámetro llamado “max-gap”, que se usará para definir los *clusters* (ver más adelante).

Para implementar Cluster Locator modelamos el genoma como una colección de segmentos independientes o unidades de replicación, que según el organismo y la versión del genoma que se utilice, pueden corresponder a los cromosomas, los brazos cromosómicos o los *scaffolds* (depende del genoma de referencia de cada análisis). En estos segmentos, los genes se consideran como ubicados uno al lado del otro, sin regiones intergénicas ni solapeo. Definimos la brecha entre dos genes, o *gap*, como la cantidad de genes que se encuentra entre ambos (Roy et al. 2002). Dada una lista de genes y cierto *gap* máximo permitido (*max-gap)*, un cluster se define como un conjunto maximal de genes de la lista tales que la brecha (*gap* en inglés) entre cualquier par sucesivo de esos genes nunca es mayor al *max-gap* fijado por el usuario. El tamaño del *cluster* es igual a la cantidad de genes que lo forman. Dada una lista de genes y un valor de *max-gap* determinado, el porcentaje de agrupamiento de la lista es el porcentaje de sus genes que quedan incluidos en algún *cluster*. Un *cluster* solo se puede formar por genes que pertenecen al mismo segmento. En la Figura 2.1 se ilustran estas definiciones y se dan algunos ejemplos.

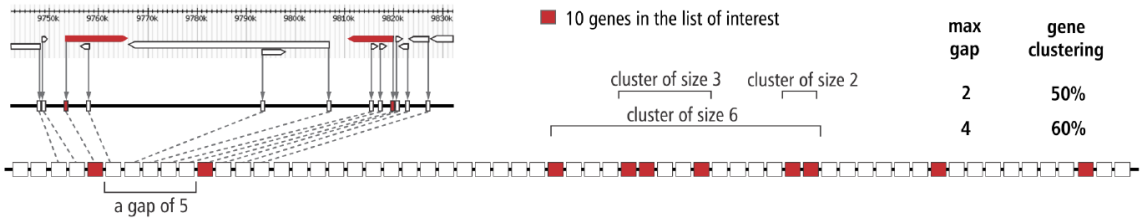


Figura 2.1 – Modelo de genoma usado en Cluster Locator. *Cluster Locator modela cada genoma como una colección de segmentos en los cuales los genes codificantes de proteínas están ubicados uno al lado del otro, sin espacios intergénicos o solapeo. La brecha o gap entre dos genes es el número de sitios de inicio de transcripción de otros genes que se encuentre entre ellos. Una vez fijado un max-gap, un cluster se define como un conjunto maximal de genes tales que el gap entre cualquier par adyacente de genes del grupo nunca supera el max-gap. Dado un max-gap, el porcentaje de agrupamiento de una lista de genes es el porcentaje de genes de la lista que forman parte de alguno de los clusters que quedan definidos. En el ejemplo de la figura, el porcentaje de agrupamiento de la lista de genes de interés, mostrados en rojo, cambia de 50% a 60% al cambiar el max-gap de 2 a 4.*

Cluster Locator puede analizar listas de genes de cualquier organismo. Las listas pueden tener un máximo de 1.000 genes y pueden ser provistas como archivos de texto (*txt o *csv) o ser directamente pegadas en un cuadro de entrada de texto que ofrece la herramienta. En su versión actual, Cluster Locator tiene precargados los genomas de 5 organismos: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster, Caenorhabditis elegans* y *Saccharomyces cerevisiae.* Estos cinco organismos fueron seleccionados porque están entre los más estudiados, tienen genomas relativamente bien anotados, se distribuyen a lo ancho de la filogenia eucariota y son diversos en términos de número y tipo de cromosomas.

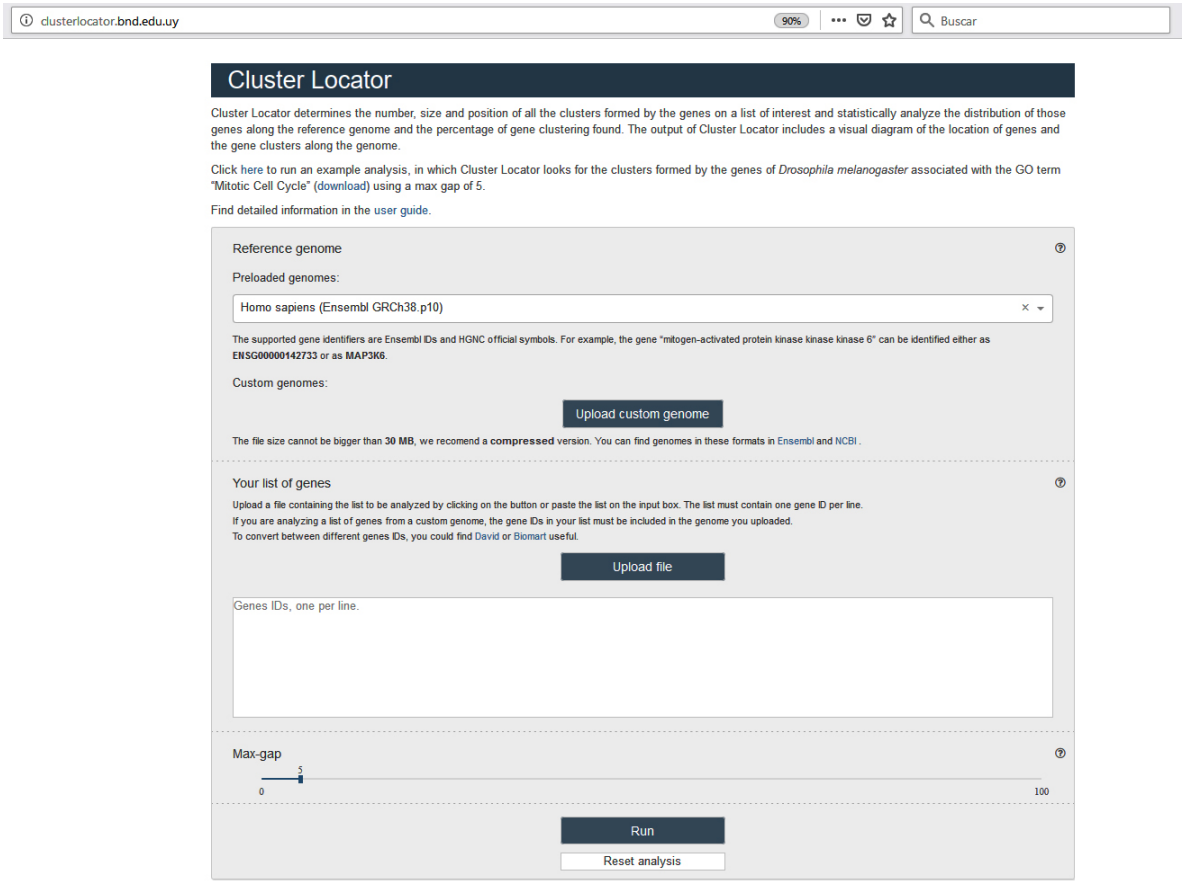
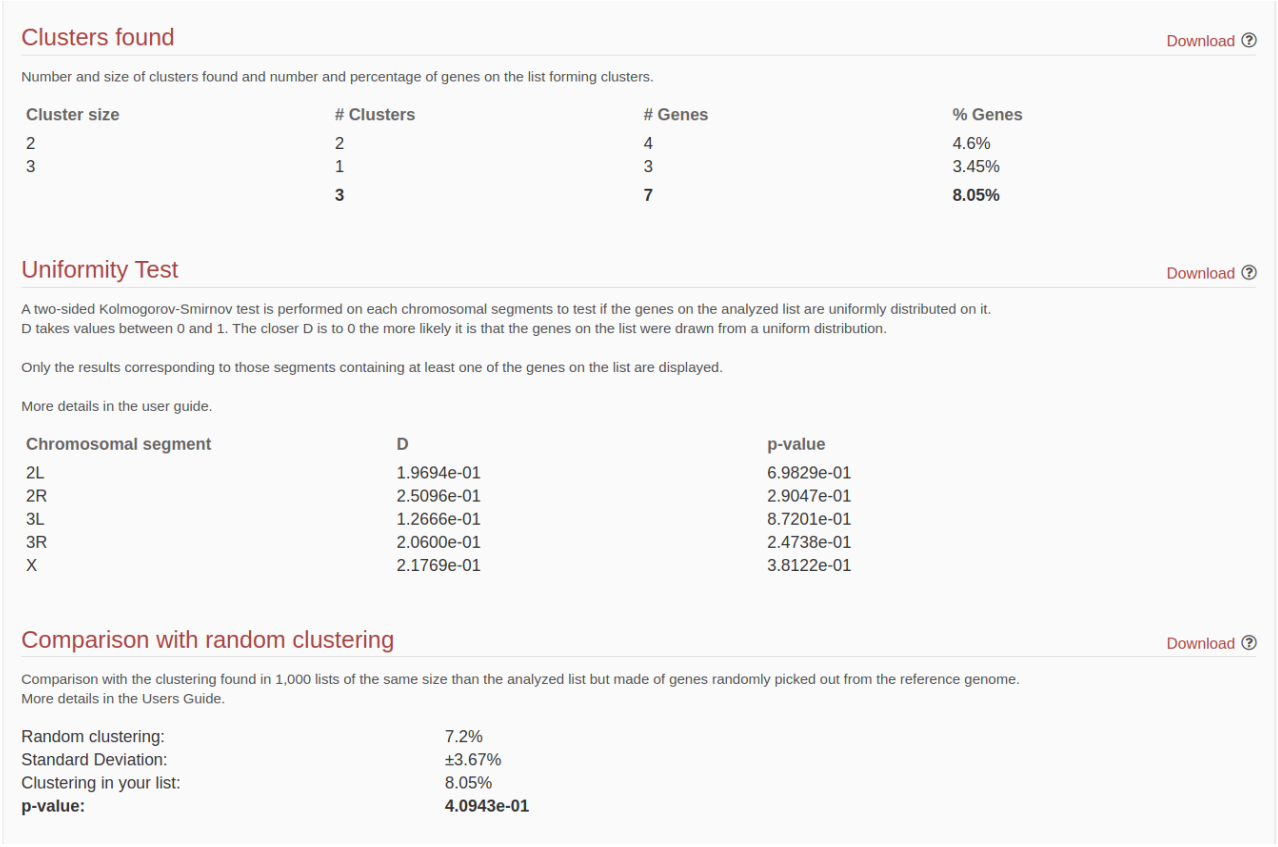


Figura 2.2 – Interfaz de Cluster Locator*. 1: Menú desplegable para seleccionar el genoma al que pertenece la lista de genes que se desea analizar, si se trata de uno de los 5 organismos precargados. 2: botón para subir un archivo con cualquier otro genoma de referencia. 3: botón para subir la lista de genes que se desea analizar, la cual también se puede pegar en 4. 5: selector de max-gap. 6: botón para iniciar el análisis.*

La herramienta puede manejar al menos dos tipos de identificadores de genes para cada uno de estos organismos y tiene en cuenta la posición de los centrómeros para definir los segmentos a lo largo de los cuales se pueden formar *clusters*. Si la lista de genes no pertenece a ninguno de los organismos precargados, el usuario puede igualmente analizarla, pero debe proveer un archivo en formato gff, gff3 o gtf con el genoma de referencia, además de la lista que desea analizar. En este caso, el análisis no considera los centrómeros para modelar las unidades de replicación y en la lista a analizar se deben utilizar los mismos identificadores de genes que en el archivo con el genoma de referencia.

La interfaz para el usuario es muy sencilla, amigable e intuitiva y no es necesario ningún conocimiento especializado previo para utilizarla (ver Figura 2.2). En primer lugar, un menú desplegable permite elegir el genoma al cual pertenecen los genes de la lista que se desea analizar (ver Figura). Si dicho genoma no es uno de los 5 genomas precargados el usuario puede cargarlo en alguno de los formatos de archivos aceptados, usando para ello la última opción del mismo menú desplegable. A continuación se debe subir la lista de genes que se desea analizar, ya sea como un archivo de texto o directamente pegándola en el campo para ingreso de texto correspondiente. Finalmente el usuario selecciona el *max-gap* que se usará para definir los *clusters* y cliquea en el botón "*run*".

El algoritmo de análisis de agrupamientos implementado en Cluster Locator requiere unos pocos segundos para generar los resultados (ver Figura 2.3), que incluyen el número, tamaño y posición de los *clusters* encontrados, la identidad y posición de los genes de cada cluster, el porcentaje de agrupamiento de la lista, así como el resultado de los análisis estadísticos. Debajo de estos resultados, que se pueden descargar como archivo *csv, se despliega una representación visual esquemática del modo en que los genes y los *clusters* de la lista analizada están distribuidos a lo largo del genoma de referencia. Por más detalles acerca de las características, la interfaz y el uso de Cluster Locator remitimos al Manual de Usuario en los anexos.



Clusters found

Download

Number and size of clusters found and number and percentage of genes on the list forming clusters.

Cluster size	# Clusters	# Genes	% Genes
2	2	4	4.6%
3	1	3	3.45%
	3	**7**	**8.05%**

Uniformity Test

Download

A two-sided Kolmogorov-Smirnov test is performed on each chromosomal segments to test if the genes on the analyzed list are uniformly distributed on it.
D takes values between 0 and 1. The closer D is to 0 the more likely it is that the genes on the list were drawn from a uniform distribution.

Only the results corresponding to those segments containing at least one of the genes on the list are displayed.

More details in the user guide.

Chromosomal segment	D	p-value
2L	1.9694e-01	6.9829e-01
2R	2.5096e-01	2.9047e-01
3L	1.2666e-01	8.7201e-01
3R	2.0600e-01	2.4738e-01
X	2.1769e-01	3.8122e-01

Comparison with random clustering

Download

Comparison with the clustering found in 1,000 lists of the same size than the analyzed list but made of genes randomly picked out from the reference genome.
More details in the Users Guide.

Random clustering:	7.2%
Standard Deviation:	±3.67%
Clustering in your list:	8.05%
p-value:	**4.0943e-01**

Figura 2.3 A *. En el panel superior se muestra la cantidad de clusters de cada tamaño encontrados en la lista analizada, así como la cantidad de genes en cada tipo de cluster. El panel intermedio muestra los resultados del Test de Uniformidad aplicado a la distribución de la lista analizada a lo largo de cada unidad de replicación o scaffold del genoma de referencia. El panel inferior muestra la comparación del porcentaje de agrupamiento encontrado en la lista analizada con el porcentaje de agrupamiento promedio encontrado en 1.000 listas de genes del mismo tamaño que la lista analizada pero formadas por genes seleccionados al azar.*

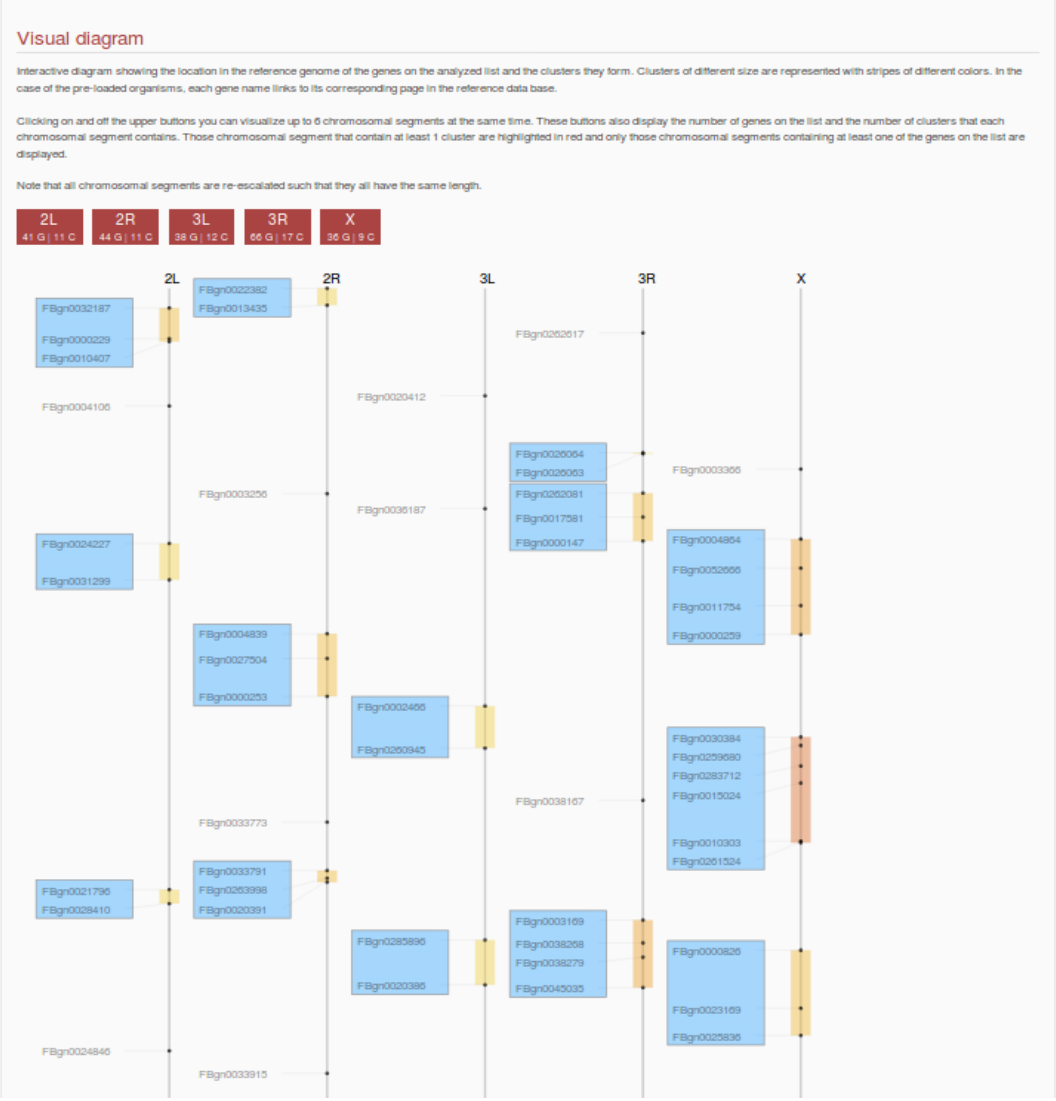


Figura 2.3 B. *Representación visual de los genes de la lista analizada y de los clusters encontrados. Cada unidad de replicación es representada por una línea recta vertical (todas del mismo largo), en las que los genes de la lista analizada aparecen representados como puntos. Una línea conecta cada punto a una etiqueta con el nombre del gen, que, en el caso de los 5 organismos precargados, enlaza a la base de datos de referencia correspondiente (Ensembl, FlyBase, WormBase o SGD). Los genes que forman parte de un cluster están rodeados por una caja cuyo color depende de la cantidad de genes en el cluster. Según el organismo, puede que no todas las unidades de replicación aparezcan en la pantalla, en la que se muestra un máximo de 6. Las unidades de replicación que se muestran pueden elegirse con los botones que aparecen en la parte superior del esquema.*

Abordaje estadístico

Cluster Locator informa dos p-valores luego de cada análisis. Uno de esos p-valores se obtiene tras realizar el test de Kolmogorov-Smirnov, cuya hipótesis nula es que los genes en la lista analizada siguen una distribución uniforme a lo largo de cada unidad de replicación. Nótese que este p-valor será siempre el mismo dada cierta lista de genes y no depende del *max-gap* utilizado para el análisis. La inclusión de este test, que entendemos no tiene ninguna utilidad en este contexto, se debe exclusivamente a una exigencia de uno de los revisores.

El otro p-valor está asociado al porcentaje de agrupamiento encontrado en la lista analizada, que a su vez depende del *max-gap* utilizado en el análisis. Este p-valor es la probabilidad de encontrar un porcentaje de agrupamiento igual o más extremo si los genes de la lista analizada estuviesen distribuidos uniformemente a lo largo del genoma. Para estimar esta probabilidad seguimos el abordaje empírico utilizado por Roy y colaboradores (Roy et al. 2002). En primer lugar Cluster Locator genera 1.000 listas de genes del mismo tamaño que la lista analizada pero formadas por genes seleccionados al azar del mismo genoma de referencia. Luego, usando el max-gap seleccionado por el usuario, Cluster Locator determina el porcentaje de agrupamiento en cada una de esas listas “random” y determina el promedio y el desvío estándar empíricos (tras confirmar que la distribución de los valores obtenidos es normal usando el test de Kolmogorv-Smirnov). Si el porcentaje de agrupamiento encontrado por azar tiene una distribución normal y contando con su promedio y su desvío estándar empíricos, Cluster Locator calcula el p-valor asociado al porcentaje de agrupamiento encontrado en la lista analizada y lo reporta (ver Figura 2.4). Nótese que incluso cuando no se puede rechazar la hipótesis nula de que los genes de la lista analizada están uniformemente distribuidos a lo largo del genoma de referencia, es posible encontrar un porcentaje de agrupamiento significativamente mayor al esperado por azar.

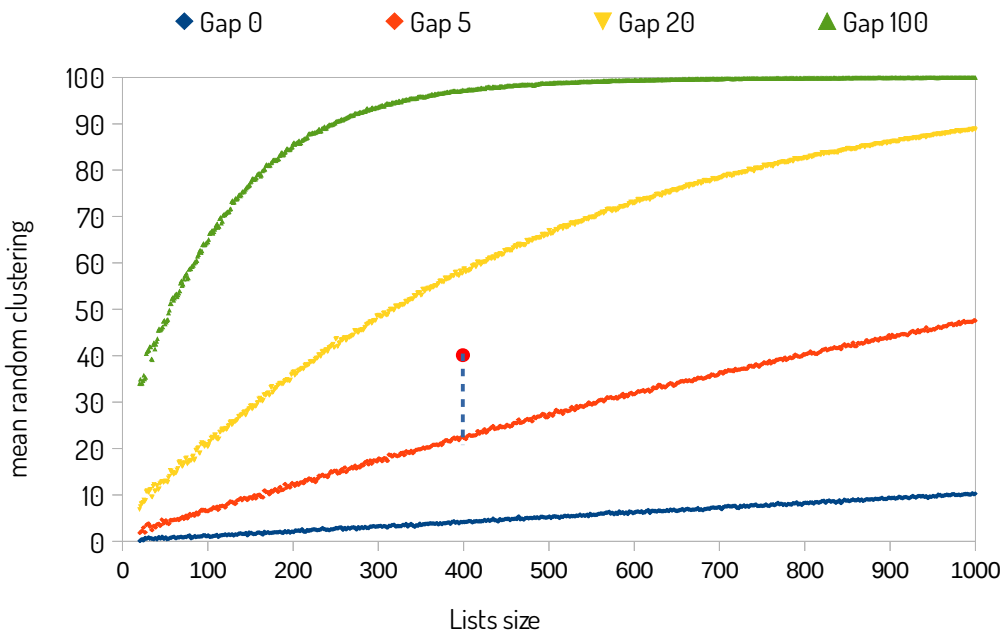


Figura 2.4 – Porcentaje de agrupamiento esperado por azar. *La figura muestra el porcentaje de agrupamiento encontrado en listas de genes seleccionados al azar del genoma humano. En el eje x el tamaño de las listas, en el eje y el porcentaje de agrupamiento encontrado. Se muestran resultados obtenidos con cuatro valores diferentes de max-gap. Dado cierto tamaño de lista, el porcentaje de agrupamiento encontrado en listas al azar tiene una distribución normal, lo cual permite evaluar estadísticamente el porcentaje encontrado en una lista de interés. En la figura se ejemplifica con un círculo rojo el porcentaje de agrupamiento encontrado en una lista de 400 genes correspondientes a cierto grupo funcional. La distancia de ese porcentaje al porcentaje encontrado en listas al azar se puede evaluar estadísticamente comparándola con el desvío estándar del agrupamiento al azar.*

Luego de desarrollar Cluster Locator y hacerlo disponible para su uso libre por la comunidad, le dimos un uso en particular: la caracterización de los patrones de distribución de cientos de grupos funcionales en *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans* y *Saccharomyces cerevisiae*.

2.3 Patrones de distribución de grupos funcionales de genes en 5 organismos

Ontología Génica y perfiles de agrupamiento

Al inicio de este capítulo nos planteamos dos preguntas: ¿Los grupos de genes con la misma función en distintos organismos tienen patrones de distribución similares? Y en un mismo organismo, ¿los grupos de genes con funciones similares tienen patrones de distribución similares? Lo que resta de este capítulo recoge resultados obtenidos en busca de respuestas a estas preguntas.

Decidimos abordar estas cuestiones utilizando Gene Ontology (Ashburner et al. 2000) y Cluster Locator. Gene Ontology (GO) es una base de datos que representa la mayor fuente de información disponible sobre las funciones de los genes conocidos. Se trata en realidad de tres ontologías, compuestas por un conjunto de términos con definición precisa organizados en tres grafos acíclicos direccionados llamados "Procesos biológicos" (PB), "Componente Celular" (CC) y "Función Molecular" (FM). Estas tres ontologías reflejan tres aspectos necesarios para caracterizar cualquier posible función de un gen: el mecanismo molecular involucrado, el proceso biológico del que forma parte y la localización celular en la que tiene lugar (ver Figura 2.5). Se puede encontrar una descripción más detallada de Gene Ontology en la sección Métodos. De aquí en más, cuando nos refiramos a “grupos funcionales de genes” nos estaremos refiriendo a la lista de genes de un organismo que han sido asociados a cierto término GO.

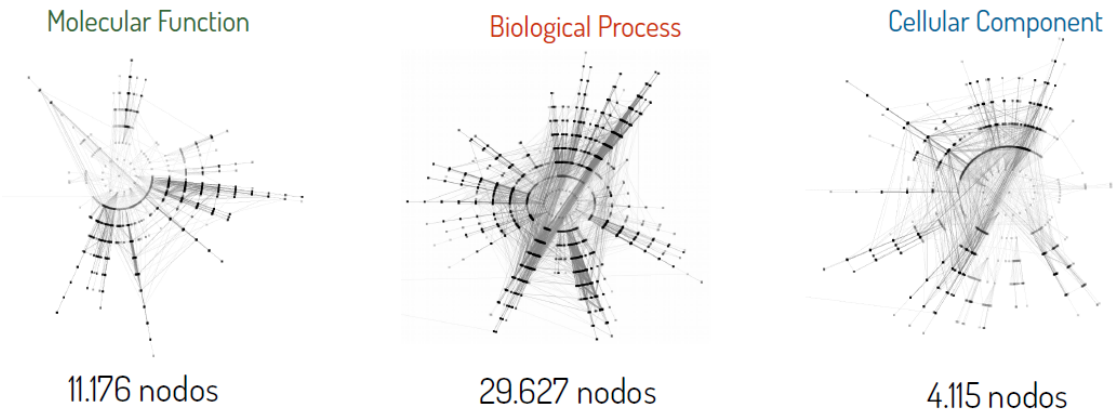


Figura 2.5 *– Gene Ontology. - Representación de los tres grafos acíclicos direccionados que conforman Gene Ontology. Cada nodo en un grafo representa un término GO y las aristas representan las relaciones entre ellos. Los grafos fueron creados por Diego Silvera con el software Cytoscape (Shannon et* al. 2003).

Utilizamos Cluster Locator para caracterizar sistemáticamente la manera en que los grupos funcionales de genes se agrupan en el genoma de 5 organismos modelo: *Homo sapiens*, *Mus musculus*, *Drosophila melanogaster*, *Caenorhabditis elegans* y *Saccharomyces cerevisiae* (los mismos 5 organismos que Cluster Locator tiene precargados). Para ello construimos lo que llamamos el "perfil de agrupamiento" de cada grupo funcional en cada uno de estos organismos. El perfil de agrupamiento de una lista de genes son las distancias encontradas, para una serie de valores de max-gap, entre el agrupamiento de la lista analizada y el agrupamiento esperable por azar. El agrupamiento esperable por azar es el promedio del agrupamiento encontrado en 1.000 listas de genes, del mismo tamaño que la lista analizada, pero seleccionados al azar. Este agrupamiento por azar tiene una distribución normal con un desvío estándar asociado, que utilizamos como unidad para medir la distancia entre el agrupamiento de la lista analizada y el agrupamiento esperable por azar.

El perfil de agrupamiento así definido es un vector que caracteriza la forma en que el agrupamiento de una lista de genes se aparta de lo esperable por azar a diferentes escalas de análisis. Al analizar dos listas de genes de tamaño diferente y con patrones de distribución diferentes utilizando un mismo valor de *max-gap* los porcentajes de agrupamiento encontrados pueden estar, por mera coincidencia, a la misma distancia de lo esperable por azar. Sin embargo, es extremadamente improbable que las distancias al azar de los porcentajes de agrupamiento de las listas sean similares para toda una serie de valores de *max-gap*, a menos que sus patrones de agrupamiento también se asemejen. La Figura 2.6 muestra el perfil de agrupamiento de una lista de 500 genes tomados al azar del genoma de *Drosophila*.

Utilizando Cluster Locator determinamos los perfiles de agrupamiento de todos los grupos funcionales con al menos 20 y no más de 1.000 genes en alguno de los 5 organismos considerados. Estos límites fueron definidos porque reducen el total de listas a analizar a un volumen manejable dejando fuera términos de menor interés en este contexto, ya sea por ser demasiado específicos o por ser demasiado generales. La Tabla 2.1 muestra cuántos términos GO cumplen estas restricciones en cada uno de los 5 organismos estudiados. La Figura 2.7 muestra, a modo de ejemplo, los perfiles de agrupamiento de los grupos de genes asociados a cuatro términos GO en los 5 organismos estudiados.

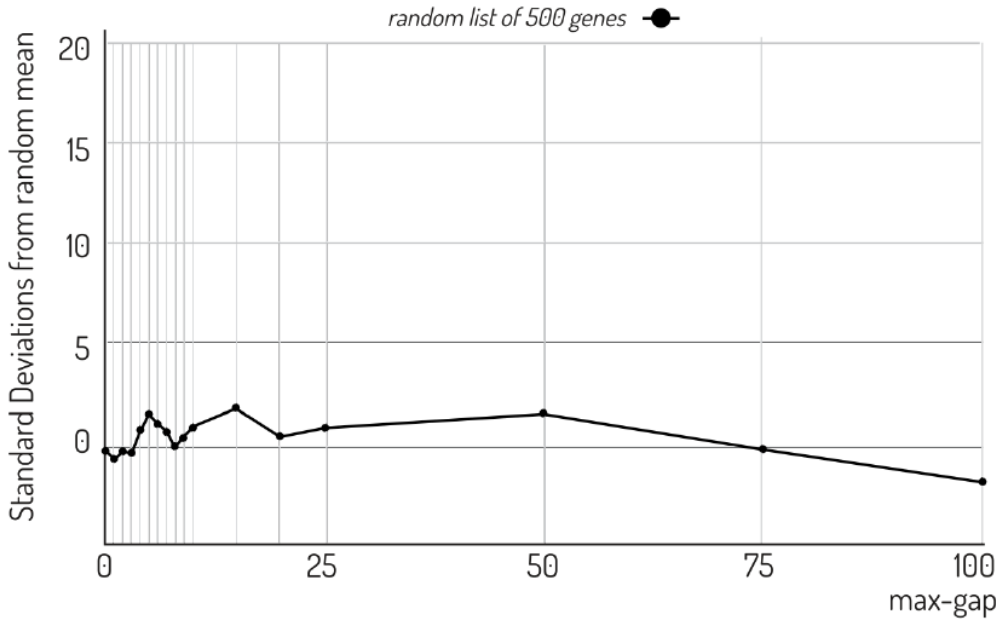


Figura 2.6 – Perfil de agrupamiento de una lista de 500 genes tomados al azar del genoma de *Drosophila melanogaster*. *En el eje de las abscisas los max-gap utilizados para definir los clusters y en el eje de las ordenadas la cantidad de desvíos estándar que separan al porcentaje de agrupamiento encontrado en en esta lista en particular del porcentaje de agrupamiento promedio encontrado en 1.000 listas del mismo tamaño también tomadas al azar. Como es de esperar, el porcentaje de agrupamiento de una lista de genes tomados al azar no se aleja significativamente del porcentaje de agrupamiento esperable por azar.*

Organismo	Ontología		
	BP	CC	MF
S. cerevisiae	1.136	244	298
C. elegans	1.166	229	340
D. melanogaster	1.804	295	425
M. musculus	3.750	506	743
H. sapiens	3.236	460	709

Tabla 2.1 *Términos GO analizados por organismo y ontología. Para cada organismo se muestra la cantidad de términos GO asociados con más de 20 pero menos de 1.000 genes.*

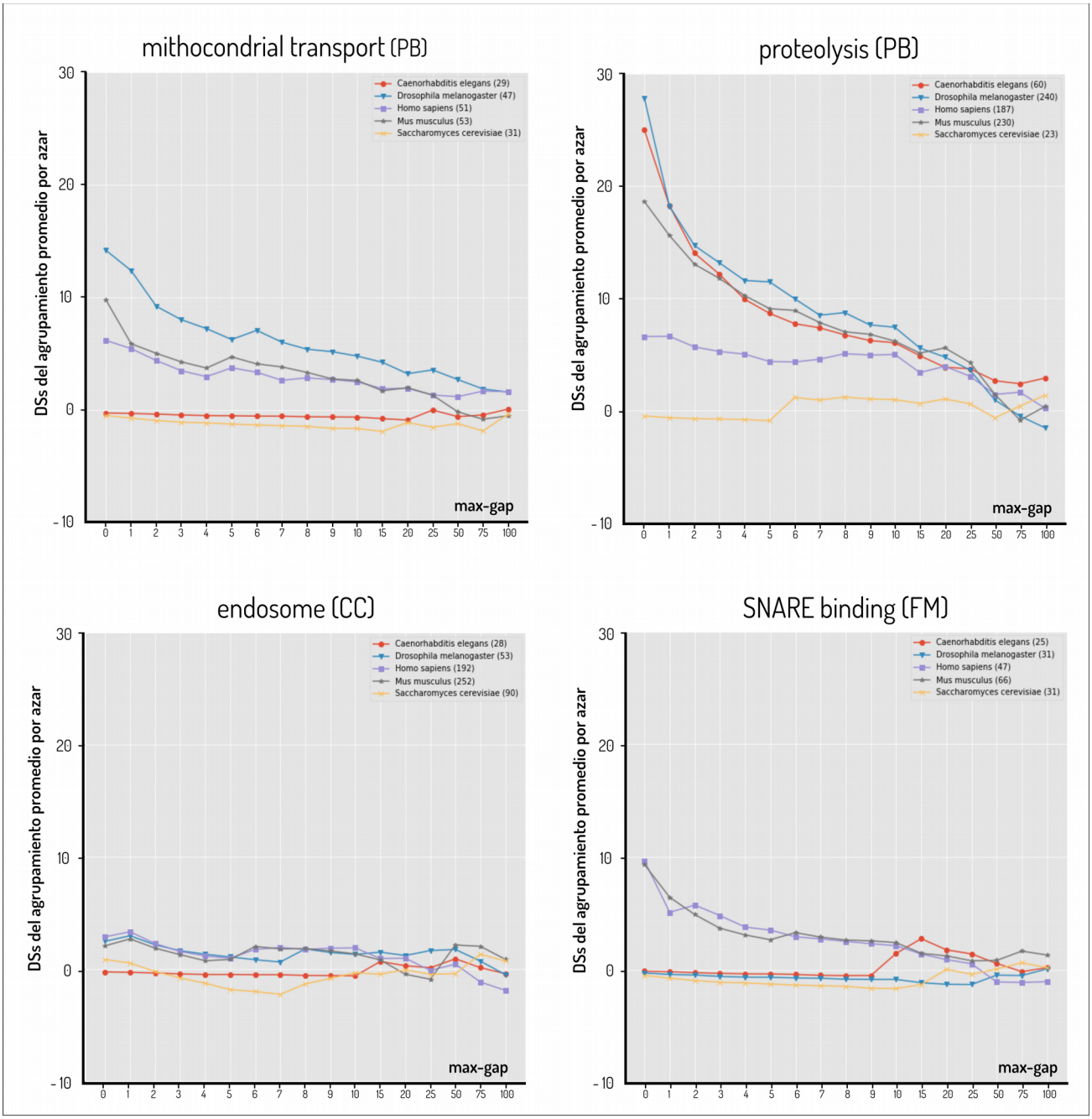


Figura 2.7 *– Ejemplos de perfiles de agrupamiento de los genes asociados a un mismo término GO en los 5 organismos considerados. En el eje de las abscisas los max-gap utilizados para definir los clusters y en el eje de las ordenadas la cantidad de desvíos estándar que separan al porcentaje de agrupamiento encontrado en la lista de interés del porcentaje de agrupamiento promedio encontrado en 1.000 listas del mismo tamaño pero tomadas al azar. PB: Procesos biológicos, CC; Componente celular, FM: Función molecular.*

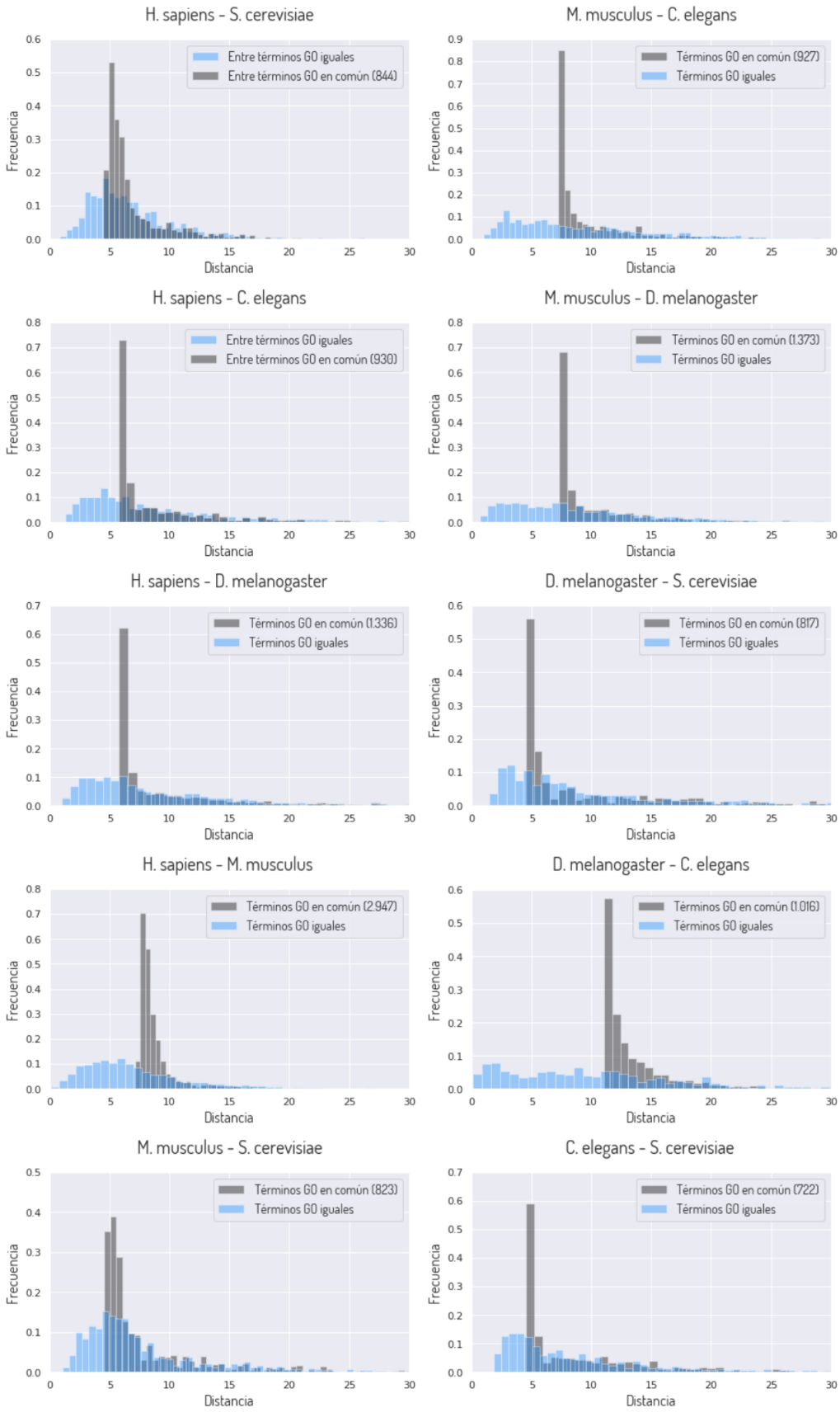
Grupos de genes en distintos organismos con la misma función y con perfiles de agrupamiento similar.

Para intentar responder a la pregunta de si los grupos de genes con la misma función tienen patrones de distribución similar en distintos organismos proponemos caracterizar la distribución de las distancias entre sus correspondientes perfiles de agrupamiento y compararla con la distribución de las distancias entre todos los perfiles de agrupamiento de ambos organismos.

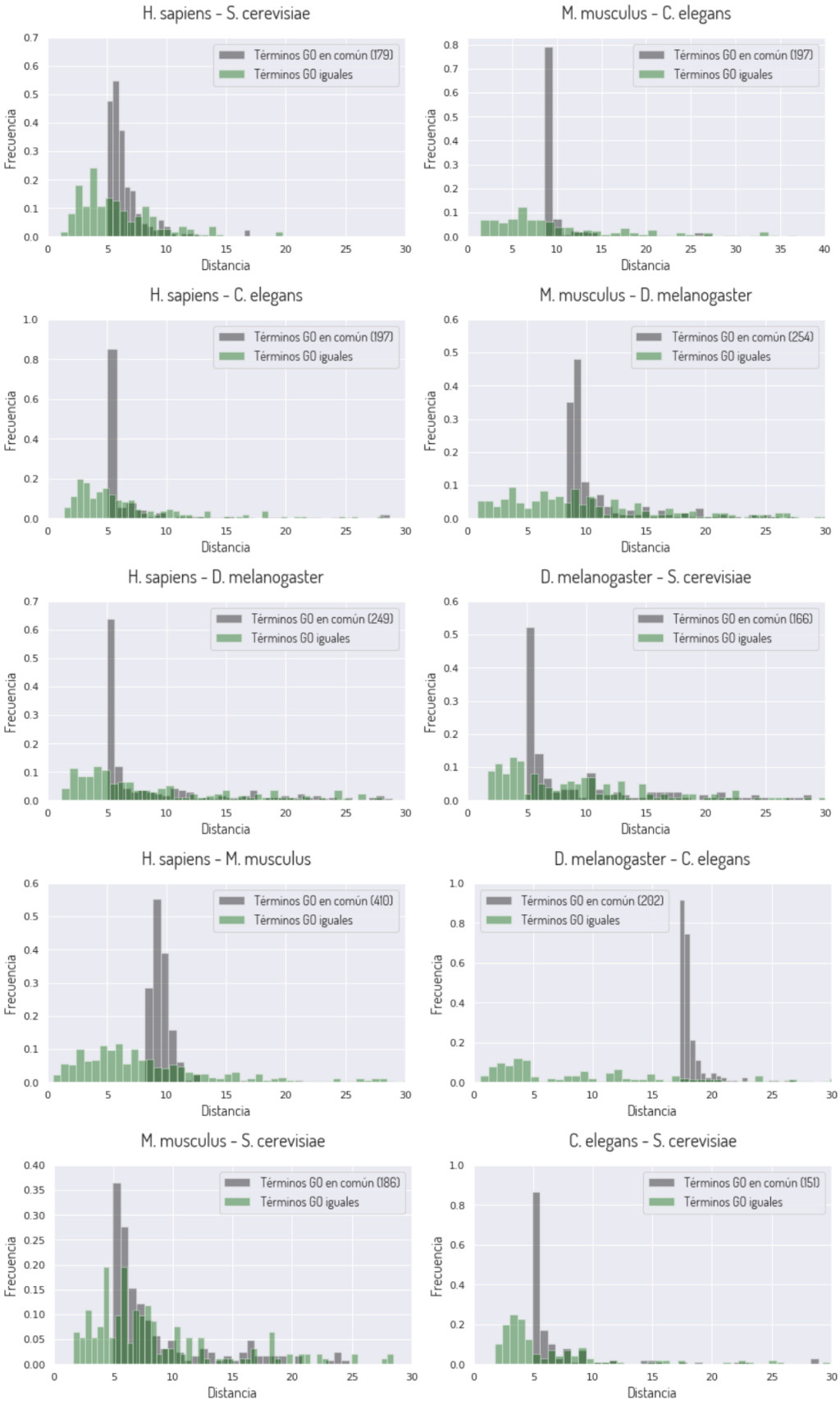
Por ejemplo, hay 844 términos GO de la ontología "Procesos Biológicos" que están asociados con al menos 20 genes tanto en *H. sapiens* como en *S. cerevisiae*, a los que llamaremos "términos GO en común". Primero calculamos la distancia entre el perfil de agrupamiento correspondiente a cada uno de estos términos GO en un organismo y todos los perfiles de agrupamiento correspondientes a los términos GO en común del otro organismo (son 844 x 844 distancias). A continuación construimos un histograma de los 844 promedios. Luego calculamos las distancias entre los perfiles de agrupamiento correspondientes al mismo término GO en uno y otro organismo (son 844 distancias) y construimos otro histograma. Comparando ambos histogramas podemos evaluar si existen diferencias en las distribuciones de estas distancias.

La Figura 2.8 muestra los resultados de este procedimiento. Los perfiles de agrupamiento correspondientes al mismo término GO en uno y otro organismo tienden a estar más cerca entre sí que la distancia a la que se encuentran pares cualquiera de perfiles de agrupamiento. Esto se cumple en las tres ontologías y entre todos los posibles pares de organismos que se forman con los cinco organismos considerados.

BP


H. sapiens - S. cerevisiae
Entre términos GO iguales
Entre términos GO en común (844)
M. musculus - C. elegans
Términos GO en común (927)
Términos GO iguales
H. sapiens - C. elegans
Entre términos GO iguales
Entre términos GO en común (930)
M. musculus - D. melanogaster
Términos GO en común (1.373)
Términos GO iguales
H. sapiens - D. melanogaster
Términos GO en común (1.336)
Términos GO iguales
D. melanogaster - S. cerevisiae
Términos GO en común (817)
Términos GO iguales
H. sapiens - M. musculus
Términos GO en común (2.947)
Términos GO iguales
D. melanogaster - C. elegans
Términos GO en común (1.016)
Términos GO iguales
M. musculus - S. cerevisiae
Términos GO en común (823)
Términos GO iguales
C. elegans - S. cerevisiae
Términos GO en común (722)
Términos GO iguales
Frecuencia
Distancia

CC


H. sapiens - S. cerevisiae
Términos GO en común (179)
Términos GO iguales
Frecuencia
Distancia
M. musculus - C. elegans
Términos GO en común (197)
Términos GO iguales
Frecuencia
Distancia
H. sapiens - C. elegans
Términos GO en común (197)
Términos GO iguales
Frecuencia
Distancia
M. musculus - D. melanogaster
Términos GO en común (254)
Términos GO iguales
Frecuencia
Distancia
H. sapiens - D. melanogaster
Términos GO en común (249)
Términos GO iguales
Frecuencia
Distancia
D. melanogaster - S. cerevisiae
Términos GO en común (166)
Términos GO iguales
Frecuencia
Distancia
H. sapiens - M. musculus
Términos GO en común (410)
Términos GO iguales
Frecuencia
Distancia
D. melanogaster - C. elegans
Términos GO en común (202)
Términos GO iguales
Frecuencia
Distancia
M. musculus - S. cerevisiae
Términos GO en común (186)
Términos GO iguales
Frecuencia
Distancia
C. elegans - S. cerevisiae
Términos GO en común (151)
Términos GO iguales
Frecuencia
Distancia

MF

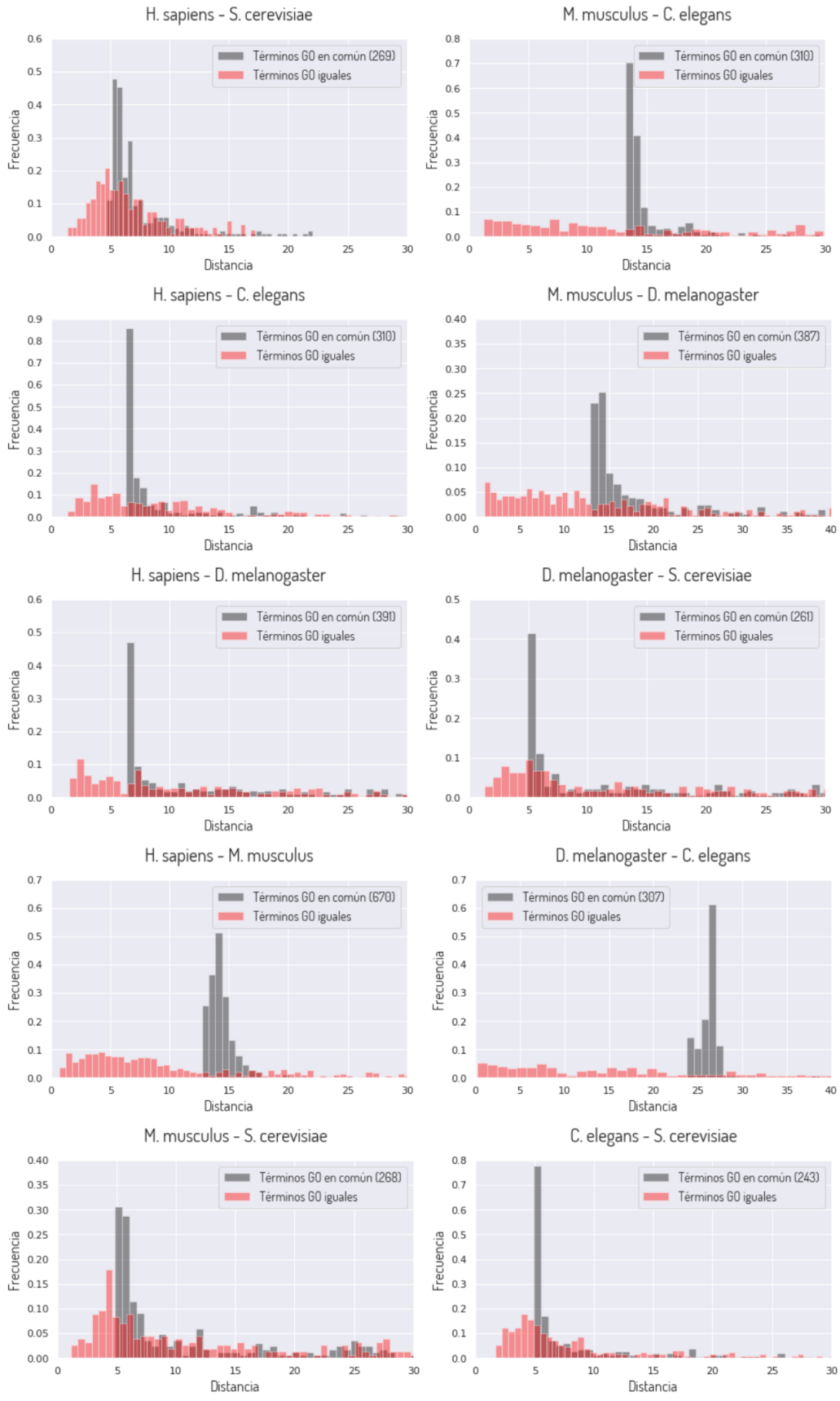

H. sapiens - S. cerevisiae
Términos GO en común (269)
Términos GO iguales
Frecuencia
Distancia
M. musculus - C. elegans
Términos GO en común (310)
Términos GO iguales
Frecuencia
Distancia
H. sapiens - C. elegans
Términos GO en común (310)
Términos GO iguales
Frecuencia
Distancia
M. musculus - D. melanogaster
Términos GO en común (387)
Términos GO iguales
Frecuencia
Distancia
H. sapiens - D. melanogaster
Términos GO en común (391)
Términos GO iguales
Frecuencia
Distancia
D. melanogaster - S. cerevisiae
Términos GO en común (261)
Términos GO iguales
Frecuencia
Distancia
H. sapiens - M. musculus
Términos GO en común (670)
Términos GO iguales
Frecuencia
Distancia
D. melanogaster - C. elegans
Términos GO en común (307)
Términos GO iguales
Frecuencia
Distancia
M. musculus - S. cerevisiae
Términos GO en común (268)
Términos GO iguales
Frecuencia
Distancia
C. elegans - S. cerevisiae
Términos GO en común (243)
Términos GO iguales
Frecuencia
Distancia

Figura 2.8 *– Histogramas de distancias entre términos GO. Se muestran por separado las tres ontologías: página 36: Procesos biológicos, página 37: Componente celular, página 38: Función molecular. Cada panel está conformado por diez histogramas, que corresponden a los 10 posibles pares que forman los cinco organismos analizados. En negro se grafica la distribución de las distancias promedio entre el perfil de agrupamiento de cada término GO en un organismo y los perfiles de agrupamiento de todos los términos GO en el otro organismo. En color azul, verde o rojo, según la ontología, se grafica la distribución de las distancias entre los perfiles de agrupamiento del mismo término GO en ambos organismos. El eje x de los histogramas está cortado, dejando fuera unos pocos casos extremos con distancias muy grandes.*

Los resultados que se muestran en la Figura 2.9 indican que los patrones de distribución correspondientes al mismo término GO en dos organismos diferentes tienden a parecerse entre sí. Según estos resultados, la primer pregunta planteada en este capítulo tiene un respuesta afirmativa.

Grupos de genes con funciones similares y perfiles de agrupamiento similar.

La segunda pregunta que nos planteamos al inicio de este capítulo fue: ¿en un organismo dado, los grupos de genes con funciones similares tienen patrones de distribución similares? Si la respuesta es afirmativa deberían existir grupos de términos GO cercanos en la ontología cuyos respectivos perfiles de agrupamiento se parezcan entre sí.

Por “grupo de términos GO cercanos en la ontología” nos referiremos aquí a un sub-grafo de la ontología con una distancia promedio entre sus nodos que es menor a la distancia que se espera encontrar en un sub-grafo formado por la misma cantidad de nodos, pero seleccionados al azar del grafo total. En el grafo de una ontología GO (Fig. 2.5), cada nodo representa un término GO, por lo que la distancia entre dos términos GO (que son dos nodos de un grafo) está bien definida: es la cantidad de aristas que separa a esos nodos. El procedimiento llevado a cabo para estos cálculos se desarrolla en la sección Métodos.

Podríamos dar una respuesta afirmativa a la pregunta planteada más arriba si encontramos una manera de dividir el grafo total de la ontología en un conjunto de subgrafos cuyos nodos sean cercanos y cuyos respectivos perfiles de agrupamiento sean parecidos. Sería deseable además que estos subgrafos contengan la mayor cantidad posible de términos GO. Se trata por lo tanto de buscar una partición de la ontología que maximice la concentración y la cantidad de nodos de cada

elemento de la partición y minimice la varianza entre los perfiles de agrupamiento correspondientes. El algoritmo desarrollado para obtener tal partición se explica en la sección Métodos.

Mediante este procedimiento obtuvimos, para cada organismo y ontología, un conjunto de subgrafos que cumplen con las características que buscábamos. La Tabla 2.2 resume los resultados obtenidos, la figura 2.9 muestra tres ejemplos de este tipo de subgrafos y la Figura 2.10 muestra el porcentaje del total de términos GO considerados que pertenece a alguno de estos subgrafos en cada organismo y para cada ontología.

	términos GO considerados	términos GO en algún subgrafo	% de términos GO en subgrafos	cantidad de subgrafos	tamaño promedio de subgrafo
BP					
S. cerevisiae	1.136	967	85	119	8
C. elegans	1.166	865	74	125	7
D. melanogaster	1.804	1.326	74	194	7
M. musculus	3.750	2.771	74	359	8
H. sapiens	3.236	2.518	78	315	8
CC					
S. cerevisiae	244	181	74	27	7
C. elegans	229	158	69	24	7
D. melanogaster	295	179	61	29	6
M. musculus	507	341	67	45	8
H. sapiens	460	343	75	41	9
MF					
S. cerevisiae	298	253	85	27	9
C. elegans	340	256	75	31	9
D. melanogaster	425	312	73	40	8
M. musculus	743	593	80	60	11
H. sapiens	709	543	77	62	9

Tabla 2.2 - Cantidad y porcentaje de los términos GO considerados que forman parte de alguno de los subgrafos encontrados, así como cantidad subgrafos y cantidad promedio de términos GO que incluyen en cada organismo y ontología.

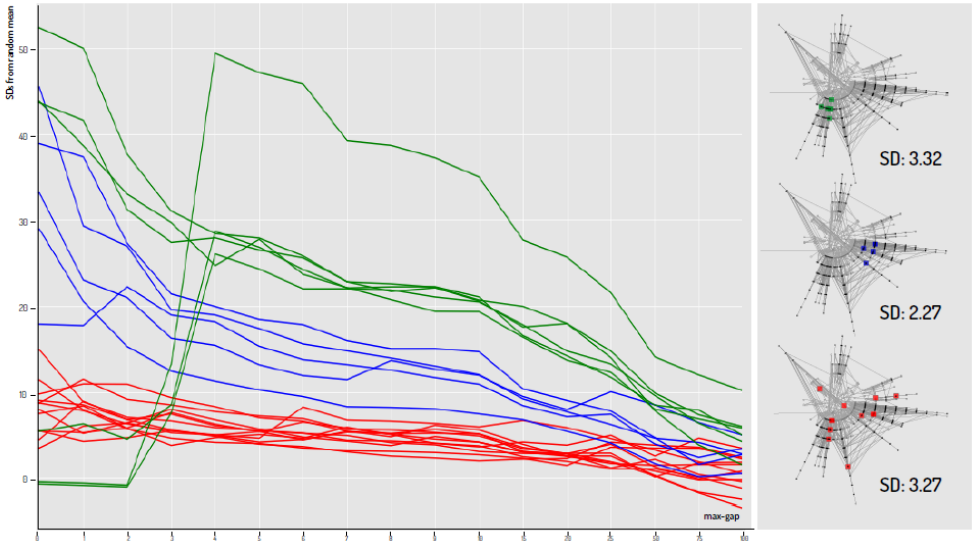


Figura 2.9 *– Ejemplos de grupos funcionales cercanos con perfiles de agrupamiento similar. En el panel de la izquierda se muestran los perfiles de agrupamiento de tres grupos de términos GO, cuya ubicación en el grafo de la ontología Función Molecular se muestra en el panel de la derecha. Al lado de cada grafo se indican los desvíos estándar encontrados entre la concentración del correspondiente subgrafo y la concentración promedio de subgrafos del mismo tamaño generados al azar.*

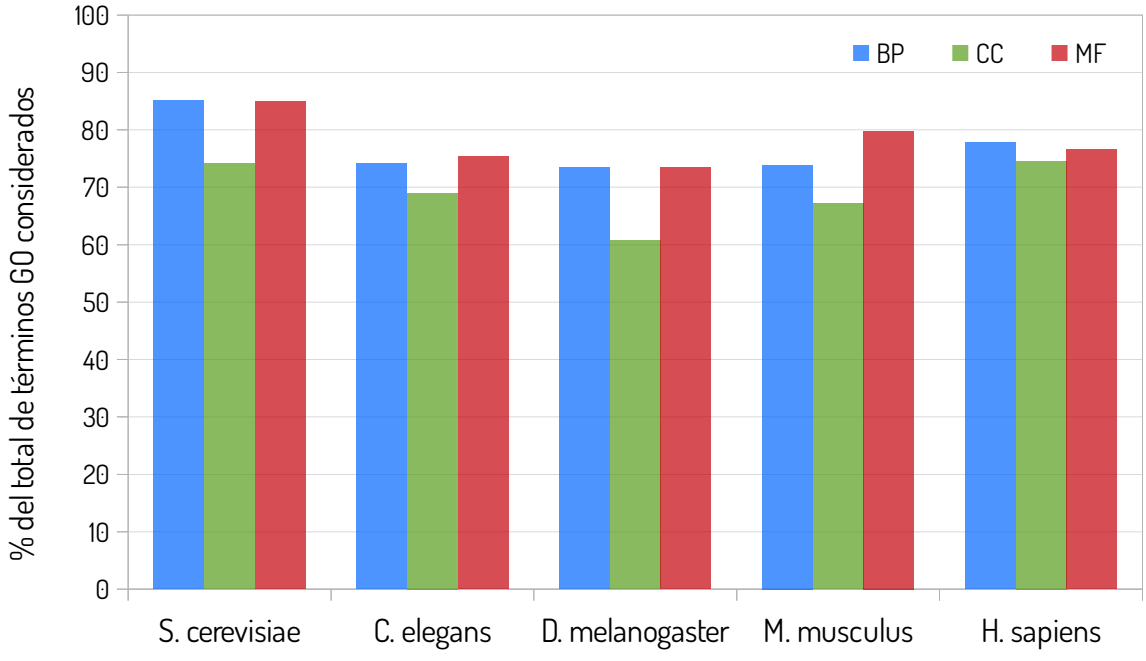


Figura 2.10 *- Proporción de términos GO que forman parte de algún elemento de la partición. Para cada organismo y ontología se muestra el porcentaje de términos GO analizados que forman parte de alguno de los subgrafos compuestos por términos GO concentrados en la ontología y con perfiles de agrupamiento significativamente similares entre sí. En los 5 organismos considerados, una proporción importante de los términos GO analizados, que va desde un 49% hasta un 83%, puede ser incluida en alguno de estos elementos de la partición.*

Los resultados de la Figura 2.10 demuestran que, en las 3 ontologías y en los 5 organismos estudiados, entre un 60% y un 85% de los grupos funcionales considerados tienen patrones de distribución que se parecen a los de grupos funcionales cercanos en la ontología.

2.4 - Discusión

Al inicio de este capítulo nos planteamos dos preguntas. ¿Existen grupos de genes con la misma función que tienen un patrón de distribución similar en distintos organismos? ¿En un organismo dado, los grupos de genes con funciones similares tienen patrones de distribución similares? Dada la falta de herramientas de análisis estadístico adecuadas, para poder abordar estas preguntas desarrollamos Cluster Locator y algunas aproximaciones y conceptos, como los perfiles de agrupamiento, que nos permitieron ahondar en dichas preguntas.

Creemos que Cluster Locator puede ser usado de muchas maneras, más allá de lo que motivó su desarrollo. En la actualidad son frecuentes los abordajes experimentales cuyo resultado principal es una lista de genes. El ulterior análisis de tales listas habitualmente se reduce a un análisis básico de enriquecimiento funcional, que determina si en la lista hay cierta función biológica que esté sobrerrepresentada. Cluster Locator permite una manera novedosa de analizar listas de genes y, al considerar únicamente la ubicación en el genoma de los genes que la componen, no se ve limitado por la cantidad de información que se disponga de los mismos. Además, por su abordaje estadístico, la herramienta permite la comparación de resultados obtenidos al analizar listas de distintos tamaños y provenientes de diferentes organismos. Todo esto convierte a Cluster Locator en una herramienta que subsana una carencia que existía en el abanico de opciones con la que se contaba para analizar listas de genes. En apoyo de esta idea se puede mencionar que la herramienta ha sido citada por varios estudios publicados por otros grupos de investigación (Wang,

Wang, y Li 2018; Planells et al. 2019; Mossman, Biancani, y Rand 2019; Shastry y Sanjay 2020).

Además, Cluster Locator se puede utilizar para poner a prueba diferentes hipótesis acerca de la evolución y organización de los genomas y de las familias multigénicas. Que los genes no se ubican aleatoriamente a lo largo de los genomas a los que pertenecen es un hecho bien establecido y se sabe que varios de los mecanismos que determinan la ubicación de un gen, como las duplicaciones en tándem y los re arreglos cromosómicos, está sujetos a presión selectiva. Análisis efectuados con varios *max-gaps* pueden arrojar luz sobre el rol que han tenido estos distintos mecanismos evolutivos en el agrupamiento de genes con diferentes funciones.

Aquí mostramos resultados obtenidos al utilizar Cluster Locator para caracterizar los patrones de distribución de grupos funcionales de genes. Como planteamos al inicio de este capítulo, si bien existe abundante evidencia acumulada que señala que los grupos funcionales de genes tienen a agruparse en los genomas, las definiciones de "*cluster*" y de "grupo funcional de genes" son muy variables, lo que dificulta la comparación de nuestros resultados con los obtenidos por otros investigadores. Existe sin embargo un antecedente con varias coincidencias metodológicas, con el cual entendemos oportuno comparar nuestros resultados.

En el año 2009 Tuller y colaboradores publicaron un estudio sobre la relación entre la localización de los genes y sus funciones biológicas en 16 organismos (Tuller et al. 2009b). Los autores, al igual que hacemos en este trabajo, consideran grupos funcionales a los grupos de genes que comparten una anotación GO y utilizan dos maneras de definir qué es un *cluster,* una de las cuales es muy semejante a la que utilizamos nosotros. Según este trabajo, los genes de cierto grupo funcional pueden formar un *cluster* si son adyacentes en el genoma o si la distancia (medida en pares de bases) entre cualquier par de genes consecutivos del *cluster* nunca supera cierto umbral. Según ambas definiciones, un gen aislado forma un *cluster* de un solo gen, por lo que un grupo de 20 genes dispersos a lo largo del genoma forma 20 *clusters*. Así definidos, cuantos menos *clusters* en total se formen entre los genes de cierto grupo funcional, más agrupado (u "organizado", según el término que utilizan los autores) se lo considera.

Para determinar el nivel de agrupamiento de un grupo funcional (o su "organización", como lo llaman los autores), el estudio compara la cantidad de *clusters* formados por los genes que lo forman con la cantidad de *clusters* que forman, en promedio, la misma cantidad de genes, pero tomados al azar del genoma. Mediante este procedimiento los autores asignan un p-valor al nivel de agrupamiento de cada grupo funcional y luego determinan, en 16 organismos, la proporción de grupos funcionales significativamente agrupados (los que tienen un p-valor < 0,05). Los organismos considerados incluyen a un procariota *(E. coli*) y a los 15 organismos eucariotas con más cantidad

de anotaciones por ese entonces, entre los cuales están los 5 organismos que abordamos en nuestro estudio.

Según los resultados obtenidos por Tuller y colaboradores, el porcentaje de los grupos funcionales que están significativamente agrupados en *S. cerevisiae* es de un 11%, en *C. elegans* de un 30%, en *D. melanogaster* de un 23%, en *M. musculus* de un 3% y en *H. sapiens* de un 10%. Es claro que no hay una correlación lineal entre el grado de complejidad de un organismo y la proporción de sus grupos funcionales que se encuentran significativamente agrupados. Lo que aquí nos propusimos no fue caracterizar, mediante un único p-valor, el nivel de agrupamiento de cada grupo funcional, sino de caracterizar y comparar la *forma* en que ese nivel de agrupamiento se aparta del azar considerando diferentes max-gaps. Sin embargo, nuestros resultados apuntan en la misma dirección. La Figura 2.8 muestra que no hay una correlación entre la distancia filogenética que separa a dos organismos y la proporción de los grupos funcionales en común que se parecen entre sí. La Figura 2.10 y la Tabla 2.2 muestran que tampoco hay correlación entre la complejidad de un organismo y la proporción de sus grupos funcionales que forman grupos de términos GO cercanos en la ontología y con perfiles de agrupamiento similares.

En el citado trabajo, los autores proponen que existe una organización "de segundo nivel" en el genoma, según la cual los grupos funcionales tienden a ubicarse unos respecto a otros de cierta manera. Para investigar esa hipótesis definen una medida de co-localización de pares de términos GO, que expresa la tendencia de genes asociados a dos términos GO diferentes a formar clusters mixtos. Luego determinan la correlación entre la co-localización de dos términos GO y su distancia en el grafo de la ontología en dos organismos: *S. cerevisiae* y *E. coli*. Esta correlación es mucho más clara en *S. cerevisiae*, por lo que los autores concluyen que al menos en este nivel de organización, el genoma eucariota está más organizado que el procariota, un resultado llamativo, ya que se suele considerar que los genomas procariotas son más organizados.

Aquí mostramos resultados que profundizan en esa misma dirección. Nuestros subgrafos de términos GO con perfiles de agrupamiento similar y cercanos en la ontología son una manera de corroborar que esa organización de "segundo nivel" del genoma también ocurre en otros 4 organismos eucariotas. Es claro que cuanto más parecidos sean los perfiles de agrupamiento de dos grupos funcionales, más clusters mixtos tenderán a formar entre sí. Las Figuras 2.9 y 2.10 muestran que, en un organismo dado, los grupos de genes con funciones similares tienden a tener patrones de distribución similares.

Entendemos que nuestros resultados son relevantes por varios motivos. Por un lado, representan evidencia clara de que una proporción significativa de los grupos de genes asociados a

la misma función en organismos filogenéticamente muy distantes han conservado ciertas características de su distribución a lo largo de los genomas. Indican también que el patrón de distribución en el genoma de un grupo funcional está vinculado a la posición que tiene en la ontología génica el término GO correspondiente, ya que términos GO cercanos tienden a tener patrones de distribución similares.

Finalmente, creemos que es importante haber basado la definición de grupo funcional en la ontología génica y haber utilizado una manera de caracterizar los patrones de distribución que permite la comparación directa entre grupos funcionales, independientemente del tamaño o del organismo. De esta manera, los resultados que hemos obtenido se podrán comparar y conjugar muy fácilmente con otros resultados que se obtengan en el futuro.

Capítulo III

Predicción de función de genes a partir de su ubicación

3.1 - Antecedentes

- La ubicación de un gen como variable predictiva de sus funciones.
- Clasificación jerárquica multiclase.

3.2 - Análisis de enriquecimiento funcional local

- Mapas del enriquecimiento funcional en 5 organismos.

3.3 - Predicción de función de genes a partir del enriquecimiento funcional local

- Implementación de los modelos predictivos.
- Evaluación.
- Resultados.

3.4 - Discusión

3.1 – Antecedentes

Los resultados reunidos en este capítulo fueron obtenidos al intentar predecir nuevas asociaciones entre genes y términos GO mediante aprendizaje automático, utilizando como únicas variables predictivas ciertas características derivadas de la ubicación de los genes en el genoma. En esta primera sección repasaremos algunos antecedentes del uso de la ubicación como variable predictiva de la función de un gen, así como algunas características particulares que tiene la clasificación jerárquica multiclase.

La ubicación de un gen como variable predictiva de sus funciones

La variable que más habitualmente se utiliza para predecir la función de un gen es su secuencia nucleotídica. El motivo resulta evidente: la secuencia del gen determina la cadena aminoacídica de la proteína que resultará de su expresión. A su vez, la cadena aminoacídica es determinante de la estructura tridimensional de la proteína, que según el paradigma actual explica

buena parte de sus propiedades bioquímicas. Determinar la secuencia de un gen, a través de las nuevas tecnologías de secuenciación de ADN, es mucho más sencillo que determinar la estructura tridimensional de una proteína. Además, la cantidad de genomas completamente secuenciados crece de forma sostenida, lo que aumenta la probabilidad de encontrar genes homólogos utilizando programas de alineamiento de secuencias.

Contar con la secuencia nucleotídica de un gen muchas veces no es suficiente para inferir su función, y como se mencionó en la introducción, también es habitual utilizar otras características de los mismos, o de las proteínas que éstos codifican, para intentar predecir sus funciones. Ejemplos de lo último son los patrones de expresión de los genes en diversas condiciones o a lo largo del tiempo, motivos conservados de la estructura tridimensional, perfiles filogenéticos o redes de interacción. En las competencias CAFA, diseñadas para comparar métodos computacionales utilizados en la predicción de función de proteínas (Radivojac et al. 2013; Jiang et al. 2016; Zhou et al. 2019) los métodos que obtienen los mejores resultados son aquellos que integran todos estos tipos de información.

Para entrenar un algoritmo de aprendizaje automático que asigne funciones a genes, se debe contar con una muestra de entrenamiento conformada por genes de función conocida y de los cuales se conocen además otras características. Para que el abordaje tenga éxito, debe cumplirse que los genes con la misma función se parezcan en términos de esas otras características. Cuando se utiliza la secuencia aminoacídica como variable predictiva, si bien se pueden usar diferentes maneras de medir la similitud entre dos secuencias, es claro que lo que se compara son las secuencias. Lo mismo sucede cuando se utilizan patrones de expresión o motivos tridimensionales.

Sin embargo, no es tan claro qué es lo que se debe comparar para utilizar la ubicación de un gen como variable predictiva. Resulta obvio que no es la ubicación *per se*, ya que, si esto fuese así, todos los genes con la misma función tenderían a estar juntos en el genoma. Este es un problema típico del ámbito del aprendizaje automático: diseñar una manera de representar la información disponible, de modo tal que le permita a un algoritmo de aprendizaje detectar patrones subyacentes que coincidan entre los ejemplos positivos de la muestra de entrenamiento. En nuestro caso, los ejemplos positivos son los genes de función conocida.

Una de las características relacionadas a la ubicación de un gen más exploradas en la bibliografía es la "vecindad genómica". Por vecindad genómica de un gen se entiende a la porción del genoma que se encuentra en su entorno inmediato. La evidencia acumulada (Ling, He, y Xin 2009; W. Yi et al. 2001; Yanai, Mellor, y DeLisi 2002; Janga, Collado-Vides, y Moreno-Hagelsieb 2005; Zheng, Roberts, y Kasif 2002) indica que la vecindad genómica es una característica

evolutivamente conservada, por lo que en organismos diferentes, las vecindades genómicas del mismo gen tenderán a ser más parecidas cuanto más cercanos filogenéticamente sean esos dos organismos. Cuando los vecindarios genómicos están conservados en múltiples genomas, las inferencias funcionales se hacen con más seguridad. Más allá de las variantes, estos métodos se basan en el hecho de que los genes con la misma función tienden a formar *clusters* a lo largo de los genomas (Huynen et al. 2000; Overbeek et al. 1999). Existe abundante evidencia de la existencia de estos *clusters* y el Capítulo II de esta tesis incluye resultados de un estudio sistemático de los mismos. Sin embargo, también sabemos que la mayoría de las veces los genes con la misma función no son adyacentes. De hecho, una de las conclusiones del Capítulo II es que los grupos funcionales de genes a menudo presentan complejos patrones de distribución conservados evolutivamente.

En esta dirección, recientemente se ha publicado una investigación (Mihelčić, Šmuc, y Supek 2019) que documenta, tanto en procariotas como en eucariotas, la tendencia de grupos de genes con funciones no relacionadas a formar *clusters* conservados evolutivamente. Evaluando el enriquecimiento de los *clusters* de genes ortólogos (COGs) ubicados en el entorno inmediato de cada gen, el trabajo evalúa la co-ocurrencia de unos 1.000 términos GO en los vecindarios genómicos de cientos de genomas procariotas y eucariotas e identifica patrones de agrupamiento que no se pueden explicar solamente por el agrupamiento de genes funcionalmente similares. Utilizando estos enriquecimientos como variables predictivas, los autores entrenan un modelo de aprendizaje automático que predice nuevas asociaciones entre genes y términos GO.

Sin embargo, al basarse en el enriquecimiento de los *clusters* de COGs, el abordaje se sostiene en información aportada por la secuencia nucleotídica, lo que impide estudiar a la ubicación *per se* como variable predictiva de la función. Como detallaremos más adelante, en este Capítulo presentamos los resultados de entrenar un modelo de aprendizaje automático utilizando como únicas variables predictivas los enriquecimientos funcionales en todos los términos GO observados en entornos de tamaño variable alrededor de cada gen. Además de ser totalmente independiente de la secuencia, nuestro abordaje tiene la ventaja de que para predecir funciones en un organismo dado solo necesita información relativa a ese organismo. Los modelos que implementamos, uno por organismo y ontología, tienen en cuenta además la estructura jerárquica de las ontologías, cuestión que abordamos a continuación.

Clasificación jerárquica multiclase

En un problema de aprendizaje automático supervisado típicamente se dispone de un conjunto de casos, representados como matrices, en el que cada caso está asociado a una o varias

clases desconocidas. Se dispone además de un conjunto de casos de clase conocida, que conforman la muestra de entrenamiento. El objetivo es encontrar una regla de clasificación utilizando la muestra de entrenamiento tal que, al observar un caso de clase desconocida, ésta se pueda predecir a partir de la matriz que lo representa. Generalmente se dispone de otro conjunto de casos de clase conocida que se reserva para evaluar la performance del modelo clasificador una vez entrenado (Hastie, Tibshirani, y Friedman 2009). En el problema que aquí tratamos, los casos son genes y las clases son términos GO. Con las asociaciones entre genes y términos GO que conocemos entrenamos un modelo de aprendizaje automático supervisado que nos permitirá predecir nuevas asociaciones entre genes y términos GO.

La predicción de nuevas asociaciones entre genes y término GO tiene ciertas características particulares. Existen miles de términos GO y cada gen puede estar asociado a muchos de ellos. Además, los términos GO guardan entre sí relaciones estructuradas en un grafo jerárquico. Planteada así, la predicción de función de genes es un problema de clasificación jerárquica multiclase (Barutcuoglu, Schapire, y Troyanskaya 2006)

La clasificación multiclase refiere a la situación en la cual existen más de dos clases y cada nuevo caso puede pertenecer a más de una de esas clases. Cuando esas clases además guardan entre sí una relación jerárquica, como sucede con los términos GO, que están organizados en un grafo acíclico direccionado, se habla de clasificación jerárquica multiclase (HMC, por su sigla en inglés). Los algoritmos que intentan resolver este tipo de problemas tienen aplicaciones en campos tan diversos como la categorización de textos, la clasificación de géneros musicales o fonemas y también en la predicción de función de genes.

Los métodos para resolver problemas de HMC se pueden dividir en dos grupos: los que tienen un abordaje global y los que tienen un abordaje local, grupo al que pertenece la mayoría de los métodos publicados (Silla y Freitas 2011). Los métodos con abordaje global generan un único modelo clasificador, que predice todas las clases a las que se asocia cada caso ejecutando el algoritmo una sola vez. Por el contrario, los abordajes locales entrenan de manera independiente un clasificador binario por cada clase y luego combinan los resultados independientes para realizar la predicción. Los abordajes locales descomponen el problema de HMC en una serie de tareas mucho más sencillas, que se pueden resolver mediante los algoritmos de clasificación binaria convencionales, pero hay dos puntos cruciales que se deben considerar cuidadosamente.

En primer lugar, la HMC es un problema de clasificación con clases muy desbalanceadas. Para las clases en los niveles más bajos de la jerarquía, en nuestro caso, los términos GO más específicos, los casos positivos (los genes asociados a ese término GO) siempre son pocos y muchos

menos que el resto de los genes, que se suelen considerar como casos negativos. El problema es que la clase minoritaria, al estar subrepresentada, es usualmente desfavorecida por los clasificadores binarios (Otero, Freitas, y Johnson 2010), por lo que se debe implementar alguna de las técnicas compensatorias que se han desarrollado para enfrentar esta situación.

En segundo lugar, las predicciones del modelo deben ser coherentes con las restricciones que impone la jerarquía. Esto es, si el modelo asigna un caso (gen) a una clase determinada, lo debería asignar también a todas las clases que son nodos ancestros de esa clase en el grafo y no debería asignarlo a ninguna clase que descienda de un clase a la cual no lo asignó (Stojanova et al. 2013). Esto no necesariamente se cumple si se entrenan muchos clasificadores independientes, uno por término GO. Se han propuesto varios métodos que corrigen esta situación (Valentini 2014). Además de estas consideraciones generales relativas a los problemas de HMC, al abordar la predicción de función de genes se debe tener en cuenta que la jerarquía que organiza los términos GO no es un árbol, sino un grafo acíclico direccionado y que son pocos los métodos de HMC que pueden lidiar con este tipo de estructura de clases, más compleja que un árbol (Silla y Freitas 2011).

Aquí implementamos, con algunas modificaciones, el algoritmo propuesto en (Feng, Fu, and Zheng 2017; 2018). Se trata de un abordaje local que propone el "método de interacción de nodos" para conseguir la coherencia de las predicciones con la jerarquía. El método aplica la "política de los hermanos", que consiste en considerar como casos negativos a los casos positivos de los nodos hermanos (o tíos) del nodo en cuestión (Silla and Freitas 2011). Antes de detallar la arquitectura general de los modelos implementados, exponemos el modo en que construimos las variables predictivas.

3.2 Análisis de enriquecimiento funcional local [2]

Nuestra estrategia consiste en entrenar un modelo de aprendizaje automático para cada organismo y ontología, utilizando como variables predictivas los enriquecimientos funcionales en todos los términos GO observados en entornos de tamaño variable alrededor de cada gen. El análisis de enriquecimiento funcional es un procedimiento habitual para caracterizar listas de genes. Tiene como objetivo determinar si en una lista de genes de interés, cierta función biológica está sobrerepresentada, eso es, si la lista analizada incluye más genes con esa función de lo que se espera

2 - La implementación computacional de buena parte de lo que resta del Capítulo, en preparación para su publicación, es autoría del Lic. Diego Silvera, estudiante de la Maestría en Ingeniería Matemática a quien co-dirijo junto al Dr. Gustavo Guerberoff.

encontrar en una lista con la misma cantidad de genes pero tomados al azar del genoma al que pertenece la lista analizada. Para una definición formal y una explicación más detallada ver la sección Métodos. Proponemos caracterizar el entorno de un gen desde el punto de vista funcional llevando a cabo un análisis de enriquecimiento funcional de su entorno, al que llamaremos “análisis de enriquecimiento funcional local” (LEA, por la sigla en inglés). La estrategia consiste en construir matrices que representan a cada gen con los resultados obtenidos con LEA.

Por LEA nos referimos al análisis de enriquecimiento funcional en el que la lista de genes analizada es una lista de genes contiguos en el genoma, centrada en el gen de interés. Utilizamos aquí el mismo modelo de genoma descrito en el Capítulo II, para el cual cada genoma es una colección de segmentos independientes en los cuales los genes codificantes de proteínas están ubicados uno al lado del otro, sin espacios intergénicos o solapeo. Recorriendo cada cromosoma gen a gen, determinamos, para una serie de distintos tamaños de ventana, el enriquecimiento funcional local (es decir, el enriquecimiento en un cierto GO en la lista de genes definida por la "ventana") respecto a todos los términos GO asociados con más de 20 y menos de 1.000 genes en el organismo en consideración. El tamaño de la ventana es la cantidad de genes que se considera hacia un lado y otro del gen en cuestión. Este abordaje tiene un antecedente en la bibliografía (Tiirikka, Siermala, y Vihinen 2014), en el que los autores llevan a cabo un procedimiento de este tipo con el objetivo de ubicar *clusters* de términos GO en el genoma de 7 organismos modelo.

Nuestro interés aquí no es ubicar *clusters* de genes asociados al mismo término GO, sino caracterizar el entorno de cada gen desde el punto de vista del enriquecimiento funcional con el objetivo de entrenar un modelo predictivo de función génica. Para construir las variables predictivas asociadas a cada gen, concatenamos los valores de enriquecimiento encontrados en el entorno de ese gen para ciertos términos GO, que dependen del término GO que se va a predecir (ver más adelante). Este procedimiento resulta en una matriz asociada a cada gen, que tiene en sus columnas el enriquecimiento encontrado para una serie de términos GO y en sus filas los distintos tamaños de ventana analizados.

Estas matrices pueden ser interpretadas como una representación del "paisaje funcional" a un lado y otro de cada gen. Utilizar estas matrices para predecir funciones de genes solo puede dar buenos resultados si las matrices asociadas a genes que tienen la misma función, se parecen entre sí de alguna manera. Esa es nuestra hipótesis: los paisajes funcionales obtenidos por medio de LEA para genes con la misma función tienen características en común que pueden ser capturadas por un modelo de aprendizaje automático.

Mapas del enriquecimiento funcional local en 5 organismos

Al entrenar un modelo de aprendizaje automático no se deben utilizar los casos de prueba, que se deben reservar para evaluar la performance de los clasificadores de manera independiente. Teniendo esto en cuenta, para entrenar los modelos predictivos no determinamos LEA considerando el total de genes anotados con cada término GO sino que consideramos solamente aquellos genes que fueron reservados para entrenamiento. El objetivo de este procedimiento, detallado más adelante, es evitar sesgar al modelo clasificador, utilizando en su entrenamiento información derivada de casos que se usaran luego para su evaluación. Sin embargo, por considerar que son resultados que tienen valor por sí mismos, en una fase previa e independiente al entrenamiento de los modelos predictivos, calculamos LEA para cada organismo y ontología utilizando el total de asociaciones disponibles entre genes y términos GO, generando lo que denominamos "mapas de enriquecimiento funcional local".

Concretamente, llevamos a cabo el análisis LEA en el genoma de cada uno de los 5 organismos estudiados en el Capítulo II, considerando, en cada organismo, a todos aquellos términos GO que estuviesen asociados con al menos 20 genes. La tabla 3.1 resume las cantidades de términos GO con al menos una asociación y con al menos 20 asociaciones en cada organismo y ontología.

Si en lugar de centrarnos en los genes, como hacemos para obtener las matrices de "paisaje funcional" descriptas en la sección anterior, nos centramos en los términos GO, obtenemos lo que llamamos "mapas de enriquecimiento funcional", uno para cada término GO. En cada mapa se ubican las zonas del genoma enriquecidas en genes asociados al correspondiente término GO, utilizando diferentes tamaños de ventana. Para cada organismo y ontología obtuvimos tantos de estos mapas como términos GO (ver Tabla 3.1). A modo de ejemplo, las Figuras 3.1 y 3.2 muestran el enriquecimiento funcional local encontrado a lo largo del genoma de *Drosophila melanogaster* en 3 términos GO de la misma rama, para cada una de las tres ontologías.

	BP	CC	MF
S. cerevisiae	1.167	261	303
C. elegans	1.189	243	345
D. melanogaster	1.855	313	435
M. Musculs	3.936	549	770
H. sapiens	3.389	503	734

Tabla 3.1 - Cantidad de términos GO con al menos 20 anotaciones por organismo y ontología.

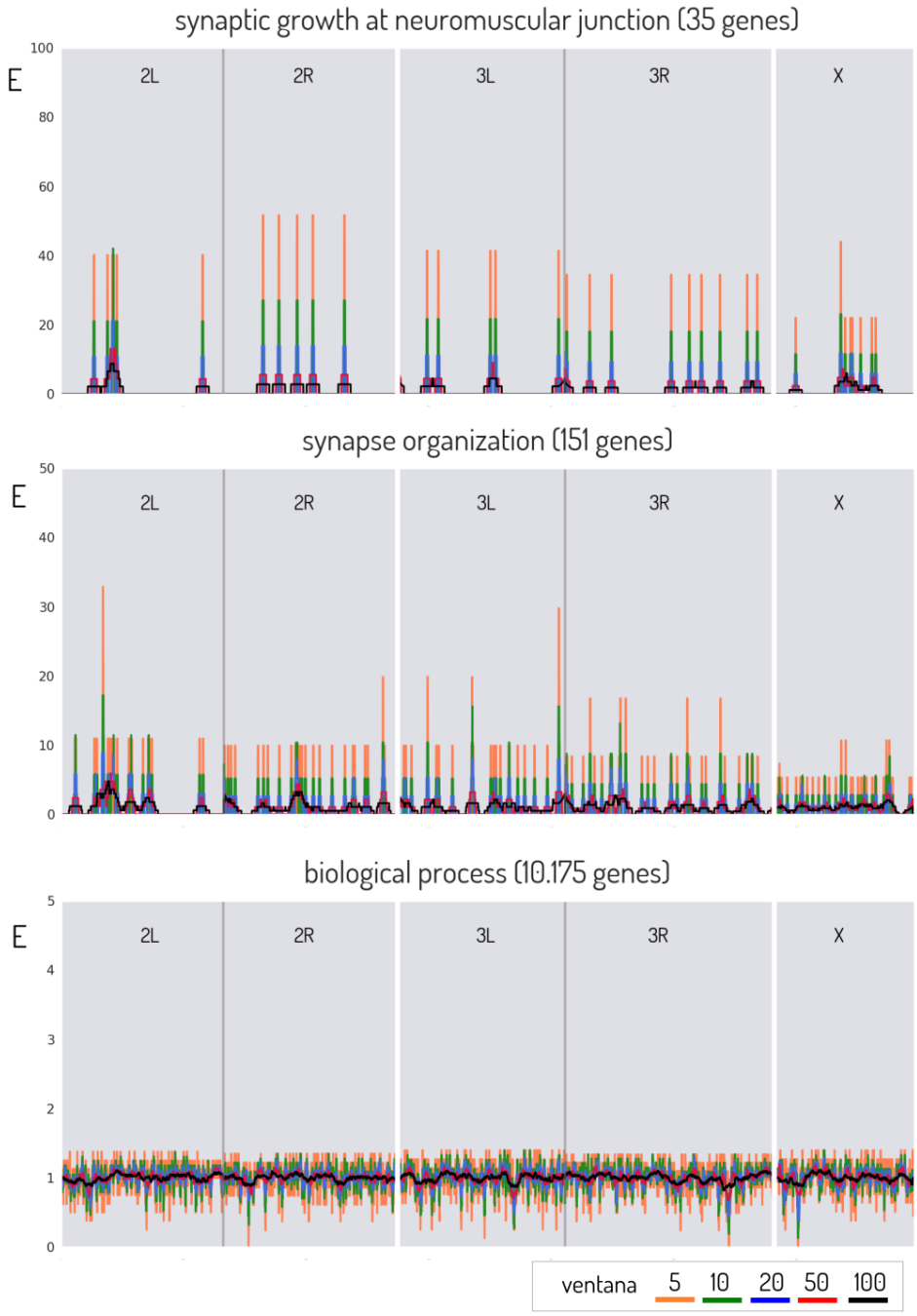


Figura 3.1 *- Enriquecimiento funcional local a lo largo del genoma de Drosophila melanogaster en tres términos GO de la misma rama de la ontología "Procesos biológicos". El término "biological process" (gráfico inferior) es el nodo raíz de la ontología, contiene a "synapse organization" (gráfico central), que a su vez contiene a "synaptic growth at neuromuscular junction" (gráfico superior). El número de genes que están anotados con cada término se indica*

entre paréntesis en la parte superior de cada gráfico. En el eje ***x*** *se representan los genes que conforman cada brazo cromosómico, cuyos nombres están indicados en negro en la parte superior. Las líneas grises verticales indican la posición de los centrómeros y las líneas blancas separan cromosomas. En el eje* ***y*** *se representa el enriquecimiento (E) encontrado en el entorno de cada gen utilizando 5 ventanas diferentes. Notar que los gráficos tienen escalas distintas en el eje* ***y****. Se excluyen los cromosomas 4 e Y por ser muy pequeños.*

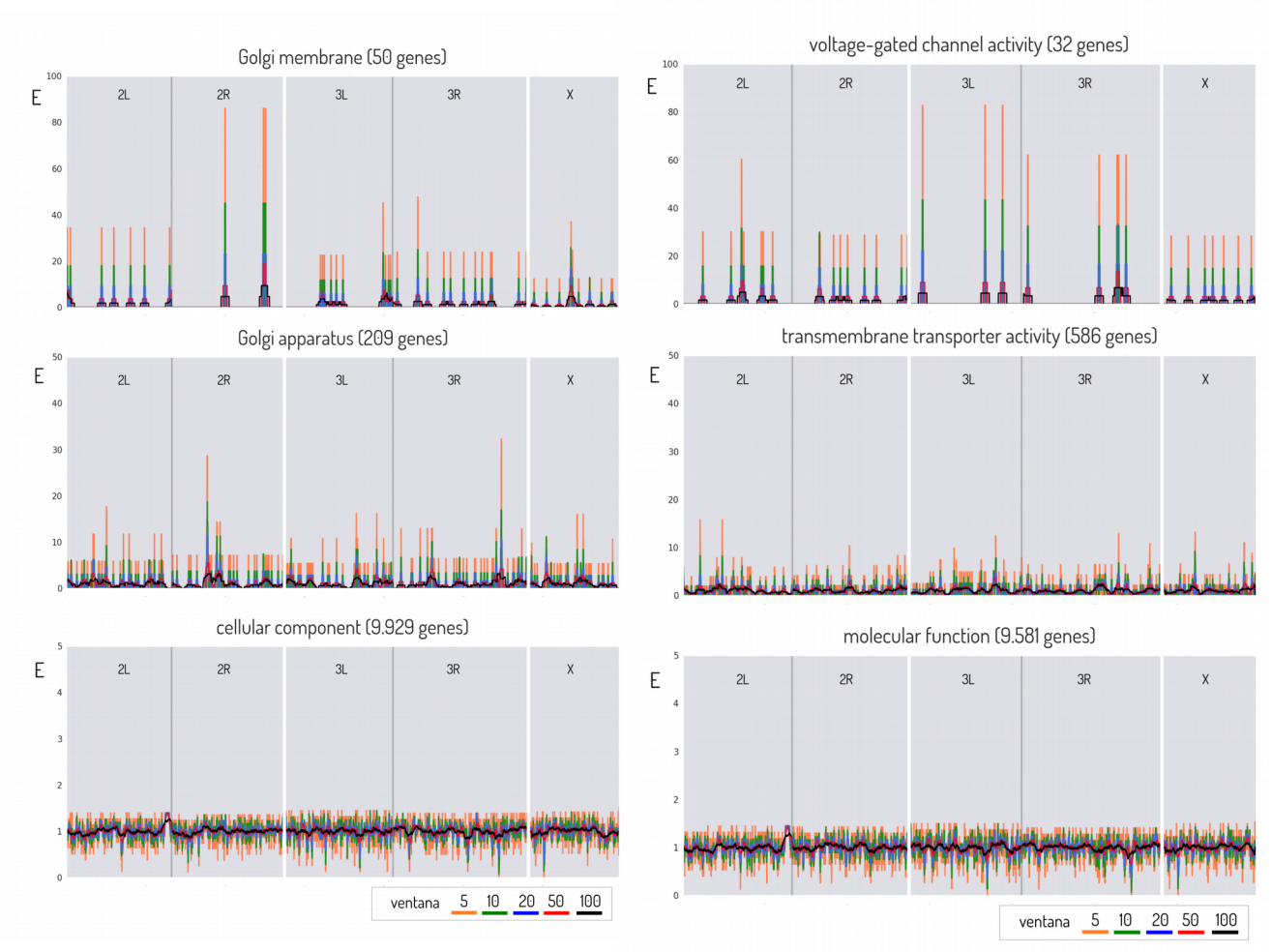


Figura 3.2 *- Enriquecimiento funcional local a lo largo del genoma de Drosophila melanogaster en tres términos GO de la misma rama de la correspondiente ontología. A la izquierda, "Componente celular" a la derecha "Función molecular". Ver leyenda de la Figura 3.1.*

Todos los mapas de enriquecimiento funcional obtenidos se pueden descargar aquí: https://github.com/IIBCE-BND/gfpml-datasets/tree/master/lea.

3.3 - Predicción de función de genes a partir del enriquecimiento funcional local

Considerando los aspectos reseñados en los antecedentes de este capítulo, decidimos implementar, con algunas modificaciones, el algoritmo propuesto por Feng, Fu y Zheng (Feng, Fu, y Zheng 2017; 2018), que implemente un abordaje local, pues entrena un clasificador para cada término GO. Para conseguir la coherencia de las predicciones con la jerarquía de la ontología, los autores proponen el "método de interacción de nodos". Como manera de seleccionar los casos negativos de cada muestra de entrenamiento se aplica la "política de los hermanos", que consiste en considerar como casos negativos a los casos positivos de los nodos hermanos (o tíos) del nodo en cuestión (Silla y Freitas 2011). A diferencia de Feng, Fu y Zheng, en lugar de SVM utilizamos Random Forests como algoritmo clasificador, que se suele desempeñar igual de bien, pero cuyo entrenamiento tiene un menor costo computacional. Por la misma razón optamos por no aplicar SMOTE, una técnica que se usa para generar artificialmente nuevos casos etiquetados cuando se dispone de muestras de entrenamiento pequeñas (Chawla et al. 2002).

Implementación de los modelos predictivos de función génica

Para cada uno de los 5 organismos considerados y para cada una de las tres ontologías, implementamos un modelo que asigna asociaciones entre genes y términos GO. La Tabla 3.2 muestra el total de genes considerados en cada uno de los 5 organismos. A lo largo de este Capítulo utilizamos versiones de las ontologías en las que, siguiendo la "true path rule", propagamos todas las anotaciones hacia la raíz del árbol (Valentini 2011). Esto quiere decir que si un gen está asociado a cierto término GO de la ontología, lo asociamos automáticamente a todos los términos GO correspondientes a los nodos ancestros hasta llegar a la raíz. Para definir las muestras de entrenamiento y las muestras de prueba, y para fijar los hiperparámetros de los modelos y evaluarlos, seguimos el procedimiento que detallamos a continuación.

Partición del genoma - En primer lugar, dividimos aleatoriamente en dos grupos a los genes de cada genoma. Uno de esos grupos, al que llamaremos **E**, incluye al 80% de los genes y se asigna a entrenamiento. El otro grupo, al que llamaremos **P**, incluye al restante 20% de los genes y se asigna a prueba. Los genes de **P** serán clasificados por el modelo una vez entrenado y como conocemos al menos algunas de sus clases, podremos evaluar cuánto se equivoca el modelo.

Cálculo de las variables predictivas - Calculamos LEA para todos los términos GO con más de 20 anotaciones y todos los genes del genoma, pero solamente considerando las anotaciones de

los genes que pertenecen a **E**. La razón de esto es evitar entrenar a los clasificadores con variables derivadas de los genes de las muestras de prueba, que se usarán para su evaluación, y no sesgar nuestro posterior análisis de performance.

Construcción de las muestras de entrenamiento y de prueba - Definimos muestras de entrenamiento y muestras de prueba para todos aquellos términos GO asociados con al menos 40 genes en **E** y con al menos 10 genes en **P** (en el caso de *M. musculus* y *H. sapiens* en la ontología de procesos biológicos estos mínimos fueron 80 y 20). Las cantidades de términos GO que cumplen con estas condiciones en cada organismo y ontología se resumen en la Tabla 3.3. Al construir cada muestra de entrenamiento y cada muestra de prueba incluimos como casos positivos a todos los genes asociados con el término GO en consideración y como casos negativos a todos los genes asociados a términos GO hermanos (o tíos) pero no al término GO en consideración (Silla y Freitas 2011). Las variables predictivas asociadas a cada gen de la muestra de entrenamiento y la muestra de prueba también dependen del término GO que se esté considerando. A cada gen se le asociaron los valores de LEA (usando 5 tamaños de ventana) correspondientes al término GO que se va a clasificar, a su término GO padre y a todos sus términos GO hijos y hermanos.

	Genes
S. cerevisiae	6.572
C. elegans	20.210
D. melanogaster	13.913
M. Musculs	21.956
H. sapiens	19.908

Tabla 3.2 - Cantidad de genes considerados por organismo.

	BP	CC	MF
S. cerevisiae	524	136	136
C. elegans	550	116	150
D. melanogaster	879	175	211
M. Musculs	1.211	337	368
H. sapiens	1.039	284	363

Tabla 3.3 - Cantidad de términos GO predichos por organismo y ontología.

Entrenamiento de los clasificadores - Con las muestras de entrenamiento correspondientes a cada término GO entrenamos los respectivos clasificadores binarios. Optamos por usar Random Forests (Breiman 2001), con hipermarámetros (profundidad y cantidad de los árboles y medida de

disimilitud) definidos tras realizar una búsqueda en grilla y validación cruzada.

Corrección de las probabilidades - Una vez entrenados, con cada uno de estos clasificadores clasificamos las correspondientes muestras de prueba, obteniendo, para cada uno de sus genes, una probabilidad de estar asociado al término GO en cuestión. Luego ajustamos estas probabilidades mediante el método de interacción de nodos, para respetar las restricciones impuestas por la jerarquía de la ontología.

Mediante este procedimiento obtuvimos un modelo entrenado para cada par organismo/ontología, cada uno de ellos constituido por un conjunto de clasificadores binarios que asignan probabilidades de asociación entre genes y cierto término GO.

Evaluación de los modelos de predicción génica

A continuación, evaluamos la performance global del modelo correspondiente a cada par organismo/ontología para una serie de umbrales de clasificación utilizando varias métricas estándar (precisión, sensibilidad y el score F1, ver Métodos). Los valores obtenidos se pueden encontrar en el Anexo 3. Además de permitirnos evaluar la performance de los modelos de manera independiente (usando los genes del grupo **P**), este procedimiento nos permitió seleccionar para cada modelo el umbral de clasificación que diese lugar al máximo valor para el *score* F1, esto es, el F-max. La Tabla 3.4 resume los umbrales de clasificación seleccionados para cada par organismo/ontología así como los valores de precisión, sensibilidad y F-max de los correspondientes modelos.

Tanto en la Tabla 3.4 como en las tablas del Anexo 3 se pueden encontrar además, para cada par organismo/ontología, las métricas de evaluación de performance de un modelo permutado, que generamos para evaluar si nuestros modelos entrenados predicen mejor que el azar. Como no conocemos la verdadera distribución de probabilidades de cada término GO no es claro de qué manera se deben generar las predicciones al azar que permitan una evaluación válida de nuestras predicciones. Decidimos generar un modelo, al que llamamos "Modelo Permutado", que toma las probabilidades asignadas por cada clasificador entrenado y las reasigna al azar entre todos los genes. Este procedimiento asegura que las probabilidades asignadas a cada gen tienen la misma distribución que las probabilidades asignadas por los modelos entrenados, lo cual permite una comparación justa, sobretodo teniendo en cuenta el gran desbalance en el tamaño de las clases. Tal como se muestra en la Tabla 3.4, en todos los organismos y ontologías nuestros modelos clasificadores se desempeñan mejor que el modelo permutado.

Organismo			Modelo			Permutación		
	Ont.	Umbral	Precisión	Sensibilidad	F-max	Precisión	Sensibilidad	F-max
S. cerevisiae	BP	0,2	0,52	0,31	**0,39**	0,41	0,03	**0,06**
	CC	0,1	0,55	0,67	**0,61**	0,51	0,09	**0,15**
	MF	0,1	0,44	0,53	**0,48**	0,36	0,14	**0,20**
C. elegans	BP	0,4	0,40	0,29	**0,34**	0,25	0,03	**0,05**
	CC	0,1	0,42	0,49	**0,45**	0,29	0,07	**0,11**
	MF	0,7	0,34	0,32	**0,33**	0,23	0,07	**0,11**
D. melanogaster	BP	0,1	0,40	0,32	**0,36**	0,30	0,03	**0,06**
	CC	0,7	0,52	0,40	**0,45**	0,42	0,03	**0,06**
	MF	0,2	0,46	0,33	**0,38**	0,37	0,06	**0,10**
M. musculus	BP	0,4	0,46	0,31	**0,37**	0,33	0,02	**0,03**
	CC	0,6	0,52	0,40	**0,45**	0,43	0,02	**0,04**
	MF	0,6	0,50	0,30	**0,38**	0,46	0,03	**0,06**
H. sapiens	BP	0,2	0,47	0,22	**0,30**	0,29	0,01	**0,02**
	CC	0,1	0,53	0,60	**0,57**	0,36	0,04	**0,08**
	MF	0,1	0,51	0,34	**0,41**	0,34	0,04	**0,06**

Tabla 3.4 - Precisión, sensibilidad y F-max de los modelos seleccionados para cada par organismo/ontología y de los correspondientes modelos permutados.

Podemos utilizar la razón entre el F-max del modelo entrenado y del correspondiente modelo permutado en cada organismo y ontología para estimar que tanto mejor que el azar son las predicciones que obtenemos con nuestro abordaje. La Tabla 3.5 resume esas razones, así como sus promedios por organismo y por ontología. La Figura 3.3 permite visualizar estos resultados.

	BP	CC	MF	Promedio
S. cerevisiae	6,5	4,0	2,4	**4,3**
C. elegans	7,1	4,0	3,1	**4,7**
D. melanogaster	5,9	7,1	3,8	**5,6**
M. musculus	12,0	10,2	6,1	**9,4**
H. sapiens	14,2	7,4	6,4	**9,3**
Promedio	**9,1**	**6,5**	**4,3**	

Tabla 3.5 - Razón entre los F-max de cada modelo entrenado y los correspondientes modelos permutados.

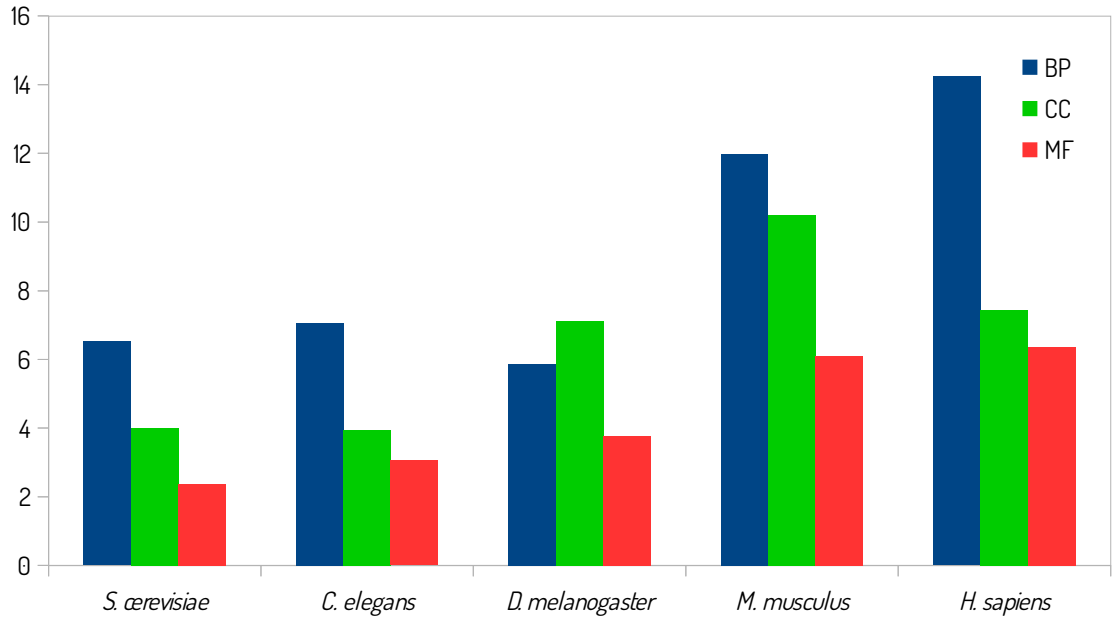


Figura 3.3 *- Razón entre el F-max del modelo entrenado y del modelo permutado para cada par organismo/ontología.*

Comparación con otros métodos

En la última edición de las competencias CAFA (Zhou et al. 2019) se evaluaron las predicciones de los métodos participantes sobre un conjunto de unas 4.300 proteínas pertenecientes a varios organismos. El método que logró el mejor desempeño global (evaluado mediante el valor de F-max) fue GOLabeler (You et al. 2018). GOLabeler utiliza una combinación de 5 clasificadores entrenados con distintas variables predictivas: la frecuencia de los términos GO, alineaciones de secuencia, trigramas de aminoácidos, presencia de dominios y motivos proteicos y propiedades biofísicas. En estas competencias se utilizan dos métodos para establecer una línea de base, uno de los cuales se basa en BLAST para asignar funciones directamente por similitud de secuencias. La Tabla 3.6 muestra los valores promedio de F-max que alcanzan en cada ontología GOLabeler y el método de referencia basado en BLAST, así como los valores que alcanzan en promedio nuestros modelos.

Modelo	BP	CC	MF
GoLabeler	0,40	0,61	0,62
BLAST	0,26	0,46	0,42
Nuestro modelo (promedio de los 5 organismos)	0,35	0,51	0,40

Tabla 3.6 - F- max por ontología de "GoLabeler" (método ganador de la CAFA 3), de uno de sus métodos de referencia (basado en BLAST) y el promedio, para cada ontología, de nuestros cinco modelos.

Finalmente, con cada uno de los modelos seleccionados en función del *score* F1, clasificamos al resto de los genes del genoma. Así obtuvimos, para cada gen, la probabilidad de estar asociado a cada término GO. Estas probabilidades, que son el resultado final de todo este procedimiento, están disponibles aquí: http://gfpml.bnd.edu.uy. En ese enlace se accede a una tabla en la que se pueden consultar las predicciones de cada modelo, buscando por organismo, ontología, cromosoma, gen o término GO.

3.4 Discusión

Los resultados reunidos en este capítulo fueron obtenidos al intentar predecir nuevas asociaciones entre genes y términos GO mediante aprendizaje automático, utilizando como única variable predictiva el enriquecimiento funcional local en el entorno de cada gen. Los antecedentes del uso de la ubicación de los genes para predecir sus funciones en organismos en eucariotas son pocos y no encontramos en la bibliografía ejemplos en los que las variables predictivas utilizadas dependan exclusivamente de la ubicación de los genes.

Aquí proponemos una manera de representar la información acerca de la ubicación de los genes y sus funciones a la que llamamos "matrices de paisaje funcional". Estas matrices de paisaje funcional se obtienen a partir de los valores de enriquecimiento funcional local que se encuentran al

recorrer el genoma gen a gen, con varios tamaños de ventana. Nuestra hipótesis era que el patrón de funciones enriquecidas en el entorno de un gen se asemeja al patrón de funciones enriquecidas en el entorno de otro gen con la misma función. Si esta hipótesis fuese cierta, con las matrices de paisaje funcional se debería poder entrenar un algoritmo de aprendizaje automático para que reconozca asociaciones aun no establecidas entre funciones y genes.

Nuestros resultados demuestran que éste es el caso.

Más allá de comparar las métricas evaluación de performance de nuestros modelos con las de otros modelos de predicción de función de genes, entendemos que lo más relevante aquí es que nuestros modelos, entrenados únicamente con información derivada de la ubicación relativa de los genes en el genoma al que pertenecen tienen un poder predictivo varias veces superior al azar (ver Tabla 3.5). Si bien no seria razonable esperar que nuestros modelos superen a otros que integran diversos tipos de información, llama la atención su buen desempeño.

Particularmente llamativo es que en las ontologías "Procesos biológicos" y "Componentes celulares" nuestros modelos tengan un mejor desempeño global que el método que asigna funciones directamente por similitud de secuencias. Es cierto que en el caso de la ontología de "Procesos biológicos" BLAST tiene un desempeño particularmente pobre. Una posible explicación para esto es que las secuencias asociadas a funciones representadas por la ontología "Procesos biológicos" no estén tan conservadas evolutivamente (Radivojac et al. 2013b; Jiang et al. 2016; Zhou et al. 2019). Los métodos basados en similitud de secuencia tienden a desempeñarse mejor al transferir anotaciones relacionadas a la bioquímica básica de las proteínas, que están mejor representadas en las otras dos ontologías. Aún así, nuestros modelos también superan a BLAST en la ontología "Componente celular".

La relevancia de este resultado está dada por el hecho de que, aún cuando es habitual que los modelos de predicción de función de genes integren múltiples tipos de información, la información derivada de la ubicación de los genes muy rara vez es tenida en cuenta. Más aún, según los organizadores de las CAFA, la inclusión de variables predictivas más variadas podría conducir a mejoras sustantivas en la predicción de función (Zhou et al. 2019). Teniendo esto en cuenta y a la luz de nuestros resultados, entendemos que la información derivada de la ubicación de los genes debería ser incluida rutinariamente en cualquier abordaje que integre distintos tipos de información para predecir función de genes.

La Figura 3.3 permite apreciar que en todos los organismos, salvo en *D. melanogaster*, la ontología en la que nuestros modelos se alejan más del azar, esto es, en la que logran hacer mejores predicciones, es la ontología de "Procesos biológicos". Este es un resultado llamativo, pues es en

esa ontología en la que todos los métodos logran sus peores desempeños en las competencias CAFA (Zhou et al. 2019). Nuestros resultados sugieren que la inclusión de las matrices de paisaje funcional en los métodos de predicción de función que integran múltiples fuentes de información sería particularmente beneficioso para hacer predicciones en la ontología más difícil de predecir.

Además de demostrar que es posible predecir nuevas funciones de genes exclusivamente a partir de su ubicación, nuestros resultados son relevantes desde otro punto de vista. La existencia en organismos eucariotas de patrones de distribución de grupos funcionales de genes tan bien definidos como para predecir nuevas funciones biológicas indica niveles de organización de esos genomas que se solía pensar eran exclusivos de los genomas procariotas, organizados en operones (Diament y Tuller 2016).

La razón entre el F-max del modelo entrenado y el F-max del modelo permutado para cada par organismo/ontología puede interpretarse como una estimación del nivel de organización lineal del genoma respecto al grafo de las ontologías. Si la organización lineal del genoma fuese totalmente independiente de las funciones biológicas que expresa, la ubicación de un gen en particular no sería informativa de sus posibles funciones biológicas y la probabilidad de que tuviese una en particular no cambiaría si cambiase su ubicación. Si así estuviesen organizados los genomas, haciendo predicciones al azar acertaríamos cierta cantidad de veces.

Aquí mostramos que, basándonos exclusivamente en información relativa a la organización lineal del genoma. acertamos varias veces esa cantidad. Según muestra la Figura 3.3, en este nivel de organización, los organismos más complejos tienen los genomas más organizados. Este resultado coincide con los obtenidos por Tuller y colaboradores que comentáramos en la discusión del Capítulo II. Estos investigadores encontraron que el grado de co-localización de términos GO distantes, al que llaman organización de segundo nivel del genoma, es mayor en un organismo ecuariota (*S. cerevisiae*) que en uno procariota (*E. coli*) (Tuller et al. 2009a) .

Esta organización de los genomas eucariotas respecto a las funciones de sus genes recientemente ha comenzado a documentarse también en el espacio tridimensional (Diament, Pinter, y Tuller 2014; Bonev y Cavalli 2016; Schwartz y Cavalli 2017). Si bien nuestros resultados demuestran que el análisis en una dimensión es capaz de capturar información acerca de esa organización, la inclusión en nuestro abordaje de las distancias en el espacio tridimensional es una perspectiva prometedora.

Conclusiones generales

Desde el punto de vista de la biología, el aporte del Capítulo I es un nuevo catálogo de genes potencialmente sinápticos en *D. melanogsater.* Este catálogo es una versión mejorada del original que, publicado hace 5 años, contenía a uno de cada tres genes sinápticos experimentalmente confirmados desde entonces. Desde el punto de vista del aprendizaje automático, para mejorar la predicción original proponemos una estrategia de entrenamiento que podría ser útil (como lo es en este caso) en otros problemas con muestras de entrenamiento pequeñas y sesgadas.

Los aportes del Capítulo II incluyen una nueva herramienta de análisis, Cluster Locator, que permite analizar estadísticamente la distribución de cualquier lista de genes en el genoma al que pertenece. Por su abordaje estadístico, la herramienta permite la comparación de los patrones de distribución en el genoma de listas de diferentes organismos. Utilizando Cluster Locator y aprendizaje automático no supervisado, demostramos la existencia de grupos de genes con la misma función que tienen patrones de distribución similares en diferentes organismos, así como la existencia, en un mismo organismo, de grupos de genes con funciones similares y patrones de distribución similares.

El Capítulo III presenta una serie de modelos basados en aprendizaje automático supervisado que predicen nuevas asociaciones entre genes y términos GO en 5 organismos modelo. Lo más novedoso aquí es que los modelos fueron entrenados con variables predictivas exclusivamente derivadas de la ubicación de cada gen en el genoma al que pertenece, y que nuestros resultados indican que esto es relevante desde el punto de vista biológico. Los "mapas de enriquecimiento funcional local" que elaboramos para entrenar esos modelos tienen interés en sí mismos y están disponibles aquí: github.com/IIBCE-BND/gfpml-datasets. Según métricas estándar de evaluación, en dos de las tres ontologías biológicas nuestros modelos hacen mejores predicciones que un método de referencia basado exclusivamente en la similitud de secuencias. Todas las predicciones que resultan de nuestros modelos están disponibles aquí: gfpml.bnd.edu.uy

Métodos

Métodos Generales

- Aprendizaje automático supervisado.
- Análisis de enriquecimiento.
- Gene Ontology.
- Python - pandas - scikit-learn.

Métodos específicos al Capítulo I

- Datos usados para la predicción de función sináptica.
- Construcción de la muestra de entrenamiento original.
- Nuevo esquema de entrenamiento.
- Hiperparámetros de los algoritmos.
- Evaluación de los clasificadores.

Métodos específicos al Capítulo II

- Cluster Locator.
- Medida de la concentración de un sub-grafo.
- Búsqueda de una partición de la ontología que maximice la concentración y minimice la varianza de los perfiles de agrupamiento.

Métodos específicos al Capítulo III

- Análisis de enriquecimiento funcional local.
- Entrenamiento de los modelos y asignación de probabilidades.

Métodos generales

Aprendizaje automático supervisado (Hastie, Tibshirani, y Friedman 2009)

Podemos representar un conjunto de datos como un conjunto de vectores en el que cada vector está asociado a una clase conocida

$(X_1 , Y_1), ..., (X_n , Y_n) /$

$X_i = (X_{i1} , ..., X_{ir}) \in X \subset R\ r$

Y_i toma valores en $Y = \{0, 1\}$ para $(i = 1, ..., n)$.

Al conjunto de datos de clase conocida se lo conoce como muestra de entrenamiento. El objetivo del aprendizaje automático supervisado es encontrar una función

$f : X \rightarrow Y$

que sirva como regla de clasificación, tal que, al observar una nueva X, se pueda predecir Y con f (X). La tasa de error del clasificador f es:

$P\{f (X) \neq Y \}$

que se puede estimar clasificando los casos conocidos y contando los errores

En general, el problema es encontrar

$f^* : R^r \rightarrow \{0, 1\} /$

$P\{f^* (X) \neq Y \} = \text{mín}_{f :R\ r \rightarrow \{0,1\}} P\{f (X) \neq Y \}$

esto es, encontrar la regla de clasificación con la menor tasa de error (aunque en la práctica, la mínima tasa de error puede indicar un sobreajuste a los datos de entrenamiento).

Análisis de enriquecimiento (Boyle et al. 2004)

El análisis de enriquecimiento se puede utilizar para evaluar estadísticamente la sobrerrepresentación de cierta característica en una lista de elementos. Si llamamos:

- N al número de elementos en la lista de referencia, esto es, el conjunto de elementos del cual se extrajo la lista que se está analizando,
- M al número de elementos de la lista de referencia que presentan la característica de interés,
- n al número de elementos en la lista analizada y

- k al número de elementos en la lista analizada asociados a la característica de interés,

El enriquecimiento de la lista de analizada en la característica de interés se define como:

((k/n) / (M/N))

Asumiendo una distribución hipergeométrica, a este enriquecimiento se le puede asociar un p-valor que se define como:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i}\binom{N-M}{n-i}}{\binom{N}{i}}.$$

Una variante habitual de este método es el análisis de enriquecimiento funcional, en el que se evalúa si en una lista de genes hay funciones biológicas que estén sobrerrepresentadas. En un ejemplo típico, se tiene una lista de genes como resultado de algún experimento y se quiere saber si la lista contiene más genes asociados a cierta función biológica de los que cabría esperar por azar. Encontrar que la lista analizada está enriquecida en cierta función biológica ayuda muchas veces a interpretar el resultado del experimento que dio lugar a esa lista.

Gene Ontology (Ashburner et al. 2000)

La "ontología génica" ("Gene Ontology" en inglés, en adelante GO) busca ser una representación computacional de la totalidad del conocimiento científico actual acerca de las funciones de los genes. Fue creada y es mantenida y actualizada por un consorcio de científicos expertos en tres organismos modelo (*Drosophila melanogaster, Mus musculus* y *Saccharomyces cerevisiae*), que en 1998 acordaron comenzar trabajar en una clasificación común para las funciones de los genes. Al día de hoy son miles los organismos incluidos en GO, lo cual permite describir las funciones de los genes de forma comparable a lo largo de todo el espectro filogenético. GO es ampliamente usada y ha sido citada en decenas de miles de publicaciones científicas («Gene Ontology Resource» s. f.).

GO tiene dos componentes:

- La ontología en sí misma; un conjunto de términos con definición precisa, organizado en tres grafos acíclicos direccionados. Estos tres grafos pretenden capturar tres aspectos necesarios para

caracterizar cualquier posible función de un gen: el mecanismo molecular involucrado, el proceso biológico del que forma parte y la localización celular en la que tiene lugar. Cada término de la ontología, o "término GO", representa un nodo en uno de los grafos, cuyas aristas representan las diversas maneras en las que los términos se relacionan entre sí.

- El conjunto de "anotaciones GO"; todas las asociaciones que se han establecido entre un término GO y un producto de la expresión de un gen. Cada anotación incluye una referencia rastreable a su origen y una indicación del tipo de evidencia en la que se basa.

La versión actual de GO lleva incorporados hallazgos experimentales de más de 150.000 artículos científicos, que se traducen en unas 700,000 anotaciones con evidencia experimental, que dan a su vez lugar a más de 6 millones de anotaciones inferidas por distintos métodos computacionales.

Python, NumPy, Pandas, Scikit-learn y Jupyter

Python es un lenguaje de programación interpretado multiparadigma. NumPy es una extensión de Python, que le agrega mayor soporte para vectores y matrices. Pandas es una extensión de NumPy que ofrece estructuras de datos y operaciones para manipular tablas numéricas y series temporales. Scikit-learn (Pedregosa et al. 2011) es una biblioteca de software de aprendizaje automático para el lenguaje de programación Python.

El Proyecto Jupyter es una organización sin ánimo de lucro creada para "desarrollar software de código abierto, estándares abiertos y servicios para computación interactiva en docenas de lenguajes de programación". Un documento de Jupyter Notebook es un documento JSON, que sigue un esquema versionado y que contiene una lista ordenada de celdas de entrada/salida que pueden contener código, texto gráficos y texto enriquecidos, generalmente terminado con la extensión ".ipynb".

Métodos específicos al Capítulo I

Datos usados para la predicción de función sináptica

Para la predicción de función sináptica en genes de *Drosophila melanogaster* utilizamos datos publicados por el consorcio modENCODE en el año 2011 (Graveley et al. 2011). Los datos recogen los niveles de transcripción de todos los GCP en 24 momentos del desarrollo del organismo. En este conjunto de datos cada muestra consiste en ARN poliAAA aislado de 30 organismos completos. El conjunto de datos original comprende los niveles de transcripción de 15.398 genes, expresados en fragmentos por kilo base de exón por millón de fragmentos mapeados (FPKM). De ese conjunto original excluimos 1.756 genes que solo exhiben niveles de transcripción por encima de cero en etapas adultas pero que no se transcriben durante el desarrollo, la etapa cuando se forman las neuronas con sus sinapsis (en el embrión, la larva y la pupa). El perfil temporal de transcripción de cada gen fue normalizado, dividiendo cada valor de transcripción de la serie temporal por el máximo valor alcanzado por ese gen, obteniendo así para cada gen una serie de 24 valores que oscilan entre 0 y 1 (Pazos Obregón et al. 2015).

Construcción de la muestra de entrenamiento original (Pazos Obregón et al. 2015)

Para obtener nuestra predicción original entrenamos 3 algoritmos de aprendizaje con una muestra de entrenamiento desbalanceada, en la que había muchos más ejemplos negativos que positivos. Los 92 casos positivos fueron seleccionados luego de una revisión bibliográfica exhaustiva, buscando genes cuya función en la formación, maduración, plasticidad o mantenimiento de la sinapsis o la neurotransmisión tuviese evidencia experimental firme. Como casos negativos seleccionamos 397 genes que cumplían con al menos uno de dos criterios biológicos definidos *ad hoc:*

- Niveles de transcripción extremadamente diferentes entre machos y hembras (ya que esperamos que los genes sinápticos sean esencialmente los mismos en ambos sexos).
- Muy bajo nivel relativo de transcripción en el sistema nervioso central en momentos conocidos de sinaptogénesis masiva en *Drosophila*.

Nuevo esquema de entrenamiento

El nuevo esquema de entrenamiento propone definir 5 muestras de entrenamiento más pequeñas y parcialmente distintas entre sí a partir de la muestra de entrenamiento original. Para construir cada una de esas muestras de entrenamiento más pequeñas, se divide en quintos a los ejemplos positivos y negativos y se deja fuera un quinto diferente cada vez. Con el quinto de ejemplos positivos y negativos que se dejó fuera al definir cada muestra de entrenamiento, se define una muestra de prueba complementaria.

Con cada muestra de entrenamiento se entrena a los tres algoritmos: kNN, SVM y Random Forest, y se obtienen 15 clasificadores. Los hiperparámetros de cada clasificador se determinaron por búsqueda en grilla combinada con validación cruzada décuple sobre la propia muestra de entrenamiento. Cada muestra de prueba se usó para evaluar independientemente al clasificador que se entrenó con la muestra de entrenamiento complementaria, determinando su precisión, el área bajo su curva ROC y su F1.

Cada clasificador asignó a cada gen una probabilidad diferente de ser sináptico. Para obtener nuestros catálogos consideramos, para cada posible umbral de clasificación, la intersección de los 15 resultados, esto es, el conjunto de genes a los cuales los 15 clasificadores habían asignado una probabilidad de ser sináptico por encima del umbral. Para cada gen en la intersección tomamos el promedio de todas las probabilidades asignadas.

Hiperparámetros de los algoritmos

En cada caso, los hiperparámetros de cada clasificador se fijaron mediante un método estándar: búsqueda en grilla combinada con validación cruzada décuple.

Evaluación de los clasificadores («Receiver Operating Characteristic» 2020)

Los clasificadores fueron evaluados mediante tres métricas de performance estándar: la exactitud, el *score* F1 y área bajo la curva ROC. Todas se definen a partir de la cantidad y del tipo

de aciertos y errores que comete el modelo al clasificar una muestra de prueba conformada por ejemplos de clase conocida. Si llamamos

VP = a los verdaderos positivos;

FP = a los falsos positivos;

VN = a los verdaderos negativos;

FN = a los falsos negativos;

la exactitud se define como:

Exactitud = (TP+TN) / (TP+TN+FP+FN)

y mide la proporción entre los aciertos y el total de casos clasificados por el clasificador.

El *score* F1se define como:

F1 = 2 * (Precisión * Sensibilidad) / (Precisión + Sensibilidad)

que es el promedio harmónico de la Precisión y la Sensibilidad del clasificador

La Precisión es la proporción de los casos que fueron predichos como positivos que realmente eran positivos y se define como:

Precisión = VP / (VP + FP)

La Sensibilidad es la proporción de los casos verdaderamente positivos que fueron clasificados como positivos y se define como:

Sensibilidad = VP / (VP + FN)

La curva ROC (del inglés R*eceiver Operating Characteristic*) es una representación gráfica de la performance de un clasificador según se varía su umbral de clasificación. Se construye graficando la sensibilidad (la tasa de verdaderos positivos) en función del inverso de la especificidad (la tasa de verdaderos negativos) para varios umbrales de clasificación. El área bajo la curva ROC se interpreta como la probabilidad de que el clasificador le asigne a un caso verdaderamente positivo y tomado el azar, una probabilidad de pertenecer a la clase positiva mayor que la que le asigne a un caso verdaderamente negativo, también tomado al azar. El área bajo la curva ROC de un clasificador perfecto es igual a 1, y la de un clasificador que clasifica al azar es igual a 0,5.

Métodos específicos al Capítulo II

Cluster Locator

Cluster Locator es una aplicación web cuyo *backend* está implementado en Python 2.7 y se despliega en AWS Lambda. El *frontend* utiliza ReactJS y librerías D3js y sus archivos estáticos son almacenados en AWS S3 Storage. La programación e implementación de la herramienta estuvieron a cargo del Bach. Pablo Soto, integrante de nuestro equipo de investigación.

Medida de la concentración de un sub-grafo (*)

Con el fin de medir la cercanía en la ontología de un conjunto de términos GO dado, definimos una medida de concentración que se detalla a continuación.

Dado un grafo **G** y un subconjunto de sus nodos **H** $\subset$ **G**, de tamaño *k*, queremos saber si la distancia promedio entre los nodos de **H** es significativamente diferente a la distancia que se espera encontrar por azar en subconjuntos aleatorios de tamaño *k* tomados de G.

Si llamamos **M(H)** a la distancia promedio entre los nodos que pertenecen a **H**,

$$M(H) := \frac{1}{C_2^k} \sum_{\substack{v,w \in H \\ v \neq w}} d(v,w)$$

Para un tamaño dado, la distancia promedio entre los nodos de **G** es igual al promedio de las distancias promedio de todos los posibles subconjuntos de **G**. Esto implica que dado un grafo **G** de tamaño *n* y $k \leq n$, se tiene que:

$$\frac{1}{C_2^n} \sum_{\substack{v,w \in G \\ v \neq w}} d(v,w) = \frac{1}{C_k^n} \sum_{\substack{H \subset G \\ |H|=k}} \left(\frac{1}{C_2^k} \sum_{\substack{v,w \in H \\ v \neq w}} d(v,w) \right)$$

Por lo tanto, dado un conjunto H de términos GO de tamaño *k* primero calculamos M(H) , es decir, la distancia promedio entre todos los posibles pares de nodos de H y la comparamos con M(G). Para evaluar la significación estadística de la distancia entre M(H) y M(G) necesitamos saber

el desvío estándar asociado a la distancia promedio esperada en sub-grafos aleatorios de tamaño k . Estimamos este desvío estándar (sd_k) generando 1000 sub-grafos aleatorios de tamaño k y calculando el promedio de sus distancias promedio y el desvío estándar asociado.

A continuación definimos la concentración de H como:

$$C(H) = M(G) – M(H) / sd_k$$

Aquí debemos hacer una observación: aun si consideráramos el total de anotaciones en cualquiera de los 5 organismos que estamos analizando, las mismas abarcarían solamente una parte del total de términos GO y además sólo consideramos aquellos términos GO asociados con al menos 20 y con no más de 1.000 genes (ver Tabla 1). Por esta razón, para cada uno de los 5 organismos y cada una de las 3 ontologías, el conjunto de perfiles de agrupamiento del que disponemos es un subconjunto del total de términos GO. Ese subconjunto de nodos conforma varios componentes no conexos dentro del grafo total de la ontología. Como este hecho impide determinar todas las distancias entre los términos GO analizados, para medir la distancia entre nodos utilizamos el grafo completo de la ontología.

Búsqueda de una partición de la ontología que maximice la concentración y minimice la varianza de los perfiles de agrupamiento

El trabajo aquí expuesto fue desarrollado conjuntamente con el Lic. Diego Silvera, estudiante de la maestría en Ingeniería Matemática de la Facultad de Ingeniería de la UDELAR, a quien co-dirijo con el Dr. Gustavo Guerberoff. Con el fin de encontrar una manera de dividir el grafo total de la ontología en un conjunto de sub-grafos cuyos nodos estén concentrados y cuyos respectivos perfiles de agrupamiento sean parecidos, desarrollamos el procedimiento que se explica a continuación. Es deseable además que estos grupos de términos GO sean lo más grandes posible. El problema es entonces buscar una partición de la ontología que maximice la concentración y la cantidad de nodos de cada elemento de la partición y minimice la varianza entre los perfiles de agrupamiento correspondientes.

Con este propósito definimos la siguiente función, con la cual podemos asignar un *score* a cada elemento de la partición:

$$\textbf{Score}(H) = \log |H| * C(H) / \text{var } |H|$$

Donde var (H) denota la varianza entre los perfiles de agrupamiento correspondientes a los nodos de H.

Utilizando un algoritmo de *clustering* aglomerativo (*Agglomerative Clustering*) construimos un dendograma basado en el grafo de la ontología. Optamos por el *clustering* aglomerativo porque forma los grupos teniendo en cuenta la estructura del grafo. Por dendograma basado en un grafo G entendemos un árbol binario con raíz G, en el que cada nodo es un sub-grafo de G y el conjunto que contiene a los hijos de cualquier nodo H es una partición de H. En el caso que nos ocupa, cada nodo del dendograma es un conjunto de términos GO. El problema ahora es seleccionar un subconjunto de los nodos del dendograma tal que cada nodo incluya la mayor cantidad posible de términos GO, que los mismos estén concentrados en la ontología y que tengan perfiles de agrupamiento parecidos.

Para ello seguimos el procedimiento que se detalla a continuación:

- Con la función definida más arriba, asignamos un score a cada nodo del dendograma
- Eliminamos todos los nodos poco concentrados o muy pequeños, aplicando dos restricciones: $C(H) > 2$ y $|H| > 5$
- De entre los nodos restantes, seleccionamos el nodo H con el mayor score
- Continuamos seleccionando nodos que no sean hijos de algún nodo ya seleccionado hasta que no queden nodos seleccionables

Nótese que no necesariamente todos los términos GO considerados (aquellos asociados con entre 20 y 1.000 genes) habrán sido seleccionados tras aplicar este procedimiento.

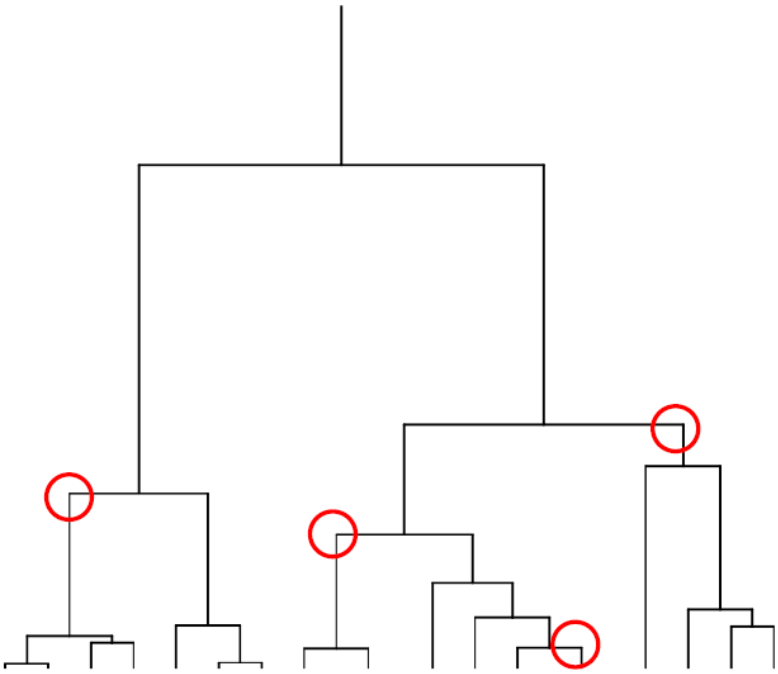


Figura 11 *– Procedimiento para seleccionar una partición de la ontología*

Métodos específicos al Capítulo III

Análisis de enriquecimiento funcional local

Llamamos análisis de enriquecimiento funcional local (LEA, por Local Enrichment Analysis) al análisis de la sobrepresentación de funciones biológicas en el entorno de cierto gen. Como se detalla en la sección "Métodos generales", una variante habitual del análisis de sobrerepresentación es el análisis de enriquecimiento funcional, que evalúa si en una lista de genes hay funciones biológicas que estén sobrerrepresentadas. Aquí implementamos este tipo de análisis con un enfoque particular: la lista de genes que se evalúa es la lista de los genes que se encuentran a cierta distancia a un lado y otro del gen de interés, distancia que según nuestro modelo del genoma se mide en cantidad de genes.

Si denotamos como **X** a cierto gen y como **r** a cierto tamaño de ventana, calculamos el enriquecimiento en términos GO de la lista de genes comprendida entre **X - r** y **X + r.** Haciendo variar **X** y **r** se puede recorrer el genoma completo y asociar a cada gen una matriz de enriquecimientos, conformada por los enriquecimientos en cada término GO y para cada tamaño de ventana utilizado, en el entorno del gen de interés.

Entrenamiento de los modelos y asignación de probabilidades

El entrenamiento de los modelos y la asignación de probabilidades, tareas con alto costo computacional, fueron realizadas por Diego Silvera en ClusterUY (https://cluster.uy).

Colaboraciones

El trabajo que presento en esta tesis tiene un fuerte carácter interdisciplinario. Abordar las preguntas planteadas requirió herramientas de la genómica y la evolución, de la estadística y del aprendizaje automático, de la teoría de grafos, de la programación y del manejo de datos. Por ello, además de la heterogénea formación de mis tres orientadores, fue necesario colaborar con especialistas e investigadores de diversas áreas del conocimiento. Lo que sigue es una lista de las personas con las que he trabajado a lo largo de este proyecto.

- Rafael Cantera, mi orientador principal, es doctor en biología y especialista en el estudio del desarrollo del sistema nervioso.

- Patricio Yankilevich, uno de mis co-orientadores, es doctor en computación y especialista en bioinformática aplicada a problemas genómicos..

- Gustavo Guerberoff, mi otro co-orientador, es doctor en física y especialista en modelado probabilistico.

- Pablo Soto es programador en el sector privado, especialista en aprendizaje automático.

- Diego Silvera es licenciado en matemáticas y estudiante de la maestría en ingeniaría matemática (orientado por el Dr. Guerberoff y co-orientado por mi).

- Martín Palazzo es magíster en ingeniería de sistemas complejos y estudiante de doctorado del Dr. Yankilevich.

- Rosa Barrio es doctora en biología y especialista en genómica funcional.

- Ana María Aransay es doctora en biología y especialista en genética de poblaciones y genómica.

- José Luis Lavín es doctor en biología y especialista en bioinformática.

- Ana Rosa Cortázar es magíster en bioinformática y estadística.

Financiamiento

El financiador principal de esta tesis de doctorado fue el Instituto de Investigaciones Biológicas Clemente Estable (IIBCE). Durante el período de tiempo en el que llevé adelante el trabajo que se reúne aquí usufructué de un contrato como "Investigador nivel II", bajo el cual mi obligación principal fue justamente llevar adelante esta tesis. Conté además con el apoyo de la Agencia Nacional de Investigación e Innovación (ANII), que me otorgó una beca de doctorado que usufructué entre julio de 2017 y julio de 2020. Además, entre julio de 2018 y julio de 2019 la ANII, financió un proyecto de investigación del cual fui responsable científico, parte de cuyos resultados se reúnen en el Capítulo III. Finalmente, algunas actividades pudieron ser financiadas con las alícuotas de estudiante que recibí del Programa para el Desarrollo de las Ciencias Básicas (PEDECIBA).

Agradecimientos

A mis orientadores, particularmente a Rafael.
A mis compañeros de laboratorio.
A mi madre y a mi padre.
Y sobretodo, a Lucía.

Referencias

Aboukhalil, Robert, Bernard Fendler, y Gurinder S. Atwal. 2013. «Kerfuffle: A Web Tool for Multi-Species Gene Colocalization Analysis». *BMC Bioinformatics* 14: 22. https://doi.org/10.1186/1471-2105-14-22.

Alenezi, Hadeel S., y Maha H. Faisal. 2020. «Utilizing Crowdsourcing and Machine Learning in Education: Literature Review». *Education and Information Technologies*. https://doi.org/10.1007/s10639-020-10102-w.

Altman, N. S. 1992. «An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression». *The American Statistician* 46 (3): 175-85. https://doi.org/10.2307/2685209.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers, y D. J. Lipman. 1990. «Basic Local Alignment Search Tool». *Journal of Molecular Biology* 215 (3): 403-10. https://doi.org/10.1016/S0022-2836(05)80360-2.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, et al. 2000. «Gene Ontology: Tool for the Unification of Biology. The Gene Ontology Consortium.» *Nature Genetics* 25 (1): 25-29. https://doi.org/10.1038/75556.

Barua, Limon, Bo Zou, y Yan Zhou. 2020. «Machine Learning for International Freight Transportation Management: A Comprehensive Review». *Research in Transportation Business & Management*, febrero, 100453. https://doi.org/10.1016/j.rtbm.2020.100453.

Barutcuoglu, Zafer, Robert E. Schapire, y Olga G. Troyanskaya. 2006. «Hierarchical Multi-Label Prediction of Gene Function». *Bioinformatics* 22 (7): 830-36. https://doi.org/10.1093/bioinformatics/btk048.

Bernardes, Juliana S., y Carlos E. Pedreira. 2013. «A Review of Protein Function Prediction under Machine Learning Perspective». *Recent Patents on Biotechnology* 7 (2): 122-41. https://doi.org/10.2174/18722083113079990006.

Blumenthal, Thomas, Donald Evans, Christopher D. Link, Alessandro Guffanti, Daniel Lawson, Jean Thierry-Mieg, Danielle Thierry-Mieg, et al. 2002. «A Global Analysis of Caenorhabditis Elegans Operons». *Nature* 417 (6891): 851-54. https://doi.org/10.1038/nature00831.

Bonetta, Rosalin, y Gianluca Valentino. 2020. «Machine Learning Techniques for Protein Function Prediction». *Proteins* 88 (3): 397-413. https://doi.org/10.1002/prot.25832.

Bonev, Boyan, y Giacomo Cavalli. 2016. «Organization and Function of the 3D Genome». *Nature Reviews Genetics* 17 (11): 661-78. https://doi.org/10.1038/nrg.2016.112.

Boutanaev, Alexander M., Alla I. Kalmykova, Yuri Y. Shevelyov, y Dmitry I. Nurminsky. 2002. «Large Clusters of Co-Expressed Genes in the Drosophila Genome». *Nature* 420 (6916): 666-69. https://doi.org/10.1038/nature01216.

Boyle, Elizabeth I., Shuai Weng, Jeremy Gollub, Heng Jin, David Botstein, J. Michael Cherry, y Gavin Sherlock. 2004. «GO::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes». *Bioinformatics (Oxford, England)* 20 (18): 3710-15. https://doi.org/10.1093/bioinformatics/bth456.

Breiman, Leo. 2001. «Random Forests». *Machine Learning* 45 (1): 5-32. https://doi.org/10.1023/A:1010933404324.

Brunton, Steven L., Bernd R. Noack, y Petros Koumoutsakos. 2020. «Machine Learning for Fluid Mechanics». *Annual Review of Fluid Mechanics* 52 (1): 477-508. https://doi.org/10.1146/annurev-fluid-010719-060214.

Burkhardt, Pawel. 2015. «The origin and evolution of synaptic proteins – choanoflagellates lead the way». *The Journal of Experimental Biology* 218 (4): 506. https://doi.org/10.1242/jeb.110247.

Cantera, Rafael, María José Ferreiro, Ana María Aransay, y Rosa Barrio. 2014. «Global Gene Expression Shift during the Transition from Early Neural Development to Late Neuronal Differentiation in Drosophila Melanogaster». *PloS One* 9 (5): e97703. https://doi.org/10.1371/journal.pone.0097703.

Caron, H., B. van Schaik, M. van der Mee, F. Baas, G. Riggins, P. van Sluis, M. C. Hermus, et al. 2001. «The Human Transcriptome Map: Clustering of Highly Expressed Genes in Chromosomal Domains». *Science (New York, N.Y.)* 291 (5507): 1289-92. https://doi.org/10.1126/science.1056794.

Caruana, Rich, y Alexandru Niculescu-Mizil. 2006. «An empirical comparison of supervised learning algorithms». En , 161-68. Pittsburgh, Pennsylvania: ACM.

Chawla, N. V., K. W. Bowyer, L. O. Hall, y W. P. Kegelmeyer. 2002. «SMOTE: Synthetic Minority Over-Sampling Technique». *Journal of Artificial Intelligence Research* 16 (junio): 321-57. https://doi.org/10.1613/jair.953.

Cohen, B. A., R. D. Mitra, J. D. Hughes, y G. M. Church. 2000. «A Computational Analysis of Whole-Genome Expression Data Reveals Chromosomal Domains of Gene Expression». *Nature Genetics* 26 (2): 183-86. https://doi.org/10.1038/79896.

Corrales-Berjano, Marc, Aranzazu Rosado Diez, Ruggero Cortini, Joris van Arensbergen, Bas van Steensel, y Guillaume J. Filion. 2017. «Clustering of Drosophila Housekeeping Promoters Facilitates Their Expression». *Genome Research*, abril, gr.211433.116. https://doi.org/10.1101/gr.211433.116.

Dávila López, Marcela, Juan José Martínez Guerra, y Tore Samuelsson. 2010. «Analysis of Gene Order Conservation in Eukaryotes Identifies Transcriptionally and Functionally Linked Genes». *PloS One* 5 (5): e10654. https://doi.org/10.1371/journal.pone.0010654.

De, Subhajyoti, y M. Madan Babu. 2010. «Genomic Neighbourhood and the Regulation of Gene Expression». *Current Opinion in Cell Biology* 22 (3): 326-33. https://doi.org/10.1016/j.ceb.2010.04.004.

Dekker, Job, Karsten Rippe, Martijn Dekker, y Nancy Kleckner. 2002. «Capturing Chromosome Conformation». *Science* 295 (5558): 1306-11. https://doi.org/10.1126/science.1067799.

Diament, Alon, Ron Y. Pinter, y Tamir Tuller. 2014. «Three-Dimensional Eukaryotic Genomic Organization Is Strongly Correlated with Codon Usage Expression and Function». *Nature Communications* 5 (diciembre): 5876. https://doi.org/10.1038/ncomms6876.

Diament, Alon, y Tamir Tuller. 2016. «Three-Dimensional Genomic Organization of Genes' Function in Eukaryotes». En *Evolutionary Biology: Convergent Evolution, Evolution of Complex Traits, Concepts and Methods*, editado por Pierre Pontarotti, 233-52. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-41324-2_14.

Dietterich, Thomas G. 2000. «Ensemble Methods in Machine Learning». En *Multiple Classifier Systems*, 1-15. Lecture Notes in Computer Science. Springer Berlin Heidelberg.

Dottorini, Tania, Pietro Palladino, Nicola Senin, Tania Persampieri, Roberta Spaccapelo, y Andrea Crisanti. 2013. «CluGene: A Bioinformatics Framework for the Identification of Co-Localized, Co-Expressed and Co-Regulated Genes Aimed at the Investigation of Transcriptional Regulatory Networks from High-Throughput Expression Data». *PloS One* 8 (6): e66196. https://doi.org/10.1371/journal.pone.0066196.

Dutartre, Leslie, Frédérique Hilliou, y René Feyereisen. 2012. «Phylogenomics of the Benzoxazinoid Biosynthetic Pathway of Poaceae: Gene Duplications and Origin of the Bx Cluster». *BMC Evolutionary Biology* 12 (mayo): 64. https://doi.org/10.1186/1471-2148-12-64.

Eden, Eran, Roy Navon, Israel Steinfeld, Doron Lipson, y Zohar Yakhini. 2009. «GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists.» *BMC Bioinformatics* 10. https://doi.org/10.1186/1471-2105-10-48.

Eisen, Michael B., Paul T. Spellman, Patrick O. Brown, y David Botstein. 1998. «Cluster analysis and display of genome-wide expression patterns». *Proceedings of the National Academy of Sciences* 95 (25): 14863-68.

Elamrani Abou Elassad, Zouhair, Hajar Mousannif, Hassan Al Moatassime, y Aimad Karkouch. 2020. «The Application of Machine Learning Techniques for Driving Behavior Analysis: A Conceptual Framework and a Systematic Literature Review». *Engineering Applications of Artificial Intelligence* 87 (enero): 103312. https://doi.org/10.1016/j.engappai.2019.103312.

Fathima, Amreen, y K. Vaidehi. 2020. «Review on Facial Expression Recognition System Using Machine Learning Techniques». En *Advances in Decision Sciences, Image Processing, Security and Computer Vision*, editado por Suresh Chandra Satapathy, K. Srujan Raju, K. Shyamala, D. Rama Krishna, y Margarita N. Favorskaya, 608-18. Learning and Analytics in Intelligent Systems. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-24318-0_70.

Feng, Shou, Ping Fu, y Wenbin Zheng. 2017. «A Hierarchical Multi-Label Classification Algorithm for Gene Function Prediction». *Algorithms* 10 (4): 138. https://doi.org/10.3390/a10040138.

———. 2018. «A hierarchical multi-label classification method based on neural networks for gene function prediction». *Biotechnology & Biotechnological Equipment* 32 (6): 1613-21. https://doi.org/10.1080/13102818.2018.1521302.

Fernández-Delgado, Manuel, Eva Cernadas, Senén Barro, y Dinani Amorim. 2014. «Do we Need Hundreds of Classifiers to Solve Real World Classification Problems?» *Journal of Machine Learning Research* 15: 3133-81.

Feuerborn, Alexander, y Peter R. Cook. 2015. «Why the Activity of a Gene Depends on Its Neighbors». *Trends in Genetics: TIG* 31 (9): 483-90. https://doi.org/10.1016/j.tig.2015.07.001.

Frank, C. Andrew, Xinnan Wang, Catherine A. Collins, Avital A. Rodal, Quan Yuan, Patrik Verstreken, y Dion K. Dickman. 2013. «New Approaches for Studying Synaptic Development, Function, and Plasticity Using Drosophila as a Model System.» *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience* 33 (45): 17560-68. https://doi.org/10.1523/JNEUROSCI.3261-13.2013.

Fukuoka, Yutaka, Hidenori Inaoka, y Isaac S. Kohane. 2004. «Inter-Species Differences of Co-Expression of Neighboring Genes in Eukaryotic Genomes». *BMC Genomics* 5 (1): 4. https://doi.org/10.1186/1471-2164-5-4.

Gangavarapu, Tushaar, C. D. Jaidhar, y Bhabesh Chanduka. 2020. «Applicability of Machine Learning in Spam and Phishing Email Filtering: Review and Approaches». *Artificial Intelligence Review*, febrero. https://doi.org/10.1007/s10462-020-09814-9.

«Gene Ontology Resource». s. f. Gene Ontology Resource. Accedido 4 de marzo de 2020. http://geneontology.org/.

Ghanbarian, Avazeh T., y Laurence D. Hurst. 2015. «Neighboring Genes Show Correlated Evolution in Gene Expression». *Molecular Biology and Evolution* 32 (7): 1748-66. https://doi.org/10.1093/molbev/msv053.

Graveley, Brenton R., Angela N. Brooks, Joseph W. Carlson, Michael O. Duff, Jane M. Landolin, Li Yang, Carlo G. Artieri, et al. 2011. «The developmental transcriptome of Drosophila melanogaster». *Nature* 471 (7339): 473-79. https://doi.org/10.1038/nature09715.

Graves, Jordan, Jacob Byerly, Eduardo Priego, Naren Makkapati, S. Vince Parish, Brenda Medellin, y Monica Berrondo. 2020. «A Review of Deep Learning Methods for Antibodies». *Antibodies (Basel, Switzerland)* 9 (2). https://doi.org/10.3390/antib9020012.

Hartwell, Leland H., Joseph Culotti, y Brian Reid. 1970. «Genetic Control of the Cell-Division Cycle in Yeast, I. Detection of Mutants». *Proceedings of the National Academy of Sciences* 66 (2): 352-59. https://doi.org/10.1073/pnas.66.2.352.

Hastie, Trevor, Robert Tibshirani, y J. H Friedman. 2009. *The Elements of Statistical Learning Data Mining, Inference, and Prediction*. New York: Springer.

Hurst, Laurence D., Csaba Pal, y Martin J. Lercher. 2004. «The evolutionary dynamics of eukaryotic gene order». *Nat Rev Genet* 5 (4): 299-310. https://doi.org/10.1038/nrg1319.

Hurst, Laurence D., Elizabeth J. B. Williams, y Csaba Pál. 2002. «Natural Selection Promotes the Conservation of Linkage of Co-Expressed Genes». *Trends in Genetics: TIG* 18 (12): 604-6. https://doi.org/10.1016/s0168-9525(02)02813-5.

Huynen, Martijn, Berend Snel, Warren Lathe, y Peer Bork. 2000. «Predicting Protein Function by Genomic Context: Quantitative Evaluation and Qualitative Inferences». *Genome Research* 10 (8): 1204-10.

Jacob, F., y J. Monod. 1961. «Genetic Regulatory Mechanisms in the Synthesis of Proteins». *Journal of Molecular Biology* 3 (junio): 318-56. https://doi.org/10.1016/s0022-2836(61)80072-7.

Janga, Sarath Chandra, Julio Collado-Vides, y Gabriel Moreno-Hagelsieb. 2005. «Nebulon: A System for the Inference of Functional Relationships of Gene Products from the Rearrangement of Predicted Operons». *Nucleic Acids Research* 33 (8): 2521-30. https://doi.org/10.1093/nar/gki545.

Jiang, Yuxiang, Tal Ronnen Oron, Wyatt T. Clark, Asma R. Bankapur, Daniel D'Andrea, Rosalba Lepore, Christopher S. Funk, et al. 2016. «An expanded evaluation of protein function prediction methods shows an improvement in accuracy». *Genome Biology* 17 (septiembre): 184. https://doi.org/10.1186/s13059-016-1037-6.

Kacsoh, Balint Z., Casey S. Greene, y Giovanni Bosco. 2017. «Machine Learning Analysis Identifies Drosophila Grunge/Atrophin as an Important Learning and Memory Gene Required for Memory Retention and Social Learning». *G3: Genes|Genomes|Genetics* 7 (11): 3705-18. https://doi.org/10.1534/g3.117.300172.

Karathia, Hiren, Carl Kingsford, Michelle Girvan, y Sridhar Hannenhalli. 2016. «A Pathway-Centric View of Spatial Proximity in the 3D Nucleome across Cell Lines». *Scientific Reports* 6 (1): 39279. https://doi.org/10.1038/srep39279.

Kerepesi, Csaba, Bálint Daróczy, Ádám Sturm, Tibor Vellai, y András Benczúr. 2018. «Prediction and Characterization of Human Ageing-Related Proteins by Using Machine Learning». *Scientific Reports* 8 (1): 4094. https://doi.org/10.1038/s41598-018-22240-w.

Larkoski, Andrew J., Ian Moult, y Benjamin Nachman. 2020. «Jet Substructure at the Large Hadron Collider: A Review of Recent Advances in Theory and Machine Learning». *Physics Reports*, Jet substructure at the Large Hadron Collider: A review of recent advances in theory and machine learning, 841 (enero): 1-63. https://doi.org/10.1016/j.physrep.2019.11.001.

Laßek, Melanie, Jens Weingarten, y Walter Volknandt. 2015. «The synaptic proteome». *Cell and Tissue Research* 359 (1): 255-65. https://doi.org/10.1007/s00441-014-1943-4.

Lee, Jennifer M., y Erik L. L. Sonnhammer. 2003. «Genomic Gene Clustering Analysis of Pathways in Eukaryotes». *Genome Research* 13 (5): 875-82. https://doi.org/10.1101/gr.737703.

Lercher, Martin J., Thomas Blumenthal, y Laurence D. Hurst. 2003. «Coexpression of Neighboring Genes in Caenorhabditis Elegans Is Mostly Due to Operons and Duplicate Genes». *Genome Research* 13 (2): 238-43. https://doi.org/10.1101/gr.553803.

Li, Yifeng, Fang-Xiang Wu, y Alioune Ngom. 2018. «A Review on Machine Learning Principles for Multi-View Biological Data Integration». *Briefings in Bioinformatics* 19 (2): 325-40. https://doi.org/10.1093/bib/bbw113.

Libbrecht, Maxwell W., y William Stafford Noble. 2015. «Machine Learning Applications in Genetics and Genomics». *Nature Reviews Genetics* 16 (6): 321-32. https://doi.org/10.1038/nrg3920.

Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009. «Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome». *Science (New York, N.Y.)* 326 (5950): 289-93. https://doi.org/10.1126/science.1181369.

Ling, Xu, Xin He, y Dong Xin. 2009. «Detecting Gene Clusters under Evolutionary Constraint in a Large Number of Genomes». *Bioinformatics (Oxford, England)* 25 (5): 571-77. https://doi.org/10.1093/bioinformatics/btp027.

Liu, Yingli, Chen Niu, Zhuo Wang, Yong Gan, Yan Zhu, Shuhong Sun, y Tao Shen. 2020. «Machine Learning in Materials Genome Initiative: A Review». *Journal of Materials Science & Technology*, mayo. https://doi.org/10.1016/j.jmst.2020.01.067.

Michalak, Pawel. 2008. «Coexpression, coregulation, and cofunctionality of neighboring genes in eukaryotic genomes». *Genomics* 91 (3): 243-48. https://doi.org/10.1016/j.ygeno.2007.11.002.

Mihelčić, Matej, Tomislav Šmuc, y Fran Supek. 2019. «Patterns of diverse gene functions in genomic neighborhoods predict gene function and phenotype». *Scientific Reports* 9 (diciembre). https://doi.org/10.1038/s41598-019-55984-0.

Moore, Bethany M., Peipei Wang, Pengxiang Fan, Bryan Leong, Craig A. Schenck, John P. Lloyd, Melissa D. Lehti-Shiu, Robert L. Last, Eran Pichersky, y Shin-Han Shiu. 2019. «Robust Predictions of Specialized Metabolism Genes through Machine Learning». *Proceedings of the National Academy of Sciences* 116 (6): 2344-53. https://doi.org/10.1073/pnas.1817074116.

Mossman, Jim A., Leann M. Biancani, y David M. Rand. 2019. «Mitochondrial Genomic Variation Drives Differential Nuclear Gene Expression in Discrete Regions of Drosophila Gene and Protein Interaction Networks». *BMC Genomics* 20 (1): 691. https://doi.org/10.1186/s12864-019-6061-y.

Nguyen, Nam D., y Daifeng Wang. 2020. «Multiview Learning for Understanding Functional Multiomics». *PLoS Computational Biology* 16 (4): e1007677. https://doi.org/10.1371/journal.pcbi.1007677.

Nicholls, Hannah L., Christopher R. John, David S. Watson, Patricia B. Munroe, Michael R. Barnes, y Claudia P. Cabrera. 2020. «Reaching the End-Game for GWAS: Machine

Learning Approaches for the Prioritization of Complex Disease Loci». *Frontiers in Genetics* 11: 350. https://doi.org/10.3389/fgene.2020.00350.

Niehrs, C., y N. Pollet. 1999. «Synexpression Groups in Eukaryotes». *Nature* 402 (6761): 483-87. https://doi.org/10.1038/990025.

Noé, Frank, Gianni De Fabritiis, y Cecilia Clementi. 2020. «Machine Learning for Protein Folding and Dynamics». *Current Opinion in Structural Biology*, Folding and Binding ● Proteins, 60 (febrero): 77-84. https://doi.org/10.1016/j.sbi.2019.12.005.

Noé, Frank, Alexandre Tkatchenko, Klaus-Robert Müller, y Cecilia Clementi. 2020. «Machine Learning for Molecular Simulation». *Annual Review of Physical Chemistry* 71 (1): 361-90. https://doi.org/10.1146/annurev-physchem-042018-052331.

Nüsslein-Volhard, Christiane, y Eric Wieschaus. 1980. «Mutations Affecting Segment Number and Polarity in Drosophila». *Nature* 287 (5785): 795-801. https://doi.org/10.1038/287795a0.

Nützmann, Hans-Wilhelm, Claudio Scazzocchio, y Anne Osbourn. 2018. «Metabolic Gene Clusters in Eukaryotes». *Annual Review of Genetics* 52: 159-83. https://doi.org/10.1146/annurev-genet-120417-031237.

Osbourn, Anne E., y Ben Field. 2009. «Operons». *Cellular and Molecular Life Sciences* 66 (23): 3755-75. https://doi.org/10.1007/s00018-009-0114-3.

Otero, Fernando E. B., Alex A. Freitas, y Colin G. Johnson. 2010. «A Hierarchical Multi-Label Classification Ant Colony Algorithm for Protein Function Prediction». *Memetic Computing* 2 (3): 165-81. https://doi.org/10.1007/s12293-010-0045-4.

Overbeek, Ross, Michael Fonstein, Mark D'Souza, Gordon D. Pusch, y Natalia Maltsev. 1999. «The use of gene clusters to infer functional coupling». *Proceedings of the National Academy of Sciences of the United States of America* 96 (6): 2896-2901.

Paul, Abriti, Sourav Ghosh, Amit Kumar Das, Saptarsi Goswami, Sruti Das Choudhury, y Soumya Sen. 2020. «A Review on Agricultural Advancement Based on Computer Vision and Machine Learning». En *Emerging Technology in Modelling and Graphics*, editado por Jyotsna Kumar Mandal y Debika Bhattacharya, 567-81. Advances in Intelligent Systems and Computing. Singapore: Springer. https://doi.org/10.1007/978-981-13-7403-6_50.

Pazos Obregón, Flavio, Martín Palazzo, Pablo Soto, Gustavo Guerberoff, Patricio Yankilevich, y Rafael Cantera. 2019. «An improved catalogue of putative synaptic genes defined exclusively by temporal transcription profiles through an ensemble machine learning approach». *BMC Genomics* 20 (1): 1011. https://doi.org/10.1186/s12864-019-6380-z.

Pazos Obregón, Flavio, Cecilia Papalardo, Sebastián Castro, Gustavo Guerberoff, y Rafael Cantera. 2015. «Putative synaptic genes defined from a Drosophila whole body developmental transcriptome by a machine learning approach». *BMC Genomics* 16 (septiembre): 694. https://doi.org/10.1186/s12864-015-1888-3.

Pazos Obregón, Flavio, Pablo Soto, José Luis Lavín, Ana Rosa Cortázar, Rosa Barrio, Ana María Aransay, y Rafael Cantera. 2018. «Cluster Locator, Online Analysis and Visualization of Gene Clustering». *Bioinformatics (Oxford, England)* 34 (19): 3377-79. https://doi.org/10.1093/bioinformatics/bty336.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. «Scikit-learn: Machine Learning in Python». *Journal of Machine Learning Research* 12 (Oct): 2825-30.

Planells, Benjamín, Isabel Gómez-Redondo, Eva Pericuesta, Patrick Lonergan, y Alfonso Gutiérrez-Adán. 2019. «Differential isoform expression and alternative splicing in sex determination in mice». *BMC Genomics* 20 (1): 202. https://doi.org/10.1186/s12864-019-5572-x.

Price, Morgan N., Adam P. Arkin, y Eric J. Alm. 2006. «The Life-Cycle of Operons». *PLoS Genetics* 2 (6): e96. https://doi.org/10.1371/journal.pgen.0020096.

Purmann, Antje, Joern Toedling, Markus Schueler, Piero Carninci, Hans Lehrach, Yoshihide Hayashizaki, Wolfgang Huber, y Silke Sperling. 2007. «Genomic Organization of Transcriptomes in Mammals: Coregulation and Cofunctionality». *Genomics* 89 (5): 580-87. https://doi.org/10.1016/j.ygeno.2007.01.010.

Radivojac, Predrag, Wyatt T. Clark, Tal Ronnen Oron, Alexandra M. Schnoes, Tobias Wittkop, Artem Sokolov, Kiley Graim, et al. 2013a. «A Large-Scale Evaluation of Computational Protein Function Prediction». *Nature Methods* 10 (3): 221-27. https://doi.org/10.1038/nmeth.2340.

———. 2013b. «A Large-Scale Evaluation of Computational Protein Function Prediction». *Nature Methods* 10 (3): 221-27. https://doi.org/10.1038/nmeth.2340.

Raphael, Alina, Zvy Dubinsky, David Iluz, y Nathan S. Netanyahu. 2020. «Neural Network Recognition of Marine Benthos and Corals». *Diversity* 12 (1): 29. https://doi.org/10.3390/d12010029.

«Receiver Operating Characteristic». 2020. En *Wikipedia*. https://en.wikipedia.org/w/index.php?title=Receiver_operating_characteristic&oldid=953967882.

Reimegård, Johan, Snehangshu Kundu, Ali Pendle, Vivian F. Irish, Peter Shaw, Naomi Nakayama, Jens F. Sundström, y Olof Emanuelsson. 2017. «Genome-wide identification of physically

clustered genes suggests chromatin-level co-regulation in male reproductive development in Arabidopsis thaliana». *Nucleic Acids Research* 45 (6): 3253-65. https://doi.org/10.1093/nar/gkx087.

Rocha, Eduardo P. C. 2008. «The Organization of the Bacterial Genome». *Annual Review of Genetics* 42: 211-33. https://doi.org/10.1146/annurev.genet.42.110807.091653.

Roy, Peter J., Joshua M. Stuart, Jim Lund, y Stuart K. Kim. 2002. «Chromosomal Clustering of Muscle-Expressed Genes in Caenorhabditis Elegans». *Nature* 418 (6901): 975-79. https://doi.org/10.1038/nature01012.

Rubin, Alan F., y Phil Green. 2013. «Expression-Based Segmentation of the Drosophila Genome». *BMC Genomics* 14 (noviembre): 812. https://doi.org/10.1186/1471-2164-14-812.

Samuel, A. L. 1959. «Some Studies in Machine Learning Using the Game of Checkers». *IBM Journal of Research and Development* 3 (3): 210-29. https://doi.org/10.1147/rd.33.0210.

Schwalbe, Nina, y Brian Wahl. 2020. «Artificial Intelligence and the Future of Global Health». *Lancet (London, England)* 395 (10236): 1579-86. https://doi.org/10.1016/S0140-6736(20)30226-9.

Schwartz, Yuri B., y Giacomo Cavalli. 2017. «Three-Dimensional Genome Organization and Function in Drosophila». *Genetics* 205 (1): 5-24. https://doi.org/10.1534/genetics.115.185132.

Sémon, Marie, y Laurent Duret. 2006. «Evolutionary Origin and Maintenance of Coexpressed Gene Clusters in Mammals». *Molecular Biology and Evolution* 23 (9): 1715-23. https://doi.org/10.1093/molbev/msl034.

Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, y Trey Ideker. 2003. «Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks». *Genome Research* 13 (11): 2498-2504. https://doi.org/10.1101/gr.1239303.

Shastry, K. Aditya, y H. A. Sanjay. 2020. «Machine Learning for Bioinformatics». En *Statistical Modelling and Machine Learning Principles for Bioinformatics Techniques, Tools, and Applications*, editado por K. G. Srinivasa, G. M. Siddesh, y S. R. Manisekhar, 25-39. Algorithms for Intelligent Systems. Singapore: Springer. https://doi.org/10.1007/978-981-15-2445-5_3.

Shimizu, Hideyuki, y Keiichi I. Nakayama. 2020. «Artificial Intelligence in Oncology». *Cancer Science* 111 (5): 1452-60. https://doi.org/10.1111/cas.14377.

Shrestha, Ajay, y Ausif Mahmood. 2019. «Review of Deep Learning Algorithms and Architectures». *IEEE Access* 7: 53040-65. https://doi.org/10.1109/ACCESS.2019.2912200.

Silla, Carlos N., y Alex A. Freitas. 2011. «A Survey of Hierarchical Classification across Different Application Domains». *Data Mining and Knowledge Discovery* 22 (1): 31-72. https://doi.org/10.1007/s10618-010-0175-9.

Spellman, Paul T., y Gerald M. Rubin. 2002. «Evidence for Large Domains of Similarly Expressed Genes in the Drosophila Genome». *Journal of Biology* 1 (1): 5.

Stafford, I. S., M. Kellermann, E. Mossotto, R. M. Beattie, B. D. MacArthur, y S. Ennis. 2020. «A Systematic Review of the Applications of Artificial Intelligence and Machine Learning in Autoimmune Diseases». *NPJ Digital Medicine* 3: 30. https://doi.org/10.1038/s41746-020-0229-3.

Stojanova, Daniela, Michelangelo Ceci, Donato Malerba, y Saso Dzeroski. 2013. «Using PPI network autocorrelation in hierarchical multi-label classification trees for gene function prediction». *BMC Bioinformatics* 14 (1): 285. https://doi.org/10.1186/1471-2105-14-285.

Sue, Masayuki, Chihiro Nakamura, y Taiji Nomura. 2011. «Dispersed Benzoxazinone Gene Cluster: Molecular Characterization and Chromosomal Localization of Glucosyltransferase and Glucosidase Genes in Wheat and Rye». *Plant Physiology* 157 (3): 985-97. https://doi.org/10.1104/pp.111.182378.

Takos, Adam M., Camilla Knudsen, Daniela Lai, Rubini Kannangara, Lisbeth Mikkelsen, Mohammed S. Motawia, Carl E. Olsen, et al. 2011. «Genomic Clustering of Cyanogenic Glucoside Biosynthetic Genes Aids Their Identification in Lotus Japonicus and Suggests the Repeated Evolution of This Chemical Defence Pathway». *The Plant Journal: For Cell and Molecular Biology* 68 (2): 273-86. https://doi.org/10.1111/j.1365-313X.2011.04685.x.

Thévenin, Annelyse, Liat Ein-Dor, Michal Ozery-Flato, y Ron Shamir. 2014. «Functional Gene Groups Are Concentrated within Chromosomes, among Chromosomes and in the Nuclear Space of the Human Genome». *Nucleic Acids Research* 42 (15): 9854-61. https://doi.org/10.1093/nar/gku667.

Tiirikka, Timo, Markku Siermala, y Mauno Vihinen. 2014. «Clustering of Gene Ontology Terms in Genomes». *Gene* 550 (2): 155-64. https://doi.org/10.1016/j.gene.2014.06.060.

Tuller, Tamir, Udi Rubinstein, Dani Bar, Michael Gurevitch, Eytan Ruppin, y Martin Kupiec. 2009a. «Higher-Order Genomic Organization of Cellular Functions in Yeast». *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* 16 (2): 303-16. https://doi.org/10.1089/cmb.2008.15TT.

———. 2009b. «Higher-Order Genomic Organization of Cellular Functions in Yeast». Research-article. Https://Home.Liebertpub.Com/Cmb. 4 de febrero de 2009. https://doi.org/10.1089/cmb.2008.15TT.

Ullah, Zaib, Fadi Al-Turjman, Leonardo Mostarda, y Roberto Gagliardi. 2020. «Applications of Artificial Intelligence and Machine Learning in Smart Cities». *Computer Communications* 154 (marzo): 313-23. https://doi.org/10.1016/j.comcom.2020.02.069.

Valentini, Giorgio. 2011. «True Path Rule Hierarchical Ensembles for Genome-Wide Gene Function Prediction». *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 8 (3): 832-47. https://doi.org/10.1109/TCBB.2010.38.

———. 2014. «Hierarchical Ensemble Methods for Protein Function Prediction». Review Article. ISRN Bioinformatics. 2014. https://doi.org/10.1155/2014/901419.

Vapnik, Vladimir. 2000. *The Nature of Statistical Learning Theory*. 2.ª ed. Information Science and Statistics. New York: Springer-Verlag. https://doi.org/10.1007/978-1-4757-3264-1.

Versteeg, Rogier, Barbera D. C. van Schaik, Marinus F. van Batenburg, Marco Roos, Ramin Monajemi, Huib Caron, Harmen J. Bussemaker, y Antoine H. C. van Kampen. 2003. «The Human Transcriptome Map Reveals Extremes in Gene Density, Intron Length, GC Content, and Repeat Pattern for Domains of Highly and Weakly Expressed Genes». *Genome Research* 13 (9): 1998-2004. https://doi.org/10.1101/gr.1649303.

Wang, Kewei, Wenji Wang, y Mang Li. 2018. «A brief procedure for big data analysis of gene expression». *Animal Models and Experimental Medicine* 1 (3): 189-93. https://doi.org/10.1002/ame2.12028.

Webb, Sarah. 2018. «Deep Learning for Biology». *Nature* 554 (7693): 555-57. https://doi.org/10.1038/d41586-018-02174-z.

Weber, Claudia C, y Laurence D Hurst. 2011. «Support for multiple classes of local expression clusters in Drosophila melanogaster, but no evidence for gene order conservation». *Genome Biology* 12 (3): R23. https://doi.org/10.1186/gb-2011-12-3-r23.

Wei, Wu, Vicent Pelechano, Aino I. Järvelin, y Lars M. Steinmetz. 2011. «Functional Consequences of Bidirectional Promoters». *Trends in Genetics: TIG* 27 (7): 267-76. https://doi.org/10.1016/j.tig.2011.04.002.

Williams, Elizabeth J. B., y Dianna J. Bowles. 2004. «Coexpression of Neighboring Genes in the Genome of Arabidopsis Thaliana». *Genome Research* 14 (6): 1060-67. https://doi.org/10.1101/gr.2131104.

Yanai, Itai, Joseph C. Mellor, y Charles DeLisi. 2002. «Identifying Functional Links between Genes Using Conserved Chromosomal Proximity». *Trends in Genetics: TIG* 18 (4): 176-79. https://doi.org/10.1016/s0168-9525(01)02621-x.

Yi, Gangman, Sing-Hoi Sze, y Michael R. Thon. 2007. «Identifying clusters of functionally related genes in genomes». *Bioinformatics* 23 (9): 1053-60. https://doi.org/10.1093/bioinformatics/btl673.

Yi, Wolf, Rogozin Ib, Kondrashov As, y Koonin Ev. 2001. «Genome Alignment, Evolution of Prokaryotic Genome Organization, and Prediction of Gene Function Using Genomic Context». Genome Research. marzo de 2001. https://doi.org/10.1101/gr.gr-1619r.

You, Ronghui, Zihan Zhang, Yi Xiong, Fengzhu Sun, Hiroshi Mamitsuka, y Shanfeng Zhu. 2018. «GOLabeler: Improving Sequence-Based Large-Scale Protein Function Prediction by Learning to Rank». *Bioinformatics (Oxford, England)* 34 (14): 2465-73. https://doi.org/10.1093/bioinformatics/bty130.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, y Qing-Yu He. 2012. «clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters». *OMICS : a Journal of Integrative Biology* 16 (5): 284-87. https://doi.org/10.1089/omi.2011.0118.

Zerbino, Daniel R., Premanand Achuthan, Wasiu Akanni, M. Ridwan Amode, Daniel Barrell, Jyothish Bhai, Konstantinos Billis, et al. 2018. «Ensembl 2018». *Nucleic Acids Research* 46 (D1): D754-61. https://doi.org/10.1093/nar/gkx1098.

Zhao, Yingwen, Jun Wang, Jian Chen, Xiangliang Zhang, Maozu Guo, y Guoxian Yu. 2020. «A Literature Review of Gene Function Prediction by Modeling Gene Ontology». *Frontiers in Genetics* 11: 400. https://doi.org/10.3389/fgene.2020.00400.

Zheng, Yu, Richard J. Roberts, y Simon Kasif. 2002. «Genomic Functional Annotation Using Co-Evolution Profiles of Gene Clusters». *Genome Biology* 3 (11): RESEARCH0060. https://doi.org/10.1186/gb-2002-3-11-research0060.

Zhou, Naihui, Yuxiang Jiang, Timothy R. Bergquist, Alexandra J. Lee, Balint Z. Kacsoh, Alex W. Crocker, Kimberley A. Lewis, et al. 2019. «The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens». *Genome Biology* 20 (1): 244. https://doi.org/10.1186/s13059-019-1835-8.

Zou, James, Mikael Huss, Abubakar Abid, Pejman Mohammadi, Ali Torkamani, y Amalio Telenti. 2019. «A Primer on Deep Learning in Genomics». *Nature Genetics* 51 (1): 12-18. https://doi.org/10.1038/s41588-018-0295-5.