Universidad Nacional Autónoma de México

Centro de Física Aplicada y Tecnología Avanzada

Implementación de modelo matemático para la inferencia y análisis de la historia evolutiva de grandes familias de genes

T E S I S

que para obtener el título de:

Licenciado en Tecnología

presenta:

José Antonio Ramírez Rafael

Tutoras:

Dra. Maribel Hernandez Rosales
Dra. Katia Aviña Padilla

Juriquilla, Querétaro 2021

Resumen

La historia evolutiva de una familia de genes es una sucesión de eventos que explican la existencia de los miembros de tal familia dentro de una o más especies. Mediante el uso de este concepto se clasifica a dos genes como *homólogos* si comparten un ancestro en común. Esta categoría se puede refinar tomando en cuenta el evento que da origen a los genes comparados en el ancestro más reciente que comparten, siendo la *ortología* una de las relaciones más relevantes, pues corresponde a genes que surgen por un evento de especiación y se les puede entender —de forma sobresimplificada— como *el mismo gen en diferentes especies*. Las relaciones de ortología pueden ser representadas por un *cografo* [31], que ayuda a la biología matemática a determinar propiedades fundamentales que deben cumplir un conjunto de relaciones de este tipo.

Este trabajo presenta la implementación de una metodología novedosa para la reconstrucción de historias evolutivas. Para tal fin, el procedimiento implementado analiza las relaciones de ortología correspondientes a un conjunto de genes. Si dichas relaciones no son conocidas, éstas pueden ser predichas por medio de la comparación de las secuencias de los genes, para posteriormente filtrar aquellas relaciones que no cumplan con las propiedades esperadas. La herramienta generada permite analizar la evolución de genomas completos y evita sesgos asociados a otras metodologías. Para validarla se analizaron escenarios evolutivos simulados, así como los genes sin intrones de ratón y sus ortólogos en vertebrados selectos: humano, chimpancé, zarigüeya, rata, gallo, y pez cebra.

Durante la presentación de los resultados se muestra cómo la herramienta desarrollada puede obtener de forma acertada historias evolutivas dado un conjunto de relaciones de ortología y un árbol de especies. Del análisis de genes sin intrones se pudo observar que estos tienden a estar más conservados en mamíferos, lo que podría aportar evidencia a la teoría de que esta clase de genes tiene un origen *reciente*. Además se observa que los ortólogos de genes de ratón tienden a conservar la condición de tener o no intrones. Para ejemplificar estas características se muestra explícitamente la historia inferida para las β-protocadherinas de ratón, así como un resumen visual de todas las familias identificadas.

El análisis de genes sin intrones toma relevancia médica, pues la presencia de intrones está relacionada a procesos de regulación genética, y se ha identificado que tales genes se involucran en enfermedades como el cáncer, además de que se han propuesto como posibles biomarcadores en esta clase de enfermedades. El análisis de su evolución muestra que los genes sin intrones pueden ser estudiados en otros organismos modelo con la finalidad de entender su papel en enfermedades que aquejan a la especie humana.

Este proyecto se ajusta perfectamente con el perfil de un licenciado en Tecnología, pues requiere de la integración de conocimientos de diferentes áreas con la finalidad de crear una herramienta que permita abordar un problema de forma innovadora, aportando velocidad y poder de análisis. De igual forma, el presente trabajo implica desarrollo de tecnología nacional que puede ser usada en investigaciones de frontera o como base para perfeccionar metodologías existentes.

Agradecimientos

[...]
¿Cómo obligarse a ser la excepción,
un accidente,
y quedarse en la conciencia del viento?
(Que te respieren, que te piensen)
¿Cómo forzar a «natura»?, si al igual que el humano la flor es bella
—porque nace de lo incierto—
si como la flor y el fruto, el humano es perecedero y vive
para aprender
y entregarse
al singular acontecimiento
de la muerte.
Adriana Tafoya - 2013

Me siento tremendamente agradecido con todas aquellas personas que han tenido impacto en mi formación ya sea en lo académico o lo cultural. Dichas personas las he encontrado en círculos de familiares, buenos amigos, y equipos de trabajo. Estos grupos no son mutuamente excluyentes, y valoro profundamente los aportes que cada quién hizo para impulsarme a desarrollar y finalizar este trabajo, así como a quienes me hicieron compañía durante su desarrollo. En los párrafos de abajo mencionaré a los principales implicados en esta actividad.

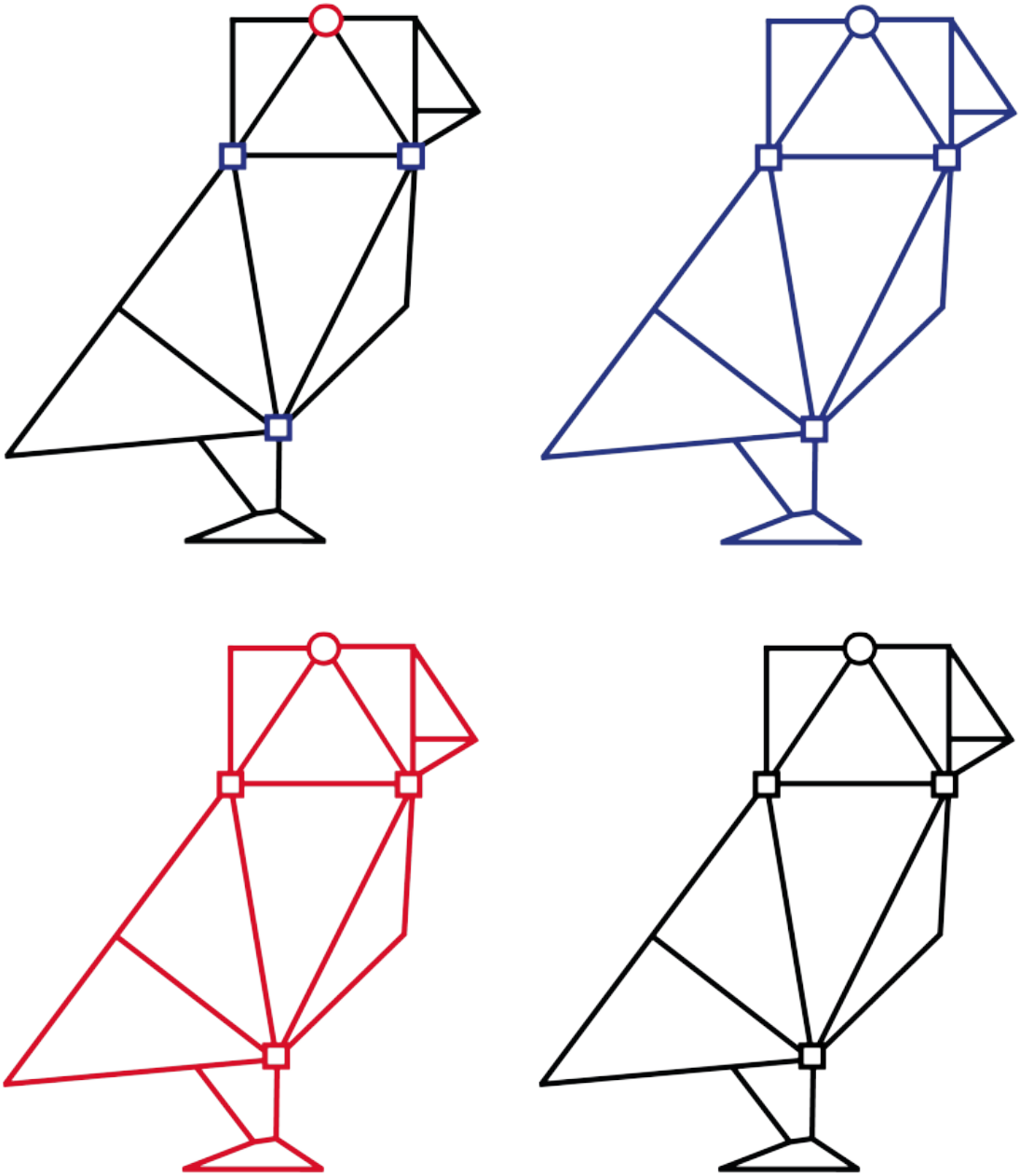
Mis familiares más cercanos, expresamente Araceli Rafael Coyol y J. Antonio Ramírez Alvarado, que a lo largo de toda mi vida me han apoyado satisfaciendo mis necesidades tanto emocionales como económicas, y que con gran entusiasmo me apoyaron durante la licenciatura y trabajo de tesis.

Mis tutoras de tesis, las doctoras Maribel Hernández Rosales y Katia Aviña Padilla, que desde mis primeras interacciones con ellas me introdujeron activamente en el mundo de la investigación científica y en el desarrollo de proyectos de frontera. De igual manera a mis sinodales: los drs. Alfredo Varela Echavarría, Bertha Guadalupe Rueda Zarazúa, Cristy Leonor Azanza Ricardo y Guillermo de Anda Jaúregui, pues se tomaron el tiempo para revisar a detalle mi tesis y sugerir cambios así como comentarios constructivos. A todos los conocí en cursos que me impartieron o en eventos científicos, situación que me motiva a seguir en esta línea de trabajo.

Grandes maestros que nutrieron mis conocimientos con cursos asombrosos, de donde me pude llevar incluso más que buen contenido técnico. Son demasiadas personas como para listarlas aquí, sin embargo no cabrá duda de la influencia que tuvieron en mi desarrollo profesional. En esta larga lista que no escribiré también me gustaría incluir innumerables

amigos y familiares que de igual forma me dieron grandes enseñanzas y buenos tiempos.

Todas las instituciones públicas que apoyaron e hicieron posible mi formación y el desarrollo del presente proyecto. En particular a la UNAM por fungir como plataforma y distribuidora de recursos para estudiantes. Al LAVIS de la UNAM y a la universidad de Leipzig por brindarme el acceso a servidores donde correr análisis, y al CONACyT por apoyarme con una beca de ayudantía para que este proyecto se vea involucrado en investigaciones alrededor de enfermedades como el cáncer.



Networks are cool, and also graphs are really awesome!

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

La evolución de las especies es un fenómeno complejo donde se ven involucrados diferentes factores como reproducción de organismos, herencia de material genético, mutación, migración, deriva génica, selección natural, entre otros. Nowak engloba estos aspectos proponiendo tres principios fundamentales que actúan sobre una población de elementos; *i*) la **replicación** de tales elementos; *ii*) **mutación** durante la replicación; y *iii*) **selección** de individuos dentro de la población [46]. Usaremos esta generalización para referirnos a la evolución de genes y organismos, aunque no debe olvidarse cuán diversas son las fuerzas que influencian tales fenómenos.

Charles Darwin fue el primero en proponer una teoría casi completa que permite entender el origen y la relación que existe entre diferentes entes vivos o sus componentes por medio de la identificación de los eventos que influenciaron a los organismos para tener las características con las que los observamos. Esto permite designar a elementos (genes, funciones, etc.) en dos organismos diferentes como *homólogos* si descienden de un ancestro en común.

La biología comparativa ha identificado estructuras homólogas por medio de la comparación de sus características, entre las que se encuentra el material genético de los organismos que las contienen, pues ahí reside el conjunto de instrucciones necesarias para desarrollar las estructuras comparadas, así como para que éstas realicen sus funciones en el sistema al que están integrados. Recientes avances tecnológicos han permitido a la comunidad científica obtener la secuencia de genomas completos de diversas especies, brindando la oportunidad de analizar la evolución por medio de la comparación de genomas.

En la perspectiva de la genómica comparativa, la necesidad de agrupar genes se da de forma natural, así como obtener una filogenia que explique su origen. Tales procedimientos derivan en el surgimiento del concepto ***familia de genes***, que se refiere a un grupo de genes que descienden de un ancestro en común, reteniendo secuencias similares y a veces funciones similares [17]. Los miembros de una misma familia pueden encontrarse en un solo genoma o estar presentes en genomas de diferentes especies.

La creciente aparición de nuevos datos genómicos, ha impulsado el desarrollo de méto-

dos y herramientas para analizar la evolución de familias de genes. Estos métodos se basan en la comparación de genes para inferir historias evolutivas o pares de genes ortólogos. Sin embargo, es común tener errores debido, por ejemplo, a que los datos son demasiado grandes o sesgados, las distancias evolutivas no son uniformes, existen tasas de mutación variable, por la selección manual de familias o por pérdida de genes. Por ello se han buscado alternativas. Al estudiar las propiedades de las relaciones de ortología se demuestra que se puede obtener la historia de los genes en términos de eventos evolutivos a partir solamente de relaciones de ortología, lo cual permite diseñar una metodología que evita errores asociados a otros métodos.

La metodología que aquí proponemos es aplicable a cualquier conjunto de genes, por lo que puede ser empleada con diferentes finalidades. Particularmente, en el presente trabajo se empleó el método para el análisis y reconstrucción de la historia evolutiva del conjunto de genes en ratón clasificados como *genes sin intrones* (IGs por sus siglas en inglés, *intronless genes*). Este conjunto de genes es de interés debido a que se ha descrito su potencial asociado a enfermedades de regulación como el cáncer. Otra característica relevante es que su evolución ha sido poco explorada. Se detalla más acerca de esta clase de genes en las secciones 2.1.5 y 5.3.

En el capítulo 2 se presentan los conceptos básicos de biología y matemáticas necesarios para la comprensión de la metodología desarrollada. Se revisarán los conceptos de *homología* y sus derivados, así como su relación con la evolución de las especies y las moléculas de la vida. El contenido matemático se concentra en teoría de grafos, que es esencial para la representación de las relaciones entre los genes y las historias evolutivas.

En el capítulo 3 se hace un recuento de los conceptos y tecnologías que se han empleado para el estudio de la evolución, mencionando brevemente el contexto histórico en que surge esta teoría, hasta la actual era de las *ómicas* donde se enfatiza la genómica comparativa y las herramientas actuales para el análisis evolutivo.

En los capítulos posteriores se presenta la metodología empleada, los resultados y la discusión. En estas partes se hace uso amplio de los conceptos definidos en el capítulo 2, se discuten mejoras que se podrían agregar a la herramienta, se analiza la evolución de genes sin intrones de ratón y finalmente se plantean análisis que se pueden desarrollar usando la información obtenida que nos permitan tratar de determinar el papel que toman estos genes en enfermedades como el cáncer.

Capítulo 2

Conceptos básicos

2.1. Bases biológicas

Una de las evidencias experimentales de la teoría de la evolución de las especies reside en la identificación de partes o estructuras en diferentes organismos que fueron heredadas de un ancestro en común de las especies comparadas. Al estudiar más a profundidad este fenómeno, se observan diferentes patrones incluso a escala molecular, es decir a nivel de los genes y proteínas que contiene un organismo. Esta sección presenta el concepto de homología desde el nivel de los organismos y lo traslada al nivel de secuencias genéticas, finalizando con las definiciones de las relaciones evolutivas que pueden existir entre pares de genes.

2.1.1. Homología enorganismos

El concepto de **homología** surgió en 3 diferentes contextos: *ideal*, *histórico*, e *iterativo*. Las bases empíricas de todos, sin embargo, es el reconocimiento de que *algunas características o partes en un cuerpo son la misma en diferentes organismos o en diferentes regiones del mismo cuerpo* [54].

La **homología ideal** presenta al fenómeno como un arquetipo o modelo estructural presente en todos los vertebrados. Este concepto es más viejo que el término *homología*, que fue formalizado por Richard Owen en 1843. La **homología histórica** (usada ampliamente desde la publicación de *El origen de las especies*, de Charles Darwin) dice que la existencia de estructuras que se corresponden en diferentes especies se explica por medio de una especie ancestral que derivó en las dos especies comparadas, heredando a ambas tales estructuras. Por otro lado, la **homología iterativa** se refiere a la correspondencia entre partes de un mismo cuerpo, por ejemplo la que existe entre las hojas de follaje y los pétalos de una flor [54]. Un ejemplo de homología estudiado por Charles Darwin es el de los pinzones, que puede observarse en la figura 2.1. Este trabajo se centra en la homología histórica.

Al estudiar la homología histórica por medio de la comparación de organismos, se puede observar que dos órganos o partes de las especies comparadas pueden tener la misma

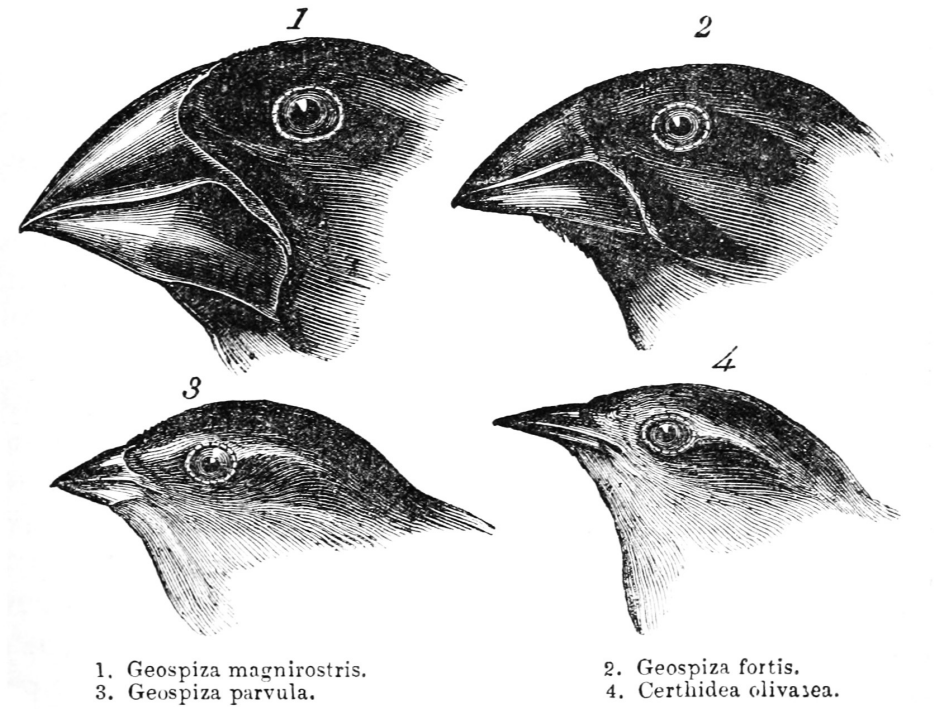


Figura 2.1: Pinzones de Darwin. Podemos ver que a pesar de sus diferencias, los cuatro organismos de la familia *Fringillidae* tienen partes que son homólogas. Un ejemplo es el pico, que fue heredado a todos por un ancestro en común, y cada organismo se adaptó para un ambiente específico. Imagen tomada de [15].

función, pero no ser homólogas, dado que los órganos correspondientes no fueron heredados de un ancestro en común, sino que se desarrollaron de forma independiente por cada especie. Cuando dos partes de un organismo cumplen con esta condición, se dice que son **análogos**, y su relación es el resultado de un evento de convergencia evolutiva [54].

Por más de 100 años el ejemplo más citado de analogía evolutiva fueron los ojos de cefalópodos y vertebrados, aunque en los últimos años la idea de que la morfología se desarrolló independientemente por cada especie ha cambiado dado el descubrimiento de una red genética conservada. En esta regulación se incluye al gen *Pax6* que gobierna la organogenésis del ojo tanto en vertebrados como en invertebrados. Sin embargo, los ojos de tipo cámara tienen un componente de convergencia en su historia evolutiva [50]. En la figura 2.2 se muestran una comparación de los ojos de pulpo y humano, así como la historia evolutiva que explica su origen.

2.1.2. Evolución de las especies

Desde tiempos remotos, se entendía a la evolución como un proceso en el que todos los organismos provienen de un ancestro en común por medio de variaciones morfológicas

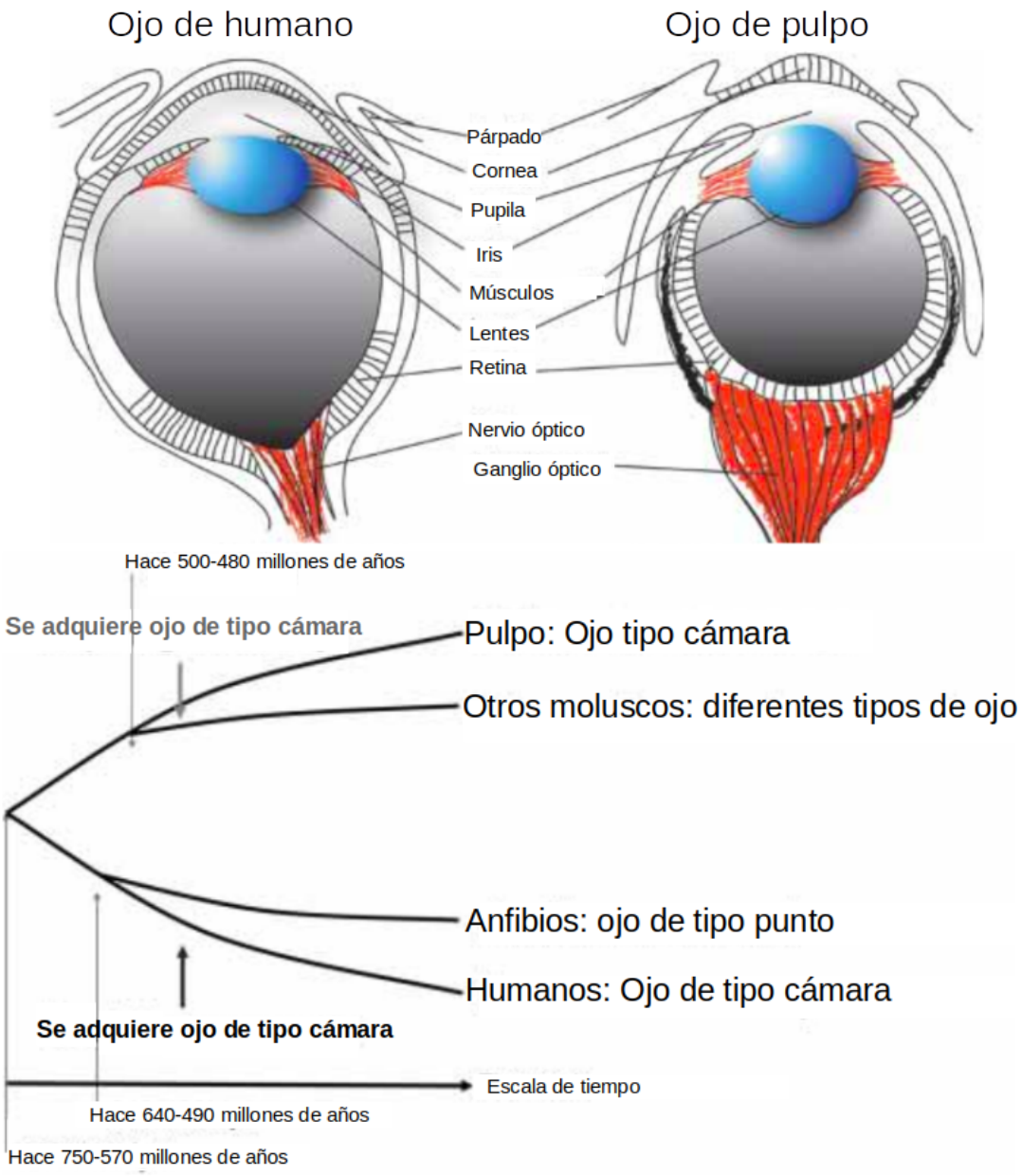


Figura 2.2: Origen evolutivo los ojos de tipo cámara del pulpo y humano. En la parte de arriba se compara la estructura de los ojos, se puede observar que se comparten la mayoría de los elementos. La diferencia radica en los procesos de desarrollo, descritos por Harris en 1997. En la parte de abajo se muestra la historia evolutiva de los ojos de cada organismo, podemos observar que cada ojo de tipo cámara tiene un origen diferente. Diagramás editados de [47].

y genómicas. Sin embargo, uno de los mecanismos fundamentales para la evolución de las especies es la **selección natural**, que fue concebido independientemente por Charles Darwin y Alfred Russel Wallace. Darwin fue el primero en resumir en su obra *El origen de las especies* un cuerpo coherente de observaciones que consolidan la evolución de los

organismos en una teoría verdaderamente científica. Con el paso del tiempo surgió el término *darwinismo*, ampliamente impulsado por Wallace [40]. Al equiparar tal concepto con *El origen de las especies* se pueden obtener 5 ideas separadas:

- La evolución como un hecho
- Teoría del ancestro común
- Gradualismo
- Multiplicación de las especies
- Selección natural

Darwin no fue capaz de explicar el mecanismo por el cual se da la diversidad biológica sobre la cual actúa la selección natural. La respuesta a este enigma se encontraba en el trabajo de Gregor Mendel, en el que estudió la herencia en plantas y propuso una unidad básica llamada **gen**. Sin embargo, su trabajo permaneció oculto en los anales de la Brno Academy of Science y no fue hasta varias décadas más tarde que su trabajo fue re-descubierto. El darwinismo y la genética mendeliana fueron unificados en la disciplina de *biología matemática*, desarrollada en las investigaciones seminales de Ronald Fisher, J. B. S. Haldane, y Sewall Wright en las décadas de 1920 y 1930 [46].

Nowak [46] determina en su libro de dinámica evolutiva que la evolución de una población de elementos se puede describir con tres principios fundamentales; *i*) la **replicación** de tales elementos; *ii*) **mutación** durante la replicación; y *iii*) **selección** de individuos dentro de la población. Según él, estos conceptos son fundamentales para definir sistemas biológicos, pues todo organismo vivo ha surgido y se por los mismos.

Los principios de arriba requieren que la evolución opere sobre poblaciones de individuos que se reproducen, por ejemplo, en el ambiente adecuado, entidades biológicas como virus, bacterias, células y organismos multicelulares son capaces de copiarse a sí mismos. El DNA y el RNA se replican y son heredados. La selección ocurre cuando diferentes tipos de individuos compiten entre sí. La reproducción no es perfecta, e implica errores o mutaciones ocasionales. La mutación puede generar nuevos tipos de individuos que serán sometidos de nuevo a procesos de selección, resultando en novedad y diversidad biológica. La selección permitirá la sobrevivencia de algunas innovaciones y otras no, y puede favorecer a la diversidad genética [46].

2.1.3. Genotipo, fenotipo y especies

En genética, la forma más común de cualquier propiedad de un organismo es llamada *forma silvestre* (**WT**), porque es la que se encuentra *en la naturaleza*, mientras que las variantes observables en los organismos que difieren de la forma silvestre son llamados ***mutantes***. El término ***fenotipo*** se refiere a cada una de las formás alternativas que puede

tomar una propiedad hereditaria [26].

Uno de los puntos que Darwin no cubrió en su teoría fue el mecanismo por el cual se da la diversidad de los fenotipos. Fue Mendel quien estudió por primera vez las relaciones matemáticas de la herencia, proponiendo al ***gen*** como su unidad básica, así como la existencia de diferentes formas de ese gen, llamadas ***alelos***, y demostrando cómo combinaciones de estos generan diferentes ***fenotipos*** [26]. En la figura 2.3 se muestra cómo variaciones en el genotipo provocan diferentes fenotipos.

Otras formas de entender estos conceptos es definiendo al fenotipo como *el organismo interactuando con su entorno*, y al genotipo como el conjunto de instrucciones para la replicación de un organismo [48]. Es de esperarse que exista una amplia conexión entre estas definiciones, pues el vehículo del fenotipo es el genotipo [48].

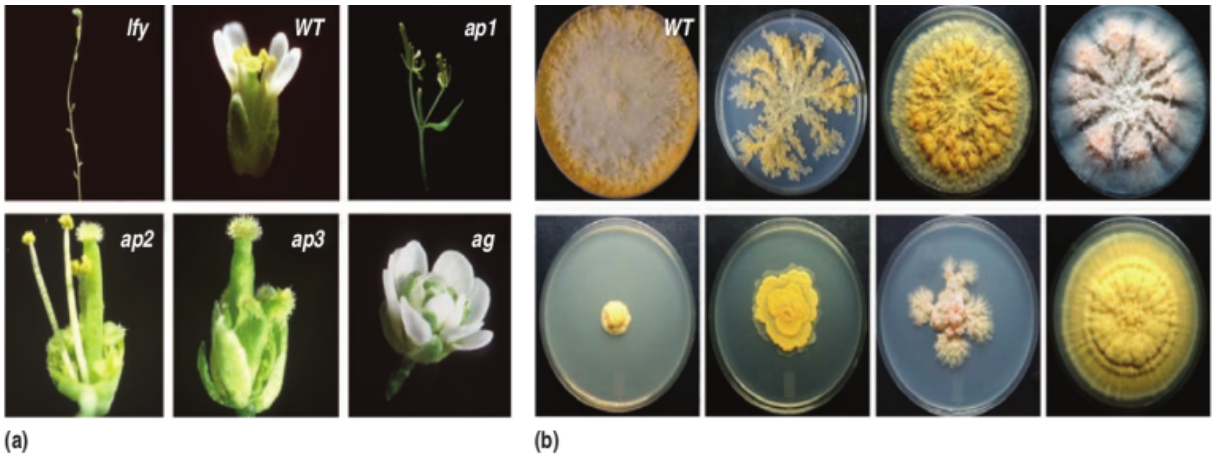


Figura 2.3: Estas fotografías muestran fenotipos mutantes, estos casos se dieron por la disección de las funciones: **(a)** desarrollo de flores en *Arabidopsis thaliana*, y **(b)** crecimiento de hifas en *Neurospora crassa*. las especies originales en su forma silvestre se denotan con **WT** , mientras que el resto de imágenes corresponden a mutantes. Imagen tomada de [26] Esto nos muestra cómo variaciones en el material genético derivan en diferentes fenotipos.

Los organismos son clasificados comúnmente como ***especies***, sin embargo no existe una definición completamente satisfactoria de tal término. Una de las mejores definiciones operacionales es la de un grupo de organismos separado de otros por *reproducción aislada*, pero sólo funciona bajo circunstancias debidamente delimitadas, pues factores geográficos, genéticos y epigenéticos pueden también tomar un rol importante en la formación de las especies. Un ejemplo de excepción a esta definición puede darse con comunidades humanas que viven aisladas y aún así se siguen considerando *Homo sapiens* [48].

Definiciones como la de arriba se vuelven aún más complicadas cuando se habla de especies que se reproducen de forma asexual (como las bacterias) pues el intercambio de material genético usualmente requiere representaciones diferentes y si no se toman en cuen-

ta pueden llevar a conclusiones erróneas. En estos casos el concepto de *reproducción aislada* es poco usado. En su lugar se usan criterios basados en la posibilidad en un intercambio significativo de material genético con otros organismos [48].

2.1.4. Moléculas de la vida y el dogma central de la biología

El (poli-)ácido desoxirribonucléico (**ADN**) es un polímero conformado por una secuencia de pares de bases de nucléotidos representados por los caracteres **A** (adenina), **T** (timina), **G** (guanina), y **C** (citosina), en el que los pares son **A-T** y **G-C**. Esta molécula es principalmente encontrada en forma de doble hélice en los organismos vivos. El ácido ribonucléico (**ARN**) es similar al ADN, aunque una de las principales diferencias es que el uracilo (**U**) remplaza a la **T** (timina), lo que hace que sus características químicas no permitan que esta molécula se encuentre en forma de doble hélice, sino que más bien como ciclos y abultamientos. Las **proteínas** toman su nombre del ser mitológico *Proteus*, que puede tomar muchas formas. Las principales funciones de las proteínas son estructurales y catalíticas. Las proteínas están formadas por más de 20 aminoácidos [48]. En la figura 2.4 se aprecian algunas conformaciones que pueden tomar estas moléculas.

Desde la primera mitad del del siglo XIX se pensaba que el ADN podría constituir las bases materiales de los genes. En 1944 Oswald Avery demostró que el material hereditario está constituido por ADN [16].

Tanto los ácidos nucléicos (ADN, ARN) y las proteínas pueden ser representadas en forma de secuencia de caracteres de un alfabeto dado (dígase nucléotidos, pares de bases o aminoácidos). El **dogma central de la biología** enunciado por Francis Crick en 1957 estudia el problema de trasferencia de información secuencial de un polímero con un alfabeto definido a otro, y se puede enunciar de la siguiente manera: *Una vez que la información secuencial ha pasado a proteína, no puede salir de esa forma.* La **hipótesis de la secuencia** es una re-formulación errónea del dogma central que dice que (en general) existe el flujo de información ADN $\rightarrow$ ADN $\rightarrow$ ARN $\rightarrow$ proteínas. Las transiciones enunciadas por esta hipótesis son llamadas como procesos de **replicación** (ADN $\rightarrow$ ADN), **trascripción** (ADN $\rightarrow$ ARN), y **traducción** (ARN $\rightarrow$ proteínas). Todas estas transiciones fueron clasificadas por Crick como **generales**, ya que se tiene evidencia directa o indirecta de que ocurren. Las transiciones **especiales** son aquellas que podrían pasar, mientras que las **desconocidas** son negadas por el dogma central, y si se llegase a encontrar una sola célula que contradiga a esta teoría, serían perturbadas las bases intelectuales de la biología molecular [14]. En la figura 2.5 se muestra un diagrama que resume al dogma central de la biología.

2.1.5. El genoma y los genes

El ***gen*** es una unidad fundamental de herencia, la cual es física y funcional, pues transporta información entre generaciones de organismos [26]. Otras definiciones mencionan de forma explícita que los genes se conforman por *un segmento de ADN que codifica para*

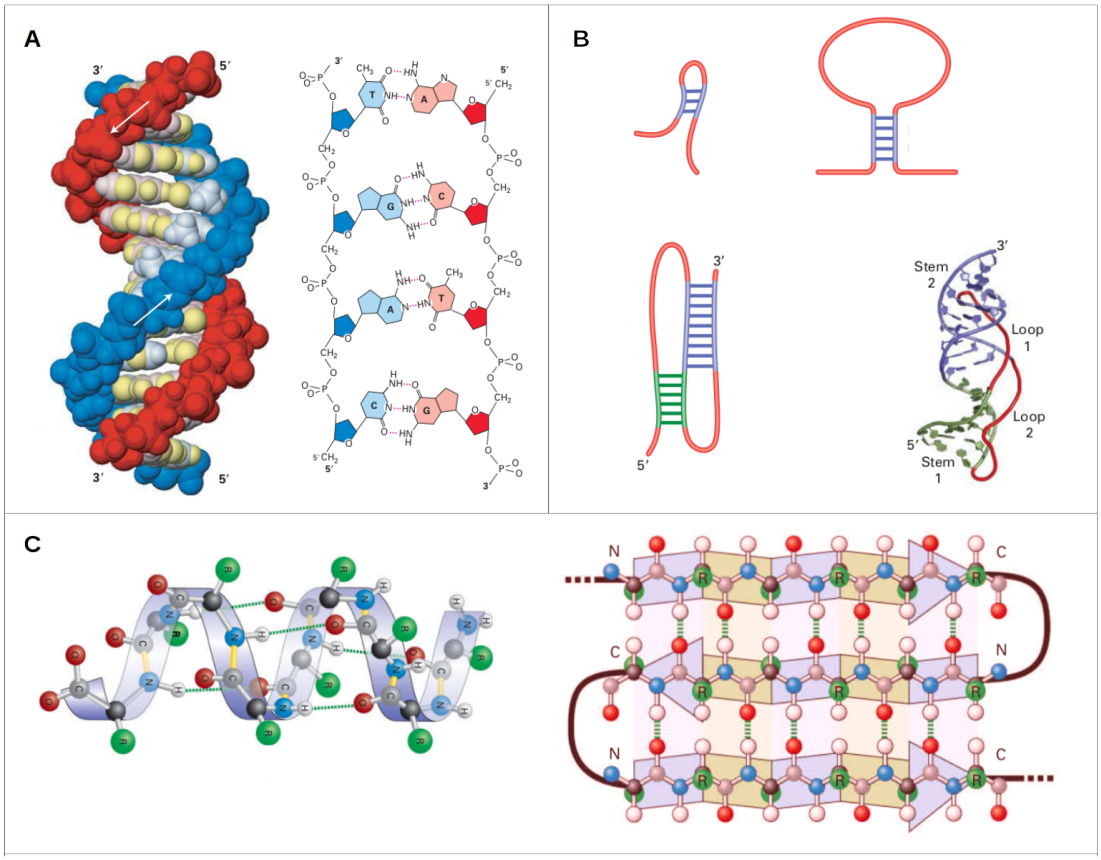


Figura 2.4: Moléculas del dogma central de la biología. En **A** se muestra la estructura tridimencional del ADN y una representación plana. Notar que esta molécula se puede representar como cadenas de caracteres, en el caso de la azul la secuencia es TGAC, mientras que del lado rojo se tiene ACTG. Otras características químicas de la molécula pueden ser usadas para determinar el orden en que es leída. En **B** se muestran algunas configuraciones que puede tomar el ARN. Esta molécua también puede ser representada por una cadena de caracteres. Finalmente, en **C** se muestran dos posibles configuraciones básicas que pueden tener las proteínas, la representación en forma de cadena de caracteres de las proteínas implican un alfabeto diferente. Esquemás tomados de [44]

ser traducido a una proteína [48]. Los genes pueden encontrarse en cromosomas, que se definen como una molécula lineal de ADN. Finalmente, el **genoma** es la inclusión de todo el material genético de un organismo en un conjunto de cromosomas [26].

El genoma de las células procarióticas está conformado por bloques de genes precedidos por secuencias reguladoras llamadas promotores. Por otro lado, los genes de las células eucarióticas se parecen a un mosaico de secuencias codificantes (llamadas **exones**), segmentos de ADN transcritos a RNA que no son traducidos a proteínas (**intrones**), regiones pequeñas de ADN (**promotores**) a los cuales se adhieren otras moléculas para regular la expresión de un gen, y **regiones intragénicas** que no son traducidas a RNA ni codifican para alguna proteína [48].

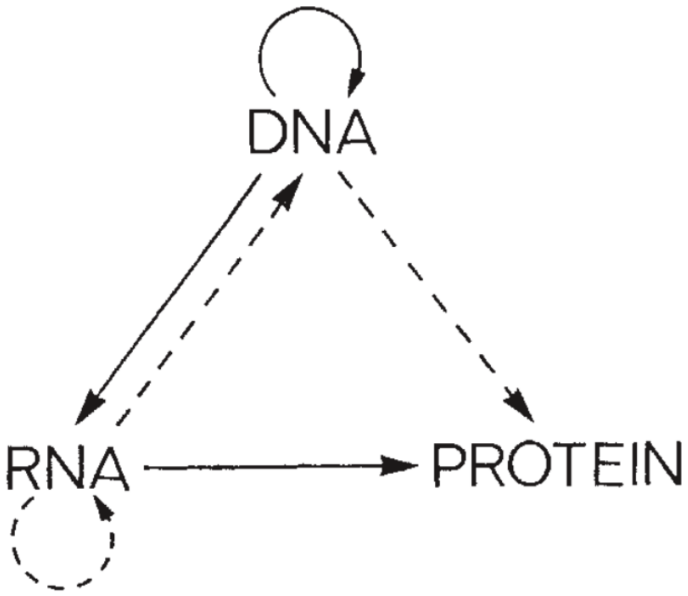


Figura 2.5: Diagrama que muestra la clasificación de las transiciones de información secuencial. Las flechas sólidas corresponden a las transiciones *generales*, las líneas punteadas son las *especiales*, mientras que las flechas que no están presentes son las *desconocidas*. Diagrama tomado de [14].

Muchos genes contienen múltiples exones separados por intrones. El uso diferencial de exones que pueden ser unidos de diferentes formas puede generar variantes de las proteínas después de la traducción [48]. En la figura 2.6 se esquematizan los componentes del genoma descritos.

2.1.6. Homología genética: ortología y paralogía

Dos **proteínas son homólogas** si divergieron de una proteína ancestral, y son **proteínas análogas** si convergieron de dos genes ancestrales diferentes. Con la llegada de la biología molecular se abrió a discusión la relevancia de estas clasificaciones, pues bajo el argumento de que *la bioquímica evolutiva ... puede mostrar la similitud de dos o más estructuras de proteínas, pero no tiene ni puede tener evidencia experimental relativa a genes ancestrales*, se aseguraba que las proteínas análogas y las homólogas son indistinguibles. Sin embargo Fitch demostró que la similitud entre proteínas existentes es consistente con la hipótesis de divergencia, y no con divergencia y convergencia mezclados [22]. Otro resultado importante de este trabajo es la identificación de dos subclases de homología, lo cual nos permite definir qué tipos de relaciones evolutivas existen entre genes:

- **Ortología**: Dos genes son ortólogos entre sí cuando su homología es el resultado de un evento de especiación.

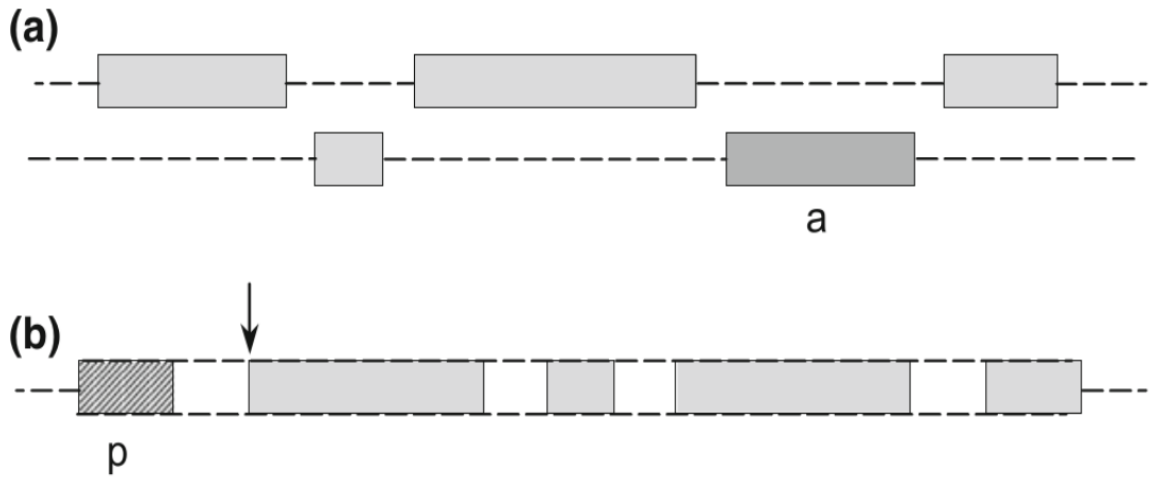


Figura 2.6: Esquema simplificado de la estructura genética de las células eukaryota. En **(a)** se muestra la estructura de doble hélice, donde los rectángulos representan genes, y las lineas punteadas son regiones intergenómicas. En **(b)** se aprecia una expanción de un sólo gen de **(a)**: los rectángulos sombreados corresponden a (exones) segmentos de ADN que son transcritos a ARN, posiblemente son unidos con otros segmentos y traducidos a proteínas, los rectángulos blancos corresponden a los intrones , mientras que **p** es un promotor.

- **Paralogía** mientras que dos genes son parálogos entre sí cuando su homología es el resultado de una duplicación de gen.

Estas definiciones no mencionan nada acerca de la función de los genes o de su distribución en los cromosomas. Sin embargo, una propiedad de los ortólogos teóricamente plausible y empíricamente soportada es que típicamente realizan funciones equivalentes en sus respectivos organismos. No obstante, la proposición inversa es mucho menos fuerte; proteínas que realizan funciones similares pueden ser catalogadas como no-ortólogas (e incluso no-homólogas) o genes que se obtuvieron por transferencia horizontal de genes entre especies tienen la misma función pero no se pueden catalogar como ortólogos. Por otro lado, los genes parálogos realizan funciones distintas, incluso si están relacionadas por su mecanismo. Esta diferenciación funcional ha sido estudiada teórica y experimentalmente en numerosos estudios. En este caso también se puede clasificar a los genes erróneamente cuando hay eventos de pérdidas o trasferencia horizontal [38].

2.2. Bases matemáticas

Las relaciones entre diferentes objetos pueden ser representados mediante un *grafo*, mientras que la teoría de grafos se encarga de caracterizar las propiedades de la definición de diferentes tipos de grafos. Si un grafo representa relaciones de ortología, entonces puede ser convertido a un árbol de genes que explica las relaciones por medio de una historia

evolutiva. En este capítulo se presentan las dos principales estructuras matemáticas usadas para la implementación de la metodología.

2.2.1. Grafos

Un **grafo** G es un conjunto V finito no vacío de objetos llamados *vértices* y un conjunto E de objetos llamados aristas, este último conjunto puede ser vacío [10]. Los vértices también son conocidos como *nodos* y las aristas como *relaciones*, conceptos que emplearemos en la descripción de este trabajo.

$$G = (V, E)$$

También es común denotar a los vértices y las aristas de G por las expresiones $V(G)$ y $E(G)$ respectivamente. Toda arista $e = (u, v)$ de G está constituida por un par de nodos u y v, y se dice que la arista e *une* o *relaciona* a tales nodos.

Un grafo puede ser representado por un diagrama como el que se muestra en la figura 2.7, en el que cada nodo (o vértice) es un punto, y hay una línea entre dos puntos si y sólo si existe una arista en $E(G)$ que una a los nodos correspondientes a los dos puntos dados.

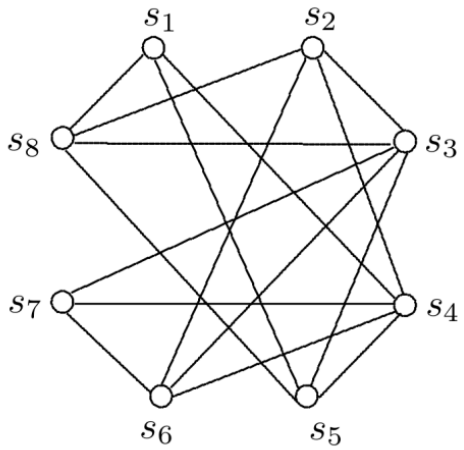


Figura 2.7: Diagrama de un grafo, tomado de [10].

Si (u, v) está en $E(G)$, entonces se dice que u y v son ***adyacentes*** o ***vecinos***. El conjunto de todos los vecinos de un nodo v es llamado la **vecindad de v**, y se denota por $N_G(v)$ o simplemente $N(v)$.

Los elementos de un grafo pueden representar muchas cosas. En el contexto específico de este trabajo usamos un **grafo de ortología**, donde los nodos corresponden a genes, y las aristas representan relaciones de ortología.

Dado un grafo $G = (V, E)$, y un subconjunto de nodos $S \subset V$, entonces el **grafo inducido** $G[S] = (S, E')$ es aquel cuyo conjunto de vértices es S, y el conjunto de aristas $E' \subseteq E$ cumple con la propiedad de que para toda arista $(u, v) \in E'$ pasa que $\{u, v\} \subseteq S$, es decir que las aristas sólo pueden existir entre miembros de S.

2.2.2. Árboles y filogenias

Un **árbol** es un grafo conexo sin ciclos $G = (V, E)$. En éste podemos distinguir i) las **hojas** $L(G)$, que son aquellos nodos con un solo vecino, y ii) los **nodos internos** $\bar{L}(G)$ conformados por todos aquellos nodos con al menos dos vecinos.

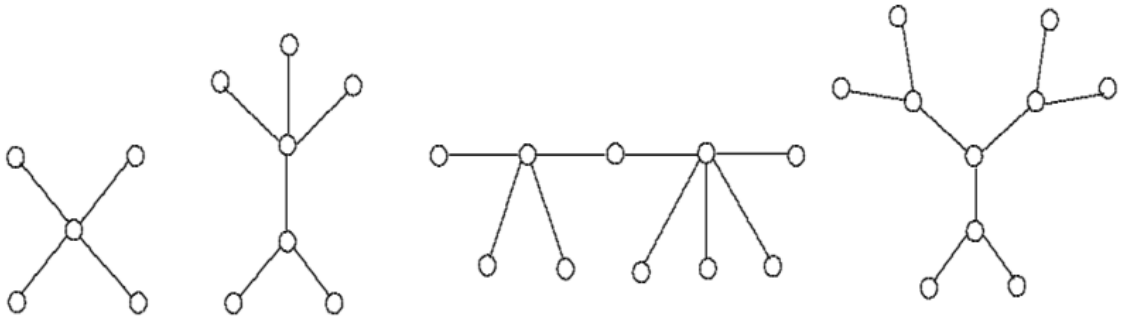


Figura 2.8: Cuatro árboles, tomados de [10].

Un *árbol con raíz* (**Rtree**) $T = (V, E; \rho)$ es un árbol con un nodo distinguido $\rho \in V$ llamado *nodo raíz*. El orden parcial $(u \preceq v)$ entre dos nodos u, v de un Rtree indica que u está en la ruta que va desde v hasta ρ, y se dice que u es **ancestro** de v. Si en este Rtree existe una arista (u, v) y además $u \preceq v$, entonces se dice que u es **padre** de v y de forma correspondiente, v es **hijo** de u. Debe notarse que el nodo raíz no tiene padre, mientras que las hojas (que no son ρ) no tienen hijos. Los RTrees pueden ser representados por diagramas como el presentado en la figura 2.9, donde el nodo raíz está hasta arriba, y las hojas están abajo. Un Rtree también puede ser representado como un grafo dirigido donde existe un arco de u a v si y sólo si u es padre de v.

Dado un árbol $T = (V, E)$, un conjunto $A \subseteq V$ de nodos y el conjunto de sus ancestros $B = \{u \in V(T) | u \preceq u' \forall u' \in A\}$, el **ancestro en común más reciente** $LCA(A)$ de A es el nodo único $b \in B | \nexists b' \in B, b \preceq b'$.

Un Rtree $T = (V, E)$ es llamado **compacto** cuando todos sus nodos internos tienen al menos dos hijos. Hay que notar que esta condición se cumple si y sólo si cada nodo interno es el ancestro común más reciente de al menos un par de hojas.

Un Rtree $T = (V, E)$ es **fechado** por un mapeo $r : V \to \mathbb{R}$ tal que

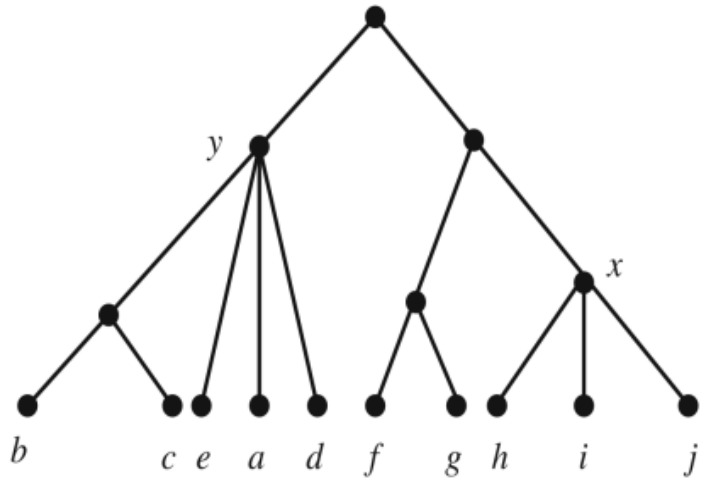


Figura 2.9: Rtree compacto. La raíz de este árbol es el nodo de arriba, mientras que las hojas son todos los de abajo. Podemos ver que $x = LCA(\{h, i, j\})$ y $y = LCA(\{a, b, c, d, e\})$. Figura tomada de [31].

I) $r(x) = 0 \forall x \in L(T)$

II) $r(u) > r(v)$ si $u \preceq v \forall (u, v) \in E$

Un **árbol filogénico** T (sobre X) es un árbol compacto con conjunto de hojas X. Si X es un conjunto de especies, entonces se tiene un **árbol de especies**, cuyos nodos internos representan especies ancestrales, mientras que si X es un conjunto de genes se tiene un **árbol de genes**.

Capítulo 3

Antecedentes

3.1. Teoría de la evolución de las especies: modelos y evidencia

En esta sección se resume el artículo [40] para presentar la teoría de la evolución de las especies, repasando las diferentes etapas en las que se realizaron aportes a la teoría, ya sea evidencia, nuevos conceptos o nuevas tecnologías, para finalizar ubicando nuestro aporte en la *síntesis post-moderna*.

Las principales proposiciones enunciadas por Darwin en su obra *El origen de las especies* se pueden resumir en la siguiente lista:

1. Las *acciones supernaturales del creador* son incompatibles con los hechos empíricos de la naturaleza.
2. Toda la vida evolucionó de uno o pocos tipos de organismos
3. Toda la vida evolucionó de variedades pre-existentes por medio de la selección natural.
4. El nacimiento de las especies es gradual y de larga duración
5. Los organismos actuales evolucionan por el mismo mecanismo que aquellos responsables del origen de las especies
6. A más grande similitud entre taxas, más estrecha es su relación evolutiva y es menor su divergencia desde su último ancestro en común.
7. La extinción es el principal resultado de la competencia inter-específica.
8. El registro geológico es incompleto: la ausencia de formas de transición entre especies y taxas superiores es debido a hoyos en nuestro conocimiento actual.

Con el paso del tiempo, surgió el término *darwinismo*, ampliamente impulsado por Wallace. Al equiparar tal concepto con *El origen de las especies*, se pueden obtener 5 ideas separadas:

- La evolución como un hecho
- Teoría del ancestro común
- Gradualismo
- Multiplicación de las especie
- Selección natural

En esos momentos los primeros dos puntos eran los que tenían menor sustento, sin embargo con el paso del tiempo diferentes científicos obtuvieron amplia evidencia para estos. Una comparación detallada de las publicaciones de Darwin y Wallace revelan que las contribuciones de Wallace fueron mucho más importantes de lo que usualmente se les atribuye, por lo que se considera apropiado llamar a la selección natural como *el mecanismo de selección natural de Darwin/Wallace*

La palabra **Darwinismo** se refiere específicamente al principio de Darwin/Wallace de selección natural como la mayor fuerza que actúa sobre la evolución. Posteriormente es anexado el término *Neo-Darwinismo*, donde A. Weismann refuta el mecanismo lamarckniano de herencia y postula que en la reproducción sexual se crea una nueva población variable de individuos por cada nueva generación. Después, en la *teoría sintética* se incorporan a la evolución hallazgos de campos como la genética, sistemática y paleontología. En la *síntesis expandida* se agregan disciplinas científicas como evolución experimental, fisiología, biología celular, etnología, paleobiología, simulación computacional, biología molecular, biología del desarrollo, sociobiología y geología, entre otras. Actualmente estamos en la *era de la post síntesis*, que comienza con el descubrimiento de la estructura del ADN y la publicación de comparación de secuencias de aminoácidos.

Al usar bases de datos que nos permiten comparar la información molecular de genomas para inferir historias evolutivas, nuestro trabajo entra dentro del campo de la síntesis postmoderna.

3.2. Genómica comparativa

Las tecnologías de secuenciación de genomas son cada vez más accesibles para la comunidad científica, a la vez que se va mejorando la precisión de sus resultados [53]. En este contexto surge la genómica comparativa, que podemos definir simplemente como el *conjunto de estudios que derivan en conocimientos biológicos por medio de la comparación de características genómicas* [55].

Un genoma tiene muchas características que pueden ser estudiadas como su secuencia, genes, el orden de los genes y secuencias de regulación. La genómica comparativa es una rama de la genómica que busca [55]:

1. Identificar las similitudes y diferencias de características genómicas y rastrear el origen, cambio y pérdida en las diferentes líneas evolutivas.

2. Entender las fuerzas evolutivas tales como la mutación, recombinación, transferencia horizontal de genes, e identificar la selección que gobierna cambios en las características genómicas.

3. Identificar cómo la evolución genómica nos puede ayudar a combatir enfermedades por medio de medicina personalizada, mejorando la salud ambiental, restableciendo el desarrollo sostenible, etc.

Se han alcanzado resultados importantes de la genómica comparativa. Por mencionar algunos ejemplos, al comparar los genomas de la mosca de fruta y del humano se encontró que cerca del 60 % de los genes están conservados y que dos tercios de los genes de humano involucrados en cáncer tienen contrapartes en la mosca de fruta [53]. La genómica comparativa estudia las similitudes y las diferencias entre especies a nivel de genes, proteínas, ARN y regiones reguladoras. Este tipo de análisis nos proporciona información de cómo la selección natural actúa en tales moléculas.

La identificación de los mecanismos evolutivos del genoma eucariota mediante genómica comparativa es uno de los objetivos importantes de esta área. Sin embargo, a menudo la genómica comparativa se sirve de organismos modelo (ratón para el estudio de humanos, por ejemplo) que son de gran importancia para los avances en nuestro conocimiento de los mecanismos generales de la evolución y de la genómica funcional. Por ejemplo en [7] se realizó la predicción de exones de humano por medio de la comparación de su genoma con el de ratón y tomando en cuenta la posición del gen; en el pez cebra se identificó una red genética conservada que parece estar involucrada en procesos de comportamiento, aprendizaje y memoria, agresión, ansiedad y sueño [45].

Dado que el uso las tecnologías de secuenciación implican cada vez un menor costo y son de fácil acceso, se encontrarán aplicaciones en la agricultura, biotecnología y zoología como una herramienta para identificar pequeñas diferencias entre diferentes especies de plantas y animales [53].

La genómica comparativa también puede traer consigo un re-arreglo de nuestro entendimiento de algunas ramas del árbol de la vida o bien nuevas estrategias para conservar especies exóticas y en peligro de extinción [53].

3.3. Análisis de la evolución de genes

Los escenarios evolutivos se dan con diferentes magnitudes que van desde eventos a pequeña escala como mutaciones puntuales e inserciones/pérdidas (**nivel nucleótido**) y a escala genética como pérdida y ganancia de genes (**nivel genético**), hasta eventos de gran escala como duplicación del genoma completo, inversiones, transposiciones y fusión/fisión de cromosomas (**nivel genómico**). Sin embargo, durante mucho tiempo, en la mayoría de los estudios de comparación de secuencias sólo se analizaron genes con una historia evolutiva simple, típicamente dónde aparece sólo una copia del gen en cada genoma (restringiendo el análisis al nivel nucleótido, pues sólo se mide disimilitud entre genes). Goodman et al.

[25] ampliaron el panorama cuando estudiaron la reconciliación de árboles de genes y de especies, que integra la evolución de familias de genes dentro del modelo de evolución de secuencias. Recientemente se han desarrollado también diversos modelos para la inferencia de árboles de genes y especies que integran la evolución de secuencias y eventos de inserción/pérdida de genes [11].

En general, los métodos actuales para el estudio evolutivo de genes consisten en su agrupación por similitud y la identificación de relaciones de ortología. Se han desarrollado diferentes métodos para la inferencia de filogenias de genes cómo la herramienta conocida como `PHYLIP` [20], que implementa un enfoque de máxima verosimilitud [21]. Variantes de este procedimiento se han implementado en algoritmos como `NJ`, `PAUP`, `PhyML`, `MrBayes` y `RAxML`, entre otros. Sin embargo estos métodos tienden a ser computacionalmente pesados y a tener errores con grandes conjuntos de datos, distancias evolutivas no uniformes, tasas de mutación variables, así como a introducir ruido con selecciones manuales de familias de genes [41, 39].

3.3.1. Inferencia de ortología

Existen diferentes métodos para abordar este problema, los cuales se basan en la definición de ortología o en sus consecuencias prácticas. Es decir, en la idea de que los genes ortólogos tienden a conservar su estructura y función dado que cumplen una función esencial en un organismo, mientras que un gen parálogo puede mutar libremente sin que el organismo pierda la función asociada a tal gen, permitiendo que con el paso del tiempo el gen adquiera otro rol en el organismo. Estas premisas no se cumplen necesariamente, sin embargo son una buena aproximación. A continuación se describen los tres principales métodos de inferencia de ortología.

Inferencia de ortología a partir de historia evolutiva Esto se hace por medio de la reconciliación de un árbol de genes con uno de especies y es la metodología teóricamente más plausible, pues su relación con la definición de Fitch es directa (ver figura 3.1). Sin embargo, la necesidad de tener a priori un árbol de genes es el factor limitante de este enfoque, pues tal proceso es propenso a tener errores con grandes conjuntos de datos, grandes distancias evolutivas e incluso los resultados se ven afectados por sesgos en los datos al incluir distancias evolutivas irregulares o con más organismos para ciertos clados [39].

Como ejemplo de este método, se presenta la metodología de la base de datos Ensembl [1]:

1. Se usa el algoritmo `DNADIST` de PHYLIP para inferir una matirz de distancias, con la que después se construye una primera filogenia usando un algoritmo de `fusión de vecinos` [24].
2. La filogenia obtenida es manipulada con el fin de maximizar la probabilidad de obtener los datos que alimentan al programa [29].

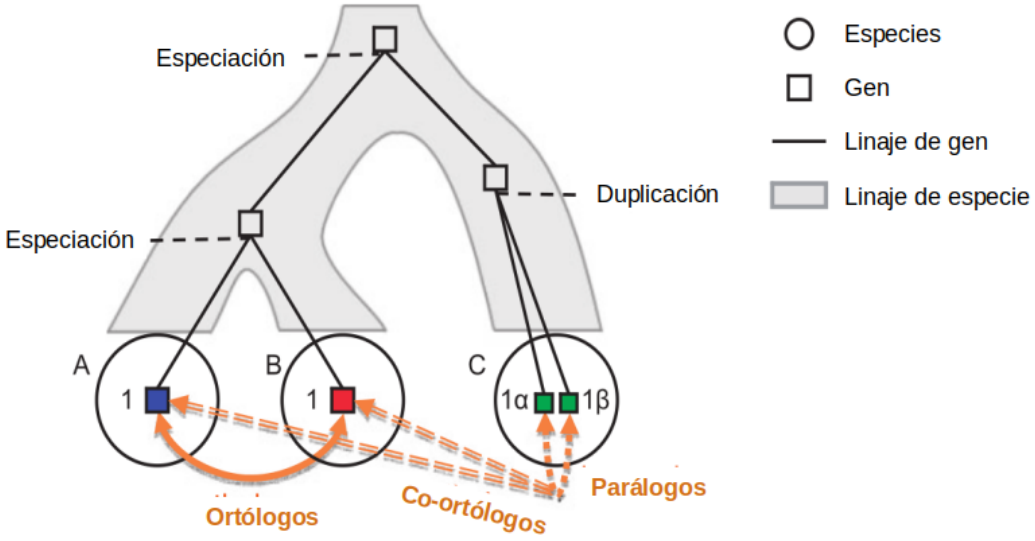


Figura 3.1: Reconciliación de árbol de genes con árbol de especies. Es fácil identificar los genes ortólogos y parálogos por la definición de Fitch, discutida en la sección 2.1.6. Imagen tomada de [39]

3. Finalmente se determina cuáles de los nodos internos del árbol de genes corresponden a eventos de duplicación y especiación por medio de la comparación de la filogenia de genes inferida y la filogenia de especies con el algoritmo de reconciliación `RAL` [19], con lo que se pueden identificar los genes ortólogos directamente por la definición de Fitch. En la figura 3.2 se muestra una forma de inferir eventos evolutivos por medio de la comparación de la topología de los árboles de genes y especies.

Inferencia de ortología a partir de comparación de pares de genes Este procedimiento hace uso del mejor alineamiento bidireccional (**BBH**), el cual no se basa en un modelo explícito de la evolución de los genes, sino en la premisa de que *los genes ortólogos son más parecidos entre sí que los que no son ortólogos*. Esto nos permite identifcar posibles ortólogos sin la necesidad de calcular un árbol de genes. Además de identificar pares de ortólogos, también existen métodos para agrupar conjuntos de pares de ortólogos. El uso de las BBHs evita los errores asociados con los métodos basados en reconciliación de árboles aunque tiene sus propias debilidades. En particular, es imposible identificar por este método la pérdida diferencial de genes, lo cual provoca que la hipótesis de la BBH sea falsa, pues genes que son parálogos aparentan ser ortólogos (ver figura 3.3). Las tasas variables de divergencia y la transferencia horizontal de genes también hacen que esta hipótesis sea falsa. En la figura 3.4 se muestran ejemplos de relaciones que se pueden inferir con este método, así como un escenario donde se pierden relaciones de ortología [39].

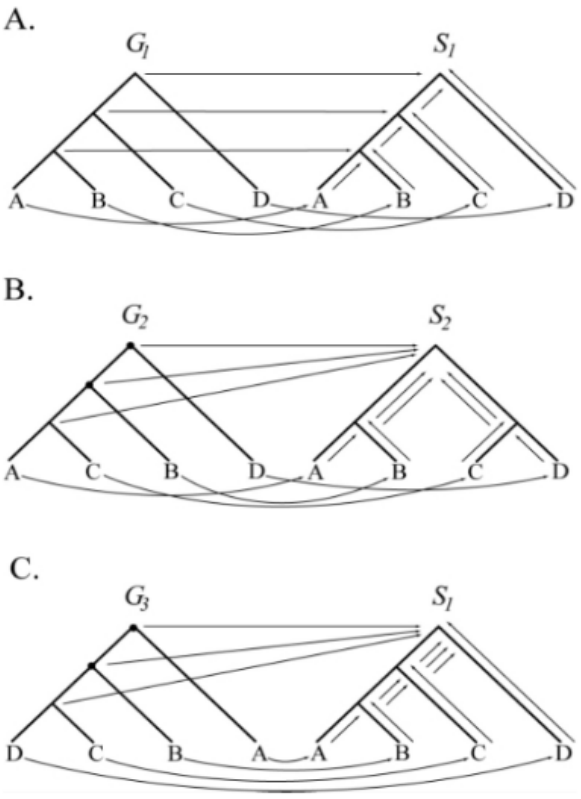


Figura 3.2: Ejemplos del proceso de inferencia de ortología y paralogía por medio de la reconciliación de un árbol de genes y uno de especies. Las filogenias G_1, G_2 y G_3 son árboles de genes, mientras que S_1, S_2 y S_3 son árboles de especies. A,B,C y D son las especies correspondientes a los genes. Podemos ver que entre los dos árboles hay un mapeo M que relaciona la topología de los árboles para determinar cuáles de los nodos de los árboles de genes corresponden necesariamente a eventos de duplicación (marcados como puntos negros). Este método es de máxima parsimonia, pues postula el mínimo número de pérdidas y duplicaciones para explicar al árbol de genes. Figura tomada de [56].

Inferencia de ortología por sintenia El análisis por sintenia se basa en la conservación del orden local de los genes, y es observada entre organismos cercanos. La posición de los genes es algo que evoluciona mucho más rápido que los genes mismos, por lo que la identificación de homología por medio de sintenia no es un método muy poderoso. Sin embargo, agrega sustento biológico a las relaciones de ortología que fueron inferidas por otros métodos, además de que puede ser usada para identificar pérdida o transferencia horizontal de genes [39].

3.3.2. Reconciliación

La historia de cada familia de genes es una serie compleja de eventos evolutivos que incluyen duplicación, pérdida y transferencia horizontal de genes. La diversidad de tales historias evolutivas posibles subraya la importancia de modelar los factores que afectan

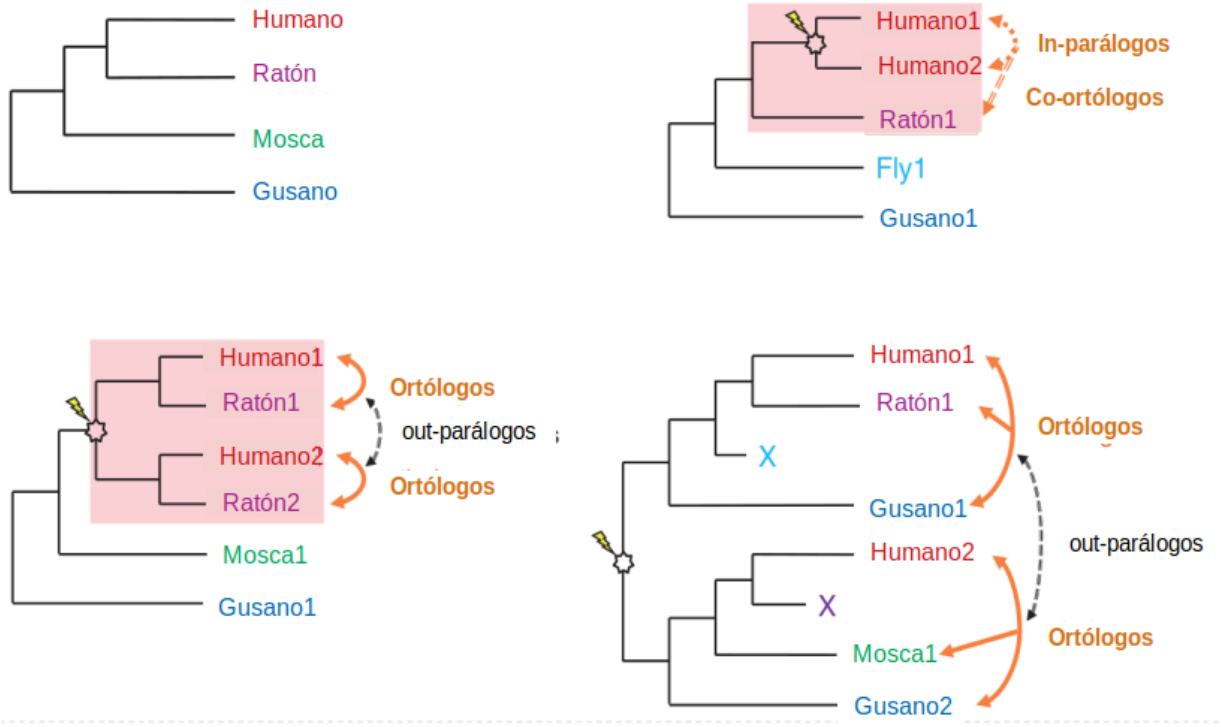


Figura 3.3: Ejemplos de posibles escenarios evolutivos en 4 especies. Arriba a la izquierda se ve una historia simple en donde existe un gen con un ortólogo en cada especie. A la derecha se observa una duplicación después de una especiación, generando in-parálogos de humano. Abajo a la izquierda se observa una duplicación antes de una especiación, formando out-parálogos, donde cada gen de humano o ratón es parálogo de un gen de humano y uno de ratón. Abajo a la derecha se muestra una duplicación ancestral seguida por pérdida diferenciada. En este último caso también se forman relaciones de ortología y paralogía, pero además hay que resaltar que en este tipo de eventos puede no cumplirse la BBH, pues los genes de mosca y humano parecerán genes ortólogos en lugar de parálogos. Diagrama tomado de [39].

la evolución de genes. **El proceso de reconciliación usualmente se usa para determinar relaciones de ortología y paralogía entre genes de una misma familia**. Los modelos de reconciliación consideran un árbol de especies dentro del cual evolucionan los genes, los cuales están asociadas a las hojas de la filogenia de especies y se explica su presencia mediante eventos específicos dentro de la misma [18].

Los diferentes modelos de reconciliación de árboles se pueden clasificar por los eventos evolutivos que contemplan, principalmente duplicación y pérdida (**DL**) o duplicación, transferencia horizontal y pérdida (**DTL**) de genes. La primera clasificación restringe los alcances de la herramienta a organismos multicelulares en los que los eventos de transferencia horizontal de genes son poco comunes [18].

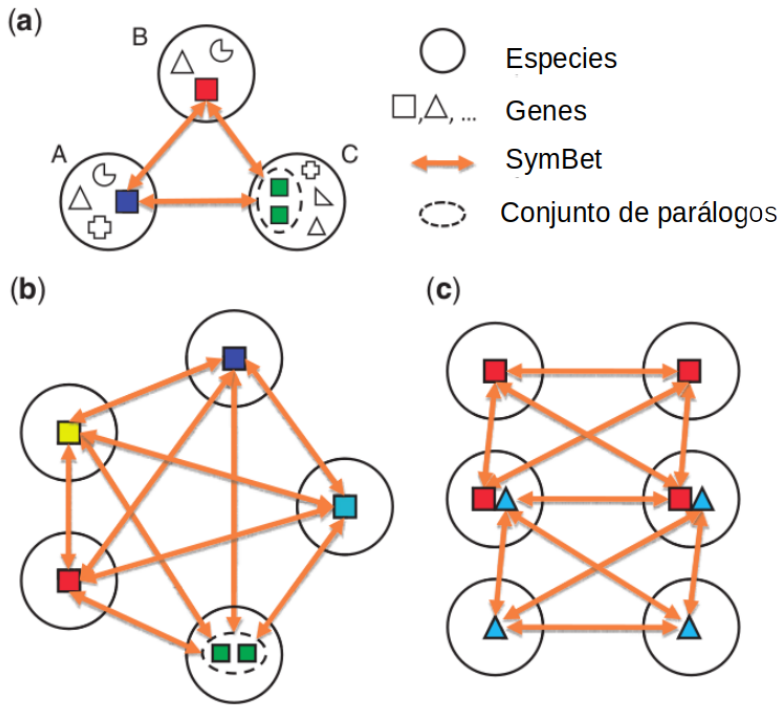


Figura 3.4: Posibles escenarios de relaciones de ortología inferidas usando la BBH. Figura tomada de [39]

Se han propuesto diferentes modelos para la reconciliación basados en parsimonia, probabilidad o fusión de vecinos (**NJ**). Los métodos basados en parsimonia obtienen una reconciliación óptima tomando en cuenta el costo de eventos evolutivos. Por otro lado, los métodos que abordan el problema desde una perspectiva probabilística usan modelos continuos para obtener una reconciliación que maximiza una función de verosimilitud.

La función de costo para los métodos basados en parsimonia toma un valor dependiendo de los eventos evolutivos considerados en una historia evolutiva dada y el objetivo de estas metodologías es encontrar el árbol con menor puntaje. En términos de velocidad computacional, estos métodos son más rápidos que los basados en probabilidad [6], algunos ejemplos son ecceTERA y `Notung` [35, 37]. La ventaja de los métodos probabilísticos es que permiten la estimación de parámetros en un marco de trabajo estadístico, donde se asocia una tasa (en lugar de costo) a los eventos evolutivos y es usada para calcular la probabilidad de diferentes escenarios. Una vez que se ha estimado un árbol de genes, éste puede ser evaluado con una función de verosimilitud, que es proporcional a la probabilidad de tener un árbol de genes dado uno de especies, lo cual plantea el problema de reconciliación como la búsqueda de parámetros del árbol de genes que maximizan dicha función [9]. Algunos ejemplos de herramientas de este tipo son `ALE_MCMC`, que explora el espacio de árboles reconciliados [34]. Por otro lado, `SPIMAP` [18] y `PrimeGSR` [18, 3] implementan

modelos Bayesianos. Otros métodos basados en **NJ** como `Treerecs` [12] usan información del árbol de especies y matrices de distancias para corregir las ramas de los árboles de genes con poco soporte experimental. Este método es determinista, por lo que tiene ventaja en complejidad computacional.

3.3.3. Inferencia de evolución a partir de ortología

La inferencia de ortología por medio de la BBH es el método más popular, el cual no necesita del conocimiento de un árbol de genes o de especies, por lo que Marc Hellmuth et al. [31] estudiaron el problema del árbol de genes de forma inversa; tratando de contestar la pregunta *¿Cuánta información acerca de los árboles de genes y especies y de su reconciliación está contenida en las relaciones de ortología de un conjunto de genes*?. Durante el desarrollo de su trabajo se demostró la equivalencia que existe entre las *ultra-métricas simbólicas* y lo cografos y discuten como se podrían usar sus resultados y algoritmos para estudiar conjuntos arbitrarios de relaciones de ortología [31].

El método anterior es capaz de inferir un árbol de genes fechado con eventos evolutivos partiendo únicamente de un conjunto de relaciones de ortología entre genes. La inferencia de relaciones de ortología por medio de la BBH tiene sus propios puntos débiles debidos a diferentes tasas de divergencia y pérdida de genes, por lo que la reconciliación del árbol de genes inferido con un árbol de especies nos puede agregar información adicional acerca de la evolución de los genes. Hernandez-Rosales et al. [32] reportaron que el problema de la reconciliación de los árboles de genes se puede inferir a partir de las relaciones de ortología. En este trabajo se describen de forma detallada el proceso de reconciliación y las condiciones bajo las cuales el árbol de genes es congruente con el árbol de especies y se muestra como estos árboles de genes etiquetados con eventos evolutivos contienen una gran cantidad de información dado un árbol de especies, incluso con grandes porcentajes de pérdida de genes [32].

3.4. Objetivos

- Crear una plataforma que implemente métodos computacionales y matemáticos para la inferencia y análisis de la evolución de familias de genes.
- Analizar datos generados por medio de historias evolutivas simuladas para validar la plataforma.
- Analizar la historia evolutiva de genes de un solo exón sin intrones del genoma de ratón para un mejor entendimiento de su papel en enfermedades de regulación, entre ellas el cáncer.

3.5. Justificación del trabajo

Los árboles de genes con eventos de especiación y duplicación así como su reconciliación con un árbol de especies aportan información relevante para el estudio evolutivo de familias de genes [32]. La inferencia de árboles de genes es uno de los puntos más débiles de los algoritmos de análisis evolutivo [32, 31], por lo que es importante desarrollar herramientas que permitan inferir y estudiar homología de forma automatizada con metodologías eficientes y precisas.

El análisis evolutivo de los genomas nos permite contrastar su estrcutura y el arreglo de genes así como sus características en diferentes especies. Dado que los genes de un solo exón que no poseen intrones (IGs) están vinculados con procesos de regulación, las mutaciones o alteraciones que sucedan en estos pueden asociarse con enfermedades severas como cáncer. Aunado a esto, es de destacar su alto potencial como posibles biomarcadores para el diagnóstico o como blancos terapéuticos en dichas enfermedades [27, 4, 43]. Pese a su relevancia biológica, la evolución de estos genes ha sido poco estudiada y las actuales bases de datos no contienen una clasificación rigurosa de este conjunto de genes [36]. En este contexto, consideramos pertinente analizar a los IGs. Los resultados obtenidos mediante este estudio podrán repercutir en el futuro en el conocimiento de su relevancia biológica, sobre todo en aquellos procesos ligados a patologías como el cáncer.

Capítulo 4

Metodología

Los tres principales pasos de la metodología propuesta son *i*) inferencia de ortología, *ii*) obtención de árboles de genes, y *iii*) reconciliación de árboles (figura 4.1). A lo largo de este capítulo se detalla la teoría y los procedimientos necesarios para realizar cada paso, así como para la identificación y corrección de ruido en las relaciones de ortología inferidas.

4.1. Predicción de grafos de ortología

`Proteinortho` es una herramienta optimizada para la inferencia de ortólogos por medio de la BBH que es capaz de trabajar con cientos de especies que contienen miles de secuencias. Esta herramienta usa algoritmos como `BLAST` para identificar pares de proteínas o genes (co-)ortólogos y posteriormente agrupar tales relaciones por medio de un algoritmo de partición espectral del grafo [42]. La figura 4.2 muestra la metodología de tal herramienta. En versiones nuevas de Proteinortho se agregan nuevos algoritmos que sustituyen a blast y al método de agrupamiento por otros más rápidos y con mayor sustento desde la perspectiva biológica [2].

Se pueden tomar las relaciones inferidas y obtener una primera aproximación del grafo de ortologia de los genes que se pasaron a Proteinortho, donde cada nodo representa a un gen, y las aristas indican que se predijo ortología entre dos genes.

Las relaciones de ortología son una propiedad que *se tiene o no se tiene*, o en otras palabras, no se puede decir que los genes a y b son más ortólogos que a y c, sin embargo, la similitud entre genes se puede definir en un espectro continuo, de tal forma que sí podríamos decir que a y b se parecen más entre sí que a y c.

La similitud entre los genes se puede determinar mediante el parámetro *bit score* arrojado para cada relación de ortología predicha por Proteinortho. Tal parámetro indica qué tan bueno fue el alineamiento de dos secuencias; entre mayor sea su valor mejor fue el alineamiento. Esta información será útil durante la edición de grafos, por lo que se guarda como el peso de las aristas.

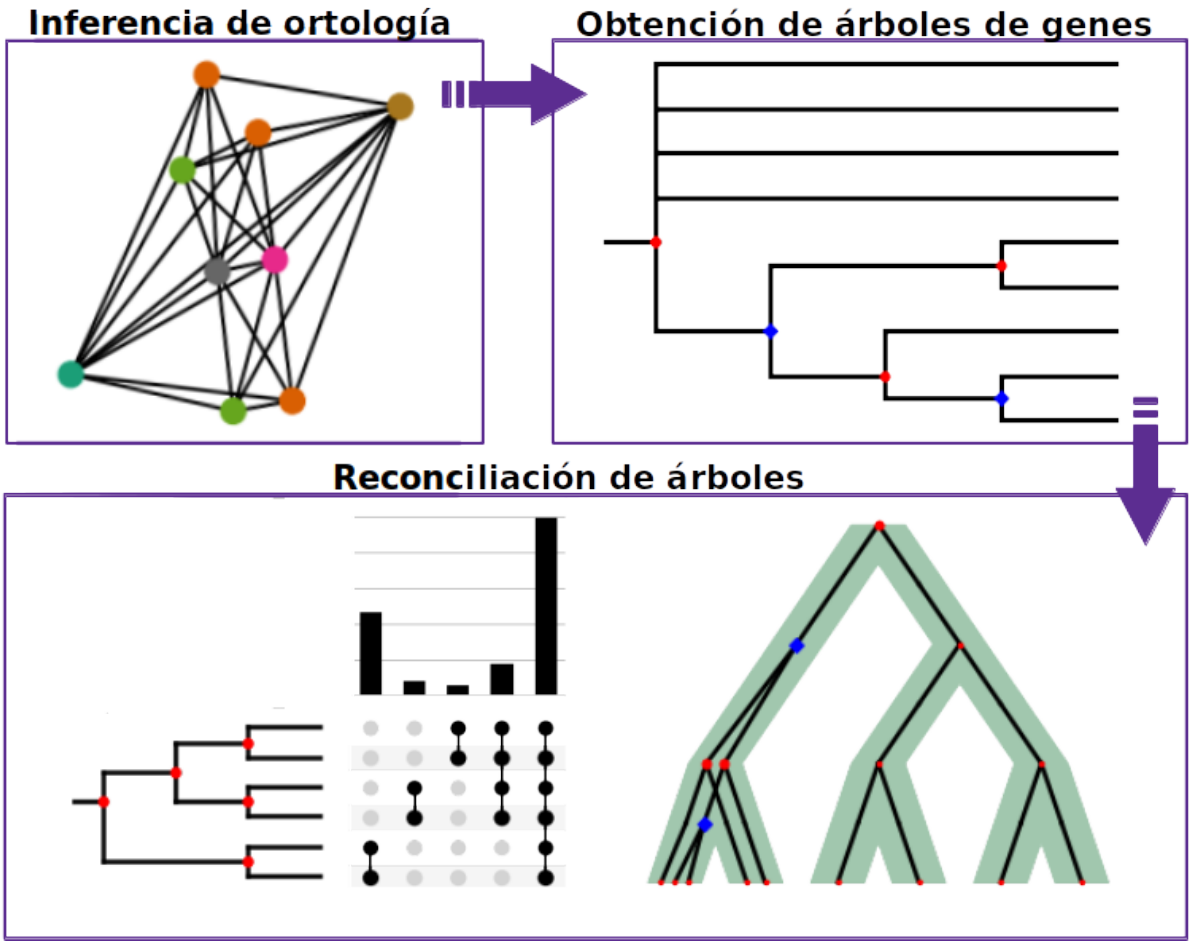


Figura 4.1: Diagrama de flujo de REvolutionH-tl

4.2. Identificación y edición de grafos de ortología

Las relaciones de ortología de un conjunto de genes siempre forman un cografo, lo cual resalta la relevancia de su identificación, que se puede realizar por diferentes métodos [13, 31]. En esta sección se detallan las características de tales estructuras, así como los métodos que existen para su identificación y transformación en otros objetos matemáticos.

4.2.1. cografos y coárboles

Los **cografos** (*complement reducible graphs*) son una familia de grafos que emergieron en diferentes áreas de las matemáticas y han sido redescubiertos independientemente por varios investigadores [13]. Esta familia de grafos puede ser construida usando únicamente la operación *complemento* aplicada a la unión disjunta de otros cografos, siendo el cografo

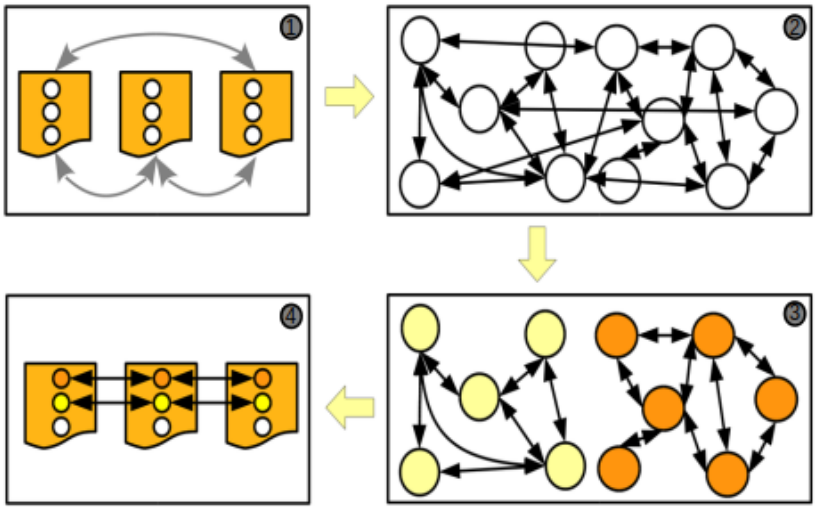


Figura 4.2: Se muestra la metodología implementada por Proteinortho: 1. Comparación de genomas, 2. Creación del grafo de mejor alineamiento, 3. Algoritmo de agrupamiento 4. Listas de ortología. Figura tomada de [2]

más simple aquel que contiene un solo nodo y ninguna arista. De forma inversa, un cografo con más de un nodo y una arista puede ser descompuesto en nodos sin aristas por medio de la obtención recursiva del complemento de sus componentes conexas. Esta última propiedad es la que asigna el nombre a dicha familia de grafos y toma relevancia porque permite representar a cualquier cografo de forma única mediante un árbol llamado **coárbol**. En la figura 4.3 se puede observar la descomposición de un cografo aplicando el método descrito arriba. Además se muestra como tal proceso resulta en la formación de un árbol. En los párrafos siguientes se describen más formalmente las propiedades de los cografos, así como la construcción del coárbol por medio de la defínción de cografo y en capítulos posteriores veremos un método más eficiente para obtener árboles incluso de no-cografos.

Abajo se definen de forma recursiva a los cografos, mientras que en la figura 4.4 se observan ejemplos de los mismos.

1. Un grafo de un único vértice es un cografo.
2. Si $G_1, G_2, G_3, ..., G_n$ son cografos, entonces su unión también es un cografo.
3. Si G es un cografo, entonces $\bar{G}$ también es cografo.

Como se mencionó arriba, esta familia de grafos fue caracterizada por las matemáticas de diferentes formas, que son equivalentes entre sí. A continuación se enumeran algunas propiedades que son de utilidad para este trabajo.

1. Un cografo no contiene un subgrafo inducido P_4 [13].
2. Todo subgrafo conexo de G tiene diámetro ≤ 2 [13].

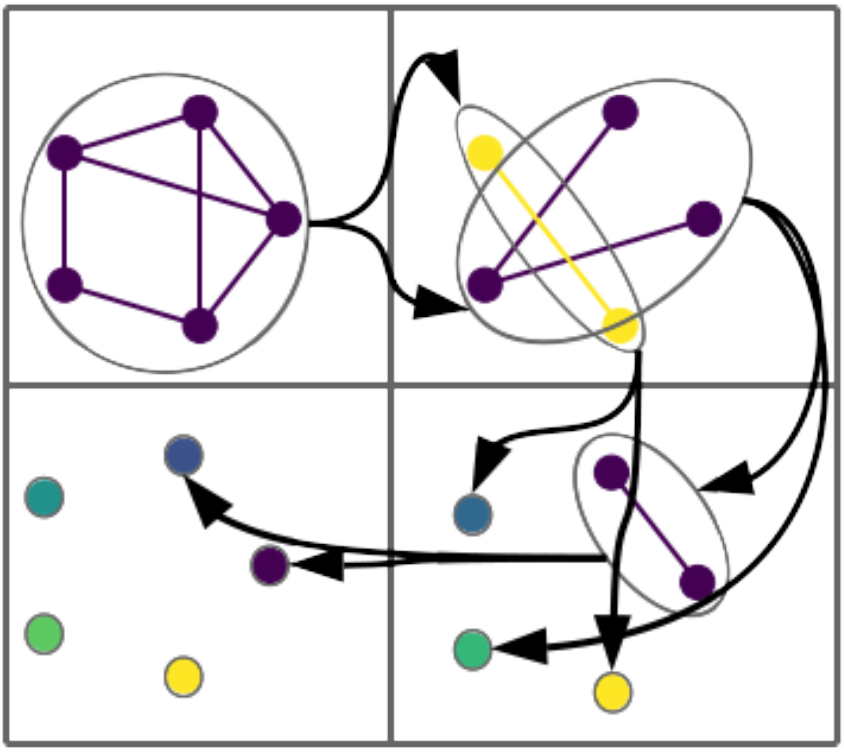


Figura 4.3: Descomposición de cografo en nodos sin aristas por medio de complemento recursivo de sus componentes conexas. Cada cuadrante gris contiene un cografo, que puede tener una o más componentes conexas, las cuales se distinguen por su color, además de que los nodos de una misma componente conexa están encerrados en una elipse gris. Las flechas negras indican cómo al obtener el complemento de un cografo conexo, éste se divide en más de una componente conexa y, de forma inversa, al obtener el complemento de un cografo no conexo, se obtiene una sola componente conexa. El grafo a descomponer en este ejemplo es el que se encuentra en el cuadrante superior izquierdo. Es importante notar que este grafo es completamente morado, pues tiene una sola componente conexa, sin embargo también se pueden descomponer grafos con más de una (por ejemplo el grafo de la esquina superior derecha). De igual manera, es posible construir el grafo morado partiendo únicamente de nodos sin aristas y siguiendo las flechas en orden contrario (obteniendo el complemento de cografos no conexos). Finalmente, resalta que se puede formar un árbol donde los nodos son las elipses grises y las aristas se representan con las flechas, siendo el nodo raíz el grafo morado, y las hojas los nodos sin aristas.

3. Todo cografo se puede representar con ultramétricas simbólicas (ver sección 4.2.2).

4. La descomposición modular de un cografo no contiene módulos primos (ver sección 4.2.3).

Las propiedades 1 y 2 nos resultarán de ayuda para implementar un método de corrección de no-cografos, el punto 3 nos brinda la relación que existe entre las relaciones de ortología y los cografos, mientras que el 4 proporciona una forma de identificar cografos de forma rápida, así como una heurística que permite simplificar el problema de edición

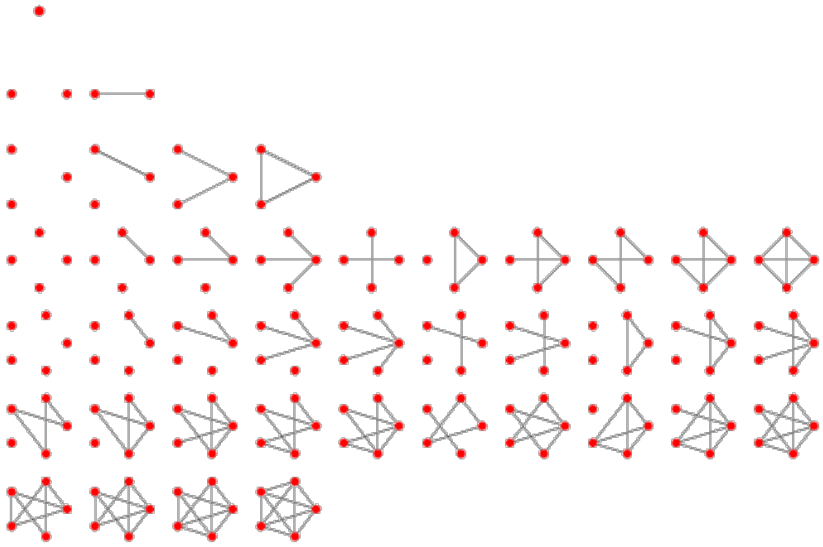


Figura 4.4: Todos los posibles cografos (salvo isomorfismos) con $n = 1, 2, 3, 4, 5$ nodos. Tomada de `https://mathworld.wolfram.com/Cograph.html`.

de no-cografos.

La definición recursiva de los cografos muestra que es posible crear cualquier cografo simplemente haciendo operaciones de unión y complemento entre cografos. Para representar cualquier cografo de forma única se introduce la noción de la *forma normalizada de un cografo*:

- **cografo conexo en forma normalizada**: Está expresado como un sólo nodo:

  $$G = (\{v\}, \{\})$$

  o como el complemento de la unión de al menos 2 cografos conexos en forma normalizada:

  $$G = \bar{\cup}(G_0, G_1, ...)$$

- **cografo no conexo en forma normalizada**: Se representa como el complemento de un cografo conexo en forma normalizada:

  $$G = G_1^C$$

El árbol con raíz que representa la estructura de un cografo en forma normalizada es llamado **coárbol**, este árbol contiene a los nodos del grafo en sus hojas, mientras que sus nodos internos representan la operación $\bar{\cup}$.

Podemos etiquetar a los nodos internos del coárbol de la siguiete manera:

- El nodo raíz tiene etiqueta 1.

- Los hijos de nodos con etiqueta 1 tienen etiqueta 0.
- Los hijos de nodos con etiqueta 0 tienen etiqueta 1.

Esto nos permite conocer algunas propiedades del grafo de forma rápida. Por ejemplo, hay una arista entre los nodos u, v del cografo si $LCA(u, v)$ tiene etiqueta 1. En la figura 4.5 podemos observar un cografo y su respectivo coárbol.

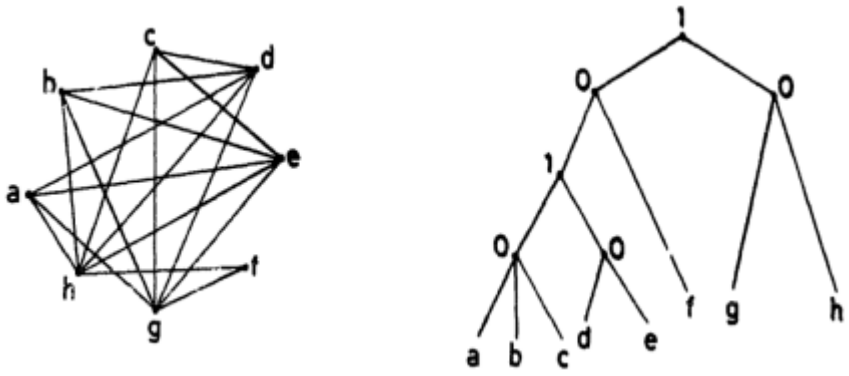


Figura 4.5: cografo y su respectivo coárbol. Figura tomada de [13].

4.2.2. Ultramétricas simbólicas, ortólogos y parálogos

Como ya se mencionó, **todo grafo de ortología es un cografo**. Esta propiedad es importante para inferir árboles a partir de las relaciones de ortología. En esta parte definimos las ultramétricas simbólicas, que nos ayudarán a representar la evolución de genes en términos de eventos evolutivos, así mismo se muestra la relación entre los árboles de genes y los grafos de ortología.

Dado un conjunto arbitrario X y un conjunto de símbolos M, designamos al símbolo $\odot$ como *no-evento*, y al conjunto $M^{\odot} = M \cup \{\odot\}$ como un conjunto de símbolos extendido que es introducido únicamente por razones técnicas. Una **ultramétrica simbólica** es una función $\delta : X \times X \to M^{\odot}$ que cumple con:

- **(U0)**: $\delta(x, y) = \odot$ si y sólo si $x = y$
- **(U1)**: $\delta(x, y) = \delta(y, x)$, es decir que δ es simétrica.
- **(U2)**: $|\{\delta(x, y), \delta(x, z), \delta(y, z)\}| \leq 2$ para todo $x, y, z \in X$
- **(U3)**: No existe un subconjunto $x, y, u, v \in \binom{X}{4}$ tal que

$$\delta(x, y) = \delta(y, u) = \delta(u, v) \neq \delta(y, v) = \delta(x, v) = \delta(x, u)$$

Dada la filogenia $T = (V, E)$ sobre X, un **fechado simbólico** sobre T es un mapeo $r : V \to M^{\odot}$ tal que $r(x) = \odot \forall x \in X$. Se dice que el fechado es **discriminante** cuando $r(u) \neq r(v) \forall (u, v) \in E$. Al par (T, r) le podemos asociar el mapeo $d_{(T,r)} : X \times X \to M^{\odot}$, definido de la siguiente manera:

$$d_{(T,r)}(x, y) = LCA(\{x, y\})$$

El par (T, r) es llamado **representación simbólica** del mapeo $\delta : X \times X \to M^{\odot}$ si $\delta(x, y) = d_{(T,r)}(x, y)$ para todo $x, y \in X$, y se dice que es discriminante si r es discriminante. En la figura 4.6 se muestra una representación simbólica. Bocker y Dress [8] demostraron que todas las ultramétricas simbólicas tienen representación simbólica discriminante y es única salvo isomorfismos.

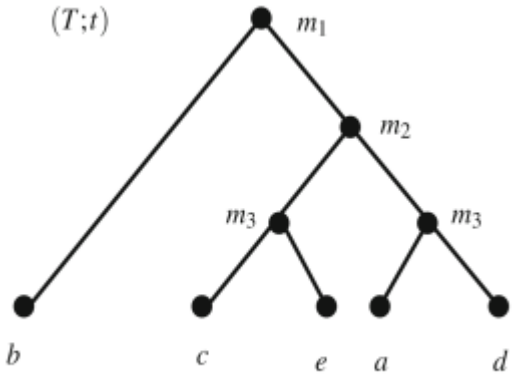


Figura 4.6: Representación simbólica (T, t) sobre el conjunto de hojas $L(T) = \{a, b, c, d, e\}$ y el fechado simbólico $t : V(T) \to \{m_1, m_2, m_3, \odot\}$. [31]

Se puede representar la historia evolutiva de genes mediante una representación simbólica (T, r), donde las hojas $L(T)$ corresponden a genes y los nodos internos $\bar{L}(T)$ corresponden a eventos evolutivos de especiación o duplicación, obteniendo un **árbol de genes fechado con eventos evolutivos**.

Todos los coárboles son una representación simbólica de un mapeo $\delta : X \times X \to \{0, 1\}$ donde $\delta(x, y) = 0$ si x y y tienen una arista en el cografo correspondiente al coárbol, y $\delta(x, y) = 1$ si x y y no tienen arista. De aquí se puede demostrar que toda representación simbólica puede ser construída a partir de coárboles, por lo que todos los árboles de genes fechados con eventos evolutivos pueden ser representados por un cografo. En la figura 4.7 se muestran los 3 coárboles que componen a la representación simbólica presentada en la figura 4.6.

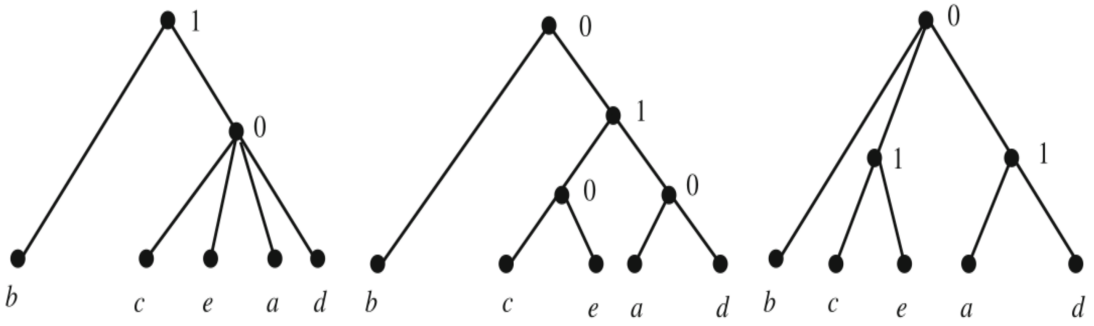


Figura 4.7: cografos que componen a la representación simbólica presentada en la figura 4.6. Figura tomada de [31].

4.2.3. Descomposición modular

La descomposición modular es un método que permite no sólo identificar cografos, sino que además provee información relevante acerca de las estructuras de no-cografos. Usamos tal información para identificar y editar las relaciones de ortología inferidas que pudieran ser erróneas.

Dado un grafo $G = (V, E)$, un subconjunto de nodos $X \subseteq V$ es un **módulo** de G si para cualquier nodo $y \notin X$ pasa que $X \subseteq N(y)$ o $X \cap N(y) = \varnothing$, es decir que los vecinos de y son un subconjunto del módulo X o ningún vecino de y forma parte de X.

Podemos ver que tanto los conjuntos de un solo nodo, como el conjunto de todos los nodos son módulos de cualquier grafo, por lo que son llamados **módulos triviales**. Si un grafo sólo contiene módulos triviales, entonces se dice que es un **grafo primo**. En la figura 4.8 podemos ver los dos grafos primos más simples con más de un nodo.

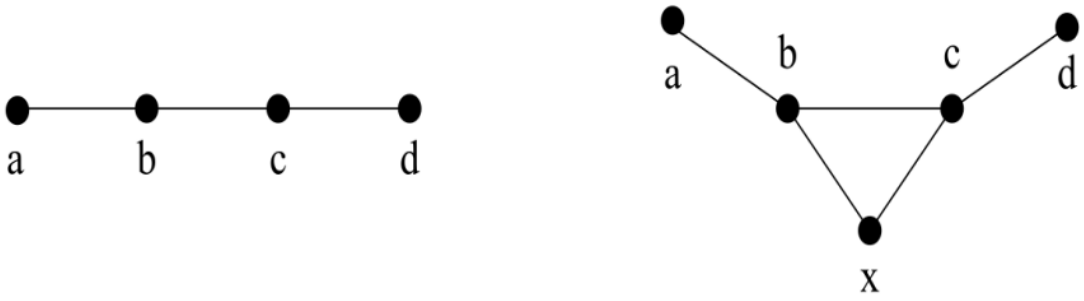


Figura 4.8: Dos grafos primos, es decir que su descomposición modular sólo contiene módulos triviales. En ambos casos podemos encontrar un $P_4 = \{\{a, b\}, \{b, c\}, \{c, d\}\}$. Del lado derecho vemos a un toro, donde el nodo x es conocido como la *nariz*. Figura tomada de [30].

Un módulo X es **fuerte** si no existe otro módulo X' tal que $X \not\subseteq X' \wedge X' \not\supset X$. Es decir, que X no se solapa con otro módulo. Un módulo X es **máximo** respecto a un conjunto S si $M \subset X$ y no existe otro módulo X' tal que $X \subset X' \subset S$. En la parte izquierda de la figura 4.9 se ejemplifican estos conceptos.

Sea $\mathcal{P}$ una partición de los nodos del grafo $G = (V, E)$. Si todos los elementos de $\mathcal{P}$ son un módulos fuertes, entonces se dice que $\mathcal{P}$ es una **partición modular** de G. Si además todos los módulos en $\mathcal{P}$ son máximos, entonces es una **partición modular máxima**.

El **árbol de descomposición modular** $\mathrm{MD}(G)$ de G es el árbol de inclusión de los módulos fuertes y máximos de G.

Dada una partición $\mathcal{P}$ de los nodos de $G = (V, E)$, para estudiar las propiedades de la descomposición modular, definimos el *quothent graph* $G_{/\mathcal{P}} = (\mathcal{P}, E')$ como el grafo donde $(X_1, X_2) \in E'$ si y sólo si para cualquier par de nodos $u \in X_1$ y $v \in X_2$ existe la arista $(u, v) \in E$.

Dada la la partición modular máxima $\mathcal{P}_G = \{X_1, X_2, ...\}$ de G, si se selecciona arbitrariamente un nodo de cada parte $P = \{x_1 \in X_1, x_2 \in X_2, ...\}$, el grafo inducido $G[P]$ siempre tendrá la misma forma que $G_{/\mathcal{P}}$. Además podemos ver que dado un nodo $X \in V(\mathrm{MD}(G))$, sus nodos hijos son la partición modular máxima $\mathcal{P}_{G[X]}$ del grafo inducido de G por X.

Todo módulo $X \in \mathrm{MD}(G)$ del grafo G se puede clasificar de la siguiente manera:

- **Paralelo** si $G[X]$ no es conexo.
- **Serie** si $G[X]^C$ no es conexo.
- **Primo** si tanto $G[X]$ como $G[X]^C$ son conexos, y el quotient $G[X]_{/\mathcal{P}}$ es un grafo primo (es decir, que sólo contiene módulos triviales), donde $\mathcal{P}$ es la partición modular máxima de $G[X]$.

Los módulos paralelos y series también son conocidos como módulos **degenerados**. En la figura 4.9 se muestra un ejemplo de esta clasificación.

Los grafos cuya descomposición modular no contiene módulos primos son cografos, donde los nodos etiquetados con '0' en el coárbol corresponden a los módulos serie, mientras los etiquetados con '1' denotan módulos paralelos. Usando las propiedades de los cografos, podemos decir que un módulo primo contiene un P_4 inducido.

Si se desea identificar qué provoca que un grafo G sea un módulo primo y se sabe que su partición modular máxima es $\mathcal{P}$, es conveniente analizar el grafo $G_{/\mathcal{P}}$, pues contiene menos nodos y aristas que G. La figura 4.10 presenta un ejemplo de esta noción.

Recordamos que si G es un grafo de ortología, entonces existe una representación simbólica (T, r) de las aristas del grafo, donde a cada nodo interno $u \in L(T)$ del árbol representa

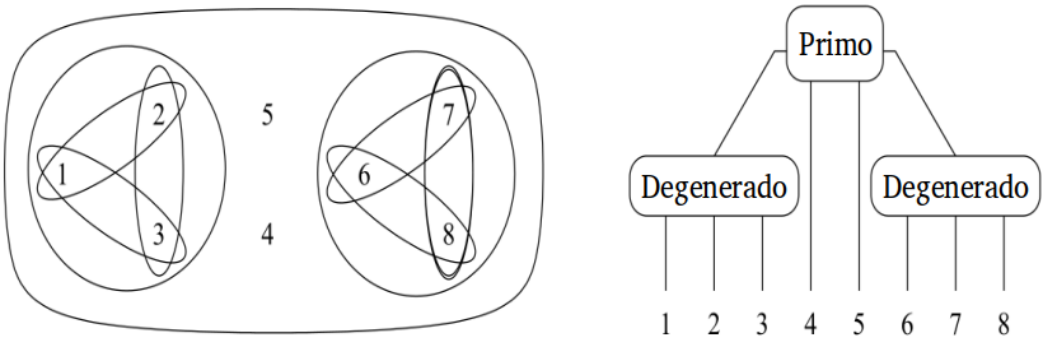


Figura 4.9: En la izquierda se puede ver la partición modular de un grafo arbitrario $G = (V, E)$ con nodos $V = \{1, 2, 3, 4, 5, 6, 7, 8\}$, y su árbol de descomposición modular $\mathrm{MD}(G)$, donde los módulos degenerados pueden ser series o paralelos. Es importante notar que la jerarquía designada por $\mathrm{MD}(G)$ sobre V no denota ningún módulo que solape a otro, es decir que no se incluyen los módulos $\{\{1, 2\}, \{2, 3\}, \{3, 1\}, \{6, 7\}, \{7, 8\}, \{8, 6\}\}$. La partición modular máxima de éste grafo es $\{\{1,2,3\},4,5,\{6, 7, 8\}\}$. Figura tomada de [30].

un evento evolutivo. En este caso los eventos de especiación corresponden a módulos series de G, mientras que los eventos de duplicación corresponden a módulos paralelos.

4.2.4. Edición de grafos a cografos

Ya vimos que si la descomposición modular $\mathrm{MD}(G)$ de un grafo G contiene un módulo primo $X \in \mathrm{MD}(G)$, y $\mathcal{P}$ es la partición modular máxima de $G[X]$, entonces el grafo $G_{/\mathcal{P}}$ es primo y tiene un P_4 inducido.

Para convertir un grafo $G = (V, E)$ a un cografo, definimos el *quotient graph* $G_{/\widetilde{\mathcal{P}}} = (\mathcal{P}, E')$ donde existe una arista $(P', P'') \in E'$ si y sólo si $\exists (u, v) \in E$ tal que $u \in P$' y $v \in P''$. Así mismo, el peso de cualquier arista $(P', P'') \in E'$ es igual a la suma del peso de las aristas $(u, v) \in E$ tales que $u \in P$' y $v \in P''$.

Dado que se busca eliminar todos los P_4s inducidos del grafo, o de forma equivalente, hacer que el diámetro del grafo sea menor o igual a dos, resulta útil aplicar el corte mínimo [51] del grafo $G_{/\widetilde{\mathcal{P}}}$. Esto no asegura que los sub-grafos resultantes sean cografos, sin embargo es de esperarse que esto pase dado que el algoritmo de corte mínimo parte al grafo, lo que puede resultar en reducir su diámetro. En la figura 4.11 se aprecia un ejemplo de esta edición. Notar que el corte mínimo de $G_{/\widetilde{\mathcal{P}}}$ es el conjunto de aristas cuya suma de pesos es mínima y que al ser retiradas parten al grafo en dos. Esto quiere decir que la edición que se implementa aquí retira aquellas aristas de G que en conjunto tienen la menor evidencia de ser ortología.

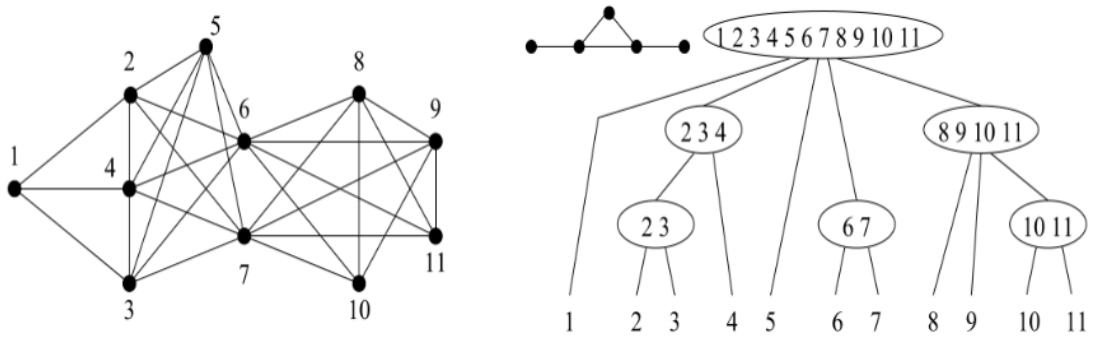


Figura 4.10: En la izquierda se ve un grafo arbitrario, a la derecha encontramos su árbol de descomposición modular MD(G). A la izquierda del nodo raíz de MD(G) podemos observar al grafo $G_{/\mathcal{P}}$, que es el grafo primo llamado *toro*, y que fue presentado en la figura 4.8. $\mathcal{P} = \{1, \{2, 3, 4\}, 5, \{6, 7\}, \{8, 9, 10, 11\}\}$ es la partición modular máxima de G, o los hijos del nodo raíz de MD(G). Figura tomada de [30].

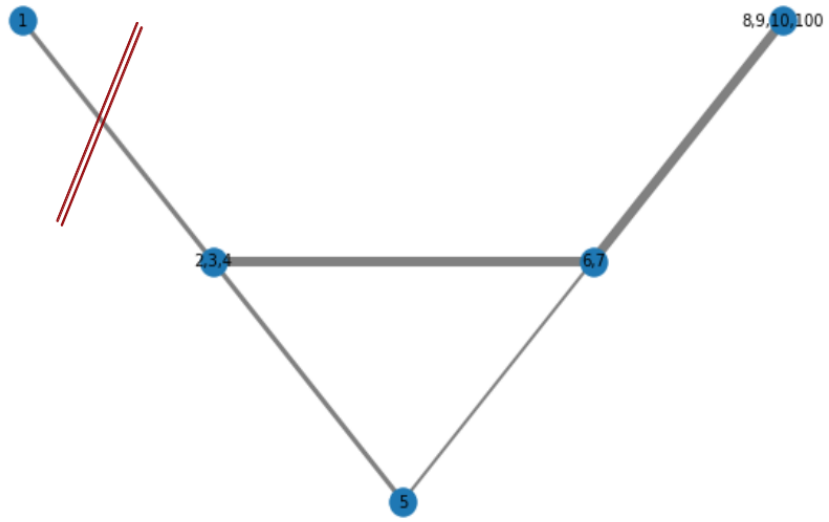


Figura 4.11: Ejemplo de corte mínimo del grafo $G_{\widetilde{/\mathcal{P}}}$ calculado a partir de la gráfica G de la figura 4.10. En este caso $G_{\widetilde{/\mathcal{P}}}$ tiene la misma forma que $G_{/\mathcal{P}}$, con la diferencia de que el mostrado aquí es un grafo pesado. Notar que la arista que conecta a los nodos (5) y (6,7) es la que tiene menor peso, sin embargo al retirarla no se obtiene un grafo no conexo, por lo que el algoritmo decide retirar (1)-(2,3,4), y el grafo resultante es un cografo. Cualquier otra combinación de aristas resultaría en un corte con más peso, es decir, que se estarían quitando relaciones de ortología con mayor evidencia.

Aplicar este proceso sobre el grafo $G_{\widetilde{/\mathcal{P}}}$ representa una ventaja sobre la carga computacional del cálculo del corte, pues se tiene una menor cantidad de nodos y aristas, por otro lado, no se editan aquellas relaciones de ortología que ya forman parte de un módulo no

primo más pequeño que el módulo primo en cuestión. Es decir que para cualquier $P \in \mathcal{P}$ pasa que si $u, v \in P$, entonces ya está determinado si existe una arista entre los nodos u, v y no necesita revisarse.

4.3. Historia evolutiva, congruencia y reconciliación de árboles

Un filogenia T en el conjunto de genes L tiene origen en una serie de eventos a lo largo de un árbol de especies S en el conjunto de especies B. Asumimos que $|L| \geq 3 \wedge |B| \geq 1$. sólo consideramos eventos de duplicación y pérdida de genes, que ocurren a lo largo de las aristas de S, mientras que los eventos de especiación son modelados como la transmisión de un gen en un linage ancestral a cada uno de sus linajes hijos. En la figura 4.12 podemos observar un ejemplo de un árbol de genes embebido dentro de uno de especies.

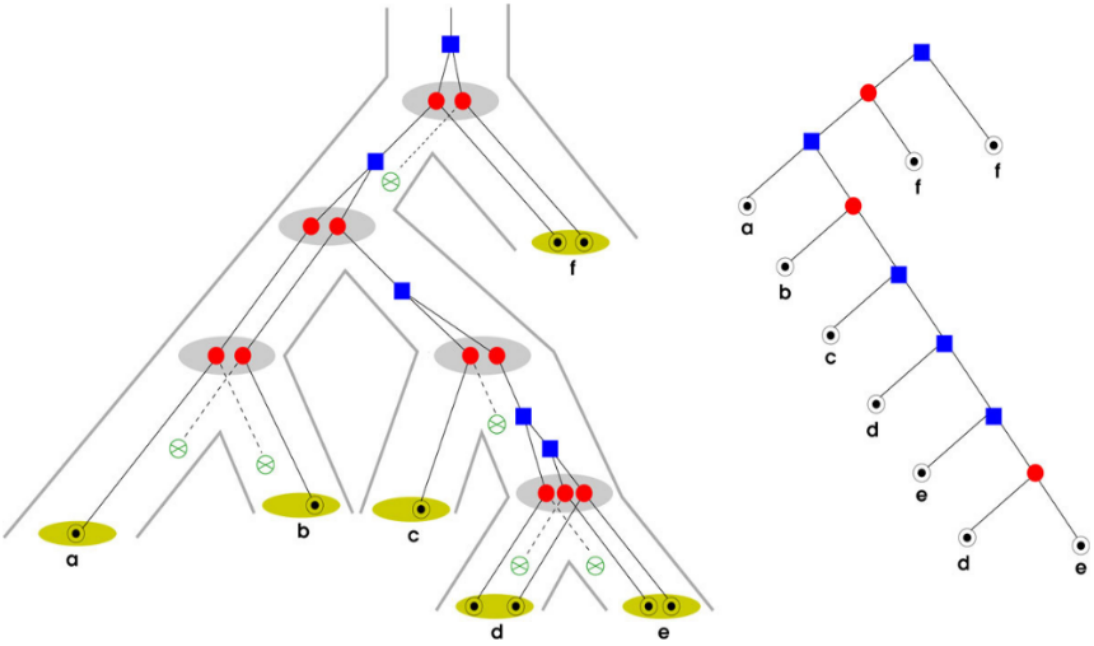


Figura 4.12: En la izquierda vemos un ejemplo de escenario evolutivo mostrando la evolución de una familia de genes correspondiente a un árbol de genes $\hat{T}$ real, que ha sido embebido en un árbol de especies $\hat{S}$. En la derecha vemos el árbol de genes observable. Los eventos de duplicación ($\diamond$) son representados por un cuadro azul, los de especiación ($\bullet$) por un círculo rojo, los eventos de pérdida de genes ($\otimes$) por un círculo tachado verde, los genes por $\odot$, y los nodos internos del árbol de especies parecen sombreados. Figura tomada de [32].

4.3.1. Historia evolutiva

Una **historia evolutiva verdadera** consta de 4 partes:

I) Un árbol de genes $\hat{T}$ sobre L

II) Un árbol de especies $\hat{S}$ sobre B

III) Un fechado simbólico $r : V(T) \to \{\bullet, \diamond, \otimes, \odot\}$

IV) Un mapeo $\mu : V(\hat{T}) \to V(\hat{S}) \cup E(\hat{S})$. Este mapeo es la **reconciliación** del árbol de T con S.

Donde los símbolos $\bullet, \diamond, \otimes$ representan respectivamente los eventos de especiación, duplicación y pérdida de gen, mientras que el símbolo $\odot$ representa a un gen observable (existente). También es útil definir el mapeo $\sigma : L(T) \to L(S)$, que asigna a cada gen del árbol de genes una especie del árbol de especies.

Cuando se analiza la evolución de genes no es posible detectar de forma directa las pérdidas de genes, ya que estos no son observables. Sin embargo, es posible determinar cuando ocurren estos eventos por medio de la reconciliación de árboles, que se describe más abajo.

4.3.2. Congruencia de árbol de genes con árbol de especies

Dados los elementos que conforman a una historia evolutiva, listados en la sub-sección anterior, se debe cumplir que para dos nodos $u, v \in V(T)$, $u \preceq v$ si y sólo si $\mu(u) \preceq \mu(v)$. Esta condición se cumple si y sólo si S contiene todas los árboles compactos inducidos de T con tres hojas correspondientes a 3 genes de diferentes especies, y cuyo nodo raíz es de tipo especiación [32]. En la figura 4.13 se ejemplifica esta condición. Dado un conjunto de tripletas, el algoritmo `BUILD` [49] puede ser usado para determinar en tiempo polinomial si un conjunto dado de tripletas es consistente, y si lo es retorna la filogenia que contiene a todas las tripletas dadas.

4.3.3. Reconciliación

El mapeo $\mu : V(T) \to V(S) \cup E(S)$ descrito en la sección 4.3.1 es una asignación de todo nodo $n_T \in V(T)$ del árbol de genes a un nodo o arista $n_T \in V(S) \cup E(S)$ del árbol de especies. En la parte izquierda de la figura 4.12 se puede observar el mapeo de forma implícita, mientras que en la figura 4.14 se muestra de forma explícita.

La reconciliación identifica el evento de especiación o divergencia más reciente en el árbol de especies donde debieron ocurrir cada uno de los escenarios descritos en el árbol de genes. Para tal fin es necesario tomar en cuenta las especies de los genes que surgieron

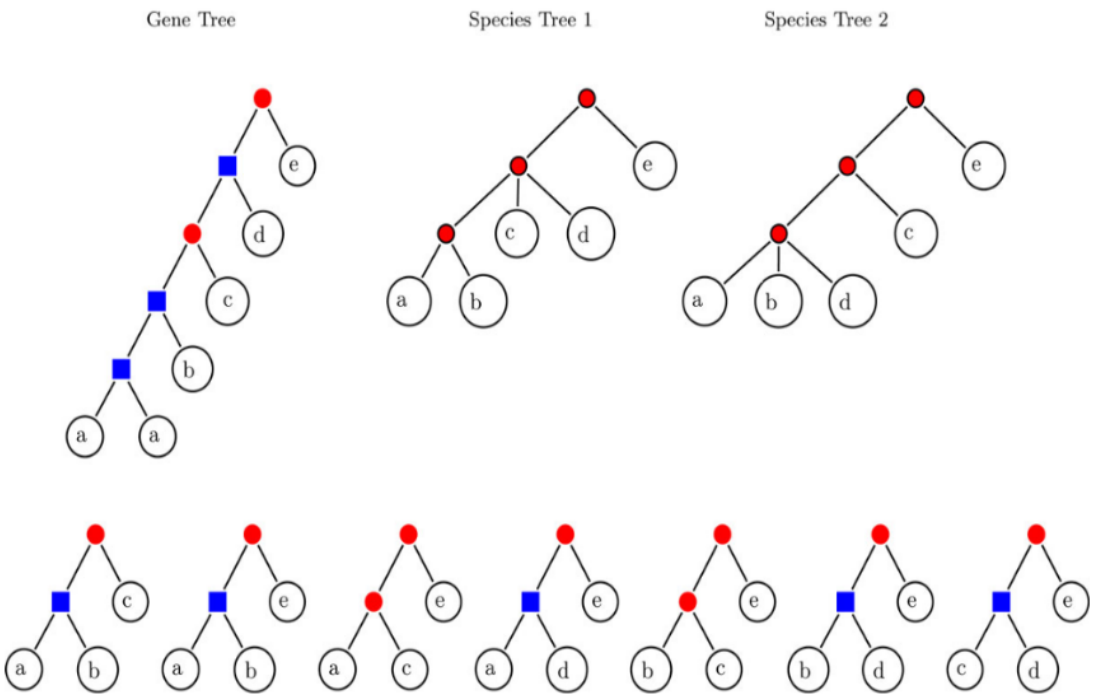


Figura 4.13: Arriba a la izquierda se muestra un árbol de genes, a la derecha de éste se pueden observar dos árboles de especies que son congruentes con el árbol de genes, mientras que en la parte inferior se muestran todas las tripletas del árbol de genes con tres hojas correspondientes a 3 genes de diferentes especies, y cuyo nodo raíz es de tipo especiación. Figura tomada de [32].

a partir de tales eventos. Es decir, el conjunto de hojas que descienden del nodo n_T, por lo que el mapeo toma la forma

$$\mu(n_T) = LCA_S(\{\sigma(v) \forall v \in L(T) | n_T \preceq v\})$$

Por medio de este mapeo es posible identificar de forma directa las ramas del árbol de especies S en las que ocurrieron los eventos de especiación y duplicación del árbol de genes. Por otro lado, si $\mu(n_T) = n_S$, es de esperarse que para cada descendiente de la especie ancestral n_S sea heredada al menos una copia de los genes de tal organismo. Sin embargo, puede pasar que ninguno de los genes descendientes de n_T correspondan a una especie descendiente de n_T, lo cual nos indica que se debió perder la copia del gen correspondiente.

La pérdida de genes se formaliza de la siguiente manera, dado que $\mu(n_T) = n_S$, denotamos a $M_T = \{\sigma(v) \forall v \in L(T) | n_T \preceq v\})$ y a $M_S = \{v \in L(S) | n_S \preceq v\})$ como los conjuntos de especies inducidas por n_T y n_S. Podemos notar que el gen correspondiente al mapeo $\mu(n_T) = n_S$ es heredado a los descendiente de n_S y a los descendientes de estos hasta llegar a las especies existentes M_S, sin embargo, si M_T no contiene a todas las especies de M_S, significa que ocurrieron pérdidas de genes en todos aquellos nodos n_S^* tal que $n_S \preceq n_S^*$ y

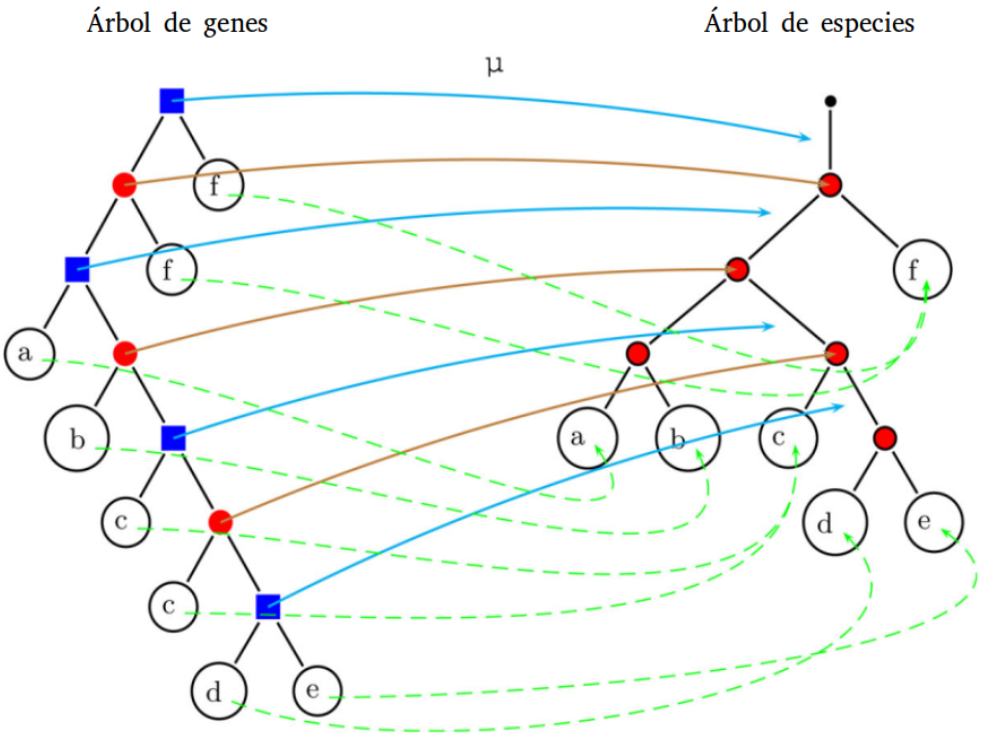


Figura 4.14: Se pueden observar el árbol de genes y el de especies del lado izquierdo y derecho respectivamente. Estos árboles se mostraron en la figura 4.12. Aquí se muestra de forma explícita el mapeo $\mu : V(T) \to V(S) \cup E(S)$. Figura tomada de [32].

$\nexists v \in M_T$ donde $n_S^* \preceq v$, y el padre de n_S^*, que denotamos por u, cumple con la condición $u \preceq v$ para cualquier $v \in M_T$.

Capítulo 5

Resultados

5.1. Herramienta

La herramienta bioinformática desarrollada se depositó en un repositorio público (`https://gitlab.com/jarr.tecn/revolutionh-tl`).En esta página web se puede acceder a la herramienta a la par de una breve descripción de la metodología, documentación, así como un tutorial que nos explica cómo realizar el análisis de genomás y de sub-conjuntos de genes. La herramienta brinda la posibilidad de visualizar árboles de genes y de especies de forma individual, así como la reconciliación explícita de uno o varios árboles de genes con el árbol de especies. Para visualizar grandes conjuntos de familias de genes se presenta un árbol con números que resume la reconciliación, junto con un gráfico *upSet*, que muestra información para cada clado. Estas visualizaciones se muestran con datos simulados y reales en las siguientes secciones.

El análisis realizado por la herramienta que hemos implementado, es a nivel familia de genes, y entra dentro de la clasificación de aquellas que utilizan los eventos de duplicación y pérdida de genes (**DL**).

5.2. Validación sintética

En [52] se estudia *in silico* la evolución de familias de genes. Como parte de sus resultados se presenta la simulación de árboles de genes y de especies [52]. Hemos empleado tales simulaciones para verificar el correcto funcionamiento de la herramienta. En la tabla 5.1 se muestran los datos que fueron usados.

Se crearon los grafos de ortología *reales* a partir de los árboles de genes simulados. Para obtener las aristas del grafo se usó la definición de cografo, el peso w que se le asignó a cada arista $e = (u, v)$ es la suma de las distancias que separan a las hojas u y v en el coárbol. Finalmente, para que el peso represente similitud se cambia por $w^* = (w_{max} - w)/w_{max}$, donde w_{max} es el peso más grande de las aristas. Notar que esta cantidad toma valores en el intervalo inclusivo $[0, 1]$, sin embargo no tiene sentido tener relaciones de ortología con peso cero, por lo que a los elementos que resultaron con este valor, se les asigna la décima

ID	Especies	Genes	Especiaciones	Duplicaciones	Pérdidas
S1	3	7	4	6	4
S2	11	19	13	12	7
S3	41	65	54	16	6
S4	46	101	101	19	20
S5	50	73	73	14	15
S6	34	64	55	15	7
S7	37	63	55	8	1
S8	28	43	29	18	5
S9	5	13	6	10	4
S10	8	22	13	10	2
S11	20	24	19	7	3

Tabla 5.1: Resumen de las historias evolutivas simuladas en [52]. La primera columna es el identificador de las historias simuladas, mientras que las otras denotan el número de especies simuladas, el número de genes, y el número de cada uno de los eventos simulados.

parte del mínimo peso de las aristas después de la trasformación.

Posteriormente se retiraron y agregaron relaciones de ortología a los grafos originales mediante el siguiente criterio: sea $G = (V, E)$ el grafo de ortología original, para toda arista $e = (u, v) \in E$ se obtiene un número aleatorio en el intervalo $[0, 1]$. Si el peso de la arista es menor al número aleatorio, entonces se retira la arista. De forma similar, para todos los pares de nodos $e' = (u', v') \notin E$ se obtiene un número aleatorio en el mismo intervalo. En esta ocasión, si el número resultante es mayor al peso promedio, la arista e' es incluida en el grafo.

Tras analizar los grafos de ortología originales fue posible obtener nuevamente los árboles de genes, así como identificar las pérdidas de genes simuladas. Al analizar los grafos de ortología con ruido fue necesario aplicar ediciones (la herramienta lo realiza de forma automática) para obtener los árboles de genes correspondientes. Sin embargo, en esta ocasión sólo se pudieron obtener árboles de genes congruentes para 4 historias evolutivas y a pesar de que son consistentes con los árboles originales, incluyen una cantidad menor de genes. En la figura 5.1 se observa la reconciliación de dos árboles de genes inferidos a partir de los grafos de ortología originales.

5.3. Análisis de genes sin intrones de ratón

Los genes de células eucariotas usualmente están conformadas por regiones codificantes (CDSs) y regiones que no son traducidas (UTRs), además de regiones intragénicas que son removidas durante el empalme de ARN durante la maduración del producto para su posterior traducción a proteínas. Los genes multiexónicos (MEGs) son aquellos que contienen dos o más exones separados por regiones intergénicas conocidas como intrones, en

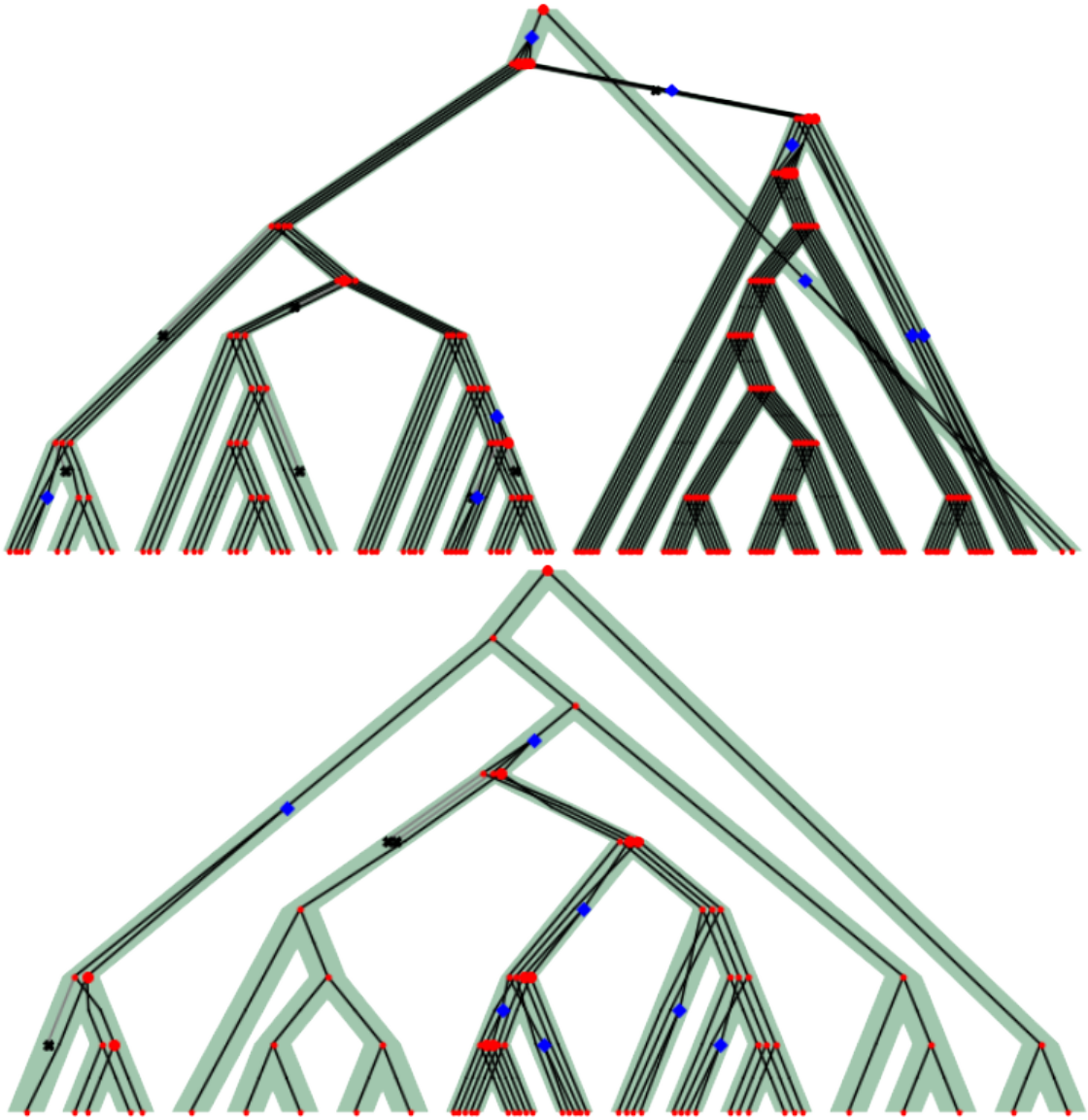


Figura 5.1: Reconciliación de historias evolutivas simuladas. Arriba se observa la historia con ID 3, y abajo la historia con ID 4. Los árboles evolutivos fueron inferidos a partir de los grafos de ortología reales y de la reconciliación de estos con un árbol de especies, obteniendo como resultado los árboles originales íntegros, incluso se pudieron determinar los clados específicos donde ocurrieron pérdida de genes.

contraste, están los genes codificantes que carecen de intrones, estos son llamados *genes de un sólo exón* (SEGs). Esta categoría de genes se subdivide en dos principales grupos: *i*) *SEGs con intrones en las regiones UTRs* (uiSEGs), y *ii*) *genes sin intrones* (IGs) [36].

Se han identificado más de 2000 IGs de humano. Es importante hacer esta distinción porque la presencia o ausencia de intrones en el 5'UTR o el 3'UTR pueden impactar en procesos de regulación transcripcional y post-transcripcional [36]. En la figura 5.2 se puede observar una comparación entre un gen sin intrones y uno multiexónico con intrones en

las regiones UTRs y CDSs.

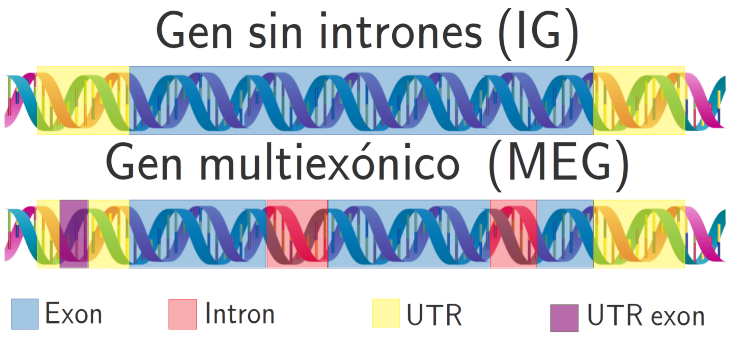


Figura 5.2: Comparación de la estructura de un IG y un MEG. Arriba se muestra un gen sin intrones, abajo un gen multiexónico con intrones en las regiones UTRs y CDSs.

Para este análisis se descargó de la base de datos de Ensembl el proteoma de las siguientes especies:

1. *Mus musculus*
2. *Homo sapiens*
3. *Pan troglodytes*
4. *Monodelphis domestica*
5. *Rattus norvegicus*
6. *Gallus gallus*
7. *Danio rerio*

En la figura 5.3 se muestra la metodología usada para la clasificación de los genes de cada especie como MEGs, uiSEGs o IGs.

Las especies que se analizan son organismos modelo de diferentes niveles taxonómicos, por lo que la conservación de genes ortólogos de ratón en especies cercanas o lejanas nos da una noción de la *antigüedad* de tales genes (ver figura del árbol de especies). A partir de esto se puede construir una escala de qué tan antiguo es un gen de ratón por medio de la búsqueda de sus ortólogos en especies cada vez más lejanas: cuando los ortólogos de un gen de ratón sólo se identificaron en rata se dice que los genes de ratón están conservados en ***Muridae***; aquellos presentes en *Muridae*, chimpancé y humano están conservados en el grupo ***Eutheria***, si incluimos al didélfido estaremos en el grupo ***Theria***, agregando al pollo llegamos a ***Tetrapoda***, y finalmente aquellos genes que tienen ortólogos en el pez cebra se dice que está conservados en ***Vertebrata***.

Especies	No. de ortólogos de IGs	No. de parálogos de IGs
RATTU	501	36
MONDO	258	9
HUMAN	397	16
PANTR	335	11
GALUS	133	2
DANRE	220	0

Tabla 5.2: Resumen de relaciones de ortología y paralogía encontradas para genes sin intrones de ratón.

Se analizaron los 7 genomas completos con la herramienta implementada, posteriormente se identificaron todos los grupos de ortólogos para los cuales se infirió una historia evolutiva congruente, estos fueron divididos en 3 subconjuntos dependiendo de los genes de ratón que incluyen: se toman todos los grupos de ortólogos que contienen al menos un gen del grupo indicado y se retiran todos aquellos genes de ratón que no pertenezcan a tal clase. Los grupos que se usaron fueron *i*) IGs de ratón, *ii*) MEGs de ratón y *iii*) β-protocaderinas de ratón, que son un subconjunto de los IGs. Para cada una de estas clasificaciones se obtuvo la distribución de genes conservados en cada grupo taxonómico y se clasificaron sus ortólogos como IGs, uiSEGs, MEGs y otros.

Se obtuvieron las distribuciones de conservación de los IGs y MEGs de ratón en las especies analizadas. Este análisis se puede ver en la figura complementaria 1 (`https://gitlab.com/jarr.tecn/revolutionh-tl/-/blob/master/docs/Mouse%20IGs/README.md`). En tal figura se compara el número de genes en especies existentes o ancestrales, así como las pérdidas y duplicaciones de genes inferidas. Mediante un gráfico *upSet* se muestra el número de genes de ratón completamente conservados por cada clado, y se puede apreciar que los genes sin intrones tienden a estar conservados en *Muridae*, mientras que los genes multiexónicos se conservan en *Vertebrata*, es decir en todas las especies analizadas.

El método empleado identificó ortólogos y parálogos de 543 IGs y 12315 MEGs de ratón, en las tablas 5.2 y 5.3 se resumen las relaciones que se encontraron para los IGs y los MEGs respectivamente. Mediante este análisis, fue posible observar que la mayoría de los IGs tienen ortólogos mayoritariamente IGs, mientras que los ortólogos de MEGs también tienden a ser MEGs. Esto se ve claramente en las tablas 5.4 y 5.5, donde se presenta cuántos ortólogos de cada especie y clase tienen los IGs y MEGs.

Posteriormente, se obtuvo la historia evolutiva de las 18 $\beta-$protocadherinas de ratón, descrita a través de 14 grupos de ortólogos. En la tabla 5.6 se muestran los tipos de relaciones que se infirieron entre estos genes y los de otras especies. Con estos resultados se colaboró en un trabajo [5] en el que se analizan los IGs de ratón desde una perspectiva evolutiva y funcional. En la figura 5.4 se muestra un resumen de resultados, en la que se puede apreciar la reconciliación explícita de los árboles de genes obtenidos.

Especies	No. de ortólogos de MEGs	No. de parálogos de MEGs
RATTU	11687	161
MONDO	8747	64
HUMAN	11429	64
PANTR	10497	51
GALUS	7266	38
DANRE	92	94

Tabla 5.3: Resumen de relaciones de ortología y paralogía encontradas para genes multiexónicos de ratón.

Genoma	IGs	uiSEGs	MEGs	Others	Total
Zebrafish	91	0	90	39	220
Chick	78	1	54	0	133
Opposum	167	1	86	4	258
Chimp	250	0	59	26	335
Human	262	0	97	38	397
Rat	442	0	57	2	501

Tabla 5.4: Clasificación de los ortólogos de genes sin intrones de ratón.

Genoma	IGs	uiSEGs	MEGs	Others	Total
Zebrafish	108	0	7602	1584	9294
Chick	122	2	7135	7	7266
Opposum	523	8	8032	184	8747
Chimp	603	0	9520	374	10497
Human	159	0	10304	966	11429
Rat	1171	0	10457	59	11687

Tabla 5.5: Clasificación de los ortólogos de genes multiexónicos de ratón.

Especies	No. de ortólogos	No. de parálogos
RATTU	9	3
MONDO	4	3
HUMAN	7	3
PANTR	9	2
GALUS	3	0
DANRE	0	0

Tabla 5.6: Relaciones de homología inferidas para las $\beta-$protocadherinas de ratón.

5.4. Cáncer y otras enfermedades

La expresión genes sin intrones está fuertemente asociada con enfermedades tales como cáncer, neuropatías y desórdenes del desarrollo. Ejemplos de IGs con relevancia clínica son el gen RPRM, relacionado con cáncer gástrico, el cual induce proliferación y posee actividad supresora de tumores [4]. Otro ejemplo son las kinasas CK2α, que aparecen *up-regulated* en todos los genomás de cáncer [33]. Otros más incluyen CLDN8 en carcinoma colorrectal y células de tumores renales, ARLTS1 en melanoma, y PURA junto con TAL2 en leucemia. [28].

Este subconjunto de genes representa un potencial clínico como biomarcadores y blancos farmacológicos [27, 4, 43]. Estudios previos describieron al gen *FZD8* como un receptor esencial de la ruta *Wnt* implicada en el desarrollo y tamaño del cerebro. Este gen está expresado ampliamente en dos lineas de cáncer humano; está relacionado con las rutas metabólicas de cáncer gástrico y de mama, así como con la metástasis de cáncer colorrectal, indicando que puede tomar un rol importante en procesos de tumorgénesis. *ASCL5* puede estar involucrado en la regulación de transcripción de ADN, además su potencial papel en cáncer de pulmón y cerebro.

Para explorar la relevancia clínica de los IGs, se identificaron las enfermedades de humano en las cuales estos genes presentan des-regulación. En la figura 5.5 se puede observar un enriquecimiento de términos de los padecimientos encontrados más representativos para este conjunto de genes. Cabe destacar que la mayoría de las enfermedades identificadas por este análisis están relacionadas con distintos tipos de cáncer. Posteriormente, se identificaron las correspondientes familias de proteínas de los IGs asociados a esta enfermedad, en la figura 5.6 se muestran las funciones más abundantes relacionadas con los IGs des-regulados y se compara con el resultado cuando se analizan todos los IGs. Resalta que en todos los casos los términos están más relacionados con los IGs des-regulados.

Una vez identificadas las familias de proteínas sobrerepresentadas en el conjunto de IGs diferencialmente expresados en cáncer, se procedió a la inferencia de la reconstrucción de la historia evolutiva. Este análisis se enfocó principalmente en las siguientes familias predominantes: factores de transcripción FOX (HMG-box), POU3, y SOX; transmembranales β-protocadherinas, CLDN (claudinas), FZD (frizzled) y, finalmente, histonas canónicas del tipo H1, H2, H3, H4. El análisis evolutivo de estas familias se realizó en las especies selectas de mamíferos: *Homo sapiens*, *Monodelphis domestica*, *Oryctolagus cuniculus*, *Equus caballus*, *Ovis aries rambouillet*, *Capra hircus*, *Bos taurus*, *Sus scrofa*, *Canis lupus familiaris* y *Felis catus*. Mediante este estudio se determinó que estas familias de genes presentan alto nivel de conservación. En la figura complementaria 2 (`https://gitlab.com/jarr.tecn/revolutionh-tl/-/blob/master/docs/Mouse%20IGs/README.md`) se muestra un panel con algunos de los árboles de reconciliación. Puede observarse que en general hay pocos eventos de pérdida y ganancia de genes.

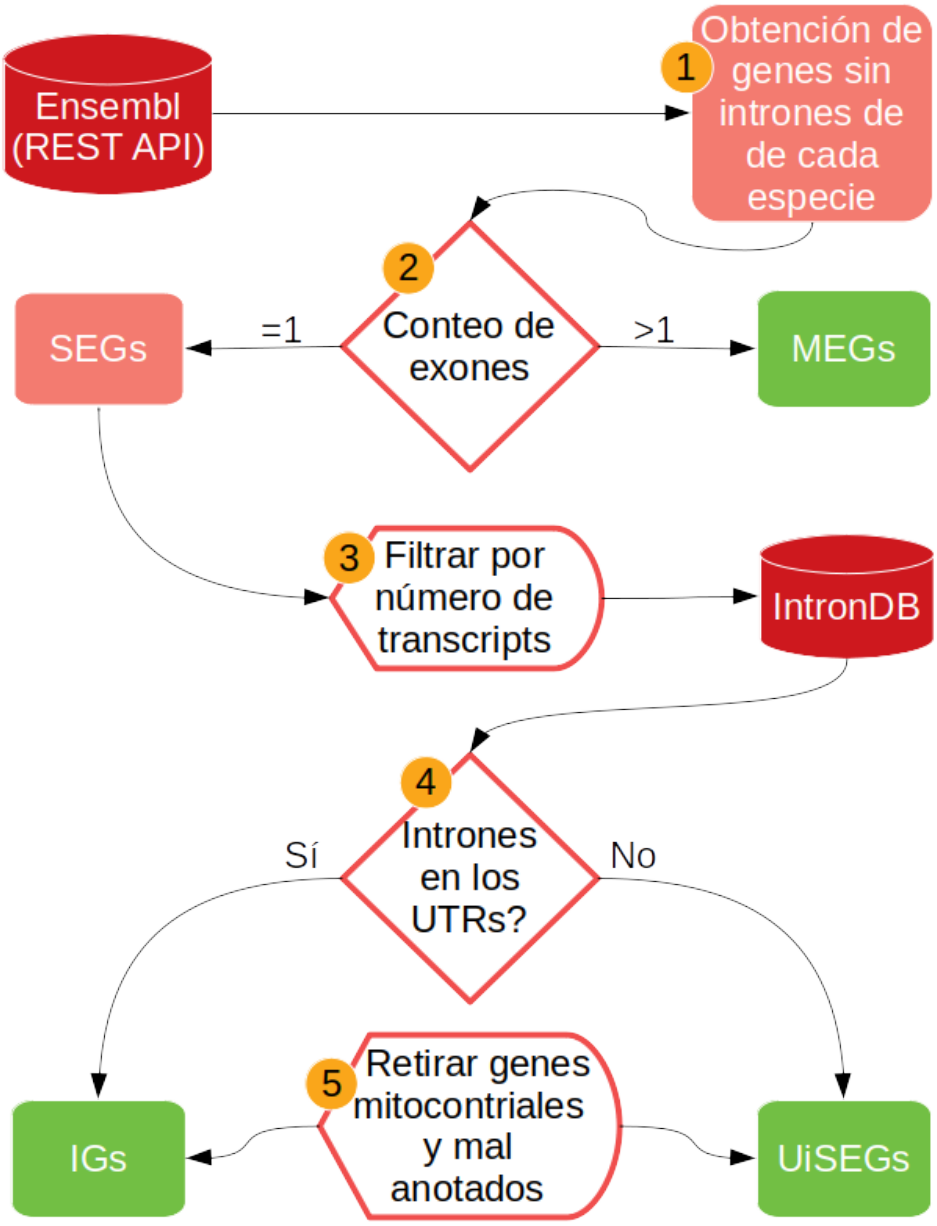


Figura 5.3: Metodología para clasificar genes como IGs, uiSEGs, y MEGs. (**1**) Se obtuvieron todas las secuencias codificantes por medio de la plataforma Ensembl REST API, los scripts están disponibles en (`https://github.com/GEmilioHO/intronless_genes`). (**2**) Aquellos genes con más de un exón se clasificaron como MEGs, y los genes de exón único como SEGs. Este último grupo se filtró (**3**) descartando los genes con más de un transcrito. Posteriormente (**4**) se revisó usando la base de datos IntronDB (`http://www.nextgenbioinformatics.org/IntronDB`) si los genes tienen intrones en sus UTRs, para finalmente (**5**) clasificarlos como IGs o uiSEGs y filtrar los genes mitocondriales o con una mala anotación de sus proteínas. En color verde se pueden apreciar los archivos de salida de la metodología.

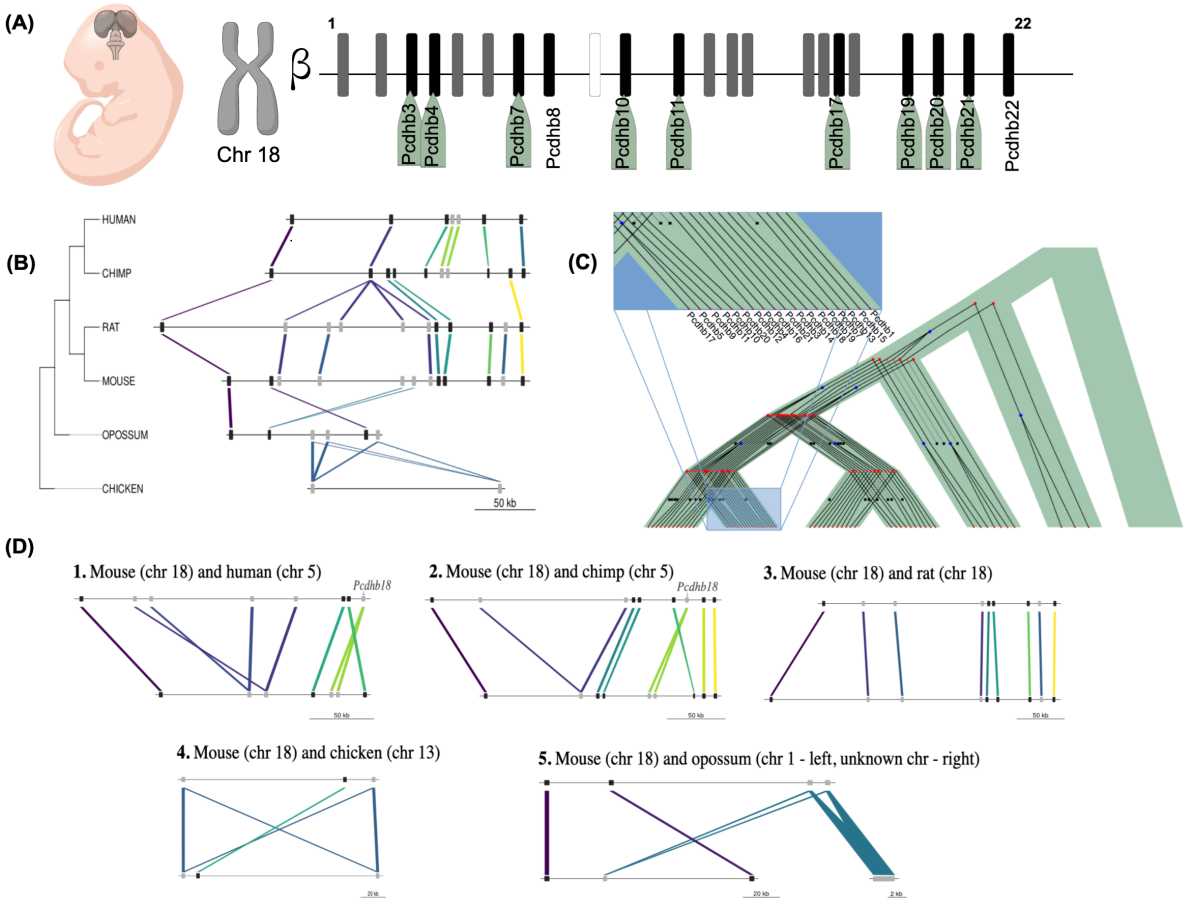


Figura 5.4: Resumen de análisis evolutivo y de expresión del grupo de $\beta-$protocadherinas de ratón. En **a** se muestra la expresión de los genes, en negro se observan genes *up-regulados* en tejido de telencéfalo de ratón entre las etapas embrionarias 9.5 y 10.5, en gris se muestran los genes *up-regulados* que no tienen cambios en su expresión en las diferentes etapas, mientras que el blanco se muestran genes no expresados. En **b** se muestra un un análisis sinténico en el que se comparan las posiciones de los genes de ratón, así como de los ortólogos inferidos. Las líneas de colores representan relaciones de ortología, donde genes conectados por un color corresponden a un mismo grupo de ortólogos. Los genes se representan por medio de rectángulos, en color negro aquellos que constan de una sola copia, y en gris los que están duplicados. En **c** se muestra la reconciliación explícita de los árboles de genes (líneas negras) y el árbol de especies (área verde). Cada árbol de genes corresponde a un grupo de ortólogos, los nodos internos representan eventos evolutivos: con círculos rojos se denotan especiaciones, mientras que en rombos azules se muestran duplicaciones de genes, y con taches negros se indican las ramas donde ocurrieron pérdidas de genes. Finalmente, en **d** se muestra una comparación sinténica a pares de los genes de ratón contra sus ortólogos en cada una de las especies. Esta comparación resalta las pérdidas y ganancias de genes que ocurrieron a lo largo del árbol de especies.

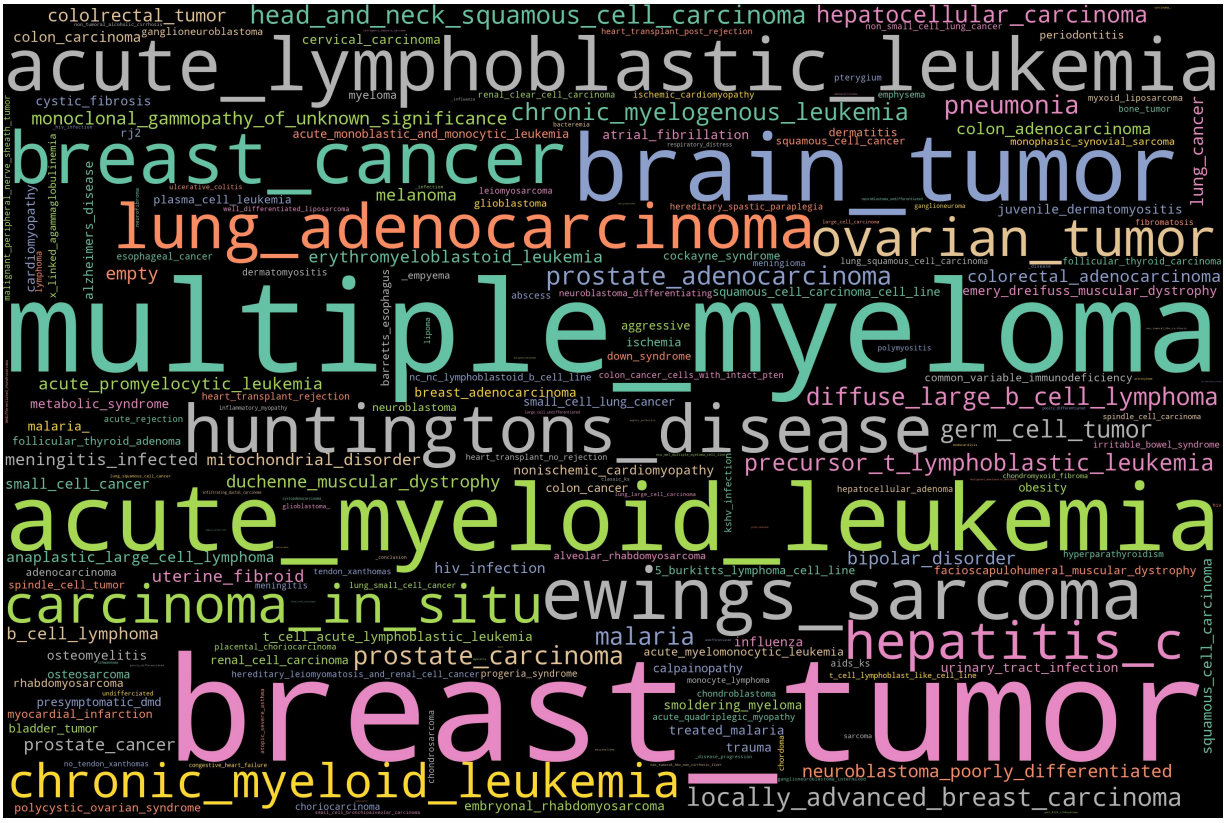


Figura 5.5: Correlación de enfermedades con IGs des-regulados. El tamaño de las palabras indica qué tan representado es un padecimiento.

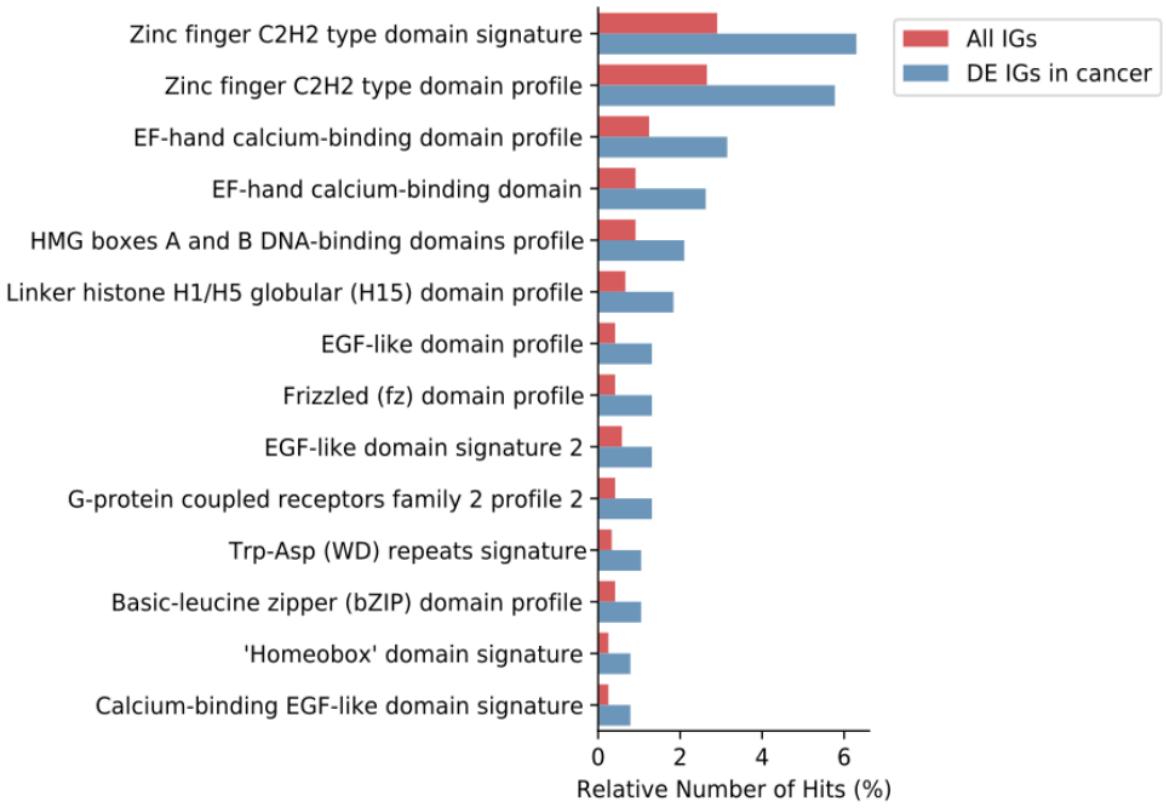


Figura 5.6: Familias de proteínas de IGs des-regulados con mayor representación en cáncer. En azul se muestra el porcentaje de IGs des-regulados que apuntan a cada familia, mientras que el rojo corresponde al porcentaje para el conjunto de todos los IGs.

Capítulo 6

Discusión

En este trabajo se implementó una metodología que usa las propiedades teóricas de las relaciones de ortología para inferir la historia evolutiva de familias de genes. Esta información puede ayudar a entender el papel fundamental que cumplen genes ortólogos en diferentes especies. La herramienta desarrollada durante el proyecto fue usada para analizar datos simulados y estudiar la evolución de genes sin intrones de ratón.

La metodología implementada es flexible, fácil de usar y es capaz de analizar genomas completos de múltiples especies. Estas características son ideales para investigadores interesados en analizar sus propios conjuntos de datos y responder preguntas específicas por lo que se piensa publicar su artículo para que otros estudios puedan utilizarla y hacer referencia a ella.

Futuras mejoras a la metodología implementada podrán calcular la probabilidad de obtener los árboles de genes dado un árbol de especies fechado y las tazas de ocurrencia de los eventos evolutivos. Por otro lado, una vez que se infirió un árbol de genes, y éste no es congruente, es natural preguntarse cuál es el cambio en las relaciones tal que se convierten en cografo y el cambio es mínimo. Ese problema es abordado en [23], donde usan la descomposición modular para editar las relaciones de ortología directamente en la descomposición modular en lugar del grafo, mediante una simple operación llamada *unión de modulo*.

El análisis de los grafos de ortología obtenidos a partir de historias simuladas muestran que el método es capaz de inferir historias evolutivas adecuadas. Al estudiar grafos con ruido también se pudieron obtener historias evolutivas acertadas, sin embargo se perdió una cantidad considerable de información dado que varios árboles de genes resultaron incongruentes, lo cual resalta la necesidad de agregar una metodología que corrija estos árboles.

La conservación de la clasificación de los genes sin intrones de ratón parece indicar que la inferencia de ortología fue apropiada, así como que el origen de estos genes se remonta a especies ancestrales antiguas. Dado que en la mayoría de familias de IGs existe una proporción de ortólogos que son genes multiexónicos, resulta de interés determinar en

qué momento se adquirieron los intrones en tales genes, así como la repercusión en sus funciones.

Al analizar la evolución de familias de IGs que aparecen desregulados en cáncer es posible observar que en general los genes están bien conservados en todas las especies analizadas, aunque casi nunca están ausentes eventos de pérdida y duplicación de genes. Esto podría indicar que en otros organismos las enfermedades como cáncer también se relacionan con los genes sin intrones, por lo que sería de interés identificar diferencias y similitudes de la expresión de sus ortólogos en otros organismos, e incluso tratar de determinar el impacto que tuvieron los eventos evolutivos inferidos en sus procesos de regulación.

Capítulo 7

Conclusión

Hoy en día existe una gran y creciente cantidad de datos biológicos públicos. Por la naturaleza de los problemas concernientes al análisis de dicha información es necesario desarrollar nuevas tecnologías que integren conocimientos de diferentes áreas de las ciencias e ingenierías, dígase de forma directa Biología, Matemáticas y Computación. Estas tecnologías tienen el potencial de impulsar el desarrollo de investigaciones con impacto internacional y sobre todo nacional, abriendo la posibilidad de nuevas líneas de investigación tanto en ciencia básica como aplicada. Dentro de este contexto, resalta el rol ideal que puede cumplir un licenciado en Tecnología durante el desarrollo de tales proyectos y particularmente en el presente trabajo, donde un claro ejemplo de esto es el uso de la herramienta desarrollada para el estudio de genes relacionados con enfermedades complejas como el cáncer, a la que son susceptibles los humanos y diferentes organismos.

La metodología presentada en este trabajo tiene la ventaja de obtener historias evolutivas de familias de genes a partir de relaciones de ortología, evitando el uso de métodos para la inferencia de filogenias basados en la comparación de distancias entre genes, pues tales procedimientos requieren mayor poder computacional y son propensos a tener errores con grandes conjuntos de datos o distancias evolutivas no uniformes.

Por otro lado, los métodos que determinan ortología entre pares de genes por medio de la hipótesis de mejor alineamiento realizan predicciones de forma rápida y sin ser propensos a los errores mencionados arriba. Sin embargo, las las predicciones de estos métodos suelen tener una mayor cantidad de falsos positivos. Este problema se afronta con `REvolutionH-tl`, pues la herramienta identifica conjuntos de relaciones de ortología sin sentido biológico (es decir, que no forman un cografo) y procede a corregirlas basándose en evidencia experimental. Además, mediante el proceso de reconciliación es posible identificar incongruencias entre los árboles de genes y el árbol de especies, sugiriendo que es posible hacer una refinación de las relaciones de ortología predichas por medio de la identificación de las tripletas incongruentes de los árboles de genes y sus correspondientes grafos de ortología.

Los resultados proporcionados por `REvolutionH-tl` son tablas de ortología, árboles de genes, árboles de reconciliación y un mapeo de los eventos evolutivos de los árboles de

genes hacia los elementos del árbol de especies. Tales resultados se entregan en archivos de texto planos. De igual manera se generan visualizaciones que resumen la gran cantidad de datos generados y muestran explícitamente los árboles de genes reconciliados con el árbol de especies. Otras herramientas de inferencia y reconciliación de árboles entregan los resultados en los mismos formatos o algunos de los mencionados, por lo que es posible usar los resultados de esta herramienta en otras metodologías.

La herramienta generada tiene la limitación de no ser capáz de detectar transferencia horizontal de genes, por lo que tratarla de usar para analizar genomas de organismos con evolución más compleja (por ejemplo, células procariotas) puede resultar en una menor precisión en las historias inferidas, generando una mayor cantidad de eventos evolutivos erróneos, así como inconsistencias entre los árboles de genes y de especies. Sin embargo la transferencia horizontal de genes es un evento poco común en organismos eucariotes, por lo que el análisis tendrá una mínima cantidad de errores debidos a dichos eventos evolutivos.

Capítulo 8

Contribuciones

Artículo en revisión Se colaboró en un estudio donde se analiza la conservación del papel regulatorio de los IGs en vertebrados desde una perspectiva evolutiva y funcional. Tal trabajo está en proceso de revisión para su publicación [5] (`https://www.biorxiv.org/content/10.1101/2021.01.13.426573v1.full`).

Escuelas de bioinformática El presente trabajo se inició durante la III Escuela de Verano de Bioinformátcia organizada por el Instituto de Matemáticas y el Instituto de Neurobiología de la UNAM Juriqulla en el año 2018. En el presente año participé como ayudante de profesor en un proyecto durante en el 2º Taller de Bioinformática Avanzada organizado por el CINVESTAV Irapuato, el Laboratorio Nacional de Visualización Científica Avanzada (LAVIS) y el INB de la UNAM Juriqulla `http://lavis.unam.mx:3000/`. En dicho taller se estudiaron las características de genes sin intrones de humano con la finalidad de identificar biomarcadores para detección diferentes tipos de cáncer y se contó con el apoyo técnico de Luis Alberto Aguilar Bautista y Alejandro de León Cuevas del LAVIS.

Presentaciones en Congresos

- Ramírez Rafael J.A, Valdivia DI., García-García A.,Herrera-Oropeza G.A , Varela-Echavarría A., Avina-Padilla K. and Maribel Hernández-Rosales. An Evolutionary Perspective of Intronless Genes highlight their "recent" highly specialized role. X ISCB Latin America SoIBio BioNetMX Symposium on Bioinformatics 2020, Mexico, Octubre, 2020. (poster presentation 62) `https://www.iscb.org/cms_addon/conferences/la2020/viewinghall.php`
- Avina-Padilla K, García-García A., Ramírez Rafael J.A, Varela-Echavarría A., and Maribel Hernández-Rosales Talk: Occurrence, distribution and role of Single Exon Genes encoding membrane proteins in embryonic development of the telencephalon in Mouse. X International Conference on Bioinformatics 2019, Uruguay, Octubre, 2019.

- Herrera-Oropeza G.A, Garcia-Garcia A., Ramirez-Rafael J.A, Hernandez-Rosales M, Varela- Echavarria A, and Avina-Padilla K. Characterization of single exon genes encoding transmembrane proteins in the mouse embryonic telencephalon, 26 Jornadas Academicas del INB, Instituto de Neurobiología UNAM-Juriquilla, INB-UNAM, Mexico, Septiembre, 2019.

- Avina-Padilla K, García-García A., Ramírez Rafael JA., Herrera- Oropeza GA, Hernández-Rosales M, VijayKumar Muley and Varela-Echavarría A. Analysis of Expression of Single Exon Genes in the Mouse Embryonic Telencephalon. III Congreso Nacional de Neurobiologia, Guanajuato, Mexico, Septiembre, 2019.

Bibliografía

[1] The ensemble database.

[2] git repository for proteinortho.

[3] Orjan Akerborga, Bengt Sennbladb, Lars Arvestada, and Jens Lagergrena. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *PNAS*, 106(14):5714–5719, 2009.

[4] Julio D. Amigo, Juan C. Opazo, Roddy Jorquera, Ignacio A. Wichmann, Benjamin A. Garcia-Bloj, Maria Alejandra Alarcon, Gareth I. Owen, and Alejandro H. Corvalán. The reprimo gene family: A novel gene lineage in gastric cancer with tumor suppressive properties. *International journal of molecular sciences*, 19(1862), 2018.

[5] Katia Aviña-Padilla, José Antonio Ramírez-Rafael, Gabriel Emilio Herrera-Oropeza, Vijaykumar Muley, Dulce I. Valdivia, Erik Díaz-Valenzuela, Andrés García-García, Alfredo Varela-Echavarría, and Maribel Hernández-Rosales. Evolutionary perspective and expression analysis of intronless genes highlight the conservation on their regulatory role. *bioRxiv*, 2021.

[6] Mukul S. Bansal, Eric J. Alm, and Manolis Kellis. Models, algorithms and programs for phylogeny reconciliation. *Bioinformatics*, 28:i283–i291, 2012.

[7] Serafim Batzoglou, Lior Pachter, Jill P. Mesirov, Bonnie Berger, and Eric S. Lander. Human and mouse gene structure: Comparative analysis and application to exon prediction. *Genome Research*, 10:950–958, 2000.

[8] Sebastian Bocker and Andreas W. M. Dress. Recovering symbolically dated, rooted trees from symbolic ultrametrics. *Advances in Mathematics*, 138:105–125, 1998.

[9] Bastien Boussau and Celine Scornavacca. *Reconciling Gene trees with Species Trees*, chapter 3.2, pages 3.2:1–3.2:23. hal-02535529f No commercial publisher | Authors open acces book, 2020.

[10] Gary ChartRand, Linda Lesniak, and Ping Zhang. *Graphs and digraphs*. Textbopressure,oks in mathematics. Taylor & Francis Group, 6 edition, 2016.

[11] Cedric Chauve, Nadia El-Mabrouk, Laurent Guéguen, Magali Semeria, and Eric Tannier. *Models and Algorithms for Genome Evolution*. Computational Biology. Springer-Verlag London, 2013.

[12] Nicolas Comte, Benoit Morel, Damir Hasic, Laurent Guéguen, Bastien Boussau, Vincent Daubin, Simon Penel, Celine Scornavacca, Manolo Gouy, Alexandros Stamatakis, Eric Tannier, and David P Parsons. Treerecs: an integrated phylogenetic tool, from sequences to reconciliations. *Bioinformatics*, 36:4822–4824, 2020.

[13] D.G. Corneil, H. Lerchs, and L. Stewart Burlingham. Complement reducible graphs. *Discrete applied mathematics*, pages 163–174, 1981.

[14] Francis Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[15] Charles darwin. Darwin's finches. *Journal of researches into the natural history and geology of the countries visited during the voyage of H.M.S. Beagle round the world, under the Command of Capt. Fitz Roy, R.N. 2d edition*, 1845.

[16] Ute Deichmann. Early responses to avery et al.'s paper on dna as hereditary material. *University of California Press*, (2):207–032, 2004.

[17] Jeffery P. Demuth and Matthew W. Hahn. The life and death of gene families. *BioEssays*, 31:29–39, 2009.

[18] Jean-Philippe Doyon, Vincent Ranwez, Vincent Daubin, and Vincent Berry. Models, algorithms and programs for phylogeny reconciliation. *Briefing in Bioinformatics*, 12:392–400, 2011.

[19] Jean-François Dufayard, Laurent Duret, Simon Penel, Manolo Gouy, François Rechenmann, and Guy Perrière. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21:2596–2603, 2005.

[20] Joseph Felsenstein. Phylip (phylogeny inference package).

[21] Joseph Felsenstein. Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

[22] Walter M. Fitch. Distinguishing homolohous from analogous proteins. *Syst. Zool*, 19:99–113, 1970.

[23] Adrian Fritz, Marc Hellmuth, Peter F. Stadler, and Nicolas Wieseke. Cograph editing: Merging modules is equivalent to editing p_4s. *Art of Discrete and Applied Mathematics*, 3(2), 2020.

[24] O. Gascuel. an improved version of the nj algorithm based on a simple model of sequence data. *Syst. Biol.*, 14:685–695, 1997.

[25] Morris Goodman, John Czelusniak, G. William Moore, A. E. Romero-Herrera, and Genji Matsuda. Fitting the gene linage into its species linage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28:132–163, 1979.

[26] Anthony JF Griffiths, Jeffrey H Miller, David T Suzuki, Richard C Lewontin, and William M Gelbart. *Introduction to Genetic Analysis*. New York: W. H. Freeman, 7 edition, 2000.

[27] Grzybowska and Ewa A. Human intronless genes: Functional groups, associated diseases, evolution, and mrna processing in absence of splicing. *Biochemical and Biophysical Research Communications*, 2012.

[28] Ewa A. Grzybowska. Human intronless genes: Functional groups, associated diseases, evolution, and mrna processing in absence of splicing. *Biochemical and Biophysical Research Communications*, 424:1–6, 2012.

[29] Stéphane Guidon and Olver Gascuel. A simple, fast and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst. Biol.*, 52:696–704, 2003.

[30] Michel Habib and Christophe Paul. A survey of the algorithmic aspects of modular decomposition. *Computer science review*, pages 41–591, 2010.

[31] Marc Hellmuth, Maribel Hernandez-Rosales, Katharina T. Huber, Vincent Moulton, Peter F. Stadler, and Nicolas Wieseke. Orthology relations, symbolic ultrametrics, and cographs. *Mathematical Biology*, 66, 2013.

[32] Maribel Hernandez-Rosales, Marc Hellmuth, Nicolas Wieseke, Katharina T. Huber, Vincent Moulton, and Peter F. Stadler. From event-labeled gene trees to species trees. *BMC Bioinformatics*, 13, 2012.

[33] Ming-Szu Hung, Yu-Ching Lin, Jian-Hua Mao, Il-Jin Kim, Zhidong Xu, Cheng-Ta Yang, David M. Jablons, and Liang You. Functional polymorphism of the ck2α intronless gene plays oncogenic roles in lung cancer. *Plos one*, 5(7), 2010.

[34] Szollosi G J, Rosikiewicz W., Boussau B., Tannier E., and Daubin V. Efficient exploration of the space of reconciled gene trees. *Systematic Biology*, 6(6):901–912, 2013.

[35] Edwin Jacox, Cedric Chauve, Gergely J. Szølløsi, Yann Ponty, and Celine Scornavacca. eccetera: comprehensive gene tree-species tree reconciliation using parsimony. *Bioinformatics*, 28:i283–i291, 2012.

[36] Roddy Jorquera, Carolina González, Philip Clausen, Bent Petersen, and David S. Holmes. Improved ontology for eukaryotic single-exon coding sequences in biological databases. *Database*, 2018.

[37] Chen Kevin, Durand Dannie, and Farach-Colton Martin. Notung: A program for dating gene duplications and optimizing gene family trees. *Bioinformatics*, 7:429–447, 2000.

[38] Eugene V. Koonin. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338, 2005.

[39] David M. Kristensen, Yuri I.Wolf, Arcady R. Mushegian, and Eugene V. Koonin. Computational methods for gene orthology inference. *Briefings in bioinformatics*, 12(5):379–391, 2011.

[40] Ulrich Kutschera and Karl J. Niklas. The modern theory of biological evolution: an expanded synthesis. *Naturwissenschaften*, 91:255–276, 2004.

[41] Manuel Lafond, Krister M. Swenson, and Nadia El-Mabrouk. *Models and Algorithms for Genome Evolution*. Computational Biology. Springer-Verlag London, 2013.

[42] Marcus Lechner, Maribel Hernandez-Rosales, Daniel Doerr, Nicolas Wieseke, Annelyse Thévenin, Jens Stoye, Roland K. Hartmann, Sonja J. Prohaska, and Peter F. Stadler. Orthology detection combining clustering and synteny for very large datasets. *PLoS ONE*, 2014.

[43] Liu, X Y, Y C Fan, S Gao, J Zhao, L Y Chen, F Li, and K Wang. Methylation of sox1 and vim promoters in serum as potential biomarkers for hepatocellular carcinoma. *Neoplasma*, pages 745—-53, 2017.

[44] Harvey Lodish, Arnold Berk, Chris A. Kaiser, Monty Krieger, Anthony Bretscher, Hidde Ploegh, Angelika Amon, and Kelsey C. Martin. *Molecular Cell Biology*. Katherine Ahr Parker, 8 edition, 2016.

[45] William Norton and Laure Bally-Cuif. Adult zebrafish as a model organism for behavioural genetics. *BMC Neurosci*, 11(90), 2010.

[46] Martin A. Nowak. *Evolutionary dynamics*. Belknap Press of Harvard University Press, 2006.

[47] Atsushi Ogura, Kazuho Ikeo, and Takashi Gojobori. Comparative analysis of gene expression for convergent evolution of camera eye between octopus and human. *Genome Res.*, (14):1555–1561, 2004.

[48] Jeremy Ramsden. *Bioinformatics*. Springer, 2009.

[49] Charles Semple and Mike Steel. *Phylogenetics*, volume 24. Oxford University Press, 2003.

[50] Jeanne M. Serb and Douglas J. Eernisse. Charting evolution's trajectory: using molluscan eye diversity to understand parallel and convergent evolution. *Springer Science*, (1):439–447, 2008.

[51] Mechthild Stoer and Frank Wagner. A simple min-cut algorithm. *Journal of the ACM*, 44(4):585–591, 1997.

[52] Alitzel López Sánchez. *Estudio computacional de escenarios evolutivos*. tesis de licenciatura, universidad nacional autónoma de méxico., 2019.

[53] Jeffrey Touchman. Comparative genomics. *Nature Education Knowledge*, 3(10):13, 2010.

[54] G. P. Wanger. The biological homology concept. *Annual Review of Ecology, Evolution, and Systematics*, (20):51–69, 1989.

[55] Xuhua Xia. *Comparative Genomics*, page 1. SpringerBriefs in Genetics, 2013.

[56] Christian M. Zmasek and Sean R. Eddy. A simple algorithm to infer gene duplication and speciation events on a gene tree. *Bioinformatics*, 17(9):821–828, 2001.