Filogenia molecular

Resumen

La filogenia molecular es la rama de la filogenia que analiza las diferencias moleculares hereditarias en las secuencias de ADN, ARN y proteínas para obtener información sobre las relaciones evolutivas de un organismo. El resultado de un análisis filogenético molecular se expresa en un árbol filogenético.

Sandra Mingo Ramírez

UAM - 2024/25

14 de septiembre de 2024 11:09

Universidad Autónoma de Madrid
Bioinformática y Biología Computacional

Código en Github

Índice general

Capítulo I

Introducción a la filogenia: principios y conceptos

La filogenia es la determinación de la historia evolutiva de los organismos. La filogenética es el estudio de la filogenia utilizando árboles filogenéticos de los distintos organismos y estudiar las relaciones entre ellos. Ha habido varias iniciativas a lo largo de la historia (como ToL Web) que han intentado lograr crear árboles de todas las especies, cuyo número ronda los 3-5 millones. Los escarabajos son los que más especies tienen, y con ellos se infieren las estimaciones sobre la biodiversidad.

La filogenia es una disciplina muy consolidada que empezó hace aproximadamente 200 años. La filogenia trabaja con árboles evolutivos, que son las **representaciones gráficas (patrones) de las relaciones ancestro-descendientes (relaciones históricas de parentescos) entre elementos**, que pueden ser especies, secuencias de genes, etc. Entender este patrón es esencial para realizar estudios comparativos de cualquier tipo, porque existen **dependencias estadísticas entre los elementos que comparten ancestros comunes**. Conforme va pasando el tiempo, se van aplicando diferentes modelos evolutivos y se van depurando. Cuantos más datos se añadan (más especies o más secuencias), se obtiene una mejor aproximación.

La filogenia sirve, entre otros, para:

- Evolución de los seres vivos
- Genómica
- Ingeniería genética
- Farmacia
- Epidemiología
- Biología de la conservación
- Control de plagas
- Lingüística

La filogenética puede ser estudiada de diversas maneras. A menudo ha sido estudiada utilizando registros fósiles, que contienen información sobre la morfología de los antepasados de especies actuales y la cronología de divergencias. Esto permite datar las filogenias. Sin embargo, el uso de registros fósiles tiene muchas limitaciones: pueden estar disponibles sólo para determinadas especies, los datos existentes de fósiles pueden estar fragmentados, la recolección de datos está limitada por la abundancia, hábitat, rango geográfico y otros factores, y las descripciones de los rasgos morfológicos son a menudo ambiguas (múltiples factores genéticos). Por todo esto, utilizar registros fósiles para determinar relaciones filogenéticas puede producir sesgos. Además, los fósiles de microorganismos son prácticamente inexistentes, imposibilitando el uso de este enfoque. Afortunadamente, los datos moleculares que están en la forma de secuencias de ADN o de proteínas pueden ser también muy útiles para proporcionar una perspectiva de la evolución de los organismos, como el ARN 16S. Debido a que los genes son el medio para registrar las mutaciones acumuladas, éstos pueden servir como "fósiles moleculares". A través del análisis comparativo de secuencias de ADN de una serie de organismos relacionados, la historia evolutiva de los genes e incluso de los organismos puede ser revelada. La ventaja de utilización de datos moleculares es que son más numerosos que los registros fósiles y más fáciles de obtener. Además, no hay ningún sesgo de muestreo, como el que hay en los registros fósiles reales. Por tanto, es posible construir árboles filogenéticos más precisos y robustos utilizando datos moleculares.

La filogenia representa un registro indirecto del proceso evolutivo al ser una reconstrucción de la evolución de caracteres. Se deben realizar test de homología para ver que los caracteres se pueden comparar entre sí al compartir un origen común. De esa forma se obtiene información para construir clasificaciones y hacer predicciones dentro de un marco temporal cuando es posible obtenerlo.

I.1. Conceptos básicos

Los árboles filogenéticos suelen ser binarios, estando compuestos por **nodos externos o terminales** y **nodos internos** unidos por **ramas** que parten de una **raíz**. A través de las diferentes ramas se van reconstruyendo las relaciones entre las especies. Los **nodos internos son hipótesis evolutivas de posibles ancestros comunes** de los cuales normalemente faltan datos para confirmar o descartar la teoría. En las distintas ramas se pueden representar la transformación de caracteres que aparecen a nivel genético y que se transmiten por herencia.

Se denominan **grupos hermanos** a los nodos terminales que parten de un mismo nodo interno, es decir, dos taxones que compartan un ancestro común no compartido por ningún otro taxón. El **grupo externo (outgroup)** es aquel que se encuentra más alejado y parte de una rama distinta desde la raíz. Normalmente, este outgroup se elige de forma consciente para poder colocar la raíz donde se estima correcto. Todas las especies que se desarrollan desde una rama de la raíz se denomina **grupo interno o ingroup**.

Los árboles filogenéticos se pueden representar sin enraizar o enraizado. Un árbol filogenético sin raíz no asume conocimiento de un ancestro común, solo posiciones de los taxones para mostrar sus relaciones relativas (no hay dirección de un camino

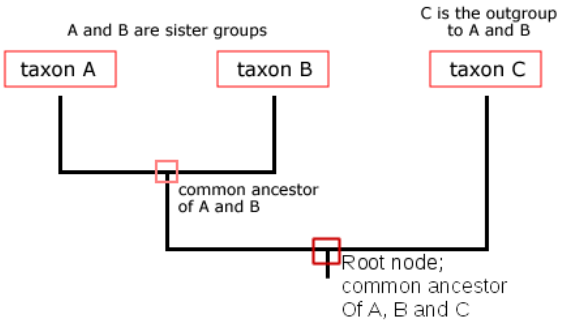

Figura I.1: *Partes de un árbol filogenético.*

evolutivo). Para definir la dirección de la evolución se necesita un árbol filogenético con raíz donde todas las secuencias bajo estudio tienen un ancestro o nodo raíz común (más informativo). Mientras que los árboles filogenéticos se centran en las relaciones evolutivas entre diferentes especies, las redes haplotípicas son representaciones gráficas sobre las relaciones evolutivas entre las diferentes poblaciones.

A la hora de visualización, hay varias formas de representar los árboles filogenéticos. Los distintos elementos no tienen un orden concreto; da igual si en un árbol los nodos terminales están en distinto orden mientras que las ramas sigan el mismo camino. En general, se suelen poner los nodos terminales de manera que sea más fácil de leer a simple vista.

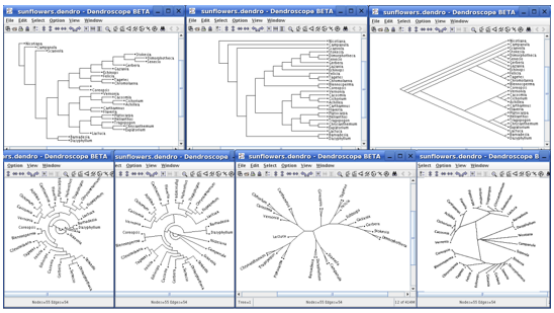

Figura I.2: *Distintas representaciones de los árboles filogenéticos.*

I.2. Politomías

Topología es la forma en que se ramifica un árbol. Cuando todas las ramas se bifurcan en un árbol filogenético, éstas son denominadas como una **dicotomía**. Por el contrario, si una rama tiene más de dos descendientes, entonces se denomina **politomía**.

Los árboles filogenéticos se consideran resueltos cuando de un nodo interno salen las distintas terminales. En la mayoría de casos, los árboles son no resueltos y tienen politomías, es decir, que desde un nodo interno no se sabe cómo han avanzado las especies. A partir de ahí solo se pueden añadir más datos, pintar uniones con un bootstrap bajo (es decir, un bajo soporte de esa bifurcación), o justificarlo como que están en el momento de especiación. Dentro de las hipótesis filogenéticas siempre hay más de una solución (se producen varios árboles igualmente óptimos), así que el árbol

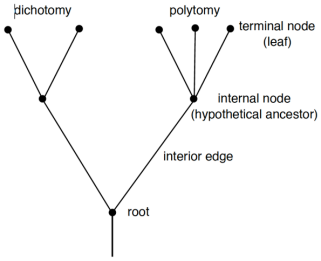

Figura I.3: *Diferencia entre dicotomía y politomía.*

final se debe elegir. Un árbol de consenso puede ser construido mostrando las porciones de bifurcación resueltas comúnmente y colapsando aquellas que no concuerdan entre los árboles. En un árbol de consenso estricto, todos los nodos en conflicto son colapsados.

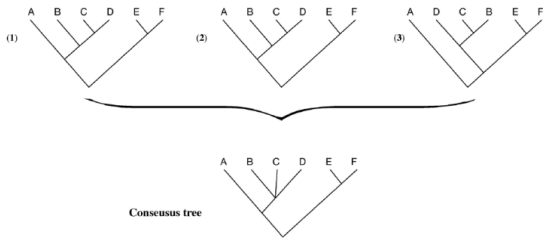

Figura I.4: *Árbol de consenso.*

I.3. Tipos de árboles filogenéticos

Existen distintos tipos de árboles filogenéticos. Los **filogramas (phylogram)** miden en las ramas los cambios que ha habido por sitio, por lo que las longitudes de las ramas representan a escala la cantidad de divergencia evolutiva. Tienen la ventaja de mostrar tanto las relaciones evolutivas como la información sobre el tiempo relativo de divergencia de las ramas. Los **cladogramas (cladogram)** muestran la similitud de los distintos elementos, pero las longitudes de sus ramas no son proporcionales al número de cambios evolutivos y, por tanto, no tienen ningún significado filogenético. Los **cronogramas (chronogram)** representan la relación de los elementos de forma temporal.

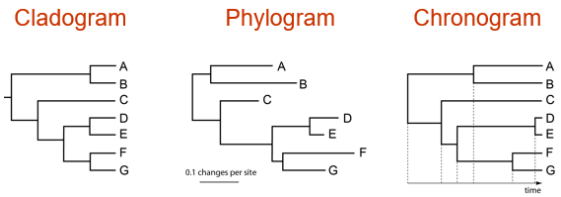

Figura I.5: *Tipos de árboles filogenéticos.*

I.4. Inferencia filogenética

Cualquier episodio histórico es, por definición, irrecuperable. La única forma que tenemos de reconstruirlo es a través del estudio de sus efectos. Por ello, la reconstrucción filogenética es un proceso de inferencia: se intenta obtener la mejor estima posible de una historia evolutiva basada en la información incompleta y con frecuencia ruidosa contenida en los datos. Las evidencias que se emplean pueden ser morfología (comparación entre caracteres de especies), ultraestructura (cortes vistos al microscopio electrónico), embriología (fases del desarrollo embrionario), paleontología (registro fósil), etología (comportamiento animal), bioquímica y moléculas.

Un **carácter** es una característica de los taxones que supuestamente es heredada (si no es heredada, no se puede utilizar la filogenia). El **estado de carácter** es el valor específico que toma un carácter en un taxón concreto. Por ejemplo, un carácter sería tener ojos y el estado de ese carácter sería 2 para humanos y 8 para algunas arañas.

I.5. Homología

La **homología** es la relación que existe entre dos partes orgánicas diferentes de dos organismos distintos cuando sus determinantes genéticos tienen el mismo origen evolutivo, es decir, cuando un mismo órgano tiene diversas formas y funciones. Los carácteres que se estudian en filogenia deben ser homólogos. Se compara la semejanza de una estructura debido a la herencia común. Por el contrario, la analogía es una estructura semejante a otra o que tiene la misma función, pero cuyo desarrollo embrionario y origen son diferentes. No se presentan en un antepasado común (como en el caso de los carácteres homólogos), si no que es fruto de convergencia evolutiva.

Dentro de la homología se distinguen dos tipos: la ortología y la paralogía. Los **genes ortólogos** son semejantes por pertenecer a dos especies que tienen un antepasado común. Los **genes parálogos** son aquellos que se encuentran en el mismo organismo y cuya semejanza revela que uno procede de la duplicación del otro (y puede adquirir funciones diferentes del gen original). La ortología requiere que se haya producido especiación, mientras que esta no es necesaria en el caso de la paralogía, que puede producirse solo en los individuos de una misma especie. Por ello, idealmente se deben comparar carácteres ortólogos para hacer las reconstrucciones filogenéticas.

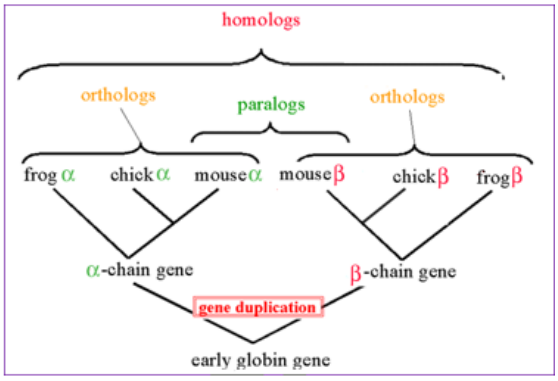

Figura I.6: *Homología en genes de la hemoglobina.*

I.5.1. Tipos de homología

En cladística, se emplean unas clasificaciones de las propiedades de organismos basándose en similitudes derivadas. Se refiere a **plesiomorfia** un estado ancestral o primitivo de un carácter al estar presente en grupos externos y en los ancestros. Cuando un carácter es evolutivamente novedoso pero deriva de otro rasgo perteneciente a un taxón ancestral filogenéticamente próximo, se denomina **apomorfia**. Así, se emplean los adjetivos plesiomófico y apomórfico en lugar de primitivo y avanzado para evitar juicios de valor sobre la evolución de los carácteres. Una **sinapomorfia** es una apomorfia (carácter exclusivo) compartida por un ancestro común y todos sus descendientes. Una **simplesiomorfia** se refiere a una plesiomorfia (carácter ancestral) compartida por dos o más taxa. Finalmente, una **autapomorfía** es un carácter único de un taxón que no aparece en el antepasado.

El prefijo "sin" viene de "compartido". Por tanto, los caracteres sinapomorfos son caracteres apomorfos compartidos, mientras que las simplesiomorfías son plesiomorfías compartidas.

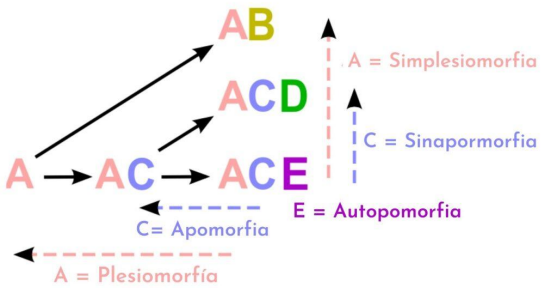

Figura I.7: *Tipos de homología en el árbol filogenético (tumbado). El carácter A es plesiomórfico al estar en el ancestro. El carácter C es apomórfico al ser una novedad evolutiva. En los nodos terminales, el carácter A se considera simplesiomórfico al estar compartido por los descendientes y ser un carácter ancestral. Por el contrario, el carácter C en los nodos terminales es sinapomórfico por ser un carácter novedoso y estar compartido en el ancestro en el que surgió y sus descendientes. Los caracteres B, D y E son autopomorfos por estar presentes en un único nodo terminal.*

I.5.2. Homoplasia

La homoplasia es el cambio evolutivo paralelo que hace que dos organismos presenten un mismo carácter adquirido independientemente. La **convergencia** se da cuando dos estructuras similares han evolucionado independientemente a partir de estructuras ancestrales distintas y por procesos de desarrollo diferentes. Se considera que el **paralelismo** involucra patrones de desarrollo similares en líneas evolutivas diferentes pero próximas. La diferencia con la convergencia es que en el paralelismo, hay un ancestro que no presenta un carácter y dos descendientes directos sí presentan esa novedad evolutiva, mientras que en la convergencia los descendientes con carácter no tienen el mismo ancestro común directo. No obstante, en la práctica, la distinción

entre convergencia y paralelismo es un tanto arbitraria porque no existe una regla exacta para limitar la antigüedad del antepasado común. Finalmente, en la **reversión**, un organismo adquiere un carácter de sus antepasados más lejanos. Esto implica que uno o más caracteres adquiridos previamente se han eliminado y se han vuelto a los más anteriores.

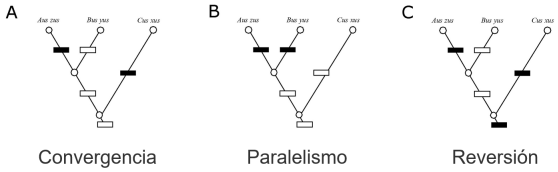

Figura I.8: *Diferencias entre convergencia, paralelismo y reversión.*

I.5.3. Agrupamientos

Un **grupo monofilético** es un clado que contiene un ancestro y todos sus descendientes, formando así un solo grupo evolutivo. Un **grupo parafilético** es similar, pero excluye a algunos de los descendientes que han sufrido cambios significativos. Un grupo con miembros de líneas evolutivas separadas se llama **polifilético**, conteniendo así grupos de especies con distintos ancestros comunes.

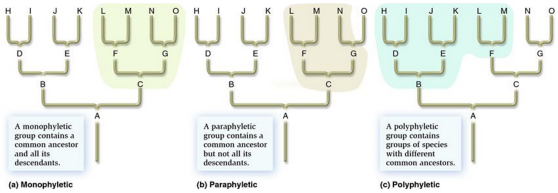

Figura I.9: *Agrupamientos de grupos monofiléticos, parafiléticos y polifiléticos.*

De esa forma, los grupos monofiléticos presentan sinapomorfía, los grupos parafiléticos presentan simplesiomorfía, y los grupos polifiléticos homoplasia.

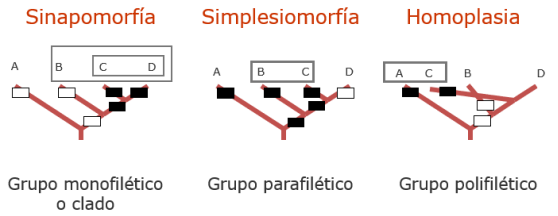

Figura I.10: *Grupos y caracteres que se apoyan.*

I.5.4. Fenotipo vs moléculas

Tradicionalmente se han empleado los **carácteres fenotípicos** para establecer las relaciones filogenéticas. Esto se debe a que suelen ser carácteres evolutivamente relevantes y complejos menos proclives a la homoplasia. Además, son los únicos carácteres disponibles en algunos casos como en fósiles o especímenes raros. No

obstante, puede haber problemas de codificación de taxones supraespecíficos como terminales (quimares) y se pueden dar casos de subjetividad en la codificación de carácteres. Además, hay un número limitado de carácteres fenotípicos y podemos encontrar taxones altamente autapomórficos.

Recientemente se están empleando **carácteres moleculares** al ser estrictamente heredables y no haber ambigüedades en la codificación. Por ello, determinar el estado de los carácteres es trivial. Hay ciertas regularidades en la evolución de los carácteres moleculares, y éstos son robustos frente a la distancia evolutiva. También son muy abundantes y ofrecen información temporal. El problema de los carácteres moleculares es que son más proclives a la homoplasia al tener solo 4 nucleótidos y 20 aminoácidos. La evolución de estos carácteres es compleja. Además, los árboles de genes no siempre coinciden con los árboles de especies. La determinación de la homología puede ser difícil por duplicación o pérdida de genes y alineamientos.

Se suelen utilizar multitud de genes separados y analizarlos de forma separada. El consenso de análisis separados es una estimación conservadora de la filogenia. Algunos métodos filogenéticos solo se pueden aplicar a ciertos tipos de datos. A nivel de especies, la concatenación de genes diferentes puede ser inapropiada si se da transferencia horizontal de genes, hibridación, duplicación de genes o coalescencia más profunda que el tiempo de divergencia. El conflicto entre caracteres se resuelve teniendo en cuenta toda la evidencia disponible y realizando análisis combinados. Diferentes tipos de datos proporcionan información a diferentes niveles filogenéticos. La señal filogenética aumenta debido a la congruencia entre caracteres de diferentes conjuntos de datos.

Es importante que el conjunto de datos sea lo más completo posible. Es necesario hacer un muestreo de taxones (incluyendo los grupos externos) y genes razonable y justificado.

Capítulo II

Alineamiento de secuencias

Decidir qué caracteres investigar, y cómo codificarlos, es un primer paso crucial en cualquier análisis filogenético.

II.1. Tipos de caracteres

Hay **sitios invariables** que no cambian en los distintos taxones. También hay **sitios filogenéticamente neutrales** que son autapomorfías. Los **sitios filogenéticamente informativos** son comunes por pares, por lo que son sinapomorfías.

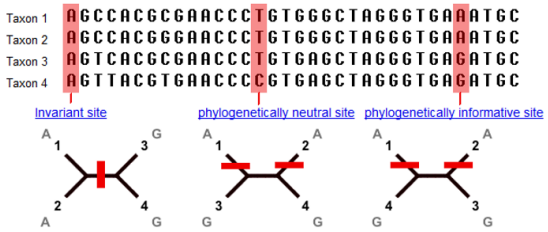

Figura II.1: *Sitios en una secuencia invariantes (no cambian entre los distintos taxones), filogenéticamente neutrales (solo cambia en un taxón) y filogenéticamente informativos (permiten dicotomía).*

Los caracteres pueden ser **binarios 0/1** (presentes o ausentes), **multiestado** o **binarios V/S** (transversiones o transiciones). Los caracteres también pueden ser **discretos o continuos**. La codificación de caracteres continuos no se pueden incluir fácilmente en las matrices de caracteres, por lo que se debe realizar una categorización arbitraria. Idealmente, se deben buscar divisiones naturales, es decir, estados discretos de un carácter de variación continua.

II.2. Ponderación de los caracteres

Se puede emplear un valor relativo de los diferentes caracteres y transformaciones como indicadores de las relaciones filogenéticas entre taxones. Se puede realizar una ponderación uniforme, que minimiza los supuestos del análisis, o una ponderación

diferencial, en la que no todas las características de un organismo tienen el mismo valor como evidencias filogenéticas.

II.2.1. Ponderación *a priori*

En la ponderación *a priori* de caracteres morfológicos, los taxónomos pueden tener muchas razones para asumir que diferentes caracteres tienen diferente importancia filogenética. Pero eso tiene dos problemas: diferentes opiniones expertas y, en caso de acuerdo, el peso proporcional que se le da a cada carácter. Se introduce precisamente el tipo de subjetividad que el análisis cladístico pretende evitar.

II.2.2. Ponderación *a posteriori*

El método más utilizado y aplicado, denominado **ponderación implícita**, se basa en Goloboff (1993): la primera vez que un carácter cambia de estado en un árbol, este cambio de estado recibe el peso «1»; los cambios posteriores son menos «costosos» y reciben pesos menores a medida que la tendencia de los caracteres a la homoplasia se hace más evidente. Los árboles que maximizan la función cóncava de homoplasia resuelven el conflicto de caracteres a favor de los caracteres que tienen más homología (menos homoplasia) e implican que el peso medio de los caracteres sea lo más alto posible.

Goloboff reconoce que los árboles con los pesos medios más elevados son los que más «respetan» los datos: un peso medio bajo implica que la mayoría de los caracteres están siendo «ignorados» por los algoritmos de construcción de árboles. Aunque originalmente se propuso con una ponderación severa de k=3, Goloboff prefiere ahora concavidades más «suaves» (por ejemplo, k = 12), que han demostrado ser más eficaces en casos simulados y del mundo real.

$$F = \Sigma fi, fi = \frac{k}{k+(s-m)}$$

m: número mínimo de pasos
s: número de pasos observados
k: constante de concavidad

II.2.3. Secuencias de ADN

Generalmente, se toma la tasa de sustitución como medida de la fiabilidad de la información filogenética del marcador. Se entiende entonces homoplasia como saturación. Las transversiones evolucionan lentamente y aumentan su frecuencia a medida que pasa el tiempo. Las transiciones se saturan a partir de cierta distancia filogenética, perdiéndose su señal.

El coste de las transformaciones se determina empíricamente mediante matrices de costes (stepmatrices).

II.3. Polaridad de los caracteres

Necesitamos conocer la polaridad de los caracteres para poder enraizar los árboles. Para ello, se debe establecer qué carácter es ancestral y qué carácter es derivado. Utilizando un **criterio ontogenético**, se ve cómo se forma el carácter durante el desarrollo para poder establecer la polaridad. En caso de que no quede claro tras ese criterio, se compara con el outgroup para establecer el estado primitivo del carácter.

II.4. Homología de los caracteres moleculares

Cuando analizamos secuencias, asumimos que son de moléculas heredadas de ancestros a descendientes (ortólogos). Cada secuencia está formada por muchos caracteres (cada posición en la secuencia). Por ello, un primer paso es determinar el estado de cada uno de esos caracteres en cada taxón de la matriz. Importante: La homología de los caracteres moleculares, como la de cualquier otro tipo de carácter, es un concepto cualitativo. Las secuencias del gen A de dos taxones son homólogas, o bien no lo son. Igualmente, la posición X en la secuencia de un taxón es homóloga de la posición Y en la secuencia de otro taxón, o bien no lo es. Pero NO puede decirse que las secuencias de dos taxones muestren mayor o menor homología (por ejemplo, en %). Podrán tener diferente porcentaje de similitud (p. ej., % de bases o aminoácidos idénticos en posiciones homólogas), pero o son homólogas o no lo son.

II.4.1. Aplicación del concepto de homología a los genes: alineamiento de secuencias

Un alineamiento es una hipótesis acerca de la homología posicional de diferentes secuencias de bases o aminoácidos. El alineamiento tiene como objetivo identificar qué posiciones son homólogas en diferentes secuencias. Cada posición de la secuencia (residuo = nucleótido o aminoácido) se interpreta como un carácter que puede tomar diferentes valores (estados de carácter: una de 4 bases, o uno de 20 aminoácidos). El alineamiento asume parsimonia: el cambio evolutivo es improbable, de modo que los segmentos de secuencia coincidentes sirven de guía para identificar posiciones homólogas. Eventualmente se identifican cambios, que cuando son compartidos por varias especies son informativos para la reconstrucción de filogenias.

Las secuencias pueden no tener la misma longitud. Los gaps son marcadores de posición que introducimos en los alineamientos para mantener la homología posicional. Representan eventos de inserción o pérdida denominados indels (del inglés insertion/deletion). La ventaja es que los indels son, en principio, menos propensos a la homoplasia que las sustituciones de bases, muy utilizadas en análisis de parsimonia. No obstante, los indels son difícilmente gestionables por la mayoría de los modelos de evolución molecular.

Es importante elegir un buen alineamiento, ya que la calidad del alineamiento influye en la calidad de la inferencia filogenética.

II.4.2. Decidir el mejor alineamiento

No existe ningún procedimiento automático para elegir objetivamente el mejor alineamiento: hay que valorar la calidad de los diferentes alineamientos posibles y elegir el que nos parezca mejor. Elegimos como mejor alineamiento el supuesto más razonable de acuerdo con un algoritmo informático y el ojo experimentado. En cualquier caso, es siempre importante examinar el resultado críticamente para valorar si tiene sentido desde un punto de vista biológico.

No todos los alineamientos son igualmente parsimoniosos. Para valorar la calidad de los alineamientos, se han propuesto diferentes mecanismos de puntuación. Se puede realizar una **puntuación por identidad**. Un alineamiento de dos secuencias puede interpretarse como una matriz con dos filas y n columnas (n = longitud del alineamiento). Las posiciones (columnas) con idéntico residuo (base o aminoácido) tienen una puntuación = 1. La puntuación del alineamiento es la suma de las puntuaciones de todas sus posiciones. El alineamiento óptimo es el que maximiza la identidad de las columnas. No obstante, los gaps no penalizan, por lo que pueden darse alineamientos con misma puntuación, pero más posiciones de diferencia. Por tanto, se pueden aplicar penalizaciones para los huecos en la secuencia, ya sea introduciendo penalizaciones por la apertura de los huecos o por la extensión de los huecos abiertos. Estos últimos son típicamente menores que las impuestas por apertura. Por ejemplo, se puede aplicar una penalización de -2 por apertura de gap y de -1 por extensión del gap abierto.

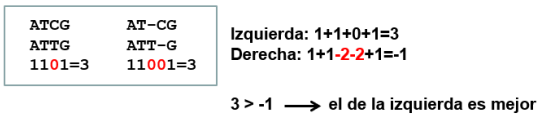

Figura II.2: *Ejemplo de cálculo del mejor alineamiento.*

No tiene mucho sentido alinear las secuencias de ADN de los genes codificantes de proteínas. Es mejor traducir las secuencias de ADN a secuencias de aminoácidos y alinear éstas últimas. Existen varios programas para alineamiento múltiple: clustal W/X/Omega, MAFFT, Muscle, T-Coffee, Dialign 2, etc.

Capítulo III

Modelos de evolución

Todos los métodos de inferencia y reconstrucción filogenética implican una serie de **supuestos**, aunque éstos no se hagan explícitos:

- Todos los sitios o posiciones cambian independientemente.
- Las tasas de evolución son constantes a lo largo del tiempo y entre linajes.
- La composición de bases es homogénea.
- La verosimilitud de los cambios de base es la misma para todos los sitios y no cambia a lo largo del tiempo.

Esto son asunciones, pero en realizar no son ciertos. Las tasas de evolución no son constantes, las posiciones no cambian independientes las unas de las otras, la composición de bases no es homogénea (hay mayor porcentaje de GC que de AT) y se pueden dar múltiples cambios en un único sitio que quedan ocultos (si el nucleótido original es C, puede que en un organismo cambie a A y en otro a G). Estos cambios ocultos hacen que las secuencias estén cada vez más saturadas: la mayoría de los sitios que cambian han cambiado antes.

En un contexto filogenético, los modelos predicen el proceso de sustitución de las secuencias a través de las ramas. Describen probabilísticamente el proceso por el que los estados de los caracteres homólogos de las secuencias (posiciones alineadas: nucleótidos o aminoácidos) cambian a lo largo del tiempo.

Los modelos implican por lo general los siguientes **parámetros**:

- **Composición:** frecuencia de las diferentes bases o aminoácidos.
- **Proceso de sustitución:** tasa de cambio de uno a otro estado de carácter.
- **Otros parámetros (heterogeneidad de tasas):** proporción de sitios invariables o agregación de los cambios a lo largo de la secuencia.

III.1. Modelos frecuentes

El modelo más sencillo es el de Jukes Cantor, el cual asume que todos los cambios son igualmente probables y que la frecuencia de todas las bases es la misma. A partir de este, la complejidad empezó a aumentar, ya que las combinaciones de parámetros son muchas. Algunos de los modelos más frecuentes son:

- **Jukes and Cantor (JC69)**: La frecuencia de todas las bases es la misma (0.25 cada una), y la tasa de cambio de una a otra base es igual.
- **Kimura 2-parámetros (K2P)**: La frecuencia de todas las bases es la misma (0.25 cada una), pero la tasa de sustitución es diferente para transiciones y transversiones.
- **Hasegawa-Kishino-Yano (HKY)**: Como K2P, pero la composición de bases varía libremente.
- **General Time Reversible (GTR)**: La composición de bases varía libremente, y todas las sustituciones posibles pueden tener distintas frecuencias.

Cada vez, los modelos son más complejos, y normalmente se utiliza el más complejo. Hay programas que ya proponen un modelo a elegir según los datos que se le proporcionen.