Prácticas 1-3: La evolución de las aves ratites

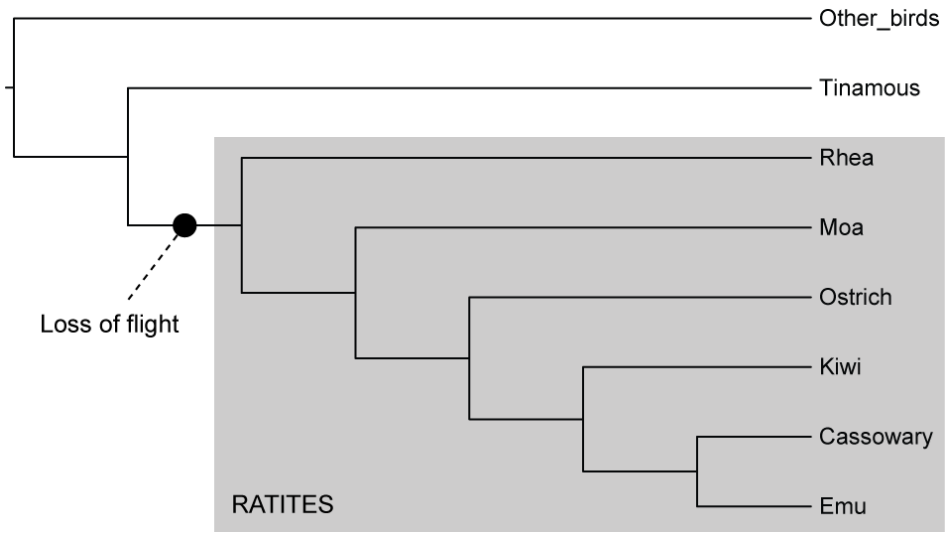
Introducción

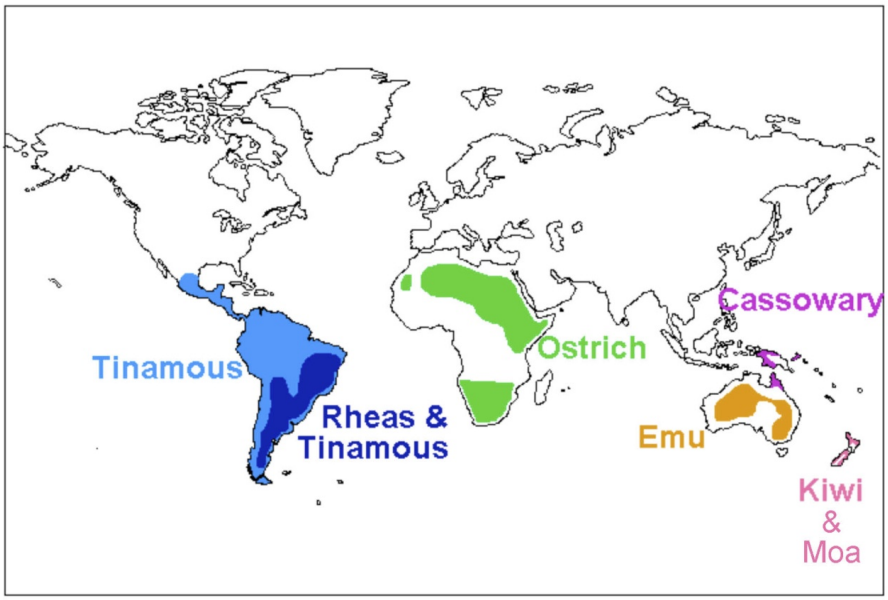
Las aves modernas están divididas en dos subclases: **Neognathae**, que incluye el >99% de todas las especies existentes de aves, y **Palaeognathae**, que incluye al tinamú y las ratites. Ésta división básica es soportada por diversas fuentes de evidencia morfológica y molecular.

Las ratities son aves que han perdido el vuelo. Incluyen la avestruz (*ostrich*; África), la rea (América del Sur), el kiwi (Nueva Zelanda), el emu (Australia), y el casouario (Australia y Nueva Guinea), al igual que las recientemente extintas moa (Nueva Zelanda) y ave elefante (Madagascár). Las ratites son generalmente grandes herbívoros u omnívoros. Los palaeognathos también incluyen a el tinamú sudamericano, un ave que tiene la capacidad de volar.

Una relación cercana entre el tinamú y las ratites es ampliamente aceptada, pero queda gran incertidumbre sobre las relaciones evolutivas entre las especies de ratites. Las aves de ratites que no tienen vuelo suelen ser vistas como un grupo monofilético, con el tinamú como el familiar más cercano. De acuerdo con esto, la interpretación mas parsimoniosa de su historia evolutiva es que la habilidad de volar se perdió en un linaje de ratite ancestral, después de haber divergdo del linaje que lleva al tinamú.

Junto con las plantas fagáceas del sur (*Nothofagus*) y ciertos grupos de peces de agua dulce, las ratites son vistas como un grupo ejemplar de **Gonduana**. En adición, a menudo son consideradas como evidencia sólida de un **modelo de biogeografía vicariante**. En otras palabras, se ha sugerido que la diversificación de las ratites es consistente con patrones de movimiento continental, y específicamente la ruptura del supercontinente de Gonduana. Una visión opuesta es que las ratites se diversificaron en un proceso de **dispersión geográfica**.





Como se puede ver en el árbol filogenético molecular en la parte superior de la página, Cooper *et al.* (2001) estimó que la rea es el grupo hermano de todas las ratites. Este descubrimiento contrasta con la hipótesis de vicarianza para explicar la evolución de las ratites. De acuerdo a vicarianza, la avestruz debe ser la hermana de el resto de las ratites porque África fue el primer continente en separase de Gonduana. En adición, los dos grupos de Nueva Zelanda (el kiwi y la moa) no aparecen como hermanos en la figura, sugiriendo que sus ancestros debieron colonizar Nueva Zelanda independientemente en lugar de haber divergido cuando los continentes se separaron.

Otra fuente de discusión ha emergido recientemente, esta vez centrada en la **relación del tinamú** a las ratites. En el pasado, se asumía que el tinamú con su capacidad de volar era el hermano de las ratities, que no vuelan. Sin embargo, análisis detallados le grandes cantidades de datos de secuencias de ADN en años recientes han producido resultados sorprendentes.

En esta practica, usaras métodos filogenéticos para investigar estas cuestiones centrales sobre la evolución de Palaeognathae. Empezarás alineando un set de secuencias de ADN. Los datos luego serán analizados usando un popular método filogenético en el programa gratuito *MEGA*.

Práctica 2: Alinear secuencias

Antes de iniciar, asegurate de tener una version reciente de *MEGA* (version 7 o X) y que tengas el archivo de ADN **ratites.fasta**. Este archivo contiene sequencias de ~2000 nucleotidos de 8 especies de aves: casouario, emu, kiwi, moa, aveztruz, rea, tinamú, y la gallina.

En esta sesión practica vas a usar el programa gratuito de filogenética *MEGA*. Este programa incluye un amplio rango de métodos de análisis de secuencias, incluyendo análisis filogenético usando el método de máxima verosimilitud.

a) Abre el programa MEGA. El programa ofrece una plataforma integrada para llevar a cabo varios tipos de análisis de datos genéticos.

b) Abre el archivo del alineamiento **ratites.fasta** y selecciona “Align”. Mira las secuencias y fíjate de sus diferencias en longitud (a excepción de la avestruz y el tinamú). La secuencia de la gallina es la más larga (2000 nt).

Q. *Es necesario alinear estas secuencias antes de llevar a cabo un análisis filogenético. Cual es el propósito del alineamiento de secuencias?*

El objetivo de alinear secuencias es usar únicamente caracteres homólogos, los cuales tienen un ancestro común. Queremos maximizar el numero de sitios para los cuales podemos confirmar homología. Esto incluye insertando espacios en algunos sitios donde creemos que hubo inserciones o perdidas de nucleótidos.

Intenta alinear estas secuencias a ojo por unos pocos minutos. Para hacer esto, usa tu cursor para seleccionar nucleótidos, y luego usa las flechas en la parte superior de la ventana del alineamiento para mover los bloques de columnas en las dos direcciones. También puedes usar tu teclado, usando la tecla alt junto con as teclas de flechas.

Como idea para comenzar, puedes seleccionar la secuencia de la rea y resaltar el nucleótido “C” en la columna 26, y muévelo 3 columnas a la derecha. Fíjate que la secuencia ahora parece mejor alineada con el resto de los datos.

c) Ahora vamos a hacer que el computador nos haga el alineamiento. En el menu **Alingnment** elige “Align by Muscle” y selecciona todas las otras secuencias. El método Muscle es uno de los dos que hay en *MEGA*. El otro es ClustalW. Estos dos son quizás los métodos más ampliamente utilizados. En esta practica nos enfocaremos en usar Muscle.

Cambia la penalización “Gap Open” a -50. Esto reduce la penalización para insertar espacios en el alineamiento. Muscle ahora estará más dispuesto a insertar espacios para hacer que las secuencias estén lo mejor alineadas posible. Haz click en “OK”.

Q. *¿Que tan largo es el alineamiento de estas secuencias?*

2118 sitios.

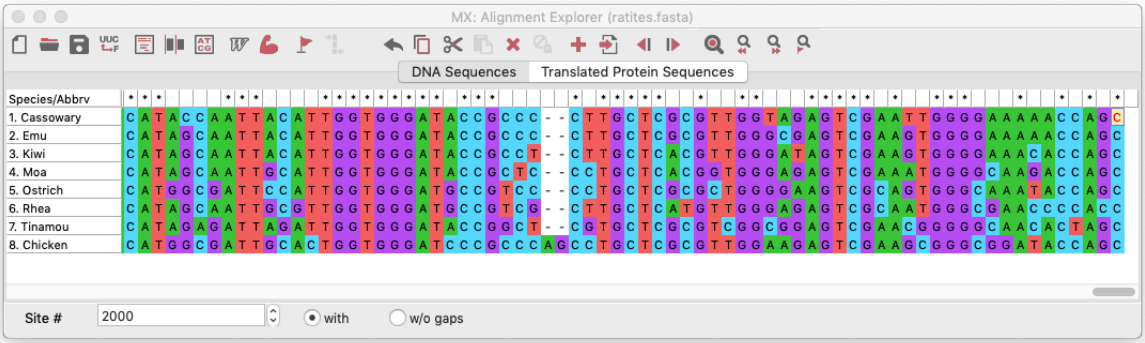
d) Ahora usa Muscle de nuevo, pero escribe en “Gap Open” una penalización de -400.

Q. *¿Que tan largo es le alineamiento de estas secuencias?*

2000 sitios.

Parece que reducir la penalización a -50 hace que Muscle inserte demasiados espacios a los datos. El segundo alineamiento, producido con una penalización de -400 parece más razonable y lo vamos a usar para nuestros análisis.

En esta práctica vamos a aceptar los resultados del alineamiento automático. Sin embargo, en un estudio formal se debe inspeccionar el alineamiento para confirmar que el método haya funcionado correctamente. En algunos casos, la inspección visual revela secciones del alineamiento que deben mejorar. Hoy día, muchos datos genéticos son demasiado grandes para que la inspección visual sea práctica, haciendo que haya una mayor dependencia en los métodos automáticos.



Q. *Si una de nuestras secuencias se moviera accidentalmente a la derecha 1 nucleótido (de manera que estaría desalineada del resto de secuencias por 1 nucleótido), ¿cuál sería la consecuencia para nuestro análisis filogenético?*

Esto llevaría a una gran sobre-estimación de la divergencia genética entre esta secuencia y las demás secuencias en los datos. Llevaría también a una sobre-estimación de la tasa de evolución molecular. En términos del árbol filogenético, la rama que lleva a la secuencia problemática sería extremadamente larga. La secuencia además estaría localizada en una posición incorrecta de la filogenia.

Ahora que hemos alineado nuestras secuencias de ADN, estamos listos para analizar nuestros datos.

Sección B: Análisis filogenético

Esta sección tiene dos partes:

- Análisis filogenético usando máxima verosimilitud
- Selección de modelos

Análisis filogenético usando máxima verosimilitud

Ahora podemos hacer estimativos de árboles filogenéticos. Lo vamos a hacer usando el método de máxima verosimilitud. Este método estadístico fue aplicado a la filogenética por primera vez en los años 70 y fue formalizado al inicio de los 80.

Nuestro objetivo en este análisis será llegar a estimativos de los parámetros (ej., el árbol, las ramas) que contengan la máxima verosimilitud. Este método usa un modelo explicito de la evolución molecular, que por ahora será el más simple posible.

a) Dirígete de nuevo a la ventana principal de *MEGA*. Del menú **Phylogeny**, selecciona "Construct Maximum Likelihood Tree" ("Construir árbol de máxima verosimilitud") y usa los datos que tienes activos. Esto abrirá una caja que contiene un rango de opciones para hacer un análisis de máxima verosimilitud.

b) Confirma que tengas las siguientes opciones seleccionadas:

Test of Phylogeny:	Bootstrap method
No. of Bootstrap Replications:	100
Substitutions Type:	Nucleotide
Model/Method:	General Time Reversible model
Rates among Sites:	Gamma distributed with invariant sites (G+I)
No of Discrete Gamma Categories:	4
Gaps/Missing Data Treatment:	Use all sites

c) Haz click en "OK" para empezar el análisis de máxima verosimilitud.

d) El estimativo de la filogenia aparecerá en una nueva ventana. Confirma que el árbol esta enraizado entre la gallina y las demás especies. Si este no es el caso, selecciona la rama que lleva a la gallina y selecciónа "Place Root on Branch" ("Colocar raíz en rama").

Q. *Hay una barra de escala debajo del árbol. ¿En qué unidades está?*

La escala de la barra esta dada en sustituciones moleculares esperadas por sitio del alineamiento.

Q. *¿Cuál es el propósito de incluir una especie externa, o outgroup (en este caso la gallina), en el análisis?*

Incluir una especie externa nos permite inferir la posición de la raíz del proceso evolutivo.

Q. *¿El árbol tiene un soporte estadístico fuerte?*

La mayoría de los valores de soporte usando bootstrap son altos (>90%), aunque uno de los nodos tiene un soporte bajo (60-70%).

Selección de modelos

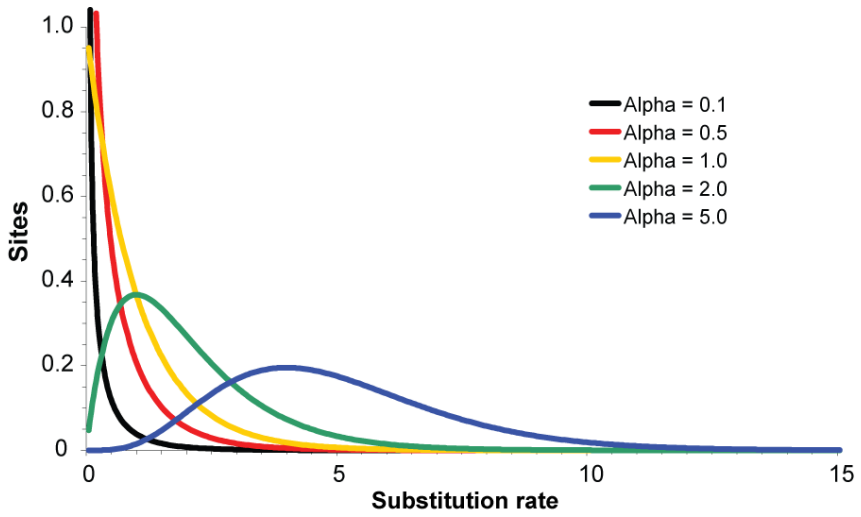
Una consideración importante en la filogenética es la selección del modelo de sustituciones. El objetivo de estos modelos es tomar en cuenta eventos de múltiples sustituciones. Quizás el modelo de sustitución más ampliamente usado es el General Time Reversible (GTR; modelo reversible en el tiempo generalizado), que permite diferentes tasas para cada uno de los tipos de sustitución. Por ejemplo, permite que las sustituciones A↔G ocurran a una tasa diferente a las de C↔G. El modelo también permite a los cuatro nucleótidos tener frecuencias ancestrales desiguales.

Q. *¿Cual es el modelo de sustitución más simple?*

El modelo Jukes-Cantor (JC), que tiene 0 parámetros libres.

También podemos permitir que la tasa evolutiva varíe entre columnas, o sitios, del alineamiento. De esta manera, asumimos que algunos sitios están restringidos (ej., por fuerzas de selección), mientras otros sitios son más libres de cambiar (ej., porque no alteran el *fitness* del organismo).

La variación en la tasa entre sitios se suele modelar usando una distribución gamma, que puede tomar una diversidad de formas. La forma de la distribución es determinada por un solo parámetro, alfa. Cuando alfa es pequeño (<1), muchos de los sitios evolucionan lentamente, pero unos pocos evolucionan rápidamente. Cuando alfa es grande (>1), la mayoría de sitios evolucionan aproximadamente a la misma tasa. Por razones computacionales usamos una distribución gamma discreta en lugar de continua. Normalmente entre 4 a 6 categorías de la tasa se usan para la distribución gamma discreta. Incrementar el numero de categorías hará que el análisis sea más demorado.



Para nuestra conveniencia, podemos comparar diferentes modelos usando *MEGA* y seleccionar el que mejor se ajuste a nuestros datos.

a) En el menú **Data** de la ventana Alignment Explorer, selecciona “Phylogenetic Analysis”. Ahora selecciona “No” cuando el programa pregunta si tienes “Protein-coding nucleotide sequence data” (“Datos de secuencias de nucleótidos que codifican a proteínas”). Estamos usando ADN no codificante, que no lleva a productos proteínicos.

b) Regresa a la ventana principal de *MEGA*. Del menú **Models**, selecciona “Find Best DNA/Protein Models (ML)” (“Encontrar el mejor modelo de ADN/proteínas”) y usa los datos que tienes activos. Esto traerá una caja que contiene una serie de opciones. Sin hacer cambios, haz click en “Compute”.

MEGA ésta calculando la verosimilitud (la probabilidad de los datos dado el modelo) para 24 modelos de sustitución diferentes. Mira los resultados del análisis. La primera columna muestra un lista de los modelos. La segunda muestra cuantos parámetros tiene cada modelo. La tercera y cuarta muestra los valores de dos criterios para la selección de modelos, el Bayesian Information Criterion (BIC) y el Akaike Information Criterion corregido (AICc), respectivamente. Para ambos criterios, valores más bajos indican que el modelo se ajusta mejor. Para más detalles puedes mirar la información debajo de la tabla.

Q. *¿Cómo se calculan el BIC y AICc? (busca explicaciones en internet)*

El BIC es igual a -2ln$L$+$k$ln(n), donde L es la verosimilitud, k es el número de parámetros libres en el modelos, y n es el número de muestras (en este caso el número de sitios en el alineamiento).

El AIC es -2ln$L$+2k, mientras el AICc esta basado en el AIC ero incluye una corrección adicional para números de muestras no-infinitos.

Q. *¿Cuál es el modelo que mejor se ajusta de acuerdo a BIC and AICc?*

El modelo GTR+I+G es el modelo que mejor se ajusta de acuerdo a ambos criterios.

En la tabla, el numero de parámetros libres de cada modelo incluye las longitudes de las ramas, porque estas se deben estimar en el análisis. Para el propósito de seleccionar un modelo, *MEGA* ha estimado el árbol filogenético usando un método rápido que no es máxima verosimilitud.

Q. *¿Cuántas ramas tiene el árbol? (pista: un árbol desenraizado de n puntas tiene 2n-3 ramas)*

Hay 8 especies en los datos, entonces hay 13 ramas en el árbol.

El modelo con más parámetros (GTR+I+G) tiene 10 parámetros libres, excluyendo las longitudes de las ramas. El modelo más simple (Jukes-Cantor o JC) tiene 0 parámetros libres.

Q. *¿Qué asunciones hace el modelo Jukes-Cantor?*

El modelo JC asume que los cuatro nucleótidos ocurren con la misma frecuencia (0.25) y que todas las tasas pareadas son idénticas (ej., la tasa de C↔G es idéntica a la tasa de C↔T). también asume que no hay sitios invariables (sin +I) y que todos los sitios evolucionan a la misma tasa (sin +G).

Q. *¿Cuáles son los 10 parámetros en el modelo GTR+I+G?*

El modelo GTR+G+I contiene 3 parámetros libres para las frecuencias de los nucleótidos (fíjate que la frecuencia del cuarto nucleótido esta restringida porque las cuatro frecuencias tienen que sumar a 1). Tiene también 5 parámetros libres para las tasas de sustitución entre nucleótidos pareadas (aunque hay 6 tipos de mutaciones pareadas, el valor de una de ellas se mantienen fijo mientras las otras 5 son estimadas relativas a la primera). Tiene 1 parámetro libre para a proporción de sitios invariables, y 1 parámetro libre para la forma de la distribución gamma que describe la variación de tasas entre sitios.

Q. *Para el modelo GTR+I+G, ¿cual es el estimativo del parámetro para la forma de la distribución gamma para estos datos? ¿Esto qué sugiere sobre la cantidad de variación en la tasa evolutiva entre sitios?*

El estimativo para el parámetro alfa de la distribución gamma es de 0.47. Esto sugiere un cantidad alta de variación en la tasa entre sitios, con muchos evolucionando lentamente y unos pocos evolucionando rápido.

Q. *¿Hay diferencias entre los resultados cuando usas el modelo JC contra el modelo GTR+G+I? ¿Qué te dice esto sobre los datos?*

Usar un modelo demasiado simple puede llevar a problemas en los resultados que queremos, como las relaciones entre organismos y las longitudes de las ramas del árbol. Específicamente, las ramas profundas del árbol suelen ser más cortas comparado a las ramas terminales cuando el modelo es demasiado simple (no observado usando lo datos de esta práctica).

Q. *¿Vez que las relaciones estimadas sean consistentes con la hipótesis de vicarianza? En otras palabras, ¿parece que las relaciones encajan con la secuencia de divergencias que esperaríamos si fueran causadas por la ruptura de Gonduana?*

Las dos aves neozelandesas (el kiwi y la moa) no se agrupan, sugiriendo que la hipótesis de vicarianza no puede explicar en exclusivo a todas las relaciones en el árbol. La localización de la Rhea también es inusual porque esta agrupada con las especies de Australia y Nueva Zelanda, en lugar de estar con el tinamú (que también es de origen suramericano).

Q. *¿Que sugiere la posición del tinamú sobre la perdida de vuelo en las aves ratities?*

El tinamú aparece agrupado con la moa y está dentro del grupo de las aves ratites. Esto sugiere que el vuelo pudo reaparecer en el linaje del tinamú, o que el tinamú retuvo la habilidad de volar mientras varios linajes de aves ratities independientemente perdieron el vuelo (aproximadamente 4 veces de acuerdo con el árbol). La segunda hipótesis es más plausible dado que es difícil evolucionar una característica tan compleja como lo es el vuelo.