Bioinformática aplicada al análisis de datos de metabarcoding de ADN ambienta

Práctico: Alineamiento y Filogenética

07 de diciembre, 2023

> "We do not like to ask, 'Is our model true or false?', since probability models in most data analyses will not be perfectly true. The more relevant question is, 'Do the model's deficiencies have a noticeable effect on the substantive inferences?'."
>
> – Gelman et al., 2004, Bayesian Data Analysis

Introducción

En este práctico continuamos analizando el set de datos de amplicones de genes 16S de ARNr generados por Griffero et al. (manuscrito en preparación). Estos datos fueron obtenidos de los arroyos afluentes de la Laguna de Rocha y la laguna de Rocha en Uruguay (https://maps.app.goo.gl/UE9kEwKJPfokc8N28), con la finalidad de estudiar la comunidad de procariotas y su respuesta a los contaminantes emergentes.

¡Comencemos!

Alineamiento

Antes de emprender análisis filogenéticos, es crucial realizar un alineamiento preciso de las secuencias genéticas. El alineamiento facilita la comparación de regiones homólogas y es esencial para inferir relaciones evolutivas entre diferentes especies o variantes genéticas. En este ejercicio, utilizaremos el programa MAFFT (Multiple Alignment using Fast Fourier Transform), conocido por su eficaz algoritmo de alineamiento. MAFFT utiliza un enfoque iterativo, para realizar un alineamiento preciso y detallado de las secuencias.

Introducción al Alineamiento con MAFFT: El procedimiento inicia con un alineamiento progresivo que organiza las secuencias en grupos básicos según su similitud. Luego, MAFFT realiza iteraciones adicionales, refinando y ajustando minuciosamente las secuencias para lograr un alineamiento más preciso. Este método garantiza una alineación exhaustiva y exacta.

Iniciaremos visualizando estas secuencias y luego ejecutaremos MAFFT para realizar el alineamiento inicial. Posteriormente, exploraremos variantes de este alineamiento, ajustando las penalidades por gaps, para evaluar cómo estos cambios impactan en el resultado.

Proceso de Alineamiento:

Visualización de las Secuencias ASV: Primero, observaremos las ASV para comprender la diversidad de las muestras.

```
less ASV.fasta
```

Alineamiento Inicial con MAFFT: Realizaremos el alineamiento utilizando MAFFT. La configuración predeterminada suele funcionar bien, pero exploraremos opciones adicionales para penalidades por gaps.

```
mafft ASV.fasta > ASV_alin.fasta
```

El archivo ASV_alin.fasta contendrá las secuencias alineadas.

Alineamientos Adicionales con Penalidades Diferentes: Para optimizar el alineamiento, produciremos dos versiones adicionales, una con una penalización aumentada para la apertura de gaps y otra para la extensión de gaps.

```
mafft --op ASV.fasta > ASV_op_alin.fasta
```

```
mafft --ep ASV.fasta > ASV_ep_alin.fasta
```

Estos ajustes nos permiten explorar diferentes configuraciones de penalidades por gaps y seleccionar la más apropiada según nuestras necesidades específicas.

Para visualizar los alineamientos utilizaremos el programa ALIVIEW (https://ormbunkar.se/aliview/).

Elección del modelo evolutivo

Antes de realizar una inferencia filogenética, es necesario llevar a cabo la selección del modelo evolutivo que mejor se ajusta a nuestros datos. Como vimos en el teórico, los modelos evolutivos describen las diferentes probabilidades de cambio de un nucleótido (o aminoácido) a otro a lo largo de un árbol filogenético, lo que nos permite elegir entre diferentes hipótesis filogenéticas para explicar los datos disponibles. Nuestro objetivo es elegir un modelo suficientemente complejo como para describir adecuadamente los datos disponibles, pero no tan complejo como para suponer más parámetros de los que los datos puedan respaldar adecuadamente.

Existe una diversidad de programas para la selección del modelo evolutivo. En éste práctico utilizaremos el ModelFinder (Subha Kalyaanamoorthy et al. 2017) implementado en IQ-TREE (Nguyen L.T., 2014). ModelFinder utiliza criterios estadísticos, como el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiana (BIC), para identificar el modelo que mejor equilibra la complejidad con la capacidad de ajuste a los datos. El modelo con los valores de AIC y BIC más bajos se selecciona como el modelo óptimo para la construcción del árbol filogenético.

Criterios de Información para la Selección de Modelos: AIC y BIC:

Los Criterios de Información, como el Criterio de Información de Akaike (AIC) y el Criterio de Información Bayesiana (BIC), son herramientas fundamentales en estadística y modelado para evaluar y seleccionar modelos estadísticos. La idea subyacente de ambos criterios es encontrar un modelo que proporcione un ajuste efectivo sin ser excesivamente complejo.

- Criterio de Información de Akaike (AIC):

El Criterio de Información de Akaike (AIC) se basa en la teoría de la información y busca encontrar un equilibrio entre el ajuste del modelo a los datos y la simplicidad del modelo. En términos generales, el AIC mide la cantidad relativa de información perdida al utilizar un modelo para aproximar la realidad, penalizando la complejidad y favoreciendo modelos que explican eficazmente los datos con un número mínimo de parámetros.

- Criterio de Información Bayesiana (BIC):

El Criterio de Información Bayesiana (BIC) está fundamentado en la estadística bayesiana y aborda la selección de modelos desde una perspectiva más conservadora. BIC penaliza la complejidad de manera más estricta que el AIC, favoreciendo modelos más simples. Considera tanto el ajuste del modelo como la cantidad de datos disponibles, incorporando el tamaño de la muestra en la penalización.

```
iqtree -s ASV_alin.fasta -m MF
```

Reconstrucción del árbol filogenético

Diferentes métodos se pueden emplear para la reconstrucción filogenética. Estos métodos se pueden dividir de forma grosera según el tipo de datos que utilizan en: basados en la distancia y basados en caracteres. Los métodos basados en caracteres utilizan las sustituciones entre las secuencias para determinar las relaciones filogenéticas más probables, mientras que los métodos basados en distancia primero calculan la distancia global entre todos los pares de secuencias y luego realizan la inferencia del árbol en base a dichas distancias. En esta práctica, utilizaremos Máxima Verosimilitud (MV), un método basado en caracteres. Como vimos en el teórico, MV busca el árbol y los parámetros del modelo asociados que maximizan la probabilidad de producir el conjunto de datos (alineamiento). Para reconstruir el árbol de MV, usaremos el programa IQ-TREE.

IQ-TREE, además de su función para la selección de modelos evolutivos, se presenta como una herramienta completa para llevar a cabo inferencias filogenéticas. La lista de todos los argumentos de la línea de comandos y cómo usarlos se encuentra en la sección "Ayuda" que se muestra después de ingresar el comando:

```
iqtree -h
```

Echa un vistazo a las opciones. ¿Cuales son los argumentos necesarios para la entrada de los datos?

En IQ-TREE estableceremos:

- el modelo de sustitución escogido para los datos a través de la búsqueda con ModelFinder (-m)*
- 100 réplicas de STANDARD NON-PARAMETRIC BOOTSTRAP (-b)

* Para hacer la búsqueda dle modelo y concatenar directamente con la inferencia filogenética se debe colocar -m MFP. El término MFP significa ModelFinder Plus, lo que indica a IQ-TREE que realice ModelFinder y el análisis restante utilizando el modelo seleccionado.

Al finalizar la ejecución, IQ-TREE generará varios archivos de salida, entre ellos:

- example.phy.iqtree: el archivo principal de informe que es autolegible. Deberías revisar este archivo para ver los resultados computacionales. También contiene una representación textual del árbol final.

- example.phy.treefile: el árbol de máxima verosimilitud en formato NEWICK (puede abrirlo con un editor de texto para ver la estructura del formato Newick), que puede visualizarse con programas de visualización de árboles compatibles como FigTree o iTOL.

- example.phy.log: archivo de registro de toda la ejecución (también mostrado en la pantalla).

Visualización del árbol filogenético

Para visualizar y editar el árbol utilizaremos el software iTOL (https://itol.embl.de) de Letunic I and Bork P (2006). Esta es una herramienta en línea para la visualización, anotación y gestión de árboles filogenéticos. Un flujo de trabajo típico implicaría:

- Subir el Árbol a iTOL: Accede a iTOL (https://itol.embl.de) y carga tu árbol, ya sea de forma anónima o a través de tu cuenta de iTOL.

- Explorar la Interfaz de iTOL: Familiarízate con la interfaz de usuario de iTOL. Explora las opciones de visualización, zoom y navegación para obtener una vista detallada del árbol.

- Como paso inicial en la visualización en iTOL, se recomienda enraizar el árbol en el punto medio. ¿Se observa alguna particularidad al enraizar el árbol de esta manera?

- Agregar Anotaciones: Utiliza las funciones de anotación de iTOL para agregar información relevante al árbol. ¿Qué les parece explorar la posibilidad de mapear el entorno de cada ASV y analizar si emergen patrones de agrupamiento que reflejen las condiciones ambientales?

- Exportar Figuras: Una vez que hayas realizado las anotaciones deseadas, exporta las figuras del árbol. iTOL proporciona opciones de exportación que te permiten guardar las representaciones visuales para su posterior análisis o presentación.