

TRABAJO DE PRÁCTICAS

Elvira Mayordomo y Jorge Álvarez

Marzo - Abril de 2018

1 Introducción

El trabajo de prácticas de la asignatura consistirá en que cada alumno realice por separado el trabajo que se describe en las siguientes secciones. Además de los resultados de cada parte, el alumno deberá realizar una memoria breve que contenga las decisiones tomadas, los pasos realizados (con las herramientas y argumentos aplicados) y las conclusiones extraídas al final de cada fase. El tiempo total no debería ser superior a las 2 sesiones de prácticas que quedan, a las cuales puede asistir el alumno si lo desea para consultar dudas a la profesora o realizar el trabajo en el laboratorio (la asistencia no es obligatoria).

2 Detalles del trabajo a realizar

El alumno deberá realizar el trabajo que se describe a continuación, el cual se ha dividido en 4 fases o secciones.

2.1 Obtención de la información biológica

Los alumnos deben utilizar uno de entre los siguientes tipos de secuencias a la hora de realizar el trabajo, según su número de DNI (en concreto, se usará la última cifra del DNI módulo 3, por ejemplo si el DNI termina en 7 se realizará la opción 1, si termina en 9 la opción 0, y si termina en 8 la opción 2):

0. Gen *neuraminidase* del virus de la gripe (**influenza.tar.gz**)

1. Gen *tubg1* de distintas especies (**tubg1.tar.gz**)
2. Proteína del gen mitocondrial *ND5* de distintas especies (**nd5.tar.gz**)

Los ficheros comprimidos correspondientes a cada conjunto de datos contienen un fichero *.rep* que contiene información relativa a la descarga de las secuencias, y un fichero *.gb* (formato GENBANK) y otro *.fasta* con las secuencias específicas y su información biológica.

2.2 Multialineamiento

Aunque se traten de fragmentos, las mutaciones y errores de secuenciación ya comentados en anteriores prácticas y en clase hacen que las secuencias no se puedan estudiar en muchos casos tras su descarga, lo que requiere un alineamiento previo. En esta parte el alumno deberá elegir la herramienta y parámetros que considere más adecuados para el tipo de datos seleccionados y el resultado deseado, teniendo en cuenta los siguientes criterios que tomarán parte de la evaluación de esta parte del trabajo:

1. La longitud final del alineamiento con respecto a la secuencia de mayor longitud del conjunto.
2. Adecuación de los argumentos utilizados a los datos de entrada (por ejemplo, utilizar *-addfragments* con Mafft no tendría sentido).
3. Coste temporal (algo razonable, no se pretende que el alumno tenga el ordenador varias horas trabajando para reducir en 1 ó 2 la longitud del alineamiento final).

2.3 Detección de mutaciones y errores de secuenciación

El estudio de mutaciones y cómo estas pueden afectar en menor o mayor medida al individuo (normalmente influyendo o causando alguna enfermedad genética) se puede hacer de diversas formas. Una de ellas es mediante el estudio del **índice de conservación**: si un nucleótido aparece en la misma posición en muchas secuencias (habitualmente observando distintas especies) es muy probable que una mutación en dicha posición afecte de forma mucho más negativa al individuo, que una mutación en una posición donde la variedad de nucleótidos es mayor. Además, este estudio no requiere de

información adicional, como podría ser un árbol filogenético, sino que simplemente se basa en el estudio del alineamiento generado en la fase anterior.

Para ello el alumno deberá implementar un algoritmo en dos fases ¹:

1. Calcular la frecuencia de cada nucleótido (o aminoácido) para cada posición del alineamiento, aplicando la siguiente fórmula: $f_v(i) = n_v(i)/n(i)$, donde v es cada nucleótido, i es cada posición (columna) del alineamiento, $n_v(i)$ es el número de secuencias en las que aparece el nucleótido v en la posición i , y $n(i)$ es el número de secuencias en las que en la posición i hay un nucleótido ².
2. Calcular la medida de conservación para cada posición según el método de entropía: $C(i) = \sum_{v=1}^4 f_v(i) \ln(f_v(i))$

Se valorará la eficiencia de la implementación así como el informe final con la información recopilada tras la aplicación del algoritmo. El lenguaje de programación a utilizar así como el entorno es libre, aunque se recomienda que el programa pueda ser probado en cualquier entorno y equipo (o que se incluyan en la memoria instrucciones claras y precisas de cómo ejecutarlo). El programa deberá adjuntarse a la memoria (como un fichero separado), por lo que se aconseja que el código sea legible y comprensible.

2.4 Construcción de una filogenia

Dos de las herramientas más usadas para la construcción de árboles filogenéticos son FastTree (<http://www.microbesonline.org/fasttree/>) y RAxML (<http://sco.h-its.org/exelixis/web/software/raxml/index.html>). Estas dos herramientas aplican el método de máxima verosimilitud, que se basa en modelos matemáticos (conocidos como modelos evolutivos) para inferir la distribución del conjunto de secuencias a lo largo de un árbol.

Será necesario estudiar ambas herramientas y evaluar su comportamiento y resultado al modificar el modelo evolutivo aplicado. Para poder evaluar cuán óptimo es un árbol, estas herramientas suelen devolver, además del árbol en formato NEWICK, su puntuación con el identificador *Log-likelihood*

¹Para más información sobre este método, el alumno puede consultar el artículo <http://bioinformatics.oxfordjournals.org/content/17/8/700.long>

²Recordad que el gap no es un nucleótido!

score. El alumno deberá adjuntar a la memoria el mejor árbol evolutivo obtenido junto con la herramienta y parámetros usados.

3 Entrega de la práctica

La práctica debe ser entregada hasta el viernes 27 de abril mediante envío por correo electrónico de un fichero comprimido (preferiblemente .zip) a elvira@unizar.es

Dicho fichero comprimido debe contener al menos:

- Una memoria detallada del trabajo realizado (preferiblemente en formato PDF) incluyendo tareas realizadas, pruebas y conclusiones.
- El código de los programas realizados.
- Un fichero de comandos o instrucciones detalladas para probar el código entregado.