Proyecto Fin de Carrera

Estudio y análisis de métodos de inferencia filogenética: del ADN a las proteínas

Autor:

Enrique Miguel Lozano

Bajo la dirección de:

Elvira Mayordomo Cámara
Jorge Álvarez Jarreta

Departamento de Informática e Ingeniería de Sistemas
Área de Lenguajes y Sistemas Informáticos
Escuela de Ingeniería y Arquitectura
Universidad de Zaragoza

Curso 2012/2013
Noviembre de 2012

Estudio y análisis de métodos de inferencia filogenética: del ADN a las proteínas

RESUMEN

Los principales objetivos de este proyecto fin de carrera son la traducción de ADN a proteínas, la construcción de árboles filogenéticos utilizando proteínas y su comparación con los árboles construidos directamente a partir de secuencias completas de ADN.

La filogenética es la disciplina que estudia las relaciones evolutivas entre distintos individuos o especies. El ADN mitocondrial es un tipo especial de ADN que está almacenado en unos orgánulos de la célula llamados mitocondrias. Parte del ADN mitocondrial codifica proteínas. Los árboles filogenéticos se construyen utilizando modelos matemáticos que intentan explicar la evolución real de los individuos. En el caso tratado de ADN mitocondrial, las filogenias son especialmente útiles a la hora de diagnosticar las mutaciones de un paciente como patógenas.

Para poder construir estos árboles es necesario identificar las proteínas dentro de la secuencia de ADN mitocondrial y extraer su información. Debido a la falta de homogeneidad en las bases de datos donde se encuentran almacenadas las secuencias es necesaria una primera fase de procesamiento para así poder localizar las proteínas.

La comparación de árboles filogenéticos es un tema abierto y candente en la filogenia computacional para el que no se conocen en la actualidad soluciones satisfactorias, por lo que las herramientas existentes no permiten un análisis profundo de los resultados. Futuros desarrollos en el área de comparación de filogenias serán el punto de partida de posteriores investigaciones a partir de las herramientas y resultados obtenidos en este PFC.

En este proyecto se trabaja con árboles filogenéticos construidos mediante proteínas, un tema novedoso en investigación por lo que se espera publicar los resultados en breve y que este proyecto sea el punto de partida de futuros estudios.

Durante todo el proyecto se ha tenido que dedicar mucho tiempo a la formación en temas de índole biológica. También se ha usado gran cantidad de herramientas bioinformáticas.

Se ha conseguido el objetivo de construir los árboles correspondientes a un total de 4.824 secuencias, un número considerado alto en filogenia computacional dado el gran coste computacional de los métodos empleados. Tras compararlos con los árboles de ADN disponibles se ha llegado a la conclusión que son muy distintos, hecho que puede tener varias explicaciones biológicas y que da pie a nuevas investigaciones para dar explicaciones a este suceso.

Agradecimientos

Parece tan cercano el día en que empecé la carrera, que lo puedo recordar como si fuera ayer. Nunca podré olvidar los primeros momentos, en los que todo era tan nuevo e ilusionante y en los que conocí a tanta gente que me ha acompañado en este largo viaje, algunos de los cuales puedo considerar verdaderos amigos.

También quiero acordarme de mis amigos de toda la vida, que siempre han estado ahí para cualquier cosa que he necesitado y con los que he tenido la suerte de compartir tantas experiencias a lo largo de estos años.

Por supuesto, agradecer a mis directores, Elvira y Jorge por su apoyo a lo largo del proyecto y por haberme ayudado a llegar al final de este camino.

Por último me gustaría agradecer a mi familia el haberme animado siempre. Pero por encima de todo, quiero dar las gracias a mis padres por todos los sacrificios que han hecho por mí. Sin ellos no sería quien soy y nunca habría podido llegar tan lejos.

Muchas gracias a todos.

Índice general

Índice de figuras

Índice de tablas

1. Introducción

1.1. Contexto del proyecto

El proyecto se ha realizado en el Departamento de Informática e Ingeniería de Sistemas de la Universidad de Zaragoza, dentro del ámbito de la bioinformática [1]. Se han utilizado los resultados del proyecto ZARAMIT [2] sobre ADN mitocondrial humano. En cuanto al contexto biológico del proyecto, se ha trabajado con las proteínas del ADN mitocondrial humano para la construcción de árboles de filogenias que reflejan la evolución de la especie humana. Dicha construcción es de gran importancia dado que el estudio filogenético del ADN mitocondrial permite detectar mutaciones asociadas a enfermedades raras, muchas de las cuales provocan la muerte prematura del individuo. En función de la localización de una secuencia en el árbol generado se puede conocer si las mutaciones presentes en dicha secuencia están asociadas a una enfermedad que pueda afectar al individuo [3]. Por otro lado, el estudio filogenético realizado sobre proteínas, en contraste con el usual estudio de filogenias de ADN, es un tema novedoso, bastante menos desarrollado en la literatura y de gran interés científico-técnico.

1.2. Objetivos

Una vez situado al lector en el contexto del proyecto, es el momento de explicar los objetivos que lo han guiado. La investigación ha estado centrada en la construcción de árboles filogenéticos a partir de proteínas, para lo que se ha trabajado con la secuencia de ADN mitocondrial humano.

El proyecto consta de tres objetivos fundamentales: introducción y formación, alineamiento de secuencias y obtención de proteínas, y obtención de árboles de filogenias creados a partir de proteínas y comparación con los árboles creados directamente a partir de la secuencia completa de ADN mitocondrial (o ADNmt).

El primer objetivo es parte de todo proyecto, ya que siempre es necesario obtener información a partir de la cual empezar a trabajar. Pero esta fase ha sido especialmente necesaria en este caso, dado que no hay formación en la carrera en el ámbito de la bioinformática y se ha tenido que adquirir una gran cantidad de conceptos básicos y no tan básicos al comenzar este proyecto. Aunque este objetivo no solo se ha llevado a cabo al inicio del trabajo, sino que prácticamente hasta la finalización del mismo se ha tenido que seguir recopilando información.

El segundo objetivo se centra en obtener los datos a partir de los cuales se han construido los árboles de filogenias. Aunque este objetivo se podría haber incluido dentro del siguiente, se ha considerado que era mejor separarlo dado el tiempo que ha durado y la importancia que tiene dentro del proyecto. Para obtener las proteínas de una secuencia de ADN es necesario traducir los nucleótidos de los genes que las codifican en aminoácidos, para lo que se ha

tenido que alinear las secuencias previamente, ya que, por un lado es necesario conocer la posición de los codones para realizar la traducción y por otro será necesario que las proteínas estén alineadas para la construcción de la filogenia

El último objetivo consiste en la construcción de los árboles de filogenias a partir de proteínas y en comparar su estructura con la de los árboles obtenidos directamente a partir de la secuencia completa de ADNmt. Se tomó como base la filogenia de 4.895 secuencias construida por el proyecto ZARAMIT. El ADNmt contiene 13 genes que codifican proteínas, por lo que hay que construir un árbol para cada una ellas. Tras constatar que el tiempo de construcción de cada uno de los árboles completos era demasiado elevado para poder realizarlo en este caso, se aprovechó la división por haplogrupos de ZARAMIT para trabajar con conjuntos de secuencias más pequeños y manejables temporalmente y estudiar cada uno de ellos por separado.

1.3. Metodología y herramientas

Como ya se ha comentado, se ha tenido que dedicar un especial esfuerzo y tiempo a la adquisición de conceptos biológicos con los que nunca se había trabajado. Se ha profundizado sobre todo en temas de alineamiento de secuencias, obtención de proteínas, evaluación de modelos evolutivos y construcción y comparación de árboles filogenéticos.

Para cada una de las soluciones diseñadas hacia los objetivos definidos se ha estudiado el marco tecnológico más apropiado. Una vez seleccionadas las herramientas y entornos adecuados se pasó a implementar los diseños.

A lo largo de las distintas fases del proyecto se han realizado pruebas para comprobar si los resultados eran correctos y en el caso de detectar errores para solucionarlos. Esto tuvo especial importancia cuando se estaban alineando las secuencias y obteniendo las proteínas, ya que las pruebas además de detectar errores en el diseño y la implementación permitían detectar anomalías en las secuencias obtenidas de GenBank y así descartarlas. Estas anomalías no tienen porqué significar que las secuencias no son válidas, ya que, aunque no sean adecuadas para el contexto de este proyecto, pueden ser interesantes para otras investigaciones. No obstante tiene gran utilidad comunicar estas incidencias a la comunidad científica e incluso proponer versiones curadas de los datos para su uso en otros trabajos.

Por último, es importante destacar que se ha trabajado con datos reales pertenecientes al proyecto ZARAMIT para la construcción de los árboles de filogenias, por lo que tanto los haplogrupos como las secuencias empleadas son utilizadas y aceptadas actualmente por los biólogos.

1.4. Software

Se ha trabajado con varias aplicaciones software a lo largo del proyecto. Una de ellas ha sido MUSCLE, una de las herramientas más usadas para alinear secuencias, en su versión 3.8.31 [4]. Como ayuda para elegir el mejor modelo evolutivo se utilizó ProtTest, en su versión 3.0 [5]. ProtTest es una aplicación que permite evaluar hasta 120 modelos para un alineamiento de

secuencias dado. Una vez elegido el mejor modelo se utilizó la herramienta RAxML, en su versión 7.0.3 [6], para construir los árboles de filogenias. Por último, se utilizó una de las herramientas que incluye el paquete Phylip (*Phylogeny Inference Package*), en su versión 3.69 [7], llamada TREEDIST, con la que se compararon los árboles filogenéticos construidos a partir de proteínas con los construidos a partir de la secuencia completa de ADNmt, disponibles en la página web de ZARAMIT.

La base de datos de secuencias de ADNmt humano con la que se ha trabajado a lo largo del proyecto se ha obtenido de GenBank [8], la gran base de datos de elementos relacionados con la genética, de uso muy extendido, de acceso público y a través de internet y a la que cualquier investigador puede aportar información. Esto provoca que los datos no estén normalizados por lo que se tuvieron que alinear todas las secuencias con la de referencia para que los genes que codifican las proteínas se encontraran en la misma posición en todas las secuencias. De ZARAMIT se obtuvo la división en haplogrupos de un conjunto de secuencias y sus árboles filogenéticos que se han comparado con los árboles obtenidos a partir de las proteínas

1.5. Entorno tecnológico

Durante la realización del proyecto se me ha dado acceso a uno de los laboratorios de investigación del Grupo de Ingeniería de Sistemas de Eventos Discretos (GISED), el L1.03b, donde he tenido a mi disposición el ordenador Duero, que tiene un procesador *Intel® Core™2 Duo* E6750 a 2.66GHz y 8GB de memoria RAM.

Adicionalmente, se me ha dado acceso al ordenador Amboto (cuyas especificaciones son las mismas que Duero) para agilizar el cálculo de los árboles de proteínas.

1.6. Estructura de la memoria

Para finalizar este primer capítulo introductorio se va a comentar brevemente el contenido del resto de capítulos y anexos que componen la memoria.

Se ha creído necesario dedicar el segundo capítulo a un breve glosario biológico en el que se incluyen conceptos y definiciones que se han considerado necesarios para la comprensión del componente biológico del proyecto.

En el tercer capítulo se explica la necesidad de realizar un procesamiento previo de los datos (secuencias de ADNmt) y en qué ha consistido.

El siguiente capítulo trata sobre los procesos de transcripción y traducción que se han efectuado sobre las secuencias una vez procesadas, para así obtener las proteínas de las secuencias.

En el quinto capítulo se detalla el trabajo realizado para la obtención de los árboles filogenéticos usando las proteínas calculadas anteriormente y la comparación de dichos árboles con los obtenidos directamente del ADN.

En el último capítulo antes de los anexos se recogen las conclusiones obtenidas del desarrollo del proyecto.

El primer apéndice incluye un diagrama de Gantt en el que se puede observar el tiempo dedicado a cada parte del proyecto.

En el segundo apéndice se han explicado más detalladamente los fundamentos biológicos que sustentan el proyecto.

En el último anexo se incluye el estudio que se ha hecho sobre la herramienta ProtTest.

Por último, se adjunta la bibliografía utilizada durante la realización del proyecto y la redacción de esta memoria.

2. Glosario biológico

Dado que a lo largo de la memoria se utilizan conceptos biológicos a los que el lector puede no estar acostumbrado, se ha considerado adecuada la inclusión del presente capítulo para explicar los conceptos más importantes para la comprensión del documento. Para obtener información más detallada se puede consultar el apéndice B.

ADN: Acrónimo de ácido desoxirribonucleico. Forma parte de la célula y almacena la información genética del individuo. Es el responsable de la transmisión de la información hereditaria y del correcto funcionamiento de la célula (lo que incluye contener información para la formación de proteínas).

ADN mitocondrial (ADNmt): Es un tipo especial de ADN que poseen unos orgánulos llamados mitocondrias que forman parte de las células. A diferencia del ADN celular, el ADN mitocondrial se hereda solo por parte materna, su tasa de mutación es muy elevada y además su longitud es mucho menor. Por sus características es muy utilizado en el estudio de filogenias.

ARN: Acrónimo de ácido ribonucleico. Es el encargado de que la síntesis de proteínas se lleve a cabo. Hay diversos tipos, entre los que destacan el ARN mensajero, el ARN de transferencia y el ARN ribosómico.

Gen: Es un conjunto ordenado de nucleótidos de una secuencia de ADN que contiene la información necesaria para sintetizar una macromolécula que desempeña una función celular específica. Dentro de estas macromoléculas se encuentran las proteínas.

Proteína: Macromolécula formada por aminoácidos. Es fundamental para la vida, ya que puede desempeñar una gran cantidad de funciones básicas para el correcto funcionamiento del organismo (estructural, inmunológica, enzimática...). El ADNmt humano contiene 37 genes en su secuencia, de los cuales 13 codifican proteínas.

Filogenética: Se puede considerar la ciencia que estudia las relaciones evolutivas entre distintos individuos y los clasifica en función de los resultados observados.

Mutación genética: Cuando dentro de una secuencia de ADN se producen alteraciones en los nucleótidos se denomina mutación genética. Estos cambios pueden provocar cambios en los aminoácidos de las proteínas, lo que puede desencadenar consecuencias graves.

Modelo evolutivo: Es un modelo matemático que pretende representar las probabilidades de mutación en el material genético de un conjunto de individuos u organismos, es decir, su evolución. Se expresan normalmente en matrices que representan cadenas de Markov.

Haplotipo: Conjunto de cambios en una base de la secuencia de ADN que están estadísticamente relacionados. A estos cambios se les suele llamar polimorfismos de nucleótido simple o SNP.

Haplogrupo: Un haplogrupo lo forman un conjunto de haplotipos con características comunes. En genética humana hay dos principales líneas de estudio de haplogrupos, la del cromosoma Y (estudio del linaje por parte del padre) y la del ADNmt (estudio del linaje materno). Este proyecto se basa en esta última.

3. Preprocesamiento de las secuencias

En este capítulo se va a exponer el tratamiento previo que se le ha dado a las secuencias de ADNmt descargadas de GenBank y se va a explicar porqué este preprocesamiento es necesario.

3.1. Estado del arte

Dado que el objetivo es construir árboles filogenéticos a partir de las proteínas de las secuencias de ADNmt, lo primero que hay que preguntarse es cómo obtener las proteínas de una secuencia.

Cada secuencia de ADNmt contiene la información para crear 13 proteínas distintas: ATP6, ATP8, ND1, ND2, ND3, ND4, ND4L, ND5, ND6, CO1, CO2, CO3 y CytB. En la página web de MITOMAP [9] se puede consultar la posición inicial y final de estas proteínas en la secuencia de referencia (rCRS [10] o *revised Cambridge Reference Sequence*), así como los nucleótidos (o aminoácidos) que la componen.

La secuencia CRS [11] fue publicada en 1981 y fue uno de los primeros pasos del proyecto del genoma humano. La rCRS, como su nombre indica, es su versión revisada. Está compuesta por 16.569 nucleótidos.

Dado que se conoce la posición de los genes que codifican proteínas en la secuencia rCRS, si se consigue que dicha posición en el resto de secuencias coincida con la de la rCRS se podrán obtener las proteínas de todas las secuencias de una forma sencilla. A este proceso se le conoce como alineamiento de secuencias.

Un alineamiento consiste en un análisis y modificación de un conjunto de secuencias con el fin de conseguir la mayor cantidad posible de nucleótidos iguales en la misma posición, para lo que puede ser necesario incluir huecos (habitualmente llamados *gaps*) en alguna de las secuencias. La elección del mejor alineamiento para un conjunto de secuencias dado es un tema que se sigue estudiando en el campo de la bioinformática por lo que para este proyecto se ha decidido utilizar una de las herramientas más utilizadas por los expertos para realizar alineamientos llamada MUSCLE [4], en su versión 3.8.31.

Se va a trabajar con las secuencias de ADNmt humano de entre 16.000 y 17.000 nucleótidos almacenadas en GenBank, lo que supone un total de 14.915 secuencias. Una vez alineadas todas las secuencias con la de referencia es de esperar que quede fija la posición de los genes que codifican las proteínas en todas ellas. En el resto del capítulo se va a comentar cómo se ha efectuado el alineamiento y los resultados obtenidos una vez finalizado.

3.2. Alineamiento de secuencias

Como se ha comentado anteriormente, se ha usado la herramienta MUSCLE para alinear las secuencias con la rCRS. Para realizar cada alineamiento se tiene que lanzar una ejecución de MUSCLE pasándole como entrada el nombre de los ficheros que contienen la secuencia rCRS y la secuencia a alinear; esto implica que hay que lanzar una ejecución por cada una de las 14.915 secuencias. Dado el excesivo coste temporal de lanzar estas ejecuciones manualmente se ha implementado un programa que llame a MUSCLE y realice un conjunto de acciones adicionales que se detallan a continuación.

Cuando MUSCLE alinea cada par de secuencias genera un fichero con extensión *.afa* en el que aparecen las dos secuencias alineadas una a continuación de la otra. El objetivo es tener un fichero en el que solo esté la secuencia alineada y procesada, por lo que el programa va a eliminar las posiciones de la secuencia a alinear que se correspondan con los *gaps* introducidos en la rCRS (a excepción del *gap* de la posición 3.107 que se mantiene por convenio) y generar el nuevo fichero con la secuencia alineada y procesada.

Por último, es importante detectar en la secuencia si se han introducido *gaps* en algún punto de una región correspondiente a un gen, ya que este evento puede estar asociado a anomalías que sería necesario estudiar. Por ello, el programa, antes de eliminar una posición, comprobará si está dentro de un gen, y si lo está, escribirá en un fichero especial la secuencia y la proteína que se codifica en el gen en el que se ha encontrado esta situación.

Cabe destacar que el alineamiento de secuencias es un proceso con un coste temporal bastante elevado. El alineamiento de las 14.915 secuencias tardó en ejecutarse aproximadamente una semana.

3.3. Análisis de resultados

Una vez terminada la ejecución del programa se han analizado dos cosas: si el alineamiento ha permitido identificar la posición de los genes en las secuencias y si, como se ha comentado en la sección anterior, han aparecido *gaps* en los genes de las secuencias alineadas.

Para comprobar si los 13 genes que codifican las proteínas empiezan en la misma posición en todas las secuencias se ha desarrollado un pequeño programa que escribe en un fichero los nombres de los archivos que contengan secuencias en las que al menos un gen no comience en la misma posición que su correspondiente en la rCRS. Al ejecutarlo se ha creado un fichero que contiene 4.808 secuencias. Este número es bastante elevado, puesto que es prácticamente un tercio del número total de secuencias.

Una vez analizados los resultados más a fondo se llega a la conclusión de que los genes que codifican las proteínas empiezan siempre entre 20 posiciones antes y 20 posiciones después que en la rCRS. Esto habrá que tenerlo en cuenta a la hora de obtener los aminoácidos que componen las proteínas (como se verá en el capítulo 4). Un estudio más profundo de las

causas por las que es necesario aplicar este umbral resultaría interesante en trabajos futuros, quedando fuera de los objetivos y alcance de este proyecto.

Como se ha comentado anteriormente, la aparición de *gaps* en los genes que codifican las proteínas puede estar asociado a anomalías en la secuencia original. Al analizar el fichero especial que se ha generado se puede observar que éste contiene 6 secuencias. Dado que representan un 0,04% del número total de secuencias recogidas y que un análisis más profundo de los motivos se aleja de los objetivos del proyecto, se ha optado por descartar las secuencias. En la tabla 3.1 se pueden observar los identificadores de GenBank de las secuencias descartadas así como la proteína que codifica el gen en el que se ha encontrado el *gap*.

Identificador	**Proteína afectada**
AB055387	CO1
JQ705682	CO2
EF184610	CytB
EF184640	ATP8
JF742208	ND2
EF660930	ND3

Tabla 3.1: Secuencias que contienen *gaps* en algún gen que codifique proteínas.

4. Transcripción y traducción

En este capítulo se va a explicar cómo una vez alineadas las secuencias se han obtenido las proteínas por medio de los procesos de transcripción y traducción.

4.1. Estado del arte

La transcripción y la traducción son dos procesos que forman parte de la expresión génica, es decir, son los mecanismos por los cuales una célula es capaz de obtener las proteínas codificadas en los distintos genes del ADN.

No se ha considerado necesaria la búsqueda de implementaciones ya existentes de estos procesos. El motivo es doble: por un lado, son mecanismos de baja complejidad cuyo objetivo es la traducción de nucleótidos a aminoácidos, y por otro, esto permite ajustar más adecuadamente el programa a las necesidades del proyecto, como la búsqueda del inicio de los genes entre 20 posiciones antes y 20 después de la posición de inicio en la rCRS.

En los siguientes apartados se resumen los procesos de transcripción y traducción y los aspectos más importantes de su implementación.

4.2. Transcripción

Básicamente, el proceso de transcripción consiste en usar la información contenida en la secuencia de ADN para sintetizar un tipo de ARN, el ARN mensajero, gracias a una enzima llamada ARN polimerasa.

Si se tienen en cuenta las consecuencias de este proceso sobre las secuencias de ADNmt con las que estamos trabajando, el programa ha de realizar dos tareas: la primera consiste en transformar la timina (“T”) en uracilo (“U”) en toda la secuencia. La segunda solo afecta al trozo de secuencia correspondiente al gen que codifica la proteína ND6. Este trozo de secuencia debe ser invertido, es decir, el primer nucleótido pasa a ser el último, el segundo el penúltimo y así sucesivamente; además, cada nucleótido debe cambiarse por su complementario teniendo en cuenta que el complementario del uracilo (“U”) es la adenina (“A”) y el de la citosina (“C”) es la guanina (“G”).

4.3. Traducción

Durante el proceso de traducción, el ARN mensajero sintetizado durante la transcripción da lugar a las proteínas. El encargado de este proceso es el orgánulo llamado ribosoma.

En este caso, la traducción provoca que cada conjunto de tres nucleótidos, denominado codón, se transforme en un aminoácido. Por ello, determinar la posición de inicio de un gen es crítico dado que la modificación en la posición de los nucleótidos dentro del codón puede generar un aminoácido distinto al que debería obtenerse.

Como se ha visto en el capítulo anterior, se conoce la posición exacta de los genes en aproximadamente dos tercios del total de las secuencias, pero para el tercio restante solamente se sabe que pueden estar a una distancia de 20 posiciones de la situación conocida. Por lo tanto, para cada secuencia hay que comprobar si una vez traducida la terna de nucleótidos, esta coincide con el inicio de la proteína (que conocemos gracias a MITOMAP) y, de no hacerlo, se ha de buscar este inicio 20 posiciones hacia atrás y 20 posiciones hacia delante.

Cabe comentar en este apartado una decisión que se ha tomado en lo relativo a la codificación de los aminoácidos. Al coger los nucleótidos en grupos de 3, hay 64 posibles codones, pero varios codones codifican el mismo aminoácido. En total hay 22 aminoácidos distintos de los cuales solo 20 aparecen en el ser humano. Se ha respetado la representación usada para los aminoácidos que forman parte de las proteínas, pero dado que los caracteres terminadores no tienen representación adecuada para este contexto, se ha decidido representarlos mediante una “X”. Se ha tomado esta decisión porque se ha observado que hay ocasiones en las que se producen mutaciones en estos caracteres terminadores, lo que provoca que la secuencia tenga un aminoácido más de lo normal si no codificamos el carácter terminador. Al añadir esta “X” se consigue que todas las secuencias tengan la misma longitud, condición necesaria para la generación de árboles de filogenias que se tratará en el siguiente capítulo.

Además, se han detectado secuencias en GenBank a las que les falta algún fragmento, pero que están almacenadas por el valor que aportan las partes que sí están completas. Para indicar esto se rellenan las posiciones desconocidas con el carácter “N” (que no representa ningún nucleótido). Para respetar la longitud de la proteína, siempre que al menos un elemento de la terna de nucleótidos a traducir sea una “N”, se representará el codón con el carácter “X”.

4.4. Resultados

Al comprobar si las proteínas se habían generado correctamente se vio que en 99 secuencias una o varias de sus proteínas no empezaban en la misma posición que la rCRS. Al investigar más detalladamente lo que pasaba se pudo comprobar que estas proteínas empezaban con una (o varias) “X”, por lo que la anomalía no era tal, sino que era producida por las secuencias incompletas.

Para estudiar el comportamiento temporal del programa se ha calculado el tiempo para los casos mejor y peor. Hay que tener en cuenta que el caso mejor se da cuando no hay que desplazarse por la secuencia durante la traducción, lo que se da en dos de cada tres secuencias. Para el cálculo del tiempo en el caso mejor se ha utilizado la secuencia HQ012252, para la que el programa tardó 0,041751 segundos. Para el caso peor se usó la secuencia EU092740, ya que en 11 de sus 13 proteínas hay que efectuar búsquedas por la secuencia y además 8 de estás búsquedas hay que hacerlas a la derecha de la posición inicial (y el programa busca primero por la izquierda). El programa tardó en este caso 0,088041 segundos. Como se puede ver el tiempo de

ejecución del programa por cada secuencia es muy bajo, ya que en ningún caso llega a una décima de segundo.

5. Árboles de filogenias

A continuación se va a detallar el trabajo realizado para la construcción de los árboles filogenéticos utilizando las proteínas que se han obtenido anteriormente. Por último se va a comentar la comparación de los resultados con los árboles obtenidos directamente a partir de las secuencias de ADN mitocondrial completas del proyecto ZARAMIT [2].

5.1. Estado del arte

Una vez cumplido el segundo objetivo propuesto, se van a utilizar los resultados obtenidos para la construcción de los árboles filogenéticos y su posterior comparación con los árboles obtenidos directamente a partir de las secuencias completas de ADNmt.

Dado que el objetivo final requería la comparación de los resultados con árboles reales, se decidió utilizar el árbol creado por el proyecto ZARAMIT a partir de 4.895 secuencias de ADNmt humano. Como se explica más adelante, el tiempo de computación para la construcción de árboles con tantas secuencias es muy elevado y se utilizó la división de haplogrupos del árbol de ZARAMIT para poder trabajar con tiempos de computación más manejables.

El proceso de construcción de árboles de filogenias es muy complejo y ha sido, y continúa siendo, un tema muy importante de estudio en el ámbito de la bioinformática [12–14], por lo que se ha optado por utilizar la herramienta más usada para la construcción de árboles filogenéticos, RAxML, en su versión 7.0.3 [6]. Esta herramienta construye los árboles en formato Newick [15], el mismo formato que se usa en el proyecto ZARAMIT, lo que será de gran ayuda en la fase de comparación de árboles.

Una vez obtenidos los árboles era necesaria la elección de una herramienta para poder compararlos. Se eligió una de las herramientas del paquete Phylip, en su versión 3.69, llamada TEEDIST [16]. Esta aplicación permite comparar la estructura de dos o más árboles filogenéticos en formato Newick, por lo que se adecúa al objetivo del proyecto perfectamente. En la última parte del capítulo se analizan los resultados obtenidos de esta comparación.

5.2. Construcción de los árboles filogenéticos

Como se ha comentado en el punto anterior, el coste temporal de la construcción de los árboles filogenéticos es muy elevado. Para que el lector se haga una idea, generar el árbol completo para la proteína ATP6, compuesta por 227 aminoácidos, tiene un coste de 107 horas, es decir, casi 4 días y medio, siendo esta la quinta proteína con menor número de aminoácidos y aproximadamente tres veces más corta que la proteína de mayor longitud. Dado que el coste temporal superaba los objetivos del proyecto, se decidió aprovechar la división en haplogrupos que había sido efectuada por el proyecto ZARAMIT para poder tratar con árboles más manejables desde el punto de

vista del coste temporal. Las 4.895 secuencias con las que trabaja ZARAMIT se dividen en 26 haplogrupos de diversos tamaños.

Una vez estudiado el funcionamiento de la herramienta que se ha utilizado para generar los árboles, RAxML, se ha visto que los ficheros de entrada tienen que tener formato Phylip, por lo que se ha desarrollado un pequeño programa que traduce los ficheros en los que están guardadas las proteínas a este formato. Adicionalmente, se han creado para cada haplogrupo 13 ficheros, uno por proteína, con las secuencias que pertenecen a cada uno de ellos. Se detectó que la secuencia con identificador FJ770972, incluida en el haplogrupo M, no tiene información de la proteína ATP8, por lo que se ha eliminado del fichero correspondiente.

Una vez generados los ficheros, hay que lanzar una ejecución de RAxML por cada uno de los 13 ficheros que contienen las proteínas de los 26 haplogrupos, lo que hace un total de 338 ejecuciones. Para evitar tener que lanzarlas manualmente se ha creado un programa que automatiza esta tarea y mide el tiempo de ejecución de cada uno de los ficheros. Se ha utilizado el modelo evolutivo MtMam ya que es el modelo que resultó más apropiado después de analizar los distintos modelos utilizando la herramienta ProtTest. Este estudio se puede consultar en el anexo C de la memoria. En la figura 5.1 se puede observar una gráfica con los tiempos de ejecución.

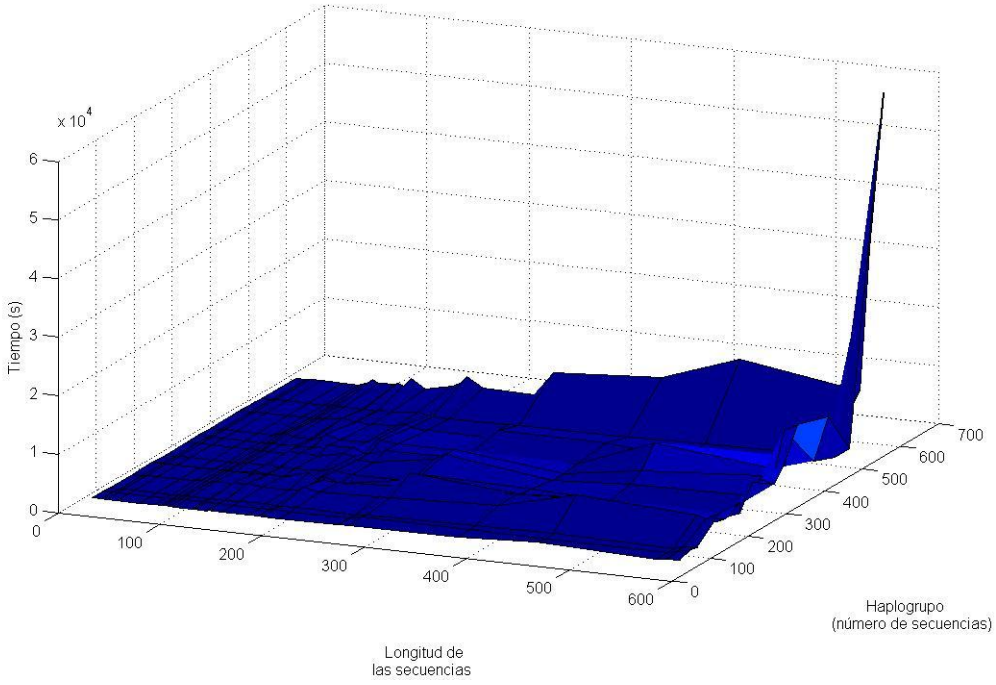

Figura 5.1: Gráfica de estudio de la evolución del coste temporal de RAxML según incrementa el número de secuencias y su longitud.

5.3. Comparación de árboles

Una vez se han generado los árboles se ha pasado a comparar sus estructuras. Con este fin se ha utilizado TREEDIST, una de las herramientas del paquete Phylip. TREEDIST posee dos métricas distintas para la comparativa, pero se ha decidido utilizar la llamada *Symmetric Difference* [17], que es más precisa con los datos obtenidos en etapas anteriores. La definición de *Symmetric Difference* entre dos árboles es el número total de cambios que hay que hacer en el primero de los ellos para poder obtener el segundo.

Teniendo en cuenta que TREEDIST sólo compara dos árboles si tienen exactamente las mismas secuencias en las hojas, se ha tenido que extraer cada uno de los subárboles de los haplogrupos del árbol de ZARAMIT, ya que en este caso el árbol está formado por las 4.895 secuencias. Para ello se ha desarrollado un programa que para cada haplogrupo crea un fichero en el que se han eliminado del árbol de ZARAMIT las secuencias que no pertenecen a dicho haplogrupo, evitando modificar en modo alguno la estructura del árbol.

Con esto se ha conseguido un fichero por cada uno de los haplogrupos, en el que está almacenado su árbol filogénico. El último paso ha sido ejecutar TREEDIST para comparar estos archivos con los que contienen los árboles de filogenias que se han generado por cada una de las 13 proteínas para cada haplogrupo.

5.4. Análisis de los resultados

Una vez evaluados los árboles se ha utilizado la información proporcionada por TREEDIST de la métrica *Symmetric Difference* para analizar los resultados. Antes de continuar, aclarar que tanto en el manual de TREEDIST como en la bibliografía sobre la métrica utilizada no se trata en ningún momento la evaluación de los resultados, dejando al investigador toda responsabilidad sobre su interpretación.

El número de cambios también se puede interpretar como el número de ramas únicas entre los dos árboles, por lo que obtener un resultado impar indica que al menos uno de los árboles no es binario. El total de ramas únicas posibles es de *2*n-6*, según la documentación de TREEDIST, donde *n* es el número de secuencias situadas en las hojas del árbol. Como se conocen los cambios entre cada par de árboles es posible calcular con estos datos el porcentaje de similitud entre ambos. Esta va a ser la medida que vamos a usar para analizar los datos. En la tabla 5.1 se pueden consultar los porcentajes obtenidos para cada uno de los haplogrupos en cada una de las proteínas. Como se puede observar, la tónica es que los árboles no se parezcan demasiado.

Estas son todas las conclusiones a las que se puede llegar con las métricas y herramientas actuales. Como se ha comentado en la memoria, la investigación sobre la construcción de árboles de proteínas es bastante escasa. Hasta ahora la investigación se ha centrado en construir árboles para estudiar proteínas específicas, pero no se ha profundizado en el estudio de las proteínas de las secuencias de ADN (o ADNmt, como en este proyecto). Como se ha mostrado, los árboles de ADNmt distan mucho de los árboles construidos

directamente a partir de las secuencias de ADN, por lo que podría ser interesante investigar la causa de este suceso.

Por último destacar la dificultad de encontrar investigaciones sobre la comparación de árboles filogenéticos. La métrica utilizada fue desarrollada en 1981 por Robinson y Foulds [18] y es la única de la que se ha encontrado la suficiente documentación y herramientas para poder aplicarla, además de utilizar como entrada los árboles en el formato utilizado anteriormente. Cabe la posibilidad de que otras métricas de las que no existe implementación aportase resultados más concretos, pero eso queda fuera de los objetivos de este proyecto.

Haplogrupo	ATP6	ATP8	ND1	ND2	ND3	ND4	ND4L	ND5	ND6	CO1	CO2	CO3	CytB
P	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	50,00%	0,00%	25,00%	0,00%	0,00%	0,00%
Y	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	7,14%	28,57%	7,14%	7,14%	0,00%	7,14%
Q	5,56%	0,00%	0,00%	11,11%	0,00%	11,11%	0,00%	5,56%	5,56%	0,00%	0,00%	0,00%	16,67%
Z	0,00%	0,00%	0,00%	3,45%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%
E	6,00%	0,00%	0,00%	2,00%	0,00%	0,00%	0,00%	0,00%	2,00%	2,00%	0,00%	0,00%	0,00%
W	0,00%	0,00%	0,00%	1,54%	1,54%	0,00%	0,00%	0,00%	0,00%	1,54%	0,00%	1,54%	0,00%
X	5,56%	1,39%	1,39%	0,00%	0,00%	1,39%	0,00%	8,33%	1,39%	1,39%	1,39%	0,00%	1,39%
HV	5,26%	2,63%	2,63%	3,95%	2,63%	3,95%	2,63%	9,21%	5,26%	2,63%	2,63%	3,95%	3,95%
F	1,27%	0,00%	1,27%	2,53%	1,27%	0,00%	0,00%	6,33%	3,80%	1,27%	0,00%	0,00%	2,53%
V	5,29%	2,94%	2,94%	7,65%	2,94%	2,94%	2,94%	5,29%	6,47%	5,29%	2,94%	2,94%	4,12%
G	0,98%	0,00%	0,98%	0,98%	0,98%	0,00%	0,00%	3,92%	0,00%	0,00%	0,00%	2,94%	5,88%
L1	0,00%	0,96%	3,85%	3,85%	2,88%	0,00%	0,00%	5,77%	0,00%	1,92%	2,88%	3,85%	0,96%
C	0,00%	0,00%	1,71%	0,00%	0,00%	0,85%	0,00%	0,85%	3,42%	0,00%	1,71%	0,85%	5,13%
L0	3,42%	4,27%	7,69%	2,56%	1,71%	1,71%	0,00%	11,11%	2,56%	5,13%	0,00%	0,85%	5,98%
T	3,05%	3,05%	3,05%	3,82%	3,82%	3,05%	3,82%	5,34%	3,05%	4,58%	6,11%	3,82%	3,05%
J	3,18%	1,91%	1,91%	1,27%	1,91%	2,55%	1,27%	2,55%	3,18%	1,91%	1,27%	1,27%	3,82%
L2	1,50%	1,50%	2,10%	1,50%	0,90%	0,90%	0,30%	5,09%	1,50%	1,50%	0,90%	1,50%	2,69%
N	4,09%	1,75%	2,34%	3,51%	1,17%	1,75%	1,17%	4,68%	4,68%	1,17%	1,75%	1,75%	3,51%
K	1,87%	1,87%	2,41%	2,41%	2,41%	2,41%	2,41%	3,48%	2,94%	2,41%	1,87%	4,01%	4,01%
A	3,27%	1,22%	0,41%	0,41%	0,41%	1,22%	0,41%	1,63%	1,22%	0,41%	0,82%	2,45%	1,63%
L3	3,32%	0,59%	0,59%	1,37%	0,98%	1,76%	0,59%	6,84%	2,15%	0,98%	0,98%	1,37%	2,54%
U	4,01%	0,97%	0,69%	2,07%	0,14%	1,24%	0,14%	1,80%	0,69%	0,41%	0,69%	1,52%	0,97%
R	4,11%	1,60%	3,20%	2,97%	1,37%	2,97%	1,14%	8,90%	2,97%	2,05%	1,37%	1,60%	3,65%
H	6,47%	5,94%	6,29%	6,65%	5,94%	6,29%	5,94%	6,65%	6,12%	6,29%	6,47%	6,29%	6,65%
D	2,65%	1,42%	2,48%	1,77%	1,24%	1,77%	1,24%	2,65%	2,48%	1,24%	1,24%	1,77%	2,30%
M	3,02%	2,85%	2,33%	1,99%	1,99%	2,68%	1,64%	4,23%	2,16%	2,16%	1,99%	1,99%	4,40%

Tabla 5.1: Porcentaje de similitudes de los árboles filogénicos construidos a partir de proteínas con los de ZARAMIT por cada haplogrupo.

6. Conclusiones

6.1. Trabajo realizado

La construcción de árboles filogenéticos a partir de proteínas, en este caso las del ADN mitocondrial humano, al igual que la comparación de árboles de filogenias son temas poco estudiados. El trabajo realizado inicia un estudio en este sentido y propone nuevas vías de investigación. En cuanto a los objetivos planteados inicialmente, se considera que se han cumplido de forma satisfactoria, obteniendo una filogenia de proteínas basada en 4.824 secuencias, un número muy alto en filogenia computacional.

Como era de esperar, a lo largo del proyecto surgieron problemas que tuvieron que solucionarse. El primero surgió con el alineamiento de secuencias, ya que se esperaba que se iba a conseguir que las proteínas de todas las secuencias empezaran y terminaran en la misma posición, pero la herramienta MUSCLE no lo consiguió y hubo que realizar modificaciones en la transcripción y la traducción para poder encontrar las proteínas en aproximadamente un tercio de las secuencias. Otro problema surgió al comparar los árboles, ya que el árbol existente (el de ZARAMIT) no estaba dividido en haplogrupos y no se encontró ninguna herramienta que permitiera obtener los subárboles que se necesitaban, por lo que se tuvo que desarrollar un programa que lo hiciera, teniendo mucho cuidado de no alterar la estructura del árbol.

También surgieron problemas de índole temporal, ya que al manejar tal cantidad de datos (unas 14.000 secuencias en la fase de transcripción y traducción y casi 5.000 en la fase de construcción de filogenias) cualquier operación tenía un coste muy elevado. Esto fue especialmente relevante en la construcción de los árboles filogenéticos, tarea para la que, a pesar de usar una división en haplogrupos, hizo falta la utilización de procesadores adicionales.

Se encuentra en fase de redacción un artículo de investigación que contiene los resultados de este PFC y será sometido a un congreso internacional de bioinformática durante el mes de diciembre.

6.2. Con vistas al futuro

Como se ha ido comentando a lo largo de la memoria, la investigación sobre árboles de filogenias obtenidos a partir de proteínas es muy escasa y está reservada al estudio de proteínas muy específicas. Con este proyecto, se ha intentado establecer una base para futuras investigaciones que vayan en este sentido. Dadas las limitaciones de un proyecto fin de carrera no se ha podido profundizar en aspectos interesantes, como podría haber sido la construcción de árboles más grandes, en vez de tener que dividirlos por haplogrupos o el estudio de distintas métricas más recientes y menos desarrolladas para la comparación de árboles filogenéticos, con las que quizás se hubiese podido obtener más información de la que se ha conseguido.

Por supuesto, hay que mencionar también el alineamiento de secuencias, ya que como se ha visto, la extendida herramienta MUSCLE no es capaz de alinear las proteínas correctamente. Este es un tema que si que se está investigando activamente en la actualidad y quizás podría ser interesante plantearse priorizar el alineamiento de las zonas de proteínas sobre el resto si se va a trabajar sólo con ellas.

6.3. De lo profesional a lo personal

Profesionalmente debo decir que ha sido una experiencia muy grata trabajar en un proyecto de investigación en este ámbito. Gracias a este trabajo he podido adentrarme en el mundo de la biología, que siempre me había generado curiosidad, y he descubierto lo que es realizar un proyecto de investigación, con su continuo proceso de aprendizaje y superación de dificultades, que te ayuda a aprender y mejorar cada día. También estoy contento de haber desarrollado un trabajo que pueda servir como base para futuras investigaciones. Y por supuesto, quiero agradecer a mis directores por haberme dado la oportunidad de trabajar en este proyecto y por haberme apoyado y ayudado a lo largo de todo el camino. Y ya entrando en lo personal, debo decir que he disfrutado mucho realizando este proyecto y no descarto dedicarme en un futuro a trabajos relacionados con la bioinformática o con la investigación.

Bibliografía

[1] A. Polanski y M. Kimmel, *Bioinformatics*, 1.ª ed. Springer, 2007.

[2] R. Blanco y E. Mayordomo, «ZARAMIT: A System for the Evolutionary Study of Human Mitochondrial DNA», in *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living*, S. Omatu, M. P. Rocha, J. Bravo, F. Fernández, E. Corchado, A. Bustillo, y J. M. Corchado, Eds. Springer Berlin Heidelberg, 2009, pp. 1139–1142.

[3] E. Ruiz-Pesini, D. Mishmar, M. Brandon, V. Procaccio, y D. C. Wallace, «Effects of Purifying and Adaptive Selection on Regional Variation in Human mtDNA», *Science*, vol. 303, nº. 5655, pp. 223–226, sep. 2004.

[4] R. C. Edgar, «MUSCLE: multiple sequence alignment with high accuracy and high throughput», *Nucl. Acids Res.*, vol. 32, nº. 5, pp. 1792–1797, ene. 2004.

[5] D. Darriba, G. L. Taboada, R. Doallo, y D. Posada, «ProtTest 3: fast selection of best-fit models of protein evolution», *Bioinformatics*, feb. 2011.

[6] A. Stamatakis, T. Ludwig, y H. Meier, «RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees», *Bioinformatics*, vol. 21, nº. 4, pp. 456–463, feb. 2005.

[7] J. D. Retief, «Phylogenetic Analysis Using PHYLIP», in *Bioinformatics Methods and Protocols*, S. Misener y S. A. Krawetz, Eds. Humana Press, 1999, pp. 243–258.

[8] D. A. Benson, M. S. Boguski, D. J. Lipman, y J. Ostell, «GenBank», *Nucl. Acids Res.*, vol. 25, nº. 1, pp. 1–6, ene. 1997.

[9] M. C. Brandon, «MITOMAP: a human mitochondrial genome database--2004 update», *Nucleic Acids Research*, vol. 33, nº. Database issue, pp. D611–D613, dic. 2004.

[10] R. M. Andrews, I. Kubacka, P. F. Chinnery, R. N. Lightowlers, D. M. Turnbull, y N. Howell, «Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA», *Nature Genetics*, vol. 23, nº. 2, pp. 147–147, 1999.

[11] S. Anderson, A. T. Bankier, B. G. Barrell, M. H. L. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. H. Smith, R. Staden, y I. G. Young, «Sequence and organization of the human mitochondrial genome», *, Published online: 09 April 1981; | doi:10.1038/290457a0*, vol. 290, nº. 5806, pp. 457–465, abr. 1981.

[12] Jorge Álvarez, «Análisis teórico-práctico de métodos de inferencia filogenética basados en selección de modelos y métodos de superárboles». Centro Politécnico Superior, Universidad de Zaragoza, 2010.

[13] S. Li, D. K. Pearl, y H. Doss, «Phylogenetic Tree Construction Using Markov Chain Monte Carlo», *Journal of the American Statistical Association*, vol. 95, nº. 450, pp. 493–508, jun. 2000.

[14] G. J. Olsen, H. Matsuda, R. Hagstrom, y R. Overbeek, «fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood», *Comput Appl Biosci*, vol. 10, nº. 1, pp. 41–48, ene. 1994.

[15] «http://evolution.genetics.washington.edu/phylip/newicktree.html. Web site donde se explica el formato Newick para árboles filologenéticos.» .

[16] «http://evolution.genetics.washington.edu/phylip/doc/treedist.html. Web site donde se explica la herramienta Treedis para la evaluación de árboles filologenéticos.» .

[17] B. L. Cantarel, H. G. Morrison, y W. Pearson, «Exploring the Relationship between Sequence Similarity and Accurate Phylogenetic Trees», *Mol Biol Evol*, vol. 23, nº. 11, pp. 2090–2100, nov. 2006.

[18] D. F. Robinson y L. R. Foulds, «Comparison of phylogenetic trees», *Mathematical Biosciences*, vol. 53, nº. 1–2, pp. 131–147, feb. 1981.

[19] A. Salas, V. Lareu, F. Calafell, J. Bertranpetit, y Á. Carracedo, «mtDNA hypervariable region II (HVII) sequences in human evolution studies», *European Journal of Human Genetics*, vol. 8, nº. 12, pp. 964–974, dic. 2000.

[20] P. Soares, L. Ermini, N. Thomson, M. Mormina, T. Rito, A. Röhl, A. Salas, S. Oppenheimer, V. Macaulay, y M. B. Richards, «Correcting for Purifying Selection: An Improved Human Mitochondrial Molecular Clock», *Am J Hum Genet*, vol. 84, nº. 6, pp. 740–759, jun. 2009.

[21] M. Patricio, F. Abascal, R. Zardoya, y D. Posada, «Accurate Selection of Models of Protein Evolution», in *Advances in Bioinformatics*, vol. 74, M. P. Rocha, F. F. Riverola, H. Shatkay, y J. M. Corchado, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 117–121.

[22] H. Piontkivska, «Efficiencies of maximum likelihood methods of phylogenetic inferences when different substitution models are used», *Molecular Phylogenetics and Evolution*, vol. 31, nº. 3, pp. 865–873, jun. 2004.

[23] D. Posada y T. R. Buckley, «Model Selection and Model Averaging in Phylogenetics: Advantages of Akaike Information Criterion and Bayesian Approaches Over Likelihood Ratio Tests», *Syst Biol*, vol. 53, nº. 5, pp. 793–808, ene. 2004.

[24] M. Steel, «The Maximum Likelihood Point for a Phylogenetic Tree is not Unique», *Systematic Biology*, vol. 43, nº. 4, p. 560, dic. 1994.

[25] S. Roch, «A Short Proof that Phylogenetic Tree Reconstruction by Maximum Likelihood Is Hard», *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, vol. 3, nº. 1, p. 92–, ene. 2006.

[26] B. Chor y T. Tuller, «Maximum Likelihood of Evolutionary Trees Is Hard», in *Research in Computational Molecular Biology*, S. Miyano, J. Mesirov, S. Kasif, S. Istrail, P. A. Pevzner, y M. Waterman, Eds. Springer Berlin Heidelberg, 2005, pp. 296–310.

[27] D. Posada y K. A. Crandall, «Selecting the Best-Fit Model of Nucleotide Substitution», *Syst Biol*, vol. 50, nº. 4, pp. 580–601, ene. 2001.

[28] F. Abascal, R. Zardoya, y D. Posada, «ProtTest: selection of best-fit models of protein evolution», *Bioinformatics*, vol. 21, n^{o}. 9, pp. 2104–2105, ene. 2005.

[29] J. P. Huelsenbeck, P. Joyce, C. Lakner, y F. Ronquist, «Bayesian analysis of amino acid substitution models», *Phil. Trans. R. Soc. B*, vol. 363, n^{o}. 1512, pp. 3941–3953, dic. 2008.

[30] F. Abascal, D. Posada, y R. Zardoya, «MtArt: A New Model of Amino Acid Replacement for Arthropoda», *Mol Biol Evol*, vol. 24, n^{o}. 1, pp. 1–5, ene. 2007.

[31] G. Talavera y J. Castresana, «Improvement of Phylogenies after Removing Divergent and Ambiguously Aligned Blocks from Protein Sequence Alignments», *Syst Biol*, vol. 56, n^{o}. 4, pp. 564–577, ene. 2007.