Proyecto Fin de Carrera

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

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

Índice general

Índice de figuras

Índice de tablas

A. Diagrama de Gantt

En la figura A.1 se puede observar el diagrama de Gantt correspondiente al proyecto. Esta herramienta permite observar gráficamente cuándo se han iniciado y finalizado las diversas tareas de un proyecto.

Se puede observar que la fase de documentación se ha llevado a cabo durante todo el proyecto, pues como ya se ha comentado a lo largo de la memoria, ha sido necesaria una formación inicial y una continua lectura de información. También cabe destacar que los primeros meses sólo se dedicaron a esta fase por estar terminando los últimos créditos de la titulación y por la necesidad de tener que asimilar conceptos totalmente nuevos.

Puede chocar también que se empezó a trabajar antes en la fase de transcripción y traducción que en la de preprocesamiento de secuencias. El motivo es que la primera era más adecuada para acostumbrarse al entorno de programación y era viable hacer una primera versión de los métodos sin necesidad de esperar a los resultados del preprocesamiento de secuencias. Una vez terminada esta fase se modificaron los métodos de transcripción y traducción con la información obtenida. Esto quiere decir que no todo el tiempo que se refleja en el diagrama se dedicó a esta fase, sino que en medio de ella tuvo lugar la fase de preprocesamiento de secuencias.

En cuanto a la fase de construcción y comparación de árboles de filogenias, se puede observar que aunque se ha comentado que esta fase ha llevado mucho trabajo, su duración es menor que la del resto. Esto ocurre porque este diagrama solo refleja la duración y no el tiempo dedicado y no queda reflejado que durante esta fase se dedicaron considerablemente más horas diarias que al resto. Por lo que sí se puede decir que esta fase ha sido a la que más tiempo se ha dedicado aunque haya estado más concentrado.

Por último comentar que se empezó a redactar la memoria paralelamente al resto de trabajo para que no se acumulara todo el trabajo al final.

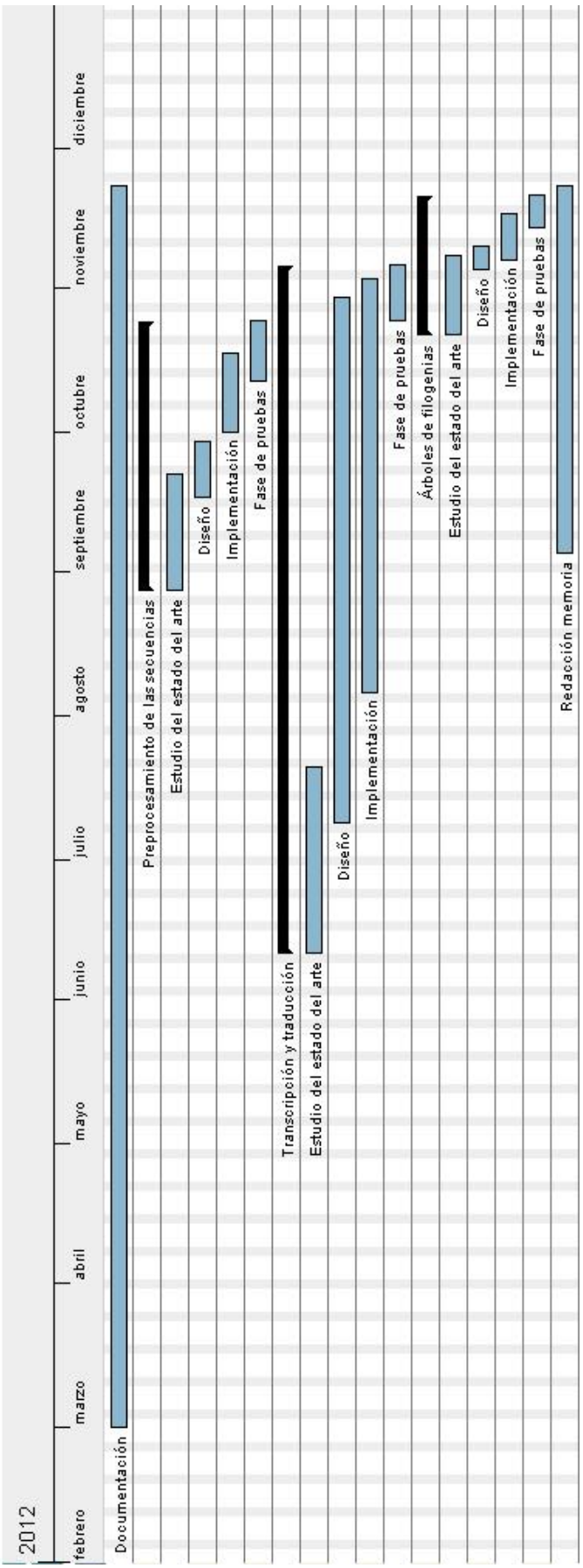

Figura A.1: Diagrama de Gantt

B. Fundamentos biológicos

En este anexo se van a detallar los términos y conceptos de índole biológica que se han considerado necesarios para la completa comprensión del presente proyecto. Antes de continuar me gustaría agradecer a uno de mis directores, Jorge Álvarez, su consentimiento para usar la memoria de su proyecto fin de carrera [12] como inspiración y base para la realización de este capítulo.

B.1. Contexto biológico

Este apartado se ha dedicado a la explicación del proyecto desde el punto de vista biológico para así poder tener una visión global del contexto en el que se ha desarrollado.

A lo largo del proyecto se trabaja con ADN mitocondrial humano (ADNmt a partir de ahora), por lo que es importante que el lector entienda sus particularidades. Como sugiere su nombre, está almacenado dentro de las mitocondrias (los orgánulos encargados de generar energía en las células). Cabe destacar que es independiente del ADN celular, lo que ha dado pie a muchos estudios sobre su origen. Pero lo que hace que sea tan importante en los estudios evolutivos de individuos de una o varias especies es su alta tasa de mutación, hasta 10 veces mayor que la del ADN celular.

Actualmente se sabe que el ADNmt lo forman entre 16.557 y 16.576 nucleótidos y contiene 37 genes, 2 de los cuales codifican ARN ribosómico (ARNr), 22 codifican ARN de transferencia (ARNt) y 13 codifican proteínas. El estudio de estas proteínas va a tener gran importancia a lo largo del proyecto. El ADNmt tiene una estructura circular (a diferencia de la estructura de doble hélice del ADN celular) que se mantiene gracias a una región de control, también denominada bucle D. Esta zona tiene una tasa de mutación muy elevada, con dos regiones hipervariables, la HVR_1 y la HVR_2, que pueden determinar el haplogrupo del organismo [19].

Un haplogrupo está formado por un conjunto de haplotipos con características estadísticas comunes. Los haplogrupos permiten clasificar a los individuos en familias en función de sus características comunes. Gracias a los haplogrupos se puede determinar el origen del linaje de una especie, lo que los hace especialmente importantes en la filogenética. Gracias a esto se ha podido determinar el ancestro matrilineal común más reciente del ser humano, que es conocido como “Eva mitocondrial” y según los últimos estudios se remonta a hace unos 200.000 años [20].

B.2. Introducción a la bioinformática

La bioinformática es un área de investigación multidisciplinar que ha surgido recientemente a causa de la complejidad creciente en muchos ámbitos de la biología. La resolución manual de muchos problemas se consideró

inviable, por lo que se introdujo el uso de herramientas y metodologías propias de la informática con el fin de resolver dichos problemas biológicos [1].

Este proyecto trata sobre la construcción de filogenias, una rama del área de análisis de secuencias. A esta área de estudio también pertenece el alineamiento de secuencias que se detalla en el punto 3.2 de la memoria y que es indispensable para la construcción de filogenias.

B.2.1. Filogenética

La filogenética es una rama de la biología que clasifica conjuntos de organismos en función de su relación evolutiva. Está muy relacionada con la cladística y muchas veces se usan como sinónimos por compartir unos mismos objetivos.

La cladística suele organizar las especies en varios grupos, entre los que destacan:

- Grupo monofilético o clado: Está formado por un organismo y todos sus descendientes, es decir, todos los organismos tienen una raíz común. En el caso de estudio de este proyecto para la especie humana la raíz es “Eva mitocondrial”.
- Grupo parafilético: Es similar al grupo monofilético pero se excluyen los descendientes que han sufrido cambios significativos.
- Grupo polifilético: Es un conjunto de grupos monofiléticos sin relación entre ellos.

Estos grupos se suelen representar mediante unos árboles llamados árboles filogenéticos, que reflejan las relaciones de parentesco entre los organismos o las especies. Estos árboles suelen ser binarios, pero hay casos en los que esto no se cumple.

B.2.2. Modelos evolutivos

Los modelos evolutivos o modelos de sustitución son modelos matemáticos que utilizan cadenas de Markov para definir la probabilidad de cambio entre nucleótidos en el caso de secuencias de ADN o ARN, o entre aminoácidos en el caso de proteínas. Estos modelos tienen en cuenta la reversibilidad del cambio a lo largo del tiempo entre otros factores. Intentan representar la evolución de las especies a lo largo del tiempo.

Estos modelos están regidos por un conjunto de parámetros, algunos fijados previamente y otros que se ajustan al evaluar las secuencias de entrada. Hay tres formas de prefijar parámetros, representadas con el signo “+” seguido de un identificador:

- “+I”: Considera que un conjunto de aminoácidos o nucleótidos son invariables.
- “+G”: Prefija la probabilidad de cambio de un nucleótido o aminoácido a

un conjunto de nucleótidos o aminoácidos.

- “+F”: Solo para aminoácidos. Se define una base de probabilidades para el cambio entre aminoácidos.

Se puede realizar cualquier combinación de los tres criterios. Cabe destacar que todos los modelos tienen en cuenta la distribución inicial de los nucleótidos o aminoácidos, ya sea indicada por el usuario o precalculada según las secuencias de entrada.

B.2.3. Selección de modelos

Hay una gran cantidad de modelos evolutivos y la elección del modelo apropiado es muy importante ya que una mala elección puede provocar cambios en el árbol filogénico de salida, disminuyendo su calidad pudiendo degradarlo hasta el punto de resultar incorrecto al representar relaciones evolutivas inexistentes [21–23].

Aunque el objetivo de los modelos evolutivos es explicar la relación real entre un conjunto de individuos, este es un problema de gran complejidad, lo que sumado al desconocimiento del funcionamiento exacto de la evolución hace que ni el mejor de los modelos represente fielmente la realidad [24]. Por este motivo se busca siempre el árbol que mejor se aproxime a la realidad aunque no la represente exactamente.

Para valorar la efectividad de los modelos se crean conjuntos de secuencias de las que se conoce su relación evolutiva “real”, así se pueden evaluar los resultados obtenidos con los de un modelo evolutivo conocido.

Hay varios métodos que evalúan la calidad de los modelos evolutivos, entre ellos el más usado es el de máxima verosimilitud. Este método utiliza los resultados observables para deducir los datos de entrada [24]. Estima la distribución de probabilidad del espacio de árboles asignando una probabilidad a las distintas mutaciones que han podido ocurrir en la secuencia. Su coste de ejecución es muy elevado, se sospecha que puede ser un problema NP-completo, aunque hasta ahora solo se haya demostrado que sea NP-duro [25], [26]. A pesar de ello es muy utilizado gracias a su gran flexibilidad.

Su mayor inconveniente es que valora mejor a los modelos más complejos por poseer estos un mayor número de parámetros. Para intentar eludir estos problemas se han desarrollado los criterios de información. Estos criterios se encargan de penalizar a los modelos más complejos para así generar una compensación entre la complejidad y la bondad de ajuste de cada modelo. Los criterios más usados son el de Akaike (AIC), el bayesiano (BIC), el de Akaike corregido (AICc) y la teoría de la decisión (DT).

C. Evaluación de ProtTest

C.1. Modelos evolutivos para proteínas

Existen varias aplicaciones diseñadas para la evaluación de modelos evolutivos, por lo que a la hora de escoger la adecuada para este trabajo se tiene que tener en cuenta si se está trabajando con ADN [27] o con proteínas [21].

Cuando se trabaja con ADN el modelo tiene que tener en cuenta las posibles mutaciones entre los 4 distintos nucleótidos que lo componen, en cambio cuando se trabaja con proteínas se deben tener en cuenta las posibles mutaciones entre los 22 aminoácidos existentes.

C.2. Descripción de la aplicación

ProtTest [28] es una aplicación que permite evaluar distintos modelos evolutivos para un alineamiento de proteínas dado, y así poder elegir el modelo que más convenga.

En su última versión, la 3.0 [5], ProtTest permite evaluar 120 modelos mediante 4 métodos distintos: el criterio Akaike (AIC), el criterio Akaike corregido (AICc), el criterio Bayesiano (BIC) y la teoría de la decisión (DT).

C.3. Estudio realizado

Para realizar las pruebas se ha utilizado siempre el mismo ordenador (*Intel® Core™2 Duo* T7100 a 1.8GHz y 2GB de memoria RAM) con una carga de trabajo similar. En cuanto a los datos, se han elegido 4 proteínas del ADNmt, la de menor longitud (ATP8), compuesta por 68 aminoácidos; la de mayor longitud (ND5), compuesta por 603 aminoácidos; y dos de longitud intermedia, la CO3, compuesta por 261 aminoácidos y la ND4, compuesta por 459 aminoácidos. En cuanto al número de secuencias de cada alineamiento, se han hecho las pruebas con 3, 5, 10, 20, 30 y 50 secuencias.

Se ha creído conveniente mostrar una tabla de resultados para cada proteína para así poder ver cómo evoluciona el tiempo en función del número de secuencias. Se han recopilado los resultados en las tablas C.1 a C.4. Se puede observar que el modelo que más veces sale elegido como el mejor es el MtMam, esto se corresponde con el resultado de los estudios que dan a este modelo como el mejor para el estudio de las proteínas en el caso del ADNmt de vertebrados [29], [30]. La figura C.1 es una representación gráfica de los resultados obtenidos.

Las secuencias usadas en las pruebas se eligieron aleatoriamente para cada conjunto de secuencias. Los identificadores de los ficheros usados (disponibles en GenBank [8]) se pueden consultar en la tabla C.5.

Nº Secuencias	Tiempo	AIC	BIC	AICc	DT
3	00m 12s	JTT+F	MtMam	MtMam	JTT+F
5	00m 18s	HIVb+F	MtMam	MtMam	MtMam MtMam+I
10	00m 29s	MtMam+F	MtMam	MtMam	MtMam MtMam+I
20	00m 49s	HIVw+F	MtMam	MtMam	MtMam MtMam+I
30	01m 17s	HIVb+F	MtMam	HIVb+F	MtMam MtMam+I
50	02m 08s	HIVw+F	MtMam	MtMam	MtMam MtMam+I

Tabla C.1: Alineamientos con el coste temporal y los modelos seleccionados por los distintos criterios para la proteína ATP8 (68 aminoácidos).

Nº Secuencias	Tiempo	AIC	BIC	AICc	DT
3	00m 14s	MtREV	MtREV	MtREV	MtREV MtREV+I
5	00m 19s	MtREV	MtREV	MtREV	JTT+F JTT+I+F
10	00m 32s	MtMam	MtMam	MtMam	MtMam MtMam+I
20	00m 59s	MtMam	MtMam	MtMam	MtMam MtMam+I
30	01m 37s	MtMam	MtMam	MtMam	MtMam MtMam+I
50	02m 18s	MtMam	MtMam	MtMam	MtMam MtMam+I

Tabla C.2: Alineamientos con el coste temporal y los modelos seleccionados por los distintos criterios para la proteína CO3 (261 aminoácidos).

Nº Secuencias	Tiempo	AIC	BIC	AICc	DT
3	00m 13s	MtMam	MtMam	MtMam	JTT
5	00m 20s	MtMam	MtMam	MtMam	MtMam MtMam+I
10	00m 34s	MtMam	MtMam	MtMam	MtMam MtMam+I
20	01m 05s	MtMam	MtMam	MtMam	MtMam MtMam+I
30	01m 41s	MtMam	MtMam	MtMam	MtMam MtMam+I
50	02m 35s	MtMam	MtMam	MtMam	MtMam MtMam+I

Tabla C.3: Alineamientos con el coste temporal y los modelos seleccionados por los distintos criterios para la proteína ND4 (459 aminoácidos).

Nº Secuencias	Tiempo	AIC	BIC	AICc	DT
3	00m 14s	MtMam	MtMam	MtMam	MtMam MtMam+I
5	00m 21s	MtMam	MtMam	MtMam	MtMam MtMam+I
10	00m 37s	MtMam	MtMam	MtMam	MtMam MtMam+I
20	01m 11s	MtREV	MtREV	MtREV	MtREV MtREV+I
30	01m 50s	MtMam	MtMam	MtMam	MtMam MtMam+I
50	03m 06s	MtMam	MtMam	MtMam	MtMam MtMam+I

Tabla C.4: Alineamientos con el coste temporal y los modelos seleccionados por los distintos criterios para la proteína ND5 (603 aminoácidos).

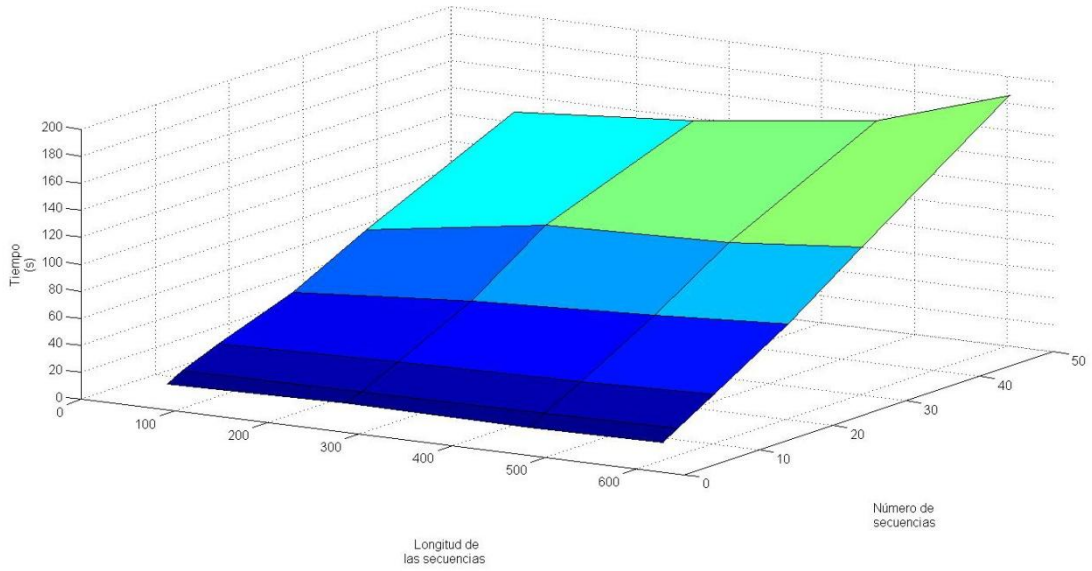

Figura C.1: Gráfica de estudio de la evolución del coste temporal de ProtTest según incrementa el número de secuencias y su longitud.

Nº secuencias	Identificadores					
3	FJ348161	JQ705965	JQ798137			
5	EF485042	EU092911	HQ593809	JQ044874	JF824915	
10	AY882389	EU092722	EU482369	AP010714	FJ940865	JQ702041
	JQ798098	JQ703313	JQ704327	JQ705740		
20	AP008624	EU007863	FJ467985	GU181350	JQ044975	JQ691414
	EF060338	EU092841	AP010988	HM852869	JQ324565	JQ703192
	EF177413	FJ384433	AY963586	JF260934	JQ324601	JQ704690
	JQ705590	JQ705925				
30	JQ705242	JQ324921	HM852799	HM346918	EU092679	DQ246811
	JQ705438	JQ324705	HM103358	FJ460531	EU092718	AP008275
	JQ702264	JN581661	HM041971	FJ348181	AY339503	FJ383196
	JQ702134	HQ384171	GU296546	EU545441	AY738976	FJ951486
	FJ348209	HM596652	GU123013	EU092803	AY714018	HQ012260
50	AY275529	DQ787109	JQ705614	JQ704105	JQ044978	JF292900
	AY714032	HQ287881	JQ705558	JQ703206	JQ044957	GU810011
	AY738981	FJ442938	JQ705288	JQ702932	JN202912	FJ467966
	AP008530	FJ467961	JQ705078	JQ702269	FJ383377	AP010734
	AP008589	JQ706005	JQ704593	JQ702063	HM050402	EU935452
	FJ951550	JQ798090	JQ704533	JQ702016	HQ661845	EF660951
	AP008714	JQ798027	JQ704303	JQ324782	HQ659686	EU092957
	HQ435319	JQ705859	GU122981	EF184599	GU122992	JX041635
	FJ383247	EF495214				

Tabla C.5: Ficheros utilizados para calcular el tiempo de ejecución de ProtTest

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, nº. 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, nº. 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, nº. 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, nº. 4, pp. 564–577, ene. 2007.